DOI: 10.5281/zenodo.7885279

Triple

GUIDELINES ON THE RESEARCH DATA IN THE HUMANITIES

PERSISTENT IDENTIFIERS

KEYWORDS

ABSTRACTS

CITATIONS

Transforming Research through Innovative
Practices for Linked Interdisciplinary Exploration

[JANUARY 2023]

Advancing Open Scholarship
D8.5 – Guidelines on the research data in the humanities

Version 1.0 – Final/PUBLIC

H2020-INFRAEOSC-2019
Grant Agreement 863420

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863420.

Disclaimer: 'The content of this publication is the sole responsibility of the TRIPLE consortium and can in no way be taken to reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains'.

GUIDELINES ON THE RESEARCH DATA IN THE HUMANITIES

Project Acronym: TRIPLE
Project Name: Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration
Grant Agreement No: 863420
Start Date: 1/10/2019
End Date: 31/03/2023
Contributing WP: WP8, WP2
WP Leader: Max Weber Stiftung
Deliverable identifier: D8.5
Contractual Delivery Date: 31/12/2022
Actual Delivery Date: 20/01/2023
Version: 1.0 Final
Dissemination level: PU

Revision History	Created/Modifier	Comments
Version 0.1	Tomasz Umerle (IBL PAN) – editor; Marta Błaszczyńska (IBL PAN), Magdalena Wnuk (IBL PAN), Mateusz Franczak (IBL PAN), Jadranka Stojanovski (University of Zadar), Cezary Rosiński (IBL PAN), Nikodem Wołczuk (IBLPAN), Agnieszka Mikołajczyk-Bareła (Voicelab), Agnieszka Karlińska (IBL PAN), Maciej Ogrodniczuk (IPI PAN), Piotr Pęzik (UŁ), Bianca Kramer (Sesame Open Science, Open Abstracts), Silvio Peroni (University of Bologna, OpenCitations), Luca De Santis (Net7)	First full version

Revision History	Created/Modifier	Comments
Version 0.2	Luca De Santis (Net7), Lorna Balkan (CESSDA), Ana Inkret, Agnieszka Karlińska, Nikodem Wołczuk, Cezary Rosiński, Tomasz Umerle, Jadranka Stojanovski, Karolina Przysiecka, Maciej Maryl, Erzsébet Tóth-Czifra	Modified full version

Revision History	Created/Modifier	Comments
Version 1	Tomasz Umerle (IBL PAN) – editor; Marta Błaszczyńska (IBL PAN), Magdalena Wnuk (IBL PAN), Mateusz Franczak (IBL PAN), Jadranka Stojanovski (University of Zadar), Cezary Rosiński (IBL PAN), Nikodem Wołczuk (IBL PAN), Agnieszka Mikołajczyk-Bareła (Voicelab), Agnieszka Karlińska (IBL PAN), Maciej Ogrodniczuk (IPI PAN), Piotr Pęzik (UŁ), Bianca Kramer (Sesame Open Science, Open Abstracts), Silvio Peroni (University of Bologna, OpenCitations), Luca De Santis (Net7).	First released version

Table of contents

List of Figures

List of tables

List of acronyms (not explicitly decoded in the main text)

APC Article Publishing Charge
API Application Programming Interface
BASE Bielefeld Academic Search Engine
BERT Bidirectional Encoder Representations from Transformers
BITS Book Interchange Tag Suite
DL Deep Learning
FAIR Findable, Accessible, Interoperable, Reusable
GLAM Galleries, Libraries, Archives, Museums
GPT Generative pre-training
ISO International Organization for Standardization
JATS Journal Article Tag Suite
JSON JavaScript Object Notation
LOD Linked Open Data
ML Machine Learning
MORESS Mapping of Research in European Social Sciences and Humanities
OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting
OCR Optical Character Recognition
OLR Optical Layout Recognition
OPERAS European Research Infrastructure for the development of open scholarly communication in the social sciences and humanities
RISIS Research Infrastructure for Science, Technology and Innovation Policy Studies
SEO Search Engine Optimization
SSH Social Sciences and Humanities
STEM Science, Technology, Engineering, Mathematics
TRIPLE Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration
UNESCO United Nations Educational, Scientific and Cultural Organization
URI Uniform Resource Identifier
URL Uniform Resource Locator
UTF Unicode Transformation Format
XML Extensible Markup Language

EXECUTIVE SUMMARY

This report focuses on metadata as a specific type of research data in the humanities, by analysing key metadata elements: persistent identifiers (PIDs), abstracts, keywords, and citations. It defines these elements, outlines the challenges of processing them within the humanities, and presents the challenges associated with GoTriple as a metadata aggregator of this kind of research data.

The assumption is that GoTriple is a specific kind of research dataset, which can, and will be reused by stakeholders such as other metadata aggregators, indexers, publishers, and information services (i.e. providers of scholarly metrics), but also scientists interested in data-driven research (cultural analytics, scientometrics, bibliometrics, etc.). This demands a good understanding of the key metadata elements which are important to GoTriple's aggregation and enrichment processes (abstracts and keywords) and their development (PIDs and citations).

Chapter 1 defines the aim of the deliverable, the context of its creation, and its audience.

Chapter 2 discusses the specificity of research data in the humanities and this report's position within the rich discussions on the topic.

Chapter 3, which is dedicated to PIDs, presents an overview of the topic and the challenges related the humanities' uptake of PIDs, such as the role of cultural heritage data in the humanities, and the importance of bibliodiversity and multilingualism (Subchapter 3.1); then it proceeds to a discussion about the processing of PIDs from GoTriple's data providers by focusing on data dispersion and heterogeneity (Subchapter 3.2).

Chapter 4, which is dedicated to keywords, begins with a typology of keywords and the standards they are expected to adhere to (Subchapter 4.1). Subchapter 4.2 tackles the issue of the automated generation of keywords and proposes various approaches which could be applicable in the context of GoTriple. In Subchapter 4.3, the current approach to keyword organisation in GoTriple is presented, focusing on a GoTriple vocabulary, which responds to the need for the LOD-ification of keywords and which can be reused in the future for automated keyword generation.

Chapter 5, which is dedicated to abstracts, starts with a comprehensive presentation of the abstract ecosystem and also offers a specific perspective on SSH. Subchapter 5.2 proposes some solutions to the issues of missing abstracts which are aimed at the needs of the GoTriple platform.

Chapter 6, which is dedicated to citations, offers an overview of the topic and its relevance to SSH. In Subchapter 6.2, an analysis of issues related to GoTriple's expression of citation data is presented (which relates especially to the challenge of processing different citation formats and to citation data quality).

Each chapter concludes with a summary of the guidelines for the specific metadata type in relation to the humanities.

1. THE AIM OF THE DELIVERABLE

This report – *Guidelines on Research Data in the Humanities* – provides a perspective on research data by focusing on **metadata**[1] **as one type of research data in the humanities,** and treating **GoTriple – a metadata aggregator – as a unique collection of humanities research data** which is being constructed through aggregation, normalisation, and enrichment procedures, and which is being made available for machine exploitation via APIs[2].

The GoTriple platform collects, enriches, and provides access to data which could and should be further re-used for building new data services, providing insights into the humanities' scientific output[3]. These insights, resulting from the analysis of GoTriple's dataset, might be produced through scientometrics, bibliometrics, cultural analytics, or other quantitative and data-driven research methods; and/or may be used by diverse stakeholders in tools and services providing statistics, metrics, or visualisations (which could be of interest to funding bodies, libraries, other GLAM institutions, and policy makers). To make this kind of reuse possible, it is important to recognise the GoTriple dataset – and the data it processes – as research data in their own right and to investigate their broader contexts, both intellectual and technological.

To further develop this dimension, it is important to investigate the GoTriple datasets' critical metadata elements and their development in two dimensions: their role in the scholarly ecosystem, and their specificity within the context of the TRIPLE project.

This deliverable focuses on four important types of metadata – **persistent identifiers [PIDs], abstracts, keywords, and citations** – and presents their specificity within the context of the humanities (and, more broadly, SSH), as discussed in this report by both TRIPLE project contributors and external experts.

Based on these comprehensive discussions, this deliverable offers a set of guidelines for using defined metadata types, which focuses on the challenges facing the humanities and GoTriple in metadata's processing and reuse. Each chapter (devoted to one metadata type) concludes with a summary of the guidelines which outlines the main tasks and challenges which need to be addressed in order to leverage the potential of a certain data type. Thanks to this, a set of guidelines has been delivered which can act as a guide for any party who is interested in building a rich, metadata-based research dataset; and which will also contribute to further discussions on the development of the GoTriple dataset. This deliverable will be of particular interest to actors such as the curators of large metadata

[1] Metadata are structured information about the form, content, and context of documents in any form (textual, graphic, musical notation, etc.) or medium (printed, electronic, etc.).

[2] L. De Santis, *TRIPLE Deliverable: D6.6 API's Development -RP3* (Draft), (Zenodo, 2022), doi.org/10.5281/zenodo.7371832

[3] See Appendix 1, below, for an overview of the current data enrichment workflow in GoTriple.

repositories, aggregators, publishers, indexes and registries, and information services (metrics providers etc.).

This deliverable has been prepared partly through the booksprint The Role of Open Metadata in the SSH Scholarly Communication, organised by IBL PAN on September 7–9, 2022[4]. During this event, external experts collaborated with TRIPLE's contributors to provide input into this deliverable by exchanging knowledge on the four key metadata types – PIDs, keywords, abstracts, and citations – as examples of research data in the humanities.

[4] See project.gotriple.eu/events/triple-booksprint/

2. RESEARCH DATA IN THE HUMANITIES

Marta Błaszczyńska, Mateusz Franczak, Magdalena Wnuk

Over the last few years, data in social sciences and humanities (SSH) has been recognised for its significance and has been growing in importance; and this has also been true in the TRIPLE project, where data has formed a significant part of the discussions and has become part of decisions about the project's foundations, planned actions, and sustainability. Before delving into the details of the data and metadata within the GoTriple platform, it is important to take a brief look at the history of data in the context of SSH more generally, and the humanities more specifically, to explore (some of) the main themes, challenges, and opportunities that the discourses have brought forth.

2.1. THE SPECIFICITY OF THE HUMANITIES

Since the nature of data and, thus, the attitudes towards it differ between the social sciences and the humanities, it is often within the latter that most heated debates have taken place. It is rare that one needs to persuade a quantitative sociologist that the numbers they have been collecting for their last paper are in fact data. On the other hand, humanists have been wary of using the word 'data' to describe the resources they have collected, produced, analysed, and published within their academic workflows. Indeed, it has often been the language which has proved to be the biggest barrier to bringing data into the humanities, where specifically selected semantic choices such as 'primary sources, secondary sources, theoretical documents, bibliographies, critical editions, annotations, [and] notes[5] have discouraged scholars from describing their research processes as simply 'data'.

At the same time, and partly stemming from this, there is often the fear of simplification or of missing out on important contexts and nuances which enter the humanities' discussions together with the concept of data. For example, big data approaches are frequently accused of being reductionist and underrepresenting the richness of the information[6]. Moreover, historians and literary scholars often perceive their work as interpretive rather than data-driven. To speak in terms of data may not seem to capture the whole nature of their work. Indeed, it may also be the nature of the humanities, where the relationship between the scholar and the resources they may be studying is different to the natural sciences. Jennifer Edmond and Erzsébet Tóth-Czifra stress that the issue is, therefore, also

[5] Jennifer Edmond and Erzsébet Tóth-Czifra, *Open Data for Humanists, a Pragmatic Guide*. (Zenodo, 2018), doi.org/10.5281/zenodo.2657248

[6] J. Edmond, N. Horsley, R. Kalnins, J. Lehman, M. Priddy, and T. Stodulka, *Big Data & Complex Knowledge. Observations and Recommendations for Research from the Knowledge Complexity Project*, 8–9, (K-PLEX. University College Dublin, 2018), kplexproject.files.wordpress.com/2018/04/trinity-big-data-report-jklr_04.pdf

a 'material one'. People who conduct humanities research rarely produce data, they often do not really own it, basing their work on the historical and cultural foundations of existing interpretations[7].

So what should be understood as data in the humanities? Definitions and typologies differ, and indeed are (and should be) discipline and workflow specific[8]. Importantly, providing a comprehensive answer to this complex question is not the aim of this deliverable. Several general specificities of humanities data ought to be highlighted here, however (in light of the metadata considerations which come later in this document). First, an aspect which is especially relevant to the TRIPLE project – data in the humanities, in addition to scholarly sources, often include cultural heritage resources. Thus, collaboration with the GLAM (galleries, libraries, archives, museums) sector needs to be established and fostered for successful studies to be carried out[9].

Second, the importance of multilingualism in the humanities cannot be underestimated. As has often been argued, topics and issues which are important to national and local cultural phenomena often ought to be discussed in the languages which are understandable to the people who are most directly affected. The same goes for data at a more general level – they may sometimes be translated into English or other languages but a strong understanding of the local context is needed for analysis and interpretation[10].

Third, data standards and general guidelines always need to be reevaluated in the context of the humanities. For instance, while the FAIR principles (for data to be findable, accessible, interoperable, and reusable) entered the humanities a while back, discussions need to

[7] Edmond and Tóth-Czifra, *Open Data*..., 1.

[8] See: B. Gualandi, L. Pareschi, S. Peroni. 'What Do We Mean by "Data"? A Proposed Classification of Data Types in the Arts and Humanities', (arXiv, 15 July 2022). See doi.org/10.48550/arXiv.2205.06764 for a proposal on 13 data types, based on interviews with researchers in philology and literary criticism, language and linguistics, the history of art, computer science, and archival studies; conducted at the Department of Classical Philology and Italian Studies at the University of Bologna. See also M. Maryl, M. Błaszczyńska, B. Szleszyński, and T. Umerle, 'Dane badawcze w literaturoznawstwie', *Teksty Drugie. Teoria literatury, krytyka, interpretacja*, 2 (1 March 2021): 13–44, for an attempt to design a coherent typology for data in Polish literary studies. The data stories and other discipline-focused work ongoing in the context of DARIAH-EU Research Data Management Working Group are also worth mentioning in this context [see, e.g., E. Tóth-Czifra and N. Truan, 'Creating and Analyzing Multilingual Parliamentary Corpora', *Research Data Management Workflows* Volume 1 (2021), halshs.archives-ouvertes.fr/halshs-03366486]. Its members are currently preparing a publication in the context of the Research Data Management for Arts and Humanities: Integrating Voices of the Community project under the DARIAH Working Groups Funding Scheme 2021–2023 that 'covers and provides practical know-how for both researchers and the new research support professionals (data stewards, subject librarians, open science officers, etc.) working with them.' (see: www.dariah.eu/2022/05/23/dariah-working-groups-funding-scheme-2021-2023-meet-the-winning-projects/).

[9] G. Angelaki, K. Badzmierowska, D. Brown, V. Chiquet, J. Colla, J. Finlay-McAlester, K. Grabowska, et al., *How to Facilitate Cooperation between Humanities Researchers and Cultural Heritage Institutions. Guidelines*, (Warsaw, Poland: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 10 March 2019), doi.org/10.5281/zenodo.2587481

[10] For a more detailed discussion on multilingualism in the humanities see: A. Balula, L. Caliman, S. Fiorini, S. Jarmelo, D. Leão, P. Mounier, J-F. Nomine, et al., *Innovative Models of Bibliodiversity in Scholarly Publications: OPERAS Special Interest Group Multilingualism* (White Paper, 8 November 2021), doi.org/10.5281/zenodo.5653084

constantly reassess what being FAIR means in terms of the humanities’ resources[11], and will, for example, be understood differently in the context of critical literary editions in comparison to philosophical sources or anthropological notes. On the other hand, the straightforward reproducibility of studies is often seen as irrelevant in the humanities because, as mentioned above, a large amount of research relies on interpretative work[12]. However, it is still important to allow our readers to have as full an understanding of the resources we have been using as possible, even though readers may have completely different perceptions of their meaning and relevance to our arguments.

2.2. HUMANITIES METADATA AS RESEARCH DATA

While the humanities community has different approaches to what research data mean for their disciplines, it needs to be emphasised that the metadata types which are emphasised in this report – PIDs, keywords, abstracts, and citations – are just as valuable to humanists as they are to other disciplines (social sciences, but also STEM).

Hence, it is even more important to address the critical challenges in humanities research data in these fields, especially when the humanities are lagging behind in their adherence to the state-of-the-art or pose very specific problems. Of course, some of the challenges we will be discussing below are common to both the humanities and social sciences (rather than being strictly specific to the humanities), in which case we discuss them jointly. However, in all cases where the humanities perspective differs, and therefore demands a very different approach, they will be showcased and detailed so that the humanities community can benefit from this deliverable.

In the following chapters (each devoted to a separate metadata type) a comprehensive overview of the topic is presented with a specific focus on SSH, and the humanities specifically. Additionally, each chapter discusses these issues using concrete examples from the GoTriple dataset and/or GoTriple’s data providers to facilitate the future development of the platform.

[11] See N. Harrower, M. Maryl, T. Biro, B. Immenhauser, (ALLEA Working Group E-Humanities), *Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities*, (Berlin: ALLEA - All European Academies, February 2020), Digital Repository of Ireland, repository.dri.ie/catalog/tq582c863

[12] There has been, however, a lot of debate on the topic. See R. Peels, ‘Replicability and Replication in the Humanities’, *Research Integrity and Peer Review* 4, 2 (2019), doi.org/10.1186/s41073-018-0060-4; J. Britt Holbrook, B. Penders, and S. de Rijcke, ‘The Humanities do not Need a Replication Drive’, *CWTS Blog* (archive), (21 January 2019), www.cwts.nl/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive; R. Peels, L. Bouter, and R. van Woudenberg, ‘Do the Humanities Need a Replication Drive? A Debate Rages on’, *Retraction Watch*, (13 February 2019), retractionwatch.com/2019/02/13/do-the-humanities-need-a-replication-drive-a-debate-rages-on/; J. O’Sullivan, ‘The Humanities have a “Reproducibility” Problem’, *Talking Humanities*, (9 July 2019), talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/; among others.

3. PERSISTENT IDENTIFIERS

3.1. PIDS IN SSH – CURRENT STATE AND UPCOMING CHALLENGES

Jadranka Stojanovski[13]

> *"An identifier is an opaque or explicit number or alphanumeric label which is machine or human readable. It uniquely and permanently identifies and retrieves an object, a document, person, place, organisation, or any entity, in the real world and on the Internet."*
>
> www.ouvrirlascience.fr/open-identifiers-for-open-science/

WHY DO WE NEED PERSISTENT IDENTIFIERS (PIDS)?

Persistent identifiers (PIDs) are unique entity names which are supported by organisational commitment and technical infrastructure to sustain them indefinitely[14]. PIDs could also be defined as a unique identification code attached to a digital object and registered at an agreed location[15]. For PIDs, these resources must be registered in trusted repositories and stored in such a way that they never change, and which can be referenced and cited in this way. Furthermore, these references must be stable, even when the underlying repositories continuously change their hardware, software, physical location, and format. PIDs must be guaranteed to remain functional and provide access to the resource as they move from one location to another. PIDs are not only persistent but also actionable – they can be plugged into a web browser, which will then take the user to the identified source.

Persistent identifiers preserve scientific resources over the long-term in order to ensure their long-lasting accessibility. PIDs serve as pointers which are used to identify and retrieve various resources (publications/documents, software, datasets, bibliographic records/metadata files, multimedia materials, or projects/grants), but can also be applied to physical research components, such as research institutions, funders, people/researchers, samples, artefacts, reagents, or instruments. Due to the identifier resolver network, they are able to provide a resolution mechanism providing direct access to the identified data. They are globally unique, with infinite lifespans, and can be resolved to a physical resource.

According to a UNESCO science report (2021), science spending increased worldwide by 19% over four years, while the number of scientists grew by 13.7% to 8.8 million. In addition,

[13] University of Zadar, orcid.org/0000-0001-7399-522X.
[14] socialhistoryportal.org
[15] www.ncdd.nl

publication output reached 2.9 million articles in 2020, with over 90% of this total from countries with high-income and upper-middle-income economies. Furthermore, the publication output's compound annual growth rate increased by 5% between 2017 and 2020[16]. These numbers do not include other types of research output like preprints, software, research data, or multimedia content; so we can conclude that activities in the area of research and their output is growing inexorably.

Depending on the discipline, scholarly research can consist of complex processes such as research planning and design, data creation and collection, data analysis, reporting of findings, dissemination and sharing, and access and reuse. Although the research process steps have not changed, traditional research has not involved a large amount of collected data and has required only limited use of technologies. On the other hand, today's research is data-intensive and is characterised by large amounts of data, the collection and analysis of which requires complex materials, equipment, infrastructure, tools and software, and new data use and reuse paradigms. In addition, research is interdisciplinary, and research teams are now larger, sometimes consisting of large collaborative teams of researchers from different disciplines and countries. In the open science world, data are available concerning the research funder, institution/lab (where the research is carried out), and instruments/equipment. Lab notes, protocols, datasets, and publications are shared publicly. Research is pre-registered, specifying the research plan before seeing the research outcomes.

Despite the significant advancements generated by findability, accessibility, interoperability, and reusability (FAIR) principles, many of the relations between the research process and its outcomes are lost in the publication-centric landscape of scholarly publishing. Even in cases where data about research infrastructure is available, these data do not have adequate levels of interoperability. Traditional identifiers, like the International Standard Book Number (ISBN) and the International Standard Serial Number (ISSN), which provide unique and persistent identifiers for specific types of resource, were designed for printed resources and are not actionable on the Internet nor interpreted as hyperlinks by web browsers. Therefore, new identifiers have been developed for digital data items in order to provide persistent links to these resources. In practice, the persistent identifier is mapped to up-to-date locators, facilitating access to the physical manifestation of the resource[17]. Based on principles of the level of indirection (separating the name from the particular instance addressed) offered by being able to resolve[18] the 'landing page' and/or the resource

[16] ncses.nsf.gov/pubs/nsb20214

[17] J. Hakala, 'Persistent Identifiers: An Overview', *KIM Technology Watch Report* (2010), www.persid.org/downloads/PI-intro-2010-09-22.pdf

[18] N. Paskin, 'Digital Object Identifier (DOI®) System', *Encyclopedia of Library and Information Sciences* 3 (2010): 1586–1592, 0-www.doi.org.oasis.unisa.ac.za/topics/020210_CSTI.pdf

itself, and the use of metadata at the registry level to describe the objects being identified, appropriate holders are provided for all services which secure reliable interoperability.

The appropriate use of PIDs support the discovery of digital resources, citations, reuse, interoperability and collaboration across facilities, disciplines, institutions and countries, the evaluation of impact through citation tracking, trust, efficiency, scalability, and innovation[19].

A successful and trustworthy PID system should be built on four pillars:

- An independent identifier for any particular technology or organisation.
- Delivery of essential PID functions: a) issuing identifiers (uniqueness, ownership, editable metadata), b) storing identifiers (scalability, integrity, interpretability, versioning), and c) resolving identifiers.
- Separate from data delivery: direct access vs landing web page.
- Employing policies for change, including technological change, social change, identifier abandonment, financial sustainability, and decommissioning[20].

The relations between the publication and the authors, affiliations, and parent publications (e.g. journals) are mainly established by using persistent identifiers. However, the publications' relations with software, research data, protocols, references, other versions of papers (e.g. preprints), the type of paper, article processing charges (APC) data, the project within which the research was conducted, the funder, and, most importantly, data concerning the peer review process and the reviewers and their reports, are mostly lost (Figure 3.1.1). Therefore, they require a broader application of persistent identifiers, supported by the first FAIR principle: '(Meta)data are assigned a globally unique and persistent identifier'.

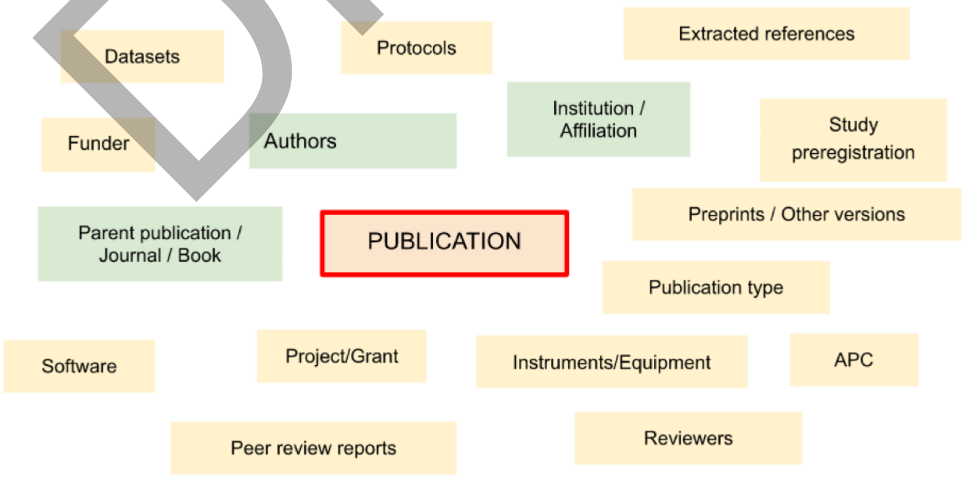


Figure 3.1.1. Entities in the publishing system

[19] Frances Madden, 'Why Use Persistent Identifiers?', *The PID Forum* (2019), pidforum.org/t/why-use-persistent-identifiers/714

[20] N. Car, P. Golodoniuc, and J. Klump, 'The Challenge of Ensuring Persistency of Identifier Systems in the World of Ever-Changing Technology', *Data Science Journal* 16 (2017): 13, doi.org/10.5334/dsj-2017-013

Persistent identifiers were first mentioned in the Data Seal of Approval (DSA), which was issued by Data Archiving and Networked Services (DANS) in 2008, and represents a certification system with 16 guidelines for increasing trust in repositories. The discussions were continued, and this set of criteria were further developed. Finally, in 2016, the FAIR principles were published, summarising the importance of persistent identifiers[21]. According to these principles, data should be 'Findable' online using a persistent identifier, which enables citation and data tracking. Metadata, or information about the data, must be 'Accessible'. To be 'Interoperable' with other data, the data must be in widely accepted file formats, preferably open file formats, and be characterised using standard vocabularies. The data should be made 'Reusable' for other researchers by being accompanied by the appropriate documentation and user licences, thus promoting collaboration and maximising the effect of study outcomes[22].

3.1.1. AN OVERVIEW OF WIDELY ADOPTED PIDS FOR PUBLISHED CONTENT, AUTHORS AND INSTITUTIONS

Machine-readable PIDs such as DOIs, ORCIDs, and ROR are widely accepted, and represent valuable advantages for enabling information sharing across systems.

DIGITAL OBJECT IDENTIFIER (DOI)

Digital object identifiers are the most widely adopted form of PIDs for the various kinds of research objects and publications. A DOI is a unique, persistent digital identifier linked to a digital or physical object, and is used extensively in scholarly publishing. DOIs promote discovery and interlinking and can be assigned to a preprint, journal, journal article, book, book chapter, conference proceeding paper, dataset, presentation, image, table, etc. DOIs provide a persistent link to an object and the standard metadata for that object.

The DOI system is managed by the International DOI Foundation (IDF), a not-for-profit membership organisation which forms the governance and management body for the federation of Registration Agencies, which provide DOI services and registration, and is also the registration authority for the DOI system's ISO standard (ISO 26324)[23].

The DOI system implements the Handle System[24] (a general-purpose global name service enabling secure name resolution over the Internet) and the index Framework[25] (a generic

[21] P. Wittenburg, 'From Persistent Identifiers to Digital Objects to Make Data Science More Efficient', *Data Intelligence* 1 (2019): 6–21, doi.org/10.1162/dint_a_00004

[22] E. Plomp, 'Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments', *Data Science Journal* 19(1) (2020): 46, doi.org/10.5334/dsj_2020_046

[23] www.doi.org

[24] www.handle.net

[25] www.doi.org/factsheets/indecs_factsheet.html

ontology-based contextual data model structure). According to the *DOI Handbook*[26], the DOI system provides a specified standard numbering syntax, a resolution service, a data model incorporating a data dictionary, and an implementation mechanism through a social infrastructure of organisations, policies, and procedures for the governance and registration of DOI names.

The DOI name syntax is structured in the following way: prefix/suffix; for example, 10.1016/159, where ‘10’ is the DOI identifier within the Handle System, ‘1016’ is the registrant code of the organisation which assigned the DOI, and the suffix ‘159’, separated by a ‘/’, identifies the resource. Each suffix is unique to the prefix element which precedes it. DOIs are usually expressed on the web as a URL: ‘dx.doi.org/10.1016/159’. The prefix and suffix can be subdivided further since DOIs are an opaque string with no embedded meaning or limits on the length. A DOI name may be assigned to any entity, which must be precisely defined through structured metadata. The DOI name itself remains persistent through ownership changes and cannot be altered once assigned.

Noteworthy among the group of DOI Registration Agencies are Crossref[27] and DataCite[28]. Crossref is an organisation run by the Publishers International Linking Association (PILA), and was registered in the United States in 2000. Today Crossref works with 17,000 members from 140+ countries, assigning DOIs to their current content (130+ million records). Crossref assigns DOIs to the following resources: journal articles, books and book chapters, conference proceedings and conference papers, technical reports and working papers, theses and dissertations, peer reviews, grants, preprints, standards, databases and datasets, and components. DataCite was founded in the UK in 2009 to improve data citation, establish easier access to research data, increase acceptance of research data, and to support data archiving.

3.1.2. OPEN RESEARCHER AND CONTRIBUTOR ID (ORCID)

Problems with one name assigned to multiple persons[29], multiple names assigned to one person, overuse of the initials and abbreviations instead of full names which stay unresolved, pseudonyms, missing names/surnames, misspelt names, added names, merged names, ordering of given names and surnames and changed names have been present in scholarly publishing for centuries. Finally, inconsistent journal practices and inappropriate cataloguing rules heighten the confusion.

[26] www.doi.org/doi_handbook

[27] www.crossref.org

[28] datacite.org

[29] Estimates by China's Ministry of Public Security suggest that more than 1.1 billion people – around 85% of China's population – share just 129 surnames

Different attempts have been made to solve these problems, but only in limited environments. Some bibliographic databases employ their own identifiers, such as ResearcherID (Web of Science Core Collection, Clarivate) or the Scopus Author ID (Scopus, Elsevier), as do some repositories and archives (e.g. arXivID). Researchers could have an identifier which is unique inside one country. Still, such solutions are not open or globally recognised.

The globally accepted and unique Google Scholar ID identifier has become very popular across the academic community, improving researchers' visibility. However, we should not forget that Google is a commercial company. Google Scholar will only remain available for as long as they believe it to be a successful, or at the very least, not an overly expensive component of their business strategy. Therefore, the Google Scholar ID cannot be considered a persistent identifier.

ORCID (Open Researcher and Contributor ID) is an open-source, cross-national identification system which provides persistent digital identifiers (ORCID ID) which the researcher owns and controls, distinguishing her or him from every other researcher[30]. ORCID provides people with a unique identity for engaging in research, scholarship, and innovation activities. Its goal is to enable transparent and reliable links between researchers, their publications, peer review and other contributions, grants, and affiliations. Researchers may include their ORCID identifier when they write a data management plan, deposit a dataset into a repository, or access a dataset for analysis purposes[31]. Since its founding in 2012, ORCID has benefited the research community by making it possible for persistent identifiers and metadata to be collected, connected, and reused under the complete authority of the researchers who use them[32]. ORCID provides a free, non-proprietary registry of persistent unique identifiers for researchers, scholars, and analysts, together with APIs which enable the interoperable exchange of information between systems in order to embed identifiers in research systems and workflows.

ORCID ID profiles can be connected with other unique identifiers assigned by different services like Google Scholar ID (Google), ResearcherID (Clarivate), Scopus Author ID (Elsevier), etc. This does not mean that the connections between different PIDs which are for the same person are explicitly stated by the various stakeholders, hence the need to harmonise, deduplicate, and normalise this type of data.

[30] orcid.org

[31] L. L. Haak, 'Persistent Identifiers Can Improve Provenance and Attribution and Encourage Sharing of Research Results', (1 Jan. 2014): 93–96, doi.org/10.3233/ISU-140736

[32] Chris Shillum, Julie Anne Petro, Tom Demeranville, Ivo Wijnbergen, Sarah Hershberger, Will Simpson, *From Vision to Value*: *ORCID's 2022–2025 Strategic Plan*, (ORCID, 2021), Online resource, doi.org/10.23640/07243.16687207.v1

3.1.3. RESEARCH ORGANIZATION REGISTRY (ROR)

ROR is a community-led project which develops an open, sustainable, usable, and unique identifier for every research organisation in the world by providing identifiers which are globally unique, stable, discoverable, and resolvable. In addition, ROR develops appropriate metadata schema for organisations and explores interoperability with other identifiers through relationship metadata[33]. ROR is intended for use by the research community to increase the use of organisation identifiers and to enable connections between organisations' records in various systems.

Access to organisations which manage ROR records is granted via permission. ROR focuses on the organisational levels most pertinent to affiliation use cases (those who employ, educate, fund, etc.). According to the ROR documentation, required metadata, and Open Definition conformant licence, an organisation should provide metadata elements sufficient to uniquely identify the organisation in both human- and machine-readable formats. ROR also provides open criteria and documented processes for inclusion/exclusion, creating, merging, or deprecating an institution's records. Changes to records are tracked and recorded using an open provenance model. ROR also claims to have a robust customer support system and an open knowledge base to maintain good relations with the community's technical teams.

Other relevant PIDs enable the linking of different aspects of SSH research, for example: the Archival Resource Key (ARK)[34]; Dewey[35]; Entertainment Identifier Registry Association (EIDR)[36]; VIAF[37]; Open Funder Registry (FundRef)[38]; Contributor Roles Taxonomy (CredIiT)[39], developing a taxonomy for contributors to research output; Persistent Identification of Instruments (PIDINST)[40]; the International Geo Sample Number (IGSN)[41]; Research Resource Identifiers (RRID)[42]; and the Research Activity Identifier (RAID)[43].

CHALLENGES FOR THE HUMANITIES

Although using PIDs has solved many problems related to finding digital sources, there are many challenges concerning their wider adoption, especially in SSH, and the humanities specifically.

First of all, humanities data are closely related to cultural heritage data. For example, many of the documents humanities researchers analyse originate from GLAM collections

[33] ror.org/
[34] arks.org/about/
[35] gutenberg.org/files/12513/12513-h/12513-h.htm
[36] www.eidr.org/about-us/
[37] www.oclc.org/en/viaf.html/
[38] gitlab.com/crossref/open_funder_registry
[39] credit.niso.org/
[40] www.rd-alliance.org/groups/persistent-identification-instruments-wg
[41] www.igsn.org/
[42] www.rrids.org/
[43] www.raid.org.au/

(i.e. archival documents, librarian catalogues, literary works, musical notations etc.); historically it is not always clear who is considered a researcher and who a creator. Finally, many institutions in the humanities have hybrid profiles (i.e. publishers who publish cultural books and research monographs). This impacts the set of PIDs which are relevant to the humanities. A good selection of humanities-specific PIDs can be found in *Developing Identifiers for Heritage Collections*[44].

Second, there is a specificity to the humanities publishing environment. For humanities stakeholders it is important, and necessary, to cherish the bibliodiversity[45] of the output. For example, books[46] play a large role in the humanities – to a greater extent than in other disciplines – which poses a set of specific issues for the humanities. One of these is the issue of granularity – should PIDs be attributed at the level of a whole book or at the level of the chapter (and how to define all the complex variations of this challenge)? Another type of humanities output which adds to the diversity is cultural heritage publications, especially those originating from smaller publishers such as cultural magazines and newspapers, or citizen science publications (i.e. blogs). These publishers, on many occasions, do not have access to digital infrastructures providing PIDs, or do not possess the relevant know-how to be able to address this issue. Finally, historical output is even more relevant to humanities research than to other sciences, where the most current findings seem to circulate more dynamically. This is exemplified by humanities publications which reference older publications, authors, and non-digitised content (and for these entities, PID attribution is not equally incentivised).

Third, humanities research output is multilingual[47]. For certain types of PIDs – like thesauri, controlled vocabularies, taxonomies etc. – this poses a serious challenge to making the resources understandable in local languages. For example, the Library of Congress's subject headings are mostly accessible in major European languages (German, French), but not minor ones (Czech, Polish, or Croatian). The lack of effort put into providing these kinds of resources in a truly multilingual fashion will limit the uptake of PIDs in smaller countries and in less widely-used languages.

Fourth, some challenges are related to the creation of PIDs and to the costs involved, which especially influence the humanities and their set of smaller stakeholders, who are not always incentivised or have the capacity to implement PID systems. An individual

[44] R. Kotarski et al., *Developing Identifiers for Heritage Collections*, (Zenodo, 2021), doi.org/10.5281/zenodo.5205757

[45] K. Shearer, L. Chan, I. Kuchma, and P. Mounier, *Fostering Bibliodiversity in Scholarly Communications: A Call for Action*, (Zenodo. 2020), doi.org/10.5281/zenodo.3752923

[46] Eelco Ferwerda, Frances Pinter, and Niels Stern, *A Landscape Study on Open Access and Monographs: Policies, Funding and Publishing in Eight European Countries*, (Zenodo, 2017), doi.org/10.5281/zenodo.815932.

[47] D. Leão, M.Angelaki, A. Bertino, S. Dumouchel, and F. Vidal, OPERAS *Multilingualism White Paper*, (Zenodo, 2018), doi.org/10.5281/zenodo.1324026.

researcher can obtain an ORCID ID and have control over the content associated with it and what information will be publicly available and displayed. Still, the institution must pay for institutional membership to access enhanced services. However, the process of obtaining a DOI is different, since DOIs must be registered via a registration agency, which needs to be a member of Crossref, DataCite, or another institution. Membership fees are not cheap, especially for small academic institutions and learned societies. Furthermore, each assigned DOI is charged separately. Together with DOI policy requirements that recommend which types of objects can be associated with DOIs, the additional fees lead to 'savings' on permanent identifiers, even for those institutions and journals which can afford annual membership. To become more aligned with the equity, diversity, and inclusion principles of open science, the PID providers' pricing should be tiered more appropriately. For example, a ROR ID could be created free of charge.

Besides the specific humanities context, there are still uncertainties and challenges which impact the whole PID ecosystem (including the humanities). Existing security, reliability, and resilience issues related to unauthorised changes of PID databases and insufficient maintenance strategies have been recorded. Accordingly, although persistency is presumed when considering PID use, some studies have shown that persistency is not warranted, and that scholarly content providers respond differently to varying request methods and network environments, and even change their response to requests relating to the same DOI[48]. Furthermore, there is also a need for sustainable services for non-data resources, a global resolution service for all types of PIDs, and a common mechanism for complex querying across PID systems[49].

Some groups of experts are already working on establishing a data infrastructure based on a digital object access protocol, which represents a universal exchange protocol for digital objects stored in repositories using different data models and organisations[50]. According to Wittenburg (2019), this approach can solve some of the fundamental problems in data management and processing.

[48] M. Klein and L. Balakireva, 'On the Persistence of Persistent Identifiers of the Scholarly Web', (arXiv, 2020), arxiv.org/abs/2004.03011

[49] European Commission, *PID Architecture for the European Open Science Cloud. Report from the EOSC Executive Board Working Group (WG) Architecture PID Task Force (TF)*, (2020), doi.org/10.2777/525581

[50] Wittenburg, 'From Persistent Identifiers to Digital Objects…' 6–21, doi.org/10.1162/dint_a_00004

CONCLUSION

The uptake of PIDs for publications, data, software, researchers, and research organisations has increased in recent years. However, the widespread adoption of PIDs which are relevant to other aspects of research has yet to be realised.

If we look at the specificity of the disciplines in the field of SSH, it is clear that it will be challenging to establish standards which can be easily applied across all disciplines and sub-disciplines. It is to be assumed that common PIDs will be extended to incorporate discipline-specific standards, especially with persistent identifiers for physical samples, artefacts, reagents, and instruments[51]. Therefore, some PIDs will have to adapt their metadata and categorisation schemas, registries, controlled vocabularies, and ontologies to support a broader diversity of research from different disciplines.

[51] E. Plomp, 'Going Digital…', 46. doi.org/10.5334/dsj-2020-046

3.2. GOTRIPLE, DUBLIN CORE AND PERSISTENT IDENTIFIERS

Cezary Rosiński[52], Tomasz Umerle[53], Nikodem Wołczuk[54]

3.2.1. PIDS IN GOTRIPLE

The data model for GoTriple is formalised using Schema.org as a base ontology. However, the data extraction also relies on exploiting the external data models' integration with external data models and formats, such as Isidore, OpenAIRE, the Europeana Data Model, and Dublin Core (DC). As a large proportion of the current data is harvested by OAI-PMH/ DC sources, as well as the fact that metadata issues which are interesting for SSH aggregation are particularly visible in this format, this chapter will focus solely on the problems with persistent identifiers found in the DC format delivered through OAI-PMH. An OAI-PMH implementation must be able to display metadata in the Dublin Core format, however, it may also have the capability to support additional formats. Therefore, the default format for the OAI-PMH protocol is OAI-DC, which is also very common among data providers (e.g. the Directory of Open Access Journals[55], Biblioteka Nauk [Library of Science][56], EKT[57]) and is easily readable by both humans and XML parsing tools, but very limited in terms of its ability to properly represent complex data relations. DC facilitates the aggregation of a large number of sources; but this choice has consequences for attributing PIDs in GoTriple, as DC is a 'flat' format, which makes it difficult to handle relationships between values and external identifiers[58].

It is not necessarily the case that GoTriple data providers do not offer PID-enriched data, but it is common that the DC-based expression of their data is not equipped with PIDs. One example is the Polish Bibliotekanauki.pl, where records are exposed in three formats: DC, JATS, and BWMETA. Of the referenced formats, only JATS (not DC!) contains additional identifiers, like ORCID and ISSN, separately, located inside the intended field. It is noteworthy that the ISSN identifier does exist in DC records, but it is placed within tags which also convey different types of information (such as 'source' where the name of the journal is placed or 'identifier' where the DOI can be stored). The statistics in the figure below (Figure 3.2.1) show the coverage of some of the data identifiers from bibliotekanauki.pl in the JATS format.

[52] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-6136-7186

[53] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-7335-0568

[54] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-4303-2016

[55] doaj.org/

[56] bibliotekanauki.pl/

[57] www.ekt.gr/en/index

[58] D. M. Vogel, 'Qualified Dublin Core and the Scholarly Works Application Profile: A Practical Comparison', *Library Philosophy and Practice (e-journal)*, 1085 (2014), digitalcommons.unl.edu/cgi/viewcontent.cgi?article=2685&context=libphilprac

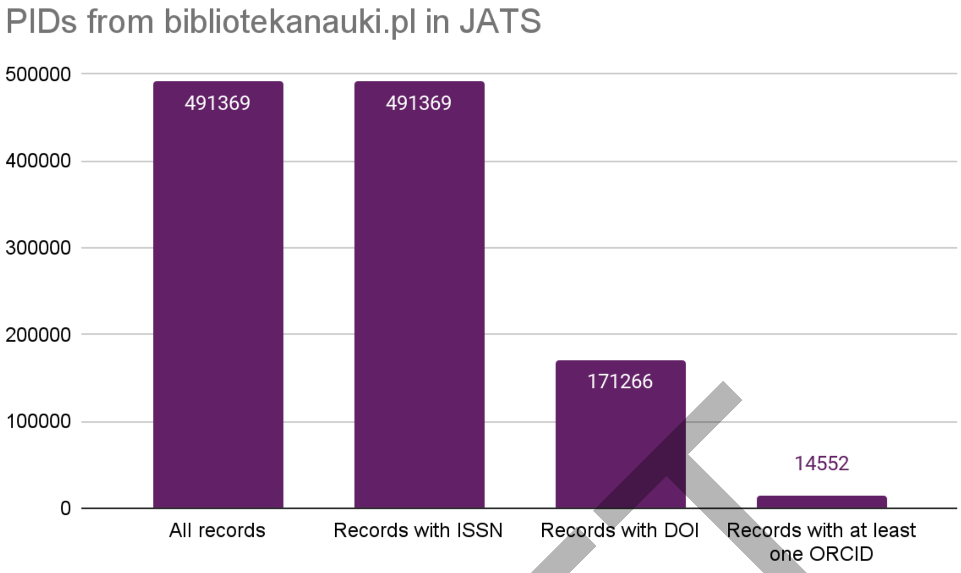


Figure 3.2.1. PIDs in the JATS format from the Library of Science (bibliotekanauki.pl, September 2022).

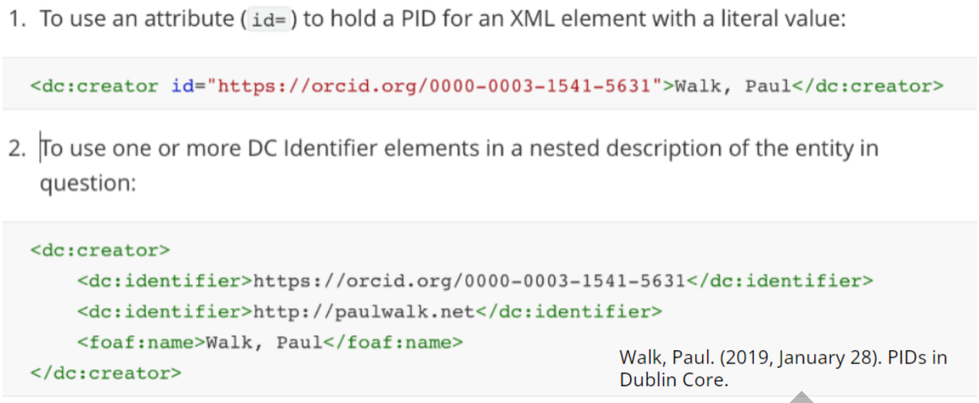
In this text we will focus on the possibilities of enriching PIDs for GoTriple data, but not on discussing which PIDs should be chosen and implemented for this enrichment. For this discussion we will, on the one hand, present the current development of the DC format and, on the other, present possible broader strategies for PID enrichment, as the DC-based approach will not be sufficient in the near future.

3.2.2. PIDS IN DUBLIN CORE – WILL DC EVER BE GOOD ENOUGH?

DC provides solutions for preserving PIDs within the tag 'dc:identifier', which mostly contains a document identifier such as a DOI. Other PIDs presented in DC are, usually, ISSN for journals and ISBN for books. However, this leads to two kinds of issues. First, all of the identifiers for the different types of data (e.g. document vs. journal) are described using multiple copies of the same tag. Second, in cases where there are multiple entities of the same type, DC does not allow these entities to be identified by connecting labels to the IDs. proaches are possible, of which DCMI considers two.

The Dublin Core Metadata Initiative (DCMI), and especially the DCMI's PIDs in Dublin Core Working Group[59], has been conducting work to analyse the existing practices and opportunities related to the presentation of identifiers. Their aim is to develop recommendations on how to correctly represent identifiers with their corresponding strings (the problem concerns, for example, information about the author of the publication) using DC format tags and attributes. Various approaches are possible, of which the DCMI are considering two.

[59] www.dublincore.org/groups/pids-in-dc-wg/



1. To use an attribute (`id=`) to hold a PID for an XML element with a literal value:

```
<dc:creator id="https://orcid.org/0000-0003-1541-5631">Walk, Paul</dc:creator>
```

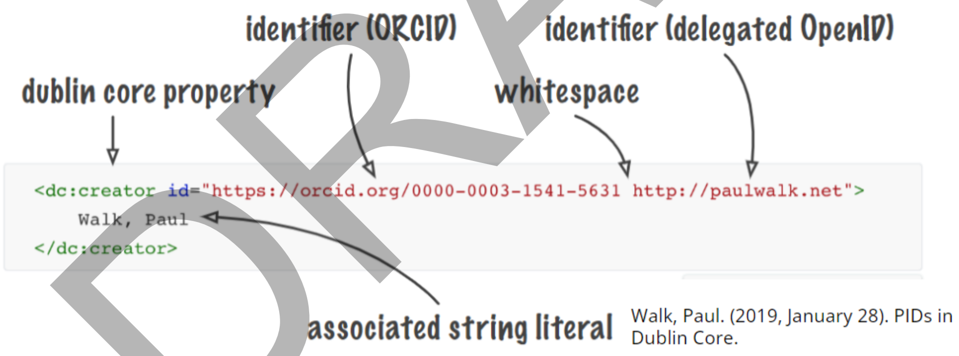
2. To use one or more DC Identifier elements in a nested description of the entity in question:

```
<dc:creator>
    <dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
    <dc:identifier>http://paulwalk.net</dc:identifier>
    <foaf:name>Walk, Paul</foaf:name>
</dc:creator>
```

Walk, Paul. (2019, January 28). PIDs in Dublin Core.

Figure 3.2.2. The approaches considered by DCMI for storing PIDs in DC [Paul Walk, PIDs in Dublin Core, (Zenodo, January 28, 2019) doi.org/10.5281/zenodo.2551181)].

One of the suggested solutions presented above equips the DC piece of information with additional structural elements such as an attribute. The attribute contains a key-value construction, where external identifiers may be presented as a property. The other solution extends the somewhat flat structure by providing nested properties for identifiers, where additional levels of the XML tree may be built.



```
<dc:creator id="https://orcid.org/0000-0003-1541-5631 http://paulwalk.net">
    Walk, Paul
</dc:creator>
```

Walk, Paul. (2019, January 28). PIDs in Dublin Core.

Figure 3.2.3. Using tag attributes to store PIDs [Paul Walk, PIDs in Dublin Core, (Zenodo, January 28, 2019) doi.org/10.5281/zenodo.2551181)]

The proposed solutions could solve the problems in the representation of identifiers mentioned earlier, but would still require extending the simplified format used by the OAI-PMH protocols, which is not a simple task. This requires interfering with the protocol and the providers' actions in order to adjust the data provided so far.

However, as long as the main data format used in GoTriple to aggregate data is the simplified OAI-DC – unconnected to DC development – the problem seems impossible to solve. The main goal of the platform in this area should be to move toward more flexible data exchange formats which will allow for the appropriate extraction and storage of identifier data – such as JATS and BITS.

3.2.3. HOW TO PROPAGATE PIDS IN GOTRIPLE IN THE FUTURE?

Once GoTriple's data format has been adapted so it can properly store and represent persistent identifiers, and the exchange protocols have been made flexible enough to handle additional data, it is important to take future steps to enrich this data with missing identifiers. It seems that GoTriple can apply three different strategies to enrich the aggregated data; but each comes with additional challenges.

- Different formats

The first strategy is to use other, often richer, formats offered by providers. An example is the Polish provider Biblioteka Nauki, which presents much richer data for articles in the JATS format than in the DC format. The GoTriple platform has already started processing data in additional formats and consistently expands these formats' diversity (EDM for OAI-PMH; soon, DC extended using BASE – named BASE-DC*; and for data dumps, the ISIDORE and OpenAIRE formats). While it is important to significantly expand the range of data being acquired, one should not forget that this can only come at the cost of adapting the service to support larger numbers of differing formats. In fact, implementing support for such large variability requires extensive, substantive, and technical work.

- Dispersed data and providers/aggregators with richer data

The first solution does not address the issue of data which is dispersed among several data sources, i.e., the same records (e.g. publications) may be stored in many places and be described differently by various databases. For the sake of the state-of-the-art aggregation of SSH data – which is the aim of GoTriple – a preprocessing mechanism should be in constant operation to identify the same content presented across multiple sources, so as to obtain the biggest possible PID coverage. This facilitates a number of other data processing-related tasks, such as the normalisation (e.g. deduplication), enrichment, and harmonisation of the data. It is easier to handle the general heterogeneity of GoTriple's contents with PID-rich metadata for documents (even though the PIDs are coming from different resources).

- Using internal identifiers

The last strategy is to use one's own persistent identifiers. While this approach avoids the problems associated with identifying resources elsewhere or acquiring data in new formats, it raises a number of other challenges. Assigning internal identifiers subsequently requires there to be continuous control of the identifiers and maintenance of additional workflow. An approach where a new data schema is created instead of using already well-established and widely used solutions could be considered inappropriate. However, the use of internal identifiers is only useful within the closed ecosystem of the service. It is of little use deviating from current Linked Open Data practices until other users adapt it for their own use.

These solutions seem to be a good way of thinking about data enrichment; but they require serious action by both providers and the GoTriple platform. Trying to implement these new solutions with multiple providers at the same time can be a major challenge. A large dataset like GoTriple demands preprocessing and tracking of different sources due to the current data ecosystem. It is necessary to at least monitor and assess the data quality of multiple providers in order to properly adapt to the changes. Appropriately, the TRIPLE consortium is securing both the sustainability of current data processing solutions and IT work for maintaining GoTriple, but is also actively looking to future development which will allow for the constant improvement and flexibility of data workflows.

3.3. GUIDELINES FOR PIDS AS HUMANITIES RESEARCH DATA IN THE CONTEXT OF TRIPLE – A SUMMARY

1. Machine-readable PIDs for people, organisations, and publications are widely accepted and implemented routinely throughout the scholarly data ecosystem. However, PIDs should be extended and adapted to discipline-specific demands. For the humanities, this means, especially, **considering the application of PIDs outside the scientific ecosystem** – such as PIDs for use in the cultural heritage domain, which is common in the cultural heritage sector (e.g. in libraries). As PIDs are currently not only used as identifiers, but also as connectors which **enable linking between different resources**, it is important to leverage the PID's potential to connect scientific and cultural output, which is crucial for the humanities.
2. The specific needs of the humanities also stem from the importance of **bibliodiversity and multilingualism**, which calls for humanities-specific actions and projects which will properly adapt PID standards to different publication genres and/or local contexts. Additionally, the diversity of the actors involved in the humanities' ecosystem – i.e. cultural institutions, NGOs etc. – might make the uptake of PIDs more challenging (due to the lack of resources and the know-how for cross-sectoral interoperability).
3. PID coverage grows in the humanities, which is more dependent on smaller data providers. This does not mean that this data is present in a way which the aggregator mechanism finds most efficient, sustainable, or proper. To build a PID-rich and reliable humanities research dataset through metadata aggregation – which is the goal of GoTriple as long as it aims to extract any knowledge from aggregated resources – a **complex approach is needed which either takes advantage of the multiplicity of providers' endpoints or is able to enrich data with external identifiers**.

4. The dispersion of documents scattered among multiple providers demands deduplication, which is also sensitive to the fact that the relevant PIDs can be provided by only one (or some) of the providers. **Building a large humanities research dataset based on publications' metadata calls for a dynamic and continuous analysis of multiple data sources** as a prerequisite to the fuller enrichment of the dataset.

4. KEYWORDS

4.1. KEYWORDS IN SSH – CURRENT STATE AND UPCOMING CHALLENGES

Cezary Rosiński[60]

4.1.1. DEFINITION AND APPLICATION OF KEYWORDS

Keywords are a central issue not only in scientific production, but also in storing the results of researchers' work. On the one hand, they are regarded as meaningful words taken from the title or text of a document to represent its content and are used in most scholarly articles to describe or summarise the articles' content. The most common sources of keywords are the author, the indexer (e.g. humans or machines who provide metadata enrichment to scholarly indexing services such as bibliographers or library catalogue systems), or both. On the other hand, and just as importantly, keywords capture the essence of the topic, make it easier to classify content, and lead users to resources. In this sense, keywords are a tool for communication between researchers and the communities interested in the results of their work, and they therefore play a crucial role in research discovery systems. However, in practice, due to the number of articles and books published, sets of keywords which describe a specific resource but are not drawn from a controlled vocabulary get lost among other metadata. Keywords are then just a collection of unrelated and unstructured phrases which do not lend themselves to reuse.

Keywords serve two purposes: describing resources and cataloguing them[61]. These are, to a degree, contradictory use cases. 'Describing' means the characterisation of a document with proper terminology (resulting in highly detailed headings which correspond as closely as possible to the content of the document), while 'cataloguing' is a process which helps find documents in a database (thus, accepting mechanisms for aggregating information and applying a degree of generalisation to provide search terms which apply across publications with similar content). Both use cases require specific competencies. How to make authors familiar with classifications (when they are researchers, not catalogers working with a specific database) versus how to make indexers more competent in different fields of research (in fact, very detailed research)?

[60] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-6136-7186

[61] C. Rockelle Strader, 'Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations', *Library Resources & Technical Services 53*, 4 (2009): 250.

4.1.2. TYPOLOGY OF KEYWORDS

Understanding the role of keywords within the information search process requires consideration of their form and origin in addition to their functionality. An initial typology of keywords is based on their form, including **strings/lists of strings, subject headings**, and **controlled vocabularies**. It is best to consider these three types as stages in the development of knowledge presentation and search mechanisms.

String keywords are a collection of phrases lifted directly from the text which are not linked to any external information source. This means that not only can they not be utilised for searching, but they also cannot be used to filter content or find content which is semantically similar but expressed in a different form. Subject headings offer slightly more possibilities. All occurrences of the same keyword are linked, making it easier to filter content and obtain search results which match certain parameters; subject headings are selected from a list of subject headings which include preferred forms for terms. Subject headings are also used in bibliographies and indexes, acting as an access point and allowing users to search for a work by subject in a library/bibliographic cataloguing database. The most advanced environment for keywords is the controlled vocabulary. This is a structured collection of words and phrases (subject headings, and nomenclature from persons and institutions) and the relations between them, which are used for content indexing and retrieval. It includes preferred as well as non-preferred terms, where preferred terms are used for indexing and have a specific scope or describe a specific domain. The use of a controlled vocabulary improves search results, and its implementation organises information and provides access to structured system resources.

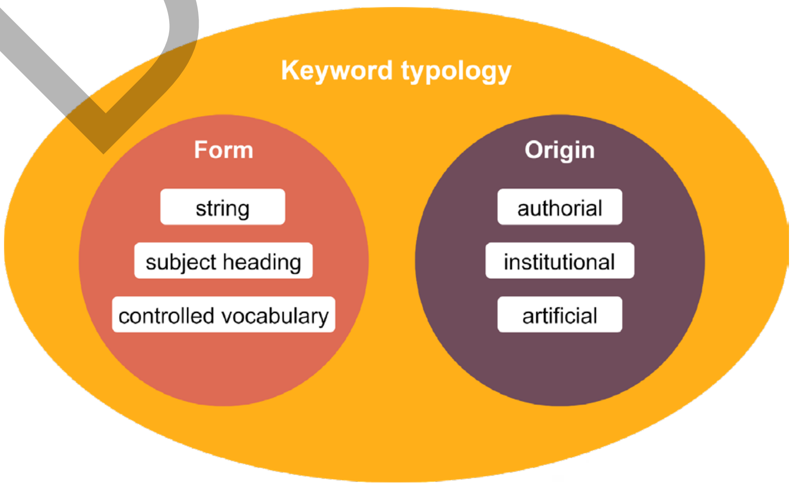


Figure 4.1.1. Keyword typology

An alternative typology of keywords includes information on their source and includes **authorial**, **institutional**, and **artificial origins**. Authorial keywords are assigned by scholars who are subject specialists, and can provide the highest degree of descriptive detail. At the same time, authorial keywords display the highest sensitivity to nuances, as authors are alert to any changes or shifts in terminology. In addition, they are characterised by a focus on the most relevant subjects, with no restrictions on the type or number of keywords. Author generated keywords can come from a controlled vocabulary (for instance, selecting a discipline from a drop-down menu while depositing content in a repository) or, much more commonly, from free text.

Institutional keywords are often assigned by general bibliographers rather than field-specific bibliographers, thus shifting the focus away from the best way of describing a text towards the best way of documenting it within its given database in order to maximise information retrieval. A vocabulary is oriented towards cataloguing, therefore it is also focused on looking for similarities. Because they need to be able to fit into a database framework means institutional keywords must, first and foremost, adhere to an existing structure, making it more difficult to capture nuance and change, while making it easier to capture the similarities to existing data.

The last – and latest – way of creating keywords is to use artificial intelligence[62]. This issue will be exhaustively discussed in the section 'Keyword generation in SSH', while here it will only be outlined. Artificial keywords are the most resource efficient way of describing content, as they are generated on a massive scale and can handle the ever-increasing number of scientific texts. However, it should be noted that they introduce additional problems into the keyword creation environment: they are difficult to verify, are limited by language models, and there are often problems balancing precision and recall. Keywords created in this way usually appear as unrelated strings with no link to controlled vocabularies. An example of an automatic tool used in cataloguing is the Finnish software Annif[63] – a tool for automated subject indexing and classification. Annif employs 'a combination of existing natural language processing and machine learning tools, including TensorFlow, Omikuji, fastText and Gensim. It is multilingual and can support any subject vocabulary (in SKOS or a simple TSV format). It provides a command-line interface, a simple Web UI and a microservice-style REST API'[64].

[62] K. Golub, 'Evaluating Automatic Subject Indexing: A Framework', *keynote speech at the 7th ISKO Italy Meeting Bologna*, 20 April 2015, www.iskoi.org/doc/bologna15/golub.htm

[63] O. Suominen, J. Inkinen, and M. Lehtinen. 'Annif and Finto AI: Developing and Implementing Automated Subject Indexing', *JLIS.It* 13, 1 (2022): 265-82. www.jlis.it/index.php/jlis/article/view/437

[64] annif.org

4.1.3. KEYWORDS AND LINKED OPEN DATA

Linked Open Data[65] (LOD) – according to the World Wide Web Consortium definition – is a vision of globally accessible and linked data on the Internet, based on the Semantic Web's Resource Description Framework (RDF) standards[66]. LOD is often thought of as a virtual data cloud where anyone can access any data they are authorised to see and may also add to any data without disturbing the original data source. This provides an open environment where data can be created, connected, and consumed on the scale of the Internet. A basic theory in LOD is that data have more value if they can be connected to other data. Data, in this context, is any structured web-based information. It has been proposed that LOD be the basis for open data governance and solving many of the data integration issues. LOD helps build bridges between different formats and allows different interoperable information sources to be connected together. This can be achieved by linking a resource to the entity of a well-known value (e.g. a Wikidata/Wikipedia URI). As a result, data integration and viewing complex data becomes easier and more efficient.

[65] F. Bauer, M. Kaltenböck, *Linked Open Data: The Essentials*, (Vienna: edition mono/monochrom, 2012), cdn.semantic-web.com/wp-content/uploads/2017/05/LOD-the-Essentials_0.pdf.

[66] www.w3.org/RDF/

4.2. KEYWORD GENERATION IN SSH

Agnieszka Mikołajczyk-Bareła[67], Agnieszka Karlińska[68],
Maciej Ogrodniczuk[69], Piotr Pęzik[70]

INTRODUCTION

Keywords used as descriptors/metadata of scholarly texts are usually nouns or longer nominal phrases, which succinctly describe the article's content. Such keywords can be abstractive in that they do not necessarily need to occur verbatim in the running text of a document to be considered significant. An important aspect in keyword generation is the nature of the vocabulary: it can be either controlled or uncontrolled.

4.2.1. GOTRIPLE'S WISHLIST

What would the perfect keyword generation algorithm be from the GoTriple users' perspective?

- A solution which assigns keywords to articles based on either abstracts or, if available, the full text of a paper.
- A solution which generates conventional generalisations from content, for example, 'postmodernism', 'literary theory'; but at the same time extracts words and phrases specific to the paper, that is, keywords can be identified extractively, i.e., based on literal occurrences in the text, or abstractively, i.e., are a generalised description of it.
- Keywords can be single or multi-word phrases, as long as they are complete syntactically and semantically.
- Noun phrases are preferred.
- Keywords should be lemmatised and, in the case of longer phrases, the syntactic agreement between phrase components should be preserved or recovered; moreover, the correct cases for named entities should be used, e.g., 'Tatra National Park', 'Institute of National Remembrance'.
- The relevance of keyword assignments is preferred over coverage (a set of precise but incomplete keywords is generally better than a set of complete but imprecise keywords).

[67] Voicelab, orcid.org/0000-0002-8003-6243
[68] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-4846-7086
[69] Institute of Computer Science, Polish Academy of Sciences, orcid.org/0000-0002-3467-9424
[70] University of Lodz, orcid.org/0000-0003-0019-5840

- For abstracts, the number of keywords identified should average between three and five items. But for full-texts, the number of keywords returned should be determined by the user.
- The keyword assignment method should be able to yield keywords from both an open, bottom-up, uncontrolled vocabulary and a controlled vocabulary.
- Keyword assignment should perform well for a large number of languages, be both resource-rich and low-resource, and cover many scientific disciplines.

4.2.2. KEYWORD AUTOMATED GENERATION METHODS

Early works about automated keyword extraction and generation go back to the 1990s. One of the first approaches used decision trees to assign a binary label to each phrase from the paper: a keyphrase or not a keyphrase. Since then, many other approaches have emerged, such as treating keyword extraction as, for instance, a statistical task, an extreme multi-label text classification, or even text generation. Statistical methods like TfIdf or KP-Miner[71] analyse keyphrase frequency, position, and sometimes context to find those that are most relevant. Recent methods based more or less on 'deep' machine learning models often treat keyphrase extraction as an extreme classification problem[72] – they assign each section of text to a few classes (keywords). This results in thousands, or even millions of possible classes (here, meaning keyphrases) to which each text can be assigned. Since it is a multi-labeling problem (each input can be assigned to multiple classes), it results in a very complicated, 'extreme' problem. Finally, most current studies focus on text generation[73]. They use large language generation models pretrained on terabytes of data, and train them to extract and generate keyphrases based on the input text. In this section we will present, train, and test a few commonly used methods for keyword extraction based on a scientific articles corpora called CURLICAT[74]. For a more in-depth review of keyword assignment methods see Papagiannopoulou and Tsoumakas[75].

[71] S. R. El-Beltagy and A. Rafea, 'KP-Miner: Participation in SemEval-2', *Proceedings of the 5th International Workshop on Semantic Evaluation*, (Uppsala 2010): 190–193, aclanthology.org/S10-1041.pdf

[72] W-C. Chang, H-F. Yu, K. Zhong, Y. Yang, and I. D. Dhillon, 'Taming Pretrained Transformers for Extreme Multi-label Text Classification', (KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, 2020): 3163–3171, dl.acm.org/doi/abs/10.1145/3394486.3403368?casa_token=-E8vT-FHdnwAAAAA:PmQdjkC9gsvvgsZok4T3MptmbPNOzh6DyGC1MkElwOwDnwaV1S0OF5lXKsRVVol7tqqBOjG3Xmc

[73] P. Pęzik, A.Mikołajczyk, A.Wawrzyński, B. Nitoń, and M. Ogrodniczuk, *Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer*, (arXiv, 2022), ACIIDS2022, arxiv.org/abs/2209.14008

[74] Pęzik, Mikołajczyk, Wawrzyński, Nitoń, Ogrodniczuk, 'Keyword…'.

[75] E. Papagiannopoulou and G. Tsoumakas, 'A Review of Keyphrase Extraction', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2), arxiv.org/pdf/1905.05044.pdf

4.2.3. CONTROLLED VOCABULARY

KEYWORD EXTRACTION

For a brief overview of keyword extraction tools it is worth starting with ExtremeText[76], which is an extension of a popular text classification library called FastText[77]. FastText uses the vector representation of subwords to train relatively shallow neural networks. ExtremeText builds on that by using a hierarchical softmax classifier. It uses both probabilistic label trees (PLT), and loss and k-means clustering for hierarchical tree building. This allows models to be trained on very large taxonomies with hundreds of thousands of classes. The downside is that it might have low portability with regard to out-of-distribution domains which were not included in the dataset. Additionally, the vocabulary is controlled, hence it poses challenges for rapidly growing disciplines where new terminology appears every year or even every month.

4.2.4. UNCONTROLLED VOCABULARY

KEYWORD EXTRACTION

KeyBERT[78] is yet another popular approach to keyword extraction, this time instead of extreme classification focusing on the unsupervised generation of keywords. The methods adapt the BERT transformer model[79] by creating vector representations of n-grams and comparing them to the vector representation of a whole document. The ranking of key terms is calculated according to the cosine similarity between the n-gram and the document vectors. The method is purely 'extractive', as terms are straightforwardly copied and pasted from the text. It is characterised by the arbitrary operationalisation of key terms (only n-grams with no normalisation). The advantage is that there is no need to train the model as it can use the pretrained BERT which is available online. The fine-tuning of keyword extraction would give users an advantage.

[76] M. Wydmuch, K. Jasinska, M. Kuznetsov, R. Busa-Fekete, and K. Dembczyński, 'A No-Regret Generalization of Hierarchical Softmax to Extreme Multi-label Classification', *Advances in Neural Information Processing Systems* 31, (2018).

[77] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, 'Bag of Tricks for Efficient Text Classification', *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 2, Short Papers, (Association for Computational Linguistics, Valencia, Spain, 2017): 427–431.

[78] M. Grootendorst, *Keybert: Minimal Keyword Extraction with Bert*, (2020), doi.org/10.5281/zenodo.4461265

[79] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *NAACL-HLT* (1), (2019): 4171–4186.

Another keyword extraction tool worth taking into account is TermoPL[80]. It is an algorithm designed to extract terminology from corpora. It identifies terms which match syntactic patterns and produces a ranking based on, among other things, a variant of the C/NC-value measure[81]. Because it can identify, lemmatise, and score recurrent noun phrases as potential terms using a ranking function, it can be used for keyword extraction. Like KeyBERT it is a pure extraction approach.

KEYWORD GENERATION

One of the latest keyword generation models is vlT5[82], which is based on encoder-decoder architecture using transformer blocks presented by Google. The model input is text preceded by a prefix, while the output is the target text, where the prefix defines the type of task, for example, 'Translate from Polish to English:'. The vlT5 was trained on a corpus of scientific articles in order to predict a given set of keyphrases based on the concatenation of the article's abstract and title. It generates precise, yet not always complete keyphrases which describe the content of the article based only on the abstract. The biggest advantage is the transferability of the vlT5 model, as it works well on all domains and types of text. The downside is that the text length, as well as the number of keywords, is similar to the training data: the text piece of an abstract length generates approximately 3 to 5 keywords. It works both extractive and abstractively. Longer pieces of text must be split into smaller chunks, and then be propagated to the model[83]. Additionally, the model is about twice the size of models containing only an encoder (BERT) or decoder (GPT).

A model which is similar in some respects to vlT5 is MBART[84]. It is a sequence-to-sequence autoencoder model pretrained on BART objectives in many languages. VLmBART, a model designed for keyword generation, uses BART autoencoder architecture for keyword generation. Like the T5 model, it is a generative model working both abstractively and extractively, and achieves similar results to the T5 model. In general, our comparison showed that bigger models achieve greater accuracy, hence the best results were for vlT5-large, mBART-large, vlT5-base, and mBART-base (best ⟶ worst)[85].

[80] M. Marciniak, A. Mykowiecka, and P.Rychlik, 'TermoPL – A Flexible Tool for Terminology Extraction'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, eds N.Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, (Portorož, Slovenia: LREC, 2016), 2278–2284, [European Language Resources Association (ELRA)].

[81] K. Frantzi, S. Ananiadou, and H. Mima, 'Automatic Recognition of Multi-Word Terms: The Cvalue/NC-value Method', Int. *Journal on Digital Libraries* (3) (2000): 115–130.

[82] huggingface.co/t5-base

[83] P. Pęzik, A. Mikołajczyk, A. Wawrzyński, B. Nitoń, M. Ogrodniczuk, 'Keyword Extraction from Short Texts with a Text-to-Text Transfer Transformer'. In 'Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2022', eds E. Szczerbicki, K. Wojtkiewicz, S.V. Nguyen, M. Pietranik, M. Krótkiewicz, *Communications in Computer and Information Science*, vol 1716. (Springer, Singapore, 2022), doi.org/10.1007/978-981-19-8234-7_41

[84] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, 'Multilingual Denoising Pre-training for Neural Machine Translation', *Transactions of the Association for Computational Linguistics* 8, (2020): 726–742.

[85] P. Pęzik, A. Mikołajczyk, A. Wawrzyński, B. Nitoń, M. Ogrodniczuk, 'Keyword Extraction...'

4.2.5. KEYWORD GENERATION EVALUATION

QUANTITATIVE EVALUATION

Quantitative evaluation is an essential aspect of assessing the effectiveness of NLP models. In the case of keyword generation, the standard evaluation metrics include the F1-score, precision, recall, and accuracy. Measuring these metrics at different ranks, and considering various scenarios can help to accurately determine the best keyword generation approach.

While quantitative evaluation provides valuable insights, it may not capture the nuances of language and context. This is where qualitative evaluation comes in, which involves human judgement and can provide a more detailed and nuanced understanding of the results.

Qualitative evaluation is a critical component of keyword generation since it can identify errors and areas for improvement which quantitative measures may not capture. In a keyword qualitative evaluation, annotators evaluate the quality of the extracted keywords based on criteria such as their coherence with the text, readability, relevance, number of keywords, and how well they describe the content of the text. Additional criteria, such as whether the quality depends on document types are also evaluated.

However, human evaluation can be expensive and time-consuming, making it infeasible for every team or project. Nonetheless, it remains an important part of the evaluation process, providing valuable insights into the strengths and weaknesses of an algorithm.

In light of the above, to examine extraction approaches, we initiated a study which aimed to compare the performance of several baseline keyword extraction approaches, including FirstPhrases and TopicRank, PositionRank, MultipartiteRank, TextRank, KPMiner, and TfIdf, as well as several state-of-the-art models, such as vlT5, ExtremeText, TermoPL, and KeyBert. We used both quantitative and qualitative measures to provide a comprehensive evaluation.

The relevance and coverage of keyword prediction were compared using standard evaluation metrics like the F1-score, precision, recall, and accuracy. We used both micro- and macro-precision and recall values, as well as their harmonic means F1-score, averaged over the documents in the test set. We measured these scores at several ranks (k=1, 3, 5, and more) for each approach. We considered two different scenarios: a) using the full set of keywords assigned in the training and test set, and b) training and/or evaluating only those keywords which occurred at least 10 times in the stratified dataset.

The highest F1 score, 0.335 (min. 10 words) and 0.227 (no limit), was achieved for vlT5. The next position was taken by ExtremeText: 0.145 and 0.094 respectively. Next, the TermoPL algorithm achieved F1 scores equal to 0.048 and 0.056, while KeyBert ranked in last place with results below 0.01.

Additionally, we tested several other baseline keyword extraction approaches, including FirstPhrases and TopicRank[86], PositionRank[87], MultipartiteRank[88], TextRank[89], KPMiner[90], and TfIdf, with some adjustments aimed at boosting their performance (such as lemmatizing input text). The results were less than 0.025 F1 for all ranks.

The details behind the experiments are described in P. Pęzik, A. Mikołajczyk, A. Wawrzyński, et.al., *Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer, ACIIDS2022*.

QUALITATIVE EVALUATION

In a qualitative evaluation of keyword generation tools, a team of three human annotators (bibliographers and literary scholars) assessed keywords and key phrases extracted from 1,000 digitised papers sourced from a Polish scientific journal in literary studies ('Teksty Drugie') covering the period 1990–2000. The quality of the data varied and there were a lot of OCR errors in many of the texts. We did not perform additional data cleaning.

We tested two unsupervised approaches, an extractive version employing BERT embeddings and cosine similarity (KeyBERT), and an abstractive (or descriptive) version employing a ranking algorithm based on a knowledge base (Monte Carlo method, Wikipedia2Vec).

We also tested the vlT5 model on a smaller sample of texts. Summaries were generated based on the full texts.

The results from the extraction approach were the poorest, and the vast majority of keywords were not accurate. The main issue was a strong dependence on document quality (i.e. OCR quality), which led, for example, to key phrases being cut off. With the descriptive approach we obtained a far better representation of the content of the papers, which allowed them to be compared. Some results were acceptable, others not. By far the best results were obtained with plT5. The keywords were comprehensible and conveyed the main topic of the text well. The detailed results of the evaluation are presented in Table 4.2.1.

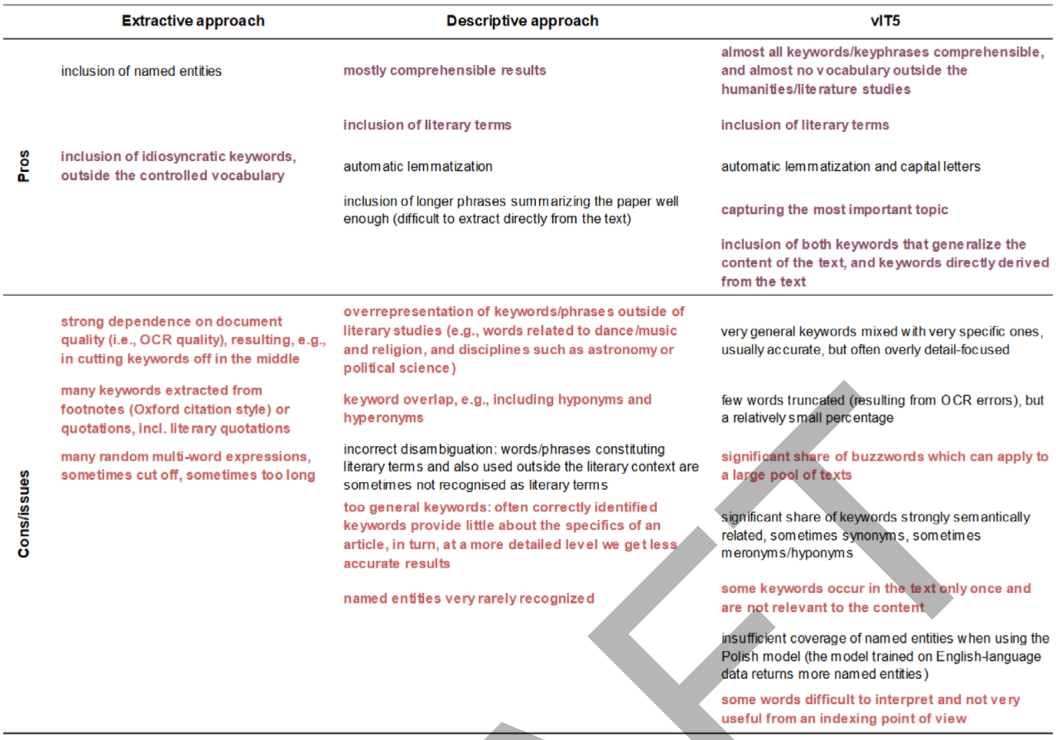
[86] A. Bougouin, F. Boudin, and B. Daille, 'TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction', *International Joint Conference on Natural Language Processing* (IJCNLP), (Nagoya, 2013): 543–551, hal.archives-ouvertes.fr/hal-00917969/

[87] C. Florescu and C. Caragea, 'PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents', Proceedings of the 55th *Annual Meeting of the Association for Computational Linguistics* Volume 1 (Long Papers), (Vancouver, 2017): 1105–1115, aclanthology.org/P17-1102/

[88] F. Boudin, 'Unsupervised Keyphrase Extraction with Multipartite Graphs', *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Volume 2 (Short Papers), (New Orleans, 2018): 667–672, arxiv.org/abs/1803.08721

[89] R. Mihalcea and P. Tarau, 'TextRank. Bringing Order into Texts', *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, 2004), aclanthology.org/W04-3252.pdf

[90] S. R. El-Beltagy and A. Rafea, 'KP-Miner...'



	Extractive approach	Descriptive approach	vIT5
Pros	inclusion of named entities	mostly comprehensible results	almost all keywords/keyphrases comprehensible, and almost no vocabulary outside the humanities/literature studies
		inclusion of literary terms	inclusion of literary terms
	inclusion of idiosyncratic keywords, outside the controlled vocabulary	automatic lemmatization	automatic lemmatization and capital letters
		inclusion of longer phrases summarizing the paper well enough (difficult to extract directly from the text)	capturing the most important topic
			inclusion of both keywords that generalize the content of the text, and keywords directly derived from the text
Cons/issues	strong dependence on document quality (i.e., OCR quality), resulting, e.g., in cutting keywords off in the middle	overrepresentation of keywords/phrases outside of literary studies (e.g., words related to dance/music and religion, and disciplines such as astronomy or political science)	very general keywords mixed with very specific ones, usually accurate, but often overly detail-focused
	many keywords extracted from footnotes (Oxford citation style) or quotations, incl. literary quotations	keyword overlap, e.g., including hyponyms and hyperonyms	few words truncated (resulting from OCR errors), but a relatively small percentage
	many random multi-word expressions, sometimes cut off, sometimes too long	incorrect disambiguation: words/phrases constituting literary terms and also used outside the literary context are sometimes not recognised as literary terms	significant share of buzzwords which can apply to a large pool of texts
		too general keywords: often correctly identified keywords provide little about the specifics of an article, in turn, at a more detailed level we get less accurate results	significant share of keywords strongly semantically related, sometimes synonyms, sometimes meronyms/hyponyms
		named entities very rarely recognized	some keywords occur in the text only once and are not relevant to the content
			insufficient coverage of named entities when using the Polish model (the model trained on English-language data returns more named entities)
			some words difficult to interpret and not very useful from an indexing point of view

Table 4.2.1. Overview of results for the qualitative evaluation of three approaches to keyword extraction for Polish

4.2.6. ISSUES AND CHALLENGES

DATA QUALITY

The key issue in creating keywords is the varying and often low quality data; that is, both the texts from which the keywords are to be extracted (i.e. scientific papers) and the training data. The keyword generation tool needs to work well for both born-digital and digitised texts, which often contain OCR errors and require extensive preprocessing. There are still too few training datasets, and those which are available differ in terms of numbers of annotated keywords and their adopted annotation schemes (sometimes described in detail, sometimes not).

Many texts, especially older ones, do not include abstracts; whereas, for more recent texts, we often only have an abstract. This is a significant difficulty, because sometimes keyphrases are not to be found explicitly in the text, or, for instance, they are available in the full-text but not in the abstract.

MULTILINGUALISM AND MULTIDISCIPLINARITY

Another problem is the multilingualism and multidisciplinarity of the datasets prepared for model training. The languages are sometimes incorrectly annotated or even mixed in both abstract and keyword fields. Thanks to the rapid development of deep learning language models, multilingualism is becoming much less of a problem, as multilingual models are being more commonly used. Moreover, new papers are more often published in English rather than in local languages.

Multidisciplinarity produces a wide variety of abstract structures and keyword description approaches. In SSH, the structure of abstracts is very heterogeneous. Unlike, for example, the medical sciences, there are usually no strict guidelines for creating abstracts in SSH scientific journals. As a result, each author may have his or her own strategy for creating an abstract and assigning keywords. The presence of specialised vocabulary also poses difficulties.

EVALUATION

Evaluating models can also be a challenge. The exact matching of keyphrases is not a good approach, however, there is no better system. Manual analysis is costly and time-consuming.

4.2.7. THE NEEDS OF THE SCIENTIFIC COMMUNITY

The results of a survey[91] of Polish scientific publishers and the editors of scientific journals on their use of digital solutions and the need for new services and tools indicated that, at least in the Polish scientific community, there is a relatively high demand for new metadata enrichment services and tools, including automatic keyword extraction. Publishers and editors of scientific journals are potential users of such solutions.

None of the respondents in the survey used automatic keyword generation tools. And a large proportion of them were not even aware of the existence of such solutions (Figure 4.2.2.).

[91] Carried out as part of the Dariah.lab project, see lab.dariah.pl/en/

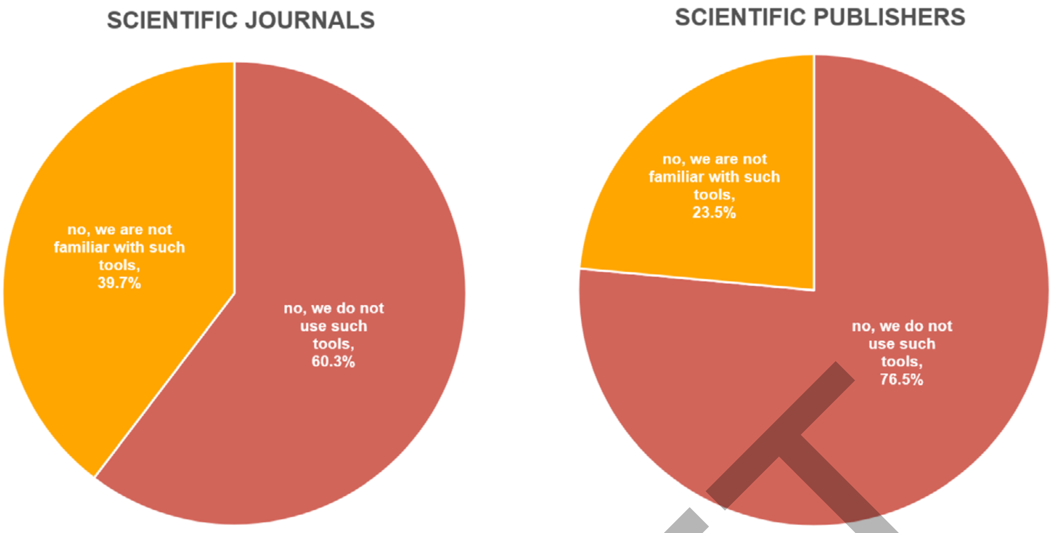


Figure 4.2.1. Use of automatic keyword extraction tools by the editors of Polish scientific journals and publishers

More than half of the editors expressed an interest in using new automatic keyword generation software, provided, however, that it was free (Figure 4.2.3.). Interest in paid software was very low among this group.

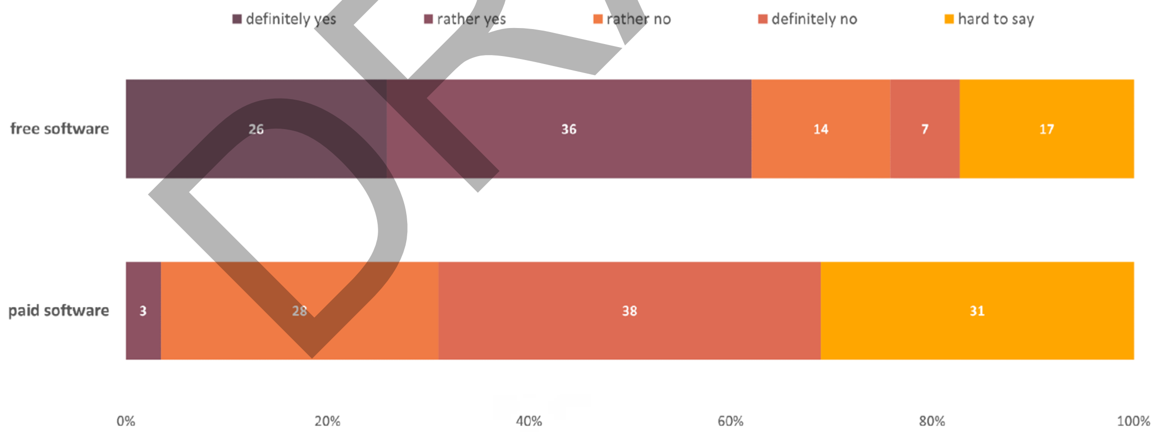


Figure 4.2.2. Interest among the editors of Polish scientific journals in using new automatic keyword extraction software.

More interest was shown by scientific publishers (Figure 4.2.4). Almost 80% of respondents said they would use free keyword extraction software, and 9% would be willing to pay for such a solution.

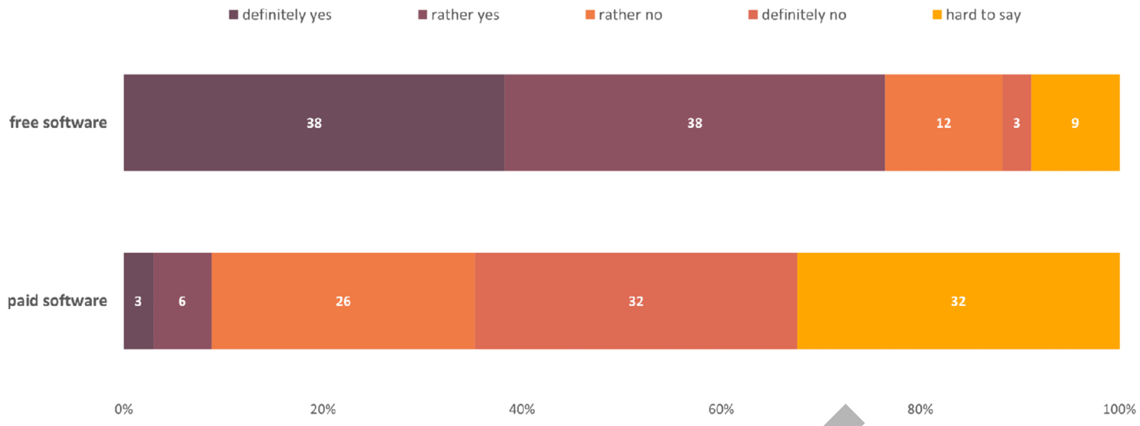


Figure 4.2.3. Interest among Polish scientific publishers in using new automatic keyword extraction software

RECOMMENDATIONS

Taking into consideration the needs of the scientific community and the evaluation of automatic keyword generation tools, we propose the following solutions:

- Consider two levels of keywords – a more general one, which allows papers to be assigned to disciplines, and a more specific one within each discipline (e.g. sub--disciplines, styles, movements, themes).
  - Use broad descriptors such as 'philosophy', 'history'
  - Extract the literary studies domain from Wordnet[92], or decide whether or not to include words/phrases based on their distance from the name of the discipline and the main subject of study; for example, 'literature', 'literary studies'
  - Use domain-specific vocabularies and extend them via Wikipedia and Wordnet;
- Add post-processing, keyword aggregation methods for keyphrases which are too specific;
- Add post-processing, keyword ranking methods which will sort keywords in order from the most to the least important;
- Include named entities, in particular, persons' names;
- Employ contrastive approaches to extract author-specific words/phrases which are absent from vocabularies (idiosyncratic terminology);
- Improve the quality (especially coherence) of the metadata in papers used to train models;
- Liaise with global, open metadata and citation data initiatives such as I4OC, and use their available sources to train models.

A key step in improving the performance of existing tools is to provide well-defined and carefully prepared training datasets – a corpora of text documents which have manually assigned 'perfect' keywords after manual correction of OCR and other errors, and checked inter-annotator agreement scores.

[92] Ch. Fellbaum, ed., *WordNet – An Electronic Lexical Database*, (Cambridge, MA: MIT Press, 1998).

4.3. KEYWORDS IN THE TRIPLE PROJECT

Cezary Rosiński[93]

In addition to aggregating keywords originally assigned to articles by authors or providers, the TRIPLE Project has also created a separate space for keyword extraction. The GoTriple Vocabulary[94] is a vocabulary of subjects in social sciences and humanities (SSH) to be used by the GoTriple platform annotation service. For the annotation mechanism to be effective for publications across all the nine languages which will be supported, the Vocabulary must contain a sufficient number of subject headings in the form of semantic concepts, and, at the same time, these concepts must have labels in as many of these languages as possible.

The project partners concluded that the most effective approach would be to establish a vocabulary by building upon an existing Linked Open Data (LOD) vocabulary. The first step was to determine which vocabulary to use as the foundation. To accomplish this, the task's contributors compiled a list of existing vocabularies[95] and their main features, such as the number of concepts, the languages supported, the subject scope, and whether they covered social sciences and humanities concepts sufficiently, as well as whether they were published as LOD.

The GoTriple Vocabulary was initially developed by identifying fourteen basic concepts from the Frascati taxonomy[96] under SSH. Based on these, thirty-seven broad concepts from the Library of Congress Subject Headings (LCSH) were identified. For each of these – using the Linked Data API of the Library of Congress[97] – the semantic Simple Knowