Table of Contents

List of Figures

Chapter 10

Chapter 11

Chapter 12

Andrea Scharnhorst
Data Archiving & Networked Services (DANS)

Richard P. Smiraglia
Institute for Knowledge Organization and Structure, Inc.

Chapter 1
The Need for Knowledge Organization
Introduction to the book *Linking Knowledge: Linked Open Data for Knowledge Organization*[†]

Abstract
This book is not restricted to semantic web (SW) technologies. An aspiration was to contribute to the awakening of a dialogue between information and documentation concerned with knowledge organization systems (KOSs), and branches in computer science with an emphasis on machines, algorithms and ontologies. The technological evolution of the last decades has not only fostered the emergence of ever more KOSs but also semantic web technologies. Both the actions of "making a KOS" and "applying existing KOSs" represent research. The design of an information layer for a knowledge domain and the design of a domain specific research process are intrinsically interwoven. We extended our intervention to KOS practices into education, by presenting a translation of existing standards and recommendations about linked open data (LOD) publishing for non-experts. The chapters describe the state of the art in providing KOSs as semantic artefacts; how the state of the art is applied in new fields; how the state of the art is pushed towards new technological solutions by being confronted with new applications; how best practices need to be tailored towards specific solutions; and what challenges occur when merging new and old ways of expressing KOSs. The linked data (LD) ecosystem represents a source of knowledge generation, acquisition, production and dissemination. The underlying discourse shows historical vision alongside the promise of linking knowledge for interaction. The already maturing ecosystems of the SW are interlocking information institutions clearly devoted to the expansion of human experience through the growth of knowledge interaction.

0.0 From the very beginning …

The title of this book is not arbitrary. While the monograph was produced in the context of a project about linked data (the Digging into the Knowledge Graph (DIKG) project[1]) the content of this book is not restricted to semantic web (SW) technologies. Instead, through the chapters, problems are addressed that prove to be almost eternal when it comes to the organization of knowledge.

From the very beginning the project—as now documented in this book's chapters—aspired to contribute to the awakening of a seemingly forlorn dialogue between those branches in the sciences of information and documentation that used to reflect about classifications, or knowledge organization systems (KOSs), and those branches in computer science that equally address the use of KOSs, but with a strong emphasis on machines, algorithms and ontologies.

[†] This work was funded by the European Commission T-AP Grant Agreement ID: 613167. We would like to acknowledge the collaboration with the VU Amsterdam Knowledge representation and reasoning group and Triply, an Amsterdam Startup. Part of this work started during visits of Richard Smiraglia at the Virtual Knowledge Studio (VKS-KNAW) and with the eHumanities group (KNAW), continued by his KNAW Visiting professorship grant and his fellowship with Data Archiving and Networked Services (DANS-KNAW).

In this introduction, we detail our motivation for why this discourse needs to be awakened and how best to do this. In doing this we rely on insights from the fields of knowledge organization (KO) and science and technology studies (STS). Most of our use cases stem from the digital humanities (DH). We also emphasise the role of visualization as means to support translation across or between different knowledge domains as part of the essential knowledge exchange.

1.0 Organizing knowledge, linking knowledge

Ordering of knowledge is as old an activity of the human mind as reflection is. One thinks here of the systems of Aristotele, Leibniz or Linnaeus (cf. Furner 2020; Kedrov 1975). Ordering systems are deeply embedded in philosophical systems and appear in all domains of human knowledge. We classify and order knowledge as we grow up in our individual ontogenesis, and we classify and order knowledge as mankind to enable orientation, navigation through growing masses of information. The ordering of knowledge is a precondition to allow for the abundance of new ideas by endless recombinations, alterations of marked cornerstones of human insights, next to flipping and breaking them and in this way pushing knowledge to a next level. In short, ordering knowledge is a prerequisite to linking knowledge.

The importance of organizing knowledge is based on its role in the functioning of the human brain. When we speak about organizing knowledge, we often refer to instruments, tools and principles that enable communication and the spread of knowledge beyond the individual. Systems (called here Knowledge Organization Systems or KOSs) are created to coordinate cognitive, communicative and social activities at the level of parts of society. Knowledge Organization is also the name of a scientific field, which grew from roots in philosophy, and emerged (arguably as a subfield in the information sciences) in the last century.

2.0 The oxymoron of abundant yet invisible KOSs

We are neither the first nor the last in struggling with ordering knowledge and with the question of how to best describe, reflect, teach and coordinate the practices of organizing knowledge. Each explosion of information almost naturally comes with visions and visioneers to make the best of large bursts of information. The emergence of the dedicated scientific domain Knowledge Organization (KO) is indispensably linked to the growth of knowledge by the last industrial revolution (end of the 19th, beginning of the 20th century), and the specific role of libraries in organizing it. In fact, classification theory (cf. Ranagathan 1973), as we know it today, had its origin in this time. New waves of technological evolution bring with them new challenges in ordering knowledge. For the information age, obviously the emergence of computers, and the emergence of networks of computers (the internet) are key.

The rise of automatization, based on computers as machines, goes together with formalization and abstractions, and KOSs of all kinds of flavour. There is no machine operation possible without strict formalization. Any data model (be it expressed in a database or a knowledge graph) relies on categorization and the definition of relations between them. Still, the more we depend on KOSs, the more they influence our lives, the less we seem to be aware of them as though the old dream of seamlessly supporting and guiding an envisioned user to information needs has become too perfectly realised. So, we experience an oxymoron: KOSs are everywhere and at the same time they seem to become more and more invisible.

To give an example from the world of scholarly communication: we have access to all libraries in the world almost, but in the design of the online public access catalog (OPAC), systematic catalog elements are often made invisible. For many purposes, their natural ordering function might be irrelevant. But, this is not always true. In particular, if one needs some context, a first orientation, an idea about which body of knowledge to consult, for those cases making existing classifications visible might be useful. But, there is more to the vanished ordering principles. Searching collections on-line is very different from visiting a physical place where a collection is held. In the physical world, the space alone conveys information. But, in front of a text-box, we often have no clue how large a collection is, how established the collection holding institution, or how this collection came about? There have been attempts to use the power of visualization to counter for that (Whitlaw 2015; Mutschke et al. 2017).

But, such invisibility of underlying KOSs does not only concern access to public or scholarly knowledge. A lot of our administrative, societal processes rely on data models and processing around them. Think about your nationality and the rights and obligations following from them. Here, categorization can deeply affect your existence. The problem augments with the mastery of artificial intelligence. We can process unprecedentedly large amounts of information, but we are concerned that we cannot even identify the algorithm principles behind them anymore, not to mention having an overview of consequences a certain design of data models and algorithms might have. To summarize, we are partly governed by hidden KOSs. This is an unsettling thought, when we remember that KOSs are always also a mirror of what is at stake in a certain society and culture at a certain time. (van den Heuvel and Zamborlini 2021, chapter 6 in this book). They are means of executing power by coordinating what to think, how to think and how to arrange access to the products of the mind (Bowker and Star 1999). They are not neutral. How can you find a position towards them, if you don't know them? To make KOSs visible and to revoke a discourse on the power of KOSs is the one aspiration of this book.

3.0 Orientation in the expanding KOS universe

Let's be clear: there is neither a way nor a reason to stop the information avalanche. Information is at the bottom of the knowledge-based economy (Leydesdorff 2006). As part of this avalanche, we as human beings will continue to organise knowledge as part of engaging in new practices of knowledge production. As KOSs are tailored towards specific practice, the more our societies differentiate and specialise internally, the more we will see different KOSs emerge naturally. So, we have to deal with this expanding KOS universe.

The rise of a great variety of KOSs can be compared to the sudden rise of innovations in certain ages, which in turn has been compared with the explosion of variants in biological evolution (Ziman 2003). Emerging new variants of KOSs can be described as mutations in such an evolutionary systems theoretical framework. But, from evolutionary theory we also know about a related thread, which has been called "mutation catastrophe" or has been referred to as "evolution window" (Rechenberg 1994). With too many variants existing, a comparison among them followed by a selection almost becomes impossible, all variants survive somehow and so evolution stops. Currently, it seems that we tend to build our own KOSs without even being aware of possibly useful already existing KOSs built by our neighbors. So, we seem to have given up on comparing and selecting also. Are we stuck in a "mutation catastrophe" of KOSs? That would be opposite to visions of Tim

Berners-Lee and others to truly connect and integrate knowledge (Berners-Le, Hendler and Lassila. 2001). So, how can we find a good balance between creating new KOSs as part of new practices and (re-)using, linking back to existing KOSs?

This question materializes not only on the level of KOSs. As the science system grows and new knowledge domains emerge at the boundaries of existing ones, there appears the need to foster, govern and organize interdisciplinarity. There is another scientific domain that formed its own epistemic framework, in particular after World War II, and this is known as Science and Technologies Studies (in short STS), a field at the crossroads of philosophy, history, sociology of science (here meant as the whole of academia) and innovation studies (Felt et al. 2017; Scharnhorst Börner and Besselaar 2012). A KOS in a knowledge domain (be it in academia or in any other societal sector) can be seen as a formalised expression of the epistemic framework, the conceptual reference, in which a domain operates. Thus, in a domain KOSs act as multipliers, coordinators of a certain worldview. Questions of how best to exchange knowledge between different domains cumulate in questions of how to bridge between different KOSs organizing those domains. The latter is part of the research in the field of KO, a field that also reflects on KOS design and that developed generic KOSs, which could govern this knowledge exchange (Smiraglia 2014b, Gnoli 2020). But, there are not many researchers combining methods from STS and KO (Smiraglia 2014a). For this book we selected chapters that combine presentation of knowledge ordering practices with a reflexive layer on those same practices.

4.0 Linking knowledge by Linked Open Data

The technological evolution of the last decades has not only fostered the emergence of ever more KOSs, it provided at the same time means to govern this explosion, namely SW technologies. As stated in Wikipedia (2020) "The Semantic Web is an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C)" with the ultimate goal of making information, data and knowledge machine readable. The visioneers, designers and engineers behind those SW-based technologies can be directly compared to pioneers in KO such as Paul Otlet and Henri La Fontaine who aimed at a universal bibliography containing all knowledge of the world and designed a system—a KOS, the Universal Decimal Classification—to navigate it (Rayward 1990; Wright 2014).

Equally, we see KOS at the heart of the Linked (Open) Data paradigm. According to Berners-Lee's (2006) 5-star system of data, linking documents on the web is only the first step toward deep linking of human knowledge (cf. Hyland et al. 2013). To enable access to the knowledge in documents is the essence behind "indexing" at the level of data. This takes the dream of Otlet and La Fontaine to a next level. Not only the works in which knowledge is documented (from music sheets to whole books, from images to technical drawings) are classified, but the "elementary units" in them such as concepts or phenomena should become (machine) referenceable. To be able to really weave those emerging network/graph representations of knowledge into one fabric of knowledge, it is necessary to also formalize the links between different KOSs (cf. Yoose and Perkins 2013). Among the ingredients that are part of this heroic effort we find: standards for expressing nodes and links in the knowledge graph and good practices (Hyland et al. 2014); an overview of existing KOSs in the LOD Cloud (cf. Linked Open Data Cloud; Vandenbussche et al. 2017); and requirements between the translation of KOSs (independently where they are published right now) into "semantic artefacts" (Le Franc et al. 2020) in a machine-readable form

adhering to FAIR principles (FAIR stands for Findable, Accessible, Interoperable and Reusable [see Wilkinson et al. 2016]).

There is an abundance of literature about how to realise the LOD paradigm and which role KOSs play in this process (cf. Heath and Bizer 2011; Antoniou et al. 2012; Hyvönen 2012). Still, there seems to be a broken line of communication between specialists from different domains who reflect about the more generic aspects of KOSification. Those who reflect about classification theory and those who reflect about knowledge representation and reasoning are not always aware of each other's traditions, epistemic frames of reference and available solutions. As a consequence, there is quite some confusion about, for instance, what concepts are and how best to express them; and if natural language or classification languages and controlled vocabulary are better suited for expressing concepts (see Slavic, Siebes and Scharnhorst 2021; chapter 5 in this book). Some experts prefer to discuss these issues in the realm of library classification, others in the realm of computer science and engineering. In both camps there is much talk about the role of the user, however the role of humans in the engineering designs and of human use of machine readable KOSs is less clear. In general, the interplay between machines and humans around KOSs remains somewhat foggy. What can machines do and where are humans indispensable? If such questions are not properly sorted out among the information specialists and professionals, they create even more uncertainty among those applying the new knowledge ordering machinery in their daily research practice. To strengthen the link between different bodies of knowledge about KO and current KO practices in the realm of the SW is another aspiration of this book.

5.0 Bridging by reflecting

The main approach of the Digging into the Knowledge Graph (DiKG) project was reflecting by engaging in new practices. This was supported by bringing together experts with the various backgrounds referred to above. Among them were Wouter Beek, one of the designers of the *LOD Laundromat*, a "cleaned, indexed version of the Linked Open Data Cloud" (Beek et al. 2014). A further member of the famous knowledge representation (KR) group at the Vrije Universiteit Amsterdam[2], to which Wouter Beek belonged, Ronald Siebes worked during the project at the KNAW-DANS collaborating partner of the DiKG consortium. Richard Smiraglia, Rick Szostak, Daniel Martínez-Ávila, and Aida Slavic represented the KO experts. Andrea Scharnhorst contributed from the side of STS, DH and complexity theory. Together, we focussed on two areas:

- The KO discourse inside of the science of information and applied librarianship; and,
- The application of SW practices in parts of the digital humanities.

We started the call for this book with the statement (Smiraglia and Scharnhorst 2019):

> The growth and population of the Semantic Web, especially the Linked Open Data (LOD) Cloud, has brought to the fore the challenges of ordering knowledge for data-mining on an unprecedented scale. The LOD Cloud is structured from billions of elements of knowledge and pointers to knowledge organization systems (KOSs) such as ontologies, taxonomies, typologies, thesauri, etc. The variant and heterogeneous knowledge areas that comprise the social sciences and humanities (SSH), including cultural heritage applications, are bringing multi-dimensional richness to the LOD Cloud. Each such application arrives with its own challenges regarding KOSs in the Cloud.

We also solicited contributions with a specific nature:

> Working from the international "Digging Into the Knowledge Graph" LOD-KOS project (http://di4kg.org/)—we aim to bring together research papers from some of the world's leading experts in the application of multi-dimensional KOSs to the LOD cloud.

Next to the multidimensionality of KOSs we called for, the final contributions highlight the relevance of practices. Analysis of practices offers a way to identify barriers in knowledge exchange. The contributions o together represent a 'trading zone' in its own rights. In STS, Gallison's concept of a trading zone describes an intellectual, social-communicative place where different epistemic perspectives meet, exchange and through this exchange lay the ground for the emergence of new ideas, innovations and possibly new fields (Galison 1997). Indispensable for such a process is an openness towards talking about one's own implicit epistemic norms and values in a way in which others can relate to it. This goes together with awareness about boundary objects: concepts that change meaning when used in different knowledge domains, and which still can serve as a carrier for mutual understanding (Star and Griesemer 1989). The very definition of a KOS is a prime example for a boundary object (see Zeng and Mayr 2021; Chapter 3 in this book).

We designed this book as an intervention to current practices, advocating the need of a specific reflection layer and the acknowledgment of temporality in our endeavours. Too often, in particular in the early stage of the adoption of a new technology or method, explorations are presented as solutions ready to be re-applied. We wanted to counter this understandable but partly misleading attitude by unfolding that both the actions of "making a KOS" as an instrument for better research and "applying existing KOSs" represent research in its own right. It might look like an accompanying, supportive information management task, but the design of an information layer for a knowledge domain and the design of a domain specific research process are intrinsically interwoven. Still, at the same time, both processes require slightly different skill sets. We extended our intervention to KOS practices into education, by presenting a translation of existing standards and recommendations about LOD publishing for non-experts (Siebes et al. 2021; Chapter 12).

The book contains chapters that describe the state of the art in providing KOSs as semantic artefacts or semantification (Zeng and Mayr 2021, Chapter 3; Siebes et al. 2021, Chapter 12); how the state of the art is applied in new fields (van den Heuvel and Zamborlini 2021, Chapter 6; Smiraglia and Szostak 2021, Chapter 7; Patuelli 2021, Chapter 8; Smiraglia, Young and van Berchum 2021, Chapter 9); how the state of the art is pushed towards new technological solutions by being confronted with new applications (Oelen Stocker and Auer 2021, Chapter 10); how best practices need to be tailored towards specific solutions (Slavic, Siebes and Scharnhorst 2021, Chapter 5); and what challenges occur when merging new and old ways of expressing KOSs (van den Heuvel and Smiraglia 2021, Chapter 11; Szostak and Renwick 2021, Chapter 2).

To summarise, in this book the contributors address the problem of linking knowledge in two different ways:

- To address fundamental issues of KO: such as presentation of concepts, roles of different KOSs (thesauri, analytico-synthetic classifications) and their representation as Linked Open Data
- To make the role of KOSs and the practices behind the design of KOSs visible in areas of scholarly communication and certain fields of humanities research.

The reader might ask, why did we choose the rather traditional form of an edited collection to document the results of this research? The main reason is the relative stability of

written documentation. Of course, the DiKG project (as well as the other projects that contributed to this book) also delivered other kinds of output: LOD, architectural designs, experimental services, and a large amount of Resource Description Framework (RDF)-modelled content. But, SW technologies, as mentioned earlier, represent also a fast moving research front with ever new approaches and corresponding tooling. As often discussed by the community itself, sustainability of published resources and solutions is an issue (cf. Benjamins et al. 2002). This can best be illustrated with the case of the LOD Laundromat. At the beginning of the DiKG project, in 2017, the LOD Laundromat (Beek et al. 2014) was still in operation. Back than one could find a whole suite of tools provided to search the crawled LOD cloud (see https://web.archive.org/web/20190103031340/http://lodlaundromat.org/ for the landing page) Currently, the website http://lod-a-lot.lod.labs.vu.nl serves as a new experimental space for researchers from the VU Amsterdam group, but the original tools are no longer available. It is important to note that the content of the LOD Laundromat was archived (albeit as large dump of static linked data) with a long-term archive (DANS-EASY) (Beek et al. 2017). Still, the cleaned LOD version is no longer available as a service. The LOD Laundromat is not the only web-based resource that has experienced such life cycle changes. Also, the CMME web-resource that we used in the DiKG project changed in functionality during the project's life-time. The chapter by Slavic, Siebes and Scharnhorst (2021; Chapter 5) shows in detail the care a service provider has to take when introducing new forms of a service. In many cases, researchers or research projects are the owners of linked data (LD) solutions, and often they are described as workshop, lab or experimental material and not designed to operate as sustained production services. In order to preserve the efforts that go into creating KOSs and KOSs as LD a close collaboration with institutions that guarantee stability is needed. As illustrated in the chapter by Smiraglia, Young and van Berchum (2021; Chapter 9), publication of LOD can often also take the form of submitting content to services maintained by others. In this line of reasoning, even archives have a role, as LD in the form of RDF is no more than a very detailed "index" or description. Even if the machinery to execute operations with this index is no longer functional, it still makes sense to document and preserve the efforts behind the RDF modelling and indexing of content with it. But, of course, documenting is wise also in competition with actual executing research, and no detailed research data management strategy will ever be able to solve this dilemma.

This brings us to a last disclaimer. While we argue in favour of linking new and old cultures of documentation and KO, we put special emphasis on making time and space for reflection in explorative practices. We see reflection in processes of cross-domain communication as an indispensable means to achieve (better) results. However, we are very well aware that there always remains a tension, the tension between making and analysing; between pushing technology forward and applying existing technology; between being precise, well rooted in your own domain and reaching out to other domains. We started this introduction with an emphasis on the key role KO in general and KOSs as specific instruments. So, the bar is high. But, even here, one needs to find a balance between the efforts requested and the benefits expected. With KOSs the situation is no different from more general discourses around data. For instance, in designing new IT services for data search and sharing one needs to balance the costs with the expected benefits (cf., Gregory 2021). While, there is endless potential in making all KOSs semantic by curating and observing them, we might not be able to achieve this. No, this is not a call to give up; but a call to be

aware of it and to honestly present achievements together with limitations; to be explicit about what should, could, and what most probably will be realised.

6.0 Content of the book

6.1 Overview

The book is organised in five sections: ***Background, Foundations, Applications, New Endeavours, and Education***. In the overview, we list all chapters in those sections, and proceed further to summarise what those chapters bring to the book in the light of the goals of this book as discussed above.

Background

Andrea Scharnhorst and Richard P. Smiraglia (Chapter 1)

"The Need for Knowledge Organization: Introduction to the book *Linking Knowledge: Linked Open Data for Knowledge Organization*

Rick Szostak, Richard Smiraglia, Andrea Scharnhorst, Ronald Siebes, Aida Slavic, Daniel Martínez-Ávila and Tobias Renwick (Chapter 2)

"Classifications as Linked Open Data: Challenges and Opportunities"

Knowledge Organization and Linked Data - Foundations

Philipp Mayr and Marcia Zeng (Chapter 3)

"Knowledge Organization Systems in the Semantic Web: A Multidimensional Review"

Tobias Renwick and Rick Szostak (Chapter 4)

"A Thesaural Interface for the Basic Concepts Classification"

Aida Slavic, Ronald Siebes and Andrea Scharnhorst (Chapter 5)

"Publishing a Knowledge Organization System as Linked Data: the case of the Universal Decimal Classification"

Application of Linked Data in the Digital Humanities

Charles van den Heuvel and Veruska Zamborlini (Chapter 6)

"Modeling and Visualizing Storylines of Historical Interactions: Kubler's *Shape of Time* and Rembrandt's *Night Watch*"

Richard P. Smiraglia and Rick Szostak (Chapter 7)

"Identifying and Classifying the Phenomena of Music"

M. Cristina Patuelli (Chapter 8)

"Graphing out Communities and Cultures in the Archives: Methods and Tools"

Richard P. Smiraglia, J. Bradford Young and Marnix van Berchum (Chapter 9)

"Digging into the Mensural Music Knowledge Graph: Renaissance Polyphony meets Linked Open Data"

Knowledge Organization and Linked Data - New Endeavours

Allard Oelen, Mohamad Yaser Jaradeh, Sören Auer and Markus Stocker (Chapter 10)

"Organizing Scholarly Knowledge leveraging Crowdsourcing, Expert Curation and Automated Techniques"

Charles van den Heuvel and Richard P. Smiraglia (Chapter 11)

"Knowledge Spaces: Visualizing and Interacting with Dimensionality"

Knowledge Organization and Linked Data - Education

Ronald Siebes, Gerard Coen, Kathleen Gregory and Andrea Scharnhorst (Chapter 12)

"Publishing Linked Open Data: A Recipe"

6.2 Background

In the background section the reader finds next to this introduction, a reprint of the paper "Classifications as Linked Open Data: Challenges and Opportunities" (Szostak et. al 2020). This paper summarises the achievements of the DIKG project. It discusses in particular the challenges that emerge when classifications designated for the bibliographic domain, be they of newer or older provenance, are prepared to be interwoven into the LOD Cloud. The cases of the Basic Concept Classification (BCC) and the Universal Decimal Classification (UDC) will be discussed in further detail in chapters in the next section.

6.3 Foundations

The foundations section starts with the chapter "Knowledge Organization systems in the Semantic Web: A Multidimensional Review." Zeng and Mayr unfold how complex and to a large extent not yet fixed the terminology is when it comes to questions of "what is a KOS?" "how a KOS as a model of knowledge can be made machine readable" and "how KOSs can be used in machine readable statements." What makes this contribution so special is that it sheds light on the different actors involved in making, providing and using KOSs in the context of the SW. The authors do this by designing personas or proto-personas, an approach from experience design. One large group of those personas is the providers of LD services of KOS. Providers can operate country-wide or deliver just one individual vocabulary. Services providing KOSs as semantic artefacts (Le Franc et al. 2020) also include middleware for end-users or registries. The KOS service providers are just one part of the wider landscape of consumers and producers of KOSs. Similar to what has been found for users of data (Borgman et al. 2019), and in many other studies on knowledge production practices (cf. Wouters et al. 2013) roles are usually mixed in practice. So, one and the same organization, group or person can operate as different proto-personas depending on their actual activities. Next to the service providers Zeng and Mayr introduce the dataset producer using LOD principles, the vocabulary producer, research groups as end-user, website and tool developer. All of them can operate on various geographical levels and in or across different knowledge domains. The authors derive those archetypical personas from a rich empirical analysis of the field as it stands now. Two aspects here are striking: first how experimental the stage of semantic KOSs still is, how fluid, how much in development (this is, by the way, a thread through all the chapters in this book) and second the gap between makers of KOSs and makers of KOSs in LOD form.

The second chapter "A Thesaural Interface for the Basic Concepts Classification" (reprint of Renwick and Szostak 2020) discusses how to design an interface that can guide a human executing a classifying task (indexer or classifier) through the BCC controlled vocabulary. The paper departs from fundamental issues of language-based classifications and tries to bridge between keyword and subject search activities by the design of an interface. The final aspirations of this use case of the DIKG project are higher, namely to support user queries formulated in sentences.

The last foundations chapter, "Publishing a Knowledge Organization System as Linked Data: the case of the Universal Decimal Classification," documents the efforts of a KOS service provider to make one of the standard bibliographic KOSs available as LD. The Universal Decimal Classification (UDC) is an analytico-synthetic and faceted classification whose origins go back to the end of the 19th century. Paul Otlet and Henri La Fontaine started in 1896 an international project intended to cover all information sources published

in human history, in any form or language, anywhere in the world (Wright 2014). The UDC design as a synthetic indexing language and its use in practice over a long period of time has not only influenced further design improvements of the UDC (Smiraglia et al. 2013; Slavic and Davies 2017), but contributed to the large amount and variation of UDC codes and their combinations in bibliographic metadata (Scharnhorst et al. 2016). How, in general, KOS expressions are created by local practices when a specific KOS is applied is a complex process in itself (Tennis 2012). It is also beyond the control of KOS editors and publishers. But, the KOS service providers have to address the large user base of their KOSs, and this problem augments, if both KOS service and KOS use become part of the LOD cloud. For bibliographical metadata we can observe that Machine Readable Cataloging Records[3] increasingly become available as LOD. In the case of the UDC, to further be able to use its analytical power, it is necessary to build the UDC as a semantic artefact (Le Franc et al. 2020) in a way that preserves both the structure of the UDC and its provenance over time. Slavic, Siebes and Scharnhorst present a detailed discussion of those issues and unravel how those influence the final architectural design for a new LD/LOD publishing service of the UDC.

6.4 Applications

As indicated above, this book discusses not only generic issues of KO when it comes to LOD, but it also zooms into practices currently applied in the DH domain. In the applications section two areas are covered: KO and knowledge graph designs in the prestigious Golden Agents[4] project, and classification issues around works and practices in music and musicology.

Van den Heuvel and Zamborlini contirbute "Modeling and Visualizing Storylines of Historical Interactions: Kubler's *Shape of Time* and Rembrandt's *Night Watch*." They describe how discourses and controversies in art history come to new life when building a knowledge graph that enables weaving different historical sources into one information fabric. One aspiration of the Golden Agents project is to be able to understand what we would call today "creative industries" during the Dutch Golden Age (ca. 1581-1672) in their entirety, covering different sectors and their different products, and describing the role of different actors (producers and consumers) (cf. Idrissou et al. 2019). Biographies of makers, networks of their interactions and traces of objects in space and time will all come together. Naturally, harmonization of information about agents and objects and dealing with a large variety of KOS standards used in the different branches of all knowledge domains involved are at the heart of this project. But, in this chapter, the authors focus on how temporality of events and processes should be captured in a way that allows for different stories about the past to appear. Naively, to pinpoint everything to an external arrow of time seems to be the obvious solution. However, the real challenge lies in the selection and later standardised description of what to connect to which point in time in which way. Processes come with their own temporal signature, objects can be found in different manifestations, stories about both deliver additional information but again come with their own temporal provenance. The chapter details how a model emerges that allows describing this complexity in a way that machine-based information processing as well as linking to other sources becomes possible. Here the emphasis is on standardization and formalization. A better retrieval of information is a very tangible outcome of the project. Yet, and this is the real focus of this chapter, the model should also be flexible enough to still support the

hermeneutic, interpretative, explorative research practice that produces new, fresh insights beyond agreed standards. In discussing the decisions behind the eventual chosen model a trading zone for concepts, epistemics frames and language to describe them becomes visible.

The next paper, "Identifying and Classifying the Phenomena of Music" (reprinted from Smiraglia and Szostak 2020) continues the discourse around how to best (for certain purposes) represent artifacts from the past. Smiraglia and Szostak focus on the phenomena of music, and call for an extension of the usually documented features in music retrieval. This discussion is actually based on mimicking or envisioning future research behaviour of musicologists. As always in the history of documentation, documentation of resources for research goes hand in hand with existing (but not answerable) and newly envisioned research questions. New designs for KOSs should be best on an analysis of current research practices in a certain knowledge domain. But indexers base their indexing on the content of the work as well as on their imagination of a use and a user. The KO providers (see Zheng and Mayr 2021; Chapter 3) the designer of KOSs that only partly overlap with those later using the KOS are both faithful to observations and visionary. Smiraglia and Szostak use a generic classification, the BCC, to identify facets that might become candidates for further standardised documentation.

The next chapter, "Graphing out Communities and Cultures in the Archives: Methods and Tools," puts such considerations in action. Patuelli describes the project *Linked Jazz*[5], which explores the power of LOD technologies when applied to the history of jazz. Similar to the Van den Heuvel and Zamborlini (2021; Chapter 6), the quest is again for a web that connects entities, people, objects, facts and concepts in new and unprecedented ways, across disparate domains and beyond repository boundaries. More specifically, the chapter (and the project) uses an oral history approach departing from existing interviews with jazz musicians in certain repositories. Using the power of the Wikidata platform, a network of information emerges that allows different perspectives, such as the social ego network of an artist, or the location history of the emergence of certain genres and styles.

Pushing boundaries of existing KOSs or recombine existing KOSs into something new have been topics of KO through history. Equally, sustaining insights inscribed in KOSs has been achieved by institutions issuing authoritative resources, and by guarding workflows around those authoritative resources. "Digging into the Mensural Music Knowledge Graph: Renaissance Polyphony meets Linked Open Data," is a description of just this. An important contribution concerns the enrichment of existing information about composers, works and sources in the Virtual International Authority (http://viaf.org/) by incorporating information from a very specific research-driven curatorial project about mensural music. But, Smiraglia, Young and van Berchum (2021; Chapter 9) also proudly report that for the first time (173) "a corporate cultural heritage entity that was not a cataloging library [has] been allowed to participate in LC/NACO [Library of Congress]/… to enter authority records directly." It is by such pathways for integrating digital humanities results with authoritative, stable (KOS) service providers that sustainability becomes most effectively achieved. The mensural music project also serves to demonstrate that producing LD does not always mean to lift the whole of a resource to the LOD cloud; but rather that one can select different levels when it comes to the process of LOD publishing (Siebes et al. 2021; Chapter 12). Again, it very much depends on what your role in the KOS universe might be (Zeng and Mayr 2021; Chapter 3).

6.5 New endeavours

The new endeavours section sheds light on two specific dilemmata in current linking-knowledge practices. The first concerns the co-evolution of a technology and social practices of its adoption. Automatic indexing is ever evolving towards finer granularity, from indexing documents (or websites) to indexing bits of content in documents in even more precise ways. SW technology is just the fuel to enabling linking in unprecedented larger (all encompassing) and deeper (more detailed) levels. Allard Oelen, Mohamad Yaser Jaradeh, Markus Stocker and Sören Auer have contributed chapter 10 "Organizing Scholarly Knowledge leveraging Crowdsourcing: Expert Curation and Automated Techniques." They have addressed the question of how to enable indexing on the level of methods, so that in turn, a user can more quickly gain an overview about methodological achievements. In the process of KOS creation and application for automatic indexing experts are still indispensable. Machines might be able to suggest a KOS structure, but validation requires human intervention. It is up to the users based on their best practices to suggest, select, test, apply and re-design the KOS in question repeatedly. In this sense, this chapter once more re-emphasizes the position of the user in machine-based/automatic classification.

The dream of a knowledge graph that enables a new organization of human knowledge—the old Otlet/La Fontaine dream so to say—also encapsulates another dilemma. To be able to scale up, to connect across knowledge domain boundaries, one needs to identify elements that are generic and can act as bridges between those domains. Inevitable, that means that concreteness, semantic embeddedness in one specific context will need to be washed out. In other words, one has to find a formalization that enables the trading zone discourse, which we have so often pointed to in this introduction. This dilemma—generic versus specific—is an eternal problem, and answers to it eventually relate back to fundamental philosophical stances: is it ever possible to find common ground or are we doomed to be stay confined inside our own individual, local world views? Is there an objective reality, and if so how do we gauge and evaluate different representations of it over and against each other? The last chapter is this section (Chapter 11) is by van den Heuvel and Smiraglia, who address exactly these questions seeking a conceptual framework for thinking about possible solutions. In "Knowledge Spaces: Visualizing and Interacting with Dimensionality," they discuss how best to enable access to different perspectives, or representations. While socially constructed, such perspectives are not arbitrary. The authors go one step further and discuss how to turn the problem (the existence of different perspectives) into a solution (enabling a better understanding). "Ordering the Ordering Systems" and "Using Visualizations" are part of their answer. In essence, they seek (201) "a more instrumental use of multidimensional knowledge spaces to organise and to interact with concepts."

6.6 Education

The last section of the book brings focus to the aforementioned dilemma resulting from the co-evolution of technology and its use in a different way. The virtue of SW technology—its innovative character and ever-new emerging possibilities—can turn into a fault when it comes to its adoption, and in particular when the adoption is not properly managed. One has to acknowledge that when operating at the boundary of different fields, investment needs to be made in the translation process concerning concepts and approaches as well as eventual experimentation and implementation. Concerning LD one can observe all kinds

of myths circulating. There is sometimes a naive belief that once data are transferred to a LOD format they automatically become interwoven into larger knowledge graphs, and sometimes the same belief can transfer into fear of losing control over one's own data, argument or research. Both beliefs arise from incomplete information, and the inhibit further wider adoption. The professional organization W3C[6] of the SW field creates extensive documentation to foster standardization processes. But, their main addressants are the experts and professionals in the field, not primarily adopters from other fields. So, while there are ample recommendations on how to publish LOD, there is still a need for educational material. This observation, made in many projects, motivated Siebes, Coen, Gregory and Scharnhorst to engage in a so-called "sprint" organised by the Mozilla Library Carpet movement and to write a guide for LOD publishing for everyone. Based on existing W3C recommendations, steps (called "things" in the Library Carpet format) were identified that need to be pursued in bringing an information resource to the LOD cloud. This guide also informed the architectural design of a new L(O)D service of the UDC (Slavic Siebes and Scharnhorst 2021; Chapter 5). However, in applying the guide to a use case of our own we once more experienced that formal workflows (as such a guide represents) in research practice are really thinking tools rather than automatic fabrication tools. In each application context, those steps to follow will be different. Having said this, one important message of this chapter remains: namely many steps in the process of publishing LOD require thinking prior to programming, and could be executed with pen and paper. In other words, each LOD project needs a blueprint as well as a machine. To produce the first, "translation work" is needed, to produce the second, a specialist from the SW community is needed.

7.0 Linking knowledge: The synergy of knowledge interaction

The title of this book was not arbitrary. Not a bit. The explorers who have collaborated in this book are a team devoted to moving beyond the simple concatenation of RDF triples into a realm where the linking of data represents a true linking of knowledge, which itself becomes a linking for knowledge interaction. We all set out on a journey to make sense of the chaos in the World Wide Web, no less than did our predecessors over the past three centuries try to make sense of the chaos unleashed by the printing press. We have laid out a path much more useful than the cookie crumbs of Hansel and Gretel. We have followed a path set out for us over centuries of work on bringing to fruition the most possibly useful organization of knowledge. We humbly present this book as evidence of our journey.

The useful linking of knowledge across spectra is an eternal human dream. From ancient stargazers to scientists of the LOD Cloud, like all those who have contributed to this book, the goal is the realization of two mid-20th century scholarly dreams: facilitating what Patrick Wilson ([1968] 1978) called exploitative power, or the power to synthesize knowledge with laser-like precision; and to do so by bringing together what Don Swanson (1986) called undiscovered public knowledge, in other words facts related in as yet undiscovered ways. These two goals meet in what van den Heuvel and Smiraglia (2021; Chapter 11) call "knowledge interaction." The rise of the idea of the SW represents the realization of these two dreams in much the same way as the automation of bibliographic control in the last quarter of the 20th century led to one realization of the Otletian dream of universal control of research, first through bibliographic utilities like OCLC, Inc., and then subsequently through the rise of the World Wide Web. That is to say, while the SW promises us the power of both exploitative ability and unfettered synthesis, still it also represents a

divergence from the imaginable technologies of the past and therefore is a pathway to new and as yet unimagined exploration. The authors contributing to this book—our explorers—have reported their observations and wisdom concerning the forging of this new dream, with the especially delicious twist of a focus on the social sciences and humanities (SSH).

Along the way our explorers have discovered and here reported on the parameters of what some of them call a new ecosystem. Pattuelli (2021; Chapter 8) refers explicitly to the "linked data (LD) ecosystem." The ecosystem is bounded by RDF triples, which are themselves the outgrowth of a universe of data. Similarly, this ecosystem is populated by realized knowledge infrastructure in the form of knowledge organization systems (KOSs) that Zeng and Mayr (2021; Chapter 3) refer to variously as communities, researchers, producers and users as well as colonies. Van den Heuvel and Smiraglia (2021; Chapter 11) relate this populated ecosystem to Otlet's visualizations of the dichotomous "Self (le Moi)" and "Societies (Societés)" as coexisting realities of perception of the organized knowledge universe. What kind of thing, then, is this new SW reality?

Smiraglia (2014a) wrote about the potential synergies of information institutions as social realities. Information institutions are defined as those (1) "that preserve, conserve and disseminate information objects and their informative content." The commonality among information institutions lies in (2ff.) their shared mission to disseminate knowledge by means of some sort of query-response system, and that by virtue of these they manifest a form of *gravitas*. Cultural synergy "is the combination of perception- and behavior-shaping knowledge, within, between, and among groups that contributes to the now realized virtual reality of a common information-sharing interface." It seems obvious that a knowledge-sharing environment as rich and lively as the semantic linking ecosystem, populated and colonized by communities of researchers, producers and users both constitutes and is comprised of information institutions. The LD ecosystem(s), the LOD Cloud, the SW and their constituent LOD KOSs and LOD knowledge graphs (KGs) all qualify as information institutions. The synergies among them are the real thesis of this book.

There is a critical element that we can use to help us comprehend the evolving cultural synergies shaping the LD ecosystems and that is the notion of social epistemology (6): "information institutions arise culturally from social forces of the cultures they inhabit, and … their purpose is to disseminate that culture." Certainly, the interwoven layers of data ecosystems, but in particular LOD KGs, populated by the "societies" of communities of researchers, users and producers demonstrate in the action of colonization around "approved" (cf. Zeng and Mayr 2021; Chapter 3) LOD KOSs are entirely creatures of the cultures from which they have sprung and are determined disseminators of their culturally requisite knowledge stores. One synergy is immediately apparent and that is the intermingling of cultural realities of the LD producing community of computer and information scientists, on the one hand, and the rich SSH communities of researchers and users, on the other. Examples in this book are LOD KOSs such as the Basic Concepts Classification and the Universal Decimal Classification, the Golden Agents, Mensural Music, Linked Jazz and Open Research knowledge graphs. In each case the KG is the intermingled product of interdisciplinary interaction between the SW and SSH communities. The synergy is the dualistic social epistemology—these KGs are disseminators not only of their research content but of their constantly evolving SW ecosystems as well.

Specific synergies exist also in the infrastructural elements of the LD ecosystem. Vectors in knowledge space, described by van den Heuvel and Smiraglia (2021; Chapter 11),

are essentially syndetic connectors that cross pathways intersecting not only specifically linked data but also the data ecosystems surrounding each such linkage. The vectors are synergistic vehicles navigating undiscovered related conceptual space creating knowledge interaction. The knowledge space in which the vectors operate is the synergistic multidimensional knowledge space of a universe of KGs, themselves connected epistemologically by the methods underlying their construction and by the very fuzzy nature—noted by Renwick and Szostak (2021; Chapter 4)) as well as by van den Heuvel and Zamborlini (2021; Chapter 6)—of SSH domains where research relies on inexact linkage to generate useful matches.

Another synergy is the power of the LD ecosystem to merge the historical record, including evidence of the products of creative action. These elements meet, indeed suffuse one another, in the LD world of the bibliographic authorities for designated creators (authors, composers, etc.) where the library driven Virtual International Authority File (VIAF) bumps against DBPedia, but lesser known creators ranging from jazz musicians to contributing librarians to artisans of the Dutch Golden Age are identified through crowdsourced Named Entity Recognition (NER) modules. Pattuelli (2021: Chapter 8) identifies the importance of this synergy by reminding us that (162) "the full potential of LD is reached when heterogeneous data from different sources are interlinked providing unified access to data and the possibility to seamlessly query multiple graphs."

Classification is perhaps the most powerful tool ever devised by science. Its emergence in the LD ecosystem as the queen of the LOD KOS is testimony to its virtue for both gathering and disambiguation. Divergent philosophies underlie potential universal (i.e., general) classifications and therefore their potential synergistic effect when used in combination in the LD ecosystem. The discipline-based UDC has the power of over a century of application in the linking of the documentary evidence of recorded knowledge (cf. Slavic, Siebes and Scharnhorst 2021; Chapter 5). A late 20th century competitor was the Information Coding Classification of Dahlberg (cf. van den Heuvel and Smiraglia 2021; Chapter 11), which is liberated from the constraints imposed on the UDC by replacing disciplines with ontical structures. The phenomenon-based Basic Concepts Classification (cf. Renwick and Szostak 2021, Chapter 4; Smiraglia and Szostak 2021, Chapter 7) is designed to promote interdisciplinarity by structuring phenomena in causal relation sequences. There is an emerging synergy produced by the use of any and all of these classifications (see Szostak, Smiraglia, Scharnhorst, Slavic, Martínez-Ávila and Renwick 2021; Chapter 2) not only as LD themselves but in conjunction with each other as descriptors linked to points representing concepts in the LD Cloud. It is as though each classification represents a distinct dimension in the knowledge universe. The points in each dimensional knowledge space representing classified concepts or phenomena become additional vectors crossing the many dimensions to create synergistic knowledge interaction.

Our intrepid explorers (the authors who contributed to this book) did not embark on this frontier unprepared. We can partially observe the manifold provisions for this journey by analyzing the discourse they share. Discourse analysis is an evolving methodology of domain analysis in KO, seeking identification of the conversation, or "discourse," to reveal underlying points of view shared by authors in a domain. According to Smiraglia (2015, 15): discourse analysis is one means of revealing the interacting symbolic contexts in the discourse that are affecting perception … [by] selecting key elements of discourse in a

domain." Whereas other methods of domain analysis reveal the ontology at work in a domain, discourse analysis helps narrate the collective theoretical framework. Techniques for discourse analysis vary from ethnographic narrative analysis to informetric analyses. Smiraglia (2018) demonstrated the use of bibliometric analysis to reveal the contours of domain discourse. For the purpose of discovering the discourse present among the authors whose work appears in this book we have compiled and analyzed the reference lists from all twelve chapters.

For example, there are 351 references to works cited in the twelve chapters, of which 25 references are cited three or more times making up one third of all references. Not surprisingly, the most-cited authors are among the contributors: Smiraglia (25), Szostak (12), van den Heuvel (11), Slavic (10) and Pattuelli (6). Although there is some self-citation, which is common on a research front where the authors are reporting sequential new research, there also is a fair bit of cross citation. That is, these contributors know, rely upon, and perhaps most importantly acknowledge their reliance upn each other's work. Works cited three or more times give a clue to the community discourse. These are (Table 1):

Authors	Title
	The LOD Cloud
Berners-Lee (2006)	"Linked Data"
Idrissou, Zamborlini, van Harmelen and Latronico (2019)	"Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers"
Rayward (1990)	*International Organization and Dessemination of Knowledge: Selected Essays of Paul Otlet*
Smiraglia and van den Heuvel (2013)	"Classifications and Concepts: Towards an Elementary Theory of Knowledge Interaction."
Szostak, Scharnhorst, Beek and Smiraglia (2018)	"Connecting KOSs and the LOD Cloud"

Table 1. Works cited three times or more in this volume.

A very interesting backdrop to the shared discourse emerges. First, we have the actual living LOD Cloud, which is clearly in every mindset. Alongside that visualization of the SW we have two historical outposts—essays by Paul Otlet, the 19th century visionary who postulated something like a semantic universe that might be technologically feasible, and musings on the technicality of LD by the SWs own 21st century visionary Berners Lee. The three remaining works include an extensive theoretical essay on visualizing knowledge interaction (Smiraglia and van den Heuvel 2013), a paper on the essence of bringing data from the SSH into the LOD Cloud—the inexactitudinous nature of SSH data, which often requires the use of inexact matching for interpretation (Idrissou et al. 2019), and the opening salvo from the Digging Into the Knowledge Graph team concerning the necessity and processes for connecting the LOD Cloud to traditional KOSs (Szostak et al. 2018).

Author co-citation analysis is a technique by which all pairs of authors cited together in a domain are mapped. In general, co-citation indicates perceived association (e.g., semantic, thematic, epistemological, etc.) on the part of the citing author between the members of a pair. That is, a co-citation map shows how citing authors perceive associations among cited works. In domain analysis the technique is useful for visualizing theoretical poles in a specific domain, or we might also say nodes of discourse, represented by the perceived

associations. Visualization takes place by using multi-dimensional scaling (MDS) to generate a network map of co-cited authors. Each referenced author forms a node in the network, and edges between authors occur if they are co-cited. The weight of the edge indicates the strength of the perceived association; thicker edges represent more frequent co-citation, thus revealing more influential theoretical poles. As part of our discourse analysis we plotted co-citation across the twelve chapters of this book; Figures 1 and 2 present two views of the author co-citation network in this volume.

Figure 1 shows a Gephi plot of author co-citation among those authors most cited in this volume. This gives a clue to the shared discourse, or conversation, among the citing authors, who are (of course) the authors contributing to this volume. Here we ask the question, what theoretical poles have influenced the work underlying the collective contributions to the idea of linking knowledge.

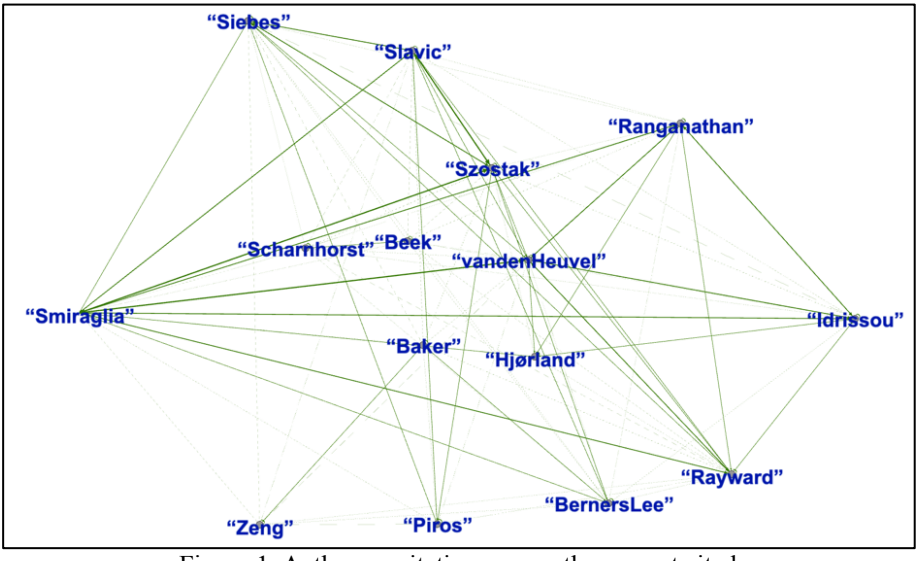


Figure 1. Author co-citation among those most cited.

At the core of this network we find a cluster including key contributors to this volume, but the core is informed by work by Ranganathan and Rayward (Otlet), which is evidence of the historical grounding of the discourse. Also interesting is the prominence of applications from Beek, Piros, Siebes and Idrissou representing the keen importance of the specific technologies necessitated for connecting KOSs and the SSH to the LD ecosystem. The strongest connection shown by the heaviest edges is the network among Smiraglia Szostak Slavic and van den Heuvel.

A slightly different view of the discourse can be generated by restricting the analysis to only those authors who are contributors to the volume. In other words, we now ask, how do these authors view each other's theoretical contributions to the notion of linking knowledge? Figure 2 shows a Gephi plot of author co-citation among authors contributing to this volume.

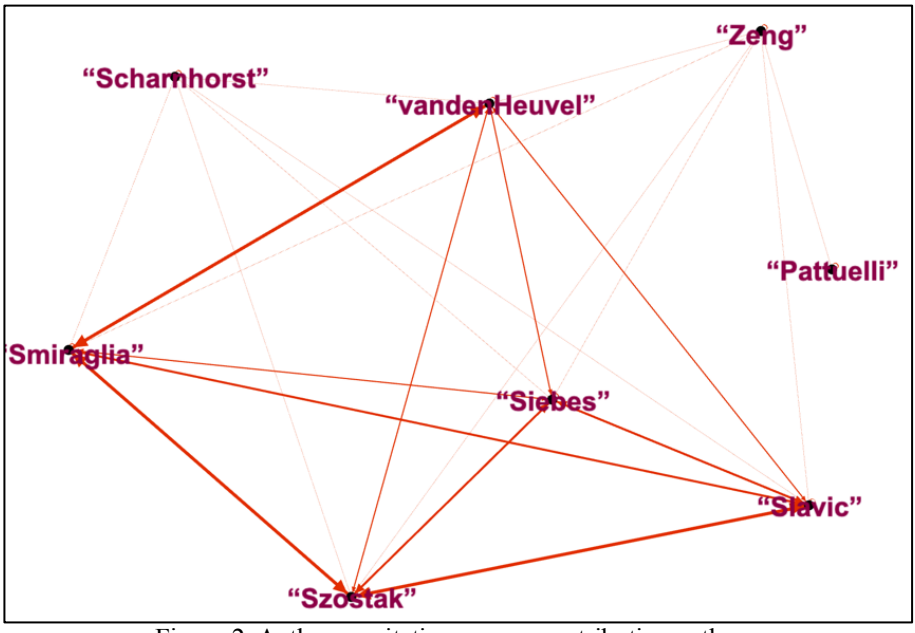


Figure 2. Author co-citation among contributing authors.

Obviously, authors from Figure 1 not represented in this plot were not co-cited in the volume, leaving a very explicit imprint of the loose but apparent discourse at work. Here we see the degree to which the core authors—our most intrepid explorers—rely on each other's work. The theoretical core overlaps that from Figure 1—the principal core related to the linking of knowledge in SSH via KOSs is anchored in reliance on specific LD technologies and buttressed by a strong connection to ideas about visualizing knowledge interactions.

Co-word analysis of the same dataset (the titles of the works cited in the volume) can be used both as a form of methodological triangulation and as a means of informing the interpretation of the discourse visualization. The Provalis ProSuite's WordStat module was used to help to visualize the core concepts represented in the research cited by our contributing authors. Figure 3 is an MDS plot of the most frequently occurring keywords and Figure 4 is a plot of the most frequently occurring two to five-word phrases.

Figure 3 shows the boundaries of the discourse at play: the core cluster is a combination of "classification" "information" and "knowledge" orbited by the SW, LD and historical memory. Figure 4 gives more breadth to the discourse by showing the core of SW LD and cultural heritage orbited by iterations of knowledge graphs and the fascinating cluster including a "universe of knowledge" and "information retrieval." We also see the prominence of the phrase "contextual entity disambiguation in domains." Thus, there is synergy in the discourse across historical and immediate imperatives driven by interdisciplinary approaches to knowledge interaction. That is the grace of this book.

Let us then recount the ways in which our non-arbitrary project has produced non-trivial synergies:

•Intermingling of cultural realities of the LD producing community of computer and information scientists and the rich SSH communities of researchers and users;

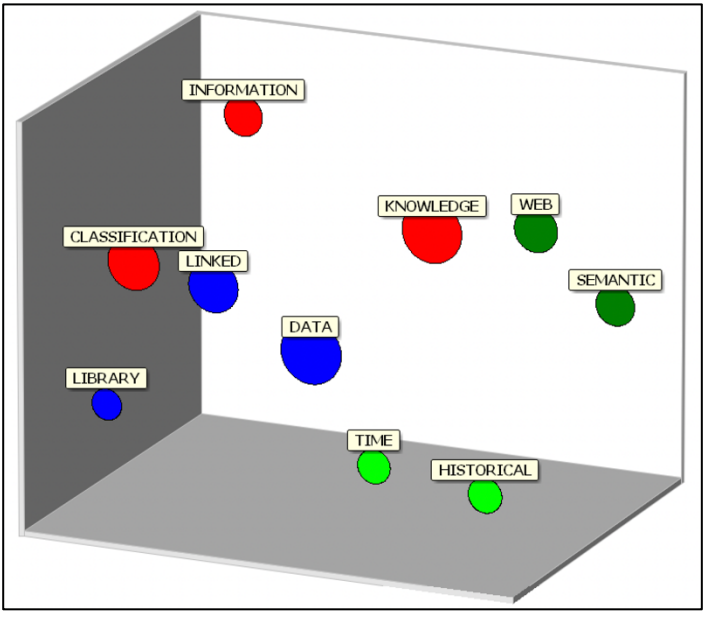


Figure 3. Most frequently occurring keywords (stress = .017138 R^2 = .9602).

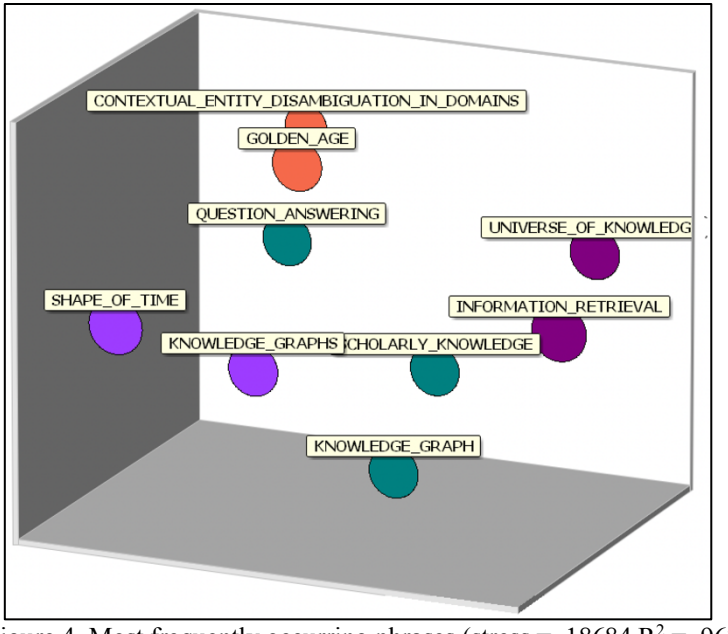


Figure 4. Most frequently occurring phrases (stress = .18684 R^2 = .9670).

•Vectors in knowledge space are synergistic vehicles navigating undiscovered related conceptual space creating knowledge interaction;

•The power of the LD ecosystem to merge the historical record, including evidence of the products of creative action;

•Divergent philosophies underlie universal (i.e., general) classifications and therefore offer a potential synergistic effect when used in combination in the LD ecosystem;

•There is an emerging synergy produced by the use of any and all of these classifications not only as LD themselves but in conjunction with each other as descriptors linked to points representing concepts in the LOD Cloud.; and,

•There is synergy in the discourse across historical and immediate imperatives driven by interdisciplinary approaches to knowledge interaction.

Ultimately the LD ecosystem explored and documented so eloquently by the contributors to this volume represents a potentially unbridled source of knowledge generation, acquisition, production and dissemination. The underlying discourse shows the extent to which our explorers are firmly grounded by historical vision yet equally firmly dedicated to the promise of linking knowledge for interaction. This new SW reality is an exciting frontier of fascination, expansion and growth. The already maturing ecosystems of the SW are interlocking information institutions clearly devoted to the expansion of human experience through the growth of knowledge interaction.

Notes

1. Digging Into the Knowledge Graph (DIKG). https://diggingintodata.org/awards/2016/project/digging-knowledge-graph
2. https://www.cs.vu.nl/~frankh/#
3. https://www.loc.gov/marc/umb/um01to06.html
4. https://www.goldenagents.org
5. https://linkedjazz.org
6. https://www.w3.org

References

Antoniou, Grigoris, Paul Groth, Frank van Harmelen and Rinke Hoekstra. 2012. *A Semantic Web Primer*. 3rd ed. Massachusetts: MIT Press.

Beek, Wouter, Laurens Rietveld, Hamid R. Bazoobandi, Jan Wielemaker, and Stefan Schlobach. 2014. "LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data." In *The Semantic Web ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, ed. Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz and Carole Goble. Lecture Notes in Computer Science 8796. Cham: Springer, 213-28. https://doi.org/10.1007/978-3-319-11964-9_14

Beek, W.G.J., L. Rietveld and S. Schlobach. 2016. "LOD Laundromat (Archival Package 2016/06)." DANS. https://doi.org/10.17026/dans-znh-bcg3

Benjamins, Victor Richard, Jesús Contreras, Oscar Corcho and Asunción Gómez-Pérez. 2002. "The Six Challenges of the Semantic Web." In "KR2002 Semantic Web Workshop." http://oa.upm.es/5668/1/Workshop06.KRR2002.pdf (accessed on 29 December 2020).

Berners-Lee, Tim. 2006. "Linked Data." https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284, no. 5: 34-43. http://www.jstor.org/stable/26059207

Borgman, Christine L. Andrea Scharnhorst and Milena S. Golshan. 2019. “Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse.” *Journal of the Association for Information Science and Technology* 70: 888-904. DOI: 10.1002/asi.24172

Bowker, Geoffrey C. and Susan Leigh Star. 1999. *Sorting Things Out: Classification and its Consequences*. Cambridge, Mass., MIT Press.

Felt, Ulrike, Rayvon Fouché, Clark A. Miller and Laurel Smith-Doerr, eds. 2017. *The Handbook of Science and Technology Studies*. 4th ed. Cambridge Mass.: MIT Press.

Furner, Jonathan. 2020. “Classification of the Sciences in Greco-Roman Antiquity.” In: *Encyclopedia of Knowledge Organization*, ed. Birger Hjørland and Claudio Gnoli. N.p.: International Society for Knowledge Organization. https://www.isko.org/cyclo/greco-roman

Galison, Peter. 1997. *Image & Logic: A Material Culture of Microphysics*. Chicago: The University of Chicago Press.

Gnoli, Claudio. 2020. *Introduction to Knowledge Organization*. London: Facet.

Gregory, Kathleen. 2021. “Findable and Re-usable? Data Discovery Practices in Research.” PhD diss. Maastricht University.

Heath, Tom and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. San Rafael, Calif.: Morgan & Claypool Publishers.

Hyland, Bernadette, Ghislain Atemezing, Michael Pendleton and Biplav Srivastava, eds. 2013. *Linked Data Glossary*. N.p.: W3C Governement Linked Data Working Group. https://www.w3.org/TR/ld-glossary/

Hyland, Bernadette, Ghislain Atemezing and Boris Villazón-Terrazas, eds. 2014. *Best Practices for Publishing Linked Data*. N.P.: W3C Working Group Note. http://www.w3.org/TR/ld-bp/

Idrissou, Al, Veruska Zamborlini, Frank van Harmelen and Chiara Latronico. 2019. “Contextual Entity Disambiguation in Domains with Weak Identity Criteria.” In *Proceedings of the 10th International Conference on Knowledge Capture*. New York, NY, USA: ACM, 259–62. https://doi.org/10.1145/3360901.3364440

Kedrow, B.M. 1975. *Klassifizierung der Wissenschaften*. Berlin: Akademie Verlag.

Le Franc, Yann, Jessica Parland-von Essen, Luiz Bonino,Heikki Lehväslaiho, Gerard Coen and Christine Staiger. 2020. “D2.2 FAIR Semantics: First recommendations (Version 1.0).” FAIRsFAIR. DOI: 10.5281/zenodo.3707985

Leydesdorff, Loet. 2006. *The Knowledge-Based Economy: Modeled, Measured, Simulated*. Boca Raton, FL: Universal Publishers.

“The Linked Open Data Cloud.” https://lod-cloud.net/#diagram. https://web.archive.org/web/20201215080839/https://lod-cloud.net/.

Mayr, Philipp and Marcia Lei Zeng. 2021. “Knowledge Organization Systems in the Semantic Web: A Multidimensional Review.” In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhordt. Baden-Baden: Ergon Verlag, 35-64.

Mutschke, Peter, Andrea Scharnhorst, Nicholas J. Belkin, André Skupin and Philipp Mayr, eds. . 2917. “Introduction to the Special Issue on Knowledge Maps and Information Retrieval (KMIR).” *International Journal on Digital Libraries* 18: 1-3. DOI: 10.1007/s00799-016-0204-4

Oelen, Allard, Mohamad Yaser Jaradeh, Markus Stocker and Sören Auer. 2021. “Organizing Scholarly Knowledge leveraging Crowdsourcing, Expert Curation and Automated Techniques.” In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 182-99.

Pattuelli, M. Cristina. 2021. “Graphing Out Communities and Cultures in the Archives: Methods and Tools.” In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 144-67.

Ranganathan, Shiyali Ramamrita. 1973. *Documentation: Genesis and Development*. Delhi, Vikas Publ. House

Rechenberg, Ingo. 1994. *Evolutionsstrategie ‘94*. Stuttgart: Frommann-Holzboog.

Rayward, W. Boyd. 1990. *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Amsterdam: Elsevier.
Renwick, Tobias and Rick Szostak. 2021. "A Thesaural Interface for the Basic Concepts Classification." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 65-69.
Scharnhorst, Andrea, Richard P. Smiraglia, Christophe Guéret and Alkim Almila Akdag Salah. 2016. "Knowledge Maps of the UDC: Uses and Use Cases." *Knowledge Organization* 43: 641-54.
Scharnhorst, Andrea, Katy Börner and Peter Besselaar, eds . 2012. *Models of Science Dynamics*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23068-4.
Siebes, Ronald, Gerard Coen, Kathleen Gregory and Andrea Scharnhorst. 2019. "Linked Open Data. 10 Things toward the LOD Realm: The "I" in FAIR in a Semantic Way." Zenodo https://doi.org/10.5281/zenodo.3471806
Siebes, Ronald, Gerard Coen, Kathleen Gregory and Andrea Scharnhorst. 2021. "Publishing Linked Open Data: A Recipe." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 219-33.
Slavic, Aida and Sylvie Davies. 2017. "Facet Analysis in the UDC: Questions of çunctionality and çormality. *Knowledge Organization* 44: 425-35.
Slavic, Aida, Ronald Siebes and Andrea Scharnhorst. 2021. "Publishing a Knowledge Organization System as Linked Data." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 70-99.
Smiraglia, Richard P. 2014a. *Cultural Synergy in Information Institutions*. Cham: Springer.
Smiraglia, Richard P. 2014b. *The Elements of Knowledge Organization*. Cham: Springer.
Smiraglia, Richard P. 2015. *Domain Analysis for Knowledge Organization: Tools for Ontology Extraction.* Chandos Information Professional Series. Oxford: Elsevier/Chandos.
Smiraglia, Richard P. 2018. "The Evolution of the Concept: A Case Study from *American Documentation*." *Canadian Journal of Information and Library Science* 42: 113-34.
Smiraglia, Richard P. and Andrea Scharnhorst. 2019. "Call for Contributions to a Monograph. Title: *Linking Knowledge*." Email.
Smiraglia, Richard P. and Charles van den Heuvel. 2013. "Classifications and Concepts: Towards an Elementary Theory of Knowledge Interaction." *Journal of Documentation* 69: 360-83.
Smiraglia, Richard P. and Rick Szostak. 2021. "Identifying and Classifying the Phenomena of Music. In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 143-67.
Smiraglia, Richard P., Andrea Scharnhorst, Almila Akdag Salah and Cheng Gao. 2013. "UDC in Action." In *Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands*, ed. Aida Slavic, Almila Akdag Salah and Sylvie Davies. Würzburg: Ergon Verlag, 259-72.
Smiraglia, Richard P., James Bradford Young and Marnix van Berchum. 2021. "Digging into the Mensural Music Knowledge Graph: Renaissance Polyphony meets Linked Open Data." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 168-81.
Star, Susan Leigh and James Griesemer. 1989. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19: 387–420. doi:10.1177/030631289019003001
Swanson, Don R. 1986. "Undiscovered Public Knowledge." *The Library Quarterly* 56: 103-18.
Szostak, Rick, Andrea Scharnhorst, Wouter Beek and Richard P. Smiraglia. 2018. "Connecting KOSs and the LOD Cloud." In *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal*, ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Baden-Baden: Ergon, 521-29.

Szostak, Rick, Richard P. Smiraglia, Andrea Scharnhorst, Ronald Siebes, Aida Slavic, Daniel Martínez-Ávila and Tobias Renwick. 2020. "Classifications as Linked Open Data: Challenges and Opportunities." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 25-34.

Tennis, Joseph T. 2012. "The Strange Case of Eugenics: A Subject's Ontogeny in a Long-Lived Classification Scheme and the Question of Collocative Integrity." *Journal of the American Society for Information Science and Technology* 63:1350-59.

Van den Heuvel, Charles and Veruska Zamborlini. 2016. "Modeling and Visualizing Storylines of Historical Interactions: Kubler's *Shape of Time* and Rembrandt's *Night Watch*." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 100-42.

Van den Heuvel, Charles and Richard P. Smiraglia. 2021. "Knowledge Spaces: Visualizing and Interacting with Dimensionality." In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 200-18.

Vandenbussche, Pierre-Yves, Ghislain A. Atemezing, Maria Poveda-Villalón, and Bernard Vatant. 2017. "Linked Open Vocabularies (LOV): A Gateway to Reusable Semantic Vocabularies on the Web." *Semantic Web* 8:437-52. https://doi.org/10.3233/SW-160213

Whitelaw, Mitchell. 2015. "Generous Interfaces for Digital Cultural Collections." *DHQ: Digital Humanities Quarterly* 9, no. 1. http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

Wikipedia. 2020. S.v. "Semantic web." https://en.wikipedia.org/wiki/Semantic_Web

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Nature* 3:160018. https://doi.org/10.1038/sdata.2016.18

Wilson, Patrick. (1968) 1978. *Two Kinds of Power: An Essay on Bibliographical Control*. California Library reprint series ed. University of California Publications; Librarianship 5. Berkeley: University of California Press.

Wouters, Paul, Anne Beaulieu, Andrea Scharnhorstcand Sally Wyatt, eds. 2013. *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*. Cambridge, Mass.: MIT Press.

Wright, Alex. 2014. *Cataloging the World: Paul Otlet and the Birth of the Information Age*. Oxford: Oxford University Press.

Yoose, Becky and Jody Perkins. 2013. "The Linked Open Data Landscape in Libraries and Beyond." *Journal of Library Metadata* 13, nos. 2/3:197-211. http://www.tandfonline.com/doi/full/10.1080/19386389.2013.826075

Ziman, John M., ed. 2003. *Technological Innovation as an Evolutionary Process*. Cambridge: Cambridge University Press.

Rick Szostak
University of Alberta

Richard P. Smiraglia
University of Wisconsin-Milwaukee

Andrea Scharnhorst
Data Archiving & Networked Services (DANS)

Aida Slavic
UDC Consortium

Daniel Martínez-Ávila
Universidad Carlos III de Madrid

Tobias Renwick
University of Alberta

Chapter 2
Classifications as Linked Open Data
Challenges and Opportunities[‡]

Abstract
Linked Data (LD) as a web-based technology enables in principle the seamless, machine-supported integration, interplay and augmentation of all kinds of knowledge, into what has been labeled a huge knowledge graph. Despite decades of web technology and, more recently, the LD approach, the task to fully exploit these new technologies in the public domain is only commencing. One specific challenge is to transfer techniques developed pre-web to order our knowledge into the realm of Linked Open Data (LOD). This paper illustrates two different models in which a general analytico-synthetic classification can be published and made available as LD. In both cases, an LD solution deals with the intricacies of a pre-coordinated indexing language. The Universal Decimal Classification (UDC) approach illustrates a more complex solution driven by the practical requirements that the LD model is expected to fulfill in the bibliographic domain, and within the constraints of copyright protection. The Basic Concepts Classification (BCC) is a new classification with a novel approach to classification structure and syntax for which LD is an important vehicle for increasing the scheme's visibility and usability. The report on these two cases illustrate some of the challenges of the representation of knowledge organization systems as LD and the possibilities that analytico-synthetic and interdisciplinary or phenomenon-based systems present for the representation of knowledge using LD.

[‡] Reprinted with minor editorial emendations by permission from *Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark*, ed. Marianne Lykke, Tanja Svarre, Mette Skov and Daniel Martínez-Ávila. Advances in Knowledge Organization 17. Baden-Baden: Ergon Verlag, 436-45.

1.0 Introduction

There is much excitement about the introduction of formal systems of knowledge organization (KO) into the infrastructure of the Linked Data (LD) and especially Linked Open Data (LOD) cloud. Expectations are grounded in the fact that LD connect phenomena with shared (controlled) vocabularies. In theory, meaningful links from specific points in the cloud-based knowledge graph to normalized concepts in formal classifications can help to strengthen a shared conceptual infrastructure—not simply meaningful semantics but also effective syndetic routing among concepts. This objective was the core research question of the "Digging Into the Knowledge Graph" research project.[1]

Szostak et al. (2018, 527-28) pointed out how three major challenges comprised sorting concepts, translating across domains and publishing knowledge organization systems (KOSs) as LOD. The Universal Decimal Classification (UDC) and the Basic Concepts Classification (BCC) —one disciplinary and the other phenomenon-based—were chosen as case studies to explore what problems might emerge along the journey of making KOSs available as LOD. As Siebes at al. (2019) detailed, the process of moving into the realm of LD is composed of stages of conceptual and technological explorations and decisions. Under the former fall questions such as what information to make available in a machine-readable form, to which extent existing vocabularies should be re-used, and whether or how to already enrich your LD prior to publication. Under the latter we find questions such as which web domain to use, how to design the URL's, but also how to guarantee stability over time and how to document provenance during possible editions or versions of the LD publication.

Special challenges arise from the formal representation of KOSs as LD which are at once semantic and logistical. Semantic issues arise due to terminological diversity in the unorchestrated, self-organized nature of the LD cloud itself. The job of linkage from specifically well-defined points in a classification to a potential of semantic relations in the cloud is a non-trivial research task. Methodologically, and when dealing with Linked Open data (LOD), different routes for interlinking exist: point-to-point explorations in the process of publishing a resource as LOD (Siebes et al. 2019); inspection of LOD clusters as literary warrant (Martínez-Ávila et al. 2018, 10); and translation between knowledge domains (Eito-Brun 2018; Marcondes 2018).

For KOSs that come with an extended legacy (a long history of well curated editions), such as the UDC, the choice of the appropriate namespace is non-trivial. We report approaches taken to publish the UDC and BCC as LOD enabling seamless integration into the cloud. Problems tackled in the process encompass data modelling, design of applied web technology (e.g., URI design), versioning (instantiating), licensing, extending KOSs published as LOD, and other possibilities to disseminate, exploit and enhance KOSs. Publication of a KOS as LD is not trivial; rather, it requires a whole process of which many parts need to be accomplished first off-line.

The task of translating a KOS into LOD is challenging in many ways. In the aforementioned conceptual stage, selectivity is one aspect seldom discussed. The first task is not to transfer the whole of the KOS to the new (Resource Description Framework or RDF) data model, but to choose those parts of the KOS that are most importantly available in a machine-readable LD format. In this process, the use of already existing vocabularies is reco

mmended to express selected features from the KOS in the new data model. This translation may allow, or at least facilitate, a translator to translate KOS terminology into items already mapped into LOD schemas. A second task is that of mapping connections from one RDF schema to another. Both tasks are far from being a mechanical mapping process, but rather require research as exemplified further below.

More particularly, this paper illustrates two different models in which a general analytico-synthetic classification can be published and made available as LD. In both cases, an LD solution deals with the intricacies of a pre-coordinated indexing language. The UDC approach illustrates a more complex solution driven by the practical requirements that the classification LD are expected to fulfil in the bibliographic domain, and within the constraints of copyright protection. The BCC, interdisciplinary in nature, is a new classification with a novel approach to classification structure and syntax for which LD are an important vehicle for increasing the scheme's visibility and usability.

2.0 The BCC linked data publishing model

The Basic Concepts Classification (BCC)[2] was created by Rick Szostak for the purpose of providing structured direct access by phenomenon to documents (and the ideas expressed in them). The BCC grew by the addition of schedules of mostly verb-like relators and adjectival/adverbial properties added to the original schedule of phenomena. Documents (objects, ideas, concepts) can be expressed with combinations of phenomena, relators and properties, either in symbolic notation (classified form) or in natural-language sentence style.

The primary difficulty in mapping a universal (i.e. a general) KOS such as the BCC to LOD is that the BCC is intended to be able to classify almost anything (see Szostak 2019 for an overview of the BCC). The LOD cloud is also universal in extent, but achieves this universality with millions of distinct terms of varying degrees of specificity. A perfect mapping of the BCC to LOD would be able to encompass the entire cloud, but only by expanding BCC classes to such an extent that they would cease to be useful for classificatory purposes. The translator is left with the task of selecting points in the LOD cloud that hopefully encompass as much related information as possible.

The first task of the translator is to understand the relations, overlap and accepted usage among the current LOD cloud schemas. The initial impression on the translator is a bewildering array of options, some new, growing and maintained (e.g., Wikidata, DBPedia, OWL, SKOS, FOAF) and others suffering from disuse, abandonment or deprecation (e.g., Freebase). This array of options is a strength of LOD, for anyone can say anything about any topic (this is the so-called AAA rule that governs the semantic web), but for the translator it is very daunting to try to figure out whether someone else is trying to say the same thing as your KOS.

Within the BCC there are essentially nouns (phenomena) and verbs (relators). There are also adjective-like Properties that can be treated in much the same way as nouns. Phenomena are significantly simpler to map as the translator needs to choose a sufficiently large schema and map terms in the BCC directly to those matching entities. As an example, we have mapped the phenomenon of "art" (http://purl.org/basic/a-art), using the relation of "sameAs" from OWL (http://www.w3.org/2002/07/owl/sameAs), to the DBPedia entry on Art (http://dbpedia.org/resource/Art). This is a reasonable mapping, and it implies that anything anyone has classified as art using DBPedia is also classified as art in the BCC. Note

that this implies to a graph query engine that the terms of the BCC and DBPedia are identical and can be merged. This property may not be ideal. In our example, within DBPedia a movie is not art, but rather it is a subclass of work (also it is identified as equivalent to schema.org "movie" where it is a creative work). In the BCC, film is indeed a subclass of art (http://purl.org/basic/ar2-film) meaning that the classification is now disjoint and films classified as art in the BCC will map to DBPedia incorrectly (at least according to DBPedia's definition). This is a general problem, for a controlled vocabulary such as the BCC is generally of greater breadth (that is, each term has a broader meaning) than an uncontrolled vocabulary such as the LOD Cloud.

One of the first points that the translator needs to comprehend is that it is possible to indicate the cardinality of relationships within LOD. For example, within SKOS there are classifications for broader and narrower, where the former implies that the object of the triple is broader than the subject, and narrower implies the inverse. For BCC "art" one might be tempted to say that "DBPedia:art SKOS:broader BCC:art", which indicates that everything DBPedia considers art is art in the BCC, but not everything the BCC considers art is in DBpedia art. The downside of using broader and narrower is that the mapping of the reciprocal is ambiguous (there is no way to know whether a BCC-Art object should be DBPedia Art). Further, there are likely examples mapped to DBPedia-Art that are not in BCC-Art. The true cardinality of the relation is that there is a significant amount of overlap between DBPedia-Art and BCC-Art, and therefore we reasonably consider them the same, given our goals. That is, we allow some small degree of inaccuracy in translation in order to indicate a broad overlap in meaning.

For the translation of relators the task is compounded as relators are used in the BCC to tie phenomena to one another, but in a more lexical way than LOD. In the BCC, relators can be used in conjunction with phenomena to add specificity to the classification. In terms of LOD, the word "visual" could be represented as "by pictures (/T7p)," where "by" is a relator and "pictures" is a phenomenon, but the idea represented by the two terms is smaller in scope than either term together. The translator may want to consider the effect of mapping the word "by" to any other definition, as while they may appear to be similar, what this implies is that for an object already mapped in LOD to be mapped to BCC it would have to link to both items in some way, which is unlikely. Again, here we were faced with a decision of imprecision and decided to create an independent classification for the relator "by."

The goal of translating a KOS to LOD is not to achieve perfection, but rather to cast a broad enough net so that the first iteration of the KOS can bring in terms that are related closely enough to its topics to test whether the KOS is capable of their classification. To this end, we begin with accepting the imperfect and hoping that it allows for iterative improvement.

3.0 The UDC linked data publishing model

The UDC has been one of the most widely used KOSs in the bibliographic domain for over a century. It is often used in conjunction with and complementary to thesauri, subject heading systems, and special classifications. During its lifetime, the classification has undergone many changes and has been made available in many languages and versions. The current UDC data standard, the UDC Master Reference File (UDC MRF) has had over twenty updates released since 1993, with 50% of the current 72,000 sets of classes having

been added or changed. The UDC data also include 12,000 cancelled (deprecated) classes that redirect to new classes. The scheme is currently owned, maintained, and developed by an international consortium of publishers, on a self-funding and non-profit basis.

The UDC scheme organizes concepts and subjects within traditional forms of knowledge (disciplines) allowing concepts and classes of concepts to be freely combined both within and between subject fields to express any level of complexity that information resources present. When both classification schemes and bibliographic metadata are published as linked data and are connected, they form a complex and dynamic knowledge space that shows the ways we create, interact with or utilize information. Classmarks stored in millions of bibliographic records hold valuable information about the contents of these collections. Once UDC classmarks are linked back to the classification scheme from which they originated, it is possible to capture their meaning and establish further meaningful associations within and among collections (cf. Slavic 2017). These connections made through linked data can help to:

- enrich bibliographic data to support information discovery by increasing subject access points using UDC terminology, by enabling semantic expansion (broadening); and by improving precision through contextualization;
- improve systematic presentation, grouping, and visualization of resources and collections (linear or multidimensional) to facilitate browsing and serendipitous discovery of information;
- link the classification to other KOSs to enable cross-collection information discovery; and,
- validate and update local classification data and local authority files or bypass local and obsolete classification data in information exchange.

Apart from many practical aspects of interest, UDC LD development represents a good testbed for further research especially through its interaction with other KOSs. As an example of an analytico-synthetic and faceted scheme, it provides a case study for managing the alignment between the simple codes that appear in the scheme and the complex classmarks generated through document indexing that contain unlimited numbers of combinations of UDC classmarks.

3.1 Challenges and solutions

While longevity and widespread use represent strong arguments for sharing the UDC as LD, this also requires more responsibilities and presents further difficulties. In 2011, an extract from UDC of 2,600 classes was published as LOD in SKOS format. This experiment proved to be a valuable experience. As more and more library catalogs became available as linked data, we learned about the magnitude of the incompatibilities between UDC classmarks in bibliographic records and the UDC standard data.

Library linked data (LLD) clouds that were observed contained specific and complex UDC classmarks that could only be resolved through the access to the complete UDC content. However, the main cause of mismatch between UDC namespace and LLD is in the fact that libraries continue to use deprecated UDC codes. Thus, it became clear that a UDC namespace has to include not only the complete content of the UDC MRF, but also a significant collection of historical data and concordances between cancelled and new classes. Needless to say, the UDC LD used from 2011-2019 indicated that programs utilizing the UDC namespace (or those creating them) have little awareness of the UDC data structure, semantics, syntax, provenance, versioning, and changes and might not be able to process and select UDC data accurately or make good use of them.

In order to serve its purpose in a bibliographic domain, the UDC namespace has to provide a robust solution for the linking and semantic alignment between classmarks in bibliographic records and those in the UDC LD cloud. This has to be achieved irrespectively of the fact that the classmark strings in library data include combinations of simple UDC codes or that some may be deprecated or generated through wrong local practice. In order to achieve this, important changes had to be made to the ways and format in which UDC LD is published. This included the change of the URI format and the change of the RDF schema, but most importantly, instead of a UDC LD dump we opted for a more complex UDC look-up service.

The main premise of the UDC LD service is that it ought to support practical use of the scheme as well as to protect UDC publishing in a way that its future is safeguarded. This specifically means that only a small part of the UDC data shall be published as LOD and most of the UDC LD content would be license protected, i.e., LD "behind the barrier." The UDC LD-based terminological service must support the following features:

1. Programmatic access to:
   a) One LOD set: the UDC Summary containing 3,000 classes (under CC BY-NC-ND 2.0 license); and,
   b) Two LD sets behind a UDC MRF license barrier:
      i) Abridged edition (12,000 classes); and,
      ii) UDC MRF (72,000 classes), including all twenty versions of the UDC MRF and historical data comprising 13,000 cancelled (deprecated) classes and their redirections to new classes;
2. A UDC Look-up service that:
   a) parses and resolves (interprets) a classmark originated from bibliographic data and links its components to relevant records in the RDF data store; and,
   b) upon request supplies URI(s) for UDC classmarks or the full RDF records.

The architecture of the UDC Look-up service has the following components:

1. RDF stores (three Virtuoso databases: the UDC Summary, the Abridged edition, and the UDC MRF) with SPARQL endpoints accessible only via a restricted RESTful API layer which uses pre-designed SPARQL templates for query execution.
2. Apache web server and custom written UDC parser written in PHP and Java. The Authentication process is handled by standard shared and private authentication keys. The HTTP/Get parameters and the HTTP headers inform the server about the type of desired result (e.g., HTML, RDF-Turtle, JSON).

Although the UDC Look-up service is planned primarily as an API for programmatic interaction it will also have an html interface for human interaction with the service. It is assumed that the API would be queried by programs submitting simple or complex UDC classmarks either to get correct URIs for UDC codes or to retrieve complete RDF records. The HTML interface allows humans to verify and explore the provided classmarks in which the parsetree, versions, and RDF translations are expressed. The most important part of this service is the "UDC interpreter", i.e., a program that parses complex UDC strings. This interpreter is based on a series of algorithms developed in an earlier project by Attila Piros (cf. Piros 2017). The UDC notation system allows for 100% accuracy in parsing of UDC strings using several groups of algorithms. Figure 1 shows an HTML interface in which a complex UDC number is split into components that, in this case, are all valid UDC classes. The service executes queries against the UDC Summary, the UDC Abridged edition or the UDC MRF and in the second step it generates an RDF representation of the information selected by the user/machine from the previous step. For clarity, the terms shown in bold underline font in Figure 1 are resolvable primitive UDC terms.

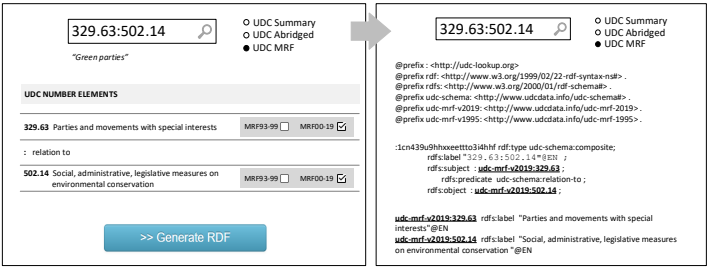


Figure 1. UDC Look-up service and interpreter

3.2 Steps in publishing UDC LD

This section outlines some of the key decisions and steps in the UDC LD service design. They broadly follow the ten step guidelines described by Siebes et al. (2019).

3.2.1 Selection of data

An important effort in this project was put into the strategic thinking and planning of UDC LD and in particular having to do with the selection of data to be published. The UDC Summary, the UDC Abridged Edition, and the UDC MRF are maintained in different MySQL databases and the same set-up is replicated for the RDF store (three Virtuoso databases). The selection of these three datasets is based on the well-established practice in UDC data use and publishing. They are representative of two kinds of access to UDC data: open access and access through a UDC MRF license requiring an authentication process based on authentication tokens (managed outside the service itself). With respect to the supported languages, the UDC Summary contains language data in 57 languages. However, in this phase the Abridged and MRF datasets are available only in English. UDC data comprise many data elements that are required for data management and publishing, for the LD we selected only 14 data elements. In terms of sequence of data release, the UDC Summary (the LOD set) was given priority due to the large community of users.

3.2.2 URI name strategy

The UDC namespace was already established in 2011 and will remain as https://udcdata.info. The UDC experience shows that the decisions regarding the URI are far from being trivial. In the 2011 LD version, we opted for URIs that had the format of the following example: "udcdata.info/068288" in which the number "068288" represented a UDC record identifier for the notation =162.3 Czech language. An important reason for not including, at the time, UDC notation in the URI was the practice of the occasional re-use of deprecated notations (usually after 10 or more years). Thus, notation on its own was considered an unreliable identifier. Once historical versions of the MRF are included in linked data, a version code can be used to contextualise the notation, so we opted for a structured URI that includes UDC notation in the following format: "udcdata.info/MRF93/=162.3." In this example, the element "MRF93" represents the earliest MRF version in which this UDC classmark appeared, i.e., the version in which it was introduced for the first time. The advantage of this approach is that it makes easier for libraries to generate classmark queries to be launched against the UDC Look-up service and allows for human control of URIs (should this be required). An inconvenience with this approach is that UDC classmarks

contain symbols and punctuations that are encoded automatically as they get processed, thus udcdata.info/MRF93/=162.3 becomes udcdata.info/MRF93/%3D162. This change of the URI format means that a new service must contain the mapping between the old 2011-2019 URIs and the new URI systems.

3.2.3 Use scenarios, serialization and resolution of UDC codes and URIs

When it comes to the Linked Data serialization of the UDC data source, we have to consider various scenarios in which the UDC namespace will be accessed. Since the service is primarily aimed for machine access, we need to have disambiguation mechanisms combined with a clear guidance to make the programmers aware of the various choices that apply. For example, often the only information libraries have about the UDC is the classmarks and the location of the UDC Look-up service. They are not aware of the UDC MRF versions, including whether classmarks contain valid or deprecated numbers or whether they have license, i.e., authentication token, to query full UDC data. Their queries may have the following format "udcdata.info/681.3(035)." The UDC Look-up service will parse and resolve the query indicating that notation 681.3 is deprecated and replaced by 004 and may return an RDF statement with sets of URIs expressing the relationship between these two numbers. If later at time an entity (machine or human) without an access key for this dataset tried to query these URI's at the UDC namespace, the authentication layer would prevent this request from being executed and return a meaningful error message, eventually combined with some sparse information about the result of the query (e.g., a superclass which the concept shares both from the MRF version and the UDC-summary version).

3.2.4 Selection of RDF schema

Following the parsing stage, URIs for individual classmark components and their grouping are generated using RDFs. For the full RDF records we use the SKOS format as it is widely used in the KOS publishing community. Equally, we wanted to maintain continuity with the 2011-2019 UDC linked data version. Below we can see the current mapping between UDC MRF data elements[3] and the SKOS schema, which is extended by UDC sub-elements (in italics):

UDC number (notation)	skos:notation	
class identifier	skos:Concept	
broader class	skos:broader	
caption	skos:prefLabel	
including note	skos:note	*udc:includingNote*
application note	skos:note	*udc:applicationNote*
scope note	skos:scopeNote	
examples	skos:example	
see also reference	skos:related	
revision history	skos:historyNote	*udc:revisionHistory*
introduction date	skos:historyNote	*udc:introductionDate*
cancellation date	skos:historyNote	*udc:cancellationDate*
replaced by	skos:historyNote	*udc:replacedBy*
last revision data	skos:historyNote	*udc:lastrevisionDate*

In the future, we plan to move towards more formalized schemas from the OWL stack. This would enable a precise formalization that allows semantic verification of classmark

strings, the vocabulary itself (e.g., when new concepts with their constraining properties are added or removed in future releases), and rich inference via transitivity, reflexivity, etc.

4.0 Conclusion and future work

As in many LD projects, the planning phase of both the UDC and BCC LD projects took more time than originally anticipated. What is often underestimated is that the translation or transference of a resource to another medium or another technology is not merely a technological enterprise but is in essence coupled to a variety of research problems. The process can be compared to the mapping of vocabularies to each other, which is also not a mere mechanical process but entails all kinds of research and editorial decisions, which in turn will influence how a KOS resource is further used. To operate on the scale of the web and with in principle unlimited outreach and spreading, the problem is only augmented. For both the UDC and BCC, key decisions had to be made through the combination of expertise in LD technologies and publishing models, on the one hand, and expertise in the UDC or BCC schemes, datasets, and publishing models, on the other. More time for reflection, research, learning and discussion than envisioned was necessary in all key stages of the project. UDC and BCC are KOSs of a different type. The BCC is newer, experimental and still growing structurally. The UDC is one of the few authorative KOSs for bibliographic databases, implemented widely, and based on a long history and fully developed KO principles of further development and implementation. Hence, the requirements for the LD publication are very different, and combining them was not part of the DiKG project. In this paper, we describe the different challenges those two KOSs are exposed to during the LD publication.

Planning and developing of the UDC namespace in the form of a Look-up service presented challenges primarily because it is both a new and a complex approach to KOS publishing, also in the realm of established semantic web practices. Web-supported access to the UDC for humans based on a multi-tier license access that combines free access to part of the resource with licensed access for experts. This needs to be mimicked in the LOD transition. In our approach, LOD and "LD behind the license barrier" models of publishing are combined and involve three different levels of classification data aimed at different audience and use scenarios. An important part of the UDC LD cloud is its historical data that will hopefully enhance the usability of UDC in the bibliographic domain where historical and obsolete classification data appear frequently. The most novel and key function to the Look-up service is the UDC interpreter. The UDC namespace is envisaged as a one-stop shop for querying and validating UDC data and it also illustrates a more complex, but hopefully more robust, model of KOS publishing as linked data. This UDC namespace offers a good environment for linked data and library linked data study and research on KOS alignments and integration.

Notes

1. Digging Into the Knowledge Graph (DIKG). https://diggingintodata.org/awards/2016/project/digging-knowledge-graph
2. Basic Concepts Classification. https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013
3. The UDC MRF data elements schema is available at: http://www.udcc.org/files/udc_data_elements_mrf11.pdf

References

Eito-Brun, Ricardo. 2018. "The Role of Knowledge Organization Tools in Open Innovation Platforms." In *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal*, ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Baden-Baden: Ergon, 666-73.

Marcondes, Carlos. H. 2018. "Culturally Relevant Relationships: Publishing and Connecting Digital Objects in Collections of Archives, Libraries, and Museums over the Web." In *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal*, ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Baden-Baden: Ergon, 539-48.

Martínez-Ávila, Daniel, Richard P. Smiraglia, Rick Szostak, Andrea Scharnhorst, Wouter Beek, Ronald Siebes, Laura Ridenour and Vanessa Schlais. 2018. "Classifying the LOD Cloud: Digging into the Knowledge Graph." *Brazilian Journal of Information Studies: Research Trends* 12, no. 4: 6-10.

Piros, Attila. 2017. "The Thought Behind the Symbol: About the Automatic Interpretation and Representation of UDC Numbers". In *Faceted Classification Today: Theory, Technology and End Users: Proceedings of the International UDC Seminar 2017, London (UK), 14-15 Septembe*r, ed. Aida Slavic and Claudio Gnoli. Würzburg: Ergon Verlag, 203-18.

Siebes, Ronald, Gerard Coen, Kathleen Gregory, and Andrea Scharnhorst. 2019. "Linked Open Data. 10 Things toward the LOD Realm: The "I" in FAIR in a Semantic Way." Zenodo https://doi.org/10.5281/zenodo.3471806

Slavic, Aida. 2017. "Klasifikacija i Library Linked Data (LLD) = Classification and Library Linked Data (LLD)". In *Predmetna Obrada: Pogled Unaprijed: Zbornik Radova*, ed. B. Purgaric and S. Spiranc. Zagreb: HKD, 13-37.

Szostak, Rick, Andrea Scharnhorst, Wouter Beek and Richard P. Smiraglia. 2018. "Connecting KOSs and the LOD Cloud." In *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal*, ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Baden-Baden: Ergon, 521-29.

Szostak, Rick. 2019. "The Basic Concepts Classification." In *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. https://www.isko.org/cyclo/bcc

Marcia Lei Zeng
Kent State University

Philipp Mayr
GESIS - Leibniz Institute for the Social Sciences

Chapter 3
Knowledge Organization Systems (KOS) in the Semantic Web
A Multi-Dimensional Review[§]

Abstract
Since the *Simple Knowledge Organization System* (SKOS) specification and its *SKOS eXtension for Labels* (SKOS-XL) became formal W3C recommendations in 2009, a significant number of conventional knowledge organization systems (KOS) (including thesauri, classification schemes, name authorities, and lists of codes and terms, produced before the arrival of the ontology-wave) have made their journeys to join the semantic web mainstream. We use "LOD KOS" as an umbrella term to refer to all of the value vocabularies and lightweight ontologies within the semantic web framework. We provide an overview of what the LOD KOS movement has brought to various communities and users. These are not limited to the colonies of the value vocabulary constructors and providers, nor the catalogers and indexers who have a long history of applying the vocabularies to their products. The LOD dataset producers and LOD service providers, the information architects and interface designers, and researchers in sciences and humanities, are also direct beneficiaries of LOD KOS. We examine a set of the collected cases (experimental or in real applications) and aim to find the usages of LOD KOS in order to share the practices and ideas among communities and users. Through the viewpoints of a number of different user groups, the functions of LOD KOS are examined from multiple dimensions. We focuss on the LOD dataset producers, vocabulary producers, and researchers as end-users.

1.0 Introduction

Conventional knowledge organization systems (KOSs—including thesauri, classification schemes, taxonomies, subject heading systems, name authorities, and lists of codes and terms, produced before the arrival of the ontology-wave) have always been quick adapters of new technologies in their publishing venues and applications. They have had timely appearances in the earliest indexing and abstracting (I&A) databases, online information services, CD-ROMs, Adobe PDF files, HTML websites, and XML databases since the 1950s. Recently they have made their journeys to join the semantic web mainstream and turned their products into Linked Open Data (LOD) datasets, along with ontologies that have been developed in the 21st century.

The Simple Knowledge Organization System (SKOS) and SKOS eXtension for Labels (SKOS-XL) became formal W3C recommendations in 2009, as a separate, lightweight, intuitive language for developing and sharing new KOSs. SKOS may be used on its own,

[§] This chapter is a slightly shortened and updated version of the paper with the same title published 2019 in *International Journal on Digital Libraries* 20:209–30. We want to thank all reviewers for their positive and constructive comments that helped to improve this paper. In addition, we thank all our co-organizers of former NKOS workshops and all participants of NKOS-related events for their continuous input and feedback that motivated us to write this paper.

or in combination with formal knowledge representation languages such as the Web Ontology Language (OWL) (W3C 2009). Eight years later, by the end of 2017, there were over 1,000 valid LOD KOS datasets registered in the DataHub (https://old.datahub.io/dataset), while many LOD KOS services also existed. The KOS products that have adopted the LOD approach using the standardized data model syntax recommended by SKOS and OWL can be found in a variety of domains and formats, from general-purpose to specialized domains, from mono-lingual to multilingual, from classification systems, thesauri and taxonomies to name-authority files, from extracted portions or a complete version of an original vocabulary to the end-products that are made from multiple vocabularies. The release of a LOD KOS product represents a turning point for the producer or provider of a vocabulary; but what are the results?

We aim to explore what the LOD KOS movement has brought to various communities and users. These are not limited to the colonies of the KOS constructors and providers, nor the catalogers and indexers who have a long history of applying the vocabularies to their products. Across domains, languages and places, the LOD dataset creators and LOD service providers, the information architects and interface designers, and researchers in sciences and humanities are also direct beneficiaries of LOD KOS. After a brief explanation of the term LOD KOS, the features of LOD KOS, and the services providing them (in Section 2 Background), we list the resources used to collect the cases and to cluster user groups based on personas (in Section 3 Methods) which are used to deliver the findings in the main body of the paper. Section 4 "Preliminary Findings" is divided into three subsections around three groups: LOD dataset producers; vocabulary producers who are involved in the development and enrichment of KOSs, and researchers who are the end-users of KOSs. Summaries were given to each of these sub-sections as well as at the end of the paper.

2.0 Background

2.1 Explanation of the term "LOD KOS"

Using the terminology of the LOD communities, KOSs are used as "value vocabularies," which are distinguished from the "property vocabularies" like metadata element sets. This term refers to its usage in the RDF-based models where the "*resource, property-type, property-value"* triples benefit from a controlled list of allowed values for an element in structured data. A value vocabulary defines resources (such as instances of topics, art styles, or named entities) that are used as values for elements in metadata records. Examples include: thesauri, code lists, term lists, classification schemes, subject heading lists, taxonomies, authority files, digital gazetteers, concept schemes and other types of KOSs (Isaac et al. 2011). It is important to remember, however, that a KOS vocabulary is more than just the source of values to be used in metadata descriptions: by modeling the underlying semantic structures of domains, KOSs act as semantic road maps and make possible a common orientation by indexers and future users, whether human or machine (Tudhope and Koch 2004; for recent special issues on NKOS see Mayr et al. 2016, Golub, Schmiede and Tudhope 2019, and Busch and Tudhope 2020).

Another notable term, "light-weight ontologies" refers to those using ontological classes and properties to express the conventional KOS. This is popular among those publishing a thesaurus with an ontology model beside SKOS. Usually they are not considered as "reference ontologies" that have rich and axiomatic theories with the focus on clarifying

the intended meanings of terms used in specific domains. In this context, lightweight ontologies are regarded as "application ontologies" which provide a minimal terminological structure to fit the needs of a specific community (Borge, Guarino and Masolo 1996; Menzel 2003). Yet the term "ontologies" has been applied to various types of vocabularies, while the approaches such as upper ontologies and hybrid ontologies have been widely applied in generating new KOSs.

In this paper, we will use "LOD KOS" as an umbrella term to refer to all of the value vocabularies and lightweight ontologies within the semantic web framework. When individual value vocabularies and lightweight ontologies are referenced, the term "vocabulary" or "vocabularies" might be used.

2.2 Features of a LOD KOS vocabulary

A LOD KOS vocabulary must follow the principles of Linked Data (Berners-Lee 2006) and must be openly available. The SKOS data model views a knowledge organization system as a "concept scheme" comprising a set of "concepts" (W3C 2009), where each concept must be named by a URI (Uniform Resource Identifier) or IRI (Internationalized Resource Identifier). Using a unique identifier to represent an entity or resource is one of the basic solutions for providing machine-processable disambiguated data. Furthermore, HTTP URIs should be used when releasing a dataset as LOD.

Data of a LOD KOS are expressed as RDF triples and may be encoded using any concrete RDF syntax such as RDF/XML, Turtle, TriG, N-Quads and JSON-LD, allowing the data to be passed between computer applications in an interoperable way, enabling a KOS to be used in distributed, decentralized metadata applications.

A LOD KOS end-product may be available as an RDF data-dump or accessed through a SPARQL endpoint. Templates for forming SPARQL queries, visualized relationships, on-the-fly mapping/matching services, and other innovative delivery methods may also enrich the presence of LOD KOS on the Web.

2.3 LOD KOS vocabulary services

The LOD KOS vocabularies are served by dedicated services. It should be noted that for KOS products, the consistency and synchronization between the original databases and the RDF stores are required. Otherwise, if a KOS's LOD version is not updated when the original data source is updated, then the quality of that product becomes questionable. The following are representatives of widely used, well-maintained service providers (SP). They have developed strategies and technologies to ensure not only the availability but also the interoperability, stability, and scalability of the contents and applications they provide.

Those services that host full content of a KOS vocabulary as well as the management data for each component updated on time are also known as vocabulary repositories. The natural languages involved could be monolingual or multi-lingual; the number of KOS vocabularies contained in a repository could range from a single one to more than 500. A dedicated portal would provide a unified point of access for KOS vocabularies hosted by a vocabulary service. Some of the services only provide the most current version of a vocabulary, while some maintain all versions. Additional functions might be available in addition to searching, browsing, displaying, and navigating. Some of them also align among vocabularies or provide direct links of data values. The following information of the service providers (SP) was collected before January 1st, 2018 and updated on May 1st, 2020.

SP-1. Individual vocabulary's provider.

•E.g., EuroVoc (http://eurovoc.europa.eu/), the multilingual thesaurus of the European Union (EU). Terms in EU languages and alignments with eight other KOSs are available on website and dump.

SP-2. Individual institution as the provider of all vocabularies produced in the institution.

•E.g., Library of Congress Linked Data Services – Authorities and Vocabularies (http://id.loc.gov/) provides access to all vocabularies promulgated by the Library of Congress including the *Library of Congress Subject Headings*, Library of Congress *Classification*, and LC Name Authority File, plus the many smaller value vocabularies such as various code lists and schemas from the MARC documentation standard, preservation vocabularies, ISO language codes, and other standards.

•E.g., Getty LOD Vocab (http://vocab.getty.edu/) provides multiple Getty vocabularies, the *Art & Architecture Thesaurus* (*AAT*), the Getty Thesaurus of Geographic Names (TGN), and the Union List of Artist Names (ULAN), through both data dump and a SPARQL endpoint, plus a comprehensive list of query templates and documentation. The contents are directly linked to the website of the vocabularies. The Cultural Objects Name Authority (CONA) is on its way to becoming LOD.

SP-3. Unified portal for a country's KOS vocabularies produced by multiple units in the country.

•E.g., The Finnish thesaurus and ontology service FINTO (http://finto.fi/en/) enables both the publication and browsing of dozens of vocabularies produced in Finland. In addition, the service offers interfaces for integrating the thesauri and ontologies into other applications and systems.

SP-4. Domain-oriented portal for collected vocabularies produced by multiple units.

•E.g., BioPortal (www.bioportal.bioontology.org) provides a Web portal enabling biomedical researchers to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice (nearly 690). Among the extra features are the mapping among the involved vocabularies, the usage data, and reviews.

•Other examples are: Ontobee (http://www.ontobee.org/) (biomedical); Planteome (http://planteome.org/) (plants); Ontology Lookup Service (OLS http://www.ebi.ac.uk/ols/index) (biomedical); GFBio terminology service (https://terminologies.gfbio.org/) (biological), and Heritage Data (http://www.heritagedata.org/blog/vocabularies-provided/)(cultural heritage).

SP-5. Middleware that provides tools for end-users to use/reuse published vocabularies.

•E.g., Skosprovider[1] provides an interface that can be included in an application to allow it to talk to different SKOS vocabularies. These vocabularies could be defined locally or accessed remotely through web services, for example, for the Getty vocabularies and the vocabularies published by EH, RCAHMS and RCAHMW at heritagedata.org.

SP-6. Upper ontology that facilitates multiple vocabularies' concept- and entity-mapping.

•E.g., Linked Open Ontology cloud KOKO[2] supports the managing and publishing of a set of interlinked Finnish core vocabularies; enables the users to use multiple ontologies as a single, interoperable, cross-domain representation instead of individual ontologies.

•E.g., Upper Mapping and Binding Exchange Layer (UMBEL) provided an UMBEL vocabulary (until October 2019) that was designed for mapping ontologies and external vocabularies (OpenCyc, DBpedia, PROTON, GeoNames, and schema.org), and provided linkages to more than 2 million Wikipedia entities.

Vocabulary registries are different from repositories because they offer information about vocabularies (i.e., metadata) instead of the vocabulary contents themselves; they are the fundamental services for locating KOS products. The metadata for vocabularies usually contain both the descriptive contents and the management and provenance information. The registry may provide the data about the reuse of ontological classes and properties among the vocabularies.

SP-7. Registry of KOS.

•E.g., BARTOC[3] (Basel Register of Thesauri, Ontologies & Classifications) currently has metadata about over 3000 KOSs in the registry, including active, inactive or historical vocabularies. Hundreds of these are available in RDF format. Furthermore, BARTOC includes the metadata of over 90 other registries.

SP-8. Registry of any vocabularies that are published with Semantic Web languages.

•E.g., LOV (Linked Open Vocabularies[4]) currently has over 600 registered vocabularies; all went through certain quality verification. Many of the vocabularies are property vocabularies. In addition to the descriptive metadata about a vocabulary, the usage metadata about properties' reuse among vocabu-

laries, the administrative metadata showing the most recent updates, and the technical metadata regarding the expressivity in terms of RDF, OWL, and RDFS are provided. The details of a vocabulary are exposed through statistics, such as the total number of classes, properties, data types, and instances.

SP-9. Registry of any LOD products, including KOSs.

•E.g., DataHub's previous version[5] (as of September 2017) was the largest registry, with over 11,273 datasets registered. Searching for various KOS types resulted in over 1,000, after verification by the authors of the paper in 2017 and 2019 (Zeng and Clunis 2020).

3.0 Methods

3.1 Sources of the study

We examined cases collected from various sources, including released LOD KOS products, journal articles, conference presentations, workshops and webinars, related tweets, blogs and posts in community-shared spaces. These sources have certain special characteristics worth mentioning here. First, many of the LOD activities are experiments, done outside of the vocabulary creator and indexer circles. Second, in cases where efforts have been initiated by and involve KOS providers, the implementation may take time to be tested, improved and officially added to the workflow. These cases are usually shared within communities and informal groups, especially at the beginning stage of the LOD products life cycle. They are most likely to be publicized through conference presentations, demos, posters and un-conference sessions, while a smaller number of formal publications appear in journals. Thus, the sources of this research are unconventional and include:

- •Sessions of KOSs at international conferences
- •Research-based journal publications
- •Theses and dissertations
- •Professional conferences and summits
- •NKOS workshops (archived at http://nkos.slis.kent.edu)
- •the NKOS bibliography project[6]

Other sources where cases were discovered include:

- •LOV[7] on Google+
- •Getty Vocab Google Group[8]
- •Getty Share[9]
- •Social media sources: tweets, blogs, Facebook groups
- •Ontolog-Forum[10]
- •LODLAM[11] challenges and un-conference-style sessions
- •GitHub entries such as OpenSKOS, NatLibFi/Skosmos, JSKOS and more.

3.2 User personas developed for communicating the preliminary findings

In an effort to classify the ideas and outcomes related to LOD KOS reported in the sources listed above, we first created personas representing typical user groups of LOD KOS in order to build a common understanding of their needs and the goals they wish to achieve. Rather than a top-down approach to collect the definitions of certain user groups, we took a bottom-up approach to group the personas that are defined through the project. Although fictional, a persona is a realistic description of a typical or target user of a product, highlighting specific details and important features of a user group. Personas have been widely used in user experience design tasks. They are user models synthesized from real-world observations and are used to incite emphatic thinking when developing a system. It is a process in which data are summarized, clustered and analyzed to discover themes, the results of which are then used to create outlines or "skeletons" of individual users that can be used for planning, design, and development (Pruitt and Adlin 2010, 156).

Proto-personas are a modification on traditional personas with the difference that they are not synthesized from data collected from interviews of users. Instead, they originate from brainstorming workshops where company participants try to encapsulate the organization's beliefs (based on their domain expertise and gut feeling) about who is using their product or service and what is motivating them to do so (Gothelf 2012; D'Amore 2016). Proto-personas can be utilized to prevent the design team from viewing themselves as the intended users, and to help guide them to create a system suitable for their intended users or user groups (Buley 2013, 132-35; Krøger, Guribye and Gjøsæter 2015).

We took the approach of proto-persona development based on our literature review and use case studies (using the data sources described above) along with user behavior observations and brainstorming working group meetings. A number of informal interviews were also conducted. We focused on the first tier of persona development defined by Dan Brown (2000): 1) requirements, 2) relationships and 3) humanization. The result is a set of personas encapsulating our understanding of who are using the LOD KOS products or services and what has motivated them to do so. A persona group, e.g., Vocabulary Producer (VP), contains multiple personas such as VP1, VP2, VP3, etc.; they are highlighting different roles of the VP group, might take in one or more projects, or in the same project over time. Among the five groups, the first three will be used in this paper:

- LOD Dataset Producer (DP) group
- Vocabulary Producer (VP) group
- Researcher (RS) group (as end-users)
- Website/Tool Developer (WD) group
- KOS Service Provider (SP) group

The formation of personas follows common practice in that they are very brief, typically bulleted lists of distinguishing data ranges for each subcategory of a user (Pruitt and Adlin 2010, 184). The resulting personas are intentionally simple and depict:

(a) who the group is, including the name and identity key of fictional characters;
(b) what are the sources of characters;
(c) which tasks they usually have;
(d) what are the contents they deal with;
(e) where and how they interact with the KOS vocabularies; and,
(f) what are the goals.

(See Appendix A for one example of the Vocabulary Producer (VP) persona document.) We consider these "skeletons" of the personas to be living documents that support this particular research, which uses unconventional data resources, while allowing the profiles to be further refined, split into narrower personas, and encompass more personas as new details are discovered at any time. They are used to provide a central point to enable us to communicate the preliminary findings and to share the cases around LOD KOS.

4.0 Preliminary findings

Through the viewpoints of different personas designed in this study, functional changes and other changes of KOS after they were released as LOD are examined from multiple dimensions. The following sections are organized around personas representing typical users of LOD KOS. Even though some specific cases are used as examples, the attention is on summarizing the general issues and benchmarks identified in this study. Best practices acknowledged by communities as well as experimental approaches are presented together with the possible challenges and hurdles.

4.1 For LOD Dataset Producers (DPs), LOD KOS vocabularies enable their data to become 4-star and 5-star LOD

In this part, the preliminary findings are presented for LOD dataset producers (DPs) facing different levels of situations when producing LOD products:

> 1) creating LOD datasets from scratch and dealing with data that have no controlled values for the named entities and topics; 2) reaching out to the datasets that might have or have not been using community standard vocabularies in their structured data; and 3) turning the existing datasets that have been using value vocabularies into 4-star and 5-star LOD.

Before looking into this section, it is necessary to revisit Tim Berners-Lee's 5-star Open Data Scheme for LOD data (Berners-Lee 2006).

★	Available on the web (whatever format) but with an open license, to be Open Data
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus, Link your data to other people's data to provide context

Among the datasets found in the Datahub, which mostly qualify to be 4-star, only about 10% were included in the LOD Cloud 2017-02 version as recognizable 5-star datasets. One of the reasons is that "the dataset is not interlinked with other datasets" (Linked Open Data Cloud 2017). The LOD KOS vocabularies are primary sources which enable datasets to become 4-star and 5-star Linked Open Data. This benefit has become the most widely acknowledged by the LOD dataset producers.

The LOD dataset producers are dedicated to exploiting existing data and delivering structured data in the RDF format. They might be dealing with already structured data such as bibliographic records, museum documentation files, clinical trial databases, etc. More often, they would make structured data out of unstructured raw data such as oral history transcripts. In order to break the silos and connect with the rich information outside of their silo boundaries, many of them took the LD approach and opened up. The linking points are primarily the concepts and named entities, i.e., the identifiable things including people, organizations, places, events, objects, concepts and virtually anything that can be represented in structured data (see a recent example in Binding and Tudhope 2016). In the RDF triples (*subject-predicate-object*), they occupy the positions of *subject* and *object*.

Nevertheless, for a dataset to become real LOD, identified entities need to be named with URIs. This is usually the first hurdle to overcome. Thus, using LOD KOS has become a best practice and popular strategy for the LOD dataset producers. Depending on the situation (see Figure 1), the usage of LOD KOS might involve multiple choices and steps.

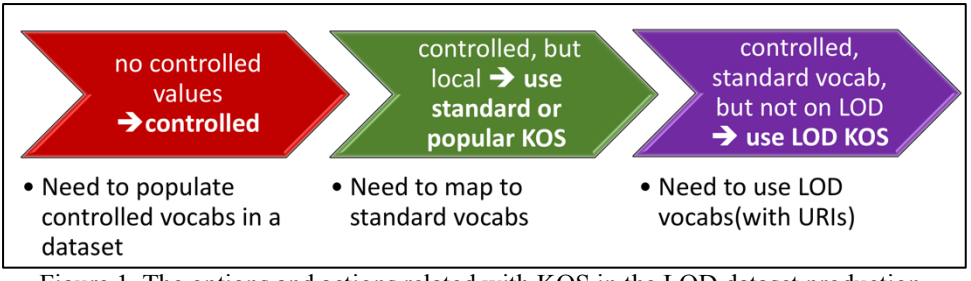


Figure 1. The options and actions related with KOS in the LOD dataset production

DP-1. Dealing with semi-structured and unstructured data that have no controlled values for the named entities and topics in order to create LOD datasets from scratch

Dataset producer DP-1 is dealing with semi-structured and unstructured data that have no controlled values for the named entities and concepts and wants to create a LOD dataset from scratch. The examples of such kinds of data include: the digitized materials (textual or non-textual) hosted in silos; archival finding aids; oral history transcripts; merged local files and others. Technologies exist to help mining data and extracting the entities. However, there are many possible issues to be faced from the beginning. Examples among those involving place related entities are:

- •Place names change through time (e.g., "Bombay" and "Mumbai");
- •Alternate names exist (e.g., "New York City," "NYC," and multilingual labels);
- •Same name is associated with multiple locations (e.g., various "St. Petersburg" in the world);
- •Unidentifiable places (e.g., places referred in a creative work but has not been identified);
- •Unnamed places;
- •Cartographic versus geographical placement; and
- •Feature typing/categorizing results are incorrect or inconsistent.

In the effort to identify and control the named entities and concepts from these semi-structured and unstructured data, and advance from digitization to datafication, these major benchmarks are to be reached before becoming 4-star data:

1. Identify the entities;
2. Put the entities into structured data;
3. Clean up the newly structured data, with local control;
4. Encode the entities with standardized KOS vocabularies (as strings);
5. Obtain URIs for entities provided by the LOD KOS datasets; and
6. Use http URIs for names of any entities.

The last three are related to LOD KOS use, in order to have high quality and trustable linkages in the RDF triples.

A well-known pioneer case is Linked Jazz[12], which concentrates on a special collection of Jazz musicians' interviews (Pattuelli 2012; Pattuelli, Provo and Thorsen 2015; Pattuelli 2021, chapter 8 in this book). Based on the data about individual musicians, the team made connections between people. Step 1 was to get the names from the transcripts and establish a name authority file with URIs. A natural language processing tool pulled entities from the transcripts of interviews with jazz musicians that mention a relationship with another jazz musician. After the process of controlling synonyms and eliminating ambiguity, the musician names were mapped to name authorities in the Virtual International Authority File (VIAF), LC Name Authority, and DBpedia, and the data about each person were obtained. If a name was not in the name authority, the team established the authority record for the person. Step 2 was to find the names in all relevant documents in the collection based on the established name authority file. Step 3 was to describe the relationships using a relationship ontology the team developed. Finally, a visualization tool was used to present a unique interactive interface.

DP-2. Reaching out to the datasets that may have or have not been using community standard vocabularies in their structured data

An effort that needs to integrate distributed data sources from outside institutions most likely will face the issues of standardization or unification on data models and value vocabularies. In this situation, the dataset producer DP-2 intends to reach out to the datasets

that might or might not have been using community standard vocabularies in their structured data. One of the key tasks involves the conversion of existing KOSs into LOD before applying them as standard value vocabularies in all datasets to be integrated.

An example of such a situation was reported by the project of Archives of France (Sibille-de Grimoüard 2014). The *Thesaurus W. Standardized Vocabularies for Describing and Indexing Local Administration Records*, developed in 1987, has been used by the French archival agencies to index descriptions of modern records created by local public services. The thesaurus and three controlled lists of terms were available as Excel sheets and PDF files on the Internet till 2008. The ability to interact with the applications used by local archival institutions would need machine-readable and machine-processable KOSs. The following needs were identified when the project initiated the LOD activity:

- •Represent the thesaurus in a machine-understandable way for automating machine-assisted indexing processes;
- •Facilitate its integration into retrieval tools;
- •Ensure the consistency of indexing even though the thesaurus evolves;
- •Facilitate the process of updating and maintaining the thesaurus (evaluating the requests for changes from users, updating terms and relationships, amending terms, customizing the display of terms, etc.);
- •Express all the concepts already represented in the thesaurus (concepts and terms, relationships between these concepts, annotations, etc.); and
- •Use standards and models related to thesauri and controlled vocabularies for interoperability purposes.

This is a very well summarized list of tasks and reflects the needed benchmarks of many projects that may deal with local and distributed sources of data. Even though the thesaurus was not considered fully compliant with ISO-25964 (2011, 2013) as a "thesaurus," the SKOSified KOSs enabled the dataset producers to reach the stated goals. This project was also an opportunity to align data with other LOD KOS and resources (e.g., RAMEAU and DBpedia) and to implement a solution for persistent identifiers of concepts of the thesaurus. Among the advantages for users were that the shared use of common vocabularies creates interoperability without any additional developments. For instance, as the thesaurus for indexing local archives provides links to RAMEAU, it would be possible to link an archival resource and a library book through these two thesauri and the links they share. A similar example of converting a thesaurus into SKOS in the Social Sciences was reported by Zapilko et al. (2013).

DP-3. Having datasets that have been using value vocabularies in structured data, turning them into 4 star and 5-star LOD

Dataset producer DP-3's objective is to turn the existing datasets into 4-star and 5-star LOD. These datasets have been using (born-with or mapped-to) value vocabularies in their structured data. Examples of such data include the national bibliographies, catalogs, special collection portals, metadata repositories, and many theme-based LOD products made in projects. A new dataset's resource may be maintained by different information systems based on traditional relational data models. In such a situation, a dataset usually has controlled the named entities and topics with KOS vocabularies. The following benchmarks are expected before becoming 4-star and 5-star data:

1. Use standardized protocols for metadata structure;
2. Enrich the original metadata, especially for those semi-structured and non-controlled fields;
3. Control the value spaces for all entities;
4. Encode the entities with standardized KOS vocabularies (as strings);
5. Use URIs for names of entities; and,
6. Use http URIs for names of any entities.

The fifth and sixth benchmarks require that the KOS vocabularies being used are LOD datasets themselves. Fortunately, most of the standardized KOS vocabularies have become LOD KOS. Otherwise the last two benchmarks might not be reachable. Nevertheless, there might be many possible issues for each of the datasets currently at the 3-star level, as summarized below:

- If it has used local controlled vocabularies, the terms used or the form representing the concepts and named entities may be different from standardized controlled vocabularies.
- If it has used pre-LOD vocabulary, there might be no URIs/IRIs yet. How to obtain the URIs/IRIs to replace the strings of a named entity or concept?
- If a decision of mapping is made, which vocabulary and how many vocabularies will be involved, since in a subject domain and a community there could be more than one standard vocabulary.
- If it needs to map the local controlled lists to a standardized LOD KOS (e.g., LCSH, EUROVOC, etc.), human resources and quality control are most critical and could be challenging.
- For a dataset formed through aggregation, in addition to the above issues, synonyms and homographs occur in the data provided by different sources. Heavy disambiguation and semantic conflict controls are needed.

There are no black-and-white answers to these questions. Many dataset producers developed their own successful products, such as the national bibliographical databases, OCLC's WorldCat, and many others that used various KOSs to become 5-star LOD data (Figure 2).

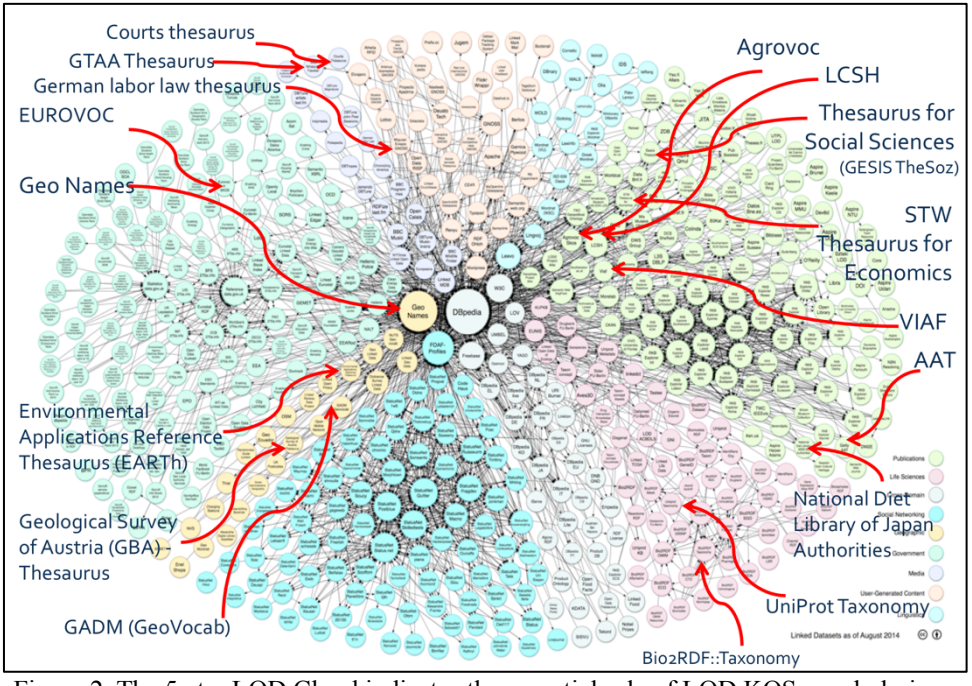


Figure 2. The 5-star LOD Cloud indicates the essential role of LOD KOS vocabularies. (Source: Annotated by the author on the LOD CLOUD 2014-08-30 image http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.png)

Summary of usages and practices (Dataset Producers)

It is clear that LOD KOS vocabularies, as the source of http URIs/IRIs for named entities and concepts used in data-transformation, enable the dataset producers to make 4-star or 5-star Open Data. In the bibliographic universe, they help the conversion from "library entities" to the "web of data" (Wallis 2014). The possibilities for linkage of high quality structured data become limitless and show the impact in the increased availability of information.

LOD KOS vocabularies empower the owners of data to convert and publish their data under the LOD principles, with high quality and trustworthy linkages in RDF triples. LOD KOS can be used to transform anyone's database into LOD datasets, even reaching 4- and 5-stars; to create machine-understandable and machine-processable data for any users, machine or human. We all understand that the creation of a KOS vocabulary involves tremendous intellectual efforts and human resources, thus, the openly available, well-established, and constantly-maintained vocabularies are invaluable engines for the LOD datasets.

4.2 For Vocabulary Producers (VPs) who are involved in the development and enrichment of KOS, LOD approaches lead to unconventional processes and results

The goals of vocabulary producers (VPs) include creating needed value vocabularies for their datasets, while also aiming at sharing the products with communities. The tasks of development and enrichment of new or existing KOS vocabularies are closely related to what LOD dataset producers usually encounter, as presented in the previous section of this paper. The value vocabulary producers to be discussed in this sub-section are considered to be different from the usual vocabulary providers such as those working for a thesaurus or classification system as editors. The following cases are presented based on five objectives:

1) creating new value vocabularies for particular project's products by extracting the components from a comprehensive KOS vocabulary;
2) creating a unified scheme for a domain based on multiple KOS vocabularies;
3) creating a heterogeneous meta-vocabulary;
4) enriching the KOS-at-hand and connecting to real things; and,
5) enhancing semantic consistency of data through shared, unconventional mashup KO activities.

Despite the fact that the presented cases resemble the approaches used in the KOS community for a long time, new methods, functions and results are observed in current approaches.

VP-1. Creating new value vocabularies for particular project's products by extracting the components from a comprehensive KOS vocabulary

When a particular project does not need to apply a full standard thesaurus, or when one existing thesaurus is not enough for the project's domain coverage, extracting components from standardized KOS vocabularies can be a relevant strategy. For example, the Government of Canada's Department of Canadian Heritage—Canadian Heritage Information Network (CHIN)'s *CHIN Guide to Museum Standards* (last updated 2019-07) provides a list of vocabularies of the terminologies for object naming; materials and techniques; disciplines; and styles, periods and cultures. Each of these terminologies can be a portion of the

Art and Architecture Thesaurus (*AAT*). A new vocabulary can be considered to be a microthesaurus, which is a designated subset of a thesaurus that is capable of functioning as a complete thesaurus (ISO25964-2:2013).

Vocabulary producer VP-1 is committed to creating new value vocabularies for particular project's products by extracting the components from a comprehensive KOS vocabulary. Whether VP-1 is extracting a whole facet from *AAT* (e.g., *Object Facet*), a sub-category under a guided term (e.g., < *Object genres by function*>), or a specific group [(e.g., *ceremonial objects* or *vessels (containers)*], the creation of a microthesaurus, with all the components and their RDF triples and URIs, can be obtained by querying *AAT* through a SPARQL query endpoint, using a template already provided to trace data of "Descendants of a Given Parent." The dataset can be obtained in about two seconds after a query is submitted (Garcia, Zeng and Ward 2017; Zeng 2017) (see Figures 3 and 4).

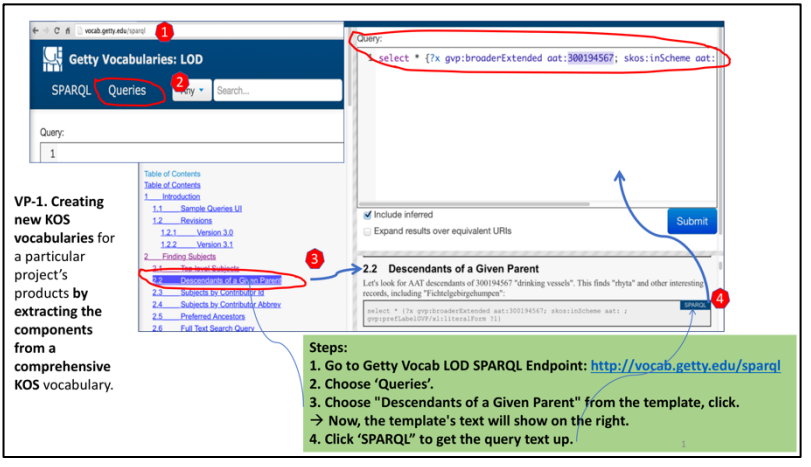


Figure 3. Using a template provided to trace data of "Descendants of a Given Parent" for "<costume by function>" (AAT concept ID 300212133). (Source: http://vocab.getty.edu/queries#Descendants_of_a_Given_Parent)

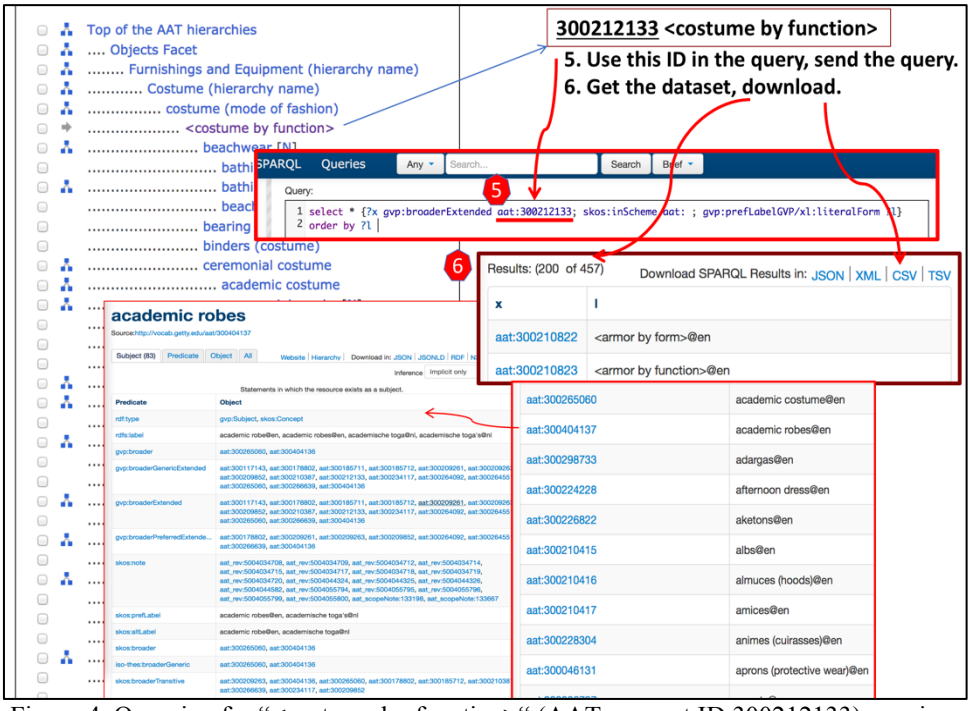


Figure 4. Querying for "<costume by function>" (AAT concept ID 300212133), receiving and downloading the datasets to make a microthesaurus.

Based on the sources of the study, especially at the Q&A portion of the conference sessions and community shared spaces, some vocabulary producers expressed concerns regarding limited knowledge of the new semantic technologies such as: #1, dealing with SPARQL queries and using the endpoints; and #2, handling the vocabularies in RDF formats. Aiming at the #1 concern, some middleware (e.g., Skosprovider) provide tools for end-users to use/reuse published vocabularies. Other LOD KOS service providers (SPs) mentioned in Section 2.3 of this paper also provide various tools for constructing, reusing and enhancing vocabularies. The cases to be discussed in Section 4.3 for researcher (RS) end users might be the best solutions to help these vocabulary producers. The United Nations' UNESCO Vocabularies SPARQL Service[13] provides over 100 microthesauri using user-friendly query templates. The #2 concern regarding handling the vocabularies in RDF formats is common, since a VP would need to organize and edit the selected concepts and terms before finalizing a set of entries to form a needed vocabulary. To solve this issue, the LOD KOS services usually offer multiple downloading formats to be selected by an end-user. One of the commonly used non-proprietary formats is CSV, a comma-separated values (CSV) format. A CSV file stores tabular data (numbers and text) in plain text, which allows a user to open the file from a spreadsheet to work on it directly. CSV is also the preferred form for visualization tools such as Google Fusion Tables and Gephi.

VP-2. Creating a unified scheme for a domain based on multiple KOS vocabularies
The Semantic Web encourages the sharing and reuse of data, including the components of KOS vocabularies. The query example shown above for VP-1 is applicable when obtaining any components of a LOD KOS vocabulary. It is also practical and common to form a new vocabulary based on more than one source, as the vocabulary producer VP-2 is engaged. The following cases demonstrate innovative approaches and results.

Thesaurus of Plant characteristics (*TOP*) was committed to the harmonization and formalization of concepts for plant characteristics widely used in ecology. It was built on previous initiatives and vocabularies for several aspects, including its model, entities and qualities, and concept definitions. *TOP* included names, definitions, formal units and synonyms for more than 700 plant characteristics (Garnier et al. 2017).

Motivated by the notion that open data need common semantics for linking diverse information, the *Global Agricultural Concept Scheme* (*GACS*) project of Agrisemantics aims to create a shared concept scheme by integrating existing standard vocabularies in agriculture and environment (Baker et al. 2016a). Agrisemantics is an emerging community network of semantic assets relevant to agriculture and food security. *GACS* functions as a multilingual thesaurus hub that includes interoperable concepts related to agriculture from several large KOSs: *AGROVOC* of the Food and Agriculture Organization of the UN, the *CAB Thesaurus* by CAB International of UK, and the U.S. *National Agricultural Library (NAL) Thesaurus*, all maintained by different institutions. *GACS* would facilitate search across databases, thereby improving the semantic reach of their databases by supporting queries that freely draw on terms from any mapped thesaurus, and achieving economies of scale from joint maintenance. The latest *GACS* beta version provides mappings for 15,000 concepts and over 350,000 terms in 28 languages as of its May 2016 release (Baker et al. 2016a). The case reveals unique processes and designs:

1) The mappings focused on three sets of frequently used concepts (10,000) from each of the three partners.
2) Mappings were automatically extracted and then manually evaluated by experts through discussions and manually corrected.
3) A classification scheme that was developed jointly in the 1990s was revised to tag concepts by thematic group (chemical, geographical, organisms, products, or topics).
4) Alongside generic thesaurus relations to broader, narrower, and related concepts, organisms will be related to relevant products.

Around the world, activities of creating a unified scheme for a domain, focusing on generating multilingual labels by using SKOS-XL, have proven successful, as reported by many other cases.

VP-3. Creating a heterogeneous meta-vocabulary
Vocabulary producer VP-3's task is similar to VP-2's task discussed above in generating a product based on multiple existing vocabularies. However, the situation involves creating a heterogeneous meta-vocabulary that supports the representation of changes and differing opinions of certain concepts. The case used here is a taxonomic meta-ontology *TaxMeOn,* built by Tuominen, Laurenne, and Hyvönen (2011). *TaxMeOn*[14] is an ontology schema for biological names, containing 12 ontological classes with 49 subclasses. The datasets utilized in the study consist of 20 published species checklists that cover mainly northern European mammals, birds and several groups of insects, resulting in about 78,000 taxon names. The difference between *TaxMeOn* and the cases shared with VP-2 is

that the representation of the dataset encompasses these contents: 1) the different conceptions of a taxon, 2) the temporal order of the changes, and 3) the references to scientific publications whose results justify these changes. The rationale is that the positions of species and the nomenclature in scientific taxonomies involve a lot of changes, which directly impacts the access to the publications and data associated with them in different time periods.

The direct application of the taxon meta-ontology model that allows multilingual, different opinions for the biological taxonomy concept and nomenclature in a unified view can be beneficial to the researchers of biology. The detailed data can be further linked to other datasets with less taxonomic information, such as species checklists, and provide users with more precise information. The data model enables managing heterogeneous biological name collections and is not tied to a single database system (Tuominen, Laurenn and Hyvönen 2011). More importantly, this modeling method and the model itself can be extended in a flexible way and integrated with other data sources.

VP-4. Enriching the KOS-at-hand and connecting to real things

Vocabulary producer VP-4 has a SKOSified thesaurus at hand and is investigating how and when to link a concept in the thesaurus to the URIs provided by name authorities and Wikipedia so as to fully benefit from LOD and enrich an existing KOS-at-hand. Another question is how to take advantage of such processes to allow any organization to improve and expand the data with other relevant sources the organization does not own. For years, there have been discussions about whether name authorities should be maintained separately from concept-based subject heading lists, thesauri and classification schemes that also contain named entities. In the LD movement, there have been confusing and incorrect applications of *skos:exactMatch* and *owl:sameAs* to align "a real thing" (e.g., a person, institution, or place) to the concepts, names or photos that "represent the thing."

FAST[15] (Faceted Application of Subject Terminology), a joint vocabulary effort of OCLC and the Library of Congress, based on *LCSH*, reported using *foaf:focus* to allow FAST's controlled terms (representing instances of *skos:Concept*) to be connected to URIs that identify real-world entities specified at GeoNames and DBpedia. With the correct coding of properties, machines can understand (reason) that a FAST-controlled term is related to a real-world entity and allows humans to gather more information about the entity that is being described (O'Neill and Mixter 2013). As Schema.org[16] grows, classes and properties defined by it are also being applied to FAST. The enrichment allows FAST terms to take advantage of all of the various string values included in VIAF (containing dozens multilingual name authorities) without having to manually include the values in the RDF triples for the specific term in FAST. The DBpedia identifiers allow FAST terms to include detailed information that is usually excluded in authority records.

Bensmann, Zapilko and Mayr (2017) reported another large-scale interlinking project in Swissbib[17], a provider for bibliographic data in Switzerland. Data available in Marc21 XML were extracted from the Swissbib system and transformed into an RDF/XML representation. From approximately 21 million monolithic records, the author information was extracted and interlinked with authority files from the VIAF and DBpedia. A main obstacle was the amount of data and the necessity of day-to-day (partial) updates. As a result, the team has developed procedures for extracting and shaping the data into a more suitable form, e.g., data are reduced to the necessary properties and blocked (see Figure 5).

The approach could establish 30,773 links to DBpedia and 20,714 links to VIAF and both link sets show high precision values and could be generated in reasonable expenditures of time, according to the authors.

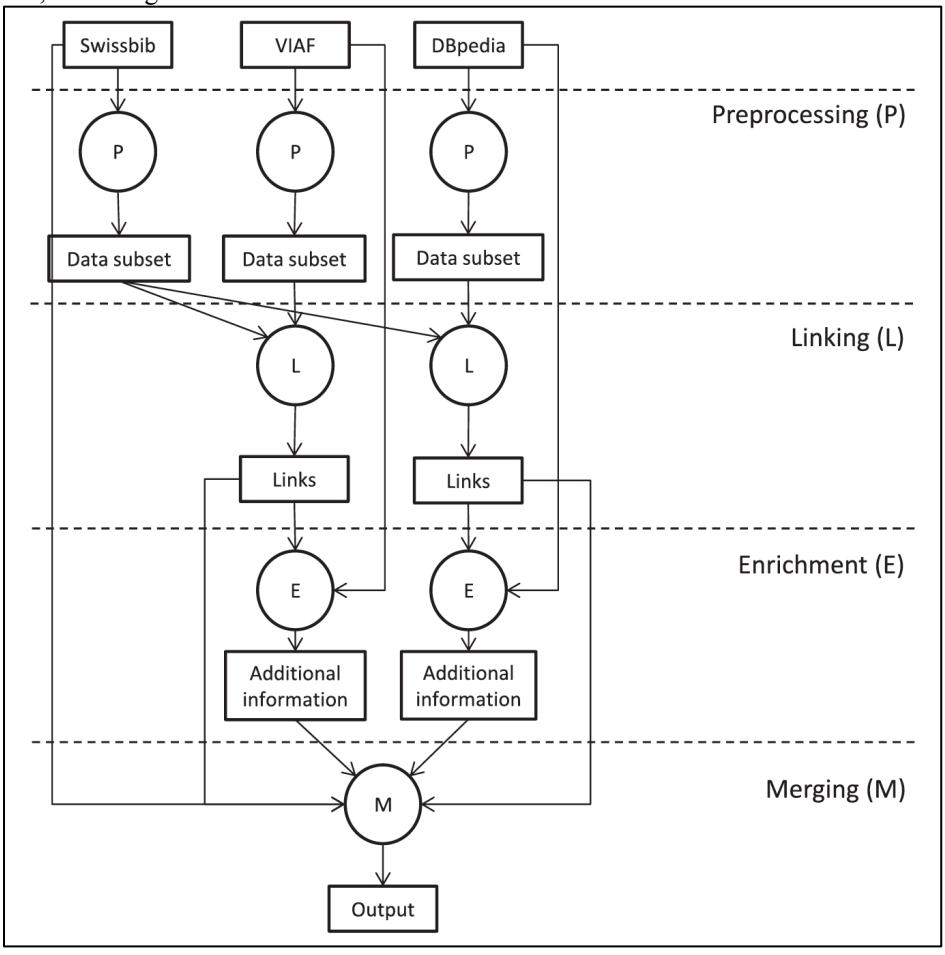


Figure 5. Data flow diagram of the interlinking procedure in the Swissbib project (Source: Bensmann et al. (2017, 8 Figure 4).

VP-5. Enhancing semantic consistency of data through shared, unconventional mashup KO activities

Vocabulary producer VP-5 is involved in the new efforts to enhance semantic consistency and interoperability through shared data which have already shown great potential for dataset producers and KOS vocabulary producers. In Web development, the term "mashup" denotes a combination of data or functionality from two or more external sources to create a new service. "Mashup culture" puts a cultural dimension into the foreground, as these developments permeate through almost all cultural techniques and practices on a global scale (Sonvilla-Weiss 2011).

The most obvious cases are the added authority identifiers and categories in *Wikipedia* entries. The "authority control" section has been added to many Wikipedia pages providing the identifiers from name authorities such as WorldCat Identities, VIAF, ISNI, ULAN, etc.

The more systematic activities can be found around *Wikidata* (https://www.wikidata.org), which functions as the authority files of named entities, but increasingly more abstract concepts have been added by volunteers. For instance, the Wikidata Visual Arts project, which intends to provide a knowledge base of reusable multilingual facts that can be used in Wikipedia and any other sites, provides the *Visual Arts Item Structures* as the guideline and classification for describing information related to visual arts. Each entry for an entity has a URI and the classes align with AAT (mostly the <Objects> facet), using AAT URIs as identifiers. Identifiers from other KOS and collections are also found for various concepts[18]. Similar projects can be found across Wikidata, Wikimedia, and other shared efforts.

Although the overall quality, coverage and mapping accuracy have not been systematically measured or proved, and the sustainability and consistency applied to each concept and named entity are not standardized, these unconventional, shared KO activities certainly provide a good reference source and quick access to LOD KOS products, filling up the gaps between currently existing KOS coverages and real world needs. The mash-up culture, a symptom of a wide paradigm shift in our engagement with information, seemed to be perfect for the data-driven cultural techniques and practices of knowledge organization (Voss 2013, Sonvilla-Weiss 2011, Bensmann, Zapilok and Mayr 2017).

Summary of usages and practices (Vocabulary Producers)

The cases presented so far for the Vocabulary Producers (VPs) seemed to resemble KOS methods developed prior to the 21st century. From conceptual and structural points of view, the newly generated vocabularies, derived from the existing ones, took similar approaches such as making microthesauri and satellite vocabularies, creating a super structure, direct mapping or employing a switching system, crowd-sourcing, post-control, etc.

The new functions and differences observed in current approaches are the results of applying LOD principles. Each "thing" included in all the new products is named with a URI, and has a domain name prefix that directly indicates its origin, thus, maintaining the original semantics and linguistic decisions while being reusable. The cases also benefit from semantic technologies and the available open tools. For example, the new microthesauri or satellite vocabularies can be generated through modifiable SPARQL queries and obtain datasets in a minute. The variety of downloadable formats available allows for easy integration with other data and visualization using open tools.

For vocabulary producers, the LOD KOS vocabularies are the resources for creating, maintaining, enriching, extending, and translating a value vocabulary that complies with LOD principles. The data-driven, shared editing and publishing workflow also facilitates the capture of administrative, provenance, and uses metadata for the whole vocabulary and its components. With an increasing number of KOSs published in standardized, machine-understandable formats, it becomes necessary for organizations to improve and expand the KOS data that they already have by using other relevant sources.

The most important achievement is the reusability of any of these new vocabularies in LOD or non-LOD databases. As the Agrisemantics project team determined (Baker et al. 2016b):

•Open-access semantics are easy to re-use;
•Mapping the semantics promotes cooperation and reduces duplication; and
•Coherent semantics benefit research, innovation systems, and value chains.

4.3 For researchers (RS) who are end-users of KOS, LOD KOS products can become knowledge bases and provide semantic-rich discoveries

It is very common that real end-users (i.e., those other than the creators and publishers of KOS products) may not be familiar with KOSs and may not be tech-savvy. The question of how to attract users and extend beneficiaries further than the dataset producer (DP) and vocabulary producer (VP) groups is a major challenge for the LOD KOS vocabulary service providers. Especially they seek to demonstrate the societal value of their efforts of converting KOSs into LOD format and providing services such as free data dumps and SPARQL endpoints (which may add extra costs). For this reason, they need users and supporters from all disciplines.

What is more, the scalability of LOD approaches in relation to KOSs must be addressed. The data dumps (which are the most popular for LOD KOS) and SPARQL endpoints seem not to be applicable for end-users whose jobs are not related to semantic technologies. Technologically, in addition to the access issues related to finding, browsing, and navigating within or across KOS vocabularies, the challenge arises as to how the LOD KOS can be used as more than traditional "controlled vocabularies" or can function as more than just being "value vocabularies" in the semantic web.

The cases collected in this section demonstrate some innovative ideas that could be followed as relevant approaches to enhance the LOD KOS usage. Note that in this section we are not discussing semantic search and content discovery in a database or a website that is enabled by using KOSs; here the cases are about the KOSs themselves. They illustrate how LOD KOS can be potentially useful to researchers among the end-users, as found in the following situations:

1) using well-developed KOS products, high quality and relevant knowledge bases are now easily available for researchers;
2) name authorities could offer foundational structured data for network analyses; and,
3) user-friendly displays of KOSs provide visually enriched understanding.

RS-1. Accessing and using KOS-based knowledge bases

Researcher RS-1 needs to access and obtain information resources that could help answering sophisticated questions through a user-friendly workflow and tool. RS-1 has little knowledge of RDF or SPARQL. Fortunately, a countable number of innovative LOD KOS providers have provided user-friendly templates for querying their LOD KOS data. From the following examples, it is clear that researcher RS-1 can use these templates to obtain special graphs or datasets for very complicated questions.

The first example is from the Universal Protein Resource (UniProt[19]), a comprehensive resource for protein sequence and annotation data (see Figure 6). Organisms are classified in a hierarchical tree structure. The taxonomy database contains every node (taxon) of the tree. Top nodes are "Archaea," "Bacteria," "Eukaryota" and "Viruses." The UniProtKB taxonomy data is manually curated: next to manually verified organism names, a

selection of external links, organism strains and viral host information are provided. Using the template, for example, in question #8, one can find all preferred gene name and disease annotation of all human UniProt entries that are known to be involved in a disease. This is much more complicated than the question #2, "Select all bacterial taxa, and their scientific name, from the UniProt taxonomy." In both cases, clicking on "show," will automatically load and make the query ready for use (see Figure 6).

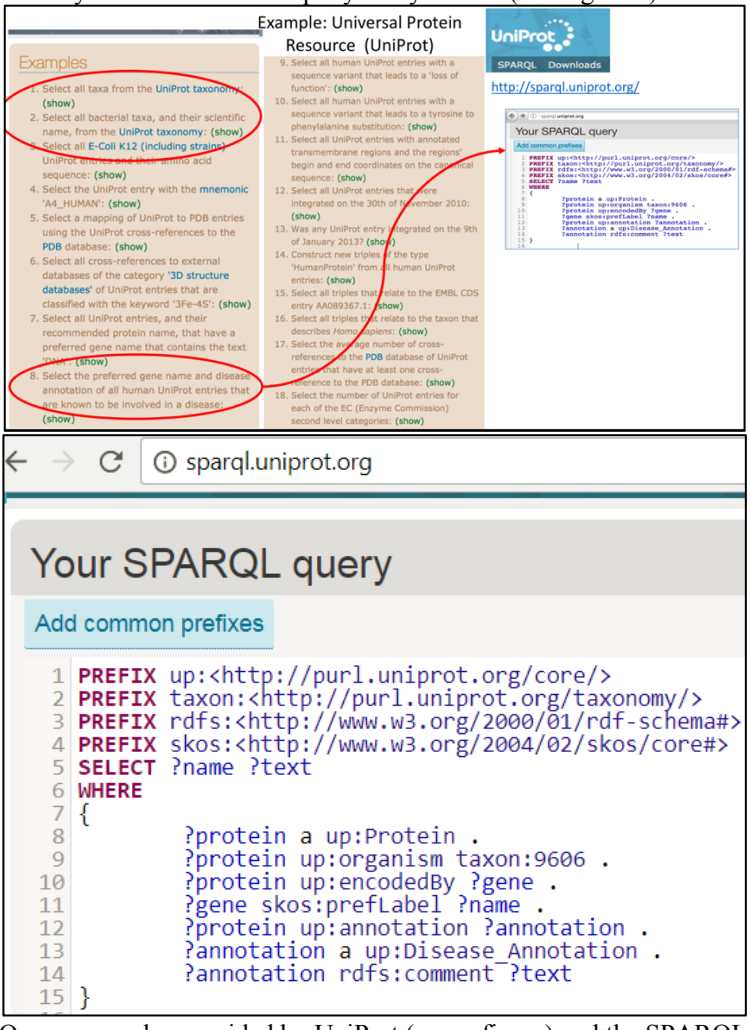


Figure 6. Query examples provided by UniProt (upper figure) and the SPARQL query for question #8, automatically "show"ed (lower figure) (Source: http://sparql.uniprot.org/).

One may argue that UniProt itself is a knowledge base, and the taxonomy is just used for organizing the information, raising the question as to whether a LOD KOS dataset itself could be considered to be a knowledge base. The next case is the *Getty Thesaurus of Geographic Names* (*TGN*) available through Getty Vocabulary LOD service[20]. The application has turned the thesaurus into a knowledge base. For example, in combination with the geographic boundary data that are available in *TGN*, queries #4.16 to #4.19 help to gather data such as places by, within or outside a coordinate bounding box, and even with further criteria such as filtering by place type and obtaining geo or column charts (see Figure 7).

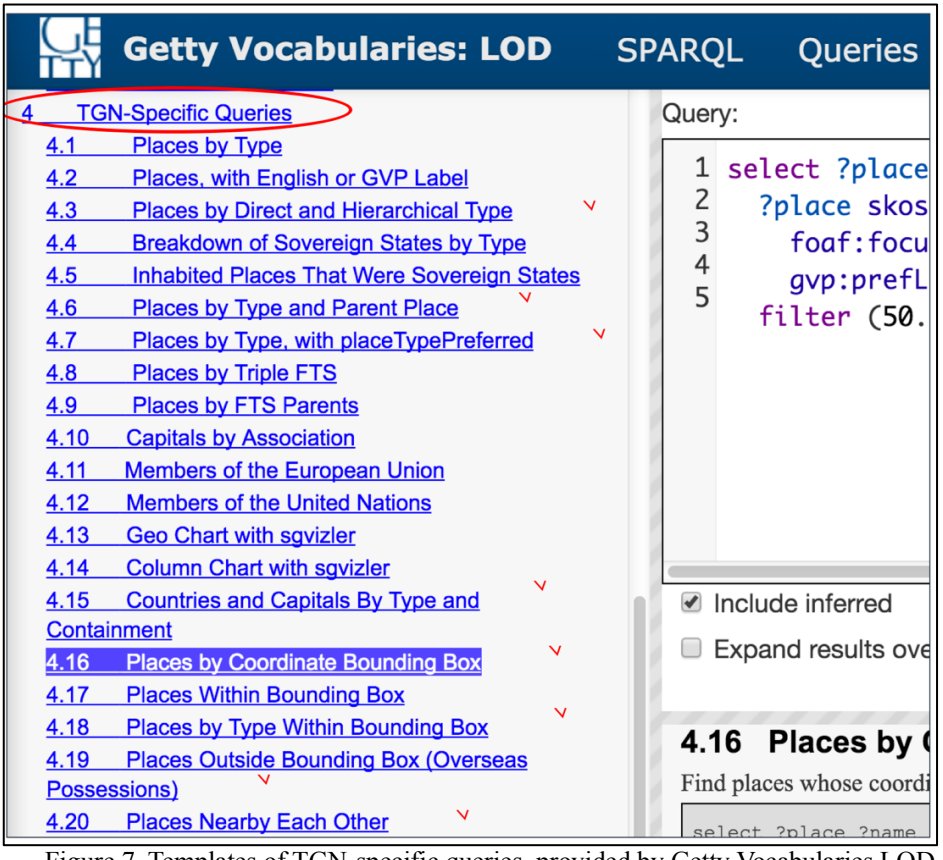


Figure 7. Templates of TGN-specific queries, provided by Getty Vocabularies LOD service (Source: http://vocab.getty.edu/queries#TGN-Specific_Queries).

To demonstrate, the screenshot of Figure 8 is an action to obtain a dataset of "Places by Type Within Bounding Box". By choosing query #4.18 (left), the query template appears accordingly (lower right) and fills in the query box on top with a single click. The example provided by the template is to look for castles around the Netherlands (within 50.787185 3.389722 53.542265 7.169019). Now, it is at the hands of the researcher RS-1 to decide what "type" and what geographic boundary box he/she would like to check. For example, at first RS-1 replaced "castles" with "caves" and marked the geo coordinators around the

ancient Silk Road, within 24.75083 28.95778 43.80722 108.92861; then RS-1 submitted the query (Figure 8 upper). The result was a dataset of over 200 caves spread in various countries (Figure 8 lower), all done within a few minutes (Zeng and Hu 2017). Each URI also brings the full data for each cave and other related information. The dataset is available for downloading with various formats.

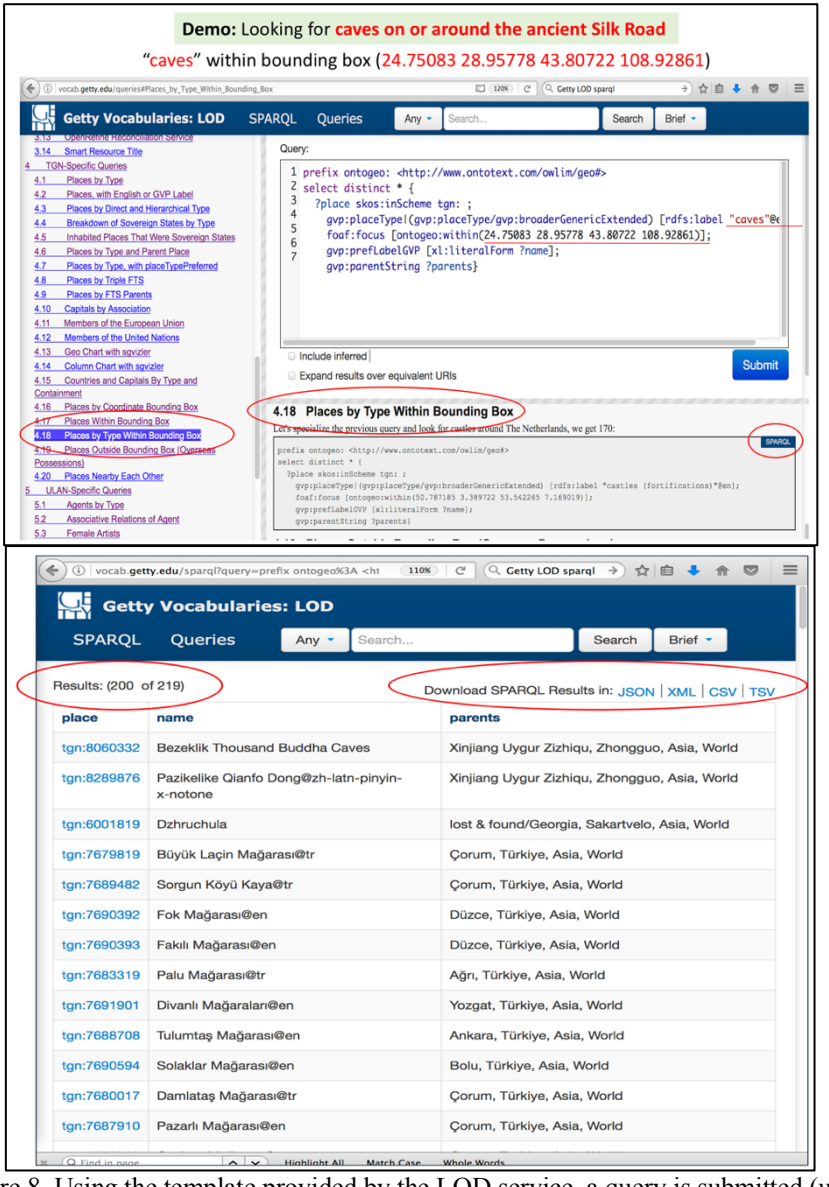


Figure 8. Using the template provided by the LOD service, a query is submitted (upper figure), resulting a dataset (lower figure) for a specific place type (e.g., caves) in a geographic boundary (Source: http://vocab.getty.edu/queries#TGN-Specific_Queries.)

RS-2. Name authorities offer foundational structured data for network analyses
Researcher RS-2's attention was on the artists who played significant roles in history. Similar to the preceding example, templates will help researcher RS-2 to use the *Union List of Artist Names* (*ULAN*) through Getty Vocabulary LOD service. The templates have also provided example queries for many complicated research questions that provide answers at impressive speeds. Checking each sample query in Figure 9 reveals without a doubt that any answer to such a question would not be possible by simply searching or browsing on a website by an end-user. For example, now RS-2 can gather the datasets for all the "Female Artists" (#5.3), for "Architects born in the 14th or 15th Century" (#5.7), for "Non-Italians who worked in Italy" (#5.9), or for all kinds of data related to an artist's network, region, time period, cultural group. These are based on the established entries that have been carefully created and quality controlled by the KOS producers; hence the results have high quality.

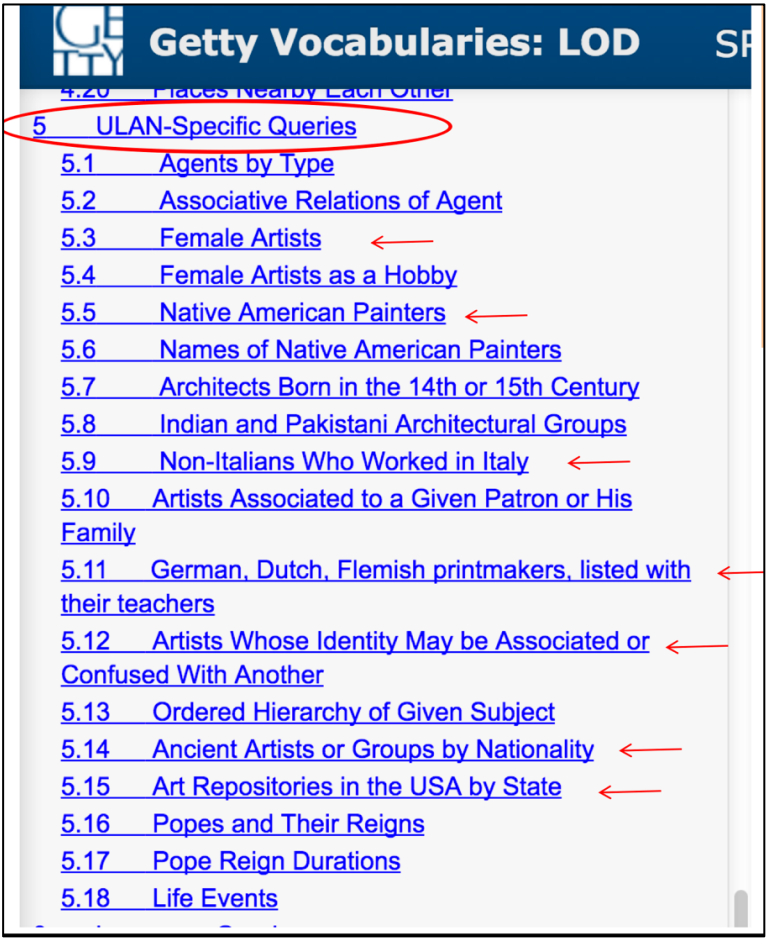


Figure 9. Templates of ULAN-specific queries, provided by Getty Vocabularies LOD service (Source: http://vocab.getty.edu/queries#ULAN-Specific_Queries.)

The following example is a query for finding associative relationships for "Wright, Frank Lloyd (American architect, 1867-1959), showing *relationship type, associated persons, each person's preferred name, preferred display biography*, and other *notes*." Again, the provided template #5.2 "Associative relations of Agent" made it possible for any end-user to just replace the URI of the artist aimed for (e.g., ulan:500020307 for Frank Lloyd Wright), submit the query (Figure 10 upper), and get the results as datasets (Figure 10 lower) (Zeng 2017).

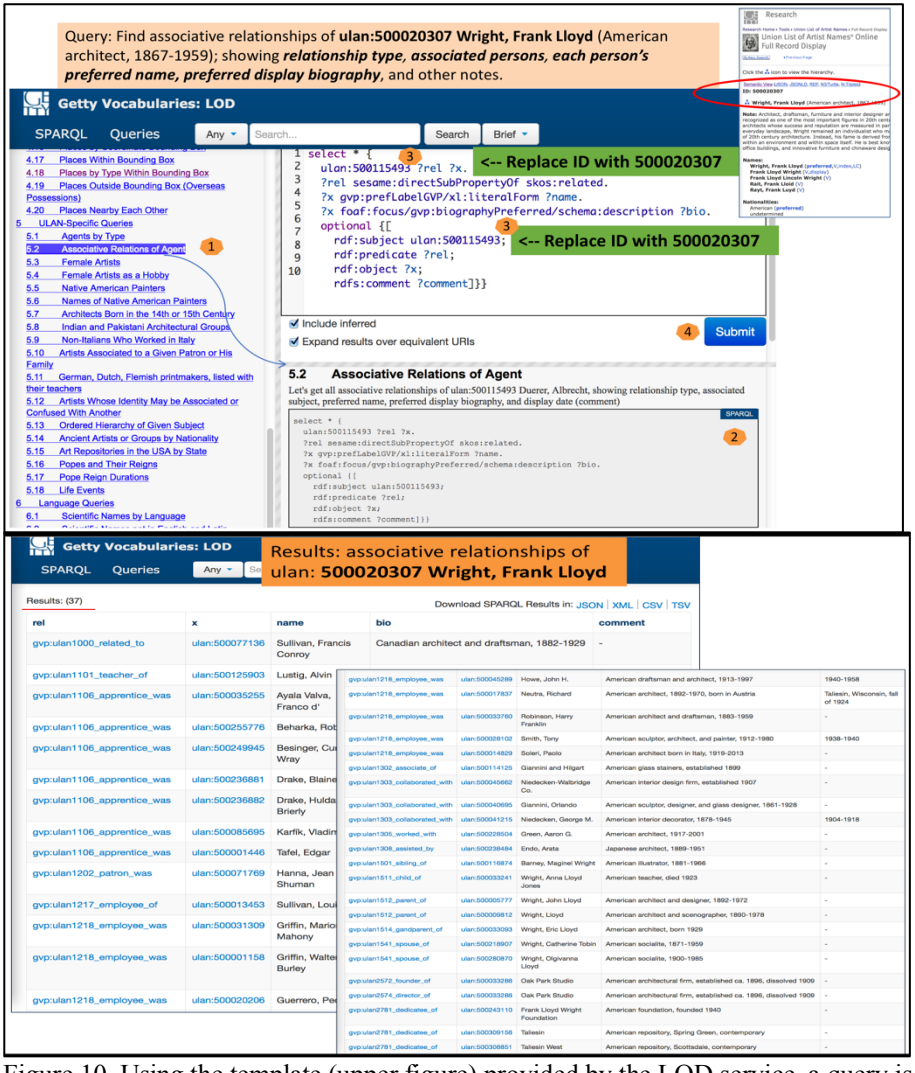


Figure 10. Using the template (upper figure) provided by the LOD service, a query is submitted to get the dataset for an artist *Wright, Frank Lloyd* and his associative relationships (lower figure) (Source: http://vocab.getty.edu/queries#ULAN-Specific_Queries).

The 37 related agents (Figure 10, lower figure) around this artist reveal the specific relationships. RS-2 or any user can further explore any of these related people, each named with a unique URI. If downloading the dataset (e.g., csv), one can also use other open tools (e.g., Google Fusion Tables, Gephi) to visualize the relationships with dynamic graphs.

We should also realize the importance of these URIs. Searching on the Web using such a URI, e.g., "ulan:500020307", the results will retrieve this artist's Wikipedia pages in all languages, the DBpedia entry, the links to the museums that host the artists' works (such as MoMA https://www.moma.org/artists/6459), and the libraries that have books about this artist (such as University of Wisconsin-Madison Libraries).

RS-3. User-friendly displays of KOS provide visually enriched understanding

To an end-user like researcher RS-3 who is not familiar with a KOS's structure and contents, a user-friendly display of KOSs may provide visually enriched understanding. The *Cadastre and Land Administration Thesaurus* (*CaLAThe*[21]), is reported to have been derived mainly from the *ISO/DIS 19152 Land Administration Domain Model* and is related to existing thesauri, primarily the *GEMET thesaurus*, the *AGROVOC thesaurus*, and the *STW Thesaurus for Economics* (Çağdaş and Stubkjær 2015). The approach is similar to the case related to VP-3. The additional effort is that the service's graphical overviews render the main groups ("Documentation," "Land," "Law," "Party" and "Activity") with thesaurus terms and relations. Individual concept searches also carry the results enriched with graphical views of the semantic relationships (see Figure 11).

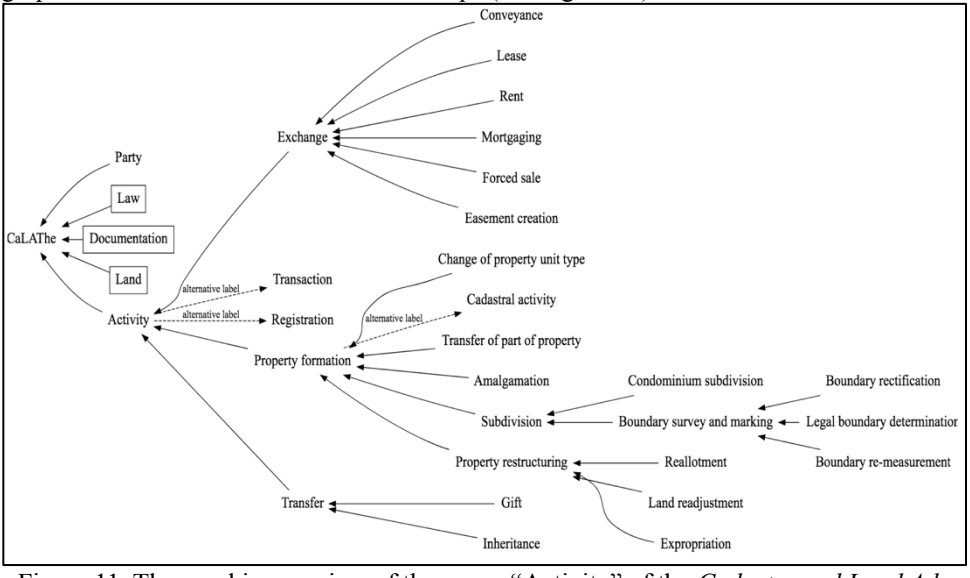


Figure 11. The graphic overview of the group "Activity" of the *Cadastre and Land Administration Thesaurus* (*CaLAThe*) (Source: http://cadastralvocabulary.org/).

Tools like SKOS-play[22] are free applications to render and visualize thesauri, taxonomies or controlled lists expressed in SKOS. For the user who is not familiar with markup languages, the tool provides a way to convert Excel spreadsheets to SKOS files plus the

visualization. Such features extend the benefits of SKOSified KOS publishing tools such as Skosmos[22] that would allow a vocabulary producer to test and verify a vocabulary during the conception phase; to exchange and communicate the vocabulary when validating it with domain experts; and to publish it when it is shared on the Web. With the added visual display function, end-users (not necessarily dataset or vocabulary producers) are able to have a visually enriched understanding of a KOS vocabulary's structure and contents.

Summary and discussion (Researchers)

The cases demonstrated in this sub-section highlight the great and endless potential of LOD KOS to be used by Researcher (RS) user group. The semantic rich structure and high-quality controlled vocabulary now can be used in an innovative manner; further than the existing controlled vocabularies or standardized name authorities.

Additionally, the appropriate practices for the implementation, extension, access, and use of these standards in the final deliverables is critical to the real extended functionality of the KOS beyond being the controlled vocabularies or standardized name authorities. There is still a long way to go to the point where KOSs are recognized as knowledge bases and semantic tools. It is important to realize the limitations of both typical web-based searching (simple, term-based) and browsing because these traditional methods are not taking the full advantages of machine-processable data that are much more powerful and useful than the previous machine-readable status.

5.0 Conclusion

This quotation from Aristotle, "The whole is greater than the sum of its parts," reminds us how much better things are together than as separate pieces. It also applies to the principles of design. All the cases presented here, as the representatives of ideas and practices, demonstrate that although it is possible to use each available component of a KOS independently, the real power lies in the skillful coordination of all. On the side of semantic technologies, the semantic web standards such as SKOS, OWL, RDFS and SPARQL have paved the way for conventional KOSs to become LOD datasets. On the side of the information and knowledge professionals, there have been tremendous and continuous needs for KOSs of all kinds, across domains and worldwide. When the two sides embrace and when KOSs join the mainstream in the 21st century, the opportunities for using the semantic-rich LOD KOS is much greater than ever before, due to the fact that LOD KOS data are machine-understandable, -processable, and -actionable (instead of just being machine-readable) in the semantic web, which connects things instead of strings.

In the effort to sort out the ideas and products related to LOD KOS (whether producing or using them) from disparate resources, we first created personas as typical users of LOD KOS, to build a common understanding of the needs and goals various user groups want to achieve. The accumulated set of cases we collected is open-ended and the sources are unconventional, as explained in Section 3.1. The research was aimed at examining the functional changes that optimize the usage of LOD KOS from multiple dimensions, in order to share the practices and ideas among related communities and users.

The findings indicate that the primary reason that LOD KOS vocabularies have become a fundamental component of the LOD building blocks is that they enable datasets to become 4- and 5-star Open Data. When trying to reach the benchmarks, every LOD Dataset Producer (DP) will realize their dependence on KOSs, which are their value vocabularies

and the sources of URIs/IRIs to be used in data-transformation. The openly available, well-established, and constantly-maintained vocabularies are invaluable engines for the LOD datasets. Common issues and benchmarks summarized can be applied to any project that LOD dataset producers might encounter.

In the section for the Vocabulary Producer (VP) group, the major conceptual and structural methodologies used by the cases resemble some found in the history of KOSs before the semantic web era. What makes them different is that the new approaches are empowered by the semantic technologies while the results comply with LOD principles. The data-driven, shared editing and publishing workflow also facilitate the capture of administrative, provenance and use metadata for the whole KOS and its components. With more KOSs being published in standardized, machine-understandable RDF format, institutions can improve and expand the KOS data that they already have with other outside sources. The most important achievement is the reusability of any of these new vocabularies.

The last section for researchers as end-users (RS) reveals great and endless potential for LOD KOS. The semantic rich structure and high-quality vocabulary now can be used integrated and innovatively, on top of being the controlled vocabularies or standardized name authorities. LOD KOS datasets should be considered as knowledge bases, as the foundation of network analyses and as the building blocks of a framework for research in the humanities and science. This might become the newest and most important function of KOSs, although such cases are still rare. We believe that the barrier resides in communication about KOSs through a delivery service rather than in the structure, format or contents of a KOS.

We would like to call for more needed collaborations between the knowledge organization communities and the semantic technology communities. Meanwhile, researchers who are real end-users will be invaluable in such collaboration because their domain expertise, information needs and information-seeking behaviors will lay out the questions that KOS knowledge bases can aim to answer, helping the growth of the KOS user communities with a variety of new objectives.

Notes

1. http://skosprovider.readthedocs.io
2. https://finto.fi/koko/en/
3. https://bartoc.org/
4. https://lov.linkeddata.es/dataset/lov
5. https://old.datahub.io/dataset
6. https://github.com/PhilippMayr/NKOS-bibliography/
7. Previously available at https://plus.google.com/u/0/communities/108509791366293651606
8. https://groups.google.com/forum/#!forum/gettyvocablod
9. https://share.getty.edu/display/ITSLODV/Home
10. https://groups.google.com/forum/#!forum/ontolog-forum
11. http://lodlam.net/
12. http://linkedjazz.org/
13. http://vocabularies.unesco.org/sparql-form/
14. http://onki.fi/onkiskos/cerambycids/
15. http://fast.oclc.org/searchfast/
16. http://schema.org/
17. http://linked.swissbib.ch
18. https://www.wikidata.org/wiki/Wikidata:WikiProject_Visual_arts/Item_structure

19. http://sparql.uniprot.org/
20. http://vocab.getty.edu
21. http://cadastralvocabulary.org
22. http://labs.sparna.fr/skos-play/
23. http://skosmos.org/

References

Baker, Thomas, Caterina Caracciolo, Anton Doroszenko, Lori Finch, Osma Suominen and Sujata Suri. 2016a. "The Global Agricultural Concept Scheme and Agrisemantics." In *Proceedings of the 2016 International Conference on Dublin Core and Metadata Applications*, *October 13-16, 2016, Copenhagen, Denmark*. https://dcpapers.dublincore.org/pubs/article/view/3847/2032

Baker, Thomas, Caterina Caracciolo, Anton Doroszenko and Osma Suominen 2016b. "GACS Core: Creation of a Global Agricultural Concept Scheme." In *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*. Springer International Publishing, 311-16.

Bensmann, Felix, Benjamin Zapilko and Philipp Mayr. 2017. "Interlinking Large-Scale Library Data with Authority Records." *Frontiers in Digital Humanities* 4, no. 5. doi:10.3389/fdigh.2017.00005

Berners-Lee, Tim. 2006. "Linked Data-Design Issues." https://www.w3.org/DesignIssues/LinkedData.html

Binding, Ceri and Douglas Tudhope. 2016. "Improving Interoperability Using Vocabulary Linked Data." *International Journal on Digital Libraries* 17, no. 1: 5-21.

Borge, Stefano, Nicola Guarino and Claudio Masolo 1996. "A Pointless Theory of Space Based on Strong Connection and Congruence." In *Proceedings of Principles of Knowledge Representation and Reasoning (KR96)*, ed. Luigia Carlucci Aiello, Jon Doyle and Stuart C. Shapiro. San Francisco, CA: Morgan Kaufmann, 220–29.

Brown, Dan M. 2010. *Communication Design: Developing Website Documentation for Design and Planning*, 2nd ed. Berkeley, CA: New Riders.

Buley, Leah 2013. *The User Experience Team of One: A Research and Design Survival Guide*. New York: Rosenfeld Media.

Busch, Joseph and Douglas Tudhope. 2020. "JDIS Special Issue on Networked Knowledge Organization Systems (NKOS)" *Journal of Data and Information Science* 5, no. 1: 1-2. http://manu47.magtech.com.cn/Jwk3_jdis/EN/volumn/volumn_60.shtml

Çağdaş, Volkan and Erik Stubkjær. 2015. "A SKOS Vocabulary for Linked Land Administration: Cadastre and Land Administration Thesaurus." *Land Use Policy* 49: 668-79.

Canadian Heritage Information Network (CHIN) [2016]. *CHIN Guide to Museum Standards*. https://www.canada.ca/en/heritage-information-network/services/collections-documentation-standards/chin-guide-museum-standards.html

D'Amore, Sebastian 2016. "Boost Empathy Quickly with Proto-personas." https://blog.mural.co/2016/05/06/boost-empathy-quickly-with-proto-personas

Garnier, Eric, Ulrike Stahl, Marie-Angélique Laporte, Jens Kattge, Isabelle Mougenot, Ingolf Kühn, Baptiste Laporte et al. 2017. "Towards a Thesaurus of Plant Characteristics: An Ecological Contribution." *Journal of Ecology* 105: 298-309.

Garcia, Gregg, Marcia Lei Zeng and Jonathan Ward. 2017. "Linked Open Data (LOD) Vocabularies: Querying, Dumping, Re-using, and Serving." How-to Session in *MW17: Museums and the Web Conference*, *April 19-22, 2017 Cleveland, Ohio, USA*. https://mw17.mwconf.org/proposal/linked-open-data-lod-vocabularies-querying-dumping-re-using-and-serving/

Golub, Koraljka, Rudi Schmiede and Douglas Tudhope. 2019. "Recent Applications of Knowledge Organization Systems: Introduction to a Special Issue." *International Journal on Digital Libraries* 20, no. 3: 205–7. https://doi.org/10.1007/s00799-018-0264-8

Gothelf, Jeff. 2012. “Using Proto-personas for Executive Alignment.” *UX Magazine* May 1: Article No:821. https://uxmag.com/articles/using-proto-personas-for-executive-alignment

Isaac, Antoine, William Waites, Jeff Young and Marcia Zeng. 2011. *Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets.* W3C Incubator Group Report. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/.

ISO 25964-1:2011 Information and Documentation -- Thesauri and Interoperability with Other Vocabularies -- Part 1: Thesauri for Information Retrieval. ISO TC 46/SC 9 Working Group on ISO 25964. Leader: Stella Dextre Clarke. Approved and published by ISO 2011-08.

ISO 25964-2:2013 Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 2. Interoperability with other vocabularies. ISO TC 46/SC 9 Working Group on ISO 25964. Leader: Stella Dextre Clarke. Approved and published by ISO 2013-03.

Krøger, Erie, Frode Guribye and Tor Gjøsæter. 2015. “Logging and Visualizing Affective Interaction for Mental Health Therapy.” *Norsk konferanse for organisasjoners bruk at IT* (*NOKOBIT)* [S.l.] 23, no. 1. http://ojs.bibsys.no/index.php/Nokobit/article/view/272

“The Linked Open Data Cloud.” https://lod-cloud.net/#diagram

Mayr, Philipp, Douglas Tudhope, Stella Dextre Clarke, Marcia Lei Zeng and Xia Lin. 2016. “Recent Applications of Knowledge Organization Systems: Introduction to a Special Issue.” *International Journal on Digital Libraries* 17, no. 1: 1-4.

Menzel, Christopher. 2003. “Reference Ontologies-Application Ontologies: Either/Or or Both/And?.” In *KI Workshop on Reference Ontologies and Application Ontologies* 16. http://ceur-ws.org/Vol-94/ki03rao_menzel.pdf

O’Neill, Ed and Jeff Mixter. 2013. “(1) The Case for Faceting (2) FAST Linked Data Mechanics.” PowerPoint slides from *76th Annual Meeting of the American Society for Information Science and Technology (ASIS&T)*, *Montreal, Canada, Nov. 2-6, 2013*. http://nkos.slis.kent.edu/ASIST2013/ONeill-Mixter.pptx

Pattuelli, M. Cristina. 2012. “Personal Name Vocabularies as Linked Open Data: A Case Study of Jazz Artist Names.” *Journal of information science* 38: 558-65.

Pattuelli, M. Cristina. 2021. “Graphing Out Communities and Cultures in the Archives: Methods and Tools.” In *Linking Knowledge: Linked Open Data for Knowledge Organization*, ed. Richard P. Smiraglia and Andrea Scharnhorst. Baden-Baden: Ergon Verlag, 144-67.

Pattuelli, M. Cristina, Alexandra Provo and Hilary Thorsen. 2015. “Ontology Building for Linked Open Data: A Pragmatic Perspective.” *Journal of Library Metadata* 15, no. 3-4: 265-94.

Pruitt, John and Tamara Adlin. 2010. “Persona Conception and Gestation.” In *User experience remastered: Your guide to getting the right design, ed.* C. Wilson. Burlington, MA: Morgan Kaufmann, 155-219.

Sibille-de Grimoüard, Claire. 2014. “The Thesaurus for French Local Archives and the Semantic Web.” *Procedia-Social and Behavioral Sciences* 147: 206-12.

Sonvilla-Weiss, Stefan, ed. 2011. *Mashup Cultures*. Heidelberg: Springer, 2010.

Tudhope, Douglas and Traugott Koch. 2004. “New Applications of Knowledge Organization Systems: Introduction to a Special Issue.” *Journal of Digital Information* 4, no. 4. https://journals.tdl.org/jodi/index.php/jodi/article/view/109/108

Tuominen, Jouni, Nina Laurenne and Eero Hyvönen. 2011. “Biological Names and Taxonomies on the Semantic Web: Managing the Change in Scientific Conception.” In *The Semantic Web: Research and Applications, 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29 – June 2, 2011, Proceedings, Part II. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg. 255-69.

Voss, Jon. 2013. “Radically open cultural heritage data on the Web.” In *Museums and the Web 2012, April 11-14, 2012. San Diego*, CA, USA. https://www.museumsandtheweb.com/mw2012/papers/radically_open_cultural_heritage_data_on_the_w

W3C. 2009. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation 18 August 2009. Ed. Alistair Miles and Sean Bechhofer. https://www.w3.org/TR/skos-reference/.

Wallis, Richard. 2014. “Linked Data: From Library Entities to the Web of Data.” Presentation slides from *American Library Association Conference, June 2014*, *Las Vegas, USA.* https://www.slideshare.net/rjw/linked-data-from-library-entities-to-the-web-of-data

Zapilko, Benjamin, Johann Schaible, Philipp Mayr and Brigitte Mathiak. 2013. “TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences.” *Semantic Web Journal (SWJ)* 4, no. 3: 257-63.

Zeng, Marcia Lei. 2017. “Create Microthesauri and other Datasets from the Getty LOD Vocabularies.” In *MW17: Museums and the Web Conference, April 19-22, 2017 Cleveland, Ohio, USA* http://www.getty.edu/research/tools/vocabularies/zeng_microthesauri_getty_lod.pdf

Zeng, Marcia Lei and Tao Hu. 2017. “Extending Exhibitions to Historical Journeys Through Data [in the Semantic Web].” In *MW17: Museums and the Web Conference, April 19-22, 2017 Cleveland, Ohio, USA.* http://www.getty.edu/research/tools/vocabularies/zeng_silk_road_tgn.pdf

Zeng, Marcia Lei and Julaine Clunis. 2020. “FAIR + FIT: Guiding Principles and Functional Metrics for Linked Open Data (LOD) KOS Products.” *Journal of Data and Information Science* 5, no. 1: 93-118. https://doi.org/10.2478/jdis-2020-0008

Zeng, Marcia Lei and Philipp Mayr. 2019. “Knowledge Organization Systems (KOS) in the Semantic Web: A Multi-Dimensional Review.” *International Journal on Digital Libraries* 20, no. 3: 209–30. https://doi.org/10.1007/s00799-018-0241-2

Appendix A. User persona document example: Vocabulary Producer (VP)

<table>
<tr><td>Name</td><td colspan="2">Vocabulary Producer</td></tr>
<tr><td>Key</td><td colspan="2">VP</td></tr>
<tr><td>Sources</td><td colspan="2"><table>
<tr><td>Original sources</td><td>Used for</td></tr>
<tr><td>LOV on Google+ https://plus.google.com/u/0/communities/108509791366293651606</td><td>VP-1</td></tr>
<tr><td>Getty Vocab Google Group https://groups.google.com/forum/#!forum/gettyvocablod</td><td>VP-1</td></tr>
<tr><td>LODLAM challenges and sessions http://lodlam.net/</td><td>VP-2</td></tr>
<tr><td>Research-based journal publications; conference and workshop presentations</td><td>VP-2, VP-3, VP-4</td></tr>
<tr><td>Theses and dissertations</td><td>VP-2, VP-3</td></tr>
<tr><td>GitHub entries such as OpenSKOS, NatLibFi/Skosmos, JSKOS</td><td>VP-4</td></tr>
<tr><td>Social media sources: tweets, blogs, Facebook groups</td><td>VP-2, VP-5</td></tr>
<tr><td>Informal interviews and local meetings</td><td>VP-1, VP-2</td></tr>
<tr><td>Mailing lists within a user group</td><td>VP-1</td></tr>
</table></td></tr>
<tr><td>Tasks</td><td colspan="2">Vocabulary producers are involved in the development, maintenance, and enrichment of new and existing KOS in a wide range of scales (e.g., micro, satellite, unified, heterogeneous, extended, enriched, or other kinds). The tasks usually include:
•Creating, developing;
•Maintaining, enriching, extending, translating;
•Integrating and unifying;
•Transforming (e.g., making an ontology from a thesaurus);
•Mapping with others;
•Sharing, reusing, contributing;
•Quality control and maintenance.</td></tr>
<tr><td>Content</td><td colspan="2">•Entries / instances -- with all property components required, including semantic and linguistic, format requirements, following standards and best practices;
•URIs – with namespace of any entry from any source;</td></tr>
</table>

	•Rights and contributors; •Provenance data; •Updates info (new concepts, terms, relations, sources, etc.); •Samples, previews, feedback, issues; •Related images; •Sources and URIs of the related real things; •Alignments coded with appropriate degrees.
Interactions	•Working platforms (spreadsheet, local database, open tool, etc.); •Desktops /Mobile Applications; •Websites (HTML, navigate-able); •API-based services; •SPARQL endpoints (with or without templates); •Datasets.
Goals	•Create and maintain high quality vocabularies; •Follow the vocabulary principles of user-warrant, literary-warrant, organizational warrant; •Follow international standards for KOS structure, components, and interoperability; •Comply with Linked Data principles; •Enrich, extend, and update contents constantly; •Share, reuse, and contribute (both in and out) in vocabulary productions.

Tobias Renwick
University of Alberta

Rick Szostak
University of Alberta

Chapter 4
A Thesaural Interface for the Basic Concepts Classification[**]

Abstract
We describe a thesaural interface that is being developed for the Basic Concepts Classification. This interface is particularly well-suited to the synthetic phenomenon-based approach to classification pursued by the BCC. We describe how the thesaural interface works, our plans to develop it further, and the advantages of this interface for both classifier and user.

1.0 Motivation

A classifier using the Basic Concepts Classification (BCC; Szostak 2019) would create a subject string combining terms from separate schedules of phenomena (mostly nouns), relators (mostly verbs or conjunctions), and properties (adjectives and adverbs). The resulting subject strings resemble sentence fragments. It is hoped that a classifier can move fairly directly from a key sentence in an abstract, book description, manuscript description, or object description to a BCC subject string.

Compared to classifying with an enumerative classification, the classifier is spared from having to find a complex enumerated subject heading that best fits a particular document or object. But the classifier now has to synthesize multiple terms, generally from two or three separate schedules. The BCC schedules are generally easy to navigate: hierarchies are flat and logically constructed for the most part. Yet a classifier seeking to synthesize several terms might nevertheless find it time-consuming to identify all of the necessary controlled vocabulary.

It has long been hoped, then, that a thesaural interface could be constructed that would guide classifiers to BCC controlled vocabulary. Importantly, such an interface might allow users also to enter a query in words of their choice and be guided quickly to controlled vocabulary. This in turn might encourage both public and university libraries to move back toward subject searching: Though keyword searching is easier for most library users, it is far less precise than subject searching (Hjørland 2012). A thesaural interface might potentially render subject searching as easy as keyword searching.

2.0 Design of the interface

We are exploring the possibility that such a thesaural interface can be developed using the Universal Sentence Encoder (USE: Cer et al. 2018). One common criticism of a synthetic

[**] Reprinted with minor editorial emendations by permission from *Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark*, ed. Marianne Lykke, Tanja Svarre, Mette Skov and Daniel Martínez-Ávila. Advances in Knowledge Organization 17. Baden-Baden: Ergon Verlag, 527-31.

(post-coordinated) approach to subject classification is that a user searching for "philosophy of history" will find many documents on "history of philosophy" (Sauperl 2009) But this is only true if the search interface does not care about word order in search queries. USE does discriminate on the basis of word order, for it places each term in the context of the phrase it is embedded within.

USE is a transformer-type deep neural network, which has been trained on very large batches of text. USE can help identify synonyms for words and phrases by embedding them into vectors in 512-dimensional space. Embeddings are modeled after the idea, "you shall know a word from the company it keeps" (Firth 1957, 11) and can be seen as a fixed length numeric representation of a text-based input. The guiding principle behind all embeddings is that if two words are often used in a similar context they likely have a similar meaning. In addition to the context of the word, USE also incorporates a token's position in the phrase to determine its meaning. This ability to discriminate words based on their position and context is created by virtue of the way USE is trained.

During the training phase, USE consists of two principal components: an encoder subgraph which builds a 512 dimension numeric encoding based on text input, and a decoder sub-graph which takes that numeric output as input and attempts to predict the next word in the sentence. USE maintains word order and positioning on the input phrase by adding a second 512D vector to the input which is built by overlapping different wavelengths of sin functions (e.g. sin x, 2sinx), which assign unique values to each position to a maximal length of 1024 tokens. The oscillating nature of the sin function allows the network to generalize shorter inputs to longer ones where it can potentially obverse a similar distance and pattern between words used. Because of this (and other contexts observed during training where these 2 phrases are used), the phrases `philosophy of history' and 'history of philosophy' have different embeddings.

After the network is trained, USE consists of only the encoder portion of the network, which then takes in a sentence in English, and outputs a 512D vector, as before, but now rather than predict the next word, we use the information present in that embedding to convey information about the input phrase (a sentence embedding). An interesting aspect of these embeddings which helps to add some intuition as to how they are constructed is that they can be shown to obey interesting properties when used mathematically. The classic example is that if you take the vector for the word 'King' and subtract the vector for 'man', you effectively remove all of the words associated with males from king, and you obtain the context that would surround a genderless royal (imagine words like crown, throne, rule, subjects etc.). Interestingly, now if you add the vector for 'woman' you will have a result which very closely matches the vector for 'Queen'. Other common examples can be illustrated by taking a country, subtracting its capital, then adding a different capital to obtain the other country's approximate vector (France – Paris + Rome = Italy).

Happily, USE can deal with phrases up to 1024 tokens in length, rather than just individual words. This will save classifiers and users from having to translate each word individually into controlled vocabulary. More importantly, phrases further clarify the meaning of the words they contain (for example clarifying whether "picture" is being used as noun or verb).

In order to make use of these embeddings, the entire terminology of the BCC (phenomena and relators), ISO Country and language codes, and UNSPSC codes used to identify goods and services within BCC have been embedded with USE (transformed into 512D

vectors of floating point numbers). Further to this, an additional embedding which contains all 2 word classifications consisting of BCC relator + BCC phenomena have also been added. These embeddings are combined into a single vector array, which allows direct comparison to an unseen embedding.

When a phrase is presented to the interface, the phrase is first checked for terms which exist directly in the BCC, and is broken into sub-phrases, which will then be translated. For example, the phrase 'a man dancing at a club' is broken into three sub-phrases by the interface, around the word 'at' ('a man dancing', 'at', 'a club') which is a BCC term (NT3). This is primarily a simple heuristic which allows for breaking up input in a predictable and reasonable way. If the phrase is too long, it will likely contain too much information to be adequately translated into a 1 or 2 word BCC classification. Therefore, the best results are obtained by using the most concise terminology possible with the translator.

Each sub-phrase is then embedded with USE and the resultant vectors are compared to the pre-calculated vector field of BCC (and related) embeddings. Classifiers and users can be (immediately) given ten possible BCC translations for each phrase block from which to choose. Technically, from the array the interface selects the ten nearest neighbors, using a measure of cosine similarity (cosine similarity is employed based on the assumption that vectors pointing in a similar direction have similar meaning, and ignores the magnitude of the vector).

A demonstration version of the interface can now be seen at

https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013/thesaural-interface-for-bcc

At present it deals best with shorter phrases. Readers can enter any phrase and be guided to appropriate BCC terminology (and given the BCC notation that goes along with that terminology). At present, they may have to search again if important terms are missing in the generated subject string.

3.0 Future work

In its current state, the thesaural interface is a helpful tool. In the future as more classifications are collected as input data, a true translator could be developed that would be able to better handle ambiguity in input data without breaking the input phrase into blocks. In the prior example 'a man dancing at a club' the 3rd sub-phrase 'a club' is ambiguous without context, and the first suggestion of the translator is incorrect (a UNPC classification for clubs), and while the correct class ('E09(901520) - nightclubs and dance halls') is included, it was returned as a more distant match.

We are also working on algorithms that can cope with larger phrases. We can also then analyze many examples of translations. We are also working on tree structures based on the hierarchies within BCC: the translator can then appreciate that the best place to look for controlled vocabulary for a type of painting is within the category "Art" rather than "Mathematical concepts." Note that our interface can be improved over time through repeated use (and selection by users or classifiers of particular options) to better select the best BCC translation of particular queries.

4.0 Discussion

Thesauri (within information science) have almost always been developed in the past to guide users toward controlled vocabulary within enumerated classifications. The thesaural

interface developed here is much better suited to a synthetic approach to classification, for it can identify the best fit by combining multiple terms within the controlled vocabulary. This thesaural interface thus reverses a potential disadvantage of a classification such as BCC. Without the thesaural interface it might be time-consuming to identify all of the controlled vocabulary necessary for a synthetic subject string (though, again, the logical and flat structure of BCC schedulers would facilitate search for controlled vocabulary). With the thesaural interface it becomes fairly straightforward to move from a key sentence in a document or object description toward a BCC subject string. This is already the case for fairly short sentences and will hopefully become feasible in future for longer phrases. Just as it is easy for a classifier to move from a document or object description toward a subject string, it should be easy for users to move from a query in their own words toward a subject string that guides them to the document or object they seek.

Documents or objects are described in sentences. User queries are generally formulated in sentences. We have in the past attempted to translate user queries into a subject heading that is not constructed grammatically, and used that ungrammatical subject heading to attempt to identify relevant documents. It is potentially both easier and more precise to translate sentence to sentence to sentence: translate the user query into a sentence-like subject string that will guide users to documents or objects that are described by a similar sentence. The thesaural interface described in this paper can hopefully guide users and classifiers to describe a document with the same (or very similar) subject string.

Users with a precise query can thus be guided to the document(s) or object(s) they seek. Users performing exploratory searches will benefit from the fact that the interface provides them with several suggested subject strings. Users can then adjust each term in their search query to identify different sets of subject strings. They start by wondering about dogs biting mail carriers, and move on to dogs biting neighbors, dogs licking mail carriers, or cats biting mail carriers. We hope to develop a visual interface that would allow users to easily adjust their search query term by term.

5.0 Concluding remarks

There is a dissonance in the field of knowledge organization between a body of theory that urges faceted classification and a body of practice around enumerated classification. One practical advantage of leading enumerated classifications is that they benefit from over a century of development. The thesaural interface discussed here can potentially allow a synthetic approach to classification such as the BCC to outperform enumerated classifications without the painstaking task of developing a thesaurus manually. The thesaural interface can facilitate the work of classifiers in moving from a key sentence in an object or document description toward a BCC subject string. It can so facilitate user queries that these can be as easy as keyword queries – but provide much more precise results.

References

Cer, Daniel, et al. 2018. Universal Sentence Encoder. *Arvix*. https://arxiv.org/pdf/1803.11175.pdf

Firth, John R. 1957. *Papers in Linguistics 1934–1951.* London: Oxford University Press.

Hjørland, Birger. 2012. Is Classification Necessary after Google? *Journal of Documentation* 68: 299-317.

Sauperl, Alenka. 2009. Precoordination or Not?: A New View of the Old Question. *Journal of Documentation* 65: 817-33.

Szostak. Rick. 2019. “Basic Concepts Classification.” *ISKO Encyclopedia of Knowledge Organization*. https://www.isko.org/cyclo/bcc
United Nations Standard Products and Services Code UNSPSC). n.d. https://www.unspsc.org

Aida Slavic
UDC Consortium

Ronald Siebes
Vrije Universiteit, Amsterdam

Andrea Scharnhorst
Data Archiving and Networked Services (KNAW)

Chapter 5
Publishing a Knowledge Organization System as Linked Data
The Case of the Universal Decimal Classification††

Abstract
Linked data (LD) technology is hailed as a long-awaited solution in web-based information exchange. Linked Open Data (LOD) bring this to another level by enabling meaningful linking of resources and creating a global, openly accessible knowledge graph. Our case is the Universal Decimal Classification (UDC) and the challenges for a KOS service provider to maintain an LD service. UDC was created during the period 1896-1904 to support systematic organization and information retrieval of a bibliography. When discussing UDC as LD we make a distinction between two types of UDC data or two provenances: UDC source data, and UDC codes as they appear in metadata. To serve the purpose of supplying semantics one has to front-end UDC LD with a service that can parse and interpret complex UDC strings. While the use of UDC is free the publishing and distributing of UDC data is protected by a licence. Publishing of UDC both as LD and as LOD must be provided for within a complex service that would allow open access as well as access through a paywall barrier for different levels of licences. The practical task of publishing the UDC as LOD was informed by the "10Things guidelines." The process includes conceptual parts and technological parts. The transition to a new technology is never a purely mechanical act but is a research endeavour in its own right. The UDC case has shown the importance of cross-domain, interdisciplinary collaboration which needs experts well situated in multiple knowledge domains.

1.0 Introduction
Linked data (LD) technology is hailed as a long-awaited solution in web-based information exchange which removes obstacles imposed by platform- and domain-dependent formats and standards. It is also seen as another way to organise information, in graph-like structures rather than in database structures. 'Open' LD, as part of the Linked Open Data (LOD) cloud, brings this to another level by enabling meaningful linking of resources and creating a global, openly accessible knowledge graph with almost unlimited potential for generating

†† Over the years, the UDC linked data project has benefited from expert help by Christophe Guéret who was first to propose the UDC linked data as a lookup service supporting a more complex model of UDC publishing, Chris Overfield for his contribution in setting up RDF stores and putting the service together and Attila Piros for developing the new UDC parser. Finally, this project benefited greatly from the DiKG project for making it possible to bring the UDC LD data project to completion.

new and unexpected associations between dispersed bits of information. The custodians of the LOD cloud, its main technologies and methods, are part of the scientific community of the semantic web. They bring together consortia and forums such as W3C, DCMI (Dublin Core Metadata Initiative), BIBFRAME (Bibliographic Framework Initiative), LD4L (Linked Data for Libraries), semantic web conferences (such as ESWC, ISWC), etc. But, the user base of LOD or LD technology is far broader, encompassing information processes and services in economy, science and society at large. Consequently, the LD technology is a very dynamic and fast-growing field. In this chapter we would like to contribute to the exchange of experiences among those who adopt this technology. Our case is the Universal Decimal Classification (UDC), and we will discuss the challenges for a knowledge organization system (KOS) service provider (in our case the UDC Consortium) to maintain an LD service.

The UDC as a showcase provides an insight into the reasoning, procedures and challenges associated with applying LD in the bibliographic domain, specifically with respect to KOSs as bibliographic tools. The expression bibliographic sector or bibliographic domain covers activities, agencies and services concerned with preserving, collecting and organizing recorded information and facilitating information discovery and access since the beginning of literacy. The bibliographic domain comprises the library sector, information and documentation centres, the publishing sector, services such as bibliographic and abstracting services, citation indexes and full text bibliographic services (e.g. Citation Index, Chemical Abstracts, Ebsco, Inspec, ARIBIB). Since the end of the 19th century the bibliographic domain has been creating international standards for describing and indexing information resources, such as cataloguing standards, KOSs, and later, in the computer era, data coding and bibliographic data standards (e.g., the MAchine-Readable Cataloging or MARC family of standards), data and service protocols, etc. These tools and standards have been enabling international exchange and cross-collection information discovery among libraries and between libraries and bibliographic services or publishers.

The UDC was one of the pioneering tools designed to meet the growing information needs of industrialisation and to support opportunities for learning and bettering the society which came with it. To better understand the challenges when publishing a bibliographical tool such as UDC as LD, in Section 2 we will remind the reader of some fundamentals of information and knowledge organization as applied in the bibliographic domain. These are very basic, but often, maybe because they are so basic, they are not articulated or made explicit. We continue in Section 3 to present the UDC as an exemplary case, outlining the main features of a classification and its own trajectory into automatization. In section 4, we elaborate on the identified challenges when it comes to sensible LD publishing of a system as complex and long lived as the UDC. Section 5 presents the technological choices made to respond to those challenges. In section 6, we conclude by summarizing our understanding of how legacy collections and tools used in information discovery can enrich and inspire the ways in which the LD technology may be applied in the future.

2.0 Information organization, knowledge organization and linked data

2.1 Visions and foundations

Until the advent of the internet and the semantic web, i.e. LD technology in particular, a bibliographic domain was a relatively isolated information space with limited options for merging with or allowing information linking with other information domains (scientific

data sources, archives, musea, etc.). Similarly, KOSs developed in the bibliographic domain such as bibliographic classifications, subject heading systems, thesauri and descriptor systems have only occasionally been used outside their field of provenance. Thus, they could not be used as a link between similar information contents dispersed in different information domains and sectors.

This is rather disheartening given that both the Mundaneum by Paul Otlet in 1910 and the Memex by Vannever Bush in 1945 were envisaged as services to enable easy access to all records of human knowledge (van den Heuvel 2008; Wright 2014). They envisaged a knowledge space where we would seamlessly access and move between primary (documents), secondary (bibliographic data) and tertiary information sources (encyclopaedias). These ambitious and visionary projects firmly rooted in the bibliographic world were destined to fail simply because they were not supported by technology similar to what is at our disposal today. The LOD cloud, as a manifestation of the semantic web, has remarkable resemblance to the visionary ideas of a place where we can access all human knowledge (The Linked Open Data Cloud 2020). If we look at the LOD cloud diagram we can discover many LD clusters, representing both KOS and contents indexed by them in different fields of human activities. We can see bibliographic data clouds, i.e., secondary sources, being connected to KOS clouds and both being linked to primary sources and scientific data. All of these clouds are connected to tertiary information sources: encyclopaedias and other reference sources. It is noteworthy, that the centre of the LOD cloud is occupied by DBpedia, the LD representation of Wikipedia. As it can be observed from the LOD cloud visualization, the semantic web operates in the realm of "everything," "universal," "all" and although bound to the digital world, it does not draw spatial, linguistical or temporal boundaries with regards to information and knowledge. This analogy was the main motivator behind making the UDC one use case in the Digging into Knowledge Graph project (DiIKG; see Martínez-Ávila et al. 2018; Szostak et al. 2018; Szostak et al. 2020).[1]

It seems natural to assume that with the help of LD technology, bibliographic data and tools such as KOSs are on their way to being fully integrated in the semantic web where they are much better placed to fulfil their role of a pivot connecting different knowledge structures and content. However, as we experienced, on the implementation level we have to deal with many details and complexities which have the potential to turn into obstacles. Once we resolve the basic technological obstacles of identifying and linking resources, everything else depends on the semantics, i.e., on the question of whether the premise upon which two things, two concepts or two resources are related is true. Identifying and connecting named entities and information objects in the Web space is relatively easy and straightforward. Connecting ideas and knowledge about these entities and preserving the many ways these may be systematized, represented and communicated in human knowledge is an entirely different plane of complexity. This is why it is important to create a shared understanding when it comes to notions such as "concept," "resource," "value," "label," "term," etc. (cf. Smiraglia and van den Heuvel 2013).

There are many "meta" levels through which recorded knowledge is communicated and there are methods and semantic models developed throughout history from the beginning of literacy. In the domain of recorded information, we manage information and knowledge by differentiating between concepts (thoughts), languages by which we communicate these thoughts, abstract bodies of work in which thoughts are organized (intellectual work), expressions we use to communicate these bodies of work (painting, speech, textbook, fiction),

the embodiments of this work into a certain physical format and recording onto some kind of carrier (book, image, recordings, file). We manage information discovery by separating information resources, with all their facets as listed above, from their content (aboutness) and we use metadata to aggregate, present and retrieve information.

2.2 Aboutness and knowledge organization systems

The expression "subject of a document," whereby subject represents a summarised body of ideas, is commonly used in information organization to denote aboutness, i.e., the content of an information resource. When populating metadata to describe what a document is about, we perform subject indexing. In doing so, due to the complexity of human knowledge and ambiguities of natural language, we have to use formalized languages, i.e., indexing languages or more broadly knowledge organization systems (KOSs). KOSs are sources of concepts and associated language terms whose meaning and position in the knowledge space is defined through semantic relationships. They tend to represent knowledge as a coherent system with an associated formal vocabulary and syntax. They are external tools for representing knowledge forms that comply with accepted scientific, educational and professional consensus in a given time and in given domains of application. In the bibliographic domain, KOSs are, conceptually, either classifications or alphabetical indexing languages (descriptors, thesauri, subject heading systems). Classification groups concepts according to their similarity into classes or class categories which are all assigned a notation (artificial code) to preserve the logical order and meaningful grouping of concepts. Alphabetical indexing languages use natural language terms to represent concepts and arrange concepts alphabetically. These two types of KOSs support different functions in indexing and retrieval and are often used in combination.

In indexing these KOSs ensure predictability and in information retrieval they can be implemented to support information browsing and semantic search expansion. The strength of KOSs is that they are standalone, self-contained, external resources that can be shared and are often developed or function as standards for the international exchange of information. The owners and curators of KOSs can be standards bodies, consortia, institutions or agencies concerned with subject access to information who take it upon themselves to maintain and keep knowledge structures and associated vocabularies up to date. More complex KOSs are kept in databases in proprietary or, in best case scenarios, in domain-specific data structures and formats (e.g., *MARC 21 Format for Authority Data* 2020) and are made available for use as printed or digital outputs or as web applications. In libraries, at the point of use, they can be integrated with library systems in the form of subject authority files. Far more frequently, indexing terms taken from a KOS will be found only as values in a dedicated field of bibliographic metadata where they appear as simple textual strings detached from the semantic context of the KOS from which they were taken (Slavic 2008). Although many KOSs are generally appreciated as authoritative sources of terminology and useful tools in providing semantic relationships, their use outside the bibliographic domain has always been limited by the lack of their accessibility and common vocabulary exchange standards.

This has all changed with LD and the availability of the Simple Knowledge Organization System (SKOS) standard, OWL (Web Ontology Language) and associated tools and standards for porting KOSs into the semantic web, which gained momentum from 2009 onwards (Slavic 2016). The discussion of exposing KOS as linked data, however, started

fairly early in the wake of the semantic web phenomenon (Zeng and Mayr 2019). Simpler controlled vocabularies such as thesauri proved to be more accessible to non-experts and easier to model by the SKOS developers. Thus, thesauri on different subjects and in a range of languages were more readily published as linked data using the SKOS standard. Larger and more complex systems such as general classification schemes traditionally dominating international information exchange within the bibliographic domain have been somewhat lagging behind due to a combination of factors related to their publishing models and limitations of SKOS (Slavic 2016). This is likely to change with the Library Linked Data (LLD) (Baker 2011; Tillet 2017) and BIBFRAME initiatives gaining prominence in the bibliographic domain and with the increasing number of bibliographic metadata collections and bibliographic tools such as name and subject authorities being published as LOD. They are creating a natural environment in which bibliographic classifications are a missing piece of the puzzle. With both bibliographic metadata and KOS data residing in the same space within a LOD cloud, it is possible to connect indexing terms in, for example bibliographic metadata, with the KOS from which these terms were taken and where they have further semantic links or to use classification as a pivot for mapping between KOSs. But this also means that some of the idiosyncrasies, procedures and approaches in supporting subject access, developed and being used in hundreds of thousands of legacy collections, are now also becoming a part of the semantic web story.

KOSs are numerous with different provenances and knowledge structures linked to various fields of application. They predate LD technology and even when kept in a machine-readable data format they are not likely to be modelled with the level of formality typical of ontologies. Thus, every KOS published as linked data has to undergo a process of converting its data model to a readily available but simplified model such as SKOS. This usually means either dumbing down sophisticated semantic features or extending the SKOS data model with elements from other web ontology standards. Further to this, publishing KOSs as LD, and in particular as LOD, is associated with many levels of decision making. In this chapter we will illustrate some of these using the example of the UDC.

2.3 Why classification?

Classifications deal with meanings and thoughts as elements of a knowledge space as a whole and, in the case of general knowledge classifications, such as UDC, this means all areas of human activity. The defining feature of classifications is that they deal with concepts, i.e. ideas, and are not concerned with the language used to express them. Classifications group and organize concepts according to their semantic proximity into a logical sequence of classes and subclasses constituting a hierarchy. Each class may comprise any number of concepts sharing the same characteristics. Classes from a classification scheme may be combined to express complex statements about meaning, following a specific set of combination rules. When we use a classification, we can group things that belong together, we can present them in a logical order and we can do so irrespective of whether we describe simple or complex subjects. This makes classifications indispensable when it comes to the logical presentation, organization and contextualization of information within a knowledge space. They can represent what is already known and what is deemed important to communicate for a given purpose in an information environment. By providing statements on how ideas are understood they are not only instructive but also helpful in

discovering new things and anomalous phenomena through the analysis of patterns, interpretation and derivation from the existing structures (Kwasnik 1999).

In the bibliographic domain classifications are used to communicate and facilitate discovery and access to knowledge. The structure of these classifications is derived from knowledge as recorded in documents, i.e., as treated in a society, culture, science and education (Smiraglia 2001). Almost all general bibliographic classifications[2] organize knowledge in a series of disciplines and subdisciplines reflecting the way knowledge is taught and used. Classifications strive to be hospitable (expandable) and extensible to accommodate new and emerging knowledge and they would probably be more successful in doing so were it not for demands for stability and resistance to change imposed by the practicalities of resource collection management (see Suchecki et al. 2012, on the evolution of categorial systems). To mitigate structural rigidity, bibliographic classifications deploy various structural features such as facet analysis, analytico-synthetic features, perspective and alternative presentations of knowledge, syndetics (lateral semantic linking), etc. All of these create difficulties in representing, i.e., expressing, these structures using formal ontological languages.

3.0 The case of the UDC

3.1 Roles and intracacies of bibliographic classifications

In bibliographic and many other knowledge classification schemes, each class is assigned a unique (alpha)numerical symbol termed a "notation,"[3] which is used to represent its particular scope of meaning and the meaning itself is described with plain text, called a "caption."[4] The expression "classification scheme" denotes a given, named and authorised reference tool containing all notations with their corresponding meaning in multiple hierarchies covering all forms of knowledge and rules for a particular knowledge organization system. Schemes can be translated, published and distributed in many world languages (Figure 1-2). In the process of document indexing, one describes subjects by assigning a notation (or their combinations) taken from a classification scheme and recording this notation in subject metadata.

notation	*caption (class description)*	
	English	*French*
51	Mathematics	Mathématiques
512	Algebra	Algèbre
512.7	Commutative algebra and algebraic geometry	Algèbre commutative et géométrie algébrique
512.73	Cohomology theory of algebraic varieties and schemes	Théorie de la cohomologie des variétés et des schémas algébriques
512.731	Classical topology and cohomology theory of complex and real algebraic varieties	Topologie classique et théorie de la cohomologie des variétés algébriques complexes et réelles

Figure 1. An excerpt from the UDC scheme hierarchy showing captions in English and French.

538.9	Condensed matter physics. Solid state physics	*English*
	Φυσική συμπυκνωμένης ύλης. Φυσική στερεάς κατάστασης	*Greek*
	Физика конденсированного состояния. Жидкое и твердое состояние	*Russian*
	凝聚态物理、固态物理	*Chinese*
	ঘনীভূত বস্তুর পদার্থবিজ্ঞান। ঘনাবস্থা পদার্থবিজ্ঞান	*Bengali*
	संघनित पदार्थ भौतिकी. ठोस पदार्थ भौतिकी	*Hindi*

Figure 2. Caption of the UDC class 538.9 in six languages and scripts (UDC Summary).

Classification notations, due to their brevity, are a practical way of labelling printed and other physical media to achieve the systematic arrangement of knowledge in a physical space (e.g., library shelves) and they have been used for this purpose for thousands of years. Equally, notations are used in managing and monitoring the acquisition and circulation of physical documents, as well as organizing and presenting collection metadata for the purpose of information browsing and searching. Because of the fact that they provide a language-independent way of expressing subjects, classifications are particularly useful in information organization and discovery in a multilingual environment. They have a long tradition in being used in this way in the bibliographic domain whether linked to institutions (libraries, museums, archives, documentation and information centres) or information services and agencies supporting research and science, and even more broadly, the publishing industry.

Classifications that have been used for a longer period of time in a larger number of information services often gain prominence and become the tools of choice in information exchange as they can help link similar content irrespective of the language, script, provenance, region, type of information resource or time of publishing. Through their widespread use, classification schemes are translated into many languages and thus gain another useful function: they serve as a reliable source of semantically rich terminology (Figure 2). As such they can be used as a basis for creating thesauri or subject headings or directly to support user friendly interfaces for browsing a knowledge space in multiple languages.

An especially important and useful feature is when a general classification scheme treats all forms and fields of knowledge and their shared commonalities as a coherent system.[5.] Such classifications are typically structured according to disciplines respecting the educational and scientific principles according to which knowledge is taught, researched and applied. In a disciplinary organization, knowledge phenomena are placed in the fields of knowledge in which they are studied. This means that many phenomena may occur in several places, i.e., where they are the subject of study, and therefore their full meaning is determined by the context. This is resolved either by linking these concepts across the entire knowledge space, thus creating a network of associative relationships called syndetic structure. If made available to users, these types of semantic relationships can be very useful in resource discovery (Figure 3).

In the following sections, we will document discussions and technological design decisions made in the process of publishing UDC as LD. We will explain the context in which these challenges emerge by providing some information about the UDC's maintenance and use. We, then, present our own approach to solving these challenges. Some of these solutions are of a general nature and applicable to different KOSs.

3.2 The origins of UDC in the context of automation

UDC was created during the period 1896-1904 to support systematic organization and information retrieval of a bibliography in the form of a card catalogue called Repertoire Bib liographique Universel (RBU). Paul Otlet and Henri La Fontaine designed RBU to hold the largest record of human knowledge ever assembled to date. The catalogue was to be organized in systematic and logical order using a knowledge classification of great flexibility and detail. They very much liked the solution of presenting knowledge in ten main

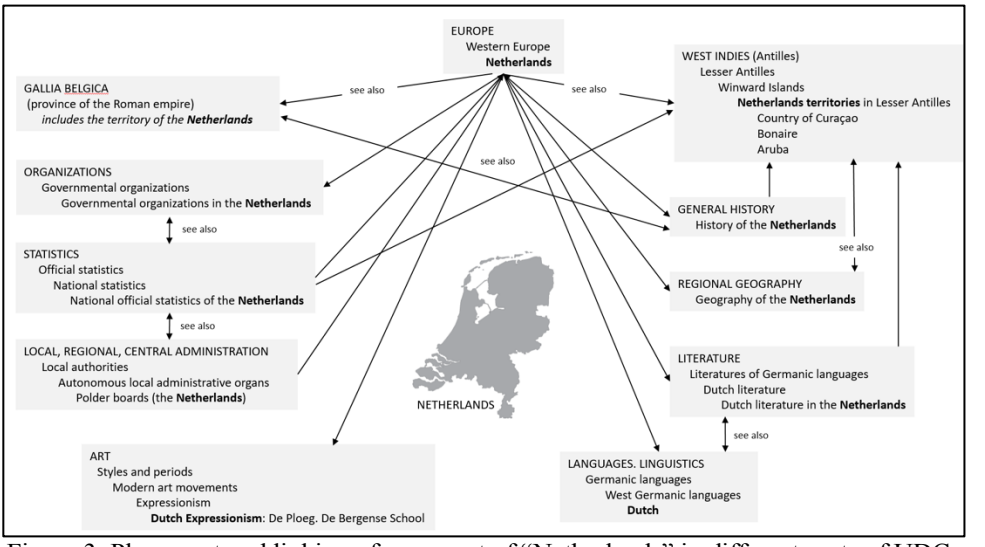


Figure 3. Placement and linking of a concept of "Netherlands" in different parts of UDC.

groups that could be subdivided as required which they found in the *Dewey Decimal Classification (*DDC*)*, but needed a very detailed indexing language capable of expressing complex subjects in science and technology. Otlet and La Fontaine obtained permission from Melvil Dewey to translate and use the basic structure of his classification and went on to develop an analytico-synthetic indexing language with vocabulary that by 1905 already exceeded the size of *DDC* by several orders of magnitude. By 1914, when it was closed down in the wake of World War I, RBU contained over fifteen million entries classified using UDC (Rayward 1990). For over a decade it supported an information service and functioned as a UDC-based "analogue search engine" answering over 1,500 information requests a year (Wright 2014). The RBU project remains an unsurpassed achievement in the field of bibliography and was entered in UNESCO's Memory of the World Register in 2013. Due to its ambition, RBU is sometimes compared to that of the internet and the semantic web (van den Heuvel 2008 and 2011; Wright 2014) and UDC remains its lasting legacy.

As a result of the international prominence of the RBU project, UDC has very quickly become the first KOS to be translated, adopted world-wide and maintained in multiple languages. Owing to its size and the amount of detail in the areas of sciences and technology it was often the choice of indexing language for scientific collections and bibliographical databases. As a consequence, UDC was not only the first classification to be used in online information retrieval (from the 1960s), but also the first to be included in information retrieval research, notably Cranfield experiments (cf. Slavic 2005, 14-28). Cranfield experiments measured the performance in the searching of UDC captions and UDC notations (implemented as an analytico-synthetic faceted notation). The UDC notation searching proved superior to other indexing languages in terms of relevance, precision and recall (Cleverdon 1962). The most important takeaway from this research, and the most

significant for LD, is that the degree of usefulness of an analytico-synthetic or faceted classification in information retrieval depends on the way it is implemented in the IR system.

There are several other automation milestones reached by UDC that are not as common when it comes to other bibliographic classifications. The English version of UDC was digitized, i.e. stored on magnetic tapes, in the 1980s and converted into a database in 1991, known as the Master Reference File (UDC MRF), containing around 60,000 classes of what has become a UDC standard. Since 1992 the UDC MRF has been distributed to users and publishers as a file export to be ingested and used within an information system or alternatively used within database software provided by the UDC Consortium. With the advent of the Internet there were several projects using UDC either to support an automatic classification of internet resources or to support browsing on information portals and gateways (Slavic 2006). In spring 2001, the standard scheme in English was published on the web, as a service which subsequently evolved into the UDC Hub in nine languages. In 2009, a selection of 2,600 UDC classes was published on the open web as UDC Summary, a database with an online translation interface made available to volunteers, that led to this database being translated into 57 languages.

In 2011, The Multilingual UDC Summary was published as LOD. In the process of publishing this first LD, the UDC Consortium did not envisage any specific purpose these data were supposed to serve. Hence, we refer to this project as an experimental UDC LD phase. Although the need to publish all UDC as LD was clear from the very beginning, requirements and functionalities the UDC LD are supposed to support from the point of view of users and publishers were rather difficult to define due to the factors we will discuss in the following section.

4.0 Challenges of publishing UDC as LD

As indicated in the previous sections on the bibliographic domain, we encounter KOSs in two forms: the system itself and KOS terms as they are being used in document metadata. In the same way, when discussing UDC as LD we make a distinction between two types of UDC data or two provenances:

- UDC source data, i.e. the UDC system itself and schedules as they are held in their native database (UDC MRF); and,
- UDC codes as they are applied in resource descriptions and appear in metadata in bibliographical databases, indexing and abstracting services, library catalogues and library shelves.

In the first, experimental phase of publishing UDC as LD from 2011-2019, there were over 2,600 UDC classes made available as SKOS exports. Whilst exposing vocabulary as a SKOS data dump with or without a SPARQL front-end was considered to be a successful completion for many value vocabularies, this was not considered to be the case with UDC.

There are four aspects of the UDC system that require a different approach and special treatment when it comes to publishing UDC as LD:

1. longevity - UDC has been continuously developed and updated for over 120 years;
2. structural complexity - UDC is an analytico-synthetic indexing language;
3. data ownership - UDC is a proprietary system with copyright protected content; and,
4. large usage base and amount of legacy UDC data – UDC is used in document indexing in a variety of bibliographic services, documentation centres and libraries in around 140 countries.

These facts influence the technical solutions that are discussed in this Chapter and deserve a more detailed introduction.

4.1 How does the longevity of a system affect LD?

Every KOS has to evolve to accommodate new concepts and terminology in science and other areas of human activity. Although knowledge and associated concepts cannot disappear from an information space or from KOSs, their status can change to obsolete or superseded and old terminology can be declared deprecated and replaced by modern terminology (cf. Tennis and Stuart 2008). Concepts can be moved to different parts of the hierarchy. The fact that the UDC is over 120 years old means that there have been many UDC versions and editions and the UDC system as a whole has a significant amount of historical UDC data and administrative documentation.

Classification schedules comprise a notation (classification codes) and associated text (class caption and notes). The revision of the scheme affects schedules in the following way:

- changes in the class caption and associated terminology affecting the scope and the meaning;
- concepts are moved from one hierarchy to another - in the process the UDC notation is cancelled (deprecated); and,
- new notations, i.e. new classes, are added.

The electronic standard version of UDC, the Master Reference File, created in 1992 comprising 60,000 classes has been undergoing regular change. Subsequent revisions of UDC affected 40% of MRF classes resulting in 12,500 cancelled notations, 22,915 new notations and over 10,000 classes affected by textual changes. The latest MRF contains around 72,000 classes. For all cancelled classes UDC provides replacement data, i.e. redirection to a new valid notation for the same content. A detailed list of changes is distributed to users and publishers with every new MRF release.

These kinds of changes produce a discrepancy between the standard, up-to-date UDC notations and notations appearing in bibliographic databases and libraries world-wide. For instance, up until 1993, UDC notation 94 represented "General Mediaeval and Modern History" and notation 930.9 was "General History. World History (chronological summation of facts)." In 1994, as a result of the UDC revision of the history class (UDC release vMRF94), notation 930.9 was cancelled and replaced by notation 94 which now has a changed description "General History." However, decades after this change there still may be bibliographies and library data world-wide using 930.9 to collocate documents on general history on what is now a non-existing UDC notation.

This is a well-known issue for all users of well-established and widely used KOSs causing permanent tension between requests from users for KOSs to be continuously updated and subsequent rejections of bibliographic agencies to introduce changes in their metadata due to a lack of resources. If both a) bibliographic databases (containing UDC codes) and b) the latest UDC version appear as an LD cloud it might not be possible to establish the link between deprecated UDC codes and those in the most recent version of the scheme. For this reason, if UDC is to serve its purpose in information exchange, it is extremely important to expose not only the most recent version of UDC but also the historical data. This affects the way we model, select the RDF schema and expose UDC data as LD.

4.2 How does structural complexity affect linked data?

UDC is an analytico-synthetic and faceted classification which enables the combination of concepts from different areas of knowledge. This feature is very important for the detailed indexing of documents and providing multiple subject access points. The main advantage of this kind of classification lies in its power to express detail and subject range with a

relatively small vocabulary. For instance, even using the UDC Summary, which contains only 3,000 classes (out of 72,000 of the complete UDC MRF), we can express the following content:

"Digital audio recordings, in mp3 format, of an anthology of the short stories of the modern Dutch literature of Suriname, in the English language":

821.124.5`06-32(883)(082)(086.7)(0.034MP3)=111

The meaning of notational elements:

821.124.5 Dutch literature (main class for literature);
`06 Modern (special auxiliary for periods);
-32 Fiction/stories (special auxiliary for literary forms and genres);
(883) Suriname (common auxiliary of place);
(082) Anthology (common auxiliary of form);
(086) Audio recordings (common auxiliary of form);
(0.034MP3) Digital documents - mp3 (common auxiliary of form);and,
=111 Documents in the English language (common auxiliary of language).

In order to function in this way, the UDC system consists of a vocabulary (classes of concepts) and syntax rules for combining classes into complex subject statements (McIlwaine 2007; Slavic and Davies 2017).[6] In practical terms, synthesis is enabled by the organization of concepts into tables (facets) and by the notational system, i.e. systems of numerical codes and syntax symbols enabling (de)composition of UDC strings. General concepts can be freely combined with themselves and with all areas of UDC, including the auxiliary tables. As a consequence, generally applicable concepts are always preceded and terminated by a certain symbol or combinations of symbols, punctuation or digits, collectively known as facet indicators, and they are all presented and used in this way throughout the schedules:

=...	Common auxiliaries of language
(0...)	Common auxiliaries of form
(1/9)	Common auxiliaries of place
(=...)	Common auxiliaries of human ancestry, ethnic grouping and nationality
"..."	Common auxiliaries of time
-02	Common auxiliaries of properties
-03	Common auxiliaries of materials
-04	Common auxiliaries of processes
-05	Common auxiliaries of persons

Thus, parentheses (followed by any digit from 1 to 9) always represent place, e.g. (492) represents Netherlands. Language will always be expressed with a number preceded by an equal sign =, e.g. =112.5 Dutch. The main table contains the main classes of disciplines, subdisciplines and fields of knowledge and these classes have a simple numerical notation (used decimally). They can be further specified through combinations with over 15,000 common auxiliary concepts (in the tables listed above), as well as a series of specialized concepts presented in special auxiliary tables that are always preceded either by -, .0 or ` (backtick or inverted comma). All UDC notations from the main tables and those from common auxiliaries can be combined among themselves using the following combination signs/symbols:

+	Coordination
:	Simple relation
::	Order-fixing
[ ]	Subgrouping

Each UDC notation, whether from the main or auxiliary tables can be specified further with alphabetical extensions, e.g. (492Delft) is used to express the Dutch city of Delft. Equally each notation can be connected using * (asterisk) for codes from other systems.

All these features make the UDC an analytico-synthetic system proper which with a relatively small number of classes can produce a very specific description of content (cf. Slavic and Davies 2017).

For instance, countries are listed only once in UDC: in a table of Common auxiliaries of place (place facet) where each country is assigned a unique notation and where the Netherlands is assigned notation (492). Disciplines of geography and history, for instance, are studies closely related to the notion of place and many classifications would need to list the geography of all countries and then the history of all countries of the world. In UDC, however, classes of 913 "Regional geography" or 94 "General history" are virtually empty, i.e., they do not need to list places of the world. Instead, in the process of indexing one combines a notation from the main table 94 and notation from the place table (492) to express 94(492) "History of the Netherlands" or 913(492) "Geography of Netherlands." This can be further extended by adding time auxiliaries, ethnicity, etc. Furthermore, if there is document content describing the relationship between regional geography and the history of the Netherlands, this can also be easily expressed with a combination 913:94(492).

The UDC's analytico-synthetic feature represented by this expressive notational system leads to complex UDC strings of various length (Smiraglia et al. 2013). Each of these strings reveals a precise and rich meaning which can be extracted from the string by accurate parsing of the UDC compound number by both humans and computers. Composition and decomposition of UDC numbers enables easier management and coordination of natural language terms when these are used instead of classification codes in the process of searching (Riesthuis 2008). But, it is also clear that for machine consumption a proper parsing of the string is of utmost importance.

For all the advantages of a partial or fully analytico-synthetic KOS such as UDC there is one important downside when it comes to LD: the UDC namespace does not contain the complex UDC strings that are created in the process of indexing locally and which may appear in bibliographic metadata in many collections world-wide. If UDC is to serve the purpose of supplying semantics, enriching and linking millions of bibliographic records one has to front-end UDC LD with a service that can parse and interpret complex UDC strings and provide adequate resolution by linking each element from the complex string to an appropriate UDC data record. Further in the text we refer to this solution as a lookup service.

4.3 How does classification ownership affect the model of LD publishing?

UDC is owned and managed by the UDC Consortium which is an association of publishers and users that operate on a non-profit, self-funded basis. UDC publishers are themselves non-commercial, publicly funded institutions such as national standard institutes and national libraries. The main source of income that sustains UDC maintenance and development comes from the sale of publishing licences or various languages and the sale of UDC schedules. Thus, while the use of UDC is free—the publishing and distributing of UDC data are protected by a licence with separate licences being issued for publishing of up to 50% and for more than 50% of UDC MRF. Clearly, although the business model is non-profit, it is impossible for the Consortium to release the complete UDC schedules as LOD without jeopardizing the future of the UDC.

In order to mitigate this situation, in 2007, the Consortium released, on the open web (under a Creative Commons licence), the UDC Summary, as previously mentioned. This

is a set of over 2,600 classes with captions translated in 57 languages. In 2011, this set was published as LOD.

In this context it should be mentioned that the complete content of the up-to-date UDC MRF is available on the web (www.udc-hub.com). Nine of the languages are available through the Consortium's UDC Hub service and two languages as a part of national services (Slovenia and Hungary). For six of these languages, national libraries are paying for a publishing licence to make UDC available free of charge in their respective countries (Croatia, Czech Republic, Hungary, Poland, Serbia and Slovakia). In these countries all the information agencies publishing metadata containing UDC as LD should be able to access and point to the UDC namespace. As time goes by, UDC in other languages is expected to be published in a similar way and free access in these countries will be regulated by publishing licences. Therefore, publishing of UDC both as LD and as LOD are options that have to be provided for, albeit within a more complex linked data service, that would allow open access as well as access through a paywall barrier for two different level of licences.

4.4 How does a large user base affect the linked data publishing model?

For over a hundred years UDC has been used in bibliographic databases, documentation centres, libraries and national bibliographies in all parts of the world. There are millions of bibliographic records containing UDC codes that may eventually be published as LOD. Due to its long history and wide-spread use, UDC functions as a *de facto* standard in information exchange. This status is closely linked to the authority and quality control enabled by the stability of its ownership.

Document content analysis and indexing using classification is a costly process and over the past decades, information services in general have fewer resources available to manage proper subject authority data. Automatic classification and indexing, where possible, are not always readily available or adequate and are associated with initial costs. But most importantly, information services have so far had difficulty in passing the benefits of knowledge access on to end users due to poor user interfaces or lack of expertise to exploit the classification data.

In general, classifications are expected to be implemented in an IR system "behind the scenes." They are supposed to support information browsing and searching without users being aware of the complexity and technicalities of an indexing language which is maintained in the background (using authority control). It is incumbent on the classification publishers to provide adequate support to bibliographic agencies and make it easier for them to keep their authorities up to date. This could support the following solution for the benefits of bibliographic agencies and their users:

- validating and updating classification data held locally (authority files), mapping deprecated notations to new notations or entirely bypassing and converting obsolete classification data in information exchange;
- enriching bibliographic metadata with additional semantics and verbal access points (additional search terminology in multiple languages, semantic expansion to broader, narrower or related subjects);
- enabling knowledge graph-based visualization, linear or multi-dimensional presentation for browsing and navigation across collections; and,
- enabling linking, i.e., mapping, to other KOSs and beyond, where classification acts as a pivot and enables cross-collection subject access.

It is, therefore, logical for the UDC owner to consider exposing the classification as LD not only as an experiment with limited value outside the semantic web community but as

a fully functional terminological service which would take on board specificities of UDC data, including legacy data and the way it has been applied in bibliographic collections.

5.0 Steps in publishing UDC as LD and LOD

Based on previous experience of publishing the UDC Summary as LD from 2011-2019 and taking on board the challenges outlined in the previous section, this new LD project was significantly larger and more complex and took a longer period of preparation. Thorough planning was especially critical given the objective of publishing different and larger UDC datasets under different access conditions. The steps in publishing UDC as LD/LOD, described in this section are to a large extent informed and follow the procedure described by Siebes et al. (2019).

5.1 UDC Summary as LOD 2011-2019: lessons learned

The first experience of publishing the UDC Summary as LD in 2011, proved an important learning step. At the time SKOS was embraced enthusiastically by the community sharing simpler controlled vocabularies such as thesauri and subject heading systems (SKOS 2009). The SKOS data model itself was designed with this kind of KOS in mind: it was simple, lightweight and easy to use and it represented a much-needed vehicle for many simple KOS systems to be published as LOD. SKOS filled in a void within cross-sector standard formats for publishing and sharing controlled vocabularies. It was also readily adopted as a data model in vocabulary management applications in enterprises and the commercial sector to support information and content management. Both KOS and the Semantic Web communities felt an urgency to expose KOSs as LOD, to secure visibility, longevity and find new purpose for traditional quality KOSs. The majority of these vocabularies have been underfunded and in danger of being superseded by advanced text retrieval technology and automatic indexing. This was, especially, the case with KOSs developed by heritage institutions, funded by the public or in the public domain.

The UDC Consortium, which, at the time, considered only a small set of data to be published as LOD (albeit in 57 languages) did not envisage any specific practical use scenario. Therefore, there were only four issues of concern:

- How to map the UDC data model into a SKOS schema?
- What to do with synthetic UDC notations?
- How to select the URI syntax?
- How to publish: as data dump or SPARQL front end?

The mapping of the UDC data structure into a SKOS schema (designed for thesauri) required a bit of tweaking. The general principle for this project was to select the minimum set of UDC data elements, leaving out all administrative and UDC data management fields. SKOS deals with concepts uniquely represented by controlled lexical terms (descriptors). Classification deals with classes of concepts and has three ways for representing and identifying classes: a) unique ID of a class (within UDC database system), b) notation, and c) caption. The most important elements are expressed as follows: the unique UDC class identifier (in the UDC MRF database) was stored as *skos:Concept*, the caption was mapped to *skos:prefLabel* and the notation was mapped to *skos:notation*. The SKOS schema was then extended by several data elements to accommodate UDC-specific notes to differentiate application notes and scope notes. After some deliberation and discussion, it was decided not to extend the SKOS schema to express the difference between simple and complex UDC notations (analytico-synthetic feature). This was left for consideration for the next

version of the UDC LD (Isaac and Slavic 2009; Slavic and Isaac 2009). The longest and most protracted discussion took place in relation to URI selection. The dilemma was whether to form a URI containing the UDC notation (the meaning of which depends on the UDC version) which is understandable to humans or whether to use a unique identifier from the UDC database. The final decision, in 2011, was to opt for a numerical identifier thus forming the following type of URI: "udcdata.info/019930" whereby the number "019930" represented a UDC record identifier for the notation 331 "Labour. Employment. Work." An important reason for not including, at the time, UDC notation in the URI was the practice of the occasional re-use of deprecated notations (usually after 10 or more years) which can be a source of ambiguity unless linked to the UDC version. Additionally, UDC notations may contain symbols and signs that according to URL standards require encoding before transmission. Another argument in favour of this decision was that the URI was meant for programs (not people), hence whether or not it contains a notation that can be read by humans makes no difference. The decision to publish UDC LOD as a data-dump and without programmatic access, i.e. SPARQL endpoint was based, primarily, on the fact that the Consortium, at the time, did not have any real life use-case scenario for UDC LOD.

Starting from 2011, the UDC Summary LOD data were available at https://udcdata.info for nine years and frequently downloaded (Figure 4). As we observed, this was mainly for the purpose of harvesting, extracting multilingual terminology and recreating and republishing UDC schedules in local systems or generating new types of proprietary vocabularies under different names. In principle, much of the LD use at the time did not consist of linking within the LOD cloud but rather of downloading, storing and processing data in local systems.

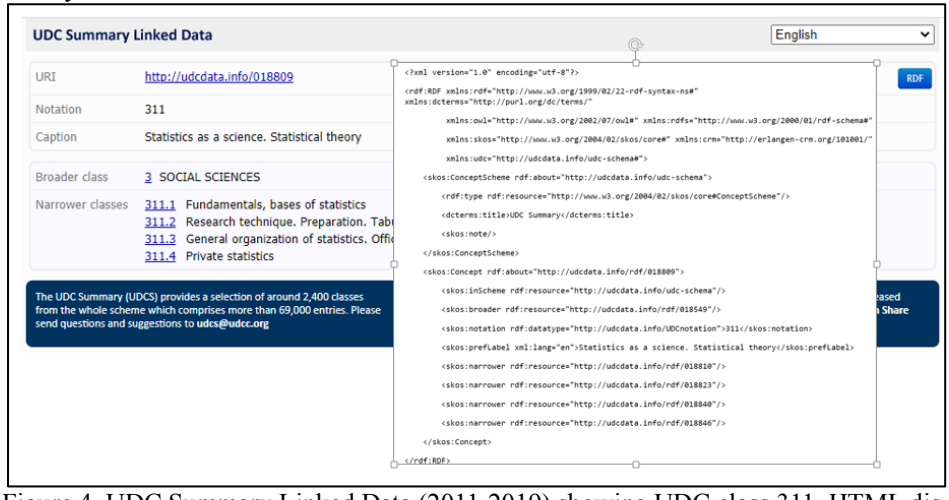


Figure 4. UDC Summary Linked Data (2011-2019) showing UDC class 311, HTML display and its RDF record.

Gradually, following the Library Linked Data (LLD) initiative, more and more library catalogues and national bibliographies became available as LOD and the importance of legacy bibliographic data and the potential use of UDC in accessing and exposing content of these collections has become more obvious. At the same time it was possible to observe

and identify potential impediments to linking a UDC standard dataset published as LD with bibliographic metadata sets containing UDC notations (Slavic et al. 2013; Slavic 2017).

LLD clouds observed from 2012 onward consisted of bibliographic metadata (e.g., catalogues of National Szechenyi Library and Library of the Norwegian University of Science and Technology-NTNU) containing very detailed and specific UDC notations. These detailed notations could only be found in the complete UDC standard and were not available within the UDC Summary LOD set. Additionally, much of UDC data in the bibliographic records consisted of complex UDC notational strings which, even if all UDC notations were available as LOD, could not have been linked easily (Figure 5). The main problem identified in almost all LLD sets, however, was the fact that libraries continue to use deprecated UDC notations that are no longer part of the standard active UDC dataset. In some instances, UDC notations found in active library records were cancelled thirty years ago. This means that a UDC namespace, if it is to serve the needs of the bibliographic domain, has to include historical data and concordances between cancelled and new notations.

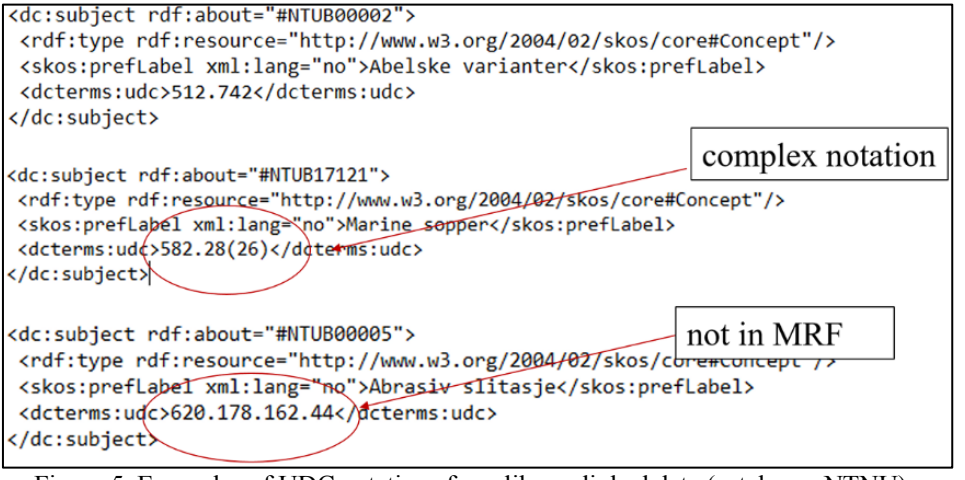


Figure 5. Examples of UDC notations from library linked data (catalogue NTNU).

5.2 New approach to publishing UDC as LD

As is evident from our observations, the UDC data found in LLD clouds consists of unstructured, simple textual strings of UDC notations, with no additional data of provenance, semantics or versioning. It is possible that the situation may be the same in local information systems from which these data were derived (cf. Slavic 2008). This means that queries launched from the LLD space to the UDC namespace will contain notational strings only. Thus, any interpretation of UDC data can only come from the UDC namespace, which needs to provide a solution for the linking and semantic alignment between UDC notations in bibliographic records and those in the UDC LD cloud. To do so all problems identified with respect to LLD so far, and as explained in the previous section, have to be addressed.

Clearly, one has to change and improve the way in which UDC LD are published and move from the simple UDC RDF repository to an LD service. This involves not only a change to the amount of data being exposed but the data format and the way these data are accessed. A new approach to publishing UDC as LD requires rethinking the URI format,

the RDF schema, but most importantly, instead of a UDC LD-dump or enabling API access, we had to opt for a more complex UDC LD look-up service. In planning this service, we took on board the aforementioned requirements and designed service components to handle the necessary functions.

The basic requirements of the UDC LD service are that, while it has to support the practical use of the scheme in the context of LLD, the service also has to protect the UDC model of publishing and secure its sustainability. The only way to handle this is having a small part of the UDC as LOD and the rest of the licence-protected UDC content available as LD behind a "paywall." In order to resolve problems observed in bibliographic data, the UDC LD-based service has to support the following:

1. Programmatic access to three sets of data:
   - •single LOD set: the UDC Summary containing 3,000 classes (under CC BY-NC-ND 2.0 licence)
   - •two LD sets behind a UDC MRF licence barrier:
     - •Abridged edition (12,000 classes)
     - •UDC MRF (72,000 classes), including all twenty versions of the UDC MRF and historical data comprising over 13,000 cancelled (deprecated) classes and their redirections to new classes;
2. UDC Look-up service that:
   - •parses and resolves (interprets) a classmark originated from bibliographic data and links its components to relevant records in the RDF data store;
   - •upon request supplies URI(s) for UDC notations or the full RDF records.

To meet these requirements the service has to have a more complex architecture and the following solution was chosen:

- •RDF stores (three Virtuoso databases: the UDC Summary, the Abridged edition, and the UDC MRF) with SPARQL endpoints accessible only via a restricted RESTful API layer which uses pre-designed SPARQL templates for query execution;
- •Web server and custom written UDC parser. The Authentication process is handled by standard shared and private authentication keys (tokens). The HTTP/Get parameters and the HTTP headers inform the server about the type of desired result (e.g., HTML, RDF-Turtle, JSON).

The architecture diagram in Figure 6 shows the UDC LD infrastructure and data flows that support the UDC notation (classmark) lookup process. The service is accessed by initially acquiring an authentication token from the authentication layer. The token received allows access to one of the following services: UDC Summary (LOD), complete MRF data (LD) or Abridged UDC LD.

For parsing and resolving a UDC notation the lookup service receives a plain text URI-encoded UDC notation query e.g. 94(492) encoded as 94%28492%29. The full-service query for, e.g., the UDC Summary, would look like this http://udcsummary.udcdata.info/api/parse/94%28492%29. The REST-API receives this query and extracts the URI encoded notation, passes this notation to the Parser which breaks down any compound, synthesized notation into constituent elements. The REST-API receives this from the Parser and retrieves UDC encoded URIs (see section 5.3.3) for each simple notation and returning the results in the required format. The user can then request the full record from the service using the returned URIs.

Other features of the UDC Look-up include an HTML interface for human interaction with the service (Figure 7). The assumption is that the API would be queried by programs submitting simple or complex UDC notations either to get correct URIs for UDC codes or to retrieve complete RDF records. In this scenario, the HTML interface allows humans to verify and explore the provided notations in which the parse tree, versions, and RDF translations are expressed.

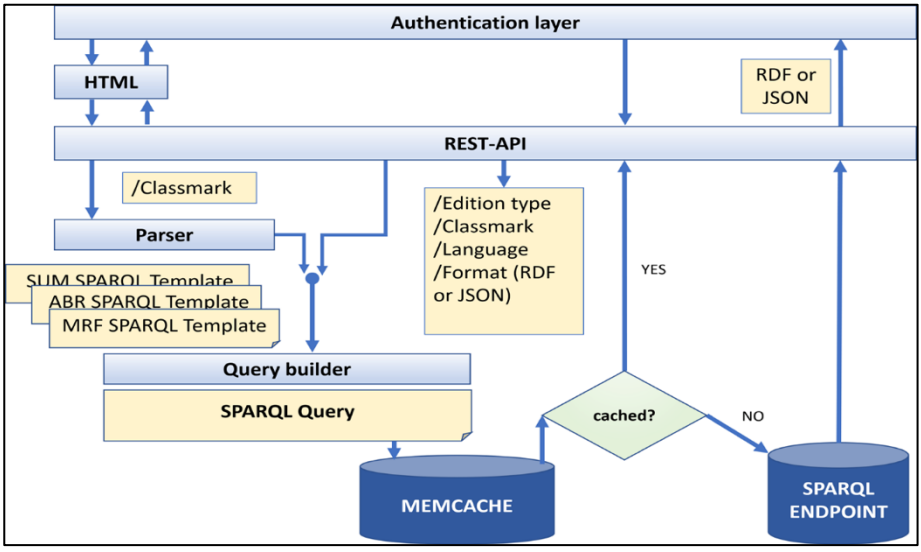


Figure 6. UDC Look-up service architecture.

5.2.1 UDC notation parser: an important component of the UDC look-up service

Within an information system, typically a library system, UDC is often managed within a subject authority file which allows managing and searching notational components and associating them with lexical terms and relevant semantic data. An authority control separates and detaches classification data from the way this is used or displayed on a searching or browsing interface. When bibliographic records are exported from a library system, in the process of information exchange or when published as LD, they show UDC notational strings, e.g., as we can see above in the NTNU RDF record 582.28(26). In order to be able to link components of complex UDC strings to their record in the UDC namespace, one has to be able to parse complex UDC notational strings. This function is one of the most important components of the UDC look-up service.

Programs for automatic UDC notation parsing were developed in the past, and as reported by Riesthuis (1997, 1998), his own program was 100% successful in splitting complex UDC strings into searchable components. This is due to the fact that UDC has an expressive notational system and uses facet indicators consisting of punctuation symbols and digits, to indicate the beginning and the termination of a notation for specific types of concepts (as explained earlier in Section 4.2). The absence of punctuation in connection to numbers is also meaningful. Thus, in the following UDC strings 94(492), 94:33 or 94“19”(492) we can clearly see that these are complex expressions consisting in the first two cases of two and in the third case three separate notational elements.

The UDC syntax rules determine which notations can be combined with which other notations and in which linear order. These rules are formalized through the use of selected digits, punctuation and characters and represent a UDC formal language. As UDC develops, over time, some syntax rules get refined or changed, and new notations and notational symbols are added. For instance, since Riesthuis wrote his programme, in 1996, two new common auxiliaries for properties and processes were added. This means that these three

characters -02 now denote the beginning of concepts from the table of properties and -04 denotes concepts of processes. Clearly, over time, new algorithms have to be added on top of the previous set of parsing algorithms. For this reason, and for the purpose of the LD Look-up service, Attila Piros has developed a new and improved parsing program that will allow for the continuous and controlled update of different sets of parsing algorithms. This was based on his previous parser known as Piros UDC-interpreter (Piros 2015 and 2017).

The UDC notation interpreter, as put by Piros, is an automata that, based on a series of algorithms, recognizes this formal UDC language and generates a tree which contains the parts of the notation (based on predefined rules) as well as connecting symbols. The basic set of parsing algorithms deals with connecting signs and notations for common auxiliary tables (general concepts) as these are the most stable part of the UDC syntax rules. This is followed by a series of smaller parsers handling subsequent rules pertaining to different parts of the UDC schedule. The very last phase of the parsing process deals with semantic analysis and UDC notations created through parallel division and application of special auxiliaries. Figure 7 shows an HTML interface in which a complex UDC number is split into components that, in this case, have valid UDC classes. The service executes queries against the UDC Summary, the UDC Abridged edition or the UDC MRF and in the second step it generates an RDF representation of the information selected by the user or program.

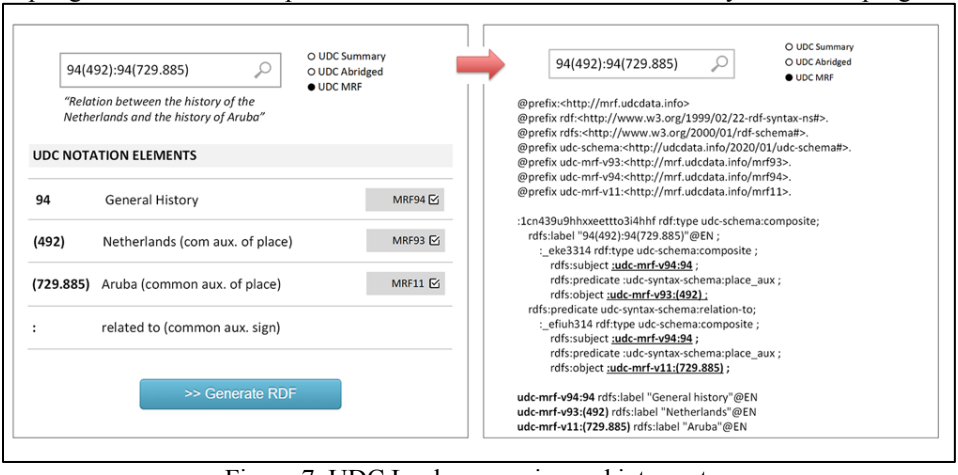


Figure 7. UDC Look-up service and interpreter.

Prior to arriving at this solution there were several important steps and key decision to be made that are relevant for many linked data projects (cf. Siebes et al. 2019).

5.3 Important steps and decisions in publishing UDC as LD

5.3.1 Selection of data

In the previous section we mentioned three different UDC datasets: UDC Summary, UDC Abridged Edition and UDC MRF (the complete UDC dataset). The main purpose of the UDC LD service, at this point in time, is to provide support in interpreting and linking content of legacy collections. In doing so, one has to balance constraints of data ownership, licensed users and related context that would enable UDC to sustain. The UDC Summary,

the UDC Abridged Edition, and the UDC MRF are maintained in different MySQL databases and the same set-up is replicated for the RDF store (three Virtuoso databases). The selection of these three datasets is based on the well-established practice in UDC data use and publishing. They are representative of two kind of access to UDC data: open access and access through a UDC MRF license requiring an authentication process based on authentication tokens (managed outside the service itself). With respect to the supported languages, the UDC Summary contains language data in 57 languages. Abridged and MRF translation databases contain 14 and 13 languages respectively. The UDC LD look-up service will be incrementally developed to include languages for which the UDC Consortium has a license clearance. In the first phase the Abridged and MRF datasets are planned to be exposed in English only. UDC data comprise many data elements that are required for data management and publishing, for the LD, we selected only 14 data elements (see 5.3.5 below). In terms of sequence of data release, the UDC Summary (the LOD set) was given priority due to the large community of users.

When it comes to the selection of UDC data, the most important and innovative aspect of the UDC LD service is that the UDC MRF dataset includes over 13,000 cancelled, non-active UDC notations. These are the result of the classification revision from 1992-2018. Further cancelled historical records will gradually be added through the process of digitization of the old UDC editions from 1905 to 1992. Once this process is completed the UDC MRF will capture the dynamics and changes in its knowledge structure from the beginning of the twentieth century to date.

To understand the significance of historical UDC data for legacy collections one has to be aware of the nature of changes that affect classifications of knowledge and notation lifecycles. Once concepts and subjects become part of human knowledge they do not disappear, but as the understanding of a field of knowledge changes this affects the concept organization of the field. A good illustration is the classification of plants and animals that evolves based on new knowledge, whereby living organisms are re-grouped in a different way. In the process, an organism can be moved to a new class of organisms and therefore is assigned a new UDC notation. The old notation for that organism is cancelled. From the point of view of UDC data maintenance, the record of the cancelled notation is marked as "not-active" and information regarding replacement notations, i.e., the new notation to which this concept was moved, is supplied.

There are millions of bibliographic records containing UDC notations that were cancelled and replaced by new notations decades ago. The possibility to query historical data, either to retrieve URI for these deprecated notations or to use these data to find redirections/replacement for these notations is very important for libraries, bibliographic services and legacy collections in general. Without this link between an old class and the new class, provided in the UDC namespace, we would not be able to establish meaningful connection between information resources dealing with the same entity in different points in time.

5.3.2 Use scenarios, serialization and resolution of UDC codes and URIs

We considered various scenarios in which the UDC namespace would likely be accessed in order to select an appropriate LD serialization of the UDC data source. The service is primarily aimed for access by programs and for this to work there is a need to have disambiguation mechanisms and clear guidance for programmers with respect to various choices that would apply. For instance, often the only information libraries have about the UDC

are the classmarks and the location of the UDC Look-up service. Libraries launching queries are not likely to be aware of the UDC MRF versions, including whether notations contain valid or cancelled (deprecated) UDC notations or whether they have licence, i.e., authentication token, to query full UDC data. Their queries may have the following format "udcdata.info/582.281.1(035)." The UDC Look-up service will parse and resolve such a query returning information that notation 582.281.1 is cancelled and replaced by 582.244 and might also return an RDF statement with sets of URIs expressing the relationship between these two numbers. If subsequently a program or person, without an access token for this dataset tries to query these URIs at the UDC namespace, the authentication layer would prevent this request from being executed and return a meaningful error message, combined with some information about the result of the query (e.g. a broader class shared both by the MRF version and the UDC Summary version, i.e. the UDC LOD set).

5.3.3 Defining the URI naming strategy

The main principle of the URI is to be durable and well structured. When it comes to UDC LD, the URI strategy involves:

- the choice of the internet domain name;
- the choice of the structure of the URI with respect to sub-domains, notation and datasets; and,
- the solution for issues caused by the URI encoding standard.

In making these decisions we took on board the architecture of the RDF store, i.e. number of separate datasets (including constraints on their access) and the ways the service will be used/queried and the use scenarios.

Domain name. The UDC LD domain name "udcdata.info" was established in 2011 and remains the same. The old LD RDF store and LD data dump were taken down in 2019 upon the release of the new UDC MRF12 version and to avoid ambiguity. Given that the look-up service will operate on three different datasets (one of which will be LOD; two are behind the data barrier and would require an authentication token. We added three subdomains in the following way:

udcsummary.udcdata.info/...
mrf.udcdata.info/...
abridged.udcdata.info/...

URI paths: As explained previously in Section 5.1, the selection of URI in the first UDC LD version back in 2011 took much deliberation. The choice, in the end, was to use an unstructured type of URI which does not contain UDC notation and the URI for class 311 was based on the database identifier and appeared as follows:

http://udcdata.info/018809

whereby 018809 represents a unique identifier in the UDC MRF database for class 311. Based on the projected use scenarios, we have established that the notation is the only element on which this kind of service can operate and it makes sense that it forms a part of the URI. This decision, however comes with the complications which follow.

Notation in UDC is not a unique identifier: Due to the size and longevity of the system a number of notations were cancelled in the past and re-used with a different meaning. Because of this, the meaning of a notation is determined by the UDC MRF version in which it is introduced.

Which MRF version was used in the URI?: Once introduced a notation will appear in many subsequent UDC MRF versions. However, the URI is formed from the name of

the MRF version in which this class was first introduced. These data are controlled through the Introduction Date field in the MRF database. The URI for class (492) will, thus, indicate that this class was first introduced in the UDC MRF release version coded as 'mrf93'.

udcsummary.udcdata.info/mrf93/(492)

The issue with URI encoding: UDC notations contain signs and symbols and these would be URI encoded automatically in the process of making an HTTP request. Thus, a URI "udcsummary.udcdata.info/mrf93/(492)" would appear encoded as "udcsummary.udcdata.info/mrf93/%28492%29." To have better control over the way UDC notations are displayed the service has its own URI encoding. For instance, parentheses (round brackets) will be replaced as follows: opening parenthesis will be "_or_" and closing parenthesis will be represented as "_cr_."

The encoded URI for class (492) in the UDC Summary LOD is as follows: "http://udcsummary.udcdata.info/mrf93/_or_492_cr_" (Figure 8). A UDC look-up service will transpose conventional notations into a required URI format. Therefore, one can query the service using a conventional UDC notation and will receive a rrelevant URI.

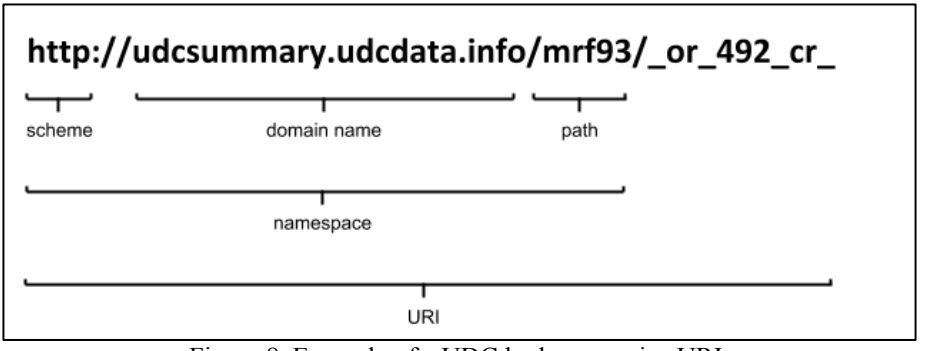


Figure 8. Example of a UDC look-up service URI.

This change of the URI format, as described above, means that a new service must contain a mapping between the old 2011-2019 URIs and the new URI systems.

5.3.4 UDC data analysis and RDF presentation schema

UDC MRF is held in a database which contains, apart from the basic UDC data, many administrative data elements relevant for system management, maintenance and publishing. Many of them deal with identification and tracking of changes and continuous revisions to the system. Rather than mimicking the MRF database structure the RDF representation and associated UDC knowledge graph capture the most relevant elements of the UDC system structure and their relationships: class identification data (identifiers, date of introduction, date of cancellation), notation, caption (lexical data associated with language attributes) and semantics. A UDC class modelled as an RDF triple is shown in Figure 9.

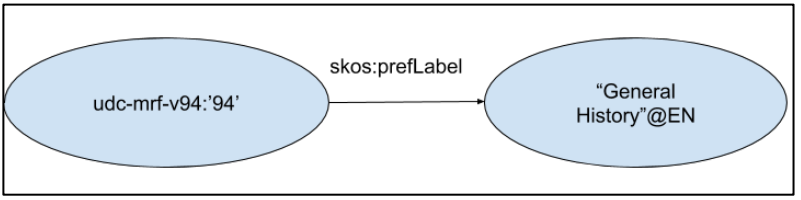


Figure 9. RDF graph representation of class 94 and its caption General History.

Figure 10, below shows a snapshot from the current web-interface of a top level UDC scheme. Next to it we show the corresponding UDC knowledge graph for class 94 General history with its broader class 9 Geography. Biography. History. Both classes were introduced in UDC MRF version v94 and "udc-mrf-v94" combined with notation 94 and 9 respectively, uniquely identifies these two classes.

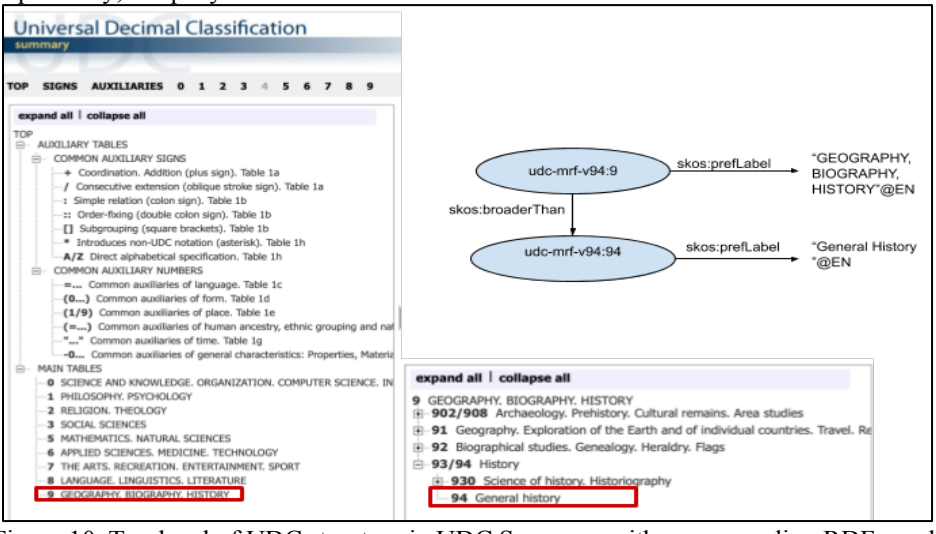


Figure 10. Top level of UDC structure in UDC Summary with corresponding RDF graph representation of class 94.

The UDC MRF contains both simple and combined, i.e., synthesized UDC notations. These combined notations are used for well-established compound subjects/topics that can only be expressed through a combination of simple UDC notations. At the point of use, in libraries and bibliographic databases the analytico-synthetic nature of UDC generates an infinite number of combinations.

Clearly, some aspects of UDC syntax and relationships could be much better managed if expressed in a more formalized way. Within the semantic web stack (RDF, RDFS, OWL) we can find formal ontology languages with a full apparatus of formal logic that have great power in executing reasoning. Our main focus, however, is not to execute reasoning but to make UDC available as a part of reasoning operations. For this purpose, it is sufficient to express UDC using SKOS and RDFS for the edge labels. At the same time, we have to find ways of expressing some aspects of the UDC syntax, in particular

parts dealing with notation synthesis, in a way that is compatible with general web reasoning operations.

For instance, Figure 11 shows how a complex UDC notation 94(492):94(729.885) meaning "The relationship between history of the Netherlands and History of Aruba" can be represented in RDF. To address the problem of the complex number syntax adequately, we deploy two solutions. First, we use "udc-syntax-scheme" to manage different types of notations denoting different types of classes in UDC (main classes, common auxiliaries, special auxiliaries, connecting symbols), as explained in section 4.2. Then we use "blank nodes" to group notational elements belonging to one complex UDC notation. A blank node is a node that has no URI label, just an internal identifier and umbrella pointer to more specific information.

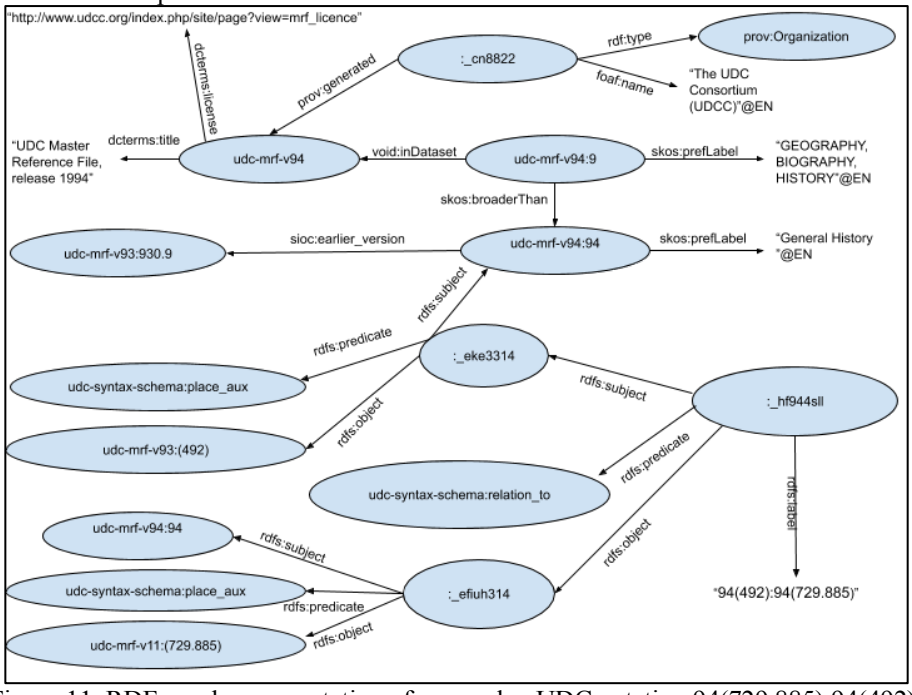


Figure 11. RDF graph representation of a complex UDC notation 94(729.885):94(492).

Our example 94(492):94(729.885) contains the following instances from the UDC scheme:

- a notation from the main table denoting subject 94 General history;
- a connecting sign : (colon) that indicates simple relationships between two subjects (i.e. their notational representation); and,
- two notations enclosed in parentheses which indicate that these are common auxiliaries of place (492) Netherlands and (729.885) Aruba respectively.

In the RDF graph in Figure 11, one can see "udc-syntax-schema:place_aux' as a predicate to the UDC common auxiliary of place (492) Netherlands and as a predicate to the other common auxiliary of place (729.885) Aruba. The syntax predicate 'udc-syntax-schema:relation to'' denotes that two UDC notations, i.e. concepts they represent, are related to each other.xThe blank node ':_hf944sl' indicates that this UDC notation 94(492):94(729.885) is a group that consists of a subject represented by the blank node ":_eke3314," one syntax

element “udc-syntax-schema:relation to” and a predicate represented by a blank node “:_efiugh314.” The node “:_eke3314” groups notational elements 94 and (492) and node “:_efiugh314” groups notational elements 94 and (729.885).

UDC notational elements in a complex number may originate from different UDC MRF versions. The URI of each notational element indicates the version in which these classes were first introduced. For instance, notation of place (492) was introduced in the version v93, and place (729.885) was first introduced in v11 of the UDC MRF.

In bibliographic collections we expect to find all kinds of new combinations of UDC strings created from simple UDC notations taken from different UDC MRF versions. The method of an “atomic” representation of synthesized UDC notations using the concept of blank nodes, as illustrated above, provides flexibility in supporting analytico-synthetic systems such as UDC.

5.3.5 Selection of the RDF schema

As mentioned previously, the UDC look-up service is expected to process requests for URI as well as requests for the full RDF records. Following the parsing stage, URIs for individual notation components and their grouping are generated using RDFs. For the full RDF records we continue to use the SKOS schema extended with several UDC sub-elements. Figure 12 shows the way we mapped the UDC data model (with more specific data elements) to a SKOS schema.

UDC number (notation)	skos:notation	
class identifier	skos:Concept	
broader class	skos:broader	
caption	skos:prefLabel	
including note	skos:note	*udc:includingNote*
application note	skos:note	*udc:applicationNote*
scope note	skos:scopeNote	
examples	skos:example	
see also reference	skos:related	
revision history	skos:historyNote	*udc:revisionHistory*
introduction date	skos:historyNote	*udc:introductionDate*
cancellation date	skos:historyNote	*udc:cancellationDate*
replaced by	skos:historyNote	*udc:replacedBy*
last revision data	skos:historyNote	*udc:lastrevisionDate*

Figure 12. Data elements in UDC LD schema.

The list of 14 data elements contains UDC-specific extensions indicated in italics. Of particular importance for managing historical UDC data, i.e. cancelled classes and their redirection to new classes are *udc:introductionDate*, *udc:cancellationDate* and *udc:replacedBy*.

6.0 Conclusion

Publishing the UDC as LD was a research endeavour which served several purposes, some of which are documented in this paper. The UDC case, as we called it within the DiKG project, is relevant beyond the LD publication of this specific KOS alone. The UDC is a representative of a long-lived, widely used KOS in the bibliographic domain. Bringing the UDC (or parts of it) to the LOD cloud entails also mending and establishing links between the design and uses of KOSs prior to the internet and the semantic web technologies available now, which in principle allow deep linking of knowledge and knowledge ordering systems on an unprecedented scale and semantic richness.

The increase in size and the number of information resources (irrespective of the format, language or provenance) and their accompanying KOSs calls for a new generation of approaches. These should allow us to relate (map) and where possible integrate KOSs and their content, with the aim of enabling cross-domain searching and eventually integration of knowledge across different domains on the level of concepts.

Publishing bibliographic classifications as LD has the following potential advantages:

- **Preserve and build on existing classifications data**: Classification notations in resource metadata may be utilised to enable access to the content of a vast number of already classified information resources, i.e., legacy collections in different formats (textual, non-textual, objects), different languages and scripts;
- **Enable navigation and orientation across knowledge spaces from different domains and across different languages and collections:**
  - internationally used classification schemes are particularly suitable to be used as a pivot to map different general and special KOSs, thus providing more opportunity for the meaningful linking of collections indexed by different KOSs;
  - hierarchical presentation of knowledge fields in bibliographical classifications enables grouping of information on a different level of specificity, and may be used to support information browsing and broadening or narrowing in the information retrieval process and to complement different types of more specific KOSs (e.g. thesauri);
  - associative relationships between different knowledge fields and disciplines may be used to enable the presentation of concept dispersion in the knowledge space as a whole and can help in semantic search expansion; and,
  - classification schemes translated and containing captions in multiple languages can help in managing connections between notations and language terms for concepts or groups of concepts in many languages; they can help in supporting mapping between classification and thesauri or subject heading systems.

Because the UDC is such an exemplary case for those advantages, we took space in the first sections of this chapter to uncover some of the foundations and related terminology relevant to the understanding of these types of KOSs and their function for information discovery and navigation.

The practical task at hand—publishing the UDC as LOD—was informed by the "10Things Guidelines" (cf. Siebes et al. 2019). As is detailed in the guidelines, each publication process includes conceptual parts (selection of what to publish as LD; how to model this selection as an RDF graph; design of namespaces and URI, etc.) and technological parts (how to ensure machine readability; setting up of the web-based service, etc.). However, as in many LD projects, when applying these principles to the UDC case, we found that the transition to a new technology is never a pure mechanical act. It is a research endeavour in its own right. While our discussions were informed by these guidelines, in the end, not totally unexpectedly, our LD publication project followed its own inner logic. In that some of the generic steps became more important, others vanished into the background, and on the whole the process was much more interwoven and iterative than it appeared in the linear description of the guidelines.

To enable the reader to follow our reasoning concerning the choices we made, we had to first explain the nature of the UDC as existing in the standard UDC data source and those enumerated manifold UDC codes as existing in bibliographical systems around the world. We also discussed four specific challenges which are important to the UDC case that are introduced in Section 4 and (Table 1). The result is the design of an LD service, based on the UDC LD model, which responses in different ways to those four challenges

From the outset, it was clear, that provision had to be made for the inclusion of historical data and for resolution of complex UDC notations, hence the "atomization approach." As discussed in this chapter at various places, general bibliographical classifications are complementary to domain-based KOS design. Their power lies in providing access to concepts (and their various historical and contextual layers) as entailed in the massive number of sources (works) indexed in our collective bibliographical past. Being part of the LOD

Challenges	**Solutions in the LD service design**
Longevity and system change over time	Inclusion of historical UDC data and concordances between old and ne classes; UDC version-based URI
Structural complexity	Provision for expressing UDC syntax in RDF (syntax scheme) and the use blank nodes to allow grouping of notational elements
Data ownership	Provision for managing both LOD data set and LD datasets behind the paywall (authentication)
Large usage base and amount of legacy data	Provision for parsing, resolving and identifying UDC notations within a look-up service

Table 1. Challenges and corresponding LD service solutions.

cloud they have the potential of being used as connectors, similar to manner in which encyclopedias are used, between concepts and their embodiment in works.

While the design of the LD look-up service is rather practical and modest, in due course, the service provides the potential to augment the UDC namespace which will become better and richer through its use. For example, we can imagine to store resolutions of complex UDC notation queries over time and morph the look-up services to a storage place (or archive) of "all subjects" and topics ever expressed by UDC. In this way, we could provide automatic concordances and conversion from expressions containing obsolete UDC numbers to current UDC expressions.

In creating the design of the new LD service, it also became evident that the LD model is only one part, and that there were other more important tasks. These include preparation of the UDC information (the parsing model) and determining how the LD service relates to the existing functioning technical UDC database structure and editorial system.

Once again, the UDC case has shown the importance of cross-domain, interdisciplinary collaboration which needs experts well situated in the two (or multiple) knowledge domains among which knowledge exchange is supposed to occur. Additionally, we also carried out a lot of work in the background which was needed to care for fruitful knowledge exchange and synergy or the, so-called "trading zone" to use Peter Galison's (1997) metaphor for collaborations in science and technology. Galison was inspired by anthropological studies on collaboration between different cultures in exchanging goods, despite differences in language and culture. This necessitates, from all those involved, a "translatory capability" that enables a cycle of understanding when it comes to each other's conceptual

frameworks and terminology, and helps reach a mutual shared understanding which is an indispensable requisite for true interdisciplinary collaboration.

Notes

1. Digging into the Knowledge Graph (DiKG) project, 2017-2019 (https://diggingintodata.org/awards/2016/project/digging-knowledge-graph) is funded by the Institute of Museum and Library Services (IMLS) grant within Round 4 of the Digging into Data Challenge grant under the umbrella Trans-Atlantic Platform for the Social Sciences and Humanities. The DiKG research focused on providing means of support for the self-organizing process of knowledge creation in the Semantic Web by enhancing findability and storage for humanities and social science Linked Open Data datasets using the artifacts and organization systems.
2. Bibliographic classifications comprise library classifications (designed for library shelf arrangement) and classifications designed for logical organization and information retrieval of entries in bibliographical services (including abstracting and indexing databases). Library classification are usually less detailed and are structurally simpler than classification designed for use in bibliographies.
3. Classification notations are sometimes called classification codes, classification symbols, classmarks or classification numbers (if notational system is numerical).
4. Classification schemes are not concerned greatly with verbal class descriptions. They differ, in that respect, from thesauri and subject headings which are primarily concerned with natural language terms used to express concepts in order to manage and control the consistent use of terms. These, for instance, provide alphabetical listing of approved natural language terms (indexing terms) to be used for certain concepts, resolving ambiguities such as homonymy, synonymy or polysemy in a certain field of knowledge but are unable to group or provide logical order of knowledge areas. For this reason, in practice, thesauri and subject headings are usually used as complementary to classifications.
5. The presentation of knowledge space as a whole is a feature of e.g., UDC and Dewey as opposed to Library of Congress *Classification* (LC*C*) or Bliss Bibliographic Classification (BC2) which function as a series of special classifications.
6. Basic principles on how UDC works are described in numerous books and articles (e.g. McIlwaine 2007). Summary instructions are provided in introduction to all printed UDC editions and an instructional text is provided in UDC Online schedules (www.udc-hub.com).

References

Baker, Thomas et al. 2011. "Library Linked Data Incubator Group: final report: W3C Incubator Group Report Report 25 October 2011." https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Cleverdon, Cyril W. 1962. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. ASLIB Cranfield Research Project. London: Aslib. https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/836/1962.pdf?sequence=2&isAllowed=y.

Galison, Peter. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago: The University of Chicago Press.

Isaac, Antoine and Aida Slavic 2009. "Using SKOS in Practice, with Examples from the Classification Domain." Presentation at *International UDC Seminar 2009, The Hague, The Netherlands, Oct. 29-30, 2009*. https://slideplayer.com/slide/4392/

Kwasnik, Barbara H. 1999. "The Role of Classification in Knowledge Representation and Discovery." *Library Trends* 48, no. 1: 22-47. https://www.ideals.illinois.edu/bitstream/handle/2142/8263/librarytrendsv48i1d_opt.pdf?sequence=1

The Linked Open Data Cloud. https://lod-cloud.net/#diagram. Web resource. Permalink https://web.archive.org/web/20201215080839/https://lod-cloud.net/

"MARC 21 Format for Authority Data," 1999 ed. Library of Congress. Update No. 31 (December 2020). https://www.loc.gov/marc/authority/

MartÍnez-Ávila, Daniel, Richard P. Smiraglia, Rick Szostak, Andrea Scharnhorst, Wouter Beek, Ronald Siebes, Laura Ridenour and Vanessa Schlais. 2018, "Classifying the LOD Cloud: Digging into the Knowledge Graph." *Brazilian Journal of Information Science: Research Trends* 12, no. 4: 6-10.

McIlwaine, Ia C. 2007. *The Universal Decimal Classification: A Guide to its Use*. The Hague: UDC Consortium.

Piros, Attila. 2015. "New Automatic Interpreter for Complex UDC Numbers." *Extensions & Corrections to the UDC* 36-37: 37-50.

Piros, Attila. 2017. "The Thought Behind the Symbol: About the Automatic Interpretation and Representation of UDC Numbers." In *Faceted Classification Today: Theory, Technology and End Users: Proceedings of the International UDC Seminar 2017, London (UK)*, 14-15 September, ed. Aida Slavic and Claudio Gnoli. Würzburg: Ergon Verlag, 203-18.

Rayward, Boyd W., ed. 1990. *International Organization and Dissemination of Knowledge: Selected Essays of Paul Otlet*. FID 864. Amsterdam: Elsevier.

Riesthuis, Gerhard J. A. 1997. "Decomposition of Complex UDC Notations." *Extensions & Corrections to the UDC* 19:13-19.

Riesthuis, Gerhard J. A. 1998. *Zoeken met woorden: hergebruik van onderwerpsontsluiting*. Amsterdam: Leerstoelgroep Boek-, Archief- en Informatiewetenschap.

Siebes, Ronald, Gerard Coen, Kathleen Gregory and Andrea Scharnhorst. 2019. "Linked Open Data. 10 Things toward the LOD Realm: The "I" in FAIR in a Semantic Way". Document Library-Mozilla Sprint May 2019. https://doi.org/10.5281/zenodo.3471806

SKOS. 2009. *Simple Knowledge Organization System: Reference*: W3C recommendation 18 August 2009, ed. Alistair Miles and Sean Bechhofer. http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Slavic, Aida. 2005. *Classification Management and Use in a Networked Environment: The Case of the Universal Decimal Classification*. PhD diss., University College London, University of London. http://discovery.ucl.ac.uk/1334914/

Slavic, Aida. 2006 "UDC in Subject Gateways: Experiment or Opportunity?" *Knowledge Organization* 3: 67-85.

Slavic, Aida. 2008. "Faceted Classification: Management and Use." *Axiomathes* 18, no. 2: 257-71.

Slavic, Aida. 2017. "Klasifikacija i Library Linked Data (LLD)." In *Predmetna obrada: pogled unaprijed: zbornik radova*, ed. B. Purgaric and S. Spiranc. Zagreb: HKD, 13-37.

Slavic, Aida and Antoine Isaac. 2009 "Identifying Management Issues in Networked KOS: Examples from Classification Schemes." Presentation at the *8th NKOS Workshop, Corfu, Greece, l October 2009*. https://slideplayer.com/slide/6385705/

Slavic, Aida and Sylvie Davies. 2017. "Facet Analysis in UDC: Questions of Structure, Functionality and Data Formality." *Knowledge Organization* 44: 425-35.

Slavic, Aida, Christophe Gueret, Chris Overfield and Andrea Scharnhorst. 2013. "Library Linked Data and its Relationship to Knowledge Organization Systems: The Case of UDC." Presentation at Workshop on User Interaction Built on Library Linked data (UILLD*).* Pre-conference to the 79th World Library and Information Conference, Singapore, 16 August 2013. https://files.dnb.de/svensson/UILLD2013/UILLD-submission-9-formatted-final.pdf

Smiraglia, Richard P. 2001. *The Nature of "A Work": Implications for the Organization of Knowledge*. Lanham, Md: Scarecrow Press.

Smiraglia, Richard P. and Charles van den Heuvel. 2013. "Classifications and Concepts: Elementary Theory of Knowledge Organization." *Journal of Documentation* 69: 360-83.

Smiraglia, Richard P., Andrea Scharnhorst, Almila Akdag Salah and Chen Gao. 2013. "UDC in Action". In *Classification and Visualization: Interfaces to Knowledge*, ed. Aida Slavic, Almila Akdag Salah and Sylvie Davies. Würzburg: Ergon Verlag: 259-270. http://arxiv.org/abs/1306.3783

Suchecki, Krzysztof, Almila Akdag Salah, Cheng Gao and Andrea Scharnhorst. 2012. "Evolution of Wikipedia's Category Structure." *Advances in Complex Systems* 15 (supp01): 1250068–1. doi:10.1142/S0219525912500683

Szostak, Rick, Andrea Scharnhorst, Wouter Beek and Richard P. Smiraglia. 2018. "Connecting KOSs and the LOD cloud." In *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto (Portugal)*, ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Baden-Baden: Ergon Verlag, 521-529.

Szostak, Rick, Richard P. Smiraglia, Andrea Scharnhorst, Ronald Siebes, Aida Slavic, Daniel MartÍnez-Ávila and Tobias Renwick. 2020. "Classifications as Linked Open Data: Challenges and Opportunities. In *Knowledge Organization at the Interface. Proceedings of the Sixteenth International ISKO Conference 6–8 July 2020 Aalborg, Denmark*, ed. Marianne Lykke, Tanja Svarre, Mette Skov, Daniel Martínez-Ávila. Advances in Knowledge Organization 17. Baden-Baden: Ergon Verlag, 436-45.

Tennis, Joseph and Stuart A. Sutton. 2008. "Extending the Simple Knowledge Organization System (SKOS) for Concept Management in Vocabulary Development Applications." *Journal of the American Society for Information Science and Technology* 59: 25-37.

Tillet, Barbara. 2015. "Complementarity of Perspectives for Resource Descriptions." In *Classification & Authority Control: Expanding Resource Discovery: Proceedings of the International UDC Seminar 2015, Lisbon (Portugal) 29-30 October 2015*, ed. Aida Slavic and Maria Ines Cordeiro. Würzburg: Ergon Verlag, 19-30.

van den Heuvel, Charles. 2008. "Building Society, Constructing Knowledge, Weaving the Web: Otlet's Visualizations of a Global Information Society and his Concept of a Universal Civilization." In *European Modernism and the Information Society: Informing the Present, Understanding the Past*, ed. W. Boyd Rayward. London: Ashgate Publishing, 127-53.

van den Heuvel, Charles and W. Boyd Rayward. 2011. "Mondothèque: A Mulitmedia Desk in a Global Internet." In *Science Maps as Visual Interfaces to Digital Libraries: 11th iteration (2011) of Places and Spaces: Mapping Science*, ed. K. Börner and M. J. Stamper. http://scimaps.org/maps/map/mondoth%C3%A8que_multimed_123/.Scott

Zeng, Marcia and Philipp Mayr. 2019. "Knowledge Organization Systems (KOS) in the Semantic Web: A Multi-dimensional Review." *International Journal on Digital Libraries* 20: 209–30.

Charles van den Heuvel
Huygens Institute for the History of the Netherlands
University of Amsterdam

Veruska Zamborlini
University of Amsterdam

Chapter 6
Modeling and Visualizing Storylines of Historical Interactions
Kubler's *Shape of Time* and Rembrandt's *Night Watch*‡‡

Abstract
The Golden Agents research infrastructure enables analyses of interactions between/within the creative industries of the Dutch Golden Age by bringing various heterogeneous (un)structured datasets of cultural heritage institutions together in linked open data. One of the challenges is the modeling of ontologies for the historical processes of the interactions between various branches, and between the production and consumption of these industries. These processes are described as multiple narratives for which we use the concept "storifying data." Here we try to demonstrate that current attempts to model temporality of historical data in linked data such as CIDOC-CRM, OWL-Time or PeriodO are too limited and that we might learn from historical conceptualisations of periodisation and duration. In particular, we will focus on George Kubler's *The Shape of Time: Remarks of the History of Things* (1962) and claim that his approach of the history of art as a system of linked historical sequences of formal relations is still relevant for modeling time and historical processes in ontologies and standards. The model "storylines of historical evidence" and the relevance of Kubler's views on duration and sequence will be demonstrated by the very rich case of the (re-)uses of Rembrandt's *Night Watch*.

1.0 Golden Agents: Creative industries and the Making of the Dutch Golden Age

During the Dutch Golden Age, Amsterdam developed into the world's center for trade, science and art, and was known for the size and scale of its creative industries, especially for paintings and book production (Rasterhoff 2017; Pettegree and Weduwen 2019). Until now, monographs have been written on famous artists and authors, but information on lesser known professions such as silversmiths, playwrights or appraisers in that period is still oblivious. We are even less informed about the consumption of cultural goods in Amsterdam during the Dutch Golden Age.

The project Golden Agents: Creative Industries and the Making of the Dutch Golden Age by using a combination of semantic web and multi-agent technologies aims at developing a sustainable infrastructure to study relations and interactions between 1) the various branches of the cultural industries and 2) between producers and consumers of creative

‡‡ The authors wish to thank Marten Jan Bok and Frans Grijzenhout (University of Amsterdam) for finding more about unknown paintings of Rembrandt mentioned in archival sources. Furthermore, they are grateful to Arianna Betti (UvA), Greta Adamo and Giancarlo Guizzardi (FBK-Trento) for fruitful discussions on the subject of this work related to its epistemological account, which may appear in future work.

goods across the long Golden Age of the Dutch Republic, in particular in Amsterdam. The project will link distributed, heterogeneous resources (both existing and new) on the production of the creative industries in the Dutch Golden Age from heritage institutions such as the Rijksmuseum, KB National Library of the Netherlands, and the RKD: The Netherlands Institute for Art History, and of academic institutions such as the data bases of painters in the Low Countries, ECARTICO and of theatre productions in Amsterdam in the 17th century ONSTAGE, both produced by the University of Amsterdam. Consumption remains an under-investigated topic with regard to the creative industries in the Dutch Golden Age. The digitisation of the enormously rich collection of the notarial acts (more specifically the probate inventories) in the Amsterdam City Archives, will provide data on the possessions of cultural goods by the inhabitants of all layers of society in Amsterdam as one of the most important global centers in the world in the 17th century. Finally, we believe that these big data of the production and consumption can provide more insight in concepts of creativity and innovation in the Dutch Golden Age and potentially contribute to the history of taste. For instance, Angela Jager (2020), in her PhD research, was able to nuance the view in the historiography of Dutch painting of the Dutch Golden Age that history paintings were the most expensive and the highest praised works of art. On the basis of prices mentioned in a few probate inventories in the notary acts she revealed that much cheaper versions were produced for the lower end of the art market. This revelation is promising because the Golden Agents has the intention of opening up the contents of 2,000,000 scans of notarial deeds such as baptism, marriage and burial registries, and other document types of the Amsterdam City Archives that give insight into the households of the more common Amsterdamer and not just of the elite culture during the Dutch Golden Age. This allows us to (partially) construct storylines about inhabitants of Amsterdam and the (type of) objects they possessed or traded.

2.0 Storifying data: Modeling historical knowledge

> Historical Truth, because it has nothing to correspond with, can only be defined as coherence with the understanding of the past (documents, including material culture) and the concepts we share with our predecessors and interlocutors (Shaw 2010, 6-7).

The Golden Agents research infrastructure enables analyses of interactions within the creative industries of the Dutch Golden Age by bringing various heterogeneous (un)structured datasets of cultural heritage institutions together in linked open data (LOD). One of the challenges is the modeling of ontologies for the historical processes of the interactions between various branches, and between the production and consumption of these industries. These processes are described as multiple narratives for which we use the concept "storifying data" (Zamborlini, Betti and Heuvel 2017).

These multiple stories developed over time in parallel orders, for instance the order in the making of an object (from idea to final product), the order of an object in the artistic life or oeuvre of its maker, the order between the original object and copies and transformations hereof and finally the order of the object within history or in fictional time depicted in paintings or described in stories. The parallel development of these multiple stories over time can be described in events to capture the historical discourses of that expanding cultural world in ontologies.

Ryan Shaw, in his Phd dissertation *Events and Periods as Concepts for Organizing Historical Knowledge*, stated eloquently that knowledge organisation (KO) is not applied to history, but that history is a form of KO. Historians produce knowledge of the past by organizing the past, by organizing documents, concepts and the systems that facilitate the processes of KO (Shaw 2010, 2 and 94). This requires not only an understanding of the applied ontologies, but also of how the historical concepts in the Golden Agent projects are organised. Shaw distinguishes three senses of the word "history:"

1) history-as-past i.e. all actions and happenings before the present time;

2) history-as-portrait as referring to some organised structure to represent the past in the form of a narrative –a story; and,

3) history-as-practice that refers to history as a discipline.

The latter also encompasses the ways historians engage with the cultural heritage of material culture and documents. Shaw rightly stresses the risk that we develop systems that portray history rather than supporting doing history. This in his view requires describing the concepts historians construct in order to describe the past and of the documents they use to describe them: i.e., history as conceptualisation (Shaw 2010, 4-5). In a recent paper Igor Frank (2019) advocates an applied ontology for digital history informed by philosophy of history to make the conceptualisations of historians explicit. His applied ontology approach to represent historical reality is directed at: 1) grasping historical processes; 2) representing multiple perspectives of different actors involved in historical events; and, 3) representing views according to different historical sources. Although all these facets of this multi-perspectival representation of knowledge make part of the Storifying Data Model, in this chapter we will in particular discuss the modeling of historical processes by focusing on time and periodisation that is not included in his discussion.

2.1 Periodisation and events in historical discourse

Frank's ([6]) warning not to commit "cliocide" by modeling away all the crucial subtleties of historical reality is well taken. However, if we follow the observation of Shaw, history itself is a form of KO (and not just applied to history); it is not sufficient to model the representation of that reality from multiple perspectives, but characteristics of history of KO should be modelled as well. One important, if not the most important, characteristic of history as KO is the preoccupation of historians with the organisation of events in time, that is, the representation of historical events in a temporal order.

The representation of time and temporal order in linked data (LD) goes beyond the common practice in applied ontology in digital history of mapping a historical event in a given place to the right (Georgian, Julian, Chinese etc.) calendar. Important is the PeriodO initiative to create a gazetteer of period definitions. However, it is not sufficient to map vague period names to more precise chronological coordinates as confined events. More flexible at first sight seems the development of the ChronOntology gazetteer (iDai.chronontology) that connects temporal (and spatial) information of "types." (Schmidle et al. 2016). In this way, for instance, the type "painting" as an object of material culture of the Italian Renaissance could be linked as (space-time) to an area described as Renaissance regardless of what we know about its extent. This allows periods, such as the "Renaissance" to take place at different times and in different regions, for instance the Renaissance in Low Countries. However, all these valiant attempts to create time models that can handle some fuzziness in periodisation in practice (regardless from the question of

how relevant it is to stylistic classifications for periodisation as we will discuss in the next section) are still calendar-focused and lack a conceptualisation of time itself. Recently, the theoretical physicist Carlo Rovelli (2018, 103) in *The Order of Time* argued that there is no need to choose a privileged variable and call it time. It would suffice to have a theory of dynamic relations that tells us how the things we see in the world vary with respect to each other. Probably these different world views and various perceptions of time in different cultures explain why so many philosophers, scientists and historians have tried to get a grip on periodisation and temporality in their disciplines. Toyoshima (2019) tried to describe the foundations of an ontology of time with a practical function in the domain of the digital humanities and opted on the comparative analysis of adherents of presentism, eternalism and the so-called growing block theory for the latter because it acknowledges in the temporal ontology the past (unlike) presentism, but not the future (unlike the eternalists).

Kauppinen et al. (2010) tried to explain the relevance of imprecise temporal intervals for information retrieval in the domain of cultural heritage. Although both studies provide some points of reference for annotation of cultural objects in cultural heritage applications in more or less precise time intervals, problems remain with the ontological representation of the co-occurence of multiple natural/real and fictional/abstract time intervals. Galton (2018) brings such problems to the front in a comparative analysis of the treatment of time in the upper ontologies BFO, DOLCE and GFO in which he points to their respective inconsistencies in modelling space-time with Einstein's relativity theory. This might seem far-fetched as bridging the gap between insights of physical theories and philosophical debates about the nature of time is not the aim of our model. However, we need to get a grip on issues of realism versus conceptualism of time and of multi-dimensional representations of space-time, with abstract or fictional notions of time when we try to model concepts of events or durations in (the making of) cultural objects. How do we model for instance the co-occurence of time of Gustave Courbet's symbolic portrayal of "L'Origine du Monde" with his depiction in close-up of the vagina of a naked woman in his provocative painting of 1866? Or how do we model the multiple events of the story of another famous painting, that of the *Adoration of the Magi* of Gentile da Fabriano of 1423, in which the three kings are appearing and disappearing behind rocks to express the (narrative) time of their journey in a (as art historians in the German language call it beautifully) "kontinuierende Darstellung" in one framed panel. We cannot discuss all these conceptualisations of time. Only those will be referred to that are relevant for modeling our concepts of events, narration and historical evidence.

One of the most classical examples of periodisation in the historiography of the historical disciplines is Fernand Braudel's conception of serial history in events (very short term); conjuncture or cyclical time (intermediate duration) and "longue durée" (structural change) that dominated the French historiography of the Annales School (Tomich 2012). Braudel's notion of time, i.e., of plural time, is interesting because it unites multi-layered geophysical-social space and historical time. His concept of conjuncture borrowed from economics that integrates correlations observed across multiple quantitative time series is of particular interest for the discussion further below of Kubler's *Shape of Time*. However, Braudel's model of time is also problematic because events are not necessarily short happenings but can vary in time and also be read in terms of narratives (Ricoeur 1980 and 1984; Shaw 2010, 53). Moreover, this interpretation of an event as something that happened over a very short period of time does not coincide with the use of historical events within the semantic

web paradigm. To this end Shaw (2013), on the basis of his dissertation, proposed a semantic tool informing users about events in historical discourse and formulated the requirements and criteria to individuate them. He distinguishes between events as concrete individual things and events as abstractions from narratives. Shaw finally defines an event as "something that happened" and stresses the point that unlike other definitions it does specify a change of state or a distinction of events from states or processes.

2.2 Narration and visualisation of historical events and processes

In the context of our model Shaw's semantic tool is not only of interest for its definition of events in relation to temporality and periodisation, but also for its role in selecting events in relation to documents. In Shaw's view (2013, 42) a document can be both a portrayal of an event and provide some evidence for some event, i.e., document-as-evidence. A document can become historical evidence after a historian has studied and made some assessments about the status of the document as a less or more trustworthy representation of the past. The latter is only possible by a portrayal or narration of the event. Since events are not explicitly named, the kind of relationship between document and the event needs to be visualised by putting it into context. In short, events need to be linked to time, place and related concepts, as well as put in the context of narratives. For understanding the historical discourse, a variety of (one may add sometimes conflicting) stories need to be told about the past (compare Shaw 2013, 45).

While the modeling of periods and events in standards such as CIDOC-CRM is not always straightforward, capturing the role of narratives in historical discourse and the role of documentation as source of evidence is even more difficult. Standards developed in the cultural heritage domain such as CIDOC-CRM support the structuring of the metadata of material artefacts and documents as cultural or bibliographic objects quite well. However, they are not always suitable for modelling (meta-)data for historical research. Frank (2019) for that reason set up a case study using Ontology Design Patterns in combination with DOLCE to explain the procedure of "colligation" to trace and to classify the relations between events based on chronological relations, mereological relations and causal relations (visualised in UML diagrams) in order to locate them in their historical context. At the same time, he explained that his Description and Situations Ontology Design Patterns (DnS) all can be expressed in CIDOC-CRM classes as well.

Similar attempts bring historians together in the DataforHistory.org consortium. It was created during a two-day meeting (23-24 November) in 2017 in Lyon on the initiative of Francesco Beretta and George Bruseker with the aim to develop ontologies for history that are complementary to the CIDOC-CRM, but still fully comply to this standard in order to guarantee optimal interoperability between the data of historical research projects and of cultural heritage institutions.

Within the Data for History consortium a working group concentrates on the modelling of storylines.[1] It was brought together by Charles van den Heuvel and includes members of the very interesting Narratives in Digital Libraries project (Bartalezi, Meghini and Metilli 2017) that models and connects narrative events in literature, but unfortunately does not allow for representing multiple time-sequences. Promising is the multiple strata (material, cultural, institutional) approach of life cycles of cultural goods that Karl Pineau presented at the 3rdData for History meeting (Pineau 2019). Alex Butterworth organised a panel at the Digital Humanitites (DH) 2019 conference in Utrecht that discussed alternative

ways of visualizing literary and historical narratives and chronotypes. (Butterworth et al. 2019). In the context of the development of our model of storylines that provide insight in historical evidence the work of another member of this working group, Regina Varniene-Janssen is relevant. She contributes to the Virtual Electronic Heritage Information System VEPIS project that develops long-term strategies to support interoperability regarding the authenticity and provenance of digital content of the National Library of Lithuania with other cultural heritage institutions such as Europeana (Varnienė-Janssen and Kuprienė 2018).

Although the creation of the model storylines of historical evidence started before the creation of the Data for History.org consortium it brings together several of the features that the "storylines" working group members are developing separately in the context of their own projects. Similar to the Narratives in Digital Libraries it connects narrative events but differently it allows for representing multiple time sequences. The latter is also the case in the life cycles of the cultural goods model of Pineau, but our model is not restricted by his three material, cultural and institutional strata, or Butterworth's macro, meso and micro levels that resonate Braudel's model of duration. Our storyline model includes in principle infinite parallel time sequences. The visualisation of these storylines is not only intended to explore and to switch between events and narratives, but also as an instrument of critical inquiry to assess the quality of the data and discourses on the basis of their provenance. In that regard we try with the model to explore the potential of the graphic arts to query knowledge production in a critical way from a humanist perspective as advocated by Drucker (2009 and 2014). However, we do not try in the model to distinguish these graphical explorations from more technical, analytical models of KO, but rather to reconcile them.

For the development of the storylines of historical interactions model, we were inspired by the work of George Kubler, *The Shape of Time: Remarks of the History of Things* (1962). This study is not only interesting for bringing in views on temporality, periodisation and narration in historical discourse in addition to those of authors discussed by Ryan Shaw (2010) in the rich historiographical overview when discussing their interpretations in his conceptualisation and modeling of periods and events in organizing historical knowledge. Kubler's *Shape of Time* is of particular of interest for our model because his discussion of the concepts of temporality, periodisation and narratives is more closely related to our aim to develop an infrastructure that can be used by researchers to use cultural heritage data and that allows cultural and art historians to deal with questions concerned with style and innovation, but also of replication to explain the boom of the creative industries of the Dutch Golden Age.

3.0 Kubler and *The Shape of Time*

> The "history of things" is intended to reunite ideas and objects under the rubric of visual forms: the term includes both artifacts and works of arts, both replicas and unique examples, both tools and expressions- in short, all materials worked by human hands under the guidance of connected ideas developed in temporal sequence. From all these things a shape in time emerges (Kubler 1962, 9).

3.1 The *Shape of Time*: Remarks on the history of things

In the preamble of the *Shape of Time* (1962), Kubler explains the motivation of his provocative work in the history of art. Instead of focusing on the work of art as a symbolic

expression of which its meaning needs to be explained, Kubler proposes another definition of art "as a system of formal relations." While in his view no meaning can be conveyed without form, structural forms can be perceived independently from meaning. The purpose of *The Shape of Time* is to (viii): "draw attention to some of the morphological problems of duration in series and sequence." Kubler's work is so much discussed by art historians because it questions and even dismisses the usefulness of their commonly used words to describe the arts, such as "style" which both is used to group objects with similar characteristics over a longer period of time and to describe often several successive changes in features within the oeuvre of an individual artist during his lifetime. In the context of the discussion above it is also interesting that Kubler discusses problems related to narration such as the limitations of biographies describing the lives of artists to describe the talent and the genius of artists. To paraphrase Kubler, both Leonardo and Raphael were talented; Romano was as well, but as a follower just had "bad luck" (7). Kubler proposes an alternative history, i.e., a history of things that consists of ideas and of objects ranging from unique artifacts to replicas all connected in temporal sequence. It is the task of the historian, similar to that of the astronomer, to collect "ancient signals" and transformations hereof in order to develop compelling theories about distance and composition. To order and class events extracted from these signals and to verify and test all their evidence is the principal task for the historian (20-21). Kubler classes things in formal sequences not so much as objects in time, but as sequences of solutions. In his example of churches built between 1140 and 1350 in Northern Europe, Kubler states (37): "The formal sequence is not 'cathedrals'. It is more like 'segmented structures with rib vaults.'" This allows him to distinguish fashions with a very brief duration as being without substantial change in the connected chain of solutions (39). The challenge is to individuate to find such sequences of solutions to find the things that shape time.

3.1.1 Things

> There are prime objects and replicas as well as the spectator's and the artist's views of the situation of the work of art in time (Kubler 1962, 39).

Things in Kubler's model include not only objects and ideas, but perceptions from multiple perspectives hereof as well. He distinguishes between prime objects and replications. Prime objects are similar to prime numbers that have no divisors as themselves and therefore cannot be decomposed in entities. Replications on the other hand comprehend an entire system of replicas, reproductions, copies, reductions and other derivations of an important work of art. Since a formal sequence can only be deduced from things we need an understanding from this system of prime objects and replications. While the number of prime objects for their uniqueness is very limited, our knowledge of sequences has to be mainly based upon replications. Therefore, most of our evidence is based on copies or other derivatives. This system of prime objects and replications has a logical order in the sense that a replication can never precede the prime object. This object however, can live on over a long period of time in all sorts of derivatives. For that reason, Kubler speaks (55) of a "systematic age of each item in a formal series according to its position in the duration." Old and new series of things coexist simultaneously at every historical moment, save the first. The reason for this is historical change in which the conditions and circumstances alter from one moment to the other. However, these processes of change and in our attitudes

towards them, shape the occurrence of things. As Kubler explains it eloquently (62): "We cultivate 'avantgardisme' together with the conservative reactions that radical innovation generates." He propagates things in processes of invention, repetition and discard. The propagation of things as processes of invention, repetition and discard needs to be measured in time.

3.1.2 Time

> Calendar time indicates nothing about the changing pace of events (Kubler 1962, 83).

Like most historians, philosophers and scientists that try to define time, Kubler tries to distinguish between absolute or solar time on the one hand and time ordered by mankind on the other. For the latter he deplores the lack of sound theories of temporal structure and speaks of "few old ways of grouping events" (96). Nevertheless, these ways of grouping events are not random, but can be measured systematically, hence the aforementioned term "systematic age." Within the historical disciplines Kubler is not so much interested in divisions in calendar time that arrange one event after the other. Similarly, he sees decades or centuries as arbitrary intervals and prefers the length of human generation as a unit. For that reason, Kubler bases his measurements not on numbers but on relations between events that express variations in duration in the lives or successive generations of artists. He analyzes variations in pace, differences between slow and fast happenings of events in tribal or urban cultures or in the lives of individual artists (86):

> The pace and tone of an artist's life can tell us much about his historical situation, although most artist's lives are uninteresting. They fall usually into routine divisions: apprenticeship, early commissions, marriage, family, mature work, pupils and followers. Sometimes the artist travels, and occasionally his path crosses those of more colorful persons.

Of particular interest are for Kubler the shape and forms of durations that last longer than a single human life (more to the point, the working life of the artist) or which require the time of more than one person for which he uses the term collective durations. He proposes to use "indiction" as the module. It is of course an approximation, but Kubler bases this module on a time span of ca. 50-60 years as the usual duration of an artist's life which can be subdivided in four stages—preparation, early, middle and late maturity—of about 15 years. Certain time intervals of linked events in the history of art—for instance to describe technical developments such as the early history of the rib-vaulted construction of Gothic architecture—according to Kubler, take intervals of doubled 60 years duration. Kubler calls it an empirical description of sequences in the history of art that allows us to avoid talking about styles of art, but instead to analyze the history of special forms among related examples occuring in limited regions (101-3). Kubler introduces new classes of duration when the series of successive events temporarily are interrupted, the so-called intermittent classes. There are two kinds of intermittent classes: those which lapse inside the same cultural grouping and those that span different cultures. In the history of art, the first kind of intermittent class is important for understanding the revival of specific forms within a specific culture, for instance the re-use of the classical architectural language in the Italian Renaissance. The second kind of intermittent class, that of transcultural diffusion, is of particular interest for the Golden Agents projects to describe the period of the cross-fertilisation between decorations on Chinese porcelain and Dutch earthenware when the art market of Amsterdam in the Golden Age opened up to the Far East. Finally, Kubler distinguishes between wandering and simultaneous series. An example of the first series is

the re-use of the same architectural ornaments of the Italian Renaissance in a later stage in the Dutch Republic that for instance were transmitted by examples in treatises and model books. Simultaneous series describe the opposite, that is different classes of specific forms in the same time interval. In short, Kubler does not provide a periodisation of one continuous timeline (compare Braudel's events, conjuncture and "longue durée") but his systematic age consists of relationships between changing classes of forms and changing classes of duration in multiple sequences.

3.1.3 Visualing the *Shape of Time*

> Instead we can imagine the flow of time as assuming the shapes of fibrous bundles with each fiber corresponding to a need upon a particular theatre of action, and the lengths of the fibers varying as to the duration of each need and the solutions to its problems (Kubler 1962, 122).

It is surprising that Kubler's art-historical analysis with the title *The Shape of Time* has only one tiny image hidden away in a footnote to the text. It concerns a visualisation of a directed graph (that is a network in which the relations (links) between the nodes are not reciprocal) provided to Kubler by his colleague at Yale University, the mathematician Øystein Ore, one of the pioneers of graph theory with whom he corresponded about the concept of series and sequences. We do not know exactly what Kubler asked but Ore's reaction was supportive, but at the same time somehow critical (33-34 n3):

> In attempting to give a systematic presentation of so complex a subject matter one would be inclined, as in the natural sciences, to look to the mathematicians for some pattern to serve as a descriptive principle. The mathematical concepts of series and sequences came to mind but after some thought these appear to be too special for the problem at hand. However, the less known field of networks or directed graphs seems to be considerably more suitable. We are concerned with the variety of stages in the creativity of the human race ... There are a variety of directions that may be selected. Some represent actual happenings. Others are only possible steps among many available ones. Similarly, each stage may have occurred among several possible steps leading to the same result ... The graphs shall be a-cyclic, that is, there exists no cyclic directed path returning to its original stage. This essentially corresponds to the observation about human progress that it never returns to the previous conditions.

The quotation from Ore's reply to Kubler (only partly represented here) is a long one, but we include it for two reasons. First of all, it is a direct reference to the expectations of the potential of graph theory in the future that we use now to model the data and agents of the Golden Agents project using semantic web and artificial intelligence technologies to which we will return later when we discuss the implications of using Kubler's model of time for our ontologies and mappings to existing ontology standards. Second, Ore's reply reveals how Kubler tried to legitimise his alternative model of time in art history with expertise from other disciplines such as, in this case, mathematics. However, it can be questioned whether he fully understood the implications of Ore's picture of the mathematical concept of directed graph or network. This might even be the reason perhaps why he just left the discussion of the network as a note. Kubler certainly imagines his model of time, at least part of it, as a network when describing the sequence of forms in duration (37-8):

> The closest definition of a formal sequence that we now can venture is to affirm it as a historical network of gradually altered repetitions of the same trait. The sequence might therefore be described as having an armature. In cross section let us say that it shows a network, a mesh or a cluster of subordinate traits; and in long section that it has a fiber-like structure of temporal stages, all recognizably similar, yet altering in their mesh from beginning to end.

When we try to envision Kubler's description it becomes clear that it is quite different from Ore's picture of a directed network. In that respect recent 3-dimensional timeline tools such as that developed by Matt Jensen (2006, fig. 4) for NewsBLIP might express Kubler's idea better (our Figure 1).

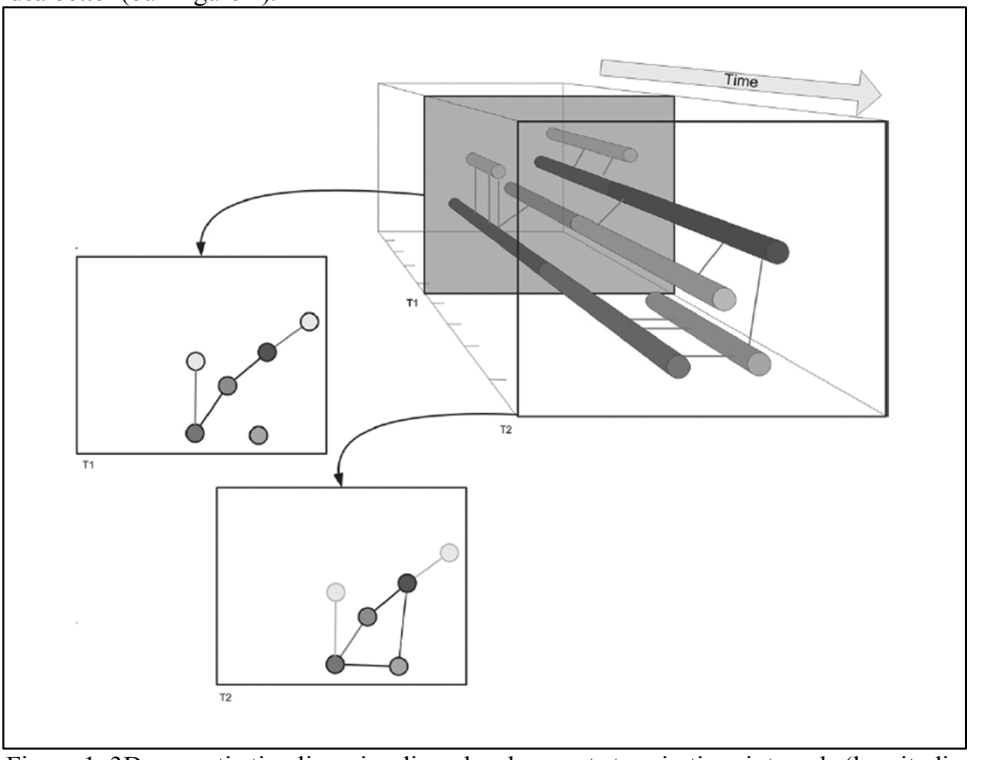


Figure 1. 3D semantic timeline-visualises development story in time-intervals (longitudinal) and network of relations between storylines (transversal) similar to Kubler's description of fibers of duration and networks in cross-section (Jensen 2006).

The limitations of Kubler's different reading of the role of networks could have in his model of duration compared to Ore's interpretation thereof becomes apparent when he tries to juxtapose his fibers of duration with the circular lenses of followers of "Strukturforschung" that tend to read the expressions of poets and artists of one place and time as radial or central patterns varying in thickness according to their antiquity (27 and 121-2). It seems that Kubler was not able to grasp the full potential of Ore's explanation of the directed network of his model by reading the formal sequences of durations just in longitudinal and transversal ways (i.e. strictly flat) instead of exploiting the full potential of the graph in which the longitudinal and transversal allow for traversing pathways in more than two dimensions.

3.2 The *Shape of Time* reconsidered: Kubler on style and historical time

> Style is like a rainbow. It is a phenomenon of perception governed by the coincidence of certain physical conditions .... Whenever we think we can grasp it, as in the work of an individual painter, it dissolves into the

farther perspectives of the work of that painter's predecessors or his followers, and it multiplies even in the painter's single works (Kubler 1962, 129).

Directly after its publication Kubler's *Shape of Time* received much attention in the world of art history, anthropology, linguistics, philosophy and other disciplines. We cannot discuss all the reviews. For our model it is interesting to see how he reacted to the various comments. Twenty years after its publication, Kubler wrote a comment (1982) with the title "*The Shape of Time* Reconsidered." In this comment he replied to some of his critics such as Priscilla Colt (1963) who had questioned whether the study of style necessarily is precluded by the study of formal sequences. In reply to her critical remarks, Kubler referred to his later publications (Kubler 1967 and [1979] 1987) with elaborations on his view on style. These later works are of interest because Kubler published herein additional "axiomas" (1967) later turned into reduced "postulates" ([1979] 1987) to explain his views on style in relation to those of art historians. Kubler formulated the following special postulates about visual style ([1979] 1987, 167).

- Style comprises acts undergoing change
- Style appears only among time-bound elements
- No human acts escape time
- Different styles coexist at the same time
- Style is more synchronic than diachronic, consisting of acts of undergoing change

Styles in the view of Kubler are historical configurations that are neither perpetual nor in random change. Style is only identifiable among time-bound elements. However, because the components are always in change the relation among them is a changing one. Although all human action has its styles, their configurations are more instantaneous and synchronic, than extended in duration. For that reason, it is best adapted to static situations in cross-cut or synchronous sections. It is unsuited to duration, because of the changing nature of every class in duration. (Kubler 1967: 855). We do not know whether Priscilla Colt was satisfied with Kubler's elaborations of the relations between style and formal sequences in historical time. It seems that Kubler, although he nuanced the tone of his formulations somehow, just tried to bring in additional arguments in particular from the natural sciences to support his case. Priscilla Colt's (1963, 79) main reservation with Kubler's theory was that it was mainly concerned with the problems of describing change rather than with explaining it. Moreover, she deplored that Kubler did not alter the methods at hand. Kubler indeed in the preamble of his *Shape of Time* immediately had set aside studies that focused on symbolic expressions and the meaning of art instead of formal relations. However, also our ontological model of storylines of historical evidence is in the first place descriptive instead of explanatory. It supports in the first place the semantic web and multi-agent technologies to link and to query data of the distributed collections of the infrastructure that allows researchers of the creative Dutch Golden Age in Amsterdam to ask questions and to test hypotheses for further interpretations and explanations.

4.0 Modeling Rembrandt's *Night Watch* in Storylines

4.1 Rembrandt thinking and painting: The *Night Watch* as a prime object

While the Golden Agents project tries to break with the canon of art history by analyzing the consumption of cultural goods in all layers of society instead of in elite culture, for the modeling the most famous painter of the Dutch Golden Age, Rembrandt, and his most famous painting *The Night Watch* of 1642 were chosen. We opted for a painter with many

pupils and copyists, for a work of art with multiple archival sources of commissions and provenance (Dudok van Heel 1987 and 2006; Remdoc), with a rich material history of production, re-use and restoration and with contemporary copies and later derivatives in other formats to make a rich model that includes as many past and present stories and perspectives as possible. Rembrandt and his *Night Watch* meet those requirements for an inclusive model.

The Rembrandt Research Project that run from 1968 until 2014 under the guidance of the expert Ernst van de Wetering and resulted in *A Corpus of Rembrandt Paintings* (Bruyn et al. 2015) in six volumes in which attributions to master and pupils changed continuously made gradually clear that connoisseurship based on stylistic criteria did not suffice and that additional material research based on methods of the natural sciences was needed to establish the corpus of 340 paintings by Rembrandt. However, additional publications by Van de Wetering, *Rembrandt the Painter at Work* (2009) and *Rembrandt The Painter Thinking* (2016) confirm the view of Kubler (things are ideas and objects) that thinking about and the practices of making paintings cannot be separated from the materiality of the painted objects. Using contemporary sources about painting materials, methods and art theory, Van de Wetering reconstructs and contextualises Rembrandt's working practices and exploration of the foundations of the art of painting in his time and explains that changes in his way of working cannot simply be attributed to stylistic evolution in his work.

Without doubt the *Night Watch* is Rembrandt's most famous and replicated work. In the traditional historiography this masterpiece might be called, in Kubler's definition, a primal object that denotes a principal invention. Several authors, referring to the comments of contemporary and later critics underlined Rembrandt's break with tradition in the composition of group portraits that focused on the faces of the individual people as recognisable entities. For instance, Rembrandt's pupil Samuel van Hoogstraten, in his *Inleyding tot de hooge schoole der schilderkonst: anders de zichtbare werelt* of 1678, praised the overall composition in which figures on the foreground were more roughly painted while those in the back more neatly draw the attention of the viewer to the whole instead of to individual parts (Wetering 2009, 181-5). However, recently Middelkoop nuanced this view (2019, 190) and stated that other lesser known painters, such as Ketel, Badens and De Keyser already used aspects of Rembrandt's composition techniques. *The Night Watch* stands in a long tradition of the so-called corporate group portraits that were produced in Amsterdam between ca. 1525 and 1850. Apparently, it was a very popular genre in the 17th century. Between 1617 and 1650, 80% of the 600 regents, guilds or arquebusiers active in Amsterdam were portrayed in such portraits (Middelkoop 2019, 717). Kubler's observation that in the wake of prime objects a whole system floats of replica's, reproductions, copies, reductions, etc., that are so important to understand the original better because they provide more evidence, seems also to be the case when we unpack the history of the *Night Watch* in multiple storylines.

4.2 *The Night Watch* in Storylines

4.2.1 Stories of *The Night Watch*: The original object

The Night Watch is not only a grand work; it is a big object which measures of 379.5 cm x 453.5 cm (149.4 x 178.5 inches), and it used to be even bigger. When *The Night Watch* changed ownership from the militia of Frans Banning Cocq who had commissioned the

work to the city of Amsterdam it was cut in 1715 to move it from its original location from the Kloveniersdoelen to the Townhall of Amsterdam.

We do not know exactly its original measurements but the system of derivatives, in Kubler's words, allows us to infer this information. A drawing in the family album of Frans Banning Cocq, a painting of 1647 attributed to the contemporary copyist Gerrit Lundens, in the Rijksmuseum on loan from the National Gallery in London, and an etching after the original of Lambertus Antonius Claessens of 1797 (see Figure 2) provides crucial contextual information to understand the original depicted scene and *The Night Watch* as an object. The copy of *The Night Watch* attributed to Lundens was painted on panel instead of canvas and was smaller in size, but it shows which parts of the scene were cut, which figures were added later and what the dimensions of the original must have been. Moreover, the smaller copy attributed to Lundens was used to make a virtual reconstruction of *The Night Watch*.

Figure 2. *Night Watch* and Derivatives: a) *Night Watch*; b) Etching Claessens 1797 after original; c) Tattoo of *Night Watch* on back Marko Bak during visit to the Rijksmuseum on 18th of May, 2019; and, d) storytelling about the composition of *The Night Watch* by the Rijksmuseum).

The research photographer Rene Gerritsen on commision of Ernst van de Wetering combined x-ray images made by Guido van der Voorde in the 1970s with digital photographs of Lundens' copy to reconstruct *The Night Watch* in its original dimensions and with a representation of the figures that Rembrandt had included in his work (Gerritsen n.d.; Middelkoop 2019) The digital *Night Watch* in its original dimensions was one of the 340

reproduced works, including those damaged and stolen included at the virtual exhibition "Discover Rembrandt: His Life and all his Paintings" (https://www.discoverrembrandt.com/en/) that opened in the RAI Amsterdam Convention Centre on the 5th of July 2019.

This attention to the original dimensions of *The Night Watch* might seem farfetched, but for the making of group portraits as Middelkoop has demonstrated, the architectural setting, or more specifically the availability of space on the wall, often determined the commissions. In the case of *The Night Watch* its original size makes part of a larger debate between art historians whether Rembrandt could have painted this big object on location in the Kloveniersdoelen or in the house at Jodenbreestraat (now the Rembrandt House Museum) that he bought shortly before the commission, or in a gallery built as an extension to this house in its courtyard. It is the beginning of a long storyline that traces the long material history of *The Night Watch* that since it was cut in 1715, was overpainted, attacked by a knife in 1911 and 1975, sprayed with a chemical in 1990 and restored several times. As we write this story, *The Night Watch* is since July 2019 once again in restoration which can be viewed live by visitors to the Rijksmuseum or by followers on line of "Operation *Night Watch*."

4.2.2 Stories of *The Night Watch* in Derivatives

Apart from this material history of the painting, the story of *The Night Watch* lived on in many other media. It inspired, for instance, Peter Greenaway to make a film, Mikhail Dronov and Alexander Taratynov to cast the arquebusiers in freestanding bronze statues and finally a theater company to bring The Shooting Company of Frans Banning Cocq to life amidst the shopping public in Amsterdam as a part of a commercial for a Dutch bank. Endearing is the story documented on the 18th of May 2019 on YouTube (https://www.youtube.com/watch?v=WJAKFjn0ODk) of the 51 year-old trucker Marko Bak, who in the making of a tattoo of *The Night Watch* on his back together with his tattooist Richard van Meerkerk, visited the Rijksmuseum to compare it with the original. Although at that time still two or three tattoo-sessions of seven hours were needed to complete the copy, the tattoo already differed considerably since Bak had asked to change some of the faces of the figures on the painting to those of his own family members and friends. Marko's mother who up to now always lamented her son's tattoos was finally proud of this one because her portrait would be included.

The sources of evidence of the very rich story of the production, re-use and restoration of *The Night Watc*h with its many copies and derivatives in other media is just one of the many stories of the history of this painting that allows us to storify data in related, partially overlapping timelines as input for modeling historical processes in knowledge graphs.

An example of how these stories of *The Night Watch* in copies and adaptations in terms of production and consumption relate to each other is visualised in Figure 3. In this figure and similar figures following, the horizontal arrowed lines represent storylines for certain entities. The arrows represent continuity (for undetermined time) in one or the other direction. The curved symbol that may connect the lines represents events in which the covered entities participate, and which are described with balloons. For convenience, some entities may be omitted, such as who resized *The Night Watch* in 1715. Observe that the events concerning the copies and adaptations (in orange) of the Night Watch are preceded by consumption events (in dark blue).

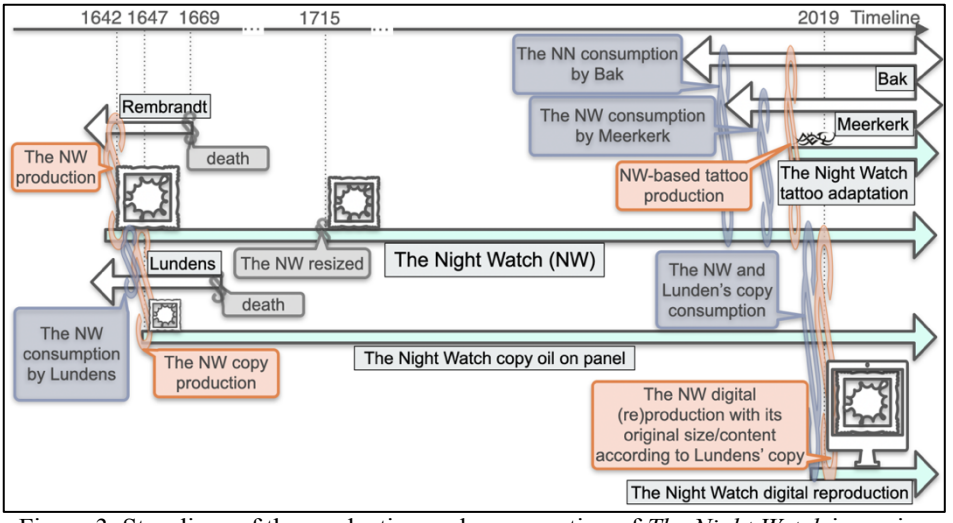


Figure 3. Storylines of the production and consumption of *The Night Watch* in copies, adaptations and digital reproductions hereof.

Additionally, we can zoom in or out on the longitudinal sections of storylines. As depicted in Figure 4, the zooming feature here proposed does not regard expanding or reducing the time frame under scrutiny, but rather allows the view more or fewer details for a particular entity, in this case, *The Night Watch*. On the left-hand side we zoom in into the details of the painting to observe the storylines of its material and immaterial parts. On the right-hand side, we zoom out to observe the Night Watch in the context of more or less contemporary paintings of Rembrandt.

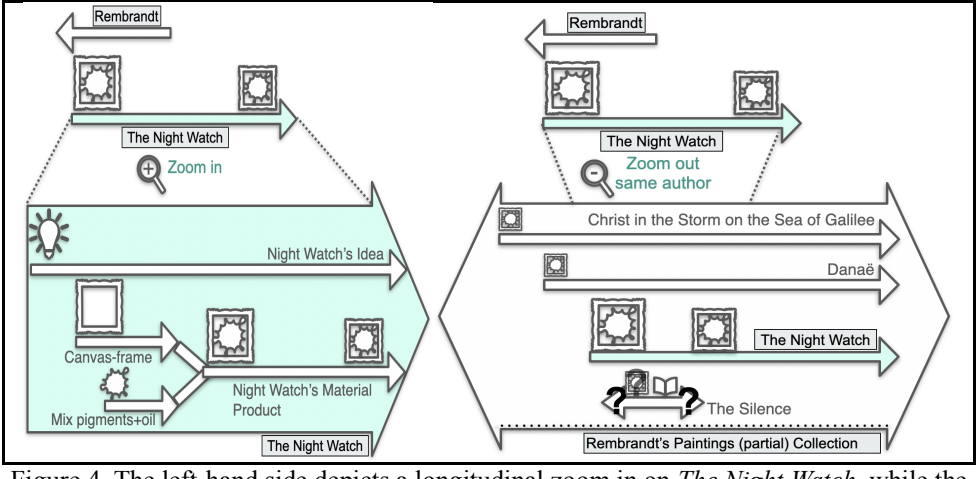


Figure 4. The left-hand side depicts a longitudinal zoom in on *The Night Watch*, while the right-hand side depicts a longitudinal zoom out showing *The Night Watch* among other paintings by Rembrandt.

We can then zoom in on the longitudinal sections of certain timelines of *The Night Watch* and its copies and adaptations, for instance, for visualizing in more detail production and consumption events regarding immaterial and material aspects of the Night Watch (Figure 5).

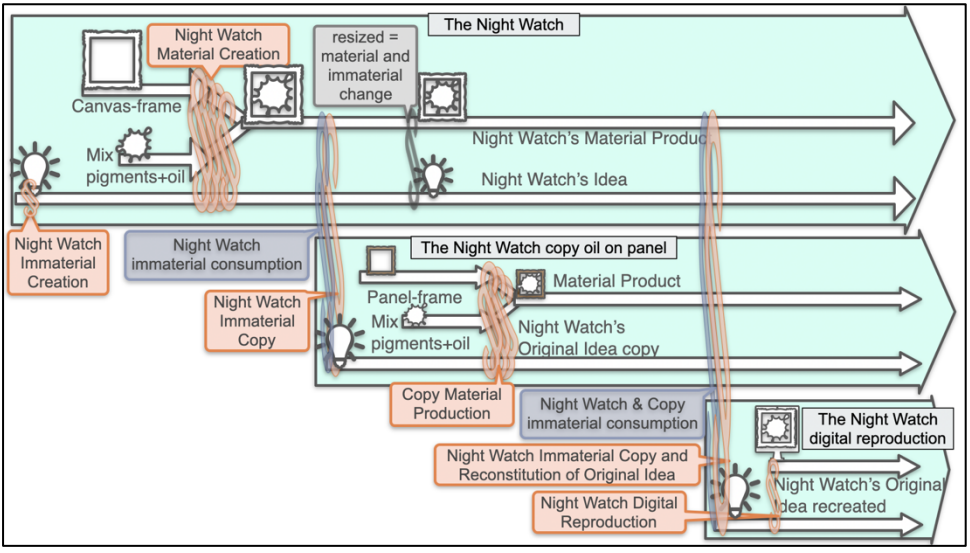


Figure 5. Zooming in on immaterial and material aspects of the production of *The Night Watch* and copies or adaptations thereof.

In cross section such longitudinal zoomings will also result in less or more detail depending on the question of whether we can see all the ends of these storylines at the same time (synchronous snapshot), or whether we get only a transversal view of some ends of these storylines, which can only be read in a meaningful way (as will be explained in more detail below) in combination with past and/or future events (asynchronous crossing). A snapshot of the unfinished tattoo of Bak on the 18th of May during his visit to Rijksmuseum can only be understood by the past and present of *The Night Watch* and by the future filling in of the blank faces for Van Meerkerk on request of Bak to make portraits of his family and friends.

4.3 Views of Rembrandt *Night Watch* and a kaleidoscope for Kubler

Earlier we noted that Kubler imagined his shape of time as a bundle of fibers instead of lenses as adherents of Strukturforschung and iconologists had done. Just now we described two moments relatively close to each other in the long history of Rembrandt's *Night Watch* in all of its contexts: the 18th of May 2019 when Marko Bak was filmed in the Rijksmuseum with the tattoo of *The Night Watch* on his back and the moment a month and half later, on the 5th of July, when the doors opened to the virtual exhibition "Discover Rembrandt: His life and all his Paintings" in the RAI, where for the first time since 1715 *The Night Watch* could be seen in its original dimensions. How would we be able to see these moments according to Kubler's *Shape of Time*? Kubler (1962, 28) describes a moment in his bundle of fibers of duration as follows:

> By this view the cross-section of the instant, taken across the full face of the moment in a given place, resembles a mosaic of pieces in different developmental states and of different ages, rather than a radial conferring in meaning upon all the pieces.

It is clear that Kubler tries to explain that if we make a slice in time we do not get a coherent picture of the whole, but rather an amalgamation of pieces that for the greater part differ in meaning because they are composed of the profiles of fibers (in our case storylines) in different stages of development. In that regard his mosaic metaphor is misleading. We can read representations of Greek gods or ferocious animals in figurative mosaics and will even be able to recognise regular patterns in non-figurative ones. The metaphor of the circular lens, varying in thickness according to the antiquity of the patterns that Kubler (122) dismissed, or the use of multiple lenses such as in a telescope, would at least allow for seeing more detail of the pattern in question. However, instead it would even be better to replace Kubler's mosaic metaphor by the one of the kaleidoscope, to explain the potential of his *Shape of Time* for the representation of the aforementioned moments in the storylines of Rembrandt's *Night Watch*.

The advantage of the kaleidoscope metaphor is that it gives depth (an extra dimension) to the view of the desired pattern. In a kaleidoscope light rays that enter from the back of the tube are reflected on mirrors that are tilted to each other in such a way that when one or more (parts of) objects are moved by rotating parts of the tube until they are aligned on one end of these mirrors these can be seen as a regular pattern.

When we return to Rembrandt we can explain and visualise Kubler's cross-section and our interpretation of his longitudinal bundles of fibers of duration as a kaleidoscope using the history of all his paintings as an example. For our visualisation in Figure 6, we include of course *The Night Watc*h and his *Danea* that stand for all his paintings that are in public or private collections in the world. However, for this historical overview it is important to realise that not all original works of Rembrandt survived. For instance, there are archival sources that point to his work that we have never seen, such as a painting with the title *de Stilte*" (*The Silence*) mentioned in a notary deed in the City Archives of Amsterdam (Dudok van Heel 1982). And there are his paintings of which we have images, but of which we do not know whether they still exist. A famous example is Rembrandt's *Storm on the Sea of Galilee* that was stolen in 1990 from the Isabella Steward Gardner collection in Boston.

Now observing the storylines (Kubler's bundle of fibers) for Rembrandt's collection transversally rather than longitudinally, we use views that could be synchronous (Kubler's cross-section) or asynchronous (kaleidoscope). Figure 7 illustrates, on the left-hand side, two ways for traversally visualising the storylines presented in Figure 6: a synchronous view as a straight line cutting the storylines in 2019, and an asynchronous view as a combination of cuttings in the storylines at the moment of their creation. The resulting views are presented on the right-hand side. The synchronous view or snapshot depicted on the top right side, only provides information on the present state of *The Night Watch* and *Danae*, meaning that *The Christ in the Storm* and *The Silence* are not accessible. In other words, it is equivalent to being able to have access to the existing paintings (in a physical sense) of Rembrandt at a chosen moment, in all public and private collections in the world.

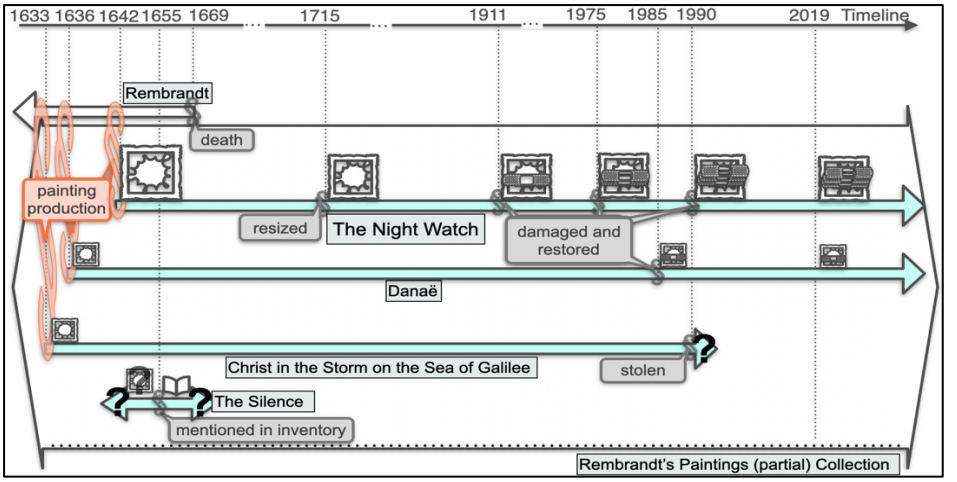


Figure 6. Storylines of Rembrandt's paintings based on information available in 2019.

Conversely, the asynchronous or kaleidoscope view of Rembrandt's painting collection as depicted on the bottom right hand side of figure 7 provides information on the state of the paintings at chosen moments in the past, which implies that *The Christ in the Storm* can be represented, as well as referred to previous paintings we only have documentary evidence of such as *The Silence*. It is equivalent to being able to have access to all paintings of Rembrandt, as close to their original version as the available information/knowledge allows for, regardless of their current condition. Hence, in this kaleidoscope view one can access all four selected paintings, including *Christ in the Storm* and *The Silence* (clearly not in the physical sense). However, using the latest virtual reproduction techniques, the exhibition "Discover Rembrandt" allowed us to virtually see the paintings resulting from a kaleidoscope view, since the paintings by Rembrandt were digitally represented and sometimes reconstructed in their original dimensions, such as *The Night Watch*. *The Silence* could not be digitally reproduced because there is no record of its appearance.

One could also consider the virtual exhibition to be a cross-section (synchronous view) of the digital reconstructions, that is historically founded in a kaleidoscope view (asynchronous view) of Rembrandt's originals. This, for the reason that the virtual reconstruction of *The Night Watch* in its original dimensions that was projected on the wall can only be understood by the historical evidence that the work was cut in 1715 and was reconstructed digitally with information about the lost part of the painting derived from the copy of Lundens. However, the pixels with which this image is built up is just an approximation of the materiality of *The Night Watch*. To get a better understanding of the materials Rembrandt used we have to manipulate the kaleidoscope—make a new alignment—in such a way that we for instance can see the pigments in the lab of the Rijksmuseum that provide evidence of other material aspects of *The Night Watch*.

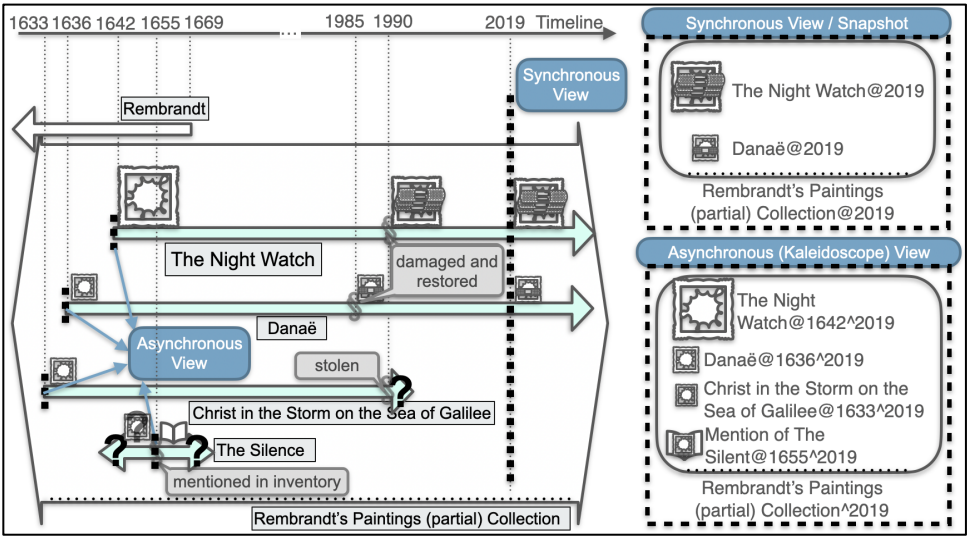


Figure 7. Storylines (longitudinal) on the left-hand side and cross-sections on the right-hand side. The one on top is a snapshot (synchronous cross-section) of Rembrandt's existing paintings in 2019 whilst the one at the bottom is a kaleidoscope view (asynchronous cross-section) of Rembrandt's paintings according to information available in 2019, similar to the digital reconstruction hereof for the Virtual exhibition "Discover Rembrandt: His Life and all his Paintings."

Similarly, the composition of the Night Watch can immediately be recognised in the tattoo on Bak's back. However, when we have a closer look at the faces of this group portrait, the photo-album of his family and friends probably provides far better contextual information to understand this dissimilarity of the tattoo with the painting. This phenomenon, that two meaningful patterns can be recognised simultaneously when aligned with multiple perspectives, is probably what Kubler tried to capture with the term "the plural present" and brings him to the conclusion that the principal object of the art historians is "to suggest other ways of aligning the main events" than style (Kubler, 1962, 129-30).

The limitations of aligning periods and events according to style and the advantages of using the kaleidoscope view of alignments of what Kubler (39) had called sequences or "chains of solutions" become evident when analysing and visualizing the term "chiaroscuro" that is often used to describe a main characteristic of several of Rembrandt's works. The term, that literally means light-dark, is comprehensive and complex. It has been used in the context of style, such as caravaggism after the Italian painter Caravaggio. This for instance to define "i caravaggisti" in Italy such as Giovanni Baglioni (accused for plagiarism by Caravaggio) or the female painter Artemesia Gentileschi but also to describe common characteristics of the Utrecht school of caravaggists with painters such as Hendrik ter Brugghen and Jan van Bijlert (*The Concert* 1635-40) or followers in France such Simon Vouet (*Fortune Teller* ca. 1620) and Georges de la Tour. It has been associated with the sub-genres of portraits and still-lifes in which faces and objects often in nocturnal scenes

are lit up against dark backgrounds by candlelight. Dirck van Baburen and Gerrit van Honthorst (*The Matchmaker* 1625) as members of the Utrecht school made small group portraits in that genre or Georges de la Tour in France who made a whole series of candle-lit portraits such as *Magdalena* with the smoking flame (c 1640). However also Rembrandt lit up portraits of himself or others—often in the act of reading or writing—by candlelight. Finally, the term chiaroscuro has been described as a technique to enhance the dramatic effects in storytelling such as in the *Crucifixion of Saint Peter* by Caravaggio (1601) or in the depiction of the same saint in prison by Rembrandt (1632), but also in far less dramatic ways such as in the composition on his *Night Watch*. The latter is important because it demonstrates that a certain technique can be applied in other styles or genres. It is generally accepted that Rembrandt who never was in Italy was indirectly influenced by Caravaggio via his teacher Pieter Lastman who visited the Mediterranean country approximately between 1604 and 1607. Nevertheless, if we compare Rembrandt's earlier work in chiaroscuro, such as *Three Singers* (1624) it differs far more in style from Caravaggio than the depiction of the musicians by Van Bijlert thirty years later in his *The Concert* produced between 1635 and 1640. Chiaroscuro is far more prominent and persistent in the sub-genres of individual or small group portraits than in the large, corporate group portraits. *The Night Watch* is one of the few exceptions in these long series of militia group portraits. Nevertheless, the contrasts between light and dark are used compared to the caravaggisti in a far subtler way (Figure 8).

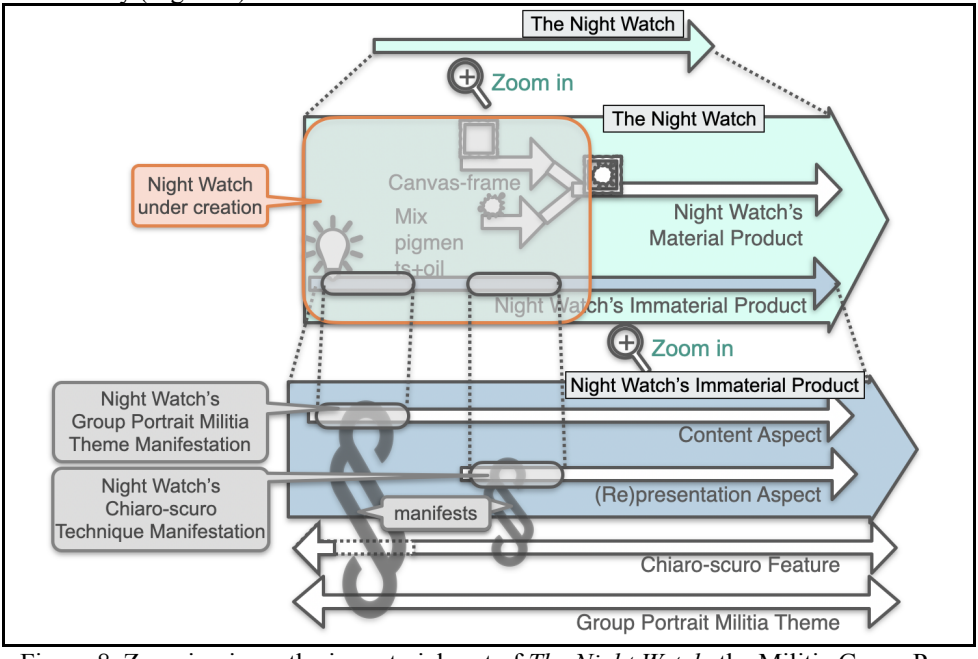


Figure 8. Zooming in on the immaterial part of *The Night Watch*, the Militia Group Portrait theme manifests as its content aspect, while the chiaroscuro Feature manifests as its (re)presentation aspect.

In short, there are overlaps between style and genre in the application of chiaroscuro, but their inconsistent sequences in time and place, as we have seen, demonstrate that they both have limitations for periodisation in the arts. Kubler is correct when he states that rather than using periods of styles (he does not discuss artistic genres in his *The Shape of Time*), it would be better to speak of chains of solutions. While only few of Rembrandt's works in which he applied chiaroscuro have some overlap with the caravagist style or the candle lit (sub-)genre, all works of Rembrandt in which he used the technique of chiaroscuro can be linked to a long chain of solutions in the use of light-dark contrasts that runs from Leonardo's *Virgin of the Rocks* (1483-86) to Stanley Kubrick's use of candle lights in the film *Barry Lyndon* (1975), to the chiaroscuro in the photographs of Christy Lee Rogers such as *Rapture* (2011). Common manifestations in genre, style, and technical solutions can be aligned (Figure 9).

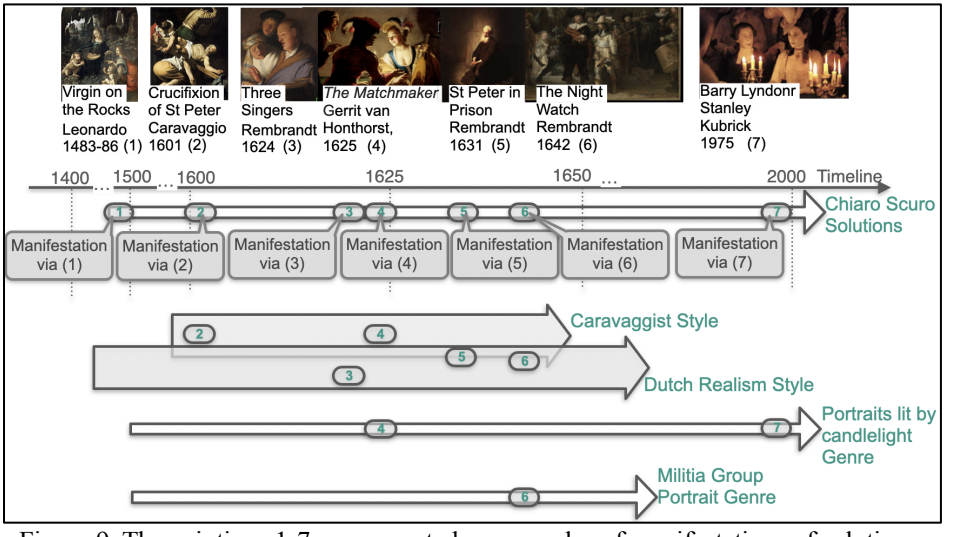


Figure 9. The paintings 1-7 are presented as examples of manifestations of solutions, styles and genres.

Some historians argue that such alignments in the kaleidoscope of history are arbitrary. For instance, Paul Veyne (1979; compare Miller ([1993] 2000, 152 and note 107) when describing Michel Foucault's approach of the past as a kaleidoscope states that the last pattern is "neither more true nor more false than those that preceded it." Indeed, with every turn of the tube a new pattern will occur. Some fragments that we observe might seem to be less relevant than others. However, similar to the idea that most people like the symmetrical patterns of the kaleidoscope for esthetic reasons, the historian in this metaphor might also be more content with one pattern over another.

In our example of chiaroscuro, the caravagist style, candlelight genre and the use of strong dark-light contrasts as a technique or "solution" can all three be aligned to explain the main characteristics of *The Matchmaker* of Gerrit Honthorst (1625). Rembrandt's *Night Watch* could only partially be aligned with the style of the "caravaggisti" given the strong

overlap with the Dutch realistic style (and be recognisable of course in the so-called Rembrandt style of followers, in the same way as Caravaggio directly corresponds with the style of the "caravaggisti"). It would fit a completely different genre, that of the militia groups running according to Middelkoop approximately between 1525 to 1800, but would fit in with all his other works in which he used light-dark contrasts in the long series of "chiaroscuro solutions" from the end of the 15th century to the present.

The use of the kaleidoscope view is not necessarily limited to visual analysis. The historian might look for fragments that fall in place when they connect to past historical evidence. Such as we have seen in our example of Rembrandt's work *The Silence,* of which we probably will never know how it looked, but which original existence still can directly be traced back to archival documents. The use of the kaleidoscope just implies dealing with less or more uncertainty in the meaning of visual patterns or in historical evidence in the interaction with these various fragments when making alignments until the moment that we recognise patterns that are deemed to be meaningful.

5.0 Toward a knowledge interaction model of historical interactions

5.1 Framework: Knowledge interaction versus KO

The Golden Agents project develops an infrastructure to analyse interactions between the production and consumption and among the various branches of the creative industries of the Dutch Golden Age. In short it should support the study of interactions. However, interactions are not only the object of study. If we follow Shaw's statements that KO is not applied to history, but that history is a form of KO and that the emphasis should not be on a (organised) portrayal of history but on supporting historians in doing history, we can argue that interactions also have methodological implications. We need a model that supports the analysis of historical knowledge interactions and interactions with historical knowledge. In earlier studies attempts have been made to formulate a theoretical framework for the analysis and visualisation of knowledge interaction between concepts in general (van den Heuvel and Smiraglia 2013; Smiraglia and van den Heuvel 2013 and 2011; Smiraglia, van den Heuvel and Dousa 2011). Similar to the way that Shaw described the requirements of a semantic tool that supports historians in the process of conceptualisation of historical discourse, we need a dynamic model to describe, analyse and visualise the interactions within the creative industries of the Dutch Golden Age. a model that we can use actively as an instrument to interact with interpretations of that past and with the documents that are used to portray historical events and to underpin those portrayals with historical evidence. The part of the ontological model that deals with historical evidence based on archival resources and expressions of uncertainties is still work in progress, but first results are and will be demonstrated (Idrissou et al. 2018 and 2019; Engelse and Wissen 2019; Zamborlini, Wissen and van den Heuvel 2020; Wissen et al. 2020; Wissen and Zamborlini 2020; Zamborlini and Wissen 2020).[2] In this chapter we focus on parts of the model that allow for describing and interacting with historical processes and discourses with the emphasis on conceptualisations of temporality and periodisation. This model needs to meet the following requirements:

Requirement 1) The model provides a framework for interactions of historical knowledge as an object of study and as a methodological instrument to interact with historical knowledge.

Requirement 2) The model supports the study of interactions between production, consumption and branches of the creative industries.

Requirement 3) The model supports conceptualisations of historical interactions with temporality and periodization.

5.2 Storifying data: Modeling historical narratives and conceptualisations of things in space/time

In a model that supports conceptualisations of interactions with historical knowledge, in our case of the creative industries of the Dutch Golden Age, things (ideas and objects) need to be linked to time, place and related concepts, as well as put in the context of narratives. Modeling things in space and time (space/time) has a long history that goes back to antiquity (Bliss 1929). In the early history of library and information science Ernest C. Richardson (1935) used the universe of knowledge metaphor to class things (which could be both ideas and physical objects) in space and time. This metaphor was followed by the universe of concepts (Ranganathan 1957; Miksa 1992; Beghtol 2008) and concepts in spacetime in the multiverse of knowledge (van den Heuvel and Smiraglia 2010; Smiraglia, van den Heuvel and Dousa 2011). van den Heuvel and Smiraglia (2010) extended the metaphor of multiverse knowledge to the laws of physics in those spaces. The "gravitational forces" in these knowledge universes were used metaphorically to explain two important concepts in the theory of classification: "likeness" and "likeliness" (Hjørland 2003; van den Heuvel and Smiraglia 2013). The latter concepts might be of interest for the understanding of the stories we tell about the stories we tell about history. The tattooist of the Night Watch was drawn between the "likeness" with the composition and colours of the painting and with the portraits of Bak's family. The "likeliness" of a meaningful pattern in the narratives depends on the weight we address to the various pieces of evidence of the relations between, in Kubler's terms, primal objects and the many different sorts of replications. The Bak's back tattoo tells multiple stories simultaneously, some finished a long time ago, others like the making of the portrait of his mother that still was a future idea for the tattoo in May 2019. This example demonstrates that the model needs to be able to handle narratives of relationships between things both in real and in fictional time in a multidimensional space for which we introduced the kaleidoscope metaphor. As Shaw states, several historians treat events as phenomena, as actual things that existed in the past. From that perspective one sees the history of the past as a kind of fabric woven of these events, and history-as-practice as the study of that fabric. According to this "unreflective view of events" historians simply describe events as a historical portrait by comparing them to an independent standard "what really happened." However, the past does not exist anymore and for that reason the best historians can do is to compare various portraits of narrations of the past. In doing so they accept or reject new and old ideas that are shaped by newly discovered documentation and that are changed by cultural changes (Shaw 2010, 45-46). This is congruent with Kubler's *Shape of Time* in which processes of change and in our attitudes towards them, shape the occurrence of things in often imprecise time intervals. It implies the remodeling of events as part of a dynamic system with sequences in different rhythms of duration instead of in calendar time (which as Kubler stated indicates nothing about the changing pace of events) and concordances hereof such as in PeriodO. However, to allow for interoperability of LD within the semantic web paradigm the remodeling of temporality of historical events must allow for mappings to other standards such as CIDOC-CRM, OWL-Time and PeriodO:

Requirement 4) In the model that supports interaction with historical knowledge, things (i.e. ideas and objects) need to be linked to time, place and related concepts, as well as put in the context of narratives.

Requirement 5) In the model that supports interaction with historical knowledge, multiple narratives of developments of ideas and objects must be represented simultaneously in a multi-dimensional way.

Requirement 6) In the model that supports interaction with historical knowledge, ontologies of events purely based on calendar time and concordances hereof need to be remodeled to describe events as part of a more empirical system based on practices of historical research. The model is calendar-agnostic.

Requirement 7) The model needs to be interoperable with ontologies/SKOS of time that are used as standards in cultural heritage.

5.3 Visualizing storylines of historical interactions

Kubler, possibly inspired by Ore as we noted, did see historical patterns as networks. Moreover, we claimed that Kubler's reading of a time instant in the fibers of duration as a mosaic perhaps better could be imagined as a kaleidoscope. In the context of this latter observation it is interesting to note that the kaleidoscope is already used as a metaphor to explore the semantic web and knowledge graphs (Haase 2019). Mackeprang et al. (2018) developed a prototype of an RDF-based data analysis tool using semantic web technologies to explore and annotate upcoming associations and ideas interactively and to link them to concepts from external knowledge graphs such as Wikidata. It is a user interface in which customizable colored dots, that function as markers of ideas generated by each SPARQL query, are distributed over a grid-pane. Unfortunately, it is therefore a two-dimensional user-interface that does not do full justice to its name, because the kaleidoscope metaphor that we envision to model and visualise our concept of storifying data inspired by Kubler's *Shape of Time* entails interactions with data in a multidimensional spacetime model. A fake news post in the satirical journal *Onion* on the 16th of July 2018 described and illustrated a $200 billion Hubble Space Kaleidoscope with brilliantly colored interlocking and rotating diamond things that captured the first images of a nebula. Such a kaleidoscope that can be used to explore the pattern of the universe does not exist. However, a combination of telescopes including the NASA/ESA Hubble Space Telescope[3] was able to produce kaleidoscope images of a galaxy cluster that reveals the effects of a phenomenon that is known as gravitational lensing. The dark matter of this observed cluster bends the light of background objects in such a way that it acts as a magnifying glass and enables astronomers to find galaxies that existed relatively shortly after the big bang. These observations make part of the Hubble Frontiers Fields program[4] that started in October 2013 when for the first time the gravitationally lensed image of a supernova was arranged four times after the alignment with a galaxy in the cluster to which it belonged. This phenomenon of gravitational lensing is of interest in the context of the aforementioned metaphor of gravitational forces in knowledge interaction based on "likeness" and "likeliness" in which alignments from multiple perspectives with "things" that are alike, increases the likelihood that patterns will be recognised that we deem to be of interest. If we replace the entering light rays of the origins of the universe that are distorted by forces but are aligned with the astronomer's recognised patterns by Kubler's fibers of historical duration we get a similar effect. By interacting through alignments with parts of history that are reflected to us we can create a pattern of the past that in a certain moment of time has a meaning that is coloured by our interactions with parts of that past. It is important to realise that we see a pattern, and not an image as in the mosaic metaphor. It is not its context in the same dimension, but the

multidimensional spacetime of history that provides the contextual information to understand this pattern. Similar to the huge task that the Time Machine project set for itself, the development of an interactive kaleidoscope to explore the multidimensional spacetime of history is still a future dream. However, there are already more concrete explorations of user interfaces that would allow us to visualise and to interact with historical storylines that actually reflect Kubler's ideas quite well. We already observed that Kubler's *Shape of Time*, consisting of a longitudinal bundle of happenings of shorter and longer duration and a transversal view of a network, could be visualised by three-dimensional timeline tools, such as Jensen's TimeVis (compare Figure 1).

Other relevant examples of multidimensional semantic timelines combined with graph visualisations are the visualisations of time in "Time-Shadows" and "Time Beads" (Morawa et al. 2014). They are of interest because these shadows and beads respectively combine interactions in zoom based on overviews with various time shapes to visualise the display of qualitative and quantitative data in different classes of durations. Similarly, the user interface to interact with time in LD as part of the EU project Smart Museum (Kauppinen et al. 2010, Figure 5) is of interest. It deals with fuzziness and uncertainty in time intervals and allows for annotations of the relevance of time periods in relation to their queries.

Requirement 8) The model allows for the visualisation of synchronous and asynchronous multiple things (ideas and objects) over time and the relations between them can be expressed in networks.

Requirement 9) The model allows for the visualisation of the multidimensionality and dynamics of these networks of things.

Requirement 10) The model allows for the visualisation of events in precise and imprecise time intervals. The GUI allows users to interact with the settings and to annotate the preciseness of the boundaries of the time intervals and to assign the relevance of time periods in relation to their queries.

6.0 A model for time in storylines of historical interactions

This section presents a conceptual model aimed at addressing most of the aforementioned requirements while leaving place for others in future work. In particular, the proposed model is meant to be calendar-agnostic but also "truth-agnostic," in the sense that it enables events to be expressed in any existing calendar regardless of its veracity, as well as in the "future" or in fictional "calendar-time," such as an Elvish Calendar. As long as one can provide a mapping from one calendar to another or create explicit formal relations among the events (such as before or during) then they can be related or compared. In future work we plan to address veracity by allowing for reported events to be provided with evidence, so that it can be believed to be true or false or even just likely, but also to address the representation of events as explicitly hypothetical or fictional.

The proposed model builds on top of a general-purpose ontology called Unified Foundational Ontology (UFO) (Guizzardi 2005; Guizzardi et al. 2013 and 2015) and its variation gUFO (Almeida et al. 2020), of which the ontological commitments are precise but also flexible enough to support our requirements. It incorporates developments from other foundational ontologies such as GFO and DOLCE in a coherent way. They are compatible with the conceptualism theory in which concepts and individuals are described according to perception. Naturally, other existing models such as CIDOC-CRM, Web Ontology Lan-

guage (OWL) and its extension for time OWL-Time, Simple Event Model (SEM) and PeriodO also partially address our requirements. The similarities and differences with respect to our proposal are discussed and reconciled when possible.

The model is presented here in several UML-like class diagrams, including some UFO concepts (in a dark-gray shade) plus newly proposed concepts (in a light-yellow shade). They also include colored references to similar concepts present in other models, which when preceded by an asterisk mean an approximation not an equivalence. Dotted lines indicate relations that are not explicitly defined in that particular diagram, but in others or in the text. Moreover, in the text the concepts will be referred to by using as prefix an acronym of the model to which it belongs (e.g. *prefix*:Concept). This is important to avoid their free interpretation as a commonsense word but also because sometimes the same term means different things in different models. For example, the reading of *UFO:*Objects should be such that, according to the UFO, a person is an object. In particular, we use the prefix *ga* (for golden agents) when describing the concepts of the model here proposed.

6.1 Perdurants and temporal extents are calendar-agnostic

Figure 10 presents some main concepts as follows: the concept *UFO*:Entity, aligned to *CIDOC:E1-CRM-Entity* and close to *owl:Thing* (which does not include literals). It comprises the universe of discourse (roughly, anything one may want to "talk about") and is divided into *UFO:Concrete* and *UFO:Abstract* entities, where the former are entities that can be "placed" in space and time directly or indirectly (e.g., a language can be situated in space and time through the people who speaks it), while the latter is not (e.g., a number).

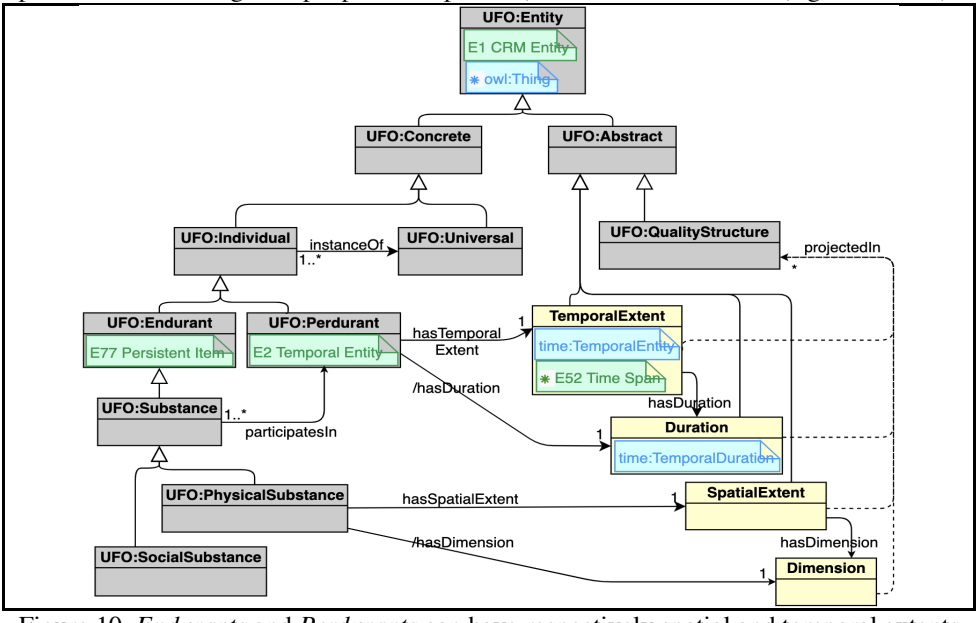


Figure 10. *Endurants* and *Perdurants* can have respectively spatial and temporal extents which are independent of a specific quality structure and can be projected in one or more of them, e.g., someone's birth date can be projected in both Gregorian and Chinese calendars.

UFO:Concrete entities are then split into *UFO*:Individual and *UFO*:Universal. The former are entities of interest (e.g., Rembrandt, *The Night Watch* or Rembrandt's role as a master instructing his pupils) while the latter, roughly, comprise ways of classifying and/or providing identity to the former (e.g., person or painting). *UFO:Individual* is split into *UFO:Endurant* and *UFO:Perdurant*. The former are entities whose essential parts are always present (e.g., a painting) while the latter's parts are not present altogether (e.g., the creation of a painting). These concepts align respectively as *CIDOC:E77-Persistent-Item* and *CIDOC:E2-Temporal-Entity*.

A particular type of *UFO:Endurant*, *UFO:Substances* are existentially independent entities said to participate in *UFO:Perdurant*. It can be split into *UFO:Physical-Substance* and *UFO:Social-Substance*. While the latter are *IMMATERIAL* entities (e.g., language), the former are *MATERIAL* entities that occupy a space, i.e. that have a *ga:Spatial-Extent* and also a *ga:Dimension*. Similarly, *UFO:Perdurant* entities have a *ga:Temporal-Extent* and also a *ga:*Duration, which is derived from the duration of its extent. Those concepts are *UFO:Abstract* entities that can be projected in a certain *UFO:Quality-Structure*, such as a calendar or a space coordinate system (to be discussed in the next subsection). Those entities are in principle independent of a quality structure, e.g., the temporal extent of a perdurant exists independently of a particular *CALENDAR SYSTEM*. Moreover, it exists regardless of our knowledge, i.e., the fact that we cannot precisely determine when an event happened does not make its temporal extent imprecise. On the other hand, some would argue that some entities' boundaries are essentially vague, such as those of a language or genre. Both cases require means to account for *UNCERTAINTY,* such as to state that the temporal extent of a language includes a smaller-precise one and is included by a bigger-precise one, thus expressing its "imprecise boundaries." Finally, observe that a perdurant is not the same as its temporal extent, since several perdurants can have exactly the same temporal extent, which is an abstract entity, meaning they happen at the same time, similarly to the manner in which several persons can have the same age or height.

Although only *UFO:Physical-Substances* and *UFO:Perdurants* are directly connected to respectively space and time, both can be indirectly connected to respectively time and space. *UFO*:*Substances* are indirectly situated in time through the perdurants in which they participate, while perdurants are indirectly placed in space through the *UFO*:*Substances* that participate in it. Naturally, in this paper we focus on perdurants and their ways of measurement.

The *OWL-Time* ontology actually concerns exactly the representation of *ga:Temporal-Extent*, where it is called *owl-time*:*Temporal-Entity*, while it does not concern perdurants or events *per se*. It does, however, consider that any entity (*owl*:*Thing*) can be attributed a temporal extent, which is not necessarily incompatible with our view if one considers that the endurants/substances can be indirectly placed in time. In turn, the "similar" concept *CIDOC:E2-Temporal-Entity* actually refers to a *UFO:Perdurant*, meaning that "temporal entity" does not mean the same in *OWL-Time* and *CIDOC*. Instead, the concept *CIDOC*:*E52-Time-Span* is close but not exactly the same as the *ga*:*Temporal-Extent* or *owl-time:Temporal-Entity*, since it does incorporate uncertainties.

6.2 Periods and durations in calendars

In Figure 11 the *UFO:Abstract* is more detailed to explain how the temporal extent and the duration are projected into a particular quality structure or, more specifically, a calendar, besides how to reconcile different interpretations of the concept period.

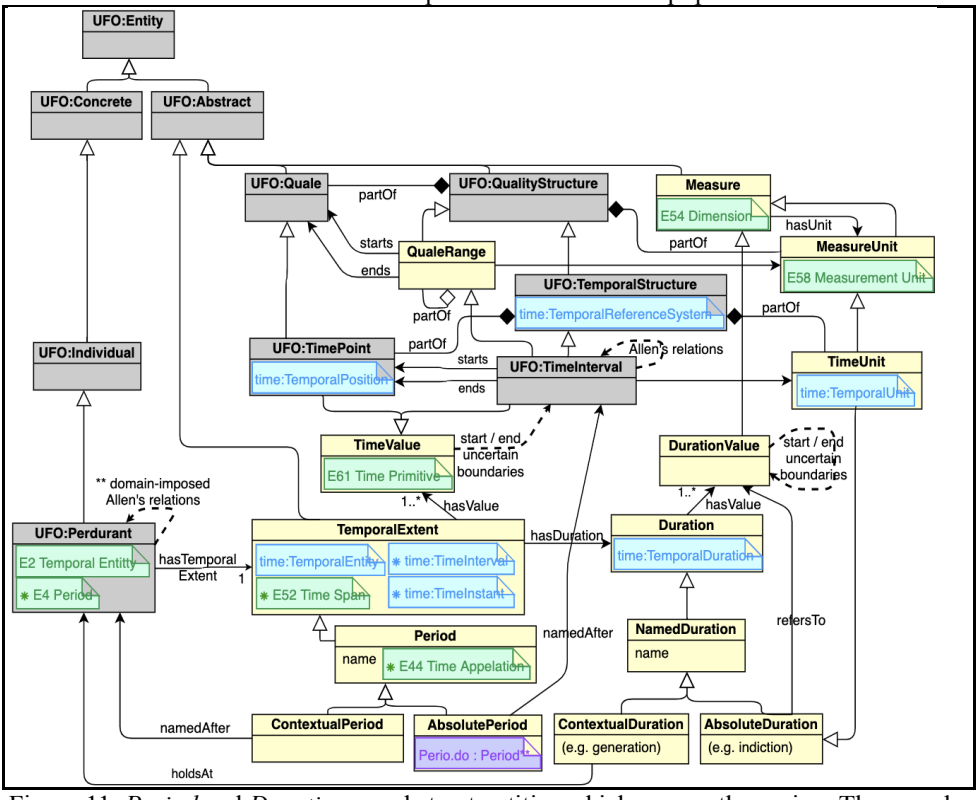


Figure 11. *Period* and *Duration* are abstract entities which are worth naming. They can be named after a specific event, e.g., the 2nd World War, or may refer to a particular time interval within a calendar, such as the 1960s or the year of the rooster.

First, a *UFO:Quality-Structure* is composed of *UFO:Quale* entities, which stands for each point in the quality structure. In a *UFO:Temporal-Structure*, which aligns with *owl-time:Temporal-Reference-System*, a quale is a *UFO:Time-Point*, which aligns with *owl-time:Temporal-Position*. In its turn, a *ga:Quale-Range* represents a subset of *UFO:Quales* and can be defined by a start- and an end-quale, e.g., a *UFO:Time-Interval* is a subset of time points. The union of time points and intervals in whatever calendar is called *ga:Time-Value*, which aligns to *CIDOC:E61-Time-Primitive*, and can be attributed to (calendar independent) *ga:Temporal-Extents*. When several values are attributed to an extent it means either projections of the extent in different calendars or a discontinuous extent. Finally, the concept *owl-time:TimeInterval* is a subset of *owl-time:Temporal-Entity* and therefore is equivalent to a subset of *ga:Temporal-Extent* whose values are *UFO:Time-Interval* in any calendar.

Special temporal algebra applies among *UFO:Time-Intervals*, also known as Allen's relations defined by Allen (1983), namely: during, starts, finishes, is equal to, overlaps, meets and takes place before. These relations can be derived between two intervals given their values. Naturally, the same relations apply to *UFO:Temporal-Extents*, although their calculation requires being able to project the extents into the same calendar system. Furthermore, equivalent relations can be inherited by perdurants/events. They can also be imposed by domain restrictions, such as a birth event must happen before the baptism. The domain restrictions allow us to state formal relations among events without knowing exactly when they have happened.

One way to allow for uncertainty is to attribute uncertain boundaries to the ga:Time-Value of a *ga:Temporal-Extent*. This allows one to express as much as is known about an event, such as the lastest start point. The *Simple Event Model* (*SEM*) provides such relations to hold between any *sem:Core* entity and a specific calendar value: *has-Earliest-Begin-Time-Stamp*, *has-Latest-Begin-Time-Stamp*, *has-Earliest-End-Time-Stamp*, h*as-Latest-End-Time-Stamp*. *CIDOC* provides a relation called *P82-at-some-time-within* describing the maximum period of time (*E61-Time-Primitive*) within which an *E52-Time-Span* falls.

A *ga:Period* is roughly a *ga:Temporal-Extent* worth naming. When the name is given after a relevant event, its temporal extent is called *ga:Contextual-Period*. Otherwise, when it is based on a time interval it is called *ga:Absolute-Period*. The latter is the case in the gazetteer *PeriodO*, where the concept period is a subset of *owl-time:Temporal-Entity*, hence a temporal extent, to which a name and other values are attributed, e.g., 1960 to 1969 is called the 1960s. However, the concept in *PeriodO* is not explicitly connected to any particular event, even if the period is called World War II. Conversely, *CIDOC:E4-Period* is a *CIDOC:E2-Temporal-Entity* which aligns with *UFO:Perdurant*. Therefore, *CIDOC:E4-Period* does not mean the same as *ga:Period*, but they are the *UFO:Perdurants* of which temporal extents are named *ga:Contextual-Periods*, such as in the previous example. Another concept called *CIDOC:E44-Time-Appelation* allows for using names to refer to a *CIDOC:E52-Time-Span*, although it is not itself a *CIDOC:E52-Time-Span* but an objectification of the naming. Finally, a *ga*:*Contextual-Period* can be associated to a place through the *UFO:Perdurant* after which it is named, while a *ga*:*Absolute-Period* has no clear connection to space.

A *ga:Measure* is an amount of *UFO:Quales* given in terms of *ga:Measure-Unit*s, which are names given to pre-defined amounts of *UFO:Quales*, e.g., *ga:Time-Units* like a second or a year. In particular, a *ga:Duration-Value* is a *ga*:*Measure* that values a *ga*:*Duration* that can represent the extension of *ga*:*Temporal-Extents*. In a similar fashion to *ga*:*Temporal-Extent*, as their *ga*:*Time-Value* can have uncertain boundaries, so can the *ga:Duration-Value* of a *ga*:*Duration*. Moreover, the *ga*:*Duration* can also be named either after a specific duration value, called *ga:Absolute-Duration* or yet after a certain duration that may change in time, called *ga:Contextual-Duration*. The former comprises all *ga:Time-Units* in any calendar such as a decade (10 years), or also Kubler's term indiction (duration of 15 years). The latter in turn comprises terms such as (human) generation, which is independent of a calendar and also may change in time, i.e., a generation 100 years ago might not correspond to the same amount of time as 100 years from now.

6.3 Modeling storylines: Fibers of duration and networks

Even though the *UFO* often uses *Perdurant* and *Event* interchangeably, we find it convenient to present them separately, since there are different correspondents in other models such as *CIDOC*. Figure 12 elaborates those concepts as well as storyline related concepts.

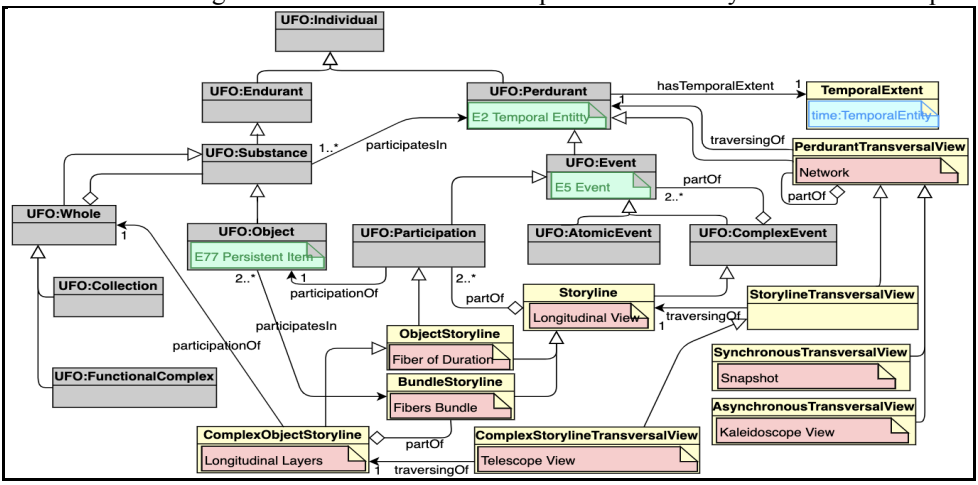


Figure 12. Storylines comprise the participations of an object/entity or of a bundle of them in events through time. A storyline transversal view is a static view or a network, which can be a synchronous view in time (e.g., Figures 2 and 7 top right) or it can be an asynchronous view (e.g., Figure 7 bottom right) as to connect objects that participate in related events at different points in time.

A *UFO:*Event, which aligns with *CIDOC:E5-Event*, is a perdurant that can be split into *UFO:ComplexEvent* and *UFO:AtomicEvent*, where the former is composed of two or more events and the latter is not. Moreover, a *UFO:Participation* is an event that conveys the participation of one single *UFO:Object*. The latter is a type of *UFO:Substance* that has a unity criterion, complementary with amount of matter, which will not be discussed in this paper.

A *ga:Storyline* is a *UFO:ComplexEvent* that can be split into (i) *ga:Object-Storyline*, which is composed of participations of a single *UFO:Object*, and (ii) *ga:Bundle-Storyline* (Kublers' "fibers of duration"), in which two or more *UFO:Objects* participate. Considering the storyline of *The Night Watch*, one can see it as a single *ga:Object-Storyline* describing only the events/participations concerning this painting (when it has been commissioned, produced, delivered, transferred, altered, etc.) or one can see a *ga:Bundle-Storyline* such as in Figure 3 connecting *The Night Watch*'s storyline to others such as Rembrandt's storylines as its creator, Lundens' copy in oil on panel, Bak's tattoo or its digital reproduction in 2019.

Furthermore, a *UFO:Object* can be a *UFO:Whole*, which means that it has two or more parts. While in a *UFO:FunctionalComplex* each part has a different "function," such as material and immaterial parts of a product, in *UFO:Collection* instead each part has the same "function," such as a collection of coins or a collection of paintings by Rembrandt. Hence, a *ga:Complex-Object-Storyline* is a *ga:Object-Storyline* as the participations of a

single *UFO:Whole*, while it is itself composed of *ga:Bundle-Storylines* in which the parts of the whole participate. This means that a particular storyline can provide a longitudinal zoom in and out from the whole to the parts and back. Since one whole-object can have parts that are themselves whole-objects several zoom levels can exist. In an example given in Figure 5 on the left-hand side, by zooming in on *The Night Watch* storyline one could see a more detailed bundle of storylines comprising both its immaterial part and the materials that were used, such as the preparation of the canvas or the pigments (more details about material versus immaterial in the next section). On the right-hand side, by zooming out from *The Night Watch* storyline, one can see the storylines of other paintings that are part of the same "whole-collection of Rembrandt's paintings."

Finally, another way to observe *UFO:Perdurants* is via a *ga:Perdurant-Transversal-View*, resulting in a "static" view of an event of interest that Kubler calls a network. It can be either (i) a *ga:Synchronous-View*, e.g., observing all the entities involved in an event at the same time like a snapshot; or (ii) a *ga:Asynchronous-View* that allows for "statically" observing a network of entities that participate in an event of interest but at different points in time, which we called a kaleidoscope-view since it allows motion back and forth through time independently for each storyline. Naturally, a *ga:Storyline-Transversal-View* is the crossing of a *ga:Storyline*. For example, Figure 7 depicts on the left-hand side the storylines of Rembrandt and some of its collection of paintings, which are crossed in two ways: (i) on the top right a snapshot of Rembrandt's painting collection in 2019, while on the bottom right a kaleidoscope view of his paintings at the time of their creation. In particular, the crossing of a *ga:Complex-Object-Storyline* allows one to zoom in and out on the parts of the whole-object but now in a transversal zoom instead of a longitudinal one, which we could call a telescope-view. For example, a transversal zoom in on the aforementioned kaleidoscope view could show the combination of the original materials used by Rembrandt in 1642 to create *The Night Watch*, while a zoom in on the snapshot of 2019 would show also the materials added due to restorations.

6.4 Modeling storylines of production and consumption

We already discussed and visualised (compare Figures 3 and 4) storylines of the production and consumption of *The Night Watch* itself and in copies and adaptations and zoomed in and out on immaterial and material aspects hereof in other paintings of Rembrandt. With these examples in mind we here model these production and consumption storylines and discuss them in relation to CIDOC CRM.

The *ga:Storyline* of a *ga:Product* is called a *ga:ProductStoryline*, which is composed of events like *ga:ProductUnderCreation* and *ga:ProductUnderConsumption* as the *UFO:Participations* of the *ga:Product* respectively in the processes of *ga:Production* and *ga:Consumption*, as depicted in Figure 13. A *UFO:Agent* is a type of *UFO:Object* with intentionality to perform actions, such as a *ga:Producer* and a *ga:Consumer*, which approximates to a *CIDOC:E39-Actor* representing (a group of) people to perform intentional actions.

For all the mappings made to CIDOC in this model, one important difference to bear in mind is that CIDOC is human centric, in the sense that all the actors are necessarily humans and the products human-made. This can be seen as a special case of our model which does not impose such restriction, so that it could cover for instance situations (real or fictional)

in which art could be created by an animal or by artificial intelligence. The CIDOC concepts are therefore subclasses of the concepts here proposed.

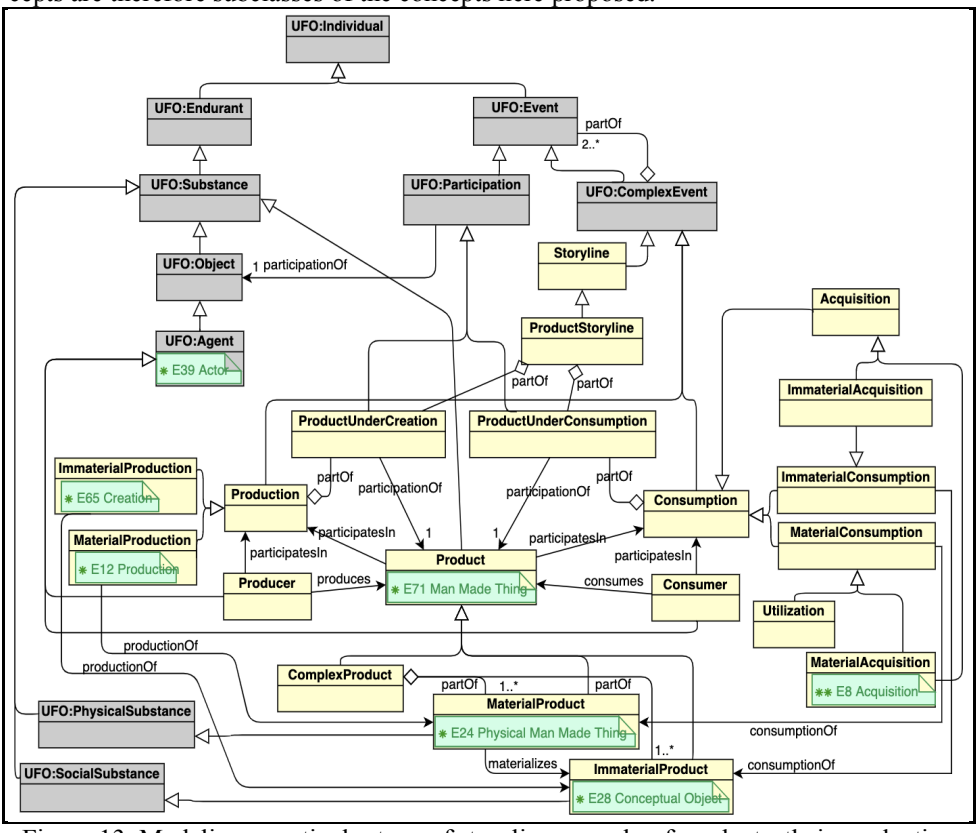


Figure 13. Modeling a particular type of storyline, namely of products, their production and consumption, material or immaterial.

A *ga:Product* can be either *ga:MaterialProduct*, *ga*:ImmaterialProduct or *ga:ComplexProduct*, where the latter has as parts entities of the former two types. Consequently, both *ga:Production* and *ga:Consumption* processed may regard some or all of those types of *ga:Product*. While *ga:Product* corresponds to *CIDOC:E71-Man-Made-Thing*, a *ga:MaterialProduct* corresponds to *CIDOC:E24-Physical-Man-Made-Thing* and a *ga*:ImmaterialProduct corresponds to a *CIDOC:E28-Conceptual-Object*. The *ga:MaterialProduction* is a *CIDOC:E12-Production* whilst the *ga:ImmaterialProduction* is a *CIDOC:E65-Creation*. Regarding *ga:Consumption*, the *ga:MaterialAcquisition* approximates to *CIDOC:E8-Acquisition*, except for the latter including loss of title due to destruction of the item.

With *The Night Watch* and its derivatives in mind the various production and consumption storylines both in an immaterial as in material sense can be modeled as follows:

(1) As a *ga:Material-Product*, the storyline starts with the materials used to create the painting, comprising the preparation of the canvas and the process of mixing the pigments and oil, the materialisation of the painting until the final touches, followed by the cuts made in order to make the painting fit into the city hall. The derivatives also have their parts as a *ga:Material-Product* which are the materialisation of their immaterial counterpart described next (see Figures 4 and 5).
(2) As a *ga:Immaterial-Product*, the storyline starts with the first conception of the idea for the painting by Rembrandt, probably after hearing the requirements set by the commissioners about its genre and who should be included in the painting, and includes the usage or adaptations of techniques such as how to mix the pigments to produce a certain effect. The immaterial part is consumed before it is copied or adapted, such as Lundens did for copying *The Night Watch*, expressing its content using different materials, or Bak's tattoo that partially preserved the content aspect, since he chose to include the faces of members of his family therefore telling a story other than that meant by Rembrandt. Finally, for the digital reproduction of the exhibition in 2019, it was necessary to includes the immaterial consumption of both the current version of *The Night Watch* and the copy by Lundens, so that the digital image could faithfully express the original *Night Watch* (see Figures 4 and 5).
(3) As a whole *ga:ComplexProduct* of which both (1) and (2) are part, as zooming out from the details in such a way that the *ga:Production* may encompass both *ga:MaterialProduction* and *ga:ImmaterialProduction*, and the *ga:Consumption* may encompass *ga:MaterialConsumption* and/or *ga:ImmaterialConsumption* (see Figure 3).

6.5 Modelling Kubler's views of periodisation via storylines of styles and solutions

According to Kubler, styles do not constitute themselves as periods as a style often cannot be represented as a single timeline (or fiber) but as multiple (parallel) ones of which the beginning and end may differ, for example, by taking the location into account for the time-period associated with the Renaissance style, which is different in Italy and in the Netherlands. In this section we discuss how this account can be addressed in our model alongside with Kubler's proposed alternative of representing chains of solutions, as depicted in Figure 14.

First of all, a *ga:Period* is the temporal extension of a *ga:Storyline* (or *UFO:Event*) that is worth naming, therefore to discuss periodisation we need to project styles and solutions into storylines. Second, since a *ga:Storyline* combines participations of entities in certain events, we need to decide which entities and which events form the storyline of styles or solutions. Clearly, it cannot simply be the collection of their exemplary products, as the storylines of the products extend throughout their existence, while the time-frame for styles and solutions is constrained around the time in which the products were created.

Regarding the nature of style and solution, when and why does something get to be called as such? Our hypothesis is that they are themselves *ga:Immaterial-Products* and *ga:Pattern-Features* that manifest by the creation of more than one *ga:Product*. This means no feature can be considered a style or solution if it is manifested only once. A (immaterial) *ga:Product* has inherent *ga:Content-Aspects* and *ga:Presentation-Aspects*, which are *UFO:Aspects* that inhere in *UFO:Objects*. While a *ga:Content-Aspect* manifests features such as a *ga:Story* or a *ga:Theme*, e.g., portrait lit by candlelight, a *ga:Presentation-Aspect* manifests features such as a *ga:Presentation-Technique*, e.g., chiaroscuro. If a technique is recurrently manifested, it can be called a *ga:Solution*, e.g., chiaroscuro. Finally, a *ga:Style*

is a combination of *ga:Solutions*. In that sense, if someone creates today a painting manifesting the set of solutions that defines the caravaggist style, it is manifesting this style (with no interference in periodisation issues).

However, deciding whether a style is being manifested might not be as straightforward as for solutions. First, since the style is composed of a number of solutions, it might bring some uncertainty regarding its manifestation, for example, on paintings that do not manifest all the expected solutions. In addition, it seems important to have as evidence a connection of the painter with other paintings of that style (assuming it is not the one who has created the style), more precisely a *ga:ImmaterialConsumption* event directly or indirectly via a teaching master. For instance, *The Night Watch* is not clearly a manifestation of the caravaggist style, but still could be somehow associated with that given (1) the chiaroscuro solution in common and (2) the knowledge of Rembrandt about other paintings in the caravaggist style, such as those of his teacher Pieter Lastman. Conversely, a painting by Leonardo da Vinci could never be taken as caravaggist since Caravaggio was not born yet nor the style created by him.

Now, how to compose the storyline(s) of a style or solution? The participation of the product in its creation, *ga:ProductUnderCreation*, comprises the creation of its content and presentation aspects. When they manifest a solution or style, this participation also includes (the creation of) their manifestation. We refer to the creation, since some may interpret the manifestation as extending through the whole existence of the product, while we need to restrict the time-frame. Therefore, their storyline consists of composing the events in which a solution or style is manifested, *ga:SolutionManifestationCreation* and *ga:StyleManifestationCreation*. As a side note, while some technique is always manifested by the time of the creation, a style or solution may be retrospectively applicable since they may be "defined" later in time, e.g., caravaggist style was not defined by the time Caravaggio was creating his paintings.

Although the aforementioned is the basis for their storyline(s), other constraints may be necessary in order to support the historical analysis. For example, one could split the storyline of a style based on the location of the products' creation, resulting in multiple storylines for a style. An additional constraint may regard a limited time-gap among manifestations, so that an isolated caravaggist painting would not interfere in the analysis.

Ergo, once one or more meaningful storylines are created for a style, their temporal extension can be considered worth naming, for example as Italian Renaissance or Utrecht Caravaggism. In other words, even though Renaissance or Caravaggism are not themselves periods, they can support the identification of relevant time-frames for historical analysis, eventually worth naming as a period. Therefore, the use of style for periodisation can, in fact, result in different periods, even different beginning-end for a period such as Italian Renaissance depending on how strict one uses the aforementioned constraints.

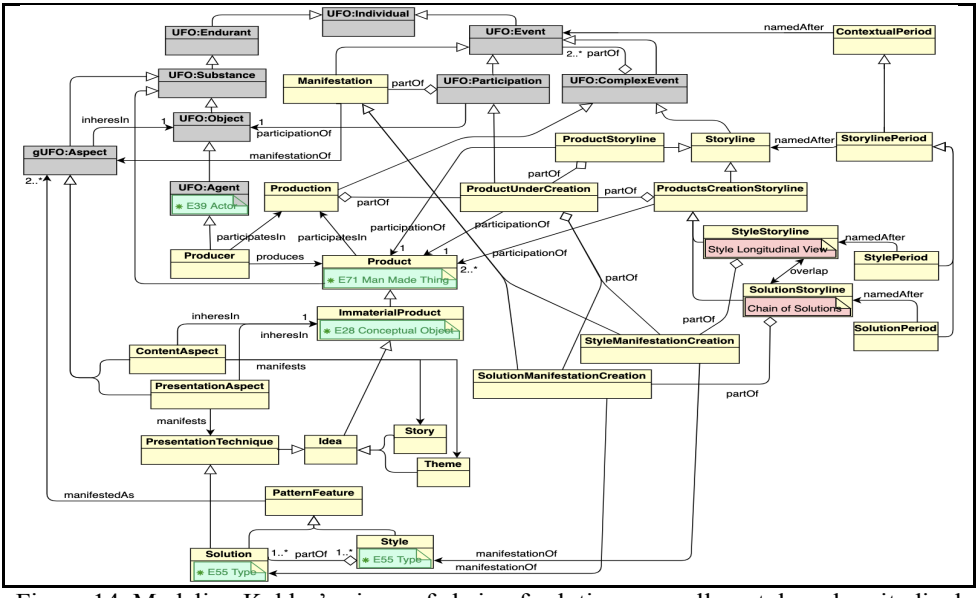


Figure 14. Modeling Kubler's views of chain of solutions as well as style as longitudinal views over the creation of products that manifest a solution or style.

CIDOC does not have specific concepts for style and solution but it does suggest means to represent them. Regarding style, two interpretations are possible according to the documentation: (1) as a *E4-Period*, which has been criticised by Kubler, and (2) as morphological object types that fall under *E55-Type*. The property *P32-used-general-technique* also has as range the *E55-Type*, which suggests that a technique (or solution) is also accounted for as such. This is compatible with our view of pattern feature, since a type is an abstraction of features expected from its instances, e.g., the type *Child* implies all its instances present as a pattern to be under a certain age limit. Finally, the concept *E55-Type* is an *E28-Conceptual-Object*, which is an *E71-Man-Made-Thing*. This means their interpretations in CIDOC are compatible with our hypothesis of them as immaterial products. Future work is to find out how human-made *CIDOC:E55-Type* relates to the supposedly equivalent *UFO:Universal*.

However, according to Kubler, a style could be better analysed via synchronous cross-section rather than longitudinally (storylines/periods). We argue that it is possible to visualise styles as storylines, although it is indeed not trivial and might not produce a unique view, as previously discussed (see Figure 9). It is not only possible, but necessary if one wants to use it for shaping the time. Nonetheless, we can also investigate how to produce Kubler's synchronous views of styles and our kaleidoscope (asynchronous) views, as well as for solutions in a similar fashion. Even though the storyline of a style or solution cannot be the storyline of its corresponding painting collection, as previously discussed, the transversal views make more sense for the latter than the former (Figure 7 illustrates transversal views). To this end, we introduce (Figure 15) the concepts *ga:StyleCollectionStoryline* and *ga:SolutionCollectionStoryline*, which are the collection of products that manifest those features. A synchronous or snapshot view of those storylines would list, at a certain time,

all (existing) products that manifest a feature, for example, all the paintings that manifest the caravaggist style in 1625 or the chiaroscuro solution in 1610. Conversely, an asynchronous or kaleidoscope view allows for accessing any of these products at different points in time, including those that were lost, for example all the paintings that were known or believed to have manifested the caravaggist style or the chiaroscuro solution at their creation time. By doing so, one could include, for example, the lost painting *The Silent* by Rembrandt in an asynchronous view of chiaroscuro solutions, if it is believed to have manifested this solution, or even the lost Caravaggio painting *Nativity with St. Francis and St. Lawrence* to a caravaggist style view.

Another interesting way that Kubler proposes is to analyse styles in terms of artists' life. This is his idea of indiction as a module of duration corresponding to the phases of an artist's life—preparation, early, middle and late maturity—lasting approximately 15 years each. Certain styles could be measured by multiple indictions of durations that are longer than single human lives or which require the time of more than one person as collective durations (Kubler 1962, 99). Naturally, it will not hold for all the cases, but we can still accommodate in the model the cases for which it does. To this end, we introduce (Figure 15) the *ga:IndictionBasedStoryline*, which has as temporal extent with a *ga:IndictionBasedDuration*, e.g., 1 indiction or 4 indictions, whereas for a style we have *ga:IndictionBasedAStyleStoryline*, corresponding to indiction-sized style storylines such as those lasting one or two successive human lives (i.e., 4 to 8 indictions, approximately 60-120 years). For example, readings of developments in Rembrandt's style in periods, such as in H.W. Janson's classic *History of Art* (1962) as middle (1636-1650) and late (1650-1669) can be compared to Kubler's indiction based on modules of maturity and late maturity. Furthermore, they can be described as (sub)storylines (parts of Rembrandt's storyline) as they last 1 indiction each (approximately 15 years). Finally, it more or less corresponds to Janson's periodisation of his outdated term "Dutch Baroque style," in his publication of the same year as Kubler's *The Shape of Time* in which he positioned Rembrandt, between ca. 1610 - 1675 as it lasts 4 indictions. Given the fact that Rembrandt's "style" hardly could be associated with the caravagist style, Janson's very arbitrary Dutch Baroque style or the very generic term Dutch Realism we can indeed question how useful it is to model style on the level of periodisation as Kubler demonstrated, although the concept "style" is still in use by art historians to get a grip on changes in the history of taste.

In conclusion, we present in Figure 15 a model that summarises the presence of Kubler's concepts (marked with a **K**) and our related adaptations/interpretations. According to Kubler, a *Fiber of Duration* or a *Bundle* of them are *Longitudinal Views* of entities through time. They can also be observed transversally as a (synchronous) *Network* or Cross-Section. He argues the latter is suitable to observe styles producing a *Style Cross-Section*, while solutions are better observed longitudinally as a *Chain of Solutions*. Other complimentary concepts are presented according to our interpretation (marked with a **GA**). A *Network* can be either *Synchronous* or *Asynchronous Transversal Views*, where the former is a *Snapshot* and the latter a *Kaleidoscope View*. In addition, a *Fiber* can consist of several "sub-fibers" that we call *Longitudinal Layers* as a *Complex Object Storyline.* This can be zoomed in and out on the layers that represent the 'internal' *Bundle of Fibers* of the parts of an object. As a consequence, a transversal view on those layers produces a *Telescope View* from which more or less details/parts can be observed. Similar to Kubler's *Chain of Solutions*, a style can also be observed in a *Style Longitudinal View* obtained by

the (relevant) manifestations of the feature during the products' creation. Their transversal counterparts, however, are obtained over the storylines of the collection of (relevant) products, for which we proposed a *Style Kaleidoscope View* complementary to *Kubler's Style Cross-Section*. And finally, Kubler's idea of *Indiction* can be used as a measure unit for storylines and their parts.

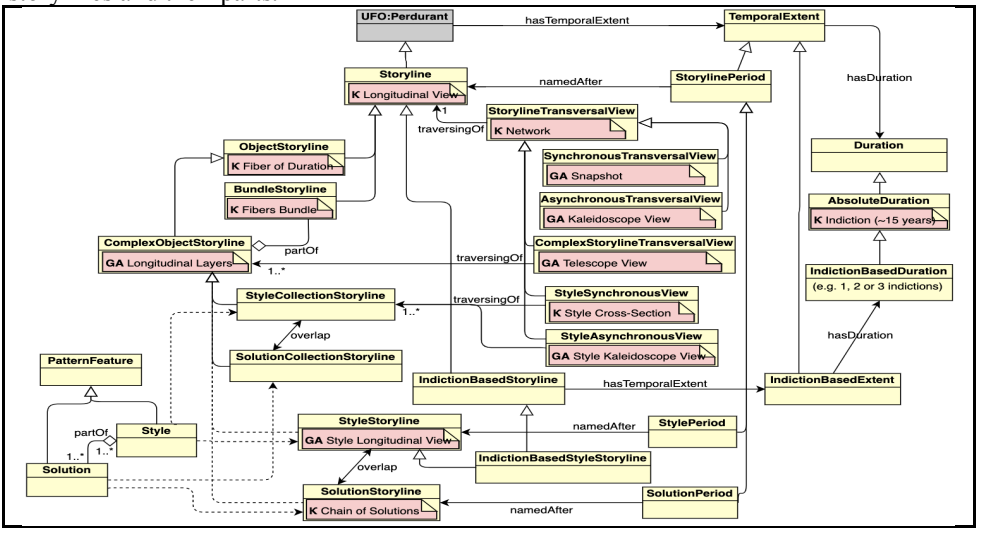


Figure 15. Modeling of Kubler's concepts and complementary interpretations related to storylines and their transversal views.

7.0 Conclusion and future work

In the context of the Golden Agents project that models historical processes of interactions between and within the creative industries of the Dutch Golden Agents as multiple narratives using the concept of "storifying data," we recognised an interesting parallel with the views on Kubler in his *Shape of Time* of 1962 on periodisation of creative production as fibers of duration based on artistic solutions instead of style. Instead of simply applying existing models of periods and events in standards such as CIDOC-CRM or PeriodO, we argued that conceptualisations of time and historical processes by historians such as that of Kubler should be taken as a point of departure for the modeling to support researchers in understanding, analysing and interacting with historical processes. We were inspired by Kubler's controversial view in the history of art that "style" is unsuitable for periodisation because different styles coexist at the same time and are in continuous flux and therefore can only be captured in an instantaneous cross-section that he described as a network. Here, we argued that Kubler had not fully grasped the potential of networks reading them in two instead of multiple dimensions and suggested for that reason to replace Kubler's own mosaic metaphor by that of a kaleidoscope to visualise his model of periodisation. Furthermore, we were interested in Kubler's empirical model of periodisation based on the life cycles of single and successive generations of artists that he brought back to modules (indictions) of (approximately) 15 years. Finally, we explored how Kubler's concept of prime objects and derivatives might be used to model the (im)material production and consumption of cultural goods in storylines in the Golden Agents project. Kubler's ideas have been

shown to be very topical, as many points are still under-addressed or partially addressed in scattered literature. Although we did not agree with all of Kubler's views they turned out to be insightful.

Therefore, we visualised Kubler's and our own perspectives using the rich history of the life and works of Rembrandt, in particular of *The Night Watch*, as a test case to formulate in total 10 requirements for our knowledge interaction model of historical interaction. Following these requirements, this historical interaction model was built on top of Unified Foundational Ontology UFO. Modelling decisions are guided herein by the rule that each introduced concept needs to fit its system of categories that makes the nature of that concept explicit. Where possible, relevant classes of CIDOC-CRM or PeriodO were mapped in the UML diagrams of the UFO-based historical interaction model. From these mappings, it became clear that several existing standard ontologies and vocabularies, such as CIDOC-CRM, OWL Time, Simple Event Model (SEM) and PeriodO did not meet our requirements in full. We believe that this not only has implications for our case study but for many semantic web applications in the humanities domain that favor data integration. One aim of our work was to find ways to reconcile concepts from the models mentioned on the basis of the formulated requirements.

All requirements for the model could be met in the parts of the historical interaction model that were visualised in UML diagrams. However, the provided visualisations of the storylines of the life and works of Rembrandt that illustrate our test case for the model of historical interaction are still static. We hope in the future to turn these static visualisations into a dynamic user interface to allow researchers to interact with the storylines in an LD paradigm including some annotation features, similar to those discussed in the cited literature on knowledge graph visualisations.

Naturally, as the proposed historical interaction model is a first attempt to materialise Kubler's ideas of time combined to our requirements, the application in practice to real data and further theoretical discussions may point out welcome improvements necessary to the model. As it is proposed, the model is truth agnostic in the sense that real or fictional events, participants and even calendars can be stated and analysed seamlessly. Important consequences of this choice are (i) likely events, as so often happens in history research for which we are not sure, can be expressed so that they can be part of the analysis that may endorse or reject them; (ii) knowingly fictional stories often mention real events or participants, which may also provide relevant input for historical research. Naturally, this position begs for (1) ways to connect the statements to one or more evidence-sources and (2) explicitly adding an epistemological layer in which statements can be taken as more or less likely facts according to someone's beliefs. An account for (1) particularly targeting archival resources are currently being developed and some preliminary results called ROAR++ can be found in van Wissen and Zamborlini 2020. The creation of an epistemological layer (2) is also under investigation for which a solution will also be proposed and published in the future.

Therefore, the conceptual model here proposed does not constitute the whole "storifying data model," which is still in development. It does provide all its different views on style, events and periodisation in relation to existing standard ontologies and vocabularies, which may require some complex modelling decisions to make important distinctions explicit. It is important to realise, however, that not all this complexity may be needed for the implementation, which will be provided in OWL also as future work.

Notes

1. The group consists of Charles van den Heuvel (Huygens ING/UvA Amsterdam, Veruska Zamborlini (University of Amsterdam), Vanessa Bartalezi Lenzi and Carlo Menghini (CNRS-Pisa), Alex Butterwoth (University of Sussex), Karl Pinneau (UTCompiegne) and Regina Varnierne-Janssen (Vilnius University).
2. For some of these papers, abstracts have been submitted but the review process has been delayed due to the coronavirus. Wissen, Latronico, Zamborlini, Reinders and van den Heuvel. 2020. "Unlocking the Archives: A Pipeline for Scanning, Transcribing, and Modelling Entities of Archival Documents into Linked Open Data." Abstract for DHBenelux2020, submitted 24 March 2020; Zamborlini, Wissen and van den Heuvel. 2020. "Reconstructions and Observations in Archival Resources: Modelling Persons, Objects and Places in the Golden Agents research Infrastructure."
3. https://www.nasa.gov/image-feature/goddard/2016/hubble-looks-into-a-cosmic-kaleidoscope
4. https://frontierfields.org

References

Allen, James F. 1983 "Maintaining Knowledge about Temporal Intervals." *Communications of the ACM*. 26: 832-43. doi:10.1145/182.358434

Almeida, João, Ricardo A. Falbo, Giancarlo Guizzardi, and Tiago P. Sales. 2020. "gUFO: A Lightweight Implementation of the Unified Foundational Ontology (UFO)." Unpublished document. http://purl.org/nemo/gufo

Bartalesi, Valentina, Carlo Meghini, and Daniele Metilli. 2017. "A Conceptualisation of Narratives and its Expression in the CRM." *International Journal of Metadata, Semantics and Ontologies* 12: 35-46.

Beghtol, Clare. 2008. "From the Universe of Knowledge to the Universe of Concepts: The Structural Revolution in Classification for Information Retrieval." *Axiomathes* 18: 131-44.

BFO (Basic Formal Ontology). 2020. https://basic-formal-ontology.org

Bliss, Henry Evelyn. 1929. *The Organization of Knowledge and the System of the Sciences,*. New York: H. Holt and Co.

Bruyn, Joshua e.a. 2015. *A Corpus of Rembrandt Paintings*, Volumes I-VI, Dordrecht: Springer Netherlands.

Butterworth, Alex, Simon Wibberley, Duncan Hay, Eirini Goudarouli, Johannes Liem, Steven Hischorn, Jo Wood, Charles Perin, Claartje Rasterhoff, Weixuan Li, Charles van den Heuvel and Duncan Speakman. 2019. "Complex Space-Time Modelling, Visualising and Performing Literary and Historical Chronotopes." In *DH 2019*. https://dev.clariah.nl/files/dh2019/boa/0635.html

CIDOC-CRM (Conceptual Reference Model). 2020. http://www.cidoc-crm.org

Colt, Priscilla. 1963. "Reviewed Work(s): The Shape of Time: Remarks on the History of Things by George Kubler." *Art Journal* 32:78-9. https://www.jstor.org/stable/774651

DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering). 2020. http://www.loa.istc.cnr.it/dolce/overview.html

Drucker, Johanna. 2009. "Temporal Modeling." In *SpecLab: Digital Aesthetics and Project in Speculative Computing*, chapter 2.1. Chicao: University of Chicago Press.

Drucker, Johanna. 2014. *Graphesis: Visual Forms of Knowledge Production (MetaLABprojects)*, Cambridge (Mass.): Harvard University Press.

Dudok van Heel, Sebastien A. C. 1982, "'Een stuck van Rembrandt genaemt de stilte' ," *Maandblad Amstelodamum* 69: 28-31.

Dudok van Heel, Sebastien A. C. 1987. *Dossier Rembrandt. Documenten, tekeningen en prenten. Catalogus van de tentoonstelling in het Museum het Rembrandthuis te Amsterdam van 17-10-1987 t/m 4-1-1988*. Amsterdam: Museum het Rembrandthuis; in samenwerking met Gemeentearchief Amsterdam.

Dudok van Heel, Sebastien A. C. 2006. *De jonge Rembrandt onder tijdgenoten: godsdienst en schilderkunst in Leiden en Amsterdam*. Rotterdam: Veenman.

Engelse, Menno den and Leon van Wissen. 2019. "Reconstructions and Observations in Archival Resources: Ontology Specification." Website. https://w3id.org/roar

Frank, Ingo. 2019. "Multi-Perspectival Representation of Historical Reality: Ontology-Based Modeling of Non-Common Conceptualizations." In *JOWO 2019: Proceedings of the Joint Ontology Workshops 2019. Episode G: The Styrian Autumn of Ontology, Graz, Austria, September 23-25, 2919*, ed. Adrien Barton, Selja Seppalä and Daniele Porello. CEUR Workshop Proceedings 2518. CEUR-WS.org/Vol-2518/paper-WODHSA3.pdf

Galton, Anthony. 2018. "The Treatment of Time in Upper Ontologies." In *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press, 33–46. doi: 10.3233/978-1-61499-910-2-33

Gerritsen, René. n.d. *Reconstructie Nachtwacht*. https://www.renegerritsen.com/reconstructie-nachtwacht-rembrandt-lundens

GFO (General Formal Ontology). 2020.: https://www.onto-med.de/ontologies/gfo

Golden, Patrick and Ryan Shaw. 2015. "Period Assertion as Nanopublication. The PeriodO Period Gazetteer." In *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web*. New York: ACM, 1013-18. https://doi.org/10.1145/2740908.2742021

Guizzardi, Giancarlo. 2005. *Ontological Foundations for Structural Conceptual Models*. Enschede: Telematica Instituut/CTIT. http://doc.utwente.nl/50826/.

Guizzardi, Giancarlo, Gerd Wagner, Ricardo Almeida Falbo, Renata S. Guizzardi and João Paulo Almeida. 2013. "Towards Ontological Foundations for the Conceptual Modeling of Events." In *ER 2013: Proceedings of the 32nd International Conference on Conceptual Modeling - Volume 8217*, ed. Wilfred Ng, Veda Storey and Juan Truillo. Berlin: Springer, 327-41. doi: 10.1007/978-3-642-41924-9_27.

Guizzardi, Giancarlo, Gerd Wagner, João Paulo Andrade Almeida and Renata S.S. Guizzardi. 2015. "Towards Ontological Foundations for Conceptual Modeling: The Unified Foundational Ontology (UFO) Story." *Applied Ontology* 10: 259-71. doi: 10.3233/AO-150157

Haase, Peter. 2019. "Knowledge Graph Kaleidoscope." Video lecture at: 16th Extended Semantic Web Conference (ESWC), Portorož 2019. http://videolectures.net/eswc2019_haase_graph_kaleidoscope/

Hjørland, Birger. 2003. "Fundamentals of Knowledge Organization." *Knowledge Organization* 30: 87-111.

iDai.chronontology. n.d. https://chronontology.dainst.org

Idrissou, Al, Veruska Zamborlini, Chiara Latronico, Frank van Harmelen and Charles van den Heuvel. 2018. "Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses." In *DH Benelux Conference 6-8 June 2018, International Institute for Social History, Amsterdam.* http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/Al-Idrissou-Chiara-Latronico_GoldenAgentsLenticularLenses_DHBenelux2018.pdf

Idrissou, Al, Veruska Zamborlini, Frank van Harmelen and Chiara Latronico. 2019. "Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers." In *K-CAP '19: Proceedings of the 10th International Conference on Knowledge Capture*, *September 2019*. New York: ACM, 259-62. https://doi.org/10.1145/3360901.3364440

Jager, Angela. 2020. "The Mass Market for History Painting in Seventeenth-century Amsterdam: Production, Distribution and Consumption." Amsterdam: Amsterdam University Press.

Janson, Horst Waldemar. 1962. *History of Art. A Survey of the Visual Arts from the Dawn of History to the Present Day*. London: Thames and Hudson.

Jensen, Matt. 2006. "Semantic Timeline Tools for History and Criticism." In: *Digital Humanities 2006. The First ADHO International Conference Université Paris-Sorbonne July 5th-July 9th. Conference Abstracts.* [Stanford, Calif.]: Stanford University Libraries, 97-100.

Kauppinen, Tomi, Glauco Mantegari, Panu Paakkarinen, Heini Kuittinen, Eero Hyvönen and Stefania Bandini. 2010. "Determining Relevance of Imprecise Temporal Intervals for Cultural

Heritage Information Retrieval." *International Journal of Human-Computer Studies* 68: 549-60.
Kubler, George. 1962. *The Shape of Time: Remarks on the History of Things*. New Haven: Yale University Press.
Kubler, George. 1967. "Style and the Representation of Historical Time." *Annals of the New York Academy of Sciences* 138: 849-55.
Kubler, George. (1979) 1987. "Toward a Reductive Theory of Style." In *The Concept of Style*. Rev. and expanded ed. by B Lang. Ithaca and London: Cornell University Press: 163-73.
Kubler, George. 1982. "The Shape of Time Reconsidered." *Perspecta*. *The Yale Architectural Journal* 19: 112-21. https://www.jstor.org/stable/1567055
Mackeprang, Maximilian, Johann Strama, Gerold Schneider, Philipp Kuhnz, Jesse Josua Benjamin and Claudia Müller-Birn. 2018. "Kaleidoscope: An RDF-based Exploratory Data Analysis Tool for Ideation Outcomes." In *UIST '18 Adjunct: The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*. New York: ACM, 75-7. https://doi.org/10.1145/3266037.3266106
Middelkoop, Norbert E. 2019. "Schutters, gildebroeders, regenten en regentessen: Het Amsterdamse corporatiestuk 1525-1850." PhD diss., University of Amsterdam.
Miksa, Francis L. 1992. "The Concept of the Universe of Knowledge and the Purpose of LIS Classification." In *Classification Research for Knowledge Representation and Organization: Proceedings of the 5th International Study Conference on Classification Research(FID),* ed. Nany Joy Williamson and Michéle Hudon.
Miller, James. (1993) 2000. *The Passion of Michel Foucault* 1st ed. Cambridge, Mass.: Harvard University Press.
Morawa, Robert, Tom Horak, Ulrike Kister, Annett Mitschick and Raimund Dachselt. 2014. "Combining Time Line and Graph Visualization." In: *ITS '14: Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces. November 2014*. New York: ACM, 345-50. https://dl.acm.org/doi/10.1145/2669485.2669544
NASA. 2016. https://www.nasa.gov/image-feature/goddard/2016/hubble-looks-into-a-cosmic-kaleidoscope
OWL-Time. 2020. Available at: https://www.w3.org/TR/owl-time
PeriodO. Available at: http://perio.do/en/
Pettegree, Andrew and Arthur der Weduwen. 2019. *The Bookshop of the World: Making and Trading books in the Dutch Golden Age*. New Haven: Yale University Press.
Pineau, Karl. 2019. "Definitions of a Life Cycle for Cultural Goods." Presentation at *RODBH* (in collaboration with Data for History org Meeting in Leipzig 4-5 April 2019).
Ranganathan, S. R. 1957. *Prologomena to Library Classification.* 2nd ed. London: The Library Association.
Rasterhoff, Claartje. 2017. *Painting and Publishing as Creative Industries: The Fabric of Creativity in the Dutch Republic, 1580-1800*. Amsterdam: Amsterdam University Press.
"RemDoc: The Rembrandt Documents Project." n.d. http://remdoc.huygens.knaw.nl/about_remdoc.html
Richardson, Ernest Cushing. 1935. *Classification: Theoretical and Practical*. 3rd ed. New York: H. W. Wilson.
Ricoeur, Paul. 1980. *The Contribution of French Historiography to the Theory of History.* Oxford: Clarendon Press.
Ricoeur, Paul. 1984. *Time and Narrative.* Trans. K. Maclaughlin and D. Pellauer. Vol. 1. Chicago: University of Chicago Press.
Rovelli, Carlo. 2018. *The Order of Time*. New York: Riverhead Books.
SEM (Simple Event Model). Available at: https://semanticweb.cs.vu.nl/2009/11/sem Accessed: 13/07/2020

Shaw, Ryan. 2010. “Events and Periods as Concepts for Organizing Historical Knowledge.” PhD diss., University of California.

Shaw, Ryan. 2013. “A Semantic Tool for Historical Events.” In *Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, ed. Eduadrd Hovy, Teruko Mitamura and Martha Palmer. W13-1205. Assoiation for Computing Linguistics. *https://www.aclweb.org/anthology/W13-1205*

Schmidle, Wolfgang, Nathalie Kallas, Sebastian Cuy and Florian Thiery. 2016. “iDai.chronOntologo.” *CAA Oslo, 30-03-2016.* https://www.academia.edu/24845165/Linking_periods_Modeling_and_utilizing_spatio-temporal_concepts_in_the_chronOntology_project

Smiraglia, Richard P. and Charles van den Heuvel. 2013. “Classifications and Concepts: Towards an Elementary Theory of Knowledge Interaction.” *Journal of Documentation* 69: 360-83.

Smiraglia, Richard P., Charles van den Heuvel and Thomas M. Dousa, 2011. “Interactions Between Elementary Structures in Universes of Knowledge.” *Classification & Ontology: Formal Approaches and Access to Knowledge: Proceedings of the International UDC Seminar 19-20 September 2011, The Hague, Netherlands*, ed. Aida Slavic and Edgardo Civallero. Würzburg: Ergon Verlag, 2011, 25-40.

Smiraglia Richard P. and Charles van den Heuvel. 2011. “Idea Collider: From a Theory of Knowledge Organization to a Theory of Knowledge Interaction.” *Bulletin of the American Society of Information Science and Technology* 37, no. 4: 43-7.

Tomich, Dale W. 2012. “The Order of Historical Time: ‘The Longue Durée’ and Microhistory.” In *The Longue Durée and World-System Analysis*, ed. R. E. Lee. Albany: State University of New York, 9-34.

Toyoshima, Fumiaki. 2019. “Ontology of Time for the Digital Humanities: A Foundational View.” In: *CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2518/paper-WODHSA10.pdf

Tussenbroek, Gabri van. 2018. *Amsterdam en de Nachtwacht*: *De mannen op het meesterwerk van Rembrandt.* Amsterdam: Uitgeverij Prometheus.

UML (Unified Modelling Language). Available at: https://www.uml.org Accessed: 13/07/2020

Van den Heuvel, Charles and Richard P. Smiraglia. 2010. “Concepts as Particles: Metaphors for the Universe of Knowledge.” In *Paradigms and Conceptual Systems in Knowledge Organization: Proceedings of the Eleventh International ISKO Conference, 23-26 February 2010 Rome Italy*., ed. Cladui Gnoli and Fulvio Mazzocchi. Würzburg: Ergon-Verlag: 50-6.

Van den Heuvel, Charles and Richard P. Smiraglia. 2013. “Visualizing Knowledge Interaction in the Multiverse of Knowledge.” In *Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands*, ed. Aida Slavic, Almila Akdag Slah and Sylvie Davies. Würzburg: Ergon Verlag, 2013, 59-72.

Van den Heuvel, Charles and Veruska Zamborlini. 2019. “Storifying Data: Modeling Historical Processes in Knowledge Graphs: Kubler’s Shape of Time Revisited.” In ADHO LOD SIG Pre Conference workshop. Ontologies for Linked Data in the Humanities–DH2019 Conference Utrecht 2019, 8 juli 2019. https://cwrc.ca/islandora/object/islandora%3A82042a5e-35bd-4986-8465-429dea5ae64e

Van Wissen, Leon and Veruska Zamborlini. 2020. “Reconstructions and Observations in Archival Sources.” YouTube video. LODLAM 2020 Challenge Entry. https://www.youtube.com/watch?v=JMO35VmUadk

Varnienė-Janssen, Regina and Jurate Kuprienė. 2018. “Authenticity and Provenance in Long-Term Digital Preservation: Analysis on the Scope of Content.” *Informacijos Mokslai* 82: 131-60. doi: https://doi.org/10.15388/Im.2018.82.9

Veyne, Paul. 1979. “Foucault révolutionne l’histoire.” In his *Comment on écrit l’histoire*. Paris: Seuil, 201-42.

Wetering, Ernst van de. 2009. *Rembrandt: The Painter at Work*, Amsterdam: Amsterdam University Press.

Wetering, Ernst van de. 2016. *Rembrandt: The Painter Thinking*, Amsterdam: Amsterdam University Press.
Zamborlini, Veruska, Arianna Betti and Charles van den Heuvel. 2017. "Toward a Core Conceptual Model for (Im)material Cultural Heritage in the Golden Agents Project.: Storifying Data." In *Semantics 2017 workshop proceedings: EVENTS September 11-14, 2017, Amsterdam.* http://ceur-ws.org/Vol-2063/events-paper1.pdf

Richard P. Smiraglia
Institute for Knowledge Organization and Structure, Inc.

Rick Szostak
University of Alberta

Chapter 7
Identifying and Classifying the Phenomena of Music[§§]

Abstract
The classification of music for information retrieval has a long history, predominantly associated with the distribution of printed music in classes based on musical medium and form. Recent research has delved into specific aspects of the classification of music such as performance and reception, in addition to the finer aspects of medium and form. Meanwhile, new input from the music information retrieval community has pointed to the potential richness of music classification that takes into account a range from simple aboutness to more auditory concepts such as listener emotion, holistic user experience, or task complexity. The extension of the classification of music in the Basic Concepts Classification requires a larger embrace of musical phenomena. A large array of musical phenomena is identified, leading to a flexible but exhaustive system of facets, and documenting the grammar of a facet analytical approach to classification of musical phenomena. A synthetic approach within a general (universal) classification can facilitate classification along diverse dimensions such as the subject of a work, the composer's intentions, and the intended audience.

1.0 Introduction: Classifying music

The classification of music for information retrieval has a long history (Smiraglia 1989; McKnight 2002). Much of the richness of the history of the creation of music classification schemes arises from the professionalization of music librarianship in the United States and United Kingdom from the early twentieth-century forward. By the mid-1950s the growth of specific practices in music libraries was synthesized as the distribution of printed music in classes based on musical medium and form (Meyer-Baer 1951 [1973]; Elmer 1957 [1973]). Meyer-Baer contrasted the broad categories (church music, vocal music, keyboard music, etc.) of the *Dewey Decimal Classification* with the granular medium-based arrays of the Library of Congress *Classification: M*, and then placed those over and against a simple pragmatic local classification that allowed the addition of style period indicators. A hallmark of music classification was the 1938 Dickinson *Classification of Musical Compositions*, originally developed at Vassar College but eventually used also at Columbia University and The Juilliard School. Dickinson's classification is medium-based, but uses a complex system of composer "book numbers" to create alpha-numeric arrays of a composers' works within a class, and also somewhat precociously makes use of what we now call facet analytical theory by permitting the addition of symbols and letters to introduce faceted indicators for arrangement, voice range, excerpt, etc. Sound recording collections, especially those in public libraries, also contributed what now might be called

[§§] Reprinted with minor editorial emendations by permission from *Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark*, ed. Marianne Lykke, Tanja Svarre, Mette Skov and Daniel Martínez-Ávila. Advances in Knowledge Organization 17. Baden-Baden: Ergon Verlag, 421-7.

"best practices" by generating highly pragmatic classifications that mimicked those of record stores, in which bins based on broad themes--e.g., Operas, Piano, Musical Shows, Jazz Music, Holiday Music, etc.--allowed library users to browse through and select among LP recordings (see "ANSCR" in Smiraglia 1989, 114ff.).

Recent research has delved into specific aspects of the classification of music such as musical performance and reception (Lee 2011, 2015), in addition to the finer aspects of medium and form, including musical medium and music ensembles (Lee 2017a, 2017b; Lee and Robinson 2017). The idea of a performance as an entity separate from the musical work, its printed instantiations, or recordings of its expressions, is critical but has received only little attention. Smiraglia (2007) demonstrated empirically the instantiation network of a set of performances, which can be thought of as "works" distinct from the musical abstractions in them. Twelve years after this paper only a few scholars have thought to distance themselves from the error of considering a performance to be a direct instantiation of a work. Cruz and Smiraglia (2020), who work with Brazilian popular music, is a notable example. For them performance of a "musical idea" instantiated through both "arrangements" and "performance expressions" is fully modeled without reference to what would be subsequently-created notated documents or recordings.

Attempts to generate more flexible systems of facets for musical concepts and rules for their combination also point to potential richness of the classification of music phenomena. The complete revision of music schedules in *DDC* in the 1980s was undertaken with a facet analytical theory in mind (Sweeney 1990). The complete set of facets arrived at included: theory, elements, techniques, character, forms, executant, composer. The use of the base *DDC* music schedules for both notated music and books about music was accomplished by designing two different citation orders; the order for music itself was "executant, forms, character," with the expectation (mirroring Dickenson) of the subsequent addition of a composer-facet symbol to create alphabetico-synthetic arrays of works under specific executants. One distinction that arose in implementation was to shift the citation order for vocal music to "forms, executants, character." A thorough analysis of the rules for generating *Library of Congress Subject Headings* (*LCSH*) for music was outlined by Young (1998). At the time, the use of *LCSH* for music required catalogers to combine terms from simple lists of medium and form with indications of number to create otherwise uncontrolled headings. Based on the contents of the cataloged artifact (score, recording), the heading could either have form or medium as lead term, to which subdivisions for the other were added. Additional subdivisions for physical form, and occasionally period of composition, were allowed. Young's detailed instructions cover every aspect of what we will later call "grammar" of music facets. A 2015 paper by Madalli, Balji and Sarangi applied ontological analytical concepts to the domain of music to generate a set of facets for a music ontology: these were "theory, person, instrument, kind, form, work." Iseminger et al. (2017, 430) describe the evolution of thesauri from elements of the former *LCSH*, revealing potential thesauro-facet arrays for the usual suspects—topical headings, genre, form [and] medium. Meanwhile, new input from the music information retrieval community has pointed to the potential richness of music classification that takes into account a range from simple aboutness to more auditory concepts such as listener emotion—e.g., amazement, solemnity, tenderness, etc.—(Aljanaki, Wiering, and Veltkamp 2016), holistic user experience (Hu et al. 2015; Downie 2003)—e.g., boring, indifferent, hopeful, circumstance, etc., or task complexity—e.g., lyrics, translation, buy or download,

etc. These IR approaches are particularly important for a classification of music that might be used for semantic web (SW) applications.

2.0 Grammar for faceting

Szostak (2017a) has described an approach to faceting that uses simple grammar to connect concepts in phenomenon-based classification. Szostak and Smiraglia (2019) reported on the exploration of this wide variety of approaches to classifying music within the Basic Concepts Classification (BCC). Szostak (2019) explores the general advantages of a synthetic approach to classification, with particular attention to the classification of music.

Since the BCC has separate schedules of things (mostly nouns), relators (verbs and conjunctions), and properties (adjectives and adverbs), the subject headings formed in BCC tend to resemble sentence fragments. Though such subject headings may surprise those used to the ungrammatical format of most subject headings in the world, there are huge advantages to a grammatical approach. First, humans spend most of their lives thinking in sentences, and can thus more readily comprehend a subject heading that is expressed in grammatical format. Second, linguists appreciate that sentences clarify the meaning of terms within a sentence. A grammatical approach thus further decreases linguistic ambiguity (and BCC terminology is generally terminology that has broadly shared understandings across disciplines and groups).

Third, the nature of a work is the ideas it expresses (see Smiraglia 2001), and these are expressed in one or more sentences, often of the form X has effect N on Y. User queries are generally also expressed in a sentence. We can do a better job of guiding users to documents if we translate the user query into a sentence-like subject heading, and likewise translate the key idea of a work into a sentence-like subject heading. We at present go from a sentence-like query to an ungrammatical subject heading to a work best defined by a sentence.

Fourth, Szostak (2017b) showed how all of the facets identified within both the Bliss Classification and the Integrative Levels Classification can be interpreted as either distinct elements of a grammatical sentence, or as distinct schedules within the BCC classification of phenomena. The BCC thus clearly expresses all facets without needing to devote notational space to facet indicators. The classifier need not explicitly perform facet analysis, but can merely translate a sentence in a document description into BCC terminology. They can, if they wish, easily check to see which facets were addressed.

3.0 Methodology: The domain analysis clinic

The extension of the classification of music in the BCC is an essential part of the Digging into the Knowledge Graph research project,[1] in which the classification of specific musical concepts rather than physical musical documents requires a larger embrace of musical phenomena. Here we describe specific work undertaken to define a larger array of musical phenomena, to generate a flexible but exhaustive system of facets, and to document the grammar of a facet analytical approach to classification of musical phenomena. In November 2019 a small group of experts in the classification of music was assembled at the Institute for Knowledge Organization and Structure, Inc. (IKOS) in Lake Oswego, Oregon (USA). The group constituted what IKOS has called a “domain analysis clinic” (DAC) on “the phenomena of music for classification.” The general outline of a DAC includes an invitation-only group, assigned “homework” to build an exhaustive corpus of

relevant research from which segments of meta-analysis are generated. When the group meets the meta-analysis is reviewed, synthesis is constructed, and follow-up assignments are fixed with the purpose of filling identified gaps in knowledge of the specific domain (Smiraglia 2019). Szostak and Smiraglia (2019) focused on how a synthetic approach within a general (universal) classification could facilitate classification along diverse dimensions such as the subject of a work, the composer's intentions, and the intended audience. Participants in addition to Szostak and Smiraglia were Deborah Lee, Richard Griscom, J. Bradford Young and Joshua A. Henry. Specific details of the meta-analysis and the generation of facets are reported in Szostak et al. (forthcoming). What follows here is the general outline of the fleshing out of schedules of musical phenomena for the BCC.

4.0 Musical phenomena in faceted arrays

Upon review of the meta-analytical data, the group arrived at the following set of musical phenomena that should be developed or extended for the BCC:

Character, occasion and function of the music

Types, forms and genres of music

Medium of performance

Commercial elements of recorded music

Format (arrangement, transcription, transformation, etc.)

In addition, consideration was given to traditionally relevant concepts such as the personal names of creative contributors (composers and librettists, but also sound editors, producers of performances, etc.) and to representations of place and time. BCC already allows synthesis of names, places and time designations.

4.1 Form, genre, etc.

It was decided to combine the elements identified above as "character, occasion, function, type, form and genre" into a single facet. The structure of this facet is to be based on the Library of Congress thesaurus for form and genre terms (LC Genre/Form Terms or LCGFT). LCGFT is maintained as linked open data (LOD) by the Library of Congress, with ongoing input from the active library community, including the Music Library Association (Library of Congress 2020):

> The Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT) is a thesaurus that describes what a work is versus what it is about. For instance, the subject heading Horror films, with appropriate subdivisions, would be assigned to a book about horror films. A cataloger assigning headings to the movie The Texas Chainsaw Massacre would also use Horror films, but it would be a genre/form term since the movie is a horror film, not a movie about horror films. The thesaurus combines both genres and forms. Form is defined as a characteristic of works with a particular format and/or purpose. A "short" is a particular form, for example, as is "animation." Genre refers to categories of works that are characterized by similar plots, themes, settings, situations, and characters. Examples of genres are westerns and thrillers. In the term Horror films "horror" is the genre and "films" is the form.

Some of the genres identified by the Library of Congress would be treated differently by the BCC. Most obviously, "humorous" is not really a distinct genre but a property that might be attributed to music from many different genres. One of the beauties of the synthetic approach taken by BCC is that terms from non-musical schedules can be used as necessary in the subject headings for works of music. "Humorous" is already a property within schedule Q of the BCC.

LCGFT is not music specific but the group easily extracted musical phenomena from the list, which can form the basis of a hierarchical array for BCC. LCGFT does include terms relating to styles and kinds of music.

A decision also was taken that occasions, functions, and character could be synthesized by adding terms from elsewhere in the BCC. The BCC already contains schedules that encompass various types of celebration, group, organization and time period. One schedule that we hope to expand upon is the schedule CR regarding religion. We intend to identify in more detail the various kinds or parts of religious services, e.g., baptism, offertory, etc.

4.2 Medium of performance

Traditionally the basis of most music classifications, medium of performance is obviously an essential facet. The BCC has imported the Hornbostel-Sachs instrument classification. This classification attempts global coverage, and provides a hierarchical structure grounded in the physical characteristics of instruments. However, the taxonomical terms used are not particularly directly the names of the "phenomena" of musical medium. For example, "flute" is embedded in a hierarchy of "aerophones," and "piano" is under "pianoforte" embedded in a hierarchy of "chordophones." The group urged incorporation of the Library of Congress Medium of Performance Thesaurus for Music (LCMPT), which like LCGFT is maintained in consultation with the Music Library Association (http://id.loc.gov/authorities/performanceMediums.html):

> The Library of Congress Medium of Performance Thesaurus (LCMPT) is a stand-alone vocabulary that provides terminology to describe the instruments, voices, etc., used in the performance of musical works …. Authorized terms and references in LCMPT generally consist of single words and phrases, but parenthetical qualifiers are occasionally employed to differentiate among homonyms. All terms and references are in the singular form ... (e.g., flute; saxophone ensemble; but Irish harp). The thesaurus has a few broadest terms as listed in the "Top Scheme Members" section. Each of the other terms is hierarchically subordinate to one or more of these terms and exhibits the class/class member relationship. Most of the authorized terms have Used For (UF) references for synonyms. Scope notes are also provided in many cases, and may describe the medium's physical structure, the time period in which it was popular, and/or its geographic origin.

For BCC, the group encouraged harmonization of the existing BCC schedule with LCMPT, and this task was assigned for work in early 2020.

4.3 Audiography

A new facet was outlined broadly with regard to input from the IR and SW communities. The general structure of the facet is to include:

> Details of capture (i.e., where and when was a performance recorded)
>
> Details of production and dissemination (release, music recording number, etc.)
>
> Physical or digital format (soundtrack, single, compilation, track number, etc.)
>
> User's purpose: settle a bet, gift, etc.
>
> Emotion invoked by the music

User studies have shown that the entities on this list are those often sought by people looking for music online. Perhaps the most controversial part of the group's discussion, this facet was tasked for detailed explanation in early 2020. It is worth noting that details of capture, emotions and purposes likely can be synthesized from existing arrays in the BCC.

5.0 Conclusion: Toward the grammar of faceted music classification

Classificationists can usefully ask what sort of queries a user might have. We might reasonably expect that users will want to search for works from a particular genre, or about a particular subject (love songs, say), or with a particular purpose (revolutionary songs), or for a particular occasion (wedding songs), or with a particular melody (maybe to accompany a particular video), or to invoke a particular emotion. And think of a group of musicians that want to play together and thus seek works designed for the particular set of instruments that they play. The simple fact is that all of these searches are either difficult or impossible within existing approaches to music classification. Szostak and Smiraglia (2019) detailed how the synthetic approach of BCC facilitated the classification of works by subject, many occasions, multiple creators (for when a work is rearranged), and many aspects of culture. The present project seeks to develop new schedules that will further enhance the classification of music. Though challenges remain, we are confident that we can satisfy user queries much better than is possible at present.

Note

1. Digging Into the Knowledge Graph. https://diggingintodata.org/awards/2016/project/digging-knowledge-graph

References

Aljanaki, Anna, Frans Wiering, and Remco C. Veltkamp 2016. "Studying Emotion Induced by Music Through a Crowdsourcing Game." *Information Processing & Management* 52: 115-28.

Cruz, Fernando and Richard P. Smiraglia. 2020. *Towards a Conceptual Model for Popular Music Representation*. Working paper. Lake Oswego, OR: Institute for Knowledge Organization and Structure.

Dickinson, George Sherman. 1938 (2002). *Classification of Musical Compositions: A Decimal-Symbol System*. Poughkeepsie, N.Y.: Vassar College. HathiTrust digitized by Univ. of Michigan. http://hdl.handle.net/2027/mdp.39015040124524

Downie, J. Stephen. 2003. "Music Information Retrieval." *Annual Review of Information Science and Technology* 37: 295-340.

Elmer, Minnie. 1957 (1973). "Classification, Cataloging, Indexing." In *Reader in Music Librarianship*, edited by Carol June Bradley. Reader Series in Library and Information Science. [Washington, D.C.]: Microcard Editions Books, 148-55.

Hu, Xiao, Jin Ha Lee, David Bainbridge, Kahyun Choi, Peter Organisciak, and J. Stephen Downie 2015. "The MIREX Grand Challenge: A Framework of Holistic User-Experience Evaluation in Music Information Retrieval." *Journal of the Association for Information Science and Technology* 68: 97-112.

Iseminger, Beth, Nancy Lorimer,. Casey Mullin, and Hermine Vermeij. 2017. "Faceted Vocabularies for Music: A New Era in Resource Discovery." *Notes* 73: 409-31.

Lee, Deborah and Lyn Robinson. 2017. "The Heart of Music Classification: Toward a Model of Classifying Musical Medium." *Journal of Documentation* 74: 258-77.

Lee, Deborah. 2011. "Classifying Musical Performance: The Application of Classification Theories to Concert Programmes." *Knowledge Organization* 38: 530-40.

Lee, Deborah. 2015. "Consumption, Criticism and Wirkung: Reception-Infused Analysis of Classification Schemes." *Knowledge Organization* 42: 508-21.

Lee, Deborah. 2017a. "Numbers, Instruments and Hands: The Impact of Faceted Analytical Theory on Classifying Music Ensembles." *Knowledge Organization* 44: 405-15.

Lee, Deborah. 2017b. "Numbers, Instruments and Hands: The Impact of Faceted Analytical Theory on Classifying Music Ensembles." In *Faceted Classification Today, Theory, Technology and*

End Users: Proceedings of the International UDC Seminar, 14-15 September 2017, London, United Kingdom, edited by Aida Slavic and Claudio Gnoli. Würzburg: Ergon Verlag, 77-92.
Library of Congress. 2020. *Library of Congress Genre/Form Terms*. Washington, D.C.: Library of Congress. http://id.loc.gov/authorities/genreForms.html
Madalli, Devika P. [et al.]. 2015. "Faceted Ontological Representation for a Music Domain." *Knowledge Organization* 42: 8-24.
McKnight, Mark. 20002. *Music Classification Systems*. Music Library Association Basic Manual Series no. 1. Lanham, Md.: Scarecrow Press and the Music Library Association.
Meyer-Baer, Kathri. 1951 (1973). "Classifications in American Music Libraries. In *Reader in Music Librarianship*, ed. Carol June Bradley. Reader Series in Library and Information Science. [Washington, D.C.]: Microcard Editions Books, 172-76.
Smiraglia, Richard P. 1989. "5. Subject Analysis," "6. Verbal Subject Analysis," "7. Classification." In *Music Cataloging: The Bibliographic Control of Printed and Recorded Music in Libraries*. Englewood, CO: Libraries Unlimited, 63-120.
Smiraglia, Richard P. 2001. *The Nature of "A Work:" Implications for the Organization of Knowledge*. Lanham, Md.: Scarecrow Press.
Smiraglia, Richard P. 2007. "Performance Works: Continuing to Comprehend Instantiation." In *Proceedings of North American Symposium on Knowledge Organization* 1: 75-86 https://journals.lib.washington.edu/index.php/nasko/article/view/12836/11317
Smiraglia, Richard P. 2019. "The IKOS Domain Analysis Clinic: Toward Critical Problem Solving." *IKOS Bulletin* 1: 5.
Sweeney, Russell. 1990. "Grand Messe des 780's (With Apologies to Berlioz)." In *In Celebration of Revised 780: Music in the* Dewey Decimal Classification *Edition 20*, edited by Richard Bruce Wursten. Canton, MA: Music Library Association, 28-38.
Szostak, Rick and Richard P. Smiraglia. 2019. "Classifying Music within the Basic Concepts Classification." In *Proceedings of the Annual Conference of CAIS; Actes du Congrès Annuel de L'acsi* 2019. https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/1064/948
Szostak, Rick. 2017a. "Theory versus Practice in Facet Analysis." In *Faceted Classification Today, Theory, Technology and End Users: Proceedings of the International UDC Seminar, 14-15 September 2017, London, United Kingdom*, edited by Aida Slavic and Claudio Gnoli. Würzburg: Ergon Verlag, 2017, 259-69.
Szostak, Rick. 2017b. "Facet Analysis Without Facet Indicators." In *Dimensions of Knowledge: Facets for Knowledge Organization*, edited by Richard P. Smiraglia and Hur-Li Lee. Würzburg: Ergon Verlag, 69-85.
Szostak, Rick. 2019. "A Synthetic approach to the Classification of Music." *El Profesional de la Information* 29, no. 1, e290105.
Szostak, Rick, Deborah Lee, Richard Griscom, Joshua Henry, J. Bradford Young, and Richard P. Smiraglia. Forthcoming. "Technical Report: The Phenomena of Music for Classification.". *IKOS Technical Reports Series* 1. Lake Oswego, OR: IKOS.
Young, J. Bradford. 1998. "Introduction to the Structure and Use of Library of Congress Subject Headings for Music and Material about Music." In *Music Subject Headings*, 2d ed., comp. Harriette Hemmasi, with technical assistance of Fred Rowley, introd. by J. Bradford Young, Foreword by R. P. Smiraglia. Soldier Creek Music Series. Lake Crystal, Minn.: Solider Creek Press, 1-28.

M. Cristina Pattuelli
Pratt Institute

Chapter 8
Graphing Out Communities and Cultures in the Archives
Methods and Tools

Abstract
Linked Jazz is a project exploring the potential of linked open data (LOD) in the area of jazz history and archives to reveal the rich web of relationships among jazz musicians. The graph-based data structure provides the framework necessary to represent the densely interconnected relationships that tie together the jazz community. While most cultural heritage efforts were converting legacy metadata, Linked Jazz mines the text of digitized primary sources (oral histories) to generate original knowledge in the form of LOD. A collection of over fifty transcribed interviews from various jazz archives across the United States served as the data source from which "native" RDF triples were generated. A crowdsourcing tool, Linked Jazz 52nd Street, served as a working prototype showing the power of leveraging volunteers' efforts to semantically augment the Linked Jazz graph. The ontological structure of the Linked Jazz knowledge graph is centered around the entity *Person* and consists of twelve predicates describing individual connections, from the predicate *"knows of"* to more specific predicates expressing various degrees of personal closeness. We enriched the biographical descriptions of musicians with the gender attribute, in order to analyze the Linked Jazz network through the lens of gender. The full potential of linked data is reached when heterogeneous data from different sources are interlinked providing unified access to data and the possibility to seamlessly query multiple graphs. Linked Jazz leveraged linked data technologies and the power of knowledge graphs to represent the community of jazz musicians whose personal and professional relationships are dense and intertwined.

1.0 Introduction
As the linked open data (LOD) initiative continues to grow making the vision of the semantic web an ever more tangible reality, the cultural heritage community has played a key role in its development. Because the fields of arts and humanities are built on complex relationships, they provide an ideal context upon which to apply methods of knowledge representation like linked data (LD) that connect entities, people, objects, facts and concepts, in new and unprecedented ways, across disparate domains and beyond repository boundaries.

This chapter describes Linked Jazz[1], a long-running project exploring the potential of linked open data in the area of jazz history and archives. The project applies a combination of automatic computational methods and LD technologies to digital content to reveal the rich web of relationships that exist among jazz musicians as recorded in primary sources, such as oral histories. The graph-based data structure that underlies the LD architecture provides the framework necessary to represent the densely interconnected relationships that tie together the jazz community. This approach is especially well suited for graphing out communities and representing interconnected person entities as social networks—a unified view of the data to draw deeper insights on otherwise fragmented and dispersed information.

Networks and knowledge graphs are an area of active research in the digital humanities. The emerging field of historical network research uses networks for the study of the past because of their ability to place historical data in complex and interconnected contexts that offer new perspectives for interpretation (Kerschbaumer et al. 2020; Morrissey 2015). When generated and powered by LOD technologies, knowledge graphs' power is magnified because they are, by design and by core principles, publicly available, transparent, shareable, and broadly reusable. Archives are a relatively new territory for LD applications, but a strategic one not only for the increased visibility of and access to primary sources collections, but, even more importantly, for the generation of novel modes of engagement with historical documents that enables the researcher to draw new and deeper insights from their content. An increasing number of projects—from large-scale initiatives like the Social Networks and Archival Context (SNAC)[2], which provides an infrastructure for discovering and connecting archival collections, to domain-specific projects like CultureSampo[3] that offers enhanced possibilities for the study of Finnish culture—have moved cultural and historical scholarship forward in important ways.

At the crossroads of digital humanities and archival research, Linked Jazz has encountered a number of different challenges that come from working in uncharted territory both technically and methodologically. The LD paradigm, with its open and boundless architecture, has redefined the boundaries of our traditional practices of information organization, from blurring the conventional lines between data and metadata, to reshaping the notion of authority control and data curation. These new methods of knowledge representation have the potential to subvert how research is practiced in archives and special collections, from data collection to the analysis and interpretation of historical data.

2.0 Text to graph

The Linked Jazz Project began in the early days of the LD initiative prompted by an interest in experimenting with what was then an evolution of the semantic web. The emerging digital semantic technologies were, from our perspective, a natural extension of the knowledge organization (KO) field. Jazz history provided a unique and engaging real-world scenario.

While most of cultural heritage efforts were concerned with converting legacy metadata into LD, Linked Jazz focused instead on the actual digital content as its source of data. In other words, we leveraged the text of digitized primary sources, oral histories in particular, to generate original knowledge in the form of LOD. A rather unique approach at that time, the process of progressing from text to knowledge graph involved digging into vast amounts of textual content using a series of methods such as automated text analysis techniques.

2.1 Methodology

A collection of over fifty transcribed interviews from various jazz archives across the United States[4] served as the data source from which "native" RDF triples were generated. Linked Jazz data were created through a development process that included named entity recognition and extraction, identity management and linking, semantic enrichment via crowdsourcing and manual annotation. Proper names of musicians were located and extracted from text and relationships were encoded through the predicate "knows of" that expresses a basic connection between an interviewee (the subject of the triple) and a musician mentioned in the text. The assumption behind each statement is that if the interviewee

cites a person, or any entity we intend to represent, we can assert with a high degree of confidence that the narrator has at least an elemental level of knowledge of that person. The outcome from processing this pool of interviews was a knowledge graph representing over 2,000 musicians interconnected through more than 9,400 relations.

With traditional research practices, the creation of such a dataset would entail a labor-intensive and time-consuming data collection process, digging through piles of documents to find relevant mentions that would then need to be annotated and compiled for analysis. The implications of using linked data technologies for archival research are even more significant when it comes to the shareability of the data generated. The datasets are intended to be publicly available and reused by scholars, as well as other developers.

The production of triples was carried out using a home-build tool called the Transcript Analyzer[5]. The tool employs open source software (i.e., Stanford NLP NLTK libraries) to support entity extraction and data linking with only partial human supervision. To make it possible to extract names of only jazz musicians from the interviews, a directory of jazz musicians in the form of RDF triples had to be created to support the identification of the relevant entities and associate a URI to them. Building such a domain-specific name directory was not an easy task due to the limitations of name authorities like the Library of Congress Name Authority File (LCNAF) and VIAF where only a small subset of jazz musicians could be found. Traditional bibliographic name authorities also fell short when it came to the inclusion of lesser-known musicians. More comprehensive linked data hubs like DBpedia, which derives its data from Wikipedia, were used instead; however, harvesting and filtering out person entity occurrences by type of occupation proved difficult because of the inconsistent classification of professions and music genres. As the LD environment continues to evolve, more sources of reliable URIs are becoming available, both general and domain-specific, including the vast and ever-growing source Wikidata[6], the free and collaborative knowledgebase.

We then performed entity resolution and reconciliation, which addressed the tasks of disambiguating homonyms, detecting inaccuracies, and assigning standard identifiers, using a dedicated application embedded in the analyzer[7]. To maximize the quality of the data output, we combined the automated approach with human supervision consisting of manually validating matches when multiple options occurred. Different iterations of the application were built over time with the goal of scaling up the identity management process that, although primarily automated, still required human assistance. This included an external data service called Ecco![8], an Italian term that emphasizes quick and effortless delivery. Ecco! was used internally but also intended for external use in outside projects. While a handful of other identity management tools for linked data existed, Ecco! was unique in that it was web-based and offered an intuitive user interface that gave users the ability to contribute in a distributed and incremental way, making identity management a cooperative and collaborative activity.

Identity management requires a whole new level of effort when it comes to entities for which a URI doesn't exist. This was the case for a number of musicians mentioned in oral histories who had not achieved a level of notability. We frequently encountered jazz artists who had not conformed to the criteria for inclusion common in name authorities because, for example, they had not been listed as a contributor on recordings or they had failed to achieve a certain level of public recognition. Nevertheless, they needed to be accounted for as they were mentioned in the source documents we were processing, so new URIs were

minted into the Linked Jazz namespace (e.g., http://linkedjazz.org/resource/Lynn_Grissett). When creating public identifiers, there are practical matters that need to be considered including ensuring that evidential documentation is provided to justify the minting and that the naming agency take responsibility to manage its local URIs for persistence and traceability. Moreover, there are cultural and socio-political implications that go along with plucking individuals from obscurity and weaving them into the fabric of LOD. Archives are filled with names of local, long-forgotten or less-prominent individuals. LOD technologies have the power to expose people and information previously overlooked, bringing them to the forefront of the historical record. We can only imagine how the inclusion of a whole new array of people and entities brought out of primary sources and incorporated into knowledge graphs could transform archival research.

2.2 Crowdsourcing

While automated techniques worked well to generate a graph based on a semantically non-committal predicate like "knows of," more specificity, such as the degree of closeness or the type of collaboration, would be difficult to express through computational methods alone. To overcome these limitations, hybrid approaches that combine automation and human intervention were employed. Crowdsourcing was the methodology adopted, meaning that the task of interpreting the nuances of the relationships was handed over to jazz experts and enthusiasts using a dedicated crowdsourcing tool called Linked Jazz 52nd Street (https://linkedjazz.org/52ndStreet/). A web-based application, this tool displays sequenced excerpts of an interview transcript along with the list of musician names mentioned. The user is asked to classify the type of relationship held between each pair of musicians based on their understanding of the text. A list of predefined predicates describing the relationship was provided. Once a predicate was selected, an RDF triple was automatically created and fed back into the existing data set. Linked Jazz 52nd Street has served as a working prototype quite effectively showing the power of leveraging volunteers' efforts to semantically augment the Linked Jazz graph. While in place only for a limited period of time during its testing phase, the crowdsourcing tool was accessed and used by more than five hundred registered users who contributed more than 9,200 annotations. The significance of the Linked Jazz 52nd Street crowdsourcing tool extended beyond its practical effectiveness to demonstrate its potential to build community and engage jazz researchers and aficionados with primary sources and archival collections.

3.0 Linked Jazz ontology

The ontological structure of the Linked Jazz knowledge graph, in its current iteration, is relatively simple. The model is centered around the entity *Person* and consists of twelve predicates describing individual connections, from the predicate *"knows of"* which serves as the most basic connector to more specific predicates expressing various degrees of personal closeness (*"has met," "acquaintance of," "friend of"*) and/or professional ties (*"influenced by," "mentor of," "collaborated with"*). Jazz experts and jazz archives users were consulted to help select the appropriate pool of relationships. Whenever possible, the predicates were derived from existing RDF vocabularies[9] to enforce consistency and facilitate interlinking. To represent more nuanced types of collaborations—a key professional relationship in our context—five original predicates were created and minted (*"played to-*

gether," "in band together," "toured with," "bandleader of," "band member") and modeled as sub-properties of *"collaborated with"*. The ontology resulted in a small application profile with local extensions needed to represent the desired degree of semantic granularity (Figure 1).[10]

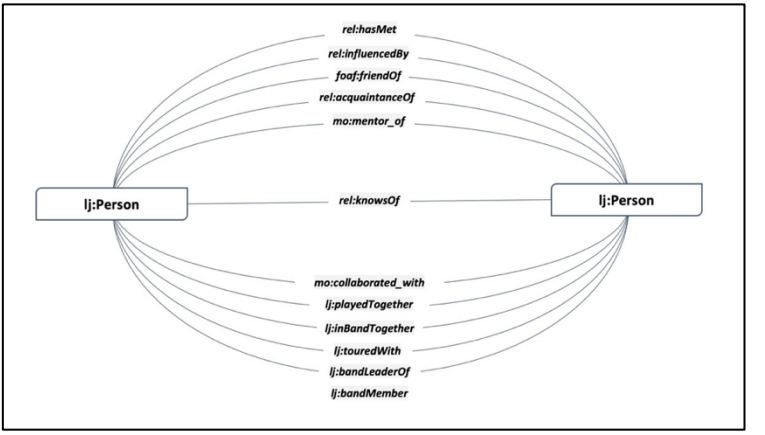


Figure 1. The Linked Jazz ontology (image by Sarah A. Adams).

The linked data paradigm has significantly altered the notion of knowledge representation as well as the practice of ontology engineering employed in traditional artificial intelligence (AI) and in the semantic web initiative. Shallow in structure, with few modeling primitives, and lightweight, with minimal constraints, LOD ontologies often take the form of application profiles, mixing and matching classes and predicates from different RDF schemas. Semantic reuse is at the core of LOD to foster interoperability and interconnectivity. For representing individuals, however, the options provided by well-established RDF vocabularies were and still are rather limited. What can be said about a person is greatly shaped by the legacy of bibliographic authority records and by the way people have been defined in a document-centered context. As Tamper et al. (2018) note, to study groups of people through their biographical data, for example for biographical and prosopographical research, multiple dimensions (biographical, familial, personal, social) would need to be represented. We experienced firsthand the lack of suitable semantics when trying to expand the views a person could be seen through and account for different contexts through which a "social identity" is constructed.

4.0 Linked Jazz data access and consumption

Processing fifty-four oral histories generated over 2,000 musicians connected through a web of more than 20,000 relationships.[11] Access to the dataset was provided via a SPARQL endpoint that allows the datasets to be queried.[12] The composition of queries in languages such as SPARQL requires an understanding of the syntax of the language and some knowledge of the data content and structure, a likely barrier for non-expert users. The knowledge graph was made available as a social network thanks to a web-based interactive visualization tool[13] (Miller et al. 2012). This mode of access and consumption of the Linked Jazz dataset relies on an intuitive and engaging public interface that has been pivotal in generating and sustaining interest in the project and in showing the value of using LOD to

different groups of potential users, from scholars to musicians and jazz enthusiasts. The visualization of the Linked Jazz dataset in the form of networks makes the data comprehensible at a glance while retaining its analytical capabilities. Various configurations of the data are possible starting from a high-level view of the entire dataset that can be explored by navigating between nodes and edges. Navigation through large networks can be overwhelming for the users, as noted in the literature (Lévesque et al. 2020), so different configurations were offered. It is possible, for example, to focus on a radial view of an individual musician displaying their ego-network of relationships, to search by the name of one or multiple musicians to discover all of the shared connections that exist between those musicians, or to infer the absence of connections (Figure 2). It is also possible to "inspect" a connection by hovering over an edge that would then trace back to its primary source, the textual passage from which the relationship originated—a useful feature for scholarly research. Also, to conduct social network analysis (SNA), Gephi files for different views can be downloaded from the use interface.

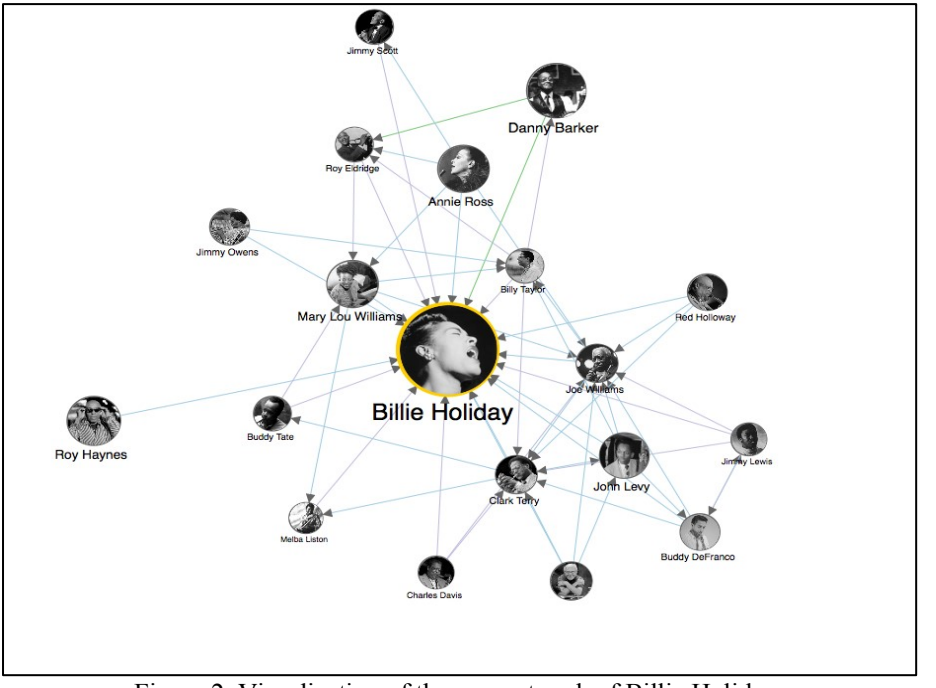


Figure 2. Visualization of the ego-network of Billie Holiday.

5.0 Technology shift

The tools and applications required to perform the full cycle of the Linked Jazz development were built in-house and have each reached the status of working prototypes. As the project has continued to grow, methods and tools have been reviewed and revised to reflect the evolution of LD practices and to experiment with new technological developments. Linked Jazz is currently going through a technology shift that includes the redesign and prototyping of a new set of tools and data services to support end-to-end graph development and the adoption to a new platform. This transition will have a deep impact on all the

steps of the production cycle—from the way we generate and manage LD to the way we store and consume them. Besides streamlining the process, the new technical stack is intended to lower barriers to entry for users seeking to create linked open data from archival material and make the entire process more distributed and collaborative.

5.1 Retooling

A new set of tools, collectively labeled DADAlytics, is currently under development with the aim of supporting the full cycle of linked data production. Building on the lessons learned and experience acquired in the first phase of Linked Jazz development, DADAlytics expands the functionality of earlier tools, while trying to respond to the demand for more intuitive and easy-to-use tools, so information professionals and humanities researchers are able to participate in LD creation without requiring specific technical expertise. A re-engineered version of the former Transcript Analyzer, DADAlytics has been designed to facilitate text analysis and identity management in order to generate RDF triples through a seamless, integrated process. Open source and web-based, DADAlytics has also been designed to accommodate a broader range of documents—virtually any kind of textual resource. It consists of two integrated modules. The first component applies Named Entity Recognition (NER) methods to digitized text for locating and extracting entities of interest in the form of name instances and entity types. It harnesses the power of six language processing tools and software applications that work in combination to detect and extract named entities using a range of statistical models, including neural networks techniques for machine learning.[14] The second module supports identity management, where a URI from a name authority is associated with the individual tokens, name variants are reconciled to remediate ambiguity, and new identifiers for local entities can be minted.

The toolchain exploits the close synergy between automatic and manual processes. While entity names can be recognized and entity types classified with limited manual intervention, mainly for rectifying incorrect and missing entities, human interpretation is still needed for typing relations between entities based on text (a sentence or a set of sentences). DADAlytics is designed to process a wider range of textual documents, the type typically found in digital archives and special collections, such as oral history transcripts, letters, diaries, personal narratives, theater booking ledgers, and so on. This opens up the possibility to represent a broader range of entity types beyond just people. The tool also offers greater flexibility in how the structure of a document is leveraged and how clusters of triples are derived from sub-sections of the document itself. More specifically, text can be manually segmented into meaningful units, each serving as the data source and contextual framework for triples derived from different configurations of entity co-occurrences within a unit.

6.0 Data Infrastructure: Wikibase

At the core of the technological reframing of the project is a new data platform that enables a radical change in the ways we create, store, manage and access linked data. Wikibase[15] is the free software that powers Wikidata. We use a locally installed instance of Wikibase[16] which makes it possible to work in a live system that holds all the utilities that Wikibase provides, while retaining control of the data. In this unified infrastructure, data sit alongside the tools and other applications that generate, curate, store and consume them. This allows for triples from different areas—from content to administrative data, including revision

tracking and history—to be integrated in a seamless data environment, blurring the lines between data and metadata. This also makes it possible for knowledge graphs to be enriched (and queried) with new types of data such as provenance data. Wikibase makes it easier for us to host and interlink old and new LD projects. It also helps them evolve and expand seamlessly. A few initiatives in the cultural heritage field have explored the potential of Wikibase including the web-based organization Rhizome[17], which relies on Wikibase for their archive of born-digital art (Fauconnier 2018) and OCLC[18]. OCLC has recently conducted a year-long pilot in the library sector that was deemed promising for "pushing the traditional notion of bibliographic records/cataloging practices/library standards challenged" (Godby et al. 2019). Our research group has begun experimenting with the new technical infrastructure and exploring the LD capabilities and ancillary utilities it offers (Miller 2018). From a KO viewpoint, data modeling is perhaps where Wikibase most impacts our development methods. As mentioned earlier, a pragmatic approach drives the creation and use of ontologies in the LD ecosystem, which have become lighter in semantics and easier to adopt and share (Pattuelli et al. 2015). Conceptual models with formal restrictions, such as those formulated in logical terms, would be more expressive, but would defy the underlying principles of LOD where interoperability and reuse are needed to interconnect large amounts of heterogeneous data in an open and distributed information space like the web. Initiatives such as Wikidata and Wikibase have furthered this shift making data modeling a more open and participatory activity (Pattuelli et al. 2019).

Because Wikidata is coupled with Wikibase, the set of capabilities and utilities of Wikidata are also passed down to our instance. As a result, we have transitioned to a Wikidata-like data structure. Wikidata is essentially a collection of entity pages that include statements. Each statement is made up of a claim in the form of a property-value pair. The building blocks of a statement are "items" and "properties", equivalent to what RDF defines as "entities" and "predicates." A richer model than RDF, Wikidata provides a way to further refine claims through optional "qualifiers." These qualifiers specify the context in which a claim is deemed valid. For example, statistics on the population of New York City are related to the year they are based on:

> New York City (item Q60) → population (property P1082) → 8,175,133 (value) → point in time (qualifier, property P585) → 2010 (value).

In addition, a claim can be annotated with references, anything from websites to datasets, providing a verifiable source for that claim:

> stated in (property P248) → 2010 United States Census (item Q523716).

Accountability through traceability is as important to archival research, as any other kind of research. In the context of our project, where social networks are built based on mentions in text, this feature allows the contextualization of each connection within the passage it was derived from, serving as a trademark on the data (Figure 3). Statements can be verified directly online, enforcing transparency and, when aggregated, enabling the possibility of deeper analysis.

Using Wikibase as our technical infrastructure involves drawing on, and taking advantage of, the Wikidata ontology model as well. As a result, our conceptual model relies on basic modeling constructs: classes and properties, and hierarchical relationships (*"subclass of," "subproperty of," "instance of"*) for taxonomic organization of instances

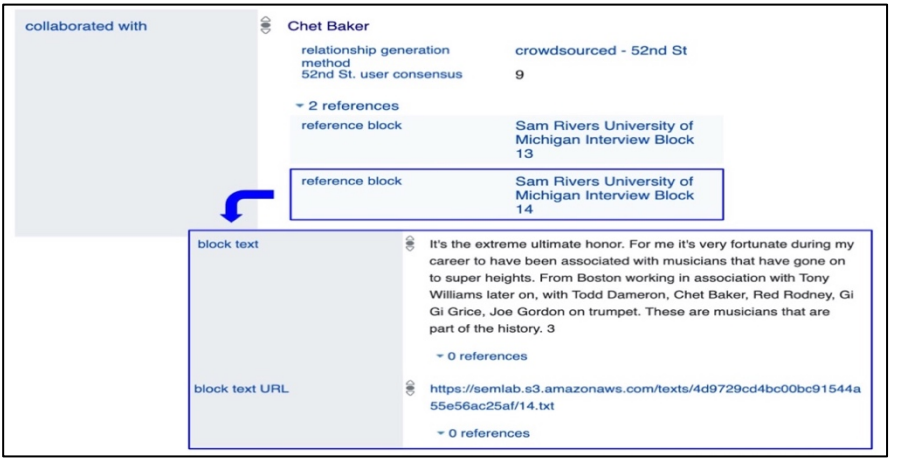


Figure 3. Properties in context on Sam River (Q26) Wikibase page.

to support subsumption in queries. The equivalence property between classes and between properties is also represented to enable schema alignments (e.g., subclass of (property P55) → equivalent property (P41)→ rdfs:subClassOf). In this new data context, there is no formal distinction between entities that are classes from entities that are simply instances, as they are all considered "items." This leveling is reflected in the morphology of the notation, as class names are not capitalized as they are in traditional practices. As for the distinction between items and properties, a clue is provided in the identifier which is prefixed by a Q for items (classes and instances) and by a P for properties. Piscopo and Simperl (2018) point out a potential drawback of the blending of classes and instances. For example, homonyms referring to both a class and an instance would not be discernable and may be a source of erroneous results when queried.

Engineered for open and distributed data environments, LOD knowledgebases have long blurred the lines between the T-box (the terminological component pertaining to the conceptual model and the vocabulary) and the A-Box (the assertion component pertaining to the instances and their property statements). Schemas and vocabularies would, however, still be stored as separate graphs and follow different data management cycles. This is no longer the case in a Wiki context where boundaries have further dissolved any distinction between models and entity occurrences.

Our entire Linked Jazz dataset was recently imported into the Wikibase platform[19] where it can be accessed and consumed using built-in utilities. Data can be queried via a SPARQL endpoint with visualization capabilities in the form of different types of charts, graphs and maps[20] (Figure 4).

The Linked Jazz ontology is expanding to represent new areas of the domain. Adding new classes and properties has become a rather straightforward task thanks to the streamlined editorial utilities that the new platform affords. As a basic LD principle, any entity (classes, properties, or instances) needs to be resolved to a web page that both machines and humans can consume. The system provides an intuitive editor, familiar to Wikidata

Figure 4. Table view of SPARQL query results.

users, for publishing entities that can be created either manually via the editable interface or via batch loads. As a result, each entity has its own Wikibase web page that can be directly inspected and easily edited, making revisions (e.g., modifying entity labels) effortless. One of the immediate advantages of this function is that semantics and data can be contributed incrementally and collaboratively without requiring specific technical knowledge. Because data and their model are interwoven, subsequent re-modeling is also possible, allowing the knowledgebase to evolve and adjust to changing demands with little impact on the system. Fit-for-use, flexibility and extensibility take precedence in LD modeling in order to foster interoperability and shareability. As discussed earlier, the expressive power that comes from logical formality is lost in these new modes of conceptualizing domains, but it has also changed the function of ontologies and the means of exploiting them. For example, knowledge discovery and creation are not expected to come from logical inferences, but instead from querying sizable volumes of interconnected data. In Sanderson's words (2020), "We don't need perfect, certain data, we need to ask appropriate questions of large amounts of imperfect data."

As new collections and new types of primary sources are processed, new sets of data need to be channeled into the knowledgebase. Not having to aim for the perfect ontology and knowing that adjustments can be made as needed, help to expand our knowledge graph dynamically and in a bottom-up fashion. New segments of the domain and multiple facets of the jazz community have been and can continue to be added in a bottom-up fashion, as the need arises, in a sort of literary warrant-based approach. We perform mappings to Wikidata entities and properties whenever possible to support future federated queries with Wikidata and then feed our data, or relevant portions of them, back into the Wikidata ecosystem.

7.0 Semantic enrichment

This technical makeover has facilitated the expansion of the Linked Jazz knowledgebase through an organic and seamless manner. Seeded with data from new kinds of archival documents, Linked Jazz will soon incorporate new types of entities, including music venues and music groups. Person entities remain, however, key to the project. In the current knowledge graph, each node is a person and a potential entry point to other collections and to the global fabric of LD cultural memory. Each node is also at the intersection of an infinite number of stories, many lost or forgotten, that semantic technologies have the ability to uncover and weave into new narratives.

7.1 Women of jazz: Gender representation

It's axiomatic to recognize that the way we model individuals has a critical impact on the type of questions we could ask the data. The attributes we choose to use to describe people add new dimensions to the knowledge graph and ultimately foster new inquiries around fundamental elements of identity construction—from gender to sexuality, from race to class—each deepening our understanding of sociopolitical aspects of the community. The modeling capabilities enabled by LOD, especially through the new collaborative data environments discussed above, provide sufficient flexibility to shape or re-shape how people are represented, leaving room for reexamination and historic contextualization.

We began enriching the biographical descriptions of musicians with the gender attribute, in order to analyze the Linked Jazz network through the lens of gender (Pattuelli et al. 2017). Women jazz musicians are a largely underrepresented segment of the jazz community, almost a footnote in music history, despite their extraordinary achievements. The scarcity of women in our dataset was evident by even the most cursory scan of the name list. However, as we processed more oral histories where women were the interviewees, more women musicians started to emerge. For many, a corresponding URI was not immediately available as they did not have an authority record or even an entry in Wikipedia, so new URIs had to be minted. When working with data, what is missing can be as revealing as what is there. We saw the disparity in the number of mentions of women versus men musicians as a call for further investigation. We began to prepare the data to enable network analysis for gender ratio and distribution and also, more broadly, to support research on the historical role of women in jazz including questions about jazz women's influence, reputation and authority within the jazz community at large. Enriching entity descriptions, what in the social sciences is called "framing," helps shine a spotlight on specific facts or facets of interest and add new dimensionality to a knowledge graph. A first step in this direction was assigning gender values to all the musicians populating the dataset. A complex and problematic construct, yet essential to define person identities, gender was a key element to make women musicians visible in our graph. It was, however, not a straightforward task revealing the challenges of dealing with the scarcity or inconsistency of biographic data in authorities or linked data sources. At the time of the study, identifying and gathering gender data to assign to person entities required significant effort and expertise as only sparse and uneven data were available. In bibliographic authorities, for example, gender values were often missing as this remains an optional attribute.[21] Other LD hubs, including DBpedia[22] and MusicBrainz[23], served as data sources. Iterative cycles of data harvesting were needed, and multiple rounds of revision and version control had to be performed to maximize data acquisition and correct errors. The process resulted in the assignment of gender to 75% of

the target list of over 2,000 musicians (Hwang 2015). It is likely that today performing this task would be easier. An expansive knowledgebase like Wikidata offers much richer biographic data for people entities to draw on[24] and would significantly mitigate the shortcomings we faced early on when attempting to represent people. While the sources we relied on earlier were limited to binary values ("male" and "female") with the only option of "unknown" to address missing or uncertain information, Wikidata offers multiple options to account for gender variance (https://www.wikidata.org/wiki/Property:P21). It also includes temporal qualifiers recognizing that gender can be a time-dependent value and to contextualize transitions. Even a little semantic enrichment, such as the addition of the gender property, has the capacity to open up new ways of examining the data. The network visualization was augmented with a faceted view where gender distribution can be surveyed at the macro level[25] (Figure 5).

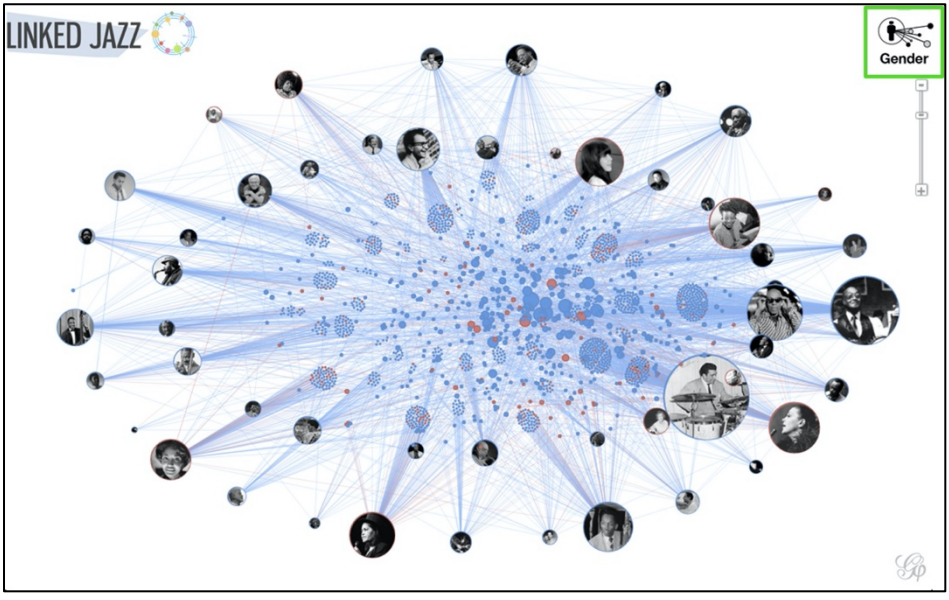


Figure 5. Gender view of the Linked Jazz graph (image by Karen Hwang).

A view of gendered relationships can be displayed at the node level. Each mention statement is now connoted with gender, so both its subject (the narrator) and object (the person mentioned) have a gender attribute. This makes it possible to do quantitative analysis of mentions by gender as well as determine the gender frequency and distribution. To conduct social network analysis a Gephi file is also available and downloadable.

7.2 Local 946: List to graph

In an effort to broaden and enrich the Linked Jazz dataset, we pursued specific archival documents and collections. Admittedly, we have not been agnostic in the choice of primary sources we have sought to process. We prioritize source documents that would help represent different perspectives, contextualize and even reframe Linked Jazz entities. Particular

attention has been paid to unearthing and documenting the lives of persons and communities traditionally underrepresented, with the aim to make them part of the global network of interrelated knowledge.

Such was the case with the Local 496 Membership Directory[26]. This document, created in the 1940s and held at the Tulane University's Hogan Jazz Archive, is the official roster of the African American chapter of the then segregated New Orleans jazz musicians' union, Local 946. The chapter, which started in 1926, later merged with the white union chapter in 1969. A unique historical document, the Local 496 directory lists 418 African American musicians working in New Orleans, including their names, residential addresses and instruments played. We converted what was essentially a static list into a knowledge graph using OpenRefine to tabulate, normalize, model, and create triples. A basic, yet powerful KO system, a list is "a means for cultural order," in Eco's words, that can also act to control and exclude (Eco 2009). Tapping into its power as an inventory, we mapped out the musician names to authority files and knowledgebases to harvest URIs. Twenty-five URIs (approximately 6% of the list) could not be found. The vacancy of data—identifiers in this instance—is a revealing indicator of the level of notability of the musicians, as discussed earlier. While they were all professional musicians, most had never reached the degree of recognition that would grant them entry into Wikipedia, let alone in LCNAF. They are, however, an integral part of the New Orleans jazz scene and their influence, cultural impact and legacy should not be ignored or forgotten. This time, missing identifiers were created directly in Wikidata where all the Local 496 directory data were entered. The Local 496 Membership Directory, the source document, was also added to Wikidata[27] to serve as the verifiable reference source and context for the data supplied (e.g., Sweet Emma Barrett (Q7655300) → member of (P64)→ American Federation of Musicians. Local 496 New Orleans, La. [Q66949304]). Using Wikidata as the public platform was motivated by the desire to engage the New Orleans community of archivists, librarians and jazz experts in creating entries and filling in the gaps with verifiable data from documents they had access to, with the ultimate goal being to build a rich and valuable research resource. Uncovering a new array of musicians previously tucked away in local archival records offered unprecedented opportunity to revisit lesser-known or overlooked segments of jazz history. As more biographical attributes are added, multiple views and access points will be available that would lend new meaning to the graph and assist with new lines of historical research. Exposing these data on a public platform such as Wikidata also helped to connect special collections with a worldwide archive of cultural memory. While deeper insights are expected to be gained once the Local 946 data become part of the broader context of the Linked Jazz network, the Local 496 graph is already queryable via the Wikidata service. Another important product to come out of the project was the geolocation of musicians based on their residencies. Address data from the document were contributed to GeoNames[28], a major linked dataset of geographic data. Plotted against different historical city maps, including the New Orleans historical districts and the predatory loan districts of the New Deal's Home Owners' Loan Corporation (HOLC)[29], the data offer researchers a view into the neighborhoods and thereby the cultural and socioeconomic lives of New Orleans jazz musicians (Figure 6).

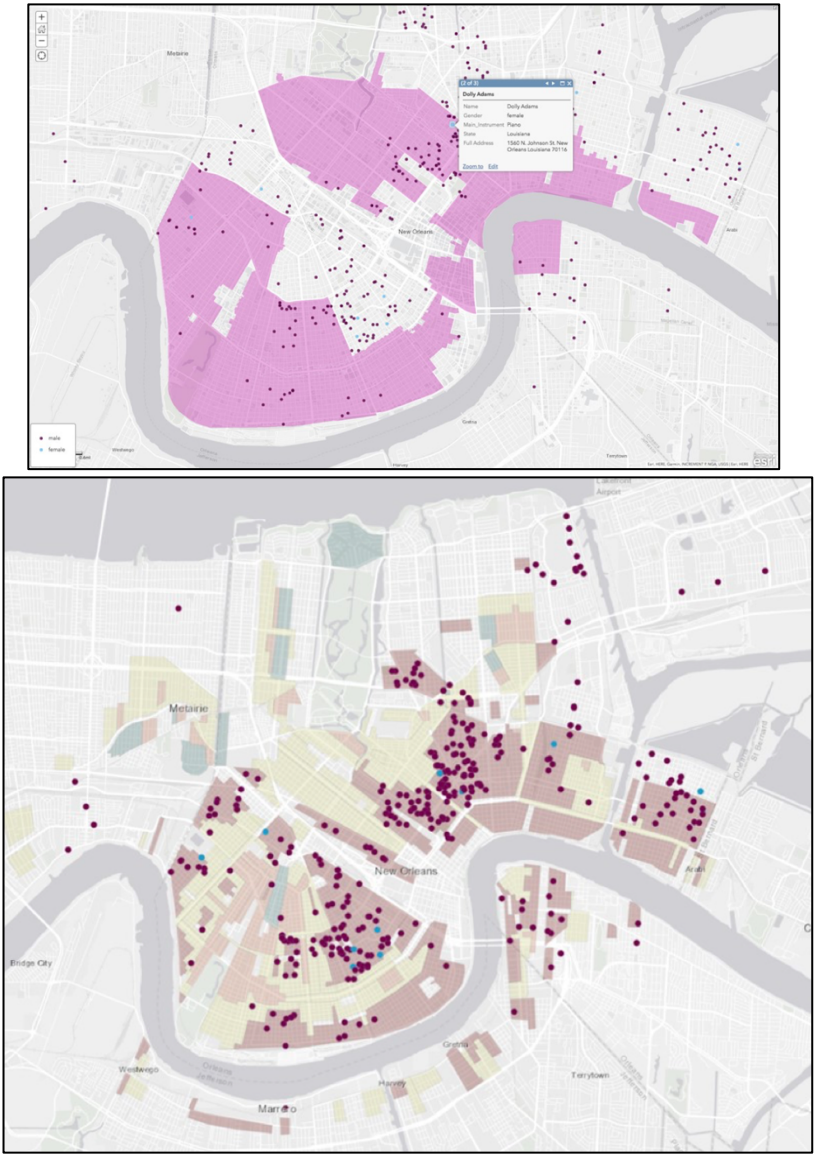


Figure 6. Map visualization of New Orleans historical districts (top) and HOLC's predatory loan districts (bottom)(images by Genvieve Milliken).

8.0 Data Integration

The full potential of LD is reached when heterogeneous data from different sources are interlinked providing unified access to data and the possibility to seamlessly query multiple graphs. To this end, the Wikibase platform serves as a common interoperability layer for the integration of diverse knowledge graphs and for a new generation of methods for data

enrichment and contextualization. Unified queries against the datasets can be executed via the SPARQL endpoint[30]. While we have just begun to populate Wikibase with data from different segments of the project, unanticipated entity overlapping has already emerged when querying across datasets that reveal data joins to be explored. A future step will be to conduct federated queries against Wikidata. This will allow us to tap in one of the most extensive bodies of public networked data available enabling us to begin asking new and more complex questions.

Data integration with Linked Jazz is also happening beyond the borders of our project. The Australian linked data project JazzCats (Jazz Collection of Aggregated Triples)[31] aggregates collections of RDF triples to trace performance history. The project combines discography and granular performance data (e.g., solos including pitch, key, and chord) with interpersonal relationship data derived from Linked Jazz to "bridge previously unconnected but complementary information about jazz music" (Bangert et al. 2017). The project JazzTube[32], a joint initiative between the Hochschule für Musik Franz Liszt Weimar (University of Music Franz Liszt Weimar) and the International Audio Laboratories Erlangen, combines annotations of jazz solos with discographies. It has incorporated data from the Linked Jazz dataset to represent the network of interpersonal relations between musicians who perform a solo[33]. These are just a few of the examples that show how the aggregation of diverse knowledge graphs can create semantic bridges across previously unrelated information spaces. Linked data integration is key to dismantling cultural data silos and opening up virtually infinite streams of connections and paving new paths of discovery and interpretation. As more linked datasets become available in the music and related domains, new opportunities arise for combining and re-contextualize data without the need for central agreement or coordination or, as has been noted (Walk 2007): "The coolest use of your data will be thought of by somebody else." We have only begun to envision the new research questions, methods of analysis, and creative scholarship that are possible when we are able to provide integrated access to cultural heritage data.

9.0 Conclusion

Linked Jazz has provided a fertile environment where to explore and experiment with making archival content semantically "understandable" and processable by machines and interconnected within and across knowledgebases. An enduring project, Linked Jazz has progressed along the evolution of the LOD initiative in the spirit of learning by doing and sharing challenges and lessons learned. In the initial phase of the project, most of our effort had to be devoted to pioneering methods and prototyping tools. Always core to the project was the goal to link people through data. More specific to our context, we leveraged LD technologies and the power of knowledge graphs to represent the community of jazz musicians whose personal and professional relationships are dense and intertwined. Using the content of archival resources, rather than just their description, as the main source of semantics has exposed rich veins of meaning yet to be mined in primary sources. A combination of approaches, from automated text analysis to human annotation, has proven to be the most effective way for us to generate linked data from text. A good deal of crafting and data preparation is involved in the production of linked data. Principles and systems of KO are at the core of the creation and management of a linked knowledgebase—from data modeling to data reconciliation. We learned through firsthand experience how the nature and functions of KO and representation systems, such as name authorities and ontologies,

are changing in the LD ecosystem. Through novel forms of hybridization with computer science methods, they are reshaping and expanding to serve as catalyst of interconnectivity.

Linked Jazz has continued to grow and is now devising a novel data infrastructure that has opened up an array of opportunities for making historical content part of the web. More than just a technical makeover, the more versatile toolkit and collaborative platform we use today have allowed us to reframe the aim and scope of the project, overcoming boundaries and expanding it in new directions. The availability of large amounts of open, distributed, and structured semantics on the web is opening up unanticipated opportunities for innovative research and narration. Its potential for digital archival practices, historical research and models of historiography is still largely untapped. Using LOD technologies as a tool for critical engagement and even storytelling, we intend to continue to graph out communities, represent different historical contexts and viewpoints to enable new insights to be drawn and ultimately foster new inquiries.

Notes

1. Linked Jazz at https://linkedjazz.org/
2. SNAC at https://snaccooperative.org/
3. CultureSampo at https://seco.cs.aalto.fi/applications/kulttuurisampo/
4. List of oral history transcripts at https://linkedjazz.org/?page_id=899
5. Transcript Analyzer at https://github.com/linkedjazz/linked-jazz-prototype-transcript
6. Wikidata at https://www.wikidata.org
7. Name Mapping and Curator Tool at https://linkedjazz.org/tools/name-mapping-tool-and-curator
8. For more information on Ecco! see https://linkedjazz.org/?page_id=719
9. Relationship Vocabulary, Friend of Friend (FOAF), and the Music Ontology.
10. Semantic Lab. (2020, February 25). Linked Jazz Applied Ontology (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.3687248
11. Semantic Lab. (2020, January 16). SemanticLab/linked-jazz-datasets: DOI Added (Version 1.1). Zenodo. http://doi.org/10.5281/zenodo.3609362
12. The current SPARQL endpoint is available at https://query.semlab.io
13. Linked Jazz visualization tool at https://linkedjazz.org/network/
14. A working prototype of the NER module is currently available in a sandbox environment at semlab.io/DADAlytics-ner-demo
15. Wikibase at http://wikiba.se
16. Project's Wikibase instance at http://base.semlab.io/
17. Rhizome at https://www.newmuseum.org/pages/view/rhizome
18. OCLC at https://www.oclc.org
19. The Linked Jazz dataset is queryable at https://tinyurl.com/tnourho
20. Outcome of a SPARQL query for musician relationships at https://tinyurl.com/slny84p
21. Gender designation has been recently a topic of discussion in the library community (Billey., Drabinski, and Roberto [2014]).
22. DBPedia at https://wiki.dbpedia.org
23. MusicBrainz at https://musicbrainz.org/
24. Wikidata: List of properties/personal life https://www.wikidata.org/wiki/Wikidata:List_of_properties/personal_life
25. Linked Jazz network visualization, gender view at https://linkedjazz.org/network/?mode=gender
26. Tulane University Libraries' record of the Local 496 Membership Directory at http://voyager.tcs.tulane.edu/vwebv/holdingsInfo?searchId=531&recCount=10&recPointer=1&bibId=966417

27. Local 496 Membership Directory, American Federation of Musicians (Q66948585) at https://www.wikidata.org/wiki/Q66948585
28. GeoNames at https://www.geonames.org
29. Mapping Inequality: Redlining in New Deal America Project at https://dsl.richmond.edu/panorama/redlining/#loc=11/40.794/-74.105
30. SPARQL query endpoint at https://query.semlab.io/
31. JazzCats at http://jazzcats.oerc.ox.ac.uk/
32. JazzTube at http://mir.audiolabs.uni-erlangen.de/jazztube/about
33. Example of artist relationships using Soloist: John Coltrane at http://mir.audiolabs.uni-erlangen.de/jazztube/soloists

References

Bangert, D, T Nurmikko-Fuller, J. S. Downie, and Y Hao. 2018. "JazzCats: Navigating an RDF Triplestore of Integrated Performance Metadata." In *DLfM 18: Proceedings of the 5th International Conference on Digital Libraries for Musicology*. New York: ACM, 74-77.

Billey, Amber, Emily Drabinski, and K. R. Roberto. 2014. "What's Gender Got to Do with It? A Critique of RDA 9.7: Cataloging & Classification Quarterly: Vol 52, No 4." *Cataloging & Classification Quarterly* 52, no. 4: 412-21.

Eco, Umberto. 2009. *The Infinity of Lists*. London: MacLehose.

Fauconnier, Sandra. 2018. "Many Faces of Wikibase: Rhizome's Archive of Born-Digital Art and Digital Preservation." Wikimedia Foundation (blog). September 6, 2018. https://wikimediafoundation.org/news/2018/09/06/rhizome-wikibase/

Gallon, K. 2016. "Making a Case for the Black Digital Humanities." *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold, Lauren F. Kleon Minneapolis: University of Minnesota Press, 42-49.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, et al. 2019. "Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage." OCLC. https://www.oclc.org/research/publications/2019/oclcresearch-creating-library-linked-data-with-wikibase-project-passage.html

Hwang, Karen Li-Lun. 2015. "Enriching the Linked Jazz Name List with Gender Information." Linked Jazz (blog). August 17, 2015. https://linkedjazz.org/?p=1007

Kerschbaumer, Florian, Linda von Keyserlingk-Rehbein, Martin Stark and Marten Düring, eds. 2020. *The Power of Networks: Prospects of Historical Network Research*. New York: Routlege.

Lévesque, François, Marielle St-Germain, Dominique Piché, Jean-François Gauvin, Michel Gagnon, Thomas Hurtut. 2020. "MusX: Online Exploring and Visualizing Graph-Based Musical Adaptations." *2020 IEEE 36th International Conference on Data Engineering*. IEEE, 1734-37. doi:10.31219/osf.io/jvtr7

Miller, M. 2018. "Wikibase for Research Infrastructure: Part 1." Medium (blog). March 19, 2018 https://medium.com/@thisismattmiller/wikibase-for-research-infrastructure-part-1-d3f640dfad34

Miller, Matthew, Jeff Walloch, and M. Cristina Pattuelli. 2012. "Visualizing Linked Jazz: A Web-Based Tool for Social Network Analysis and Exploration." Poster presentation at *American Society for Information Science and Technology (ASIS&T) Annual Meeting*. Baltimore, MD.

Morrissey, Robert Michael. 2015. "Archives of Connection." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 48, no. 2: 67-79. https://doi.org/10.1080/01615440.2014.962208

Pattuelli, M. Cristina, Matt Miller, and Sarah Ann Adams. 2019. "Ontology Shift: Cultural Heritage Ontologies in the Time of Linked Open Data." Presented at the Digital Humanities 2019 (DH2019), Ontologies for Linked Data in the Humanities Workshop, Utrecht, The Netherlands. https://cwrc.ca/islandora/object/islandora%3A34f1ac71-c799-4f24-a9a1-0732e3920e52

Pattuelli, M. Cristina, Karen Hwang and Matt Miller. 2017. "Accidental Discovery, Intentional Inquiry: Leveraging Linked Data to Uncover the Women of Jazz." *DSH: Digital Scholarship in the Humanities* 32: 918-24. https://doi.org/10.1093/llc/fqw047

Pattuelli, M. Cristina, Alexandra Provo, and Hilary Thorson. 2015. "Ontology Building for Linked Open Data: A Pragmatic Perspective." *Journal of Library Metadata* 15: 265-94. https://doi.org/10.1080/19386389.2015.1099979

Piscopo, Alessandro and Elena Simperl. 2018. "Who Models the World? Collaborative Ontology Creation and User Roles in Wikidata." In *Proceedings of the ACM on Human-Computer Interaction, 2 (CSCW)*. New York: ACM, 1–18. https://doi.org/10.1145/3274410

Sanderson, Rob. 2020. "It's 2020 … Where Is My Flying Car and Cultural Heritage Research Data Ecosystem?" Presented at the Coalition for Networked Information (CNI) Spring 2020 Membership Meeting, March 30. https://www.cni.org/wp-content/uploads/2020/03/CNI_Sanderson_keynote.pdf

Tamper, Minna, Petri Leskinen, Kasper Apajalahti and Eero Hyvönen. 2018. "Using Biographical Texts as Linked Data for Prosopographical Research and Applications." In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection: 7th International Conference, EuroMed 2018, Nicosia, Cyprus, October 29–November 3, 2018, Proceedings, Part I.* Cham: Springer, 125-37.

Walk, Paul. 2007. "Paul Walk's Web: The Coolest Thing to Do with Your Data Will Be Thought of by Someone Else." July 23, 2007. https://www.paulwalk.net/post/2007/07-23-the-coolest-thing-to-do-with-your-data-will-be-thought-of-by-someone-else/

Richard P. Smiraglia
Institute for Knowledge Organization and Structure, Inc.

James Bradford Young
Institute for Knowledge Organization and Structure, Inc.

Marnix van Berchum
Utrecht University

Chapter 9
Digging into the Mensural Music Knowledge Graph
Renaissance Polyphony meets Linked Open Data[***]

Abstract
The Semantic Web is created by a dense network of individual linkages. One challenge for a SW is to bring together online resources and metadata representing them that heretofore have been the remand of physical repositories. An excellent example is polyphonic music, the sources of which are manuscripts and early printed collections. To discover the details required to create linked data in a curated humanities environment we turned to the Computerized Mensural Music Editing project. Appropriate linkage to the cloud was effected via linking the composers and works to the Virtual International Authority File. The major contribution of our project team has been the creation of several hundred VIAF authority records for works in the CMME repository added via the Library of Congress' Name-Authority Cooperative project. For contextual enrichment we added terms for form and genre and medium of performance. We have produced a "Mensural Music Knowledge Graph" based on the content for which we have provided linkable or linked data, including explicit links to the Universal Decimal Classification. We were fortunate to be able to utilize the bibliographic community's intellectual structure for the control of musical works. The problem for LD is to move the complex systems created manually for successful clustering and disambiguation into the LOD cloud through the use of SW technologies. We have demonstrated the vast potential of the LOD Cloud to contribute to scholarship in musicology, and by extension, in other artifact-rich humanistic endeavors.

1.0 Linked Open Data, the humanities, and musicology

One of the premier advances in information technology in the twenty-first century is the so-called "semantic web," often given in upper case as though it were a formal institution and hailed by the computer science community as the forefront of knowledge dissemination. This web of semantic meaning (hereafter SW for semantic web) is created by a dense network of individual linkages, especially among names of places, persons, concepts and institutions already represented in the World Wide Web. The purpose of the linkages is both to exploit knowledge stored and to overcome the limitations of undiscovered public knowledge. One challenge for a SW is to bring together online resources and metadata representing them that heretofore have been the remand of physical repositories. An excellent example is polyphonic music, the sources of which are manuscripts and early printed collections. Some of the evidentiary artifact base consists of partial representations, e.g., one or two part-books from an original set of four or five.

[***] The authors wish to gratefully acknowledge the assistance of Ronald Siebes in producing images of the Mensural Music Knowledge Graph.

As source inventory goes, the musicological community has a long head start, ranging from nineteenth-century projects such as François Joseph Fétis' *Biographie universelle des musiciens* (Bruxelles: Meline, Cans & Comp, 1837) and Robert Eitner's *Biographisch-bibliographisches Quellen-lexikon der Musiker und Musikgelehrten der christlichen Zeitrechnung bis zur Mitte des neunzehnten Jahrhunderts* (Leipzig: Breitkopf & Härtel, 1900-04) to mid-to-late twentieth century projects such as the *Repertoire International des Sources Musicales (*RISM), it's sister projects *RILM* (Répertoire International de Littérature Musicale), *RIdIm* (Répertoire International d'Iconographie Musicale) and *RIPM* (*Répertoire international de la presse musicale*), and of course the famous *Census-Catalogue of Manuscript Sources of Polyphonic Music, 1400 1550* (Neuhausen-Stuttgart: American Institute of Musciology; Hänssler-Verlag, 1979-1988), compiled by the University of Illinois Musicological Archives for Renaissance Manuscript Studies. In addition, a long tradition of precise cataloging and especially authority control of music has created an exemplary network of data concerning musical composers and musical works, especially those now represented in the SW by VIAF (the Virtual International Authority File), which contains among others the authority records from the Library of Congress, the Bibliothéque nationale de France and the Deutsche Nationalbibliothek—among the world's most important repositories of musical sources. The challenge for the SW is to bring these excellent sources together with precise linkages and to enrich them with metadata that can help expand and support ongoing musicological scholarship.

2.0 Digging into the Knowledge Graph

In 2016 we were awarded a grant under the fourth Digging Into Data challenge program, sponsored by the Trans-Atlantic Platform (T-AP). Our project was titled Digging Into the Knowledge Graph[1]. International and interdisciplinary our project team included: 1) the University of Wisconsin, Milwaukee (UWM) from the United States, a consultant from Sao Paulo State University-Marilia (UNESP) in Brazil; these participants were classical knowledge organization scholars from iSchools; 2) The Data Archiving and Networked Services division (DANS), Royal Netherland Academy of the Arts and Sciences (national data repository) and Vrije Universiteit (VU), Amsterdam and LOD Laundromat (http://lodlaundromat.org/); these participants were data scholars and leaders in the linked open data (LOD) community; and, 3) University of Alberta (UA) in Canada; participants from UA were interdisciplinarians working with classification in the social sciences, particularly political science and economics. There was musicological expertise on both the US and Dutch teams.

The goal of our project (hereafter Di4kg) was to use LOD and Semantic Web technologies properly to produce contextual enrichment at the level of single artifacts, in effect creating an environment in which humanities and social sciences research expressed as LOD might become self-organizing in the LOD cloud. We attempted this not only by linking data in our use cases to existing cloud vocabularies (such as Wikidata) but also by converting two major classifications into LOD:

the global bibliographic Universal Decimal Classification (UDC)(http://www.udcc.org/)

and the evolving phenomenon-based Basic Concepts Classification (BCC)

(https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013)

Also by discovering semantic web ontologies in the cloud that could be reused (see Renwick and Szostak 2021; Smiraglia and Szostak 2021; and Slavic, Siebe and Scharnhorst 2021). "Contextual enrichment" means that we committed to linking curated data to curated ontologies and to reusing existing vocabularies. The five oft-stated linked data principles[2], suggest adding value in the process of creating linkages—the value we are adding is knowledge organization by explicit linkages to appropriate knowledge organization systems. Our two use cases are in musicology and economics. Our musicological case is the subject of this report.

3.0 CMME, our humanities use case

To discover the details required to create LOD in a curated humanities environment we turned to the Computerized Mensural Music Editing (CMME) project (http://cmme.org/), based at Utrecht University, a project initiated in 1999 by Theodor Dumitrescu that was released to the public in 2005-2006 (Dumitrescu and van Berchum 2009). In its own words CMME is:

> a scholarly Initiative to offer free online access to new, high-quality early music scores produced by today's leading experts. Based at Utrecht University in the Netherlands, the project represents a collaborative development effort of specialists in musicology, information science, and music retrieval. The major purpose of the enterprise is to produce and maintain an online corpus of electronic editions, in addition to software tools making them accessible to students, scholars, performers, and interested amateurs. Here, the brilliant polyphonic styles known to the modern world through the works of such masters as Dufay, Josquin, Machaut, Palestrina, and Tallis can come to life again in the central medium of the 21st century.

Our project team selected the CMME project partially because of overlap between our Dutch colleagues and the CMME editorial staff, and partially because DANS housed an archived version of the project (referred to below as a "dump") that would provide a stable starting point for the creation of LOD. Furthermore, the site is richly curated—a requirement of our grant to establish "best practices" for creating LOD in musicology, it has a simple yet elegant structure that is rich with specific, curated metadata, including:

> Entities: composers, works, sources and editorial projects; and,

a fair bit of rich non-curated metadata representing:

> Concepts: form, medium, notation, text and liturgical function.

The CMME database has specific pages for editorial projects, composers, sources, and compositions. Displays in each include hyperlinks to the others. For example, we can click under "editorial projects" on the link for "The Occo Codex ed. Jaap van Benthem, Marnix van Berchum, Anna Dieleman, Theodor Dumitrescu, and Frans Wiering" to arrive at a summary of this project including compositions, sources, an introduction, and edited compositions. We can link to the composition *Ave regina celorum a4 I* (CMME.org/database/composers/19) by Gaspar van Weerbeke (CMME.orgdatabase/composers/19). Clicking on the composer name will take us to a list of works, clicking on a hyperlinked source will take us to source data; the first source in this case is "(Milan, Archivio della Veneranda Fabbrica del Duomo, Sezione Musicale, Librone 1 (olim 2269)" (CMME.org/database/sources/138). In this example incipits are available at each node (Figure 1).

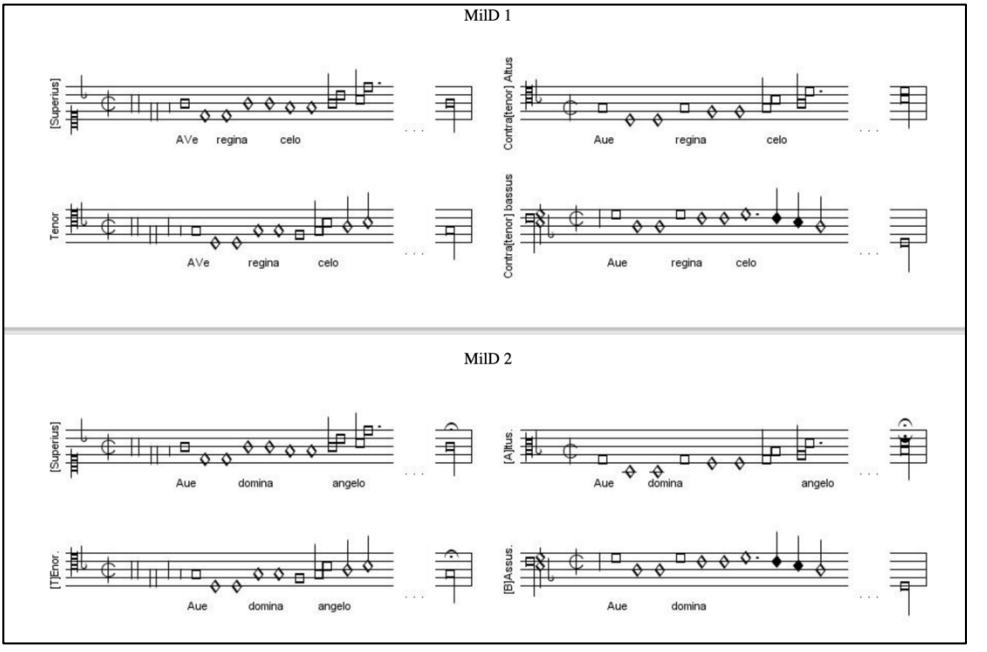


Figure 1. CMME incipits for Weerbeke *Ave regina celorum a4*.

Content analysis of the SQL "dump" from which our work began told us the internal structure of the database included 73,305 individual statements of which about half was comprised of programming statements and about half was rich mensural music metadata (all content analysis in this project was conducted using the Provalis ProSuite (https://provalisresearch.com/). In the dump there were 221 composers represented ranging from Agricola to Nicolaus Zoilo (plus Anonymous), and there were 3,671 individually identifiable musical works. It is important to bear in mind that the dump represented CMME at a specific point in the past and thus does not completely conform to the visible online content of the database.

At the beginning of the project we determined the most appropriate linkage to the cloud would be to link the composers, works and sources to VIAF (Virtual International Authority File http://viaf.org/); we will describe this phase of the project in detail below—initially we were able to match approximately 50% of the composers and about 25% of the works with existing VIAF authority records. Only a few of the manuscript sources were established in VIAF.

Our content analysis also gave us an opportunity to look at the terminological content of the database. The 3,672 titles (or incipits) were analyzed using the WordStat module of the Provalis Prosuite. This software (among other things) sorts individual words and phrases by frequency of occurrence and allows co-occurrence analysis as well. From our simple analysis we learned there were 14,512 words in the titles of which 4,784 were unique (occurred only once) and 161 occurred 10 or more times. These include individual words such as "alleluia" or "missa." Phrases of two to five words were analyzed (to better visualize terms). There were 17,511 phrases of which 62 occurred 5 or more times. Table

1 below shows the upper segment (most frequently occurring) of both keywords and phrases.

Keyword	Frequency	Phrase	Frequency
missa	368	missa de	36
de	185	je suis	18
est	165	je ne	15
et	147	mon cueur	14
domine	129	que je	14
je	126	illo tempore	14
la	120	jesu christe	12
en	96	de mon	11
qui	96	dei genitrix	11
le	87	je vous	11
que	80	salve regina	10
ave	79	te domine	10
vous	73	maria virgo	10
deus	72	ave maria	10
mon	69	de vous	10
virgo	69	mon amy	10
te	68	domine deus	10
si	67	sancta maria	9
maria	61	missa la	9
dominus	60	ego sum	9
il	60	domine jesu	8
ne	60	vray dieu	8
alleluia	55		

Table 1. Frequency distribution of keywords in CMME dump.

Of course, the purpose of this analysis was simply to gain an empirical understanding of what we already could see in the CMME website. That is, we could see a rich vocabulary of unique terms, but also terms pointing to a standard vocabulary of musical forms, most of which are or include liturgical terms. More interestingly, using a keyword in context (KWIC) function we gathered clusters of phrases including the keywords “alleluia,” “missa,” “kyrie,” “magnificat,” “gloria” and “credo”; a few of these are shown in Table 2.

	keyword	
Missa	Alleluia	
	Alleluia	Confitemini domino
	Alleluia	Beatus vir sanctus martinus
Tristitia vestra	alleluia	
	Credo	L'amour du moy
	Credo	de Villagiis
	Magnificat	[Tertii toni]
	Magnificat	Regali ex progenie
	Missa	Petrus apostolus
	Missa	La belle se siet
	Missa	Gracieuse plaisant
	Missa	Ludovicus dux wirtenbergensis
	Kyrie	Fons bonitatis - Gloria
Missa Summum	kyrie	
	Kyrie	O rex clemens
Kyrie Fons bonitatis -	Gloria	
	Gloria	tibi trinitas

Table 2. A selection of keywords in context in titles and incipits.

Our purpose in running this type of analysis was to discover the content that we wanted to represent with LOD linkages to controlled vocabularies, specifically in this case the Library of Congress Genre/Form Terms list. But we also were able to see here the presence of rich vocabulary such as that used in digital humanities research to track the evolution of perhaps potential sub-genres (Signer 2019) such as Missa “alle regretz” or Alleluia “Stabant iusti.”

In sum, the data model we developed is shown in Figure 2.

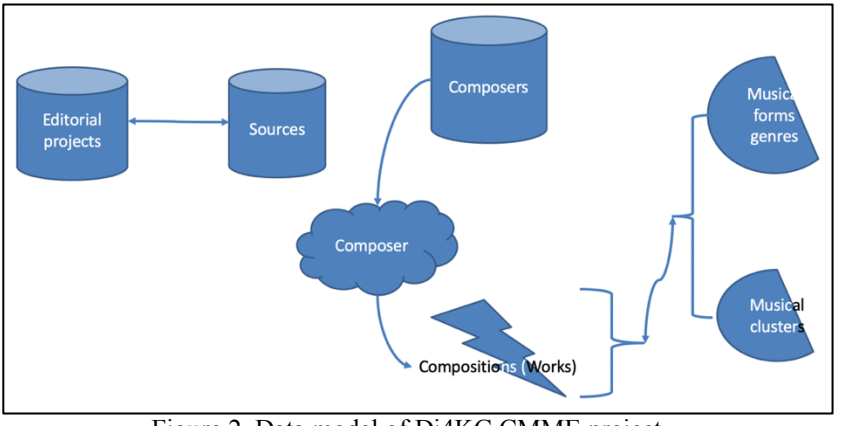


Figure 2. Data model of Di4KG CMME project.

As the model indicates, our work was to focus on making explicit linkages between data records representing composers and their musical works as represented in CMME, and available formal LOD knowledge organization systems. We would add value by making two classifications—UDC and BCC—into LOD and linking to them as well.

4.0 Musical works in VIAF for LOD

The major contribution of our project team has been the creation of several hundred authority records for works in the CMME repository. These records have been added to the Virtual International Authority File (VIAF) via the Library of Congress' Name-Authority Cooperative (NACO) project.

An authority record is a data record in a bibliographic information system that records the authorized form of a properly formulated access point, any variant forms that might be encountered in the system through bibliographic representations, and citations to sources used to create the authorized form. Figure 3 shows the text portion of an authority record for the composer Johann Buchmayer.

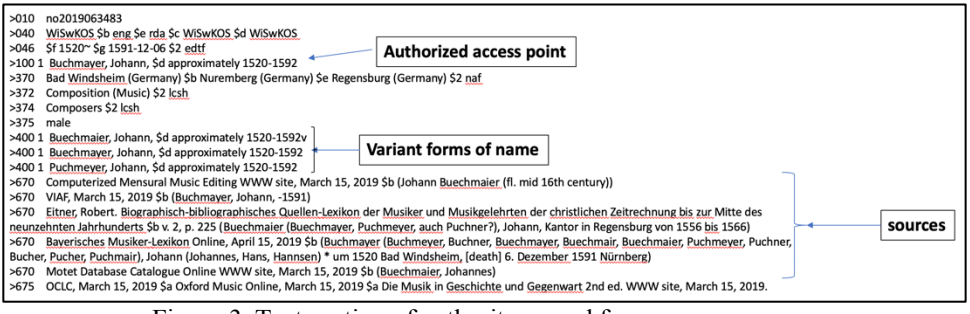


Figure 3. Text portion of authority record for composer name.

This visualization was created from the OCLC version of a record submitted by our project. The record is formatted according to the conventions of USMARC (US Machine Readable Cataloging), a standard format for bibliographic data. The authorized access point consists of the form of name found most reliably in the composer's works, to which is added identifying information concerning dates of birth and death. Variant forms that have occurred in reference sources or in publications are recorded in the source data field and then formulated to serve as "use for" references. An authority record for a work records the combination of composer name and preferred title for the work, together with any variant forms. For example, the work by Buchmayer titled *Resurrexi et adhuc* has this form:

Buchmayer, Johann, approximately 1520-1592. Resurrexi et adhuc.

in which the title portion must accompany the authorized access point for the composer.

In October 2018, The private non-profit research institute IKOS (Institute for Knowledge Organization and Structure, Inc.) became a partner in the Di4kg consortium. IKOS negotiated with the Collection Services Division of the Concordia University Libraries, Montréal and with the LC/NACO project to allow our institute to create LC/NACO authority records for the content of the CMME dump. This was an historic first, as never before had a corporate cultural heritage entity that was not a cataloging library been allowed to participate in LC/NACO, let alone to enter authority records directly. Since 2018

our project has added over 1200 records to VIAF via LC/NACO. Once entered into VIAF the records are converted into RDF for SW operability, which will be demonstrated in section 6.0.

5.0 Linking to form, genre and classification

For the purpose of contextual enrichment (the fifth open data "star") we added to every authority record whenever possible terms for form and genre and medium of performance. For these we linked to Library of Congress LOD thesauri LCGFT (Thesaurus for Form and Genre Terms) and LCMPT (Thesaurus for Medium of Performance Terms), LCGFT is maintained in consultation with the Music Library Association (Library of Congress 2020):

> The Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT) is a thesaurus that describes what a work is versus what it is about …. The thesaurus combines both genres and forms. Form is defined as a characteristic of works with a particular format and/or purpose …. Genre refers to categories of works that are characterized by similar plots, themes, settings, situations, and characters. Examples of genres are westerns and thrillers. In the term Horror films "horror" is the genre and "films" is the form.

LCMPT, like LCGFT is maintained in consultation with the Music Library Association (http://id.loc.gov/authorities/performanceMediums.html):

> The Library of Congress Medium of Performance Thesaurus (LCMPT) is a stand-alone vocabulary that provides terminology to describe the instruments, voices, etc., used in the performance of musical works …. Authorized terms and references in LCMPT generally consist of single words and phrases, but parenthetical qualifiers are occasionally employed to differentiate among homonyms. All terms and references are in the singular form ... (e.g., flute; saxophone ensemble; but Irish harp). The thesaurus has a few broadest terms as listed in the "Top Scheme Members" section. Each of the other terms is hierarchically subordinate to one or more of these terms and exhibits the class/class member relationship. Most of the authorized terms have Used For (UF) references for synonyms. Scope notes are also provided in many cases, and may describe the medium's physical structure, the time period in which it was popular, and/or its geographic origin.

In some complex cases, or in cases where LCGFT does not have an appropriate term, we also added terms from the *Library of Congress Subject Headings* (*LCSH*), which is maintained in part as LOD (https://id.loc.gov/authorities/subjects.html):

> Library of Congress Subject Headings (LCSH) has been actively maintained since 1898 to catalog materials held at the Library of Congress. By virtue of cooperative cataloging other libraries around the United States also use LCSH to provide subject access to their collections. In addition LCSH is used internationally, often in translation. LCSH in this service includes all Library of Congress Subject Headings, free-floating subdivisions (topical and form), Genre/Form headings, Children's (AC) headings, and validation strings* for which authority records have been created. The content includes a few name headings (personal and corporate), such as William Shakespeare, Jesus Christ, and Harvard University, and geographic headings that are added to LCSH as they are needed to establish subdivisions, provide a pattern for subdivision practice, or provide reference structure for other terms. This content is expanded beyond the print issue of LCSH (the "red books") with inclusion of validation strings. *Validation strings: Some authority records are for headings that have been built by adding subdivisions. These records are the result of an ongoing project to programmatically create authority records for valid subject strings from subject heading strings found in bibliographic records. The authority records for these subject strings were created so the entire string could be machine-validated. The strings do not have broader, narrower, or related terms.

Figure 4 shows an example of one such case, the *Magnificat* by Jacquet of Mantua.

>010 no2020067385
>040 WiSwKOS $b eng $e rda $c WiSwKOS
>100 1 Jacquet, $c of Mantua, $d 1483-1559. $t Magnificats, $m singers (4)
>380 Part songs $a Sacred music $2 lcgft
>380 Magnificat (Music) $2 lcsh
>382 0 singer $n 4 $s 4 $2 lcmpt
>670 Computerized Mensural Music Editing WWW site, June 15, 2020 $b (Magnificat BolC R142 18v-19r)
>670 Grove Music Online, June 15, 2020 $b (several Magnificat settings in I-CMac VIII, I-Mc S Barbara 49)
>670 Digital Image Archive of Medieval Music WWW site, June 15, 2020 $b (Magnificat: Et exultavit spiritus meus Mantua, Jacquet de (?) or Pastrana, Pedro de (1490-ca. 1558) (?) Appears on I-Bc R.142 18v-19v Number of Voices: 4 Concordances 1554/17 E-TZ 4 [i.e. 5?])
>675 OCLC, June 15, 2020 $a VIAF, June 15, 2020 $a Die Musik in Geschichte und Gegenwart 2nd ed. WWW site, June 15, 2020.

Figure 4. Authority record including term from *LCSH*.

In addition to making explicit links to form and genre and medium of performance terms, we wanted also to demonstrate the power of classification in the SW. For this reason our project created explicit links to the UDC. (Because the BCC is undergoing revision in its music schedules, we have not yet attempted explicit linkage to BCC; see Smiraglia and Szostak 2020). Where it was possible to make links to the Library of Congress thesauri via the authority records, we had to make manual links to the UDC via in the process of creating our own mensural music knowledge graph. The linkages were made simply enough to composers and musical works as expressed in the newly created authority records. For example, the Buchmayer *Resurrexi et adhuc* (Figure 3) are linked as follows to the UDC:

Resurrexi et adhuc	no2019064077	‡a Motets ‡a Sacred music ‡2 lcgft	‡a singer ‡n 4 ‡s 4 ‡2 lcmpt	783.4.64

The UDC string represents a point in the following hierarchy:

78 Music

783 Church Music. Sacred Music

783.4 Anthems. Motets and Chants not Mentioned Elsewhere, *e.g.,* in the Manner of Palestrina

783.4.64 Vocal quartet

Therefore, demonstrating the value of the order in a classification, by placing this work at 783.4 we have placed it alongside (or, collocated it with) other sacred anthems, motets and chants. Further, we have placed it adjacent to all sacred music, which itself is adjacent to all music. And finally we have collocated it by facet with all vocal quartets. It also is important to note that we are classifying concepts in this case rather than specific documents. Thus we are able to make a more precise representation of the musical concepts represented in the work. Notice also that we have captured in the classification all of the terms represented by either LCGFT or LCMPT and done so in a more expressive string.

6.0 The mensural music knowledge Graph

To demonstrate the SW capability of the musicological content of the CMME database we have produced a "Mensural Music Knowledge Graph" based on the content for which we have provided linkable or linked data. All authority records in VIAF are available in RDF. Figure 5 shows the record for the same work as in Figure 4, rendered in RDF.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix schema: <http://schema.org/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dc: <http://purl.org/dc/terms/> .

<http://viaf.org/viaf/100226104>
  rdfs:label "Jacquet," ;
  a <http://bibliograph.net/Agent> .

<http://viaf.org/viaf/4261159156323412180006/>
  foaf:primaryTopic <http://viaf.org/viaf/4261159156323412180006> ;
  void:inDataset <http://viaf.org/viaf/data> ;
  a <http://www.w3.org/2006/gen/ont#InformationResource>, foaf:Document .

<http://viaf.org/viaf/4261159156323412180006>
  schema:creator <http://viaf.org/viaf/4261159156323412180006/#Agent/jacquet> ;
  schema:sameAs <http://id.loc.gov/authorities/names/no2020067385> ;
  skos:prefLabel "Magnificats,"@en-US ;
  rdfs:comment "Warning: skos:prefLabels are not ensured against change!"@en ;
  schema:description "of Mantua"@en-US ;
  schema:name "Magnificats,"@en-US ;
  schema:author <http://viaf.org/viaf/100226104> ;
  a schema:CreativeWork ;
  dc:identifier "4261159156323412180006" .

<http://viaf.org/viaf/4261159156323412180006/#Agent/jacquet>
  rdfs:comment "This is a placeholder URI in need of further matching" ;
  schema:deathDate "1559" ;
  schema:birthDate "1483" ;
  a schema:Person, <http://bibliograph.net/Agent> ;
  schema:name "Jacquet," .

<http://viaf.org/viaf/sourceID/LC%7Cno2020067385#skos:Concept>
  foaf:focus <http://viaf.org/viaf/4261159156323412180006> ;
  skos:prefLabel "Jacquet, of Mantua, 1483-1559. | Magnificats, singers (4)" ;
  skos:inScheme <http://viaf.org/authorityScheme/LC> ;
  a skos:Concept .
```

Figure 5. RDF rendering of authority record for Jacquet of Mantua's *Magnificat*.

To create our knowledge graph, we used a crawler to grab all RDF renderings of the authority records we had created from VIAF. To these we manually added links to the UDC using our tabular data. A model appears in Figure 6.

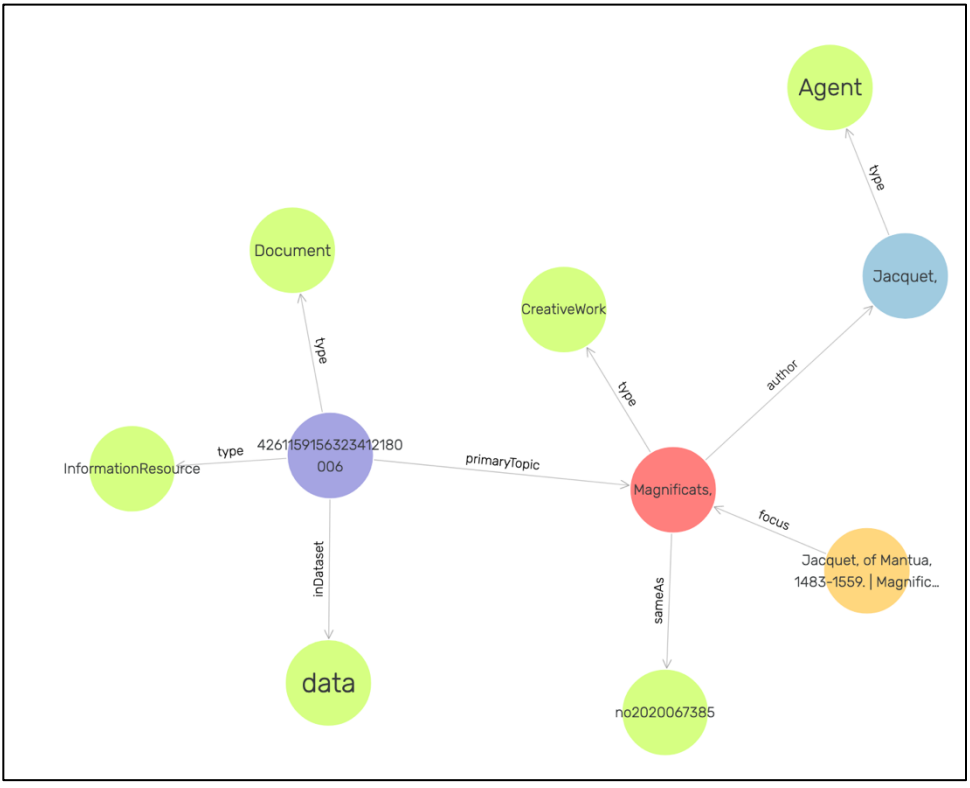


Figure 6. Model of the Mensural Music Knowledge Graph.

7.0 Conclusions

As we noted in the beginning, the goal of the CMME use case as part of the Di4kg project was to use LOD and SW technologies properly to produce contextual enrichment at the level of single artifacts (in this case composers and musical works), in effect creating an environment in which humanities research expressed as LOD might become self-organizing in the LOD cloud. Our major realization, and in effect our major accomplishment, was that we would have to find a way to convert data representing composers and their musical compositions as LD. We were able to manage this by creating a conduit for contributing authority records for the entities in the CMME database eventually into the LD environment of the VIAF.

We were fortunate to be able to utilize the bibliographic community's intellectual structure for the control of musical works. The creation of authorized access points for composers and their works is essentially a curatorial activity, discovering as much as possible about the extant manuscripts or publications of the works, including recording and controlling variant forms whenever possible. These control mechanisms, terms "preferred titles," are also accompanied by extensive records of sources consulted and are amplified by links to form and genre, subject heading, and medium of performance thesauri.

This curatorial activity makes use of the recognition shared among the musicological and bibliographic domains that musical works are mentefacts—intellectual (or mental) constructs—for which artifacts—manuscripts, scores, parts, and even recordings—exist and are controlled in the real world by information systems (Thomas and Smiraglia 1998; Smiraglia 2002; Smiraglia 2019). These mentefacts and the artifacts that represent their instantiations are in turn cultural artifacts. Meaning is abstract at every level from conception to reception to inherence. Cultural meaning, collective and individual, is mutable and constantly mutating—e.g., follow a specific work across time through its varied performance instantiations. Work identifiers—preferred titles—are historical anchors in knowledge organization systems. They serve as cultural memory triggers on the one hand, and as names of classes of instantiations on the other.

We can take a simple example from the *Symphony no. 5* of Ludwig van Beethoven. The musical signature is a cultural signal, even for people who do not know the full work (Figure 7).

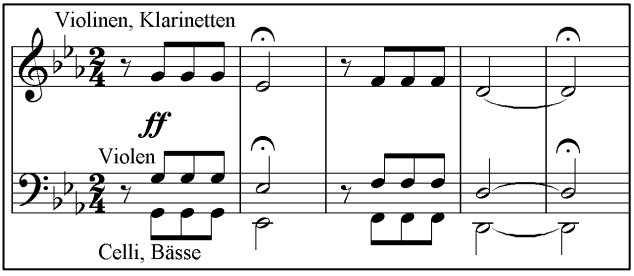


Figure 7. Opening theme of Beethoven's *Fifth Symphony*.

Information retrieval systems organize musical works by composer, and subdivide the "class" of composer by "work" preferred title. A bibliographic system model appears in Figure 8.

Beethoven, Ludwig van, 1770-1827. Symphonies, no. 5, op. 67, C minor

Symphony no. 5 / by Ludwig van Beethoven; op. 67. First performed in Vienna, December 22nd, 1808 under the direction of Beethoven. Revised by Max Unger. With foreword by Wilhelm Altmann. — London ; New York : Ernst Eulenberg, [n.d.]. — viii, miniature score (136 p.) ; 18 cm. — Edition Eulenberg ; no. 402. — Dedicated to H.H. the Reigning Prince Lobkowitz, Duke of Raudnitz and to His Excellency Count Rasumowsky.

Symphony no. 5 in C minor, op. 67 / Beethoven. — Orchestra score. — New York : Edwin F. Kalmus, [19--?]. -- 1 score (100 p.) ; 34 cm. — (Kalmus orchestra library). -- Cover title.

Figure 8. Bibliographic representation of a score of Beethoven's *Fifth Symphony*.

The problem for musicological scholarship as for knowledge organization is to simultaneously collocate or gather all instantiations of a work, and then to disambiguate the gathered cluster by distinguishing variants.

The problem for LD is to move the complex systems created manually for successful clustering and disambiguation into the LOD cloud through the use of SW technologies. In our project we were able to convert a large component of the CMME mensural music database to LD by entering each composer and musical work (mentefact) into LD authority records. These records are themselves linked to knowledge organization systems ranging from the alphabetico-classified system of composer and title indexes to the LOD thesauri of subject headings, forms, genres and medium of performance terms. In this way we have attempted to activate the self-indexing capability of the SW. For the purpose of introducing the traditional advantages of classification we have participated in the conversion of the UDC to LD, and we have linked each work in our Mensural Music Knowledge Graph to the UDC. For every work we have linked to a musical term we have created a permanent identity in the LOD cloud such that every new instantiation of that work can itself be linked to the same term. Similarly, every link to a given term will become a node in a complex network of links to the musical work. All of it is, in turn, linked to the powerful UDC.

We have demonstrated the vast potential of the LOD Cloud to contribute to scholarship in musicology, and by extension, in other artifact-rich humanistic endeavors.

Notes

1. Digging into the Knowledge Graph ; TAP: https://diggingintodata.org/awards/2016/project/digging-knowledge-graph; Project: http://di4kg.org/
2. See Berners-Lee 2006; see for example "5 Star Open Data." https://5stardata.info/en/

★ make your stuff available on the Web (whatever format) under an open license

★★ make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)

★★★★ use URIs to denote things, so that people can point at your stuff

★★★★★ link your data to other data to provide context

References

Baca, Murtha, and Gill, Melissa. Encoding Multilingual Knowledge Systems in the Digital Age: the Getty Vocabularies. *Knowledge Organization* 42: 232-43.

Berners-Lee, Tim. 2006. "Linked Data." https://www.w3.org/DesignIssues/LinkedData.html

Dumitrescu, Theodor and Marnix van Berchum. 2009 "The CMME Occo Codex Edition: Variants and Versions in Encoding and Interface." In *Digitale Edition zwischen Experiment und Standardisierung: Musik - Text - Codierung*, ed. Peter Stadler and Joachim Veit. Beihefte zu Editio 31. Tübingen: Niemeyer, 129-46.

Library of Congress. 2020. *Library of Congress Genre/Form Terms*. Washington, D.C.: Library of Congress. http://id.loc.gov/authorities/genreForms.html

Renwick, Tobias and Rick Szostak. 2021. "A Thesaural Interface for the Basic Concepts Classfication." In *Linked Open Data for Knowledge Organziation and Visualization*, ed. Andrea Scharnhorst and Richard P. Smiraglia. Baden-Baden: Ergon, xx-xx.

Signer, Emmanuel. 2019. "Ceci n'est pas un titre, Or: What We can Learn from Titles in Early Modern Printed Music?" Paper presented at Works, Work Titles, Work Authorities: Perspectives on Introducing a Work Level in RISM, Mainz, May 9-11.

Slavic, Aida, Ronald Siebes and Andrea Scharnhorst. 2021. "Publishing a Knowledge Organisation System as Linked Data: the case of the Universal Decimal Classification." In *Linked Open Data for Knowledge Organziation and Visualization*, ed. Andrea Scharnhorst and Richard P. Smiraglia. Baden-Baden: Ergon, 70-99.

Smiraglia, Richard P. 2002. “Musical Works and Information Retrieval.” *Notes: The Quarterly Journal of the Music Library Association* 58: 747-64.

Smiraglia, Richard P. 2019. “Work.“ *Knowledge Organization* 46: 308-19.

Smiraglia, Richard P. and Rick Szostak. 2018. “Converting UDC to BCC: Comparative Approaches to Interdisciplinarity.” In *Challenges and Opportunities for Knowledge Organization: Proceedings of the Fifteenth International ISKO Conference 9-11 July, Porto, Portugal,* ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Würzburg: Ergon Verlag, 30-38.

Smiraglia, Richard P. and Rick Szostak. 2021. “Identifying and Classifying the Phenomena of Music.” In *Linked Open Data for Knowledge Organziation and Visualization*, ed. Andrea Scharnhorst and Richard P. Smiraglia. Baden-Baden: Ergon, xx-xx.

Szostak, Rick, Andrea Scharnhorst, Wouter Beek, and Richard P. Smiraglia. 2018. “Connecting KOSs and the LOD Cloud.” In *Challenges and Opportunities for Knowledge Organization: Proceedings of the Fifteenth International ISKO Conference 9-11 July, Porto, Portugal,* ed. Fernanda Ribeiro and Maria Elisa Cerveira. Advances in Knowledge Organization 16. Würzburg: Ergon Verlag, 521-29.

Thomas, David H. and Richard P. Smiraglia. 1998. “Beyond the Score.” *Notes: The Quarterly Journal of the Music Library Association* 54: 649-66.

Allard Oelen
Leibniz University Hannover

Mohamad Yaser Jaradeh
Leibniz University Hannover

Markus Stocker
Leibniz Information Centre for Science and Technology

Sören Auer
Leibniz Information Centre for Science and Technology

Chapter 10
Organizing Scholarly Knowledge leveraging Crowdsourcing, Expert Curation and Automated Techniques

Abstract
Research is a fundamental pillar of societal progress. Yet we use antique methods for representing and sharing scholarly knowledge: scientific articles. Instead of representing research in static PDF articles, we work on a dynamic knowledge graph, the Open Research Knowledge Graph (ORKG), where ideas, approaches and methods are represented in machine-readable form. The core rationale of the Open Research Knowledge Graph is to facilitate the manual and automated curation of interlinked, rich semantic descriptions of research contributions. The task of converting unstructured research papers into structured papers is a cumbersome process that requires domain experts. In order to solve difficulties with crowdsourcing, a more structured approach is needed to support users in creating uniform structured paper descriptions. We propose a machine-in-the-loop approach to provide the users guidance during the process of describing a paper. The transformation from unstructured research papers to structured contributions is mainly performed via crowdsourcing. The ORKG infrastructure makes use of automated techniques to help users while adding new data or by extending the data. Another aspect of knowledge curation within the ORKG is to extract structured data from unstructured text. The ORKG system wields the power of structured data to provide yet another interface to explore and interact with scholarly knowledge using question answering.

1.0 Towards truly digital scholarly communication

Research is a fundamental pillar of societal progress. Research enabled us to be connected to the whole world using the small digital devices in our hands, to already cover almost 50% of our energy consumption from renewable energy and cure previously deadly diseases such as AIDS. We are spending worldwide more than 2 trillion esurs—a figure with 12 zeros—per year for acquiring new knowledge through research. However, currently this is not a good investment and every year a larger and larger share of this investment is wasted. The reason for this is that for representing and sharing research findings we use antique methods, which were developed many centuries ago. Since the beginning of modern science—with the publishing of the first scientific journal the *Journal des savants* in

1665—we use the same methods for representing and sharing scholarly knowledge: scientific articles.

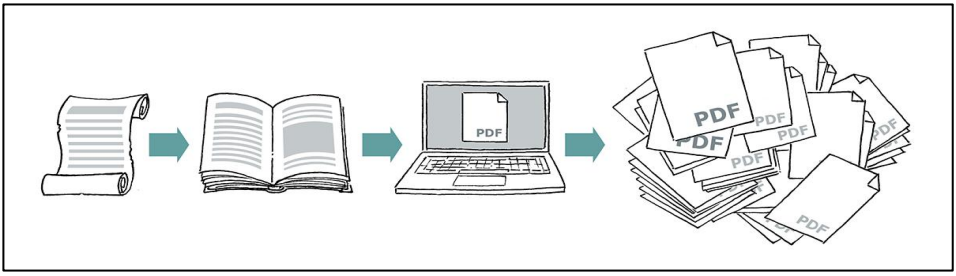


Figure 1. New knowledge, old methods: For centuries, the same method has been used to pass on research knowledge—scientific articles.

At the time of Gottfried Wilhelm Leibniz around 1700, a single researcher could still read all of published scientific literature. Today, every year 2.5 million new research articles are produced and even in a relatively narrow scientific field it is impossible to read, comprehend and make sense of all the scientific articles. For the genome editing method CRISPR/Cas9, for example, the research search engine Google Scholar lists almost a quarter million publications available as PDF articles. If a researcher is interested in how good the method is compared to other genome editing methods, what specifics it has when applied to insects and who has applied it to butterflies, a researcher needs either years of experience or is very likely not to find what is being sought. Imagine, if in order to order a new smartphone, you had to compare prices checking dozens of mail order catalogs published as PDF or to navigate to a hotel, you would need to look at a PDF scan of a street map. This is exactly how the exchange of research findings works today—the previously analog articles from scientific journals are now made available and distributed as PDF documents. The new methods of the digital world, such as filtering large amounts of data and information, integrating information from different sources or involving users via crowdsourcing to review and help organizing the information, are still largely missing in scholarly communication. Researchers are drowning in a flood of millions of pseudo-digitized PDF publications. As a result, research is seriously flawed: many research results cannot be reproduced by other researchers, peer-review is defunct and we have more and more redundancy. Major social challenges such as climate neutrality require interdisciplinarity and putting bits and pieces from different disciplines together.

Instead of representing research in static PDF articles, we work on a dynamic knowledge graph, the Open Research Knowledge Graph (ORKG), where ideas, approaches and methods are represented in machine-readable form. As a result, we can query the graph, for example, for a systematic comparison of different genome editing methods, and such a comparison can be created instantly.

Thus, researchers can easily access the state-of-the-art in a certain field and more precisely devise how their approaches go beyond. As a result, different research contributions can be seamlessly integrated and scientific discovery can be accelerated for solving grand challenges of the next decades, such as carbon neutrality or infectious diseases. Solving

such grand challenges requires interdisciplinarity and assembling bits and pieces from different disciplines, which will be dramatically simplified by a knowledge graph-based approach.

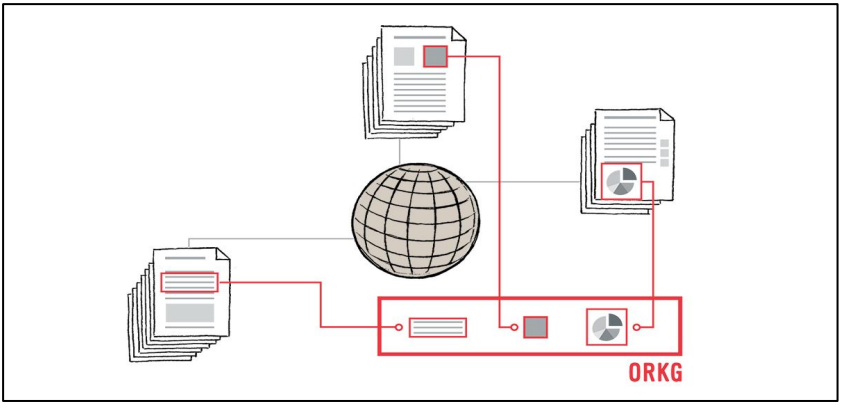


Figure 2. Connecting semantic descriptions of research contributions with various research artifacts using the Open Research Knowledge Graph.

2.0 Open Research Knowledge Graph

The core rationale of the Open Research Knowledge Graph is to facilitate the manual and automated curation of interlinked, rich semantic descriptions of research contributions. In addition, the ORKG is interlinking these semantic descriptions with further artifacts of the research life-cycle, such as data, software and visualizations. As a result of such a semantically-rich interlinked representation of research knowledge, aggregated views of this knowledge, such as comparisons of contributions or domain-specific visualizations of the state-of-the-art can be automatically generated. Before we describe the crowd-sourcing and expert curation techniques and some concrete use cases in more detail in subsequent sections, we first give an overview on the technical architecture of the ORKG.

The infrastructure of the ORKG follows a classical layered architecture, as depicted in Figure 3. A persistence layer abstracts the raw data stores that keep the data of the ORKG. Multiple models are used to cover different requirements of the system. A property graph models the main data in the ORKG, resources, statements, and literals are stored with more provenance information annotating statements and resources. The data in the property graph are synchronized with a triple store that contains the data represented in Resource Description Framework (RDF) format. Other stores (relational model) are used to store management information for users and user roles. The domain layer above the persistence layer contains the domain model, provenance management, and user authentication. Building on top of that, the application layer provides business logic that organizes input/output operations on the data and guarantees consistency. Furthermore, a REST API (Application Programming Interface) implements all the functionalities to manipulate the underlying data, and is a window to the outside world to interact with the system. For more advanced users, the ORKG provides a SPARQL endpoint that enables users to interact with the data in any manner they deem fit. At the top of the stack, the User Interface (UI) takes advantage

of all components underneath to enable various functionalities for users. Such functionalities include exploring the ORKG data, searching the content, visualizing data and graphs, contributing new data, editing existing data, and comparing resources and papers (Jaradeh et al. 2019).

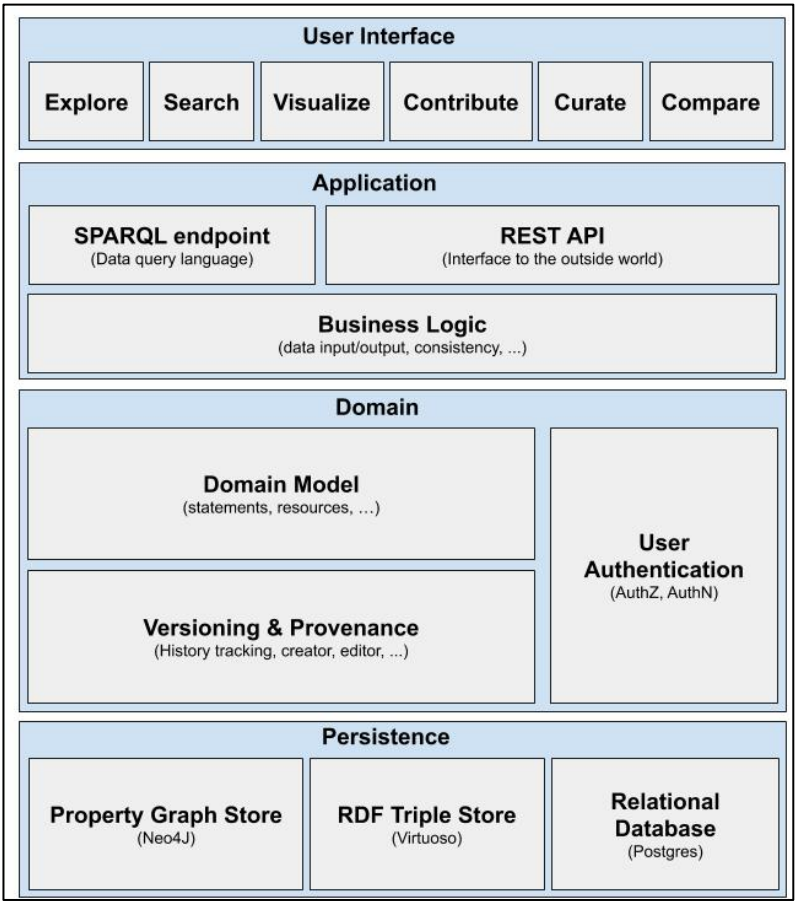


Figure 3. ORKG layered architecture from data persistence to services.

The rationale of ORKG (in analogy to the RDF, RDF-Schema and OWL knowledge representation layers) is to provide a very concise core data model (resembling RDF) and domain-specific data models (i.e., vocabularies or ontologies) realized on top of the core data model. While the core-data model is fixed, we envision the domain specific data models to evolve based on the manual collaboration on the ORKG platform. This agile-inspired knowledge organization (KO) approach deviates from the still common KO methodologies, where schema and ontology development is strictly separated from the knowledge and data acquisition and curation. In ORKG on the other hand we want to intertwine

schema and knowledge/data curation as much as possible. There are some successful examples following a similar approach, such as OpenStreetMaps, where the core data model consists simply of points, ways and relations and the actual semantics is encoded using arbitrary evolving annotations on these.

The ORKG core data model resembles and augments the RDF data model, allowing for interoperability among other things. The data model consists of nodes and edges. This model was chosen to simplify the process of adding information to the system. One of the greatest differences to RDF is that everything is modeled as an entity, i.e., it can be referenced by an identifier (ID). The data model is centered around the concept of a statement. A statement is a triple that consists of a subject and an object (nodes) that are connected by a predicate (relationship). Nodes can have one of two types: resources and literals. Resources represent a concept, such as a scientific method or an author, whereas literals represent values, such as a name of a method or author, or a measurement value. Within statements, literals can only appear in the object position of the statement.

The ORKG data model (Figure 4) appends provenance information on all elements of the data model, every entity whether it is a resource, predicate or a literal, has provenance information (when created/edited, who created/edited, etc.). Furthermore, every statement in the ORKG holds provenance information as well, by using property graphs annotating edges. Adding such provenance and further meta-information is the main difference to a pure RDF data model. Of course, our data model can be mapped and transformed into an equivalent RDF data model, which, however, would then make extensive use of debated and performance-prone reification features.

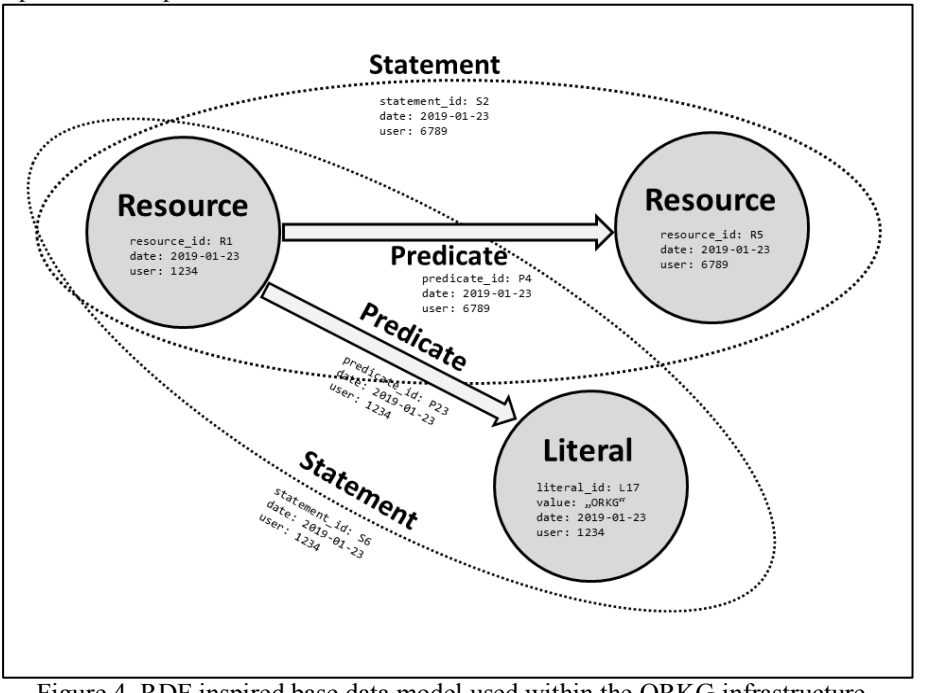


Figure 4. RDF inspired base data model used within the ORKG infrastructure.

Third party applications as well as the original ORKG user interface consume the ORKG data via the REST API. The API provides access to the main CRUD (Create, Read, Update, Delete) operations on all main components on the data model, and on some other high-level concepts (e.g., Papers, Comparisons). The API abstracts all implementation details from potential clients and applications. It also makes sure that requests and responses are in an adequate and suitable format. Furthermore, the API is the gatekeeper for some sensitive calls that would require elevated access or permissions for user roles of the system.

3.0 Crowd-sourcing and expert curation

3.1 ORKG challenges

The task of converting unstructured research papers into structured papers is a cumbersome process that requires domain experts. Since domain knowledge is required, this task cannot be performed by machine-learning or natural language processing alone. The accuracy of such techniques is currently not sufficient to generate a high-quality scholarly knowledge graph that is suitable for research. On the other hand, if this conversion is manual, for example, via crowdsourcing, the result is highly subjective and expensive. As a consequence, the resulting structured data highly differ, between contributors, which negatively impacts the quality of the knowledge graph. Some contributors might describe information on a high level while others do this on a more granular level. Furthermore, the same pieces of information might be described differently.

In order to solve the previously described difficulties with crowdsourcing, a more structured approach is needed to support users in creating uniform structured paper descriptions. We propose a machine-in-the-loop approach to provide the users guidance during the process of describing a paper. Predefined templates can be used to describe common research contribution structures. For instance, in the computer science domain, such templates can be used to describe the approach, implementation or evaluation of a research contribution. Using Named Entity Recognition (NER), relevant templates can be automatically selected for a specific paper. To further support users and to increase ontology reuse, intelligent suggestions of existing vocabularies and resources are provided as much as possible. This increases the machine actionability and the overall quality of the knowledge graph. Additionally, relevant existing research descriptions are automatically selected and serve as examples. In the end, the previously complex task of making a structured description of a paper is reduced to populating and curating the automatically selected templates. The descriptions of the templates are crowdsourced as well. To create new templates, a domain expert has to identify commonly shared patterns across papers that tackle the same research problem. A key factor is that domain experts can have discussions about the structures of the template patterns to reach consensus on how to describe specific pieces of information. When research contributions addressing a similar research problem are described with the same templates, the contributions become comparable. For instance, it is possible to select a set of papers and to automatically compare quality metrics such as F-measures to see which implementation performs best. Also, finding and comparing state-of-the-art becomes more straightforward with structured paper descriptions.

3.2 Interplay between user roles

The transformation from unstructured research papers to structured contributions is mainly performed via crowdsourcing. The process is supported by automated techniques. This results in a process that uses the best characteristics of both worlds; first, the high accuracy and intelligence of humans and second the processing power of machines (Ece et al. 2012). ORKG users can be divided into three user roles, 1) content consumers, 2) content creators and 3) content curators. The content consumers are users who consume the content from the knowledge graph. Those users include researchers who are using ORKG for finding and comparing contributions or performing data analysis. The second group, the content creators, are part of the crowdsourcing approach. Those users consist of paper authors and other crowd workers, for example librarians, who are adding individual papers to the ORKG. Finally, the content curators are responsible for managing and curating the content. These users are domain experts for a specific sub-field. Within ORKG, managing sub-fields is done via observatories. The content curators are members of an observatory for their field of expertise. The role of observatories is to manage and structure the content for a sub-field. There is collaboration between content creators and content curators. Content curators are providing guidance to content creators on how to structure content (in the form of templates). Content creators provide content that is curated by the content curators. The interplay between content creators, content curators and automated approaches is depicted in Figure 5.

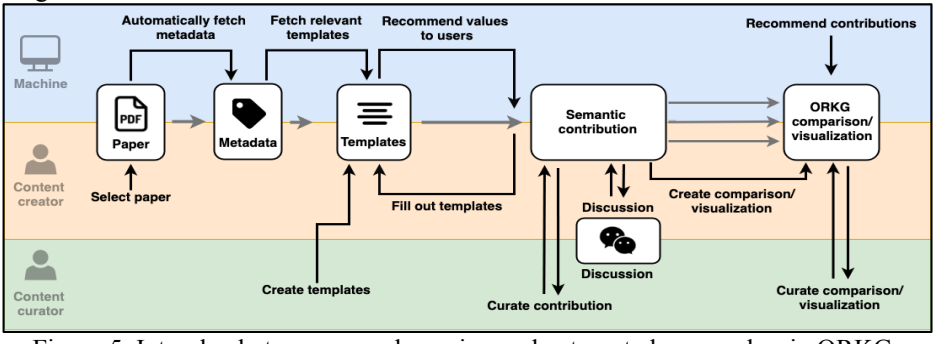


Figure 5. Interplay between crowdsourcing and automated approaches in ORKG.

3.3 Templates for generating comparable paper descriptions

When structured data are available for a set of contributions that are addressing the same research problem, a comparison can be created. Comparisons provide an overview of the state-of-the-art for a specific problem. Compared to literature comparisons (or surveys) published in static PDF articles, the ORKG approach has several benefits (Oelen et al. 2020). Firstly, comparisons can evolve over time. Meaning that new research can be added to a comparison, as soon as it is published. Secondly, it is possible to collaborate on comparisons, for example to add your own work. Thirdly, ORKG serves as a data repository for comparisons. An ORKG comparison can be generated automatically based on contributions that address the same problem and contributions that are similar in structure. Users can also manually add contributions to a comparison. Once a comparison is formatted, it can be published. Publishing a comparison ensures that the state of a comparison is saved, to support data persistency. A persistent identifier is attached to a comparison. In the future,

(DataCite) DOIs will be assigned to the comparisons. This makes reuse possible for comparisons in research articles. Furthermore, the publication of a comparison ensures that relevant metadata are attached to the comparison. This contributes to the overall machine-actionability of the ORKG data. Comparisons can be exported to multiple formats, including LaTeX, CSV, PDF and RDF. The LaTeX export is specifically focused on using comparisons in research papers. In addition to the exported comparison table, a BibTeX file is generated which contains the paper references used in the comparison. Comparisons can be created both by content creators and content curators. By default, ORKG comparisons are visualized in a tabular format. However, due to the underlying structured data, the data could be visualized in other, more appropriate formats. For example, using trend charts to visualize data over time or world maps to visualize data for different countries.

Content creators are generally users who will only add individual papers. On the other hand, content curators could also import larger numbers of papers for their field of expertise. For this, ORKG leverages survey tables presented in literature survey articles. Such tables list paper data in a (semi-)structured way and are therefore suitable to be imported in a knowledge graph. Although extracting these tables from papers will not provide complete contribution descriptions, they can serve as an initial starting point for structured contributions. This automatically extracted description could later be extended by the content creators. In order to import survey tables, content creators first have to select relevant survey articles. Afterwards they extract the table from the PDF article, adding the full information of the referenced paper and put this in a CSV file. Finally, the CSV file is imported in the ORKG.

One of the advantages of importing survey tables from existing papers is that the overall machine-actionability of the data improves. In general, scholarly articles are published in PDF format. PDF files do not contain any information about the structure of the content, only the layout and formatting is stored (Deliang and Yang 2009). As a consequence, tables within PDF files cannot simply be read by machines. The table has to be extracted first, which means extracting the text and trying to replicate the tabular structure as presented in the paper. Building a knowledge graph using the extracted tables has the direct benefit of more machine-actional data. If comparisons are published with the ORKG they become more FAIR (Findable, Accessible, Interoperable, Reusable) compared to the comparisons published in static PDF documents.

4.0 Automated techniques

4.1 Recommendation of values and papers

The ORKG infrastructure makes use of automated techniques to help users while adding new data or by extending the data. The ORKG uses machine learning for automated extraction of scientific knowledge from literature. Of particular interest are the NLP tasks NER as well as named entity classification and linking. As a first step, we trained a neural network-based machine learning model for NER using in-house developed annotations on the Elsevier Labs corpus of Science, Technology, and Medicine (STM[1]) for the following generic concepts: process, method, material and data. The ORKG system uses the Beltagy et al. (2019) NER task-specific neural architecture atop pretrained SciBERT embeddings with a CRF-based sequence tag decoder. This is a supporting part in the user curation pro-

cess, when users are adding their data into the ORKG, an abstract annotator using the mentioned model, annotates the four generic concepts, and the user can refine the extraction and add them to the contribution of the paper.

Furthermore, automatic methods in the ORKG try to find connections between existing papers in the system. The more the papers have similar structures and similar content, the more they are related and are recommended for comparisons or visualizations. Figure 6 illustrates how automatic detection of entities is done on abstracts, helping users to add their data quickly into the system.

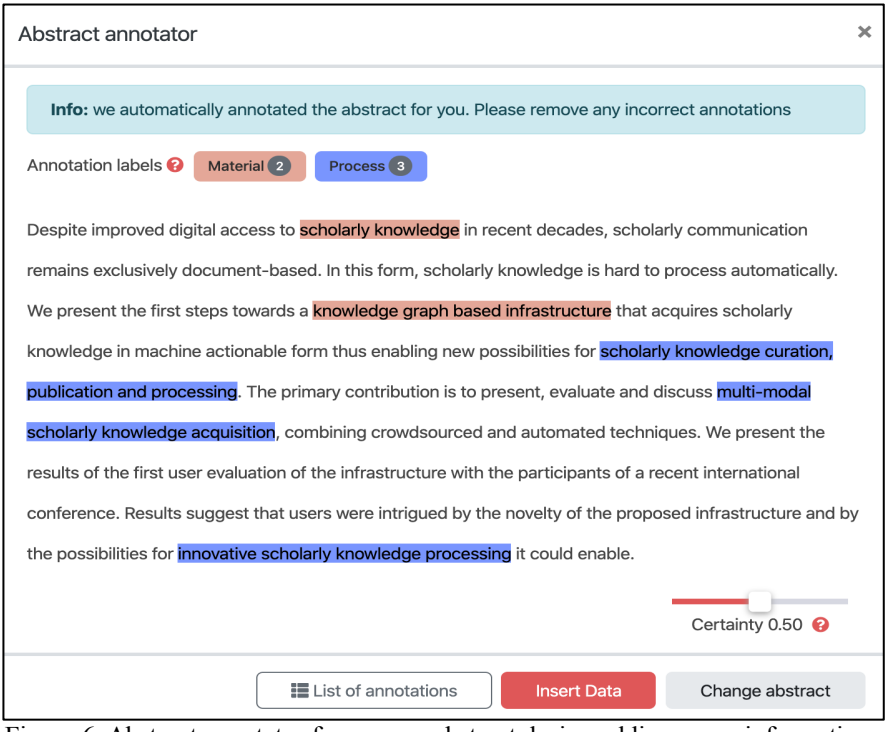


Figure 6. Abstract annotator for a paper abstract during adding paper information.

4.2 Integration of automated extraction results via an API

The ORKG also strives to be a singular point to collect existing datasets and automatically extracted results made available by the research community. To that end, the ORKG APIs set the provenance of the data when adding it (i.e., extracted from, extraction method). Moreover, the infrastructure systemizes the collection of external data by mapping it to RDF, then enriching it and importing it to the graph. This process consumes data from multiple formats (RDF, JSON, CSV, etc.) and then maps the data using different techniques (RML[2], CSVW[3], JSON-LD[4]) to an intermediate format which is then imported into the ORKG. The mapping process connects lone datasets using common predicates and resources uncovering previously unknown links between datasets at scale.

4.3 Structured data extraction from unstructured text

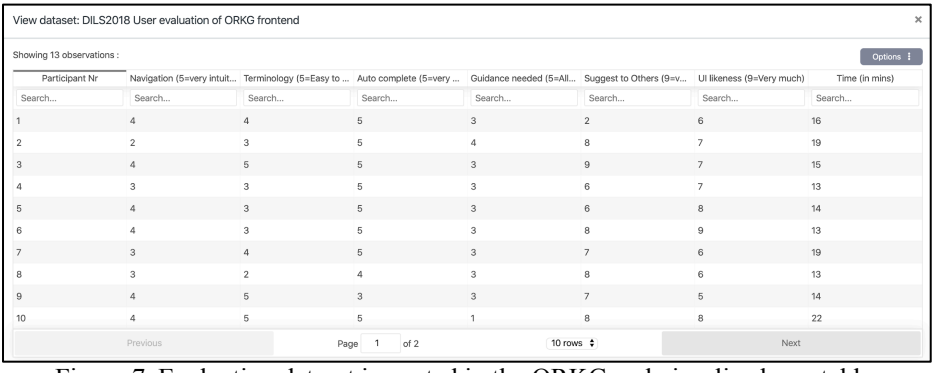
Another aspect of knowledge curation within the ORKG is to extract structured data from unstructured text. The unstructured exists in high volumes as static publications in the form of PDF or other digital representations of scientific articles (Gandomi and Haider 2015). The ORKG leverages existing tools as well as custom tailored ones to extract triples of the form <Subject, Predicate, Object> from the unstructured text. Such triples go through a pipeline of steps (components) to align them from ambiguous information into known concepts in the knowledge graph. The pipelines usually have certain steps to go through, varying from triple extraction, coreference resolution, entity extraction, entity disambiguation and linking, relation extraction, and relation disambiguation and linking (Kertkeidkachorn and Ichise 2018). Different pipelines stretch over different combinations of components and in different orders. These pipelines in the ORKG can be customized to extract certain types of information (e.g., COVID reproduction numbers, or methods and algorithms used in papers). The dynamic nature of the pipelines allows for great extensibility with new components, steps, and methods, improving the overall performance of such pipelines.

Each of the different steps that the process goes through has different objectives. Extraction components focus on creating text triples of the form subject, predicate, object from the unstructured text. The coreference resolution step mainly addresses pronouns and acronyms in the unstructured text and tries to resolve them with the original mention to the end of disambiguating the text for other components. Entity recognition as well as relation recognition work to find the named entities in the text (i.e. measure, method, problem, etc.) and similarly the relation phase strives to find the relation between said entities (i.e. utilizes, produces, etc.). Moreover, entity and relation disambiguation components work on linking mentioned entities and relation to their linked counterpart in the ORKG graph. Singh et al. (2018) mention a collection of entity and relation linkers that can be used on DBpedia knowledge graph. The collection of all steps will result in new triples (new pieces of information) to be added to the knowledge graph.

4.4 Research data integration with Jupyter NB

Data are also byproducts of research, and during the life cycle of research researchers often have data produced using Jupyter notebooks (Stocker et al. 2019). To that intent, ORKG has a python package that can be used easily to add/search/update data inside the ORKG. This is a two-way street, it can be used to add newly created datasets into an ORKG contribution or it can be used to fetch data from the ORKG papers[5] and perform more in-depth analysis or visualization in the Jupyter notebooks environment.

Figure 7 shows evaluation results imported into the ORKG from a CSV file and visualized in a tabular format via the infrastructure. The same data can be fetched from the ORKG system and then visualized in different sorts of visualizations in notebooks as well.


View dataset: DILS2018 User evaluation of ORKG frontend

Showing 13 observations :

Options

Participant Nr	Navigation (5=very intuit...	Terminology (5=Easy to ...	Auto complete (5=very ...	Guidance needed (5=All...	Suggest to Others (9=v...	UI likeness (9=Very much)	Time (in mins)
Search...	Search...	Search...	Search...	Search...	Search...	Search...	Search...
1	4	4	5	3	2	6	16
2	2	3	5	4	8	7	19
3	4	5	5	3	9	7	15
4	3	3	5	3	6	7	13
5	4	3	5	3	6	8	14
6	4	3	5	3	8	9	13
7	3	4	5	3	7	6	19
8	3	2	4	3	8	6	13
9	4	5	3	3	7	5	14
10	4	5	5	1	8	8	22

Previous Page 1 of 2 10 rows Next

Figure 7. Evaluation dataset imported in the ORKG and visualized as a table.

4.5 Question answering on scholarly data

One of the benefits of having scholarly data in a machine actionable structured manner is the ability to run automatic information retrieval methods such as Question Answering (QA). The ORKG system wields the power of structured data to provide yet another interface to explore and interact with scholarly knowledge using QA. Different types of question answering systems exist nowadays, some of them run on unstructured text, and some run on knowledge graphs and structured information (Jaradeh et al. 2020). ORKG exploits the later, with the available structured data in multiple forms within the infrastructure, different kinds of QA systems can be used. Comparison tables or datasets within the ORKG represent one aspect of the knowledge that a QA system can operate on. Figure 8 illustrates what an ORKG QA subsystem can do with information represented in tables and the different types of questions that can be answered.

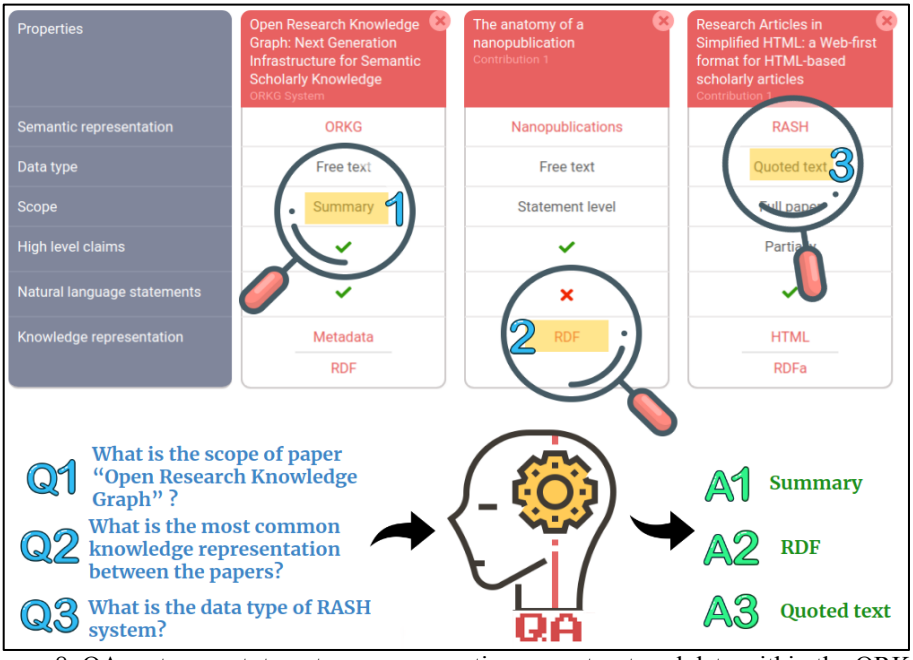


Figure 8. QA system prototype to answer question over structured data within the ORKG.

Other types of QA systems (specifically knowledge graph based) can also be implemented in the ORKG to capture more in-depth questions such as "What is the most common machine learning method used by the state-of-the-art papers addressing entity linking?". Such a question needs deep comprehension of the schema of the graph. Furthermore, it needs to locate named entities in the question (e.g., machine learning, entity linking), find the state-of-the-art papers (based on evaluation metrics and values) and then get the "most common" method of those papers, to be able answer the question. Regardless of how complicated it is to find the answer or to represent it, as long as the knowledge graph contains the information, it is possible to create tools and systems to find the answers more accurately and efficiently.

5.0 Semantic scholarly contribution examples

In this section, we demonstrate how the ORKG can be used to structure information for three different domains. For each domain, an ORKG comparison is generated to give an overview of the state-of-the-art for this particular example.

5.1 Computer science

In the computer science domain a set of recurring properties are of interest for the ORKG. Among other properties, they include evaluation results (F-measure, precision, recall), datasets, benchmarks and implementation details. For a considerable number of computer science papers these properties are applicable and of interest for many researchers. In this example, we demonstrate how these recurring properties are used in the Question Answering (QA) domain.

The objective of the QA task is to automatically provide answers to natural language questions (Soricut and Brill 2006). To organize QA related information in the ORKG, two different research problems are addressed. The first problem, "Question answering systems," relates to question answering systems in general, describing their features and which tasks are addressed. The second problem, "Question answering systems evaluation," addresses the evaluation results of the compared systems. An interesting aspect of these problems is that the compared QA systems and the evaluation results of those systems are originally not published in the same articles nor by the same authors. However, due to the dynamic aspects of the ORKG, it is possible to create links between the systems and evaluations retrospectively. In Figures 9 and 10 simplified comparisons of both the QA systems and their evaluations are depicted to further explain the relation between the different research problems. In Figure 9, multiple QA systems are compared based on the tasks performed by this system. Each QA system is presented in a different research paper. Furthermore, the tasks are grouped via well-defined subtasks for this community (e.g., Disambiguation task, Query construction task). In Figure 10, a comparison is displayed evaluating some of the QA systems from the previous comparison. The Evaluation property is used to link an evaluation to the respective QA system. For example, the resource "SemGraphQA" describes the system that has been evaluated. Other properties in this comparison are related to the evaluation itself. The Dataset and Language properties describe the characteristics of the datasets used for the evaluation. In this comparison, the datasets are the same for each evaluation, but the languages differ between them. Finally, the actual evaluation results are compared based on the properties F-measure, precision and recall.

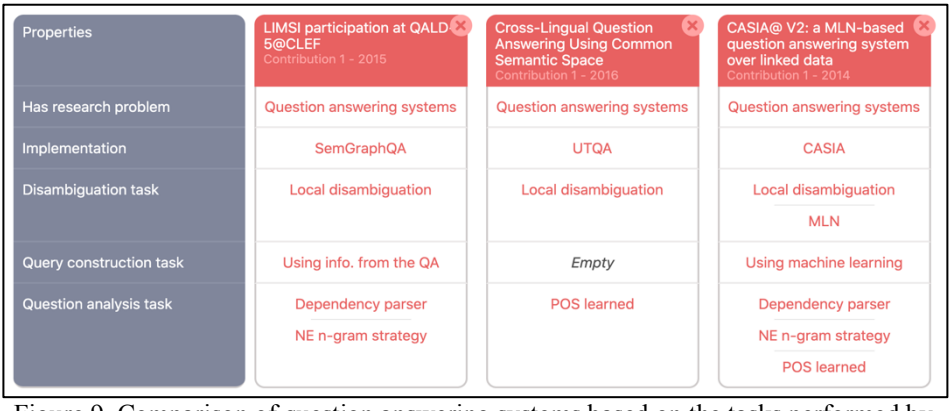


Figure 9. Comparison of question answering systems based on the tasks performed by these systems (data imported from Diefenbach et al. 2018).

Properties	6th Open Challenge on Question Answering over Linked Data (QALD-6) Contribution 4 - 2016	6th Open Challenge on Question Answering over Linked Data (QALD-6) Contribution 3 - 2016	6th Open Challenge on Question Answering over Linked Data (QALD-6) Contribution 2 - 2016
Has research problem	Question answering systems evaluation	Question answering systems evaluation	Question answering systems evaluation
On	QALD-6	QALD-6	QALD-6
Dataset	DBpedia 2015 DBpedia 2015 with abstracts LinkedSpending	DBpedia 2015 DBpedia 2015 with abstracts LinkedSpending	DBpedia 2015 DBpedia 2015 with abstracts LinkedSpending
Evaluation	SemGraphQA	UTQA	UTQA
Language	Farsi	English	Spanish
F-measure	0.37	0.65	0.68
Precision	0.70	0.70	0.76
Recall	0.25	0.61	0.62

Figure 10. Comparison of evaluation results of the question answering systems presented in Figure 9 (depicted values are imported from Diefenbach et al. 2018).

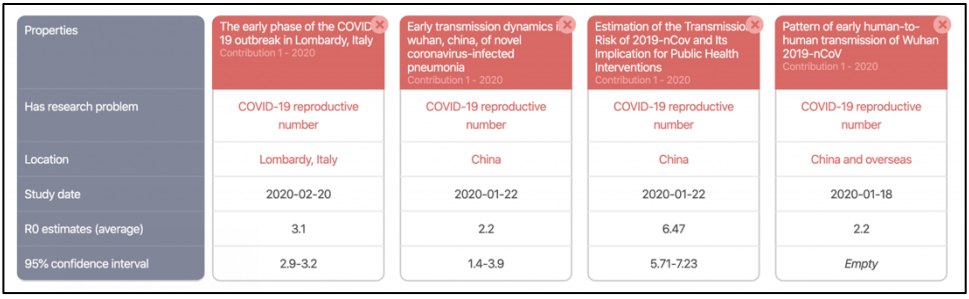
The information presented in the comparison from Figure 9 is using domain specific properties (describing the tasks addressed per system). Creating those domain specific properties is supported since the ORKG does not have a fixed data model, specifically to support such domain specific properties. The second comparison from Figure 10 uses properties that are applicable to computer science papers in general. Especially the properties related to the evaluation results (e.g., the F-measure) are of interest in comparisons. It is possible to quickly get an overview of the state-of-the-art systems and their performance. Additionally, comparing these evaluation results over time gives insights of the advancements for a specific research domain.

5.2 COVID-19

The COVID-19 crisis is driving substantial research with new articles published daily. To support COVID-19 research, many publishers decided to publish related articles via open access (Wang et al. 2020). While access is crucial, organizing information published in articles is essential for effective research, but is extremely time consuming, and time is an asset that under these circumstances is as valuable as ever. One example of information that can be organized across numerous published (preprint) papers is the COVID-19 basic reproduction number (R0), its value, 95% confidence interval, location of the population and the period of observation. Indeed, some authors (Lui et al. 2020) have already collected information on R0 from the literature and published a survey. Other information can be on case fatality rate or modelled cases and their comparison with actual cases.

Contrary to the conventional document-based publishing of such information in natural language text, tables or figures, with ORKG we can publish such information in a structured, semantic manner. Information is thus machine actionable.

Machine actionable scholarly knowledge opens a range of very interesting possibilities. As mentioned before, it is possible to automatically create literature comparisons. Figure 11 showcases this for our use case on the COVID-19 basic reproduction number. It is the structured representation of scholarly knowledge that enables the automatic creation of such comparisons in ORKG. Furthermore, contrary to review articles, ORKG comparisons can evolve. As new literature on R0 research is published, it is straightforward to extend such a comparison, which therefore continues to reflect in a comparable manner the current state of knowledge.



Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy Contribution 1 - 2020	Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia Contribution 1 - 2020	Estimation of the Transmission Risk of 2019-nCov and Its Implication for Public Health Interventions Contribution 1 - 2020	Pattern of early human-to-human transmission of Wuhan 2019-nCoV Contribution 1 - 2020
Has research problem	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number
Location	Lombardy, Italy	China	China	China and overseas
Study date	2020-02-20	2020-01-22	2020-01-22	2020-01-18
R0 estimates (average)	3.1	2.2	6.47	2.2
95% confidence interval	2.9-3.2	1.4-3.9	5.71-7.23	*Empty*

Figure 11. Automatic comparison of basic reproduction numbers published in the literature.

The real power of such ORKG comparisons, however, can be seen if they are taken as data sources. Indeed, thanks to machine actionability of both the data and the data exchange protocol, it is possible to link the ORKG and comparisons, specifically, with downstream data science. We demonstrate this by connecting Jupyter with ORKG to show how we can leverage the flexibility of data science environments and programming languages such as Python and R to visualize or otherwise process the COVID-19 comparison data. Figure 12 shows a possible visualization of R0 values and 95% confidence interval over time. Naturally, such data science activities can draw data from numerous sources, specifically several ORKG comparisons, for instance also case fatality rate. As such, downstream data science benefits not just from comparable, structured information across the literature, but also from integration of data from multiple ORKG comparisons and other sources.

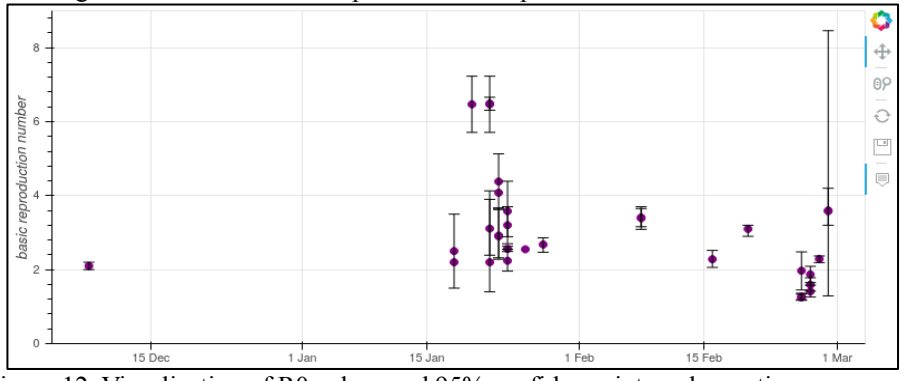
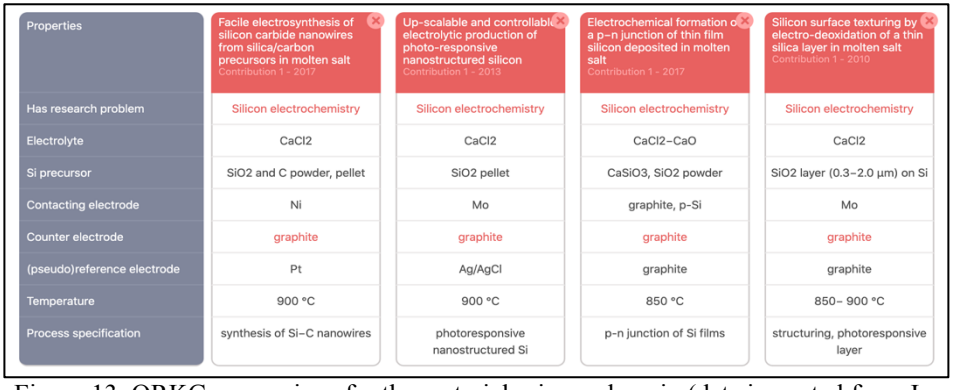


Figure 12. Visualization of R0 values and 95% confidence interval over time, as a possible output of a data science activity that reuses ORKG comparison data.

5.3 Material science

Finally, we present an example from electrochemistry, material sciences and engineering. Silicon is a major element in modern technology. It is widely used to produce metal alloys, optical fibers, solar elements, advanced ceramics, batteries, microchips and numerous other advantageous applications. For solar and electronic devices, there is a need for solar grade silicon (SoG-Si) with a purity of 99.9999% or electronics grade silicon with an even higher purity.

Silicon electrochemistry in molten salts has recently attracted considerable attention due to its potential to produce SolarGrade-Silicon with negligible carbon footprint. This comparison (Figure 13) provides a comprehensive overview of several parameters such as SiliconDioxide Precursors, electrolyte, contacting electrode, or temperature of experimental conditions of Silicon-Electrochemical reduction in molten electrolytes. In this way the researcher easily can retrieve relevant parameters used in the process specifications for the generation of silicon surface structures.



Properties	Facile electrosynthesis of silicon carbide nanowires from silica/carbon precursors in molten salt Contribution 1 - 2017	Up-scalable and controllable electrolytic production of photo-responsive nanostructured silicon Contribution 1 - 2013	Electrochemical formation of a p–n junction of thin film silicon deposited in molten salt Contribution 1 - 2017	Silicon surface texturing by electro-deoxidation of a thin silica layer in molten salt Contribution 1 - 2010
Has research problem	Silicon electrochemistry	Silicon electrochemistry	Silicon electrochemistry	Silicon electrochemistry
Electrolyte	CaCl2	CaCl2	CaCl2–CaO	CaCl2
Si precursor	SiO2 and C powder, pellet	SiO2 pellet	CaSiO3, SiO2 powder	SiO2 layer (0.3–2.0 µm) on Si
Contacting electrode	Ni	Mo	graphite, p-Si	Mo
Counter electrode	graphite	graphite	graphite	graphite
(pseudo)reference electrode	Pt	Ag/AgCl	graphite	graphite
Temperature	900 °C	900 °C	850 °C	850– 900 °C
Process specification	synthesis of Si–C nanowires	photoresponsive nanostructured Si	p-n junction of Si films	structuring, photoresponsive layer

Figure 13. ORKG comparison for the material science domain (data imported from Juzeliūnas and Gray 2019).

6.0 Conclusions

In this article, we presented the ORKG KO system (KOS). The core idea of the ORKG KOS is not to define *apriori* a fixed data model or ontology, but rather to rely on a concise core data model basically resembling RDF with comprehensive provenance and metadata. On top of this core data model, there are very few general entities, such as articles, research problems, contributions, but the vast part of the KOS is envisioned to be happening through the collaboration of the ORKG curators on the platform. We integrated features for easily defining new classes and properties to capture domain specific descriptions and properties of research contributions. Using automated techniques, we aim to facilitate the integration of knowledge in such a way that a coherent knowledge graph emerges. While we described in this article the core concepts and their implementation of the ORKG KOS approach along with some use cases, much more work needs to be done to firmly establish this concept widely in various science communities. In this regard, we work on making ORKG an architecture of participation, where continuous contributions are triggered and encouraged. For example, we are currently realizing domain-specific observatories, which are curated

by research organizations or libraries in a particular field aiming at covering a core of relevant research descriptions for their field in the ORKG. Another angle of future work is making the ORKG a home for the results of the various domain specific scientific literature knowledge extraction projects. We envision loading such datasets into the ORKG but clearly separating such automatic knowledge extraction results from the core ORKG content. Upon manual validation such automatically extracted data can then be used directly within the ORKG.

Notes

1. https://github.com/elsevierlabs/OA-STM-Corpus
2. https://rml.io/
3. https://www.w3.org/ns/csvw
4. https://json-ld.org/
5. https://gitlab.com/TIBHannover/orkg/orkg-covid19-hub/-/blob/master/R0-estimates-plot.ipynb

References

Beltagy, Iz, Kyle Lo and Arman Cohan. 2019. "SciBERT: A Pretrained Language Model for Scientific Text." *ArXiv* 1903.10676 [Cs]. http://arxiv.org/abs/1903.10676

Diefenbach, Dennis, Vanessa Lopez, Kamal Singh and Pierre Maret. 2018. "Core Techniques of Question Answering Systems over Knowledge Bases: A Survey." *Knowledge and Information Systems* 55: 529-69. https://doi.org/10.1007/s10115-017-1100-y

Gandomi, Amir and Murtaza Haider. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35: 137-44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Jaradeh, Mohamad Yaser, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker and Sören Auer. "2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge." Paper presented at K-Cap 2019. 19-21 November 2019, Marina Del Rey California, United States.

Jaradeh, Mohamad Yaser, Markus Stocker, and Sören Auer. 2020. "Question Answering on Scholarly Knowledge Graphs." *ArXiv* 2006.01527 [Cs]. http://arxiv.org/abs/2006.01527

Jiang, Deliang, and Xiaohu Yang. 2009. "Converting PDF to HTML approach based on Text Detection." In *ICIS '09: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*. New York: ACM, 982-5.

Juzeliūnas, Eimutis and Derek J. Fray 2019. "Silicon Electrochemistry in Molten Salts." *Chemical Reviews* 120: 1690-1709.

Kamar, Ece, Severin Hacker and Eric Horvitz. 2012. "Combining Human and Machine Intelligence in Large-scale Crowdsourcing." In *AAMAS '12: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*. Richmond: Systems, 467-74.

Kertkeidkachorn, Natthawut and Ryutaro Ichise. 2018. "An Automatic Knowledge Graph Creation Framework from Natural Language Text." *IEICE Transactions on Information and Systems* E101.D, no. 1: 90–98. https://doi.org/10.1587/transinf.2017SWP0006

Liu, Ying, Albert A. Gayle, Annelies Wilder-Smith and Joacim Rocklöv. 2020. "The Reproductive Number of COVID-19 is Higher Compared to SARS Coronavirus." *Journal of Travel Medicine* 27. https://doi.org/10.1093/jtm/taaa021

Mons, Barend and Jan Velterop. 2009. "Nano-Publication in the E-science Era." In *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*. CEUR Workshop Proceedings 523, 14-15. http://ceur-ws.org/Vol-523/Mons.pdf

Oelen, Allard, Mohamad Yaser Jaradeh, Kheir Eddine Farfar, Markus Stocker and Sören Auer. 2019. "Comparing Research Contributions in a Scholarly Knowledge Graph." In *K-CAP '19: Proceedings of the 10th International Conference on Knowledge Capture*, *September 2019*. New York: ACM, 21-26. http://ceur-ws.org/Vol-2526/paper3.pdf

Oelen, Allard, Mohamad Yaser Jaradeh, Markus Stocker and Sören Auer. 2020. "Generate FAIR Literature Surveys with Scholarly Knowledge Graphs." *In* JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. New York: ACM, 97-106. https://doi.org/10.1145/3383583.3398520

Van de Sompel, Herbert and Carl Lagoze. 2009. "All Aboard: Toward a Machine-friendly Scholarly Communication System." In *The Fourth Paradigm.: Data-Intensive Scientific Discovery*, ed. Tony Hay, Stewart Tansley and Kristin Tolle. Microsoft Research. https://www.semanticscholar.org/paper/All-aboard%3A-toward-a-machine-friendly-scholarly-Sompel-Lagoze/b9d52b620a89a677e729e25836dcd2eb52c1df97

Singh, Kuldeep, Christoph Lange, Maria Esther Vidal, Jens Lehmann, Sören Auer, Arun Sethupat Radhakrishna, Andreas Both, et al. 2018. "Why Reinvent the Wheel: Let's Build Question Answering Systems Together." In *WWW '18: Proceedings of the 2018 World Wide Web Conference*. Geneva: IWWWC, 1247-56. https://doi.org/10.1145/3178876.3186023

Soricut, Radu and Eric Brill. 2006. "Automatic Question Answering Using the Web: Beyond the Factoid." I*nformation Retrieval* 9: 191-206. https://doi.org/10.1007/s10791-006-7149-y

Stocker, Markus, Manuel Prinz, Fatemeh Rostami and Tibor Kempf. 2019. "Towards Research Infrastructures That Curate Scientific Information: A Use Case in Life Sciences." In *Data Integration in the Life Sciences*, ed. Sören Auer and Maria-Esther Vidal. Lecture Notes in Computer Science 11371. Cham: Springer, 61–74. https://doi.org/10.1007/978-3-030-06016-9_6

Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, DarrinEide, Kathryn Funk et al. 2004. "CORD-19: The Covid-19 Open Research Dataset." *ArXiv*. https://arxiv.org/abs/2004.10706

Charles van den Heuvel
Huygens Institute for the History of the Netherlands (KNAW)

Richard P. Smiraglia
Institute for Knowledge Organization and Structure, Inc.

Chapter 11
Knowledge Spaces
Visualizing and Interacting with Dimensionality[†††]

Abstract
Despite the full potential of visualizing linked data the representation of the semantic web (SW) is still very flat and static. We explore here the potential of visualizations of historical knowledge spaces for the SW beyond mere representations of organized knowledge as analytical instruments for knowledge interaction. Dimensionality is an aspect of even elementary knowledge structures. Conceptualizations and applications of knowledge spaces are analyzed with the focus on dimension extension in the classification, visualization and retrieval of knowledge from two-dimensional into three-dimensional knowledge spaces. Dahlberg's 1982 Information Coding Classification (ICC) represents an attempt to create a universal classification using conceptual relationships alongside the notational expressivity of existing classifications. The ICC is presented in a two-dimensional matrix of nine ontical structures and nine form categories. However, the two-dimensional matrix should be read as multi-dimensional. In 1975 Salton, Wong and Yang introduced the term "document space" as a multi-dimensional vector space model for automatic indexing. A similarly associative knowledge vector space was envisioned and visualized by Meincke and Atherton in 1976 encompassing the use of concept vectors, state vectors and representational vectors. The visualization of classifications in knowledge spaces as objects that organize and integrate knowledge is instrumental to interfaces in knowledge interaction. The history of knowledge organization and representation in combination with the spectacular recent affordances in information visualization should be included as strategies to bring more dimensionality to the SW.

1.0 Introduction: toward interactive, multi-dimensional visualizations of the Semantic Web

Despite the full potential of visualizing linked data in all its dimensions and dynamics—see for instance, Börner's (2010 and 2015) beautiful atlases of science and knowledge—it cannot be denied that the representation of the semantic web (SW) is still very flat and static. Knowledge domains and their (sub-)disciplines are represented as clusters in different color schemes. Hierarchical relationships are not visible and in order to establish the growth (let alone its dynamics) a new snapshot of the semantic web of a later moment is required. Here, we explore the potential of visualizations of historical knowledge spaces for the semantic web beyond merely representations of organized knowledge, but as analytical instruments for knowledge interaction. This exploration is driven by challenges within two semantic web projects under the supervision of the authors. In the first project: Digging into the Knowledge Graph, Richard Smiraglia, together with the co-applicants

[†††] The authors are grateful to Vanessa Schlais, University of Wisconsin-Milwaukee, for assistance with gathering sources. We are incredibly grateful to Kathryn La Barre and Pauline Cochrane for advice and reflection. We are indebted to Chiara Piccoli who on the basis of the author's instructions created the three-dimensional visualization of the ICC with Blender.

Andrea Scharnhorst (DANS, Netherlands) and Rick Szostak (University of Alberta, Canada), contextualizes Linked Open Data (LOD) of cultural artifacts of the humanities and social sciences (SSH) for inclusion in the semantic web. This, by identifying, evaluating and indexing SSH vocabularies and mapping clusters of similar meaning on to various Knowledge Organization Systems (KOSs), such as classifications. In the second one: Golden Agents: Creative Industries and the Making of the Dutch Golden Age with Charles van den Heuvel as principal investigator, a team of developers and researchers builds an infrastructure combining SW and AI (artificial intelligence) technologies to analyze interactions between the production and consumption of cultural goods and between the various branches of the creative industries of the Dutch Golden Age. While in the first project the focus is on the potential role of the organization and representation of concepts (Smiraglia 2014) in various classifications for knowledge organization (KO) of the SW (Figure 1a), in the second one the unstructured data of archival sources (notary acts, testaments, prenuptial agreements) with information about the consumption of cultural goods in the households of Amsterdam, needs to be aligned with the structured data of cultural heritage institutions on their production concepts and unstructured data are not the same, but have some properties in the process of knowledge production in common. Both in this process can be seen at the same time as preliminary and concretely instrumental. Concepts consist in the mind as a representation of comprehended or generalized information and as the formulation of a plan or action (Smiraglia 2014). Similarly, unstructured data represent information about the structured data in the making that successively can be captured and processed for further knowledge production. In Digging into the Knowledge Graph, concepts are mapped to various classifications to establish similarities in meaning, in the Golden Agents project AI is used for a pre-classification of the unstructured data (Figure 1b) (Baas, Dastani and Feelders 2019).

In short, in both cases knowledge interaction is required that combines the ontological with the epistemological to give meaning to the knowledge in the making. In order to provide meaning to knowledge there is a need for visualizations that allow perception of information from multiple perspectives and interaction with that knowledge. After a brief historical discussion of the three-dimensional visualizations in cubes, globes and combinations hereof by Paul Otlet as representations and perceptions of knowledge and as instruments of action, we discuss two spatial knowledge cubes in more detail. The first cube visualizes the potential of dimension extension to critically analyze the Information Coding Classification of Ingetraut Dahlberg. The second one is a model for an interactive knowledge vector space imagined by Peter Meincke and Pauline S. Cochrane (at that time under the author's name Pauline S. Atherton) to explore and retrieve closely and less related concepts.

2.0 Dimensionality in Knowledge Organization Systems

Knowledge organization systems (KOSs) are often conceived and designed as essentially flat, linear indexes, lists, of concepts. It is accepted that KOSs may be enumerative; that is, that they might include a term to enumerate every concept in a given domain. It also is accepted that classes are divided into divisions and subdivisions such that hierarchy pertains. This is true even in faceted (synthetic) systems, in which often the individual facets are made up of hierarchical enumerated lists of concepts. However, as we pointed out in

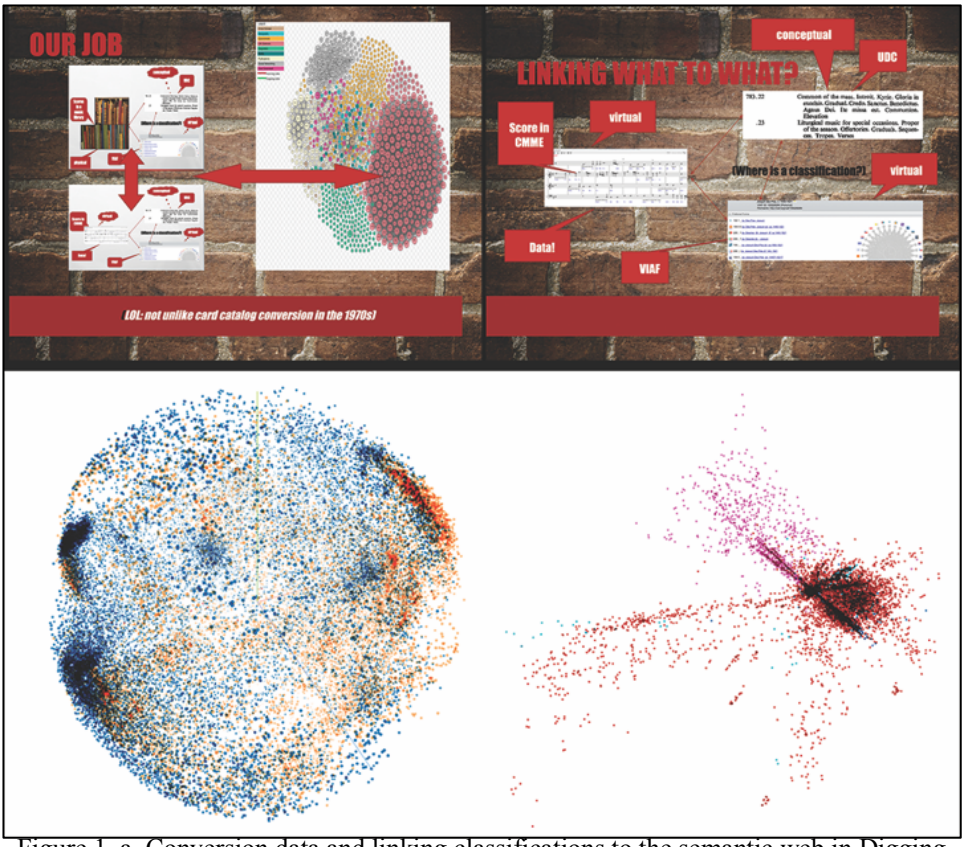


Figure 1. a. Conversion data and linking classifications to the semantic web in Digging into the Knowledge Graph project (top); b. Pre-classification of unstructured data with AI in Golden Agents project (bottom).

Smiraglia and van den Heuvel (2013, 374), dimensionality is an aspect of even elementary knowledge structures:

> The physical world is considered to be made up of objects in space and time, and knowledge of the physical world is embodied in representations …. Knowledge that is recorded can be organized using the entities named, and retrieval of that knowledge is facilitated by the syndetic pathways among them.

We discussed the role of dimensionality in the history of knowledge organization in various studies. In order to conceptualize and classify knowledge in spatial terms, mankind has been using the universe of knowledge metaphor since antiquity (Bliss 1929; van den Heuvel 2012). This metaphor was followed by the universe of concepts (Ranganathan 1957; Miksa 1992; Beghtol 2008) and the multiverse of knowledge (Smiraglia and van den Heuvel [2013 and 2011]; van den Heuvel and Smiraglia [2013]; Smiraglia, van den Heuvel and Dousa]2011]). Instead of positioning things (Richardson), Facts/Elements (Otlet), Facets (Ranganathan) or Concepts (Beghtol, Miksa) to organize knowledge in one universe, we used the metaphor of the multiverse to explore potential knowledge interaction. To this end we extended the metaphor of multiverse of knowledge to the laws of physics

in those spaces. The "gravitational forces" in these knowledge universes are used metaphorically to explain two important concepts in the theory of classification: "likeness" and "likeliness" (Hjørland 2003; van den Heuvel and Smiraglia 2010 and 2013). In this chapter we go beyond the metaphors of multidimensional knowledge universe and will focus on a more instrumental use of multidimensional knowledge spaces to organize and to interact with concepts. Although we do not intend knowledge spaces in metaphorical terms it is important to realize that we do not read them as "real" physiological spaces, but as visual spaces. As Rosar (2016, 531) pointed out this distinction is crucial in our perception and understanding of topological space and the role one of its properties, "dimensionality" herein. Historical examples of such instrumental use of such knowledge spaces for dimension reduction to facilitate information retrieval in the work of Otlet, Ranganathan, Farraday and Nelson were already discussed (van den Heuvel 2012). Here, other conceptualizations and applications of knowledge spaces will be analyzed. This time the focus will be on dimension extension in the classification, visualization and retrieval of knowledge from two-dimensional into three-dimensional knowledge spaces.

As early as 1864, Herbert Spencer (26) claimed that the sciences and their evolution were too complex to be classified in two dimensions. Paul Otlet visualized bibliographical classification and intellectual labor with a hierarchal order of the sciences as open folded cubes in his most important publication the *Traité de Documentation* (1934, 378 and 418). His lesser known book *Monde. Essai d'Universalisme. Connaissance du Monde, Sentiment du Monde, Action organisée et Plan du Monde* (1935) in which the UDC becomes an active instrument not only to organize knowledge, but the whole word in all its aspects does not have illustrations apart from some diagrams and symbolical letter and numerical formulae. However, the Archives of the Mundaneum in Mons (Belgium) preserve hundreds and hundreds of sketches scribbled by Paul Otlet in which the organization of, dynamics in and interaction with knowledge are represented in three dimensional globes, cubes or combinations thereof (van den Heuvel 2008 and 2012; Ducheyne 2009; Van Acker 2011). They not only visualize the order in the sciences and the relationships between the classes of the Universal Decimal Classification system, but also reveal a continuous striving to capture all sorts of metaphysical qualities ("nature, man, society, divinity and even the unknown") in an organized model and action plan for the world in multidimensional knowledge spaces.

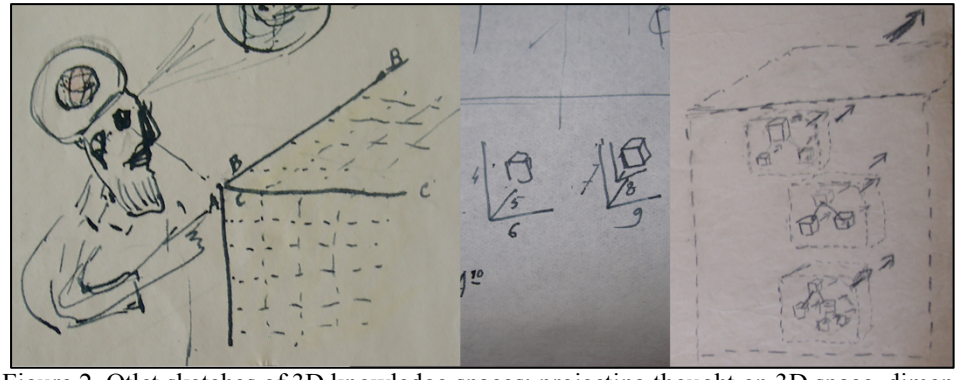


Figure 2. Otlet sketches of 3D knowledge spaces: projecting thought on 3D space, dimension expansion and capturing dynamics of 3D knowledge spaces.

It is important to note that in some of Otlet's visualizations the whole world is classified in the UDC and in others classification is outside "the Self (le Moi)" or "Societies (Sociétés)." This distinction seems to be related to the point of perception of reality. In some images "the Self" observes all knowledge of the world in a cube or globe from the outside; in others (s)he is positioned in the center. By taking an outside position, the self or societies are not classed objects themselves (yet), but related entities perceiving reality to be classed (Otlet 1896 transl. Rayward 1990, 64): "While a classification always involves the abstract point of view and deals with the objects in relation to each other, it is necessary to be aware that the two points of view constantly interact."

This awareness of perception and interactions between general points of view and relations between objects that Otlet expressed in sketches from the first decades of the 20th century onward, returns in his reflections on the evolution and future of documentation in later publications. Herein, Otlet anticipates a future development of documentation in which "Document-Instruments" and "Sense-Perception-Documents" are successive steps in a process of multidimensional "Hyper-documentation." The latter sensory-perception documents are fusions of things, ideas and associations, and comprehend not only images or sound, but also tactile, gustatory and olfactory documents. That what was unknown and imperceptible so far will become known and perceptible by the concrete mediation of the "Document-Instrument." This document developed in a previous stage (Otlet 1934, 429; van den Heuvel and Smiraglia 2013, 63) "intervenes to register directly the perception created by instruments. Documents and instruments at this point are linked in such a way that they are not two distinct things, but one single, The Document-Instrument." This notion that document and instrument can become one, that the organization of knowledge and the perception of, dynamics in and interactions with knowledge can become one, is important for the understanding of the two models that will be discussed here below: the Information Coding Classification ((ICC) developed by Ingetraut Dahlberg, and the Knowledge Space Vectors (KSV) by Peter Meincke and Pauline S. Cochrane.

3.0 ICC: Classification and operator

Dahlberg introduced the ICC formally in 1982, although it is clear from her writing that the underlying ideas arose from her dissertation research (Dahlberg 1974). It represents an attempt to create a universal (i.e., general) classification using what was for her "a new design philosophy" in combination with what had recently been learned about conceptual relationships from thesaurus construction alongside the notational expressivity of existing so-called universal classifications (Dahlberg 1982, 87). On the same page she elaborates: her new design philosophy should integrate "the application of the integrative level theory, the faceting of substructures of fields, the consistent application of concept relationships and the recurring arrangement of facets." ICC accomplishes this through its structural feature known as the "Systematifier," a sort of built-in facet operator, which Dahlberg rightly describes as (2008b, 170) "a feature known from systems theory, namely that … all hierarchical levels obey an element position plan … [such that] one can be sure to find specific concepts always at certain positions." She visualizes her ICC as a matrix, but as we will see later an extra dimension is brought in by describing an evolutionary process of integration.

Anyone who attended ISKO or other major classification conferences during Dahlberg's lifetime heard her present this classification one way or another. In almost every

instance, she presented a two-dimensional matrix, from which she read to us box by box about the nine classes and their systematic subdivision. The version from 2012 (145) is used here as Figure 3.

0 GENERAL FORM CONCEPTS	01 THEO-RIES; PRINCI-PLES	02 OBJECTS; COMPO-NENTS	03 ACTIVI-TIES PROC-ESSES	04 PROPER-TIES or 1st kind of field spe-cialty	05 PERSONS or 2nd kind of Field spe-cialty	06 INSTITU-TIONS or 3rd kind of field specialty	07 TECH-NOLOGY & PRODUC-TION	08 APPLICA-TION in other fields, DETER-MINA-TION	09 DISTRIBU-TION & SYN-THESIS
1 FORM & STRUC-TURE AREA	11 Logic	12 Mathematics	13 Statistics	14 Systemol-ogy	15 Organiza-tion Science	16 Metrology	17 Cybernetics, Control & Automation	18 Standardiza-tion	19 Testing & Monitoring
2 MATTER & ENERGY AREA	21 Mechan-ics	22 Physics of Matter	23 General & Technical Physics	24 Electronics	25 Physical Chemistry	26 Pure Chemis-try	27 Chemical Technol. & Engineer-ing	28 Energy Sci-ence & Technol-ogy	29 Electrical Engineering
3 COSMOS & EARTH AREA	31 Astron-omy & Astro-physics	32 Astronautics & Space Research	33 Basic Geo-sciences	34 Atmos-pheric Sciences & Technology	35 Hydro-spheric & Ocean. Sci-ence & Technology	36 Geological Sciences	37 Mining	38 Materials Science & Technology	39 Geography
4 BIO SPHERE	41 Basic Biological Sciences	42 Microbiol-ogy & Culti-vation	43 Plant Bi-ology & Cultiva-tion	44 Animal Bi-ology & Breeding	45 Veterinary Sciences	46 Agriculture & Horticulture	47 Forestry & Wood Sci-ence & Technology	48 Food Sci-ences & Technology	49 Ecology & Environment
5 HUMAN AREA	51 Human Biology	52 Health & Theor. Medicine	53 Pathology & Pract. Medicine	54 Clinical Medicine & Cure	55 Psychology	56 Education	57 Occupation, Labor &Leisure	58 Sports	59 Household & Home Life
6 SOCIETAL AREA	61 Sociology	62 State & Poli-tics	63 Public Admini-stration	64 Money & Finances	65 Social As-sistance, Appraisal & Survey	66 Law & Legal Science	67 Areal Planification & Urbanism	68 Structure of Defense	69 History Sci-ence & History
7 ECONOMY & TECH-NOLOGY AREA	71 General & National Econom-ics	72 Applied Economics, Business Mgt.	73 Technical Sciences	74 Mechanical & Precision Engg-	75 Building & Civil Engi-neering	76 Science of Commodities & Tech-noLofGoods	77 Vehicle Sci-ence & Technology	78 Traffic & Trans-port.Techn. & Services	79 Service Economics
8 SCIENCE & INFORMA-TION AREA	81 Science of Science	82 Information Sciences	83 Computer Science	84 Informa-tion in general	85 Communi-cation	86 Mass Communica-tion	87 Printing & Publishing	88 Tele-communica-tion	89 Semiotics
9 CULTURE AREA	91 Language & Linguis-tics	92 Literature & Philology	93 Music & Musicol-ogy	94 Fine Arts	95 Theatre	96 Culture Sci-ence (narrow sense)	97 Philosophy	98 Religion (in general)	99 Christian Religion & Theology

Figure 3. Dahlberg's ICC matrix (version 2012).

The ICC Matrix can be described as follows.

The first order on the vertical are the categories 1-9, which she names "ontical structures" (1982, 87) or categories of being (2008a, 161 and 163). These comprise three ontical groups:

I structure and matter
II living beings and
III products of man (artefacts).

Within these three groups, nine categories represent an elaboration of the ontical groups such that:

I 1. General forms and structures
2. Matter and energy
3. Aggregated matter (cosmos and earth)
II 4. Biological objects (micro-organisms, plants, animals)
5. Human Beings
6. Societal Beings
III 7. Material products of man and society (products of economy and technology)
8. Intellectual products of man and society (scientific, information and communication products)

9. Spiritual products of man and society (language, literature, music, arts, etc.)

The second order on the horizontal axis are categories of form which all together function as an operator of the categories of being in the facets which she names the Systematifier. Once again there nine of them which are divided in three groups or facets:

I 1 Theories, principles and general questions
2 Specific objects of a field and their components, kinds and properties
3 Specific operations in a field and their kinds and properties

The second three categories apply to the first three

II 4 Specific aspects/Properties of a field
5 Specific aspects/Properties of persons
6 Specific aspects/Properties of institutions

And third and final categories of form of the Systematifier deal with transdisciplinarity because knowledge is interconnected

III 7 Influences from other fields on the field in question, also its technology
8 Applications of methods and operations of a field in other fields
9 Synthesis and distribution of knowledge of a field

Despite the familiarity with Dahlberg's representation of the ICC within the KO community, what was rarely grasped at the aforementioned conferences was that the matrix in its construction was two-dimensional but should be read as multi-dimensional. This becomes clear from the theoretical underpinning of the ICC and the explanation of the matrix in many successive publications from 1982 until her death in 2017.

3.1 Dahlberg's theoretical underpinning of the ICC

Already from Dahlberg's first introductions of the ICC in 1982 it becomes clear that it cannot be represented in its full dimensions in a matrix (Dahlberg 1982, 87):

> It is easy to see that there is an evolutionary series from 1-6 and also that there is a division according to man's faculties in the three latter entity areas. It should also be evident that the entities of the levels presuppose each other or contain each other in a natural sequence. The last three levels (7-9) show the same evolutionary series with respect to the products of man and society applied on a matter-oriented, intellect-oriented and mental-spiritual oriented level. Thus, one can also say that the levels presuppose each other, they are "integrative."

Bringing in "evolutionary series" and "entities of the levels [that] presuppose each other or contain each other in a natural sequence" implies a development over time. This requires a reflection on dimensionality in the incorporation of integrative levels, in the faceting of substructures using recurring arrangements, and in the expansion of phenomena within classes according to concept relationships. This tension between fixed positions in a matrix and the concept of evolutionary series and natural sequences remains apparent in her later work, as we for instance can see both in the year 2008 in an article in *Axiomathes* (2008a) and an interview (2008b).

In the interview Dahlberg (2008b, 84) gave this succinct definition of the ICC:

> ICC is a fully faceted universal classification system of knowledge fields based not on disciplines but on universal ontical levels. It has fixed system positions at which interdisciplinary and transdisciplinary relationships to other fields of knowledge can be established according to a given rule. It also provides intra-relationship possibilities, i.e., combinations necessary for the expression of logical sentence structures within a field of knowledge.

This summary, brief as it is, is very expressive; with these words Dahlberg gets directly to her main points: a) ICC is based on ontical levels rather than academic disciplines; b) it has fixed positions that are derived systematically; and, c) it provides intra-relationship possibilities.

All these three components of the ICC can be combined in a two-dimensional matrix. However, that is not the case from other aspects of this classification described in her theoretical underpinning of the ICC in *Axiomathes* in the same year. In this article, Dahlberg provides a theoretical basis for the ICC which consists of the combination of:

1. The Integrative Level Theory following an evolutionary approach of ontical areas, and integrating on each level the aspects contained in the sequence of levels
2. The distinction between categories of being and categories of form
3. The application of a feature of System Theory (namely the element position plan)
4. The inclusion of a Concept Theory, distinguishing four kinds of relationship, originated by the kinds of characteristics (which are the elements of concepts to be derived from the statements on properties of referents of concepts).

Ad 1 The integration level theory

The integration of levels following an evolutionary approach is driven by the limitations of existing universal/bibliographical classification systems: the *Dewey Decimal Classification*, the Universal Decimal Classification (UDC), the Library of Congress *Classification*, the *Colon Classification* of Ranganathan, the Library Bibliographical Classification and the *Bliss Bibliographic Classification*, that according to Dahlberg all suffer from the major handicap that their main classes are either disciplines or groupings of disciplines that with the growing of number of disciplines inevitably results in the problem that number of main classes cannot be kept low.

Although Dahlberg considered the faceted classification the most helpful form of classification to overcome this problem this disciplinary approach had to be fundamentally reconsidered. For this she turns to lessons from philosophers and lessons from classificationists. From the philosophers Dahlberg selected and adapted the categories of being and categories of form; from the classificationists she borrowed specific features, such as general and special auxiliaries to enhance the operator functionalities of the Systematifier.

Ad 2 Categories of being and categories of form

Dahlberg's categories of being were in particular based on Aristotle's distinction in four levels of being: I dead being; II living being; III spiritual being; and, IV divine being. Furthermore, the Greek philosopher distinguished four groups with in total 9 form categories: I objects (1 substance), II Properties (2 Quality, 3 Quantity, 4 Relation), III Operations (5 Activity, 6 Passivity), IV Dimensions (7 Having Space, 8 Time, 9 Position). Based on the input of 20th century and contemporary German philosophers, Dahlberg created a new order of three groups of objects of beings on nine levels and a new order of three groups of form categories on nine levels. Aristotle's groups of being III and IV became part of a group of products of man (material, intellectual and spiritual). From the categories of form, Dahlberg left out Aristotle's group of Dimensions. For the dimensions space and time, she referred to other classifications, in particular the UDC, with elaborated schedules for spatial and temporal relationships (Dahlberg 2008, 173). The category position she based on System Theory.

Ad 3 System theory element: position plan

Dahlberg is far less elaborate on the position plan that she derived from systems theory and explains in a few sentences its essence that specific concepts can always be found at certain positions by the Systematifier device which by application is "fully faceted." For establishing the positions in this fully faceted system, Dahlberg makes use of concept theory.

Ad 4 The inclusion of Concept Theory

Dahlberg, when discussing concept theory in particular, focuses on relations and sequences of hierarchies of concepts (171):

> A further feature … is the introduction of a concept theory distinguishing among others between four kinds of concept relationships: the generic, the partitive, the complementary and the functional relationship … based on the characteristics which are the elements of concepts.

Moreover, she states that concepts are better grouped by their characteristics and by their characteristic relationships:

> The generic relationship is applied at all positions where hierarchies of concepts follow the ''is-a-kind of'' rule, the partitive one whenever a whole is to be subdivided into its parts—which occurs of course also at each subject group and its subdivisions into subject fields. The complementary relationship helps distinguish all kinds of oppositions which usually occur when forming a sequence of generic hierarchies.

Finally, Dahlberg claims that the system provides for an unlimited possibility of combinations between concepts in an organized way. She distinguishes four kinds of combinations:

1 external between different subject fields of an inter- and transdisciplinary type;
2 internal within a subject field only
3 form-categorial, when combinations become necessary with the form categories of the zero level; and,
4 additional with concepts from extra schedules for the individualizing concepts of space and time.

The first three kinds of combinations in principal should be easily made within the matrix. To the fourth one we'll turn later.

3.2 Working with the matrix and a virtual 3D model of the ICC

Dahlberg describes the matrix with the object and being in the knowledge fields and the working of the Systematifier on all nine levels and invites the reader "to accompany me box by box" (Dahlberg 2008a, 165 ff.). We quote from the first and last level:

"Level 1: Form and Structure Area" (165):

> It is the area the concepts of which can be applied in all the fields of the following levels. But why 11 Logic before 12 Mathematics? … it is simply because Logic, now still a sub-field of philosophy, is absolutely the most formal and theoretical field.

"Level 9: Culture Area" (170):

> [Which] begins with 91 Language and Linguistics, the mental faculty of mankind par excellence and one might ask, would it not logically belong at the beginning of everything? But Language has also a spiritual side, as the statements to be made with it can be true or false, and they can be even deliberately false.

Finally, she concludes, the entire system can be represented by the intersection of diagonals (170):

> Thus, the entire scheme reaches by its diagonal line from left to right from the most formal subject group of 11 Logic to the most ample subject group to the benefit of mankind under 99. If one would draw the counter diagonal from 19 to 91 the two lines would meet at 55 Psychology, which is the subject group concerning man's mental faculties and consciousness, the impulse giving and mover of all our efforts.

So far, the working of the ICC in principal can be read in a two-dimensional matrix. However, as for instances resulted from tests with the ICC used within the DANS project Knowledge Organization System Observatory (KOSo) in a comparative analysis with other classifications such as the Dutch National Academic Research and Collaborations Information System (NARCIS)(Coen and Smiraglia 2019; Coen, Smiraglia, Doorn and Scharnhorst 2019), to which we will turn later, several classes from the sciences and in particular from the social sciences and humanities did not have an ICC coding or could only be captured partly or indirectly—such as "fashion"—in combination with other codes (Coen and Smiraglia 2019, 346-7 and 348-9).

Apart from some missing or complex trajectories within the ICC, its two-dimensional form hinders the exploration of the "evolutionary series," and "entities of the levels [that] presuppose each other or contain each other in a natural sequence." Therefore, we developed a virtual 3D model of the ICC that we will discuss first before turning our attention to the four kinds of combinations dealing with concepts of space and time.

Dahlberg explains the evolutionary characteristics of the ICC as follows (171):

> The handling of any classification system is facilitated by the sense underlying its structures. It has been pointed out that the main sequence of general objects in ICC is a pattern of 3 x 3 areas in an evolutionary

sequence and that its ‘‘level character’’ makes sure that every area becomes a necessary presupposition for the existence of the following area. In the last area 9 all instances of the previous areas are contained. By using the concept theory mentioned above, one could say that the constituting characteristics of the concepts of objects of one level cumulate in the objects from one level to the next levels 1-6. The levels 7-9 build up on them, however, here, the characteristics ‘‘having life,’’ ‘‘having a soul’’ of level 4 to 6 are replaced by the characteristics ‘‘created by man,’’ ‘‘having a purpose given by man,’’ etc. Also, into the products of levels 7-9 the values are entered which man and society have implanted into them, since all products of mankind are dependent on knowledge, mastership and willpower of man, or, more precisely, the values which man and society put into their products determine their quality and durability.

Thus, the structure present is 3x3 followed by 6x6 and 9x9. The classification of categories of being only works in full once the process of evolutionary integration "based on the highest concepts possible" has been completed (and the level of spirituality has been reached). Matter exists before and can exist without beings, but once taken up by beings, can become products of various kinds, ultimately at the highest level reaching a spiritual plane.

In the direction of the Systematifier (categories of form) this seems to makes sense as well. Objects having no soul cannot apply something. In level 4 there are living beings such as plants and animals that also have products. Level 7 described by Dahlberg (163) as "material object products of mankind" presupposes the presence of human beings which happens on level 5. Level 8 cannot be used before the objects 4, 5 and 6 have been described. Applications (level 8 of categories of form) presupposes the presence of living beings. In this line of reasoning animals and human beings use tools, but stars do not. Described spirituality belongs to mankind and cannot come in place before level 5.

These sequences in both directions corresponds with the observation of Dahlberg that the diagonals also have a logical relationship within the ICC. To facilitate reading the evolutionary sequences of the ICC we enhanced its patterns 3x3 areas in the matrix three-dimensionally in such a way that the second layer of the sequence of 6x6 areas builds on the first layer and the third one of 9x9 areas on the first two.

It is important to realize that with this dimension expansion from a 2D to a 3D the evolutionary sequence of the ICC has been visualized, but that in reality we did not create a real 3D visualization that includes time (and space). The notation of the facets to create the latter visualization is lacking in the ICC. The dimension expansion provides an illusion of a three-dimensional space based on Dahlberg's description but not on the notation of the systems that describes the hierarchical relation of facets only in a two-dimensional way. In that regard it is similar to what in the data sciences often is referred to as the kernel-trick in space vector machines. That is, data are projected to a higher-dimensional space—in our case from a matrix into a series of successive expanding cubes—but in reality, the data are not mapped to a higher dimensional space, but act as though they were. Rosar explains (2016, 546): "In that dimensionality—the number of dimensions of a space—is a topological property of space, a space of a given dimensionality cannot be mapped continuously to one of a different dimensionality."

Most authors who have tried to come to an interpretation of the ICC have referred to the combination described here of four theories brought in by Dahlberg herself in her article in *Axiomathes* of 2008 to explain its "philosophical design." However, it is important to realize that the next paragraph is no less important for the understanding of the working of the ICC. In the paragraph on the combinatory functions, Dahlberg provides some disclaimers on the completeness of the ICC itself and refers to the additional value of other classifications systems to enhance it. For instance, Dahlberg explains (172):

Internal combinations which have been foreseen, however, have not as yet been elaborated, since the ICC consists so far only of subject fields, not of the subdivisions of them into theories, objects, activities, etc. except for the subject groups.

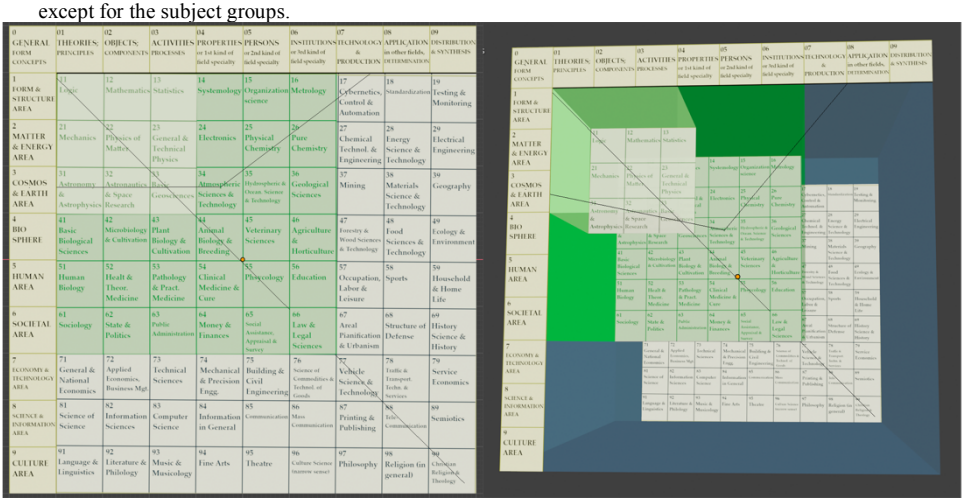


Figure 4. Dahlberg matrix and 3D reconstruction of ICC (virtual reconstruction by Chiara Piccoli).

And even more important for the interpretation of the evolution of the integrative characteristics of the ICC is her comment, that space and time are extremely well elaborated schedules in the UDC and should be recognized as a possible example when establishing additional schedules (173). Nevertheless, the use of these schedules are not the same at each level. Although it would be possible, for instance, to express the time of different historical periods in different parts in which particular artifacts (level of being 7) are produced, by using expressions of time in different calendar systems, on a higher level of the evolutionary integration with the spiritual product of man and society, such as literature or religion (level of being 9), classes for fictional time lack (compare van den Heuvel and Zamborlini 2021; chapter 6 in the present volume.)

In short, evolution can be observed in the sense of successive phases in our 3D cube, but we do not have any information about time and space in the model. For that reason, the evolutionary integration sits only partly within the ICC, but in order to represent it in time and space Dahlberg's model depends on future combinations with classifications outside the ICC. This also sheds a different light on the term "integration" as Dahlberg actually highlighted herself in the abstract of the article (161): "Further elaboration and use have been suggested, be it only as a switching language between the six existing universal classification systems at present in use internationally."

Dahlberg's assigned role of the ICC as a switching language between the six universal classification is not only of particular interest for the combinations with classificatory languages (Hutchins 1975), such as the UDC (explicitly named as a classificatory language cf. Otlet 1895-96; Smiraglia, van den Heuvel and Dousa 2011) and Ranaganathan's *Colon Classification*, but potentially also for queries in the context of a semantic web composed of machine-readable subject-predicate-object triples.

In short, to fully understand the integrative qualities of the ICC and its potential re-use in the context of the visualizing and interacting with its concepts in the context of the semantic web it should be analyzed in comparison with other classification systems.

3.3 The ICC and the SW

In October 2013, Hermann Bense a computer scientist from Dortmund, together with Dahlberg, explored the possibilities for structuring the semantic web with ICC codes. Knowledge fields to the first two levels and their possible subdivisions were included and visualized in a graphic representation (Figure 5) that can be found under Ontology4Us (https://www.ontology4.us/english/Ontologies/Science%2520Ontology/index.html).

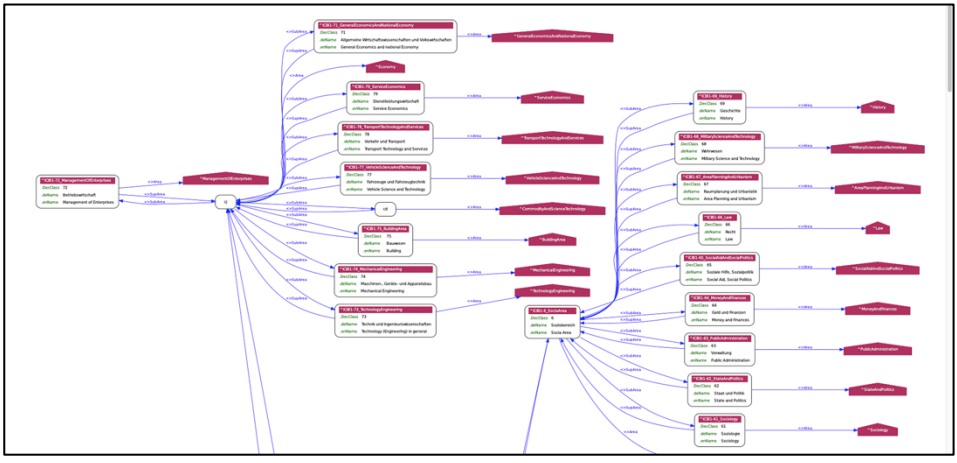


Figure 5. Representation of first two layers of ICC in SW model Ontology 4Us.

Although some hierarchies within the ICC can now be observed in an SW representation its tree-like structure is still rather flat. Moreover, the *Dewey Decimal Classification* is mentioned in the explanation of this figure but its relations are not visible. Therefore, the potential added value of the inclusion of functionalities from other classification systems remains unclear. More elaborate is the comparative analysis and visualization (albeit still in table form, see figure 1 bottom) of the classifications and KOSs with the ICC and the NARCIS classification that formed part of the aforementioned experiments with the DANS KOSo. The first experiment compared six knowledge organization system using NARCIS and ICC, three from the social sciences and humanities and ICONCLASS and three from the life sciences. This comparison explored the (mis-) matches of NARCIS and the ICC when applied to these knowledge organization systems based on the comparison of three of their characteristics:

1 coverage/precision of conceptual content;

2 population of the classes; and,

3 economy of classification.

For the conclusions in detail we need to refer to Coen and Smiraglia (2019, 352-3), but a general outcome was that despite some missing classes (such as for the aforementioned complex concept “fashion”), both the ICC and the NARCIS classification provide fairly precise coverage. The inflexibility of NARCIS makes it hard to express complex concepts

compared to the ICC and while the first provides more clarity and granularity in the representation of the sciences, the ICC allows for a better ontological structuring of human knowledge in general.

Successive tables visualize step by step matches and mismatches of concepts of the six KOSs between the NARCIS classification and the ICC. A second exploration included other examples of comparisons of the in total 132 KOSs in the DANS KOSo using the ICC and NARCIS. The outcome of this experiment was that the ICC once again showed "multidimensional flexibility for classifying knowledge" (Coen, Smiraglia, Doorn and Scharnhorst 2019, 13), while the NARCIS classification showed accuracy in the representation of the scientific fields and classes. This is interesting for the expressed aim to design a workflow to support researchers and cultural heritage institutions in data curation in order to enable them to find useful KOSs for their research questions and to assess the quality of submitted data collections. By the combination of the two classifications the DANS KOSo could be used first to explore a conceptual knowledge space in a more associative way and successively used for scientific positioning of the classes to assist end users in information retrieval. Although the authors speak of the multidimensional flexibility of the ICC, the comparison of the coverage, population and hospitality of the various knowledge systems have been visualized in separate two-dimensional tables which makes it hard to switch between and to interact with the various orders of the concepts to explore and annotate similarities and differences between concepts.

4.0 An associative knowledge space for information retrieval based on perception

In their seminal article Salton, Wong and Yang (1975) introduced the term "document space" as a multi-dimensional vector space model for automatic indexing. They explain that creating a document space that after a query distiguishes between relevant and non-relevant items is hindered by the lack of information about what relevance assessments will be made by the users over the course of time (614): "That is, the optimum configuration is difficult to generate in the absence of a priori knowledge of the complete retrieval history for the given collection." For that reason, they state that the next best thing is to achieve a maximum possible separation between the individual documents or between documents grouped into classes. In this way each document in groups of documents may be retrieved when located close to a user query without also necessarily retrieving its neighbours (614-5).

Gerard Salton (1968; 1971) had already experimented with the concepts of document space and document vectors in his SMART retrieval environment. Document vectors are all of the same dimensionality and this is determined by the number of possible concepts. Queries presented to SMART are also transformed into vectors of concepts (a query vector). The relevance of an information item to a particular request provided by the system is accomplished by determining the similarity of the query vector to each of the document vectors. Although such space vectors work well for the automatic indexing of concepts to which we give meaning afterwards, such as for use in topic modeling, the retrieved items still need to be compared with other items. In the case of the unstructured metadata in the aforementioned Golden Agents project meaning is given in a process of data-alignment using the Lenticular Lenses tool by comparing them to all variants with the "ground truth" of the structured data (Idrissou et al. 2018 and 2019) while in the case of the Digging into the Knowledge Graph project the concepts are mapped in a process of ontology-alignment

in the DANS KOSo to various classifications to establish their similarities in different epistemological contexts in order to find their proper meaning for use. This implies that differently from the space vector model of Salton, Wong and Yang focused on automatic indexing we actually need to be able to associate (and that implies to be able to see) *a priori* information and knowledge respectively of closely related metadata and concepts. This need for association, rather than indexing was already expressed by Vannevar Bush in 1945 in his famous article "As we May Think (§5):

> Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing …. The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts .... Selection by association, rather than indexing, may yet be mechanized"

Such an associative knowledge vector space was envisioned and visualized by Peter Meincke and Pauline Atherton (Cochrane) in 1976 (20):

> concepts could be thought of as having direction in the sense that they may be orthogonal to each other, if totally unrelated, and projecting in nearly the same direction in space, if closely related. The concept vectors span all of knowledge space.

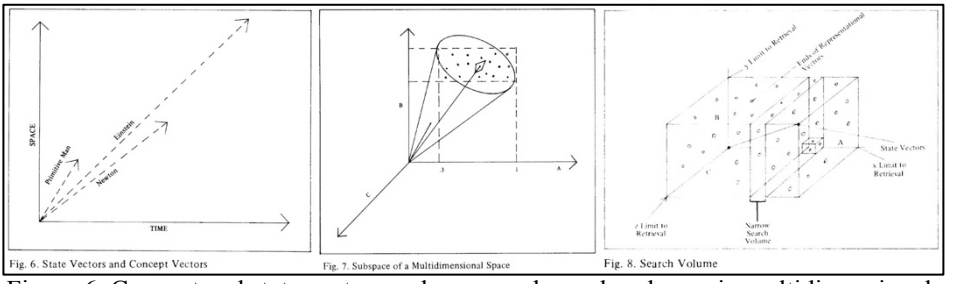


Fig. 6. State Vectors and Concept Vectors Fig. 7. Subspace of a Multidimensional Space Fig. 8. Search Volume

Figure 6. Concept and state vectors; subspace and search volumes in multidimensional knowledge space (Meincke and Atherton 1976).

It encompassed the use of three different sorts of vectors in a multidimensional space that they visualized as a cube: 1) concept vectors for a field of knowledge; 2) state vectors based on a person's understanding of these concepts; and, 3) representational vectors for information items that in a retrieval system might cover a flexible sub space of knowledge. The user of the retrieval system can expand or reduce the subspaces with the representational vectors with components on basic concept vectors similar to his state vector.

In addition to Vickery's (1960) seven methods of classification and indexing (increasing in degree of control) they proposed an additional eighth: the assignment of representational vectors as in multidimensional knowledge space with components on basic concept vectors (Meincke and Atherton 1976, 19 and 23). The presence of state vector based on a person's understanding of these concepts presupposes differently from the knowledge vector space of Salton, Wong and Yang, *a priori* knowledge. This *a priori* knowledge in the knowledge space model of Meincke and Atherton (1976, 21) could in their view, for instance, be provided by the Selective Dissemination of Information (SDI) profile, referring to the (new) resources to keep the user informed on specific topics. Although the state vector seems to reflect the state of understanding of the user of a specific concept at a certain moment in time, Meincke and Atherton foresee an incremental role of the state vector, based on the user's learning (20-21):

> Learning a new concept can be thought of as adding to the state vector in knowledge space (or rotating it), so that the person develops a component on a concept which he has just learned. This is similar to describing the change of state of an atom as change in the direction of its state vector in a multidimensional space.

Michael J. McGill set up an experiment at Syracuse University in which Meincke and Atherton's knowledge space model was compared with Salton's SMART information retrieval system to establish the implications for the knowledge vector space of the latter system if the search volumes could be increased or decreased as in the Meincke and Atherton model. A first conclusion was that the change of the volume and selection of the projection schemes within the knowledge space, whether implicit or explicit, would have impact on the performance of a retrieval system. This implies in McGill's view (209) that: "a system designed to recognize and take advantage of the dimensionalities could potentially offer enhanced performance to the user."

A second conclusion was that the state vector should be dynamic and likely to move in a given direction within the knowledge space over a specified period of time. Furthermore, McGill suggested establishing a point of view with respect to an SDI profile, i.e., that the point of view might establish not only the location of an individual within the knowledge space—but similar to what Otlet had claimed as mentioned above—the context associated with that point. Similar to the DANS KOSo that made use of the ICC and the NARCIS classifications, McGill (1976, 209) envisioned that one projection scheme, for instance, would correspond well on a generic level to represent a survey of the available literature within a field, and another one might be used as well for pinpointing a single extremely informative item.

The most important implication of the enhancement of the SMART system with features of the Meincke and Atherton model is, according to McGill (210), its potential for the comparison and evaluation of very different systems that allow for refinements for a future SMART system. However, such comparison is only possible if we take differences in the nature of knowledge spaces (visual spaces or phenomenal spaces with no metrical properties versus mathematical spaces based on collections of points) and how we perceive them (Rosar 2016, 543-5). The full implications of mixing conceptual spaces with physical metaphors and cognitive methods based on perception of those spaces in classification, on the one hand, with geometrical spaces analysed with mathematic-analytical methods in IR, on the other hand, should be an object for future studies.

5.0 Epilogue: Visualizing knowledge spaces and the SW

Full integration of knowledge in one system or network is a Utopian dream. The classifications of the UDC and ICC that we discussed are just a few examples of attempts in a very long history to organize the knowledge of the world. Despite its enormous growth, it is unlikely that the SW will succeed in fulfilling this dream completely either. For that reason, it seems that Licklider's prediction in his 1965 *Libraries of the Future* (78) that we shall have to content ourselves with partial models of the universe, based on geometry, logic or natural language, is valid for the SW as well. The need for multiple models seems also to be acknowledged in the discussions about the future of the SW.

Similar to Dahlberg's attempt to seek (the evolution of) integration within the system and in connecting to other systems, within the SW community (historical) links are explored between knowledge graphs and knowledge networks and their implications for AI

(Sheth, Pahdee and Gyrard 2019). Based on use-cases these authors individuated the following "emerging" challenges for the development of knowledge graphs:

1) Capturing context;
2) Domain specific knowledge extraction;
3) Knowledge Alignment;
4) Real Time Knowledge Graphs for Fast Data;
5) Quality and Validity of Knowledge Graphs; and,
6) Adaptive Knowledge Networks (able to adapt to multimodal spatiotemporally evolving data with change of time).

More specific to the thoughts of the SW community about its future requirements is a recent report of a Dagstuhl Seminar held in 2018 with the title "Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web" (Bonatti et al. 2019). In particular, the chapter "grand challenges" (21-24) brought forward by Paul Groth, Frank van Harmelen et al. is quite revealing. It lists:

1) challenges in the development of knowledge and data models at scale that can deal with the V's: volume, variety, veracity and velocity;
2) challenges in representing knowledge that can capture all forms of knowledge including ambiguous, incomplete, biased, approximated and context-specific knowledge, that acknowledge the evolution and change of events, languages and entities and that can reconcile symbols and sub-symbols in knowledge graphs and machine learning;
3) challenges in access and interoperability dealing, for instance, with a knowledge universe with different semantics for both humans and AI, that is open to the public according to the FAIR principles and finding methods for re-indexing knowledge as part of creating an infrastructure that facilitates repurposing global knowledge; and finally,
4) challenges in creating applications that allow for answering sophisticated questions over heterogenous knowledge graphs, translating knowledge into action and creating knowledge graphs as socio-technical systems that, for instance, represent what is missing, viewpoints and opinions and that allow knowledge graphs to be interfaces between humanity and machines and between machines and machines for humanity knowledge sharing.

When we compare these future strategies for knowledge representation on the SW with the historical cases that we explored we can observe that despite the technical differences most of these challenges can be recognized. A second observation that can be made is that in the strategies for dealing with knowledge representation, a potential instrumental role of visualization is lacking.

We discussed the visualization of classifications in knowledge spaces as objects that organize and integrate knowledge and that at the same time are instrumental as interfaces in knowledge interaction. Otlet tried to capture the knowledge of the world in multidimensional knowledge spaces, but struggled with representing its growth and dynamics hindered by the paper format of his sketches. A representation of Dahlberg's ICC matrix in a virtual three-dimensional knowledge cube taught us how visualization can play a critical role in assessing claims about the quality and evolution of systems in which knowledge is organized. Finally, the illustration of the Meincke and Atherton knowledge space made clear that knowledge interaction can be simulated. This simulation of knowledge interaction could be twofold: 1) as an automated machine-machine interaction between concepts based on likeliness, as resulted from McGill's discussion of the extension of the SMART system for automatic indexing with the Meincke and Atherton model; 2) and as human-computer

interaction with concepts based on the associations of users and their state of perceiving and understanding specific concepts at a certain moment in time and by adapting the search volumes (and consequently the dimensionality) for information retrieval.

Unfortunately, only a few historical examples of knowledge spaces could be discussed to explore their potential for a future agenda for knowledge representation on the SW. It was impossible in the context of this study to include the many ways visualization tools are used, particular in information studies and data-science, to align ontologies and metadata, to represent completeness, ambiguity, uncertainty and heterogeneity in (meta-) data and/or relevant contextual information presented as some of aforementioned grand challenges of the SW. Nevertheless, it can be argued that the history of knowledge organization and representation in combination with the spectacular recent affordances in information visualization should be included as strategies to bring in more dimensionality in the SW.

References

Baas, Jurian, Mehdi Dastani and Ad Feelders. 2019. "Graph Embeddings for Enrichment of Historical Data." In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* Wurzburg 16-20 September 2019.

Beghtol, Clare. 2008. "From the Universe of Knowledge to the Universe of Concepts: The Structural Revolution in Classification for Information Retrieval." *Axiomathes* 18: 131-44.

Bense, Hermann and Bastian Haarmann. 2013. "A Richer Notation for the Representation of Ontological Knowledge." In *ICCESSE 2013: International Conference on Computer, Electrical, and Systems Sciences, and Engineering, London, 2013* http://o4.cms2web.com/download/publications/A_richer_Notation_for_the_Representation_of_Ontological_Knowledge_07_2013.pdf

Bliss, Henry Evelyn. 1929. *The Organization of Knowledge and the System of the Sciences*. New York: H. Holt and Co.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know.* Cambridge, Mass.: The MIT Press.

Börner, Katy. 2015. *Atlas of Knowledge: Anyone Can Map.* Cambridge, Mass.: The MIT Press.

Bonatti, Piero Andrea, Michael Cochez, Stefan Decker, Axel Polleres and Valentina Presutti. 2019. "Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web. Report Dagstuhl Seminar 18371." *Dagsthul Reports* 8, no. 9:1-92. https://aic.ai.wu.ac.at/~polleres/publications/bona-etal-DagstuhlReport18371.pdf

Bush, Vannevar. 1945. "As We May Think." *Atlantic Monthly* 176, July: 101-8.

Coen, Gerard and Richard P. Smiraglia. 2019. "Toward Better Interoperability of the NARCIS Classification." *Knowledge Organization* 46: 345-53.

Coen, Gerard, Richard P. Smiraglia, Peter Doorn and Andrea Scharnhorst. 2019. "Classifying KOSs: The Use of Dahlberg's ICC in the DANS KOS Observatory (KOSo)." Paper presented at ISKO-LC 2019 (ISKO Low Countries Conference 2019) Morsels of Knowledge, Brussels 20-21 June.

Dahlberg, Ingetraut. 1974. *Grundlagen universaler Wissensordnung. Probleme und Möglichkeiten eines universalen Klassifikations-systems des Wissens*. Pullach bei München: Verlag Dokumentation.

Dahlberg, Ingetraut. 1982. "ICC—Information Coding Classification: Principles, Structure and Application Possibilities." *International Classification* 9: 87-93.

Dahlberg, Ingetraut. 2008a. "The Information Coding Classification (ICC): A Modern, Theory-based, Fully-faceted, Universal System of Knowledge Fields." *Axiomathes* 18: 161-76. DOI 10.1007/s10516-007-9026-8

Dahlberg, Ingetraut. 2008b. "Interview with Ingetraut Dahlberg, December 2007" [by Ia C. MicIlwaine and Joan S. Mitchell]. *Knowledge Organization* 35: 82-85.

Dahlberg, Ingetraut. 2012. "A Systematic New Lexicon of All Knowledge Fields based on the Information Coding Classification." *Knowledge Organization* 39: 142-50.
Ducheyne, Steffen. 2009. "'To treat of the world.' Paul Otler's Ontology and Epistemology and the Circle of Knowledge." *Journal of Documentation* 65: 223-44.
Hjørland, Birger. 2003. "Fundamentals of Knowledge Organization." *Knowledge Organization* 30: 87-111.
Idrissou, Al, Veruska Zamborlini, Chiara Latronico, Frank van Harmelen and Charles van den Heuvel. 2018. "Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses." In *DH Benelux Conference 6-8 June 2018, International Institute for Social History, Amsterdam.* http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/Al-Idrissou-Chiara-Latronico_GoldenAgentsLenticularLenses_DHBenelux2018.pdf
Idrissou, Al, Veruska Zamborlini, Frank van Harmelen and Chiara Latronico. 2019. "Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers." In *K-CAP '19: Proceedings of the 10th International Conference on Knowledge Capture*, *September 2019*. New York: ACM, 259-62. https://doi.org/10.1145/3360901.3364440
Licklider, Joseph Carl Robnett. 1965 *Libraries of the Future*. Cambridge, Mass.: The M.I.T. Press.
McGill, Michael J. 1976. "Knowledge and Information Spaces: Implications for Retrieval Systems." *Journal of the American Society for Information Science* 27: 205-10.
Meincke, Peter M. and Pauline Atherton. 1976. "Knowledge Space: A Conceptual Basis for the Organization of Knowledge." *Journal of the American Society for Information Science* 27:18-24.
Miksa, Francis L. 1992. "The Concept of the Universe of Knowledge and the Purpose of LIS Classification." In *Classification Research for Knowledge Representation and Organization: Proceedings of the 5th International Study Conference on Classification Research(FID),* ed. Nany Joy Williamson and Michéle Hudon.
Otlet, Paul. 1896. *Règles pour les développements à apporter à la Classification Décimale.* Bruxelles: Office Internationale de Biliographie.
Otlet, Paul. 1934. *Traité de documentation: le livre sur le livre: théorie et pratique*. Bruxelles: Editiones Mundaneum, Palais Mondial.
Otlet, Paul. 1935. *Monde, essai d'universalisme: connaissance du monde, sentiment du monde, action organisée et plan du monde.* Bruxelles: Editiones Mundaneum/D.van Keerberghen&Fils.
Ranganathan, Shiyali R. 1957. *Prologomena to Library Classification.* 2nd ed. London: The Library Association.
Rayward, W. Boyd. 1990. *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet,* trans. and ed. with an introd. by W. Boyd Rayward. FID 684. Amsterdam: Elsevier.
Rosar, William H. 2016. "The Dimensionality of Visual Space." *Topoi* 35: 531-70.
Salton, George. 1968. *Automatic Information Organization and Retrieval.* New York: McGraw-Hill.
Salton, George, ed. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, NJ: Prentice- Hall.
Salton George, Andrew Wong and Chungshu Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18, no. 11: 613-29.
Seth, Amit, Swati Padhee and Amelie Gyrard. 2019. "Knowledge Graphs and Knowledge Networks: The Story in Brief." Knoesis Wiki: *IEEE Internet Computing, July-Aug* https://www.researchgate.net/publication/339814197_Knowledge_Graphs_and_Knowledge_Networks_The_Story_in_Brief
Smiraglia, Richard P. 2014. "The Concept of Concepts: A Case Study from *American Documentation*." In *Connecting Across Borders: Globalization and Information Science: Proceedings if the Canadian Association for Information Science Annual Conference, St Catherine's, Ontario, May 28-30*, ed. M. Griffis, H. Julien and L. Given http://www.cais-acsi.ca/ojs/index.php/cais/issue/view/29

Smiraglia, Richard P. and Charles van den Heuvel. 2013. "Classifications and Concepts: Towards an Elementary Theory of Knowledge Interaction." *Journal of Documentation* 69: 360-83.

Smiraglia, Richard P., Charles van den Heuvel and Thomas M. Dousa, 2011. "Interactions Between Elementary Structures in Universes of Knowledge." *Classification & Ontology: Formal Approaches and Access to Knowledge: Proceedings of the International UDC Seminar 19-20 September 2011, The Hague, Netherlands*, ed. Aida Slavic and Edgardo Civallero. Würzburg: Ergon Verlag, 2011, 25-40.

Smiraglia Richard P. and Charles van Heuvel. 2011. "Idea Collider: From a Theory of Knowledge Organization to a Theory of Knowledge Interaction." *Bulletin of the American Society of Information Science and Technology* 37, no. 4: 43-7.

Spencer, Herbert. 1864. *The Classification of the Sciences: Which are Added Reasons for Dissenting from the Philosophy of M. Comte.* New York: D. Appleton and Co.

Van Acker, Wouter. 2011. "Universalism as Utopia. A Historical Study of the Schemes and Schemas of Paul Otlet (1868-1944*).*" PhD diss., Ghent University.

Van den Heuvel, Charles. 2008. "Building Society, Constructing Knowledge, Weaving the Web. Otlet's Visualizations of a Global Information Society and his Concept of a Universal Civilization." In *European Modernism and the Information Society*, ed. W. B. Rayward. London: Ashgate Publishers: 127-53.

Van den Heuvel, Charles. 2012. "Multidimensional Classifications: Past and Future Conceptualizations and Visualizations." *Knowledge Organization* 39: 446-60.

Van den Heuvel, Charles and Richard P. Smiraglia. 2010. "Concepts as Particles: Metaphors for the Universe of Knowledge." In *Paradigms and Conceptual Systems in Knowledge Organization: Proceedings of the Eleventh International ISKO Conference, 23-26 February 2010 Rome Italy*., ed. Cladui Gnoli and Fulvio Mazzocchi. Würzburg: Ergon-Verlag: 50-6.

Van den Heuvel, Charles and Richard P. Smiraglia. 2013. "Visualizing Knowledge Interaction in the Multiverse of Knowledge." In *Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands*, ed. Aida Slavic, Almila Akdag Slah and Sylvie Davies. Würzburg: Ergon Verlag, 2013, 59-72.

Van den Heuvel, Charles and Veruska Zamborlini. 2019. "Storifying Data: Modeling Historical Processes in Knowledge Graphs: Kubler's Shape of Time Revisited." In ADHO LOD SIG Pre Conference workshop. Ontologies for Linked Data in the Humanities–DH2019 Conference Utrecht 2019, 8 juli 2019. https://cwrc.ca/islandora/object/islandora%3A82042a5e-35bd-4986-8465-429dea5ae64e

Vickery, Brian Campbell. 1960. *Faceted Classification.* London: Aslib.

Ronald Siebes
Data Archiving & Networked Services (DANS)

Gerard Coen
Data Archiving & Networked Services (DANS)

Kathleen Gregory
Data Archiving & Networked Services (DANS)

Andrea Scharnhorst
Data Archiving & Networked Services (DANS)

Chapter 12
Publishing Linked Open Data
A Recipe‡‡‡

Abstract
Linked Open Data (LOD) are inherently interoperable and have the potential to play a key role in implementing interoperability. They offer great promise in helping to achieve semantic interoperability, which relies on linking data via common vocabularies or knowledge organisation systems. This document attempts to demystify LOD and presents "Ten Things" to help anyone wanting to publish LOD. We include visualisations, suggest readings and activities, and highlight other projects to make this guide understandable and usable for people across disciplines and levels of expertise.

1.0 Audience
We aim to provide a document which is understandable to non-experts, but that also provides specific technical references and does not downplay some of the complexities of LOD. Our target audiences therefore include:

- •Researchers (especially from the social sciences & humanities)
- •Anyone interested in publishing Linked Open Data (LOD)
- •Anyone interested in supporting use of LOD in research

2.0 Introduction
Linked Open Data (LOD) are inherently interoperable and have the potential to play a key role in implementing the "I," interoperability, in the FAIR data principles (Wilkinson et al. 2016). They are machine-readable, based on a standard data representation and are seen as

‡‡‡ We are grateful for pointers from and discussion with the DANS Research group, in particular Herbert van de Sompel, and for the valuable contributions of Esther Plomp, Marjan Grootveld, Stefan Dietze, Maria Poveda Villalon, Beyza Yaman and Enrico Daga who commented on earlier drafts of this document. This work has been informed by the "GO FAIR Implementation Network Manifesto: Cross-Domain Interoperability of Heterogeneous Research Data (Go Inter)," ed. by Peter Mutschke, https://www.go-fair.org/implementation-networks/overview/go-inter/ . The grants "Digging into the Knowledge Graph" (http://di4kg.org/, TAP-NWO Grant 463.17.005), "Re-search: Contextual search for scientific research data" (NWO Grant 652.001.002) and FAIRsFAIR "Fostering FAIR Data Practices In Europe" (European Union Horizon 2020 project call H2020-INFRAEOSC-2018-2020, grant agreement 831558) have enabled part of this work.

epitomizing the ideals of open data (see https://5stardata.info/en/). They offer great promise in helping to achieve a specific type of machine-executable interoperability known as semantic interoperability, which relies on linking data via common vocabularies or knowledge organisation systems (KOSs). This document attempts to demystify LOD and presents "Ten Things" to help anyone wanting to publish LOD.

Although this list of "Things" is presented in a roughly linear order, preparing and publishing LOD are iterative processes. Expect to go back and forth a bit between the Things, and take the time to double check that your progress matches your desired end result. Some "Things" can be executed in parallel; you will also notice recurring themes (e.g., sustainability and licensing concerns) that need to be considered throughout the workflow. As with any formal representation of a research process, the seeming sequence of this list can be best envisioned as an iterative and often also messy process (Beaulieu et al. 2013).

These "Things" are based on our own practical experiences in publishing LOD in various interdisciplinary settings, e.g., the Digging into the Knowledge Graph project (http://di4kg.org). Our goal is to complement existing scholarly reports on LOD implementations (e.g., Hyvönen 2012; Hyvönen 2020; Meroño-Peñuela et al. 2019), other workflow models (see https://www.w3.org/TR/ld-bp/#PREPARE), and the authoritative "Best Practices for Publishing Linked Data" of the W3C, which we cross-reference (as W3C Step #X) wherever appropriate (W3C 2014). We include visualisations, suggest readings and activities, and highlight other projects to make this guide understandable and usable for people across disciplines and levels of expertise. However, it is important to note that semantic web technology is a complex scientific field; you may need to consult a semantic web expert along the way.

3.0 Overview of Ten Things

Thing 1: Learning: Understand and practice the Semantic Web and LOD basics.
Thing 2: Exploring: Inventory of your data.
Thing 3: Defining: Define the URI (Uniform Resource Identifier) naming strategy.
Thing 4: Resolving: Consider resolvability when a person or machine visits the URI.
Thing 5: Transforming: Generate the URIs for the selected concepts and relations according to the URI naming strategy.
Thing 6: Mapping: Map your Linked Data from your newly defined namespace to similar concepts and relations within the LOD.
Thing 7: Enriching: Enrich your data with information from the LOD.
Thing 8: Exposing: Define how people can get access to your LD: a data dump, a SPARQL endpoint, or a Web API.
Thing 9: Promoting: Publish and disseminate the value of your data via visualisations and workflows.
Thing 10: Sustaining: Ensure sustainability of your data.

4.0 Things

4.1 Thing 1: Learning: Understand and practice the semantic web and LOD basics

Semantic web technology (which underlies LOD) is complex. It requires not only a new data model (Resource Description Framework, or RDF), but also infrastructures for storing and linking data as well as algorithms for retrieving, enriching and reasoning across those data.

Understanding LOD begins with understanding the Resource Description Framework. RDF is a standard format defined by the World Wide Web Consortium (W3C,

https://www.w3.org) that can be easily interpreted by machines. RDF statements are called "triples" because they contain three pieces:

the subject : predicate : object.

RDF data are modelled as a "labeled graph" which links descriptions of resources together. Subjects and objects are nodes, predicates are links. RDF is notated as a list of these statements (the triples) that describe each piece of knowledge in your dataset. This list of statements can be thought of as a large indexing file to your data.

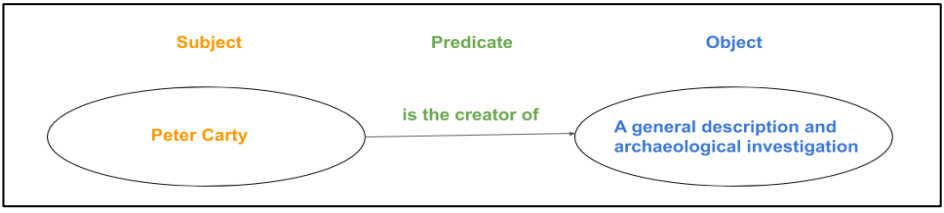


Figure 1. An example of an RDF triple.

There are different formats for creating RDF. Popular formats include RDFa (https://www.w3.org/TR/rdfa-primer/), RDF/XML (https://www.w3.org/TR/rdf-syntax-grammar/), Turtle (https://www.w3.org/TR/turtle/) and N-Triples (https://www.w3.org/TR/n-triples/). Although these formats are slightly different, the meaning of the RDF statements written with them remains the same. In our examples, we use the Turtle format and wrote the code using Atom (https://atom.io), a collaborative text editor. (See W3C Step #3: https://www.w3.org/TR/ld-bp/#MODEL).

Figure 2 shows a section of RDF representing an example dataset, which we use throughout the text.

```
1
2  # Ourxiv Records
3
4  # Prefixes
5  @prefix dc: <http://purl.org/dc/elements/1.1/> .
6  @prefix rds: <http://rdfs.org/ns/void#> .
7  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
8  @base <http://www.ourxiv.com/ui/datasets/id/ourxiv-dataset> .
9
10 ourxiv:98261 dc:identifier "ourxiv:98261" .
11 ourxiv:98261 dc:creator "Peter Carty" .
12 ourxiv:98261 dc:title "A general description and archaeological investigation" .
13 ourxiv:98261 dc:date "2002-08-18" .
14 ourxiv:98261 dc:coverage "Dublin" .
15 ourxiv:98261 dc:type "Text" .
```

Figure 2. A screenshot of an RDF representation.

The graph structure of RDF offers benefits over typical database structures. Creating new subjects and predicates is far less tedious than creating new fields and linking tables, as is common in database design. Storage in RDF is also more compact. Perhaps most importantly, RDF enables specific ways of questioning your data that are not possible with other structures. In a triple, the predicates (the links between the nodes in Figure 1) also have meaning and thus are semantically encoded; this facilitates executing more complex

operations (known as "semantic reasoning") on the graph. In our example (see Figures 1 and 2), the role of Peter Carty as creator of the dataset is spelled out, and so can be differentiated from other possible roles, such as being a contributor or a collaborator. In the end, RDF is simply another way of expressing your data.

Activity: Who better to introduce you to the concepts of LOD than Tim Berners-Lee, founder of the World Wide Web? View these videos for an overview of the topic:

- Tim Berners-Lee on "The Next Web": http://www.ted.com/talks/tim_berners_lee_on_the_next_web
- Tim Berners-Lee at the GOV 2.0 expo ("bag of crisps"): https://youtu.be/ga1aSJXCFe0
- Then, check out how Google uses this same technology in the Knowledge Graph: https://www.youtube.com/watch?v=mmQl6VGvX-c

Putting data into RDF is one ingredient in working toward LOD. RDF statements must also be expressed as Uniform Resource Identifiers (URIs, see Things 3 through 6) in order to link them to other data. It is possible to have linked data (LD) living on internal servers that are not a part of the larger LOD Cloud. In order to publish data on the LOD Cloud, URIs must be readable, or resolvable (see Thing 4), not only internally, but also to outside sources.

Activity: Visit the Linked Open Vocabularies (LOV, https://lov.linkeddata.es) and see the different types of things that can be linked to. Eventually you will be exposing and/or mapping your data to some of these vocabularies and schemas (see Thing 7). Can you identify datasets that contain concepts similar to those in your own dataset? (For example, if you have data about cities, you may want to look for datasets with information about cities, e.g. DBpedia).

Many of the terms which we use in this document (i.e. schemas, ontologies, and concepts) have a specific meaning within computer science that are different from how these terms are used in other disciplines. When we discuss "schemas" and "ontologies" here, we do so very broadly to refer to formal models used to order knowledge.

It is important, however, to understand the term "concept" in more detail. Here, as in computer science, concept refers to a class, i.e., a knowledge representation (of objects, individuals, actions, etc.). Concepts are essentially abstractions made to order things. They are represented by terms (Dextre Clarke 2019). An ensemble of concepts is often represented in the form of controlled vocabularies, schemas, or ontologies, more generally known as KOSs). (For a recent discussion on the role of KOSs in data curation see Scharnhorst et al. 2019). This means that, if your data are structured in a database format, the headers or fields represent your concepts. The actual cell values are the concrete instantiations (called "instances") or observations for these concepts. Although this is a general rule of thumb, you could have situations in which your cells are concepts themselves.

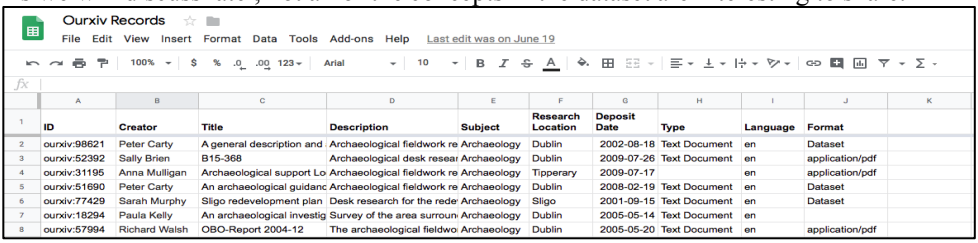
4.2 Thing 2: Exploring: Inventory of your data

4.2.1 Identify relevant concepts from your datasets that you want to expose as LOD

The first step in expressing your data in RDF is to identify the concepts in your dataset that you would eventually like to expose and link to the LOD Cloud. It will most likely not make sense to expose all of the concepts that exist in your data; you will need to be selective. (See W3C Step #2: https://www.w3.org/TR/ld-bp/#SELECT).

Activity: Take an inventory of your dataset. What type of data do you have? What structure (e.g. XML, a database) are they in? Based on your exploration of the LOD Vocabularies in Thing 1, think about which concepts it makes sense to eventually link.

We take as our example an archeological dataset based on data found in the EASY data repository of the Data Archiving and Networked Services (https://easy.dans.knaw.nl). Figure 3 shows the example dataset, which contains archaeological records with ten attributes listed for each record. The dataset has been anonymised for the purpose of this example. As we will discuss later, not all of the concepts in the dataset are interesting to share.



Ourxiv Records

File Edit View Insert Format Data Tools Add-ons Help Last edit was on June 19

	ID	Creator	Title	Description	Subject	Research Location	Deposit Date	Type	Language	Format
2	ourxiv:98621	Peter Carty	A general description and	Archaeological fieldwork re	Archaeology	Dublin	2002-08-18	Text Document	en	Dataset
3	ourxiv:52392	Sally Brien	B15-368	Archaeological desk resear	Archaeology	Dublin	2009-07-26	Text Document	en	application/pdf
4	ourxiv:31195	Anna Mulligan	Archaeological support Lo	Archaeological fieldwork re	Archaeology	Tipperary	2009-07-17		en	application/pdf
5	ourxiv:51690	Peter Carty	An archaeological guidanc	Archaeological fieldwork re	Archaeology	Dublin	2008-02-19	Text Document	en	Dataset
6	ourxiv:77429	Sarah Murphy	Sligo redevelopment plan	Desk research for the rede	Archaeology	Sligo	2001-09-15	Text Document	en	Dataset
7	ourxiv:18294	Paula Kelly	An archaeological investig	Survey of the area surroun	Archaeology	Dublin	2005-05-14	Text Document	en	
8	ourxiv:57994	Richard Walsh	OBO-Report 2004-12	The archaeological fieldwo	Archaeology	Dublin	2005-05-20	Text Document	en	application/pdf

Figure 3. The original data in tabular form.

It is also important to consider who owns the data in this step. In an archive, each dataset can have information about the license status. Are the data listed as being open, or is there a requirement for you to request permission or acknowledgement in order to use the data? Ownership and licensing are also important to consider for your own data.

Activity: Think about your own data. Are you the data rights holder for all parts of your data? If not, identify any licenses that might restrict whether you can expose and link the data to the LOD Cloud. (See W3C Step #1: https://www.w3.org/TR/ld-bp/#PREPARE).

Identifying relevant concepts and relationships is a step that is vital for everyone, both novices and experienced computer scientists. Computer scientists use visual tools to "model" their data, such as Visio (https://products.office.com/en/visio/flowchart-software) and Draw.io (https://www.draw.io) , but you could also create a simple list of the concepts that you would like to expose and link.

Activity: View the introductory tutorial for Draw.io at: https://about.draw.io/support/ . Then, experiment using the tool by visiting this link: https://www.draw.io .

Often not all of the elements in the dataset are things you wish to share. (We discuss this further in Thing 5 when we address how to filter your dataset). We have selected five concepts from our example dataset that we think are important to share (Figure 4). Figure 4 also further demonstrates the difference between concepts and instances, which we mentioned in Thing 1.

Concepts from our dataset:

ID	Creator	Title	Research Location	Date	Type

Instances of each of these concepts:

ourxiv:98621	Peter Carty	A general description and archaeological investigation	Dublin	2002-08-18	Text Document

Figure 4. Examples of concepts and instances in our example dataset.

4.2.2 Identify relevant relations from your datasets that you want to expose as LD

Having identified the concepts in your dataset, you now need to identify the relations (i.e., what later becomes predicates or links) between those concepts. The relations are often determined by the structure of the database itself; however, sometimes the column or row headers can also express relations. Figure 5 presents a data model which shows some of the relationships between the concepts in our example from 4.2.1. In this example, we can see the subject : predicate : object structure.

Figure 5. An example of our data model showing the relationships between concepts and the subject : predicate : object structure (marked in color).

Visualising your data model in this way can also help with understanding where there are relationships which may have been hidden. Once you have a model of the concepts and relationships you would like to link and expose, you are ready to begin defining them as URIs, which will be explained in Thing 3.

Activity: Examine the model of our dataset shown in Figure 6. What might a possible relationship between 'Peter Carty' and 'Dublin' be?

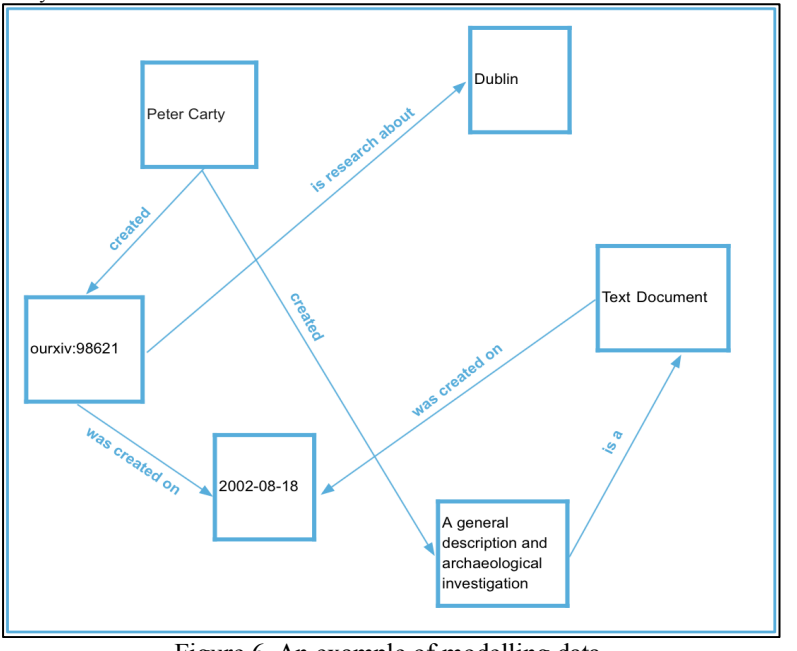


Figure 6. An example of modelling data.

4.3 Thing 3: Defining: Define the URI naming strategy

4.3.1 Define a suitable and durable namespace

The URI is an address that a machine can use to find exactly the right piece of information on the World Wide Web. (You are familiar with this idea already; think of the URL of any website). A URI consists of three parts: a scheme, a domain name and a path. The domain name plus the path are known together as the namespace (see Figure 7). Defining a namespace is extremely important in LOD, as it allows machines (and humans) to tell the difference between identically named elements from multiple datasets (see W3C Step#5: https://www.w3.org/TR/ld-bp/#HTTP-URIS).

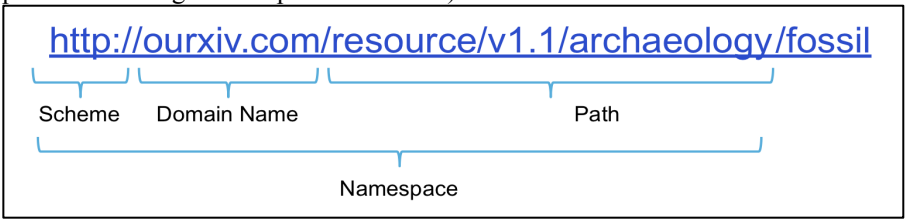


Figure 7. Example of the three parts of a URI and a namespace.

You will first need to choose a domain name to use for the URIs you will create. It is important to think about the sustainability of the domain name that you use. URIs should be persistent and not change over time. If you plan on using a domain name that is part of a project, think about how (or whether) that website will be maintained after the end of the project. It is always better to choose something that you are sure will stand the test of time (see W3C Step #5: https://www.w3.org/TR/ld-bp/#HTTP-URIS), such as institutional domains. Institutional domain names have the added benefit of conferring a sense of authority. That is, a domain name of "harvard.edu" suggests more authority than a domain name of "jane_smith.com". (If you are unsure about your institutional options, check with your local IT professional for guidance).

The remainder of the URI is the path. You can think of the slashes in the path like folders and subfolders that are used to organize information in a manner that is understandable to both people and search engines. We have further recommendations for constructing the path in 4.3.2 and 4.3.3.

Activity: To further prepare for creating your own URIs, read "The Role of Good URIs for Linked Data" from the W3C guidelines at: https://www.w3.org/TR/ld-bp/#HTTP-URIS .

4.3.2 Consider a versioning strategy that reflects past and future modifications of your LD in the URI Path

Datasets are not static; they are often updated and modified with new versions. We recommend that you include versioning as part of your namespace (in the path) to make it perfectly clear which version of your data you are referring to.

The W3C also recommends using vocabularies that provide versioning control. Vocabularies are the definition of concepts, relations and their mutual order. Vocabularies also change, as they are developed and edited. Using a vocabulary with versioning control ensures that if the vocabulary changes, you point to the correct version of it. (See W3C Step

#6: https://www.w3.org/TR/ld-bp/#VOCABULARIES). Concrete observations or instantiations of a vocabulary can also be reclassified. If this happens, you also need to indicate the change at this level.

There is no firm policy on this problem yet. Although in most cases only a small percentage of concepts and relations change between different versions, our proposal is to include the version information in the URI, as indicated by "v1.1" in Figure 8.

http://ourxiv.com/resource/**v1.1**/archaeology/fossil

Figure 8. An example of a URI containing version information.

We believe that this strategy will help your "audience," those who map to your versioned vocabularies, from doing unnecessary updates. We imagine the following scenario: when anybody uses the URI without the version tag (shown in red in Figure 8) a smart lookup service would return versioning information about the concept "fossil." It would also return the preferred version, and the versions where there were changes to this concept.

Changes in vocabulary or instantiations, but also any other changes that you make (i.e., mapping and enriching, which we will discuss in later Things) can be documented or "logged" and described with a commonly shared vocabulary (Moreau and Groth 2013). This is also called provenance information. Compared to versioning, it is like a meta operation on changes concerning the whole or parts of the graph. The versioning we discuss above concerns what has been changed in the knowledge representation structure itself, but does not focus that much on who did it, when and by which process.

4.3.3 Decide how the concepts and relations are represented by its unique identifiers which are part of the URI

We also recommend that you construct your URI in a way that it reflects the meaning of the concepts and relations that you identified in Thing 2. This will make it much easier for people to interpret the URI and understand the link. Rather than using a long string of numbers in our example URI in Figure 8, we used the URI to indicate the relationship between the thing that our data describe (a fossil) and the description of that thing. This involves thinking about how to distinguish between objects in the real world and the webpages describing those objects. Use specific patterns to represent properties, individuals, and classes. Figure 9 shows an example of how to do this:

Example URI	Type of resource
http://ourxiv.com/**resource**/v1.1/archaeology/fossil	← Thing (itself)
http://ourxiv.com/**data**/v1.1/archaeology/fossil	← RDF data (about the thing)
http://ourxiv.com/**page**/archaeology/fossil	← HTML page (about the thing)

Figure 9. Example URIs that use specific terms to represent the relationship between an object in the real world (here, a fossil) and the types of descriptions.

Activity: Using what we have discussed about best practices for creating a URI, draft a few URIs to describe your own data.

4.4 Thing 4: Resolving: Consider resolvability when a person or machine 'visits' the URI

If someone were to put your URI into a browser, what would she get back? A URI is resolvable if anyone, regardless of their own domain, can put it into a browser and see a result. Please note, that the example URI's we have constructed so far are not resolvable!

Activity: Take a look at http://example.org . Is this domain resolvable? Why or why not?

Not every domain is resolvable. The domain in the above activity is not resolvable, but is rather just a placeholder. Remember, you gain authority and trust from other users when your URIs are resolvable and lead to information.

In terms of LOD, it is important that the information that is returned describes the concept in the URI entered in the browser. The information returned could be a snippet of RDF with, for example, information about properties, classes or provenance.

A basic implementation of an RDF URI resolver is the Urisolve server (https://github.com/pharmbio/urisolve). The Urisolve server takes a URI as input and returns a simple list of triples that all have the URI somewhere in each statement. This implementation assumes that there is an HDT (Header, Dictionary, Triples) or SPARQL endpoint that hosts your RDF data. Virtuoso (http://vos.openlinksw.com/owiki/wiki/VOS) is a well known open-source RDF datastore that includes a SPARQL endpoint. HDT is a binary format for RDF which has major performance benefits.

Activity: Visit http://www.rdfhdt.org/ to learn more about HDT and supporting tools.

4.5 Thing 5: Transforming: Generate the URIs for the selected concepts and relations according to the URI naming strategy

Things 1 through 4 are primarily planning steps; in principle, you could actually do them on paper. Thing 5 requires software, tools and/or scripts to transform your data into LD. Your exact approach depends a lot on your particular situation; the format of your data, the size and the available (programming) expertise are the main factors. The following workflow suits many situations.

4.5.1 Filter your data

In Thing 2, we mentioned that it will most likely not make sense to share all of your data. Filtering your data involves creating a new temporary dataset that contains only those concepts and relations that you want to expose as LD. If your data are in a database like PostgreSQL or MySQL, it is often easiest to write a SQL command that generates one new temporal table containing the union of selected columns from the various tables. If your data are in a spreadsheet like Excel, you can create a new sheet via macros and filters. Note that you should try to keep this filtering and generation process as automated as possible and save the macros or SQL for future version conversions. Figure 10 provides an example of the filtering process.

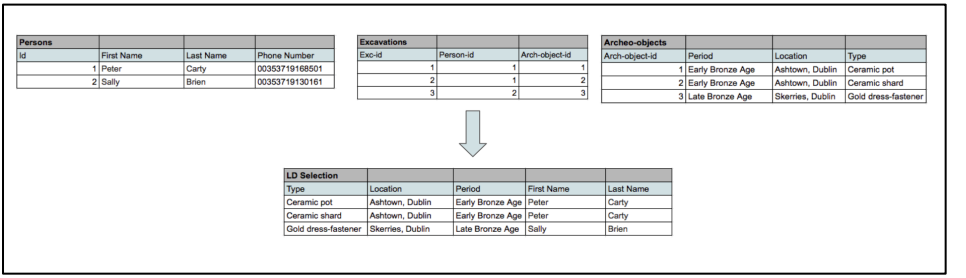


Figure 10. An example of filtering data.

Activity: Based on your work in Thing 2, create a temporary dataset containing only those concepts and relations that you want to expose as LD.

4.5.2 Bridge your prepared data to your tool

There are basically two ways for an RDF generation tool to work with the table from the previous step: 1) set up a connection between the data store and the tool, or 2) serialize the data to a format that the tool can use.

Set up a DB connection: tools like Ontop (https://ontop.inf.unibz.it/) can connect directly to your database and use transformation rules to create LD, or even a SPARQL endpoint to your live data.

Serialize your data: serialization is turning your data from the format you usually interact with to a series of bits. Based on your data format, most tools and databases have the functionality to store tables in CSV format. Please be aware that encoding can be tricky especially with special character sets.

4.5.3 Use tools to transform your serialized/connected data into LD

Depending on the previous step, the selected tool directs the way how your data will be transformed into LD.

4.6 Thing 6: Mapping: Map your LD from your newly defined namespace to similar concepts and relations within the LOD

Most likely there will be concepts and relations in your fresh LD dataset that are similar to concepts and relations in the LOD. The challenge in this step is to: 1) find them; 2) make a selection based on a quality metric; and, 3) select the schema to express these mappings. Mapping (in the sense of defining your own system in relation to other systems) sometimes involves creating an ontology (mapping out your own schema), but this is rarely necessary. In most cases it suffices to create a linkset using the SKOS vocabulary (see the SKOS section of 4.6.3 below).

4.6.1 Finding related concepts and relations

The 'Linked' aspect of LD is the focus of this point. In this exercise you browse online resources to find vocabularies and schemas that have concepts and relations similar to those you have created.

Activity:

•Explore the following public sources to find LD related to your own: The LOD Laundromat (http://lodlaundromat.org/), the Linked Open Vocabularies (https://lov.linkeddata.es/dataset/lov/) portal and BARTOC (https://bartoc.org).
•Next, look at the following domain specific resources: GeoName for locations (http://www.geonames.org/ontology/documentation.html) and Getty AAT (https://www.getty.edu/research/tools/vocabularies/aat/) for excavational objects like Etruscian Pottery (see http://vocab.getty.edu/aat/300020499.rdf).
•Are there any other domain specific sources for vocabularies that you know of that could be relevant for your data?

Note that, although preferred otherwise, the external concepts you wish to link to themselves do not need to be designed as LD . For example, a researcher mentioned in your database can have a persistent identifier in ORCID (https://orcid.org) and a publication can have a DOI (https://www.doi.org).

4.6.2 Sort and make a selection from the sources found in the previous step

The decision depends on many factors, such as your audience (if, e.g., it needs to be multilingual), the coverage with your own concepts (i.e., exact match is preferred to broad superclasses), the authority of the external source (who developed it and maintains it), etc.

4.6.3 Selection of the mapping schema

There are different 'flavours' regarding mapping concepts and relations. The choice is made primarily based on the inferencing and other logical reasoning requirements, as we detail below.

RDFS
In RDF one can specify that an instance is of a certain class, like a *cat* is a type of *animal*. This expressive power is often too limited. RDFS is an extra logical expressive schema that allows one to bind a property to a domain and range, for example the *employer* relation is between the domain: *person* and range: *organisation*. RDFS provides the means to specify subclasses, e.g., *student rdfs:subClassOf person*, and subproperties, e.g., *hasSibling rdfs:subPropertyOf hasRelative*. Unfortunately, neither RDF nor RDFS offer an option to state equality between concepts or relations. For that we have OWL and SKOS which we cover next.

The OWL variants
The W3C-OWL stack (e.g. OWL-Lite, OWL-full OWL-DL; see https://www.w3.org/OWL/) extends RDFS with additional reasoning options grounded in formal logic, which has as an advantage that more automated checking and derivations can be done but is also for many people difficult to learn and adds more computational demands to the reasoning backend. The most popular owl statement is the property *owl:sameAs* which as expected expresses equality between two instances (e.g., "Bill Clinton" and "William_Jefferson_Clinton") or classes (e.g., "Area" and "Region").

SKOS
The popularity of SKOS (https://www.w3.org/TR/skos-primer/) perhaps lies in the fact that it has no formal grounding and people use it to express all kinds of containment relations. For example the *skos:broader* property is used to express a subclass relation (*mammal skos:broader animal*), a subregion (*Texas skos:broader USA*), subperiod (*baby-boom-period skos:broader 20thCentury*), etc. Despite the lack of formal grounding, most humans do understand the inherent reasoning and can develop in retrospect applications that properly deal with these mappings.

Having a good idea about which concepts in your own data you want to expose and which concepts are already published as LD helps in the decision-making process for selecting a mapping schema. Likewise, you can revisit your data model and see if there are better ways to define the relationships between your concepts.

Activity: Study Figure 11 below. On the left is the data model from Figure 6 and on the right we can see some concepts which we have identified as relevant to map to our dataset. Can you identify which concepts on the right would be mapped to which parts of our data model on the left?

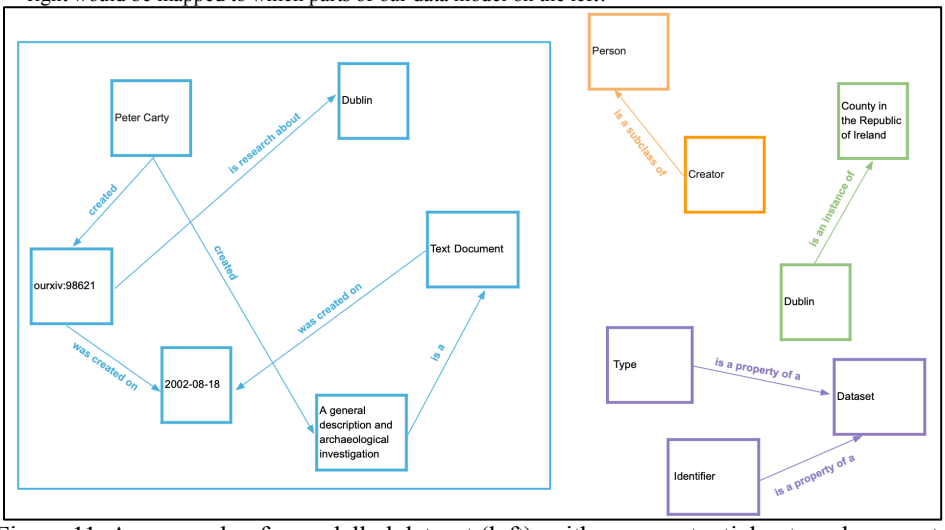


Figure 11. An example of a modelled dataset (left), with some potential external concepts for the data to be linked to (right).

4.7 Thing 7: Enriching: Enrich your data with information from the LOD

The enrichment process is very similar to the mapping, with the subtle difference that the goal of mapping is to connect your data to existing LD, and enrichment is to describe your data with LD. Although not set in stone, the mapping process uses a well-known set of properties that results in a linkset of similarities either on class level or instance level, in RDFS (e.g., *subClass*), SKOS (e.g., *exactMatch*) and OWL (e.g., *sameAs*). The enrichment process has a wider scope on both the selection of properties and objects. Key is that the enrichments are relevant for the goal of sharing your data.

Activity: Imagine that you are a producer of chemical compounds. The molecular weight, structure, boiling point, etc. for different compounds may be relevant properties for your data. Take a look at ChEMBL (https://www.ebi.ac.uk/chembl/) and explore how it could be useful to you.

Similarly, if you work for a library, you can enrich your collection with concepts from library classification systems like Library of Congress *Classification* (http://id.loc.gov/ontologies/lcc.html), Universal Decimal Classification (http://www.udcc.org/) and *Dewey Decimal Classification* (https://www.oclc.org/en/dewey.html).

Even using our tiny example (shown in Figure 12) the power of LD becomes apparent. By linking your dataset it is possible to enrich it with new meaning.

Activity: Take a close look at the figure below where we have now labelled the relationships between the two sides of the diagram. Does this match what you were thinking of in the earlier activity with this diagram? Through enriching our data, we now know that the Dublin in our dataset is Dublin, Ireland and not Dublin, Ohio (where the Dublin Core metadata schema originated).

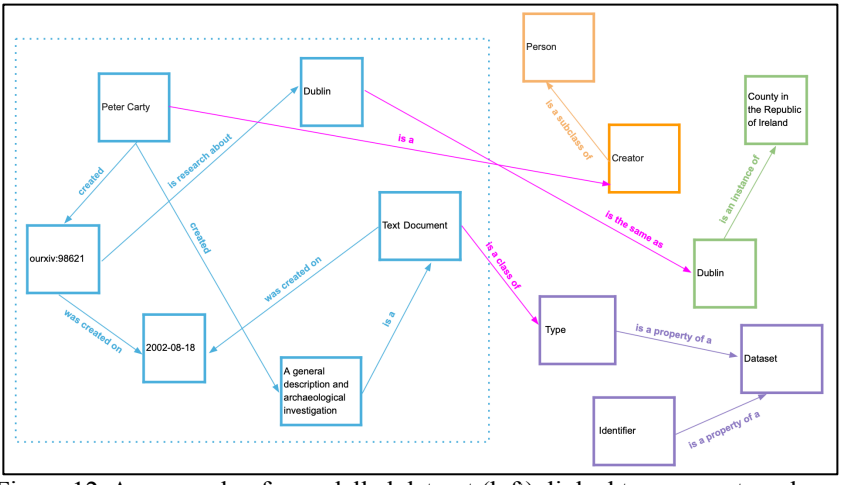


Figure 12. An example of a modelled dataset (left), linked to some external concepts (right).

4.8 Thing 8: Exposing: Define how people can get access to your LD: a data dump, a SPARQL endpoint or a Web API.

After you expose your data, the next step is to think about your intended audience and how people will use and access the data. You will need to consider whether and how you are going to expose your LD "graph" as a whole. You have a few options for exposing your data and making it part of the LOD cloud. Access to the entire LD dataset must be possible via either RDF exploration, an RDF dump or a SPARQL endpoint. These options are further described below.

- Via RDF exploration: This refers to the ability to manually navigate the graph. It allows you to save the "bread-crumb trail" links from document to document and gather the results for searching.
- As an RDF (data) dump: RDF/Turtle is a human friendly serialization format, and one can describe the graph with provenance metadata (e.g. W3C-PROV) and accessibility information (W3C-VOID).
- As a SPARQL endpoint: Be careful because SPARQL is not very easy. It requires a background in query languages and one can easily get lost in the graph; the wrong queries can also put a very heavy load on the server. Initiatives like Puelia-PHP (https://code.google.com/archive/p/puelia-php/), RISIS-SMS (http://sms.risis.eu/) and GRLC (http://grlc.io/) shield the SPARQL complexity by offering an abstraction layer (e.g., as a RESTful service) or visual components for predefined query templates.

Activity: If you are curious to learn more about how queries are formed, visit the Wikidata Query Service: https://query.wikidata.org . This service provides user friendly query examples which allow you to see how queries are formed and how the results are presented.

An alternative option is to use Linked Data Fragments (LDF). LDF (http://linkeddatafragments.org) is a conceptual framework that provides a uniform view on all possible interfaces to RDF. An LDF is characterized by a specific selector (subject URI, SPARQL query, etc.), metadata (variable names, counts, etc.), and controls (links or URIs to other fragments).

4.9 Thing 9: Promoting: Publish and disseminate the value of your data via visualisations and workflows

Once your data are out in the open, you can continue to link each of your statements (objects, subjects, predicates) to other statements in the LOD cloud (see Thing 7). But, you can also create other services on top of your data to tell the world how your data are equal, similar or different to other existing data.

Activity: Visit https://www.cedar-project.nl/about/ to learn more about how LOD are being used in the CEDAR project. Then, take a look at Figure 13, a map which Ashkan Ashkpour and Albert Meroño Peñuela created as a part of this project. To make the map, they combined LD from the Dutch census with openly available geographic data to bring their research to life.

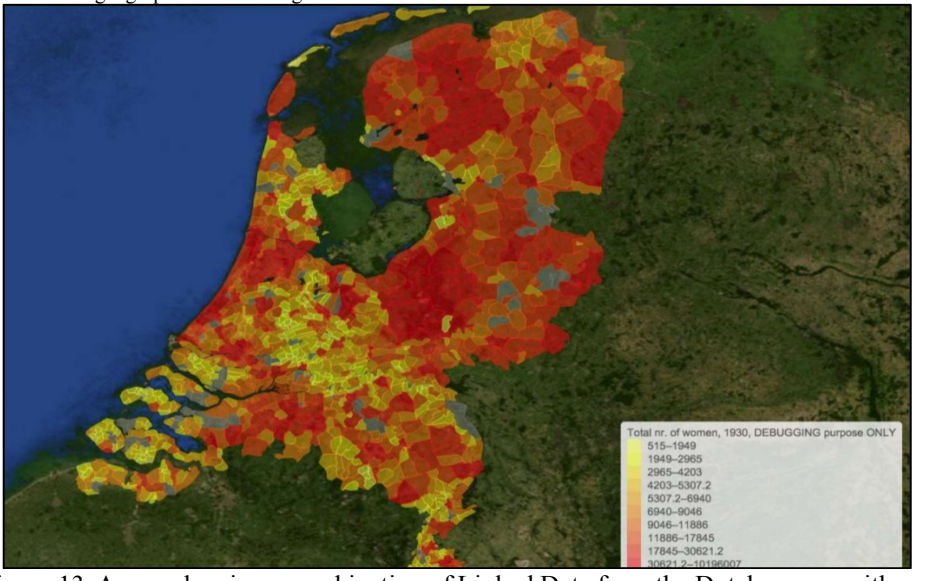


Figure 13. A map showing a combination of Linked Data from the Dutch census with geographic data; the heatmap shows the total number of female inhabitants.

4.10 Thing 10: Sustaining: Ensure sustainability of your data

Publishing LD into the LOD cloud is one specific instance of dealing with data on the web. The W3C recommendation "Data on the Web Best Practices" provides further pointers and considerations for many of the issues that we have raised here, such as the persistence of URI's, version policy, or the reuse of vocabularies (W3C 2017). Many of the best practices listed in the W3C recommendation touch upon the importance of ensuring the sustainability of data publications in the immediate, mid- and long-term. These are also important for you to consider when publishing your data to the LOD cloud.

For example, it is important to associate a clear and preferably well-known standard license with your data and to present it clearly to the audience. You should also indicate whether you maintain the right to change the license in the future. Standard content licences such as Creative Commons (https://creativecommons.org) can be used for this purpose; licence information should be included in the served content (see W3C Step #4: https://www.w3.org/TR/ld-bp/#LICENSE).

Archiving a version of your RDF dataset as a static data dump in a certified, long-term stable data repository might also be a good option to help ensure the long-term sustainability of your data. This provides a way for you to preserve and potentially reuse all of the work that you have already invested (see for an example Beek et al. 2016)

Activity: Examples of how to archive your RDF dataset can be found in the DANS EASY data archive. Explore these examples, paying particular attention to the associated readme file instructions:

The deposit of an RDF dataset from the CEDAR project: https://doi.org/10.17026/dans-xpk-wj5w

The deposit of the Laundromat dataset: https://doi.org/10.17026/dans-znh-bcg3

Publishing data as LOD is new for many researchers. Hopefully the Things, activities, recommendations and references that we have presented here will help you to begin your own journey into the realm of LOD.

References

Beaulieu, Anne, Matt Ratto and Andrea Scharnhorst. 2013. "Learning in a Landscape: Simulation—Building as Reflexive Intervention." *Mind & Society* 12: 91-112. doi:10.1007/s11299-013-0117-5

Beek, W.G.J., L. Rietveld, and S. Schlobach. 2016. LOD Laundromat (Archival Package 2016/06). DANS. https://doi.org/10.17026/dans-znh-bcg3

Dextre Clarke, Stella. 2019. "Thesaurus (for information retrieval). *Knowledge Organization* 46: 439-59.

Hyvönen, Eero. 2012. "Publishing and Using Cultural Heritage Linked Data on the Semantic Web." *Synthesis Lectures on the Semantic Web: Theory and Technology* 2: 1-159.

Hyvönen, Eero. 2020. "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery." *Semantic Web* 11: 187-93. DOI: 10.3233/SW-190386

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, Stefan and Frank van Harmelen. 2015. "Semantic Technologies for Historical Research: A Survey." *Semantic Web* 6: 539-64. DOI: 10.3233/SW-140158

Moreau, Luc and Paul Groth. 2013. *Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology* 3(4): 1-129. https://doi.org/10.2200/S00528ED1V01Y201308WBE007

Scharnhorst, Andrea, Marieke van Erp, Ronald Siebes, Christophe Guéret, Marie Dominique, Tom Crick, Vyacheslav Tykhonov, Gerard Coen, Richard P. Smiraglia, Peter K. Doorn, H. van den Berg, H., Jerry de Vries, A. Meroño-Peñuela, A. Ashkpour and Reinier De Valk, Reinier. 2019. "Curating and Archiving Linked Data Datasets from the Humanities—From Data of the Present to Data of the Future." In: *Book of Abstracts DH2019*. ADHO. https://pure.knaw.nl/portal/en/publications/curating-and-archiving-linked-data-datasets-from-the-humanities-f

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 1-9. https://doi.org/10.1038/sdata.2016.18

W3C. 2014. *Best Practices for Publishing Linked Data.* W3C. Viewed July 10, 2020. https://www.w3.org/TR/ld-bp/

W3C. 2017. *Data on the Web Best Practices*. W3C. Viewed July 10, 2020. https://www.w3.org/TR/dwbp/

Contributors

Andrea Scharnhorst, Data Arching and Networked Services (DANS), Royal Netherlands Academy of the Arts and Sciences, The Hague, The Netherlands

Richard P. Smiraglia, Institute for Knowledge Organization and Structure, Inc., Lake Oswego, Oregon, United States

Rick Szostak, University of Alberta, Edmonton, Alberta, Canada

Aida Slavic, UDC Consortium, The Hague, The Netherlands

Daniel Martínez-Ávila, Carlos III Madrid University, Madrid, Spain

Tobias Renwick, University of Alberta, Edmonton, Alberta, Canada

Marcia Lei Zeng, Kent State University, Kent, Ohio, United States

Philipp Mayr, GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany

Ronald Siebes, Vrije Universiteit, Amstedam, The Netherlands

Charles van den Heuvel, Huygens Institute for the History of the Netherlands and University of Amsterdam, Amsterdam, The Netherlands

Veruska Zamborlini, University of Amsterdam, Amsterdam, The Netherlands

M. Cristina Pattuelli. Pratt Institute, New York, New York, United States

James Bradford Young, Institute for Knowledge Organization and Structure, Inc., Lake Oswego, Oregon, United States

Marnix van Berchum, Utrecht University, Utrecht, The Netherlands

Allard Oelen, Leibniz University, Hannover, Germany

Mohamad Yaser Jaradeh, Leibniz University, Hannover, Germany

Markus Stocker, Leibniz Information Centre for Science and Technology, Hannover, Germany

Sören Auer, , Leibniz Information Centre for Science and Technology, Hannover, Germany

Gerard Coen, Data Arching and Networked Services (DANS), Royal Netherlands Academy of the Arts and Sciences, The Hague, The Netherlands

Kathleen Gregory, Data Arching and Networked Services (DANS), Royal Netherlands Academy of the Arts and Sciences, The Hague, The Netherlands

Index

R

S

T

U

W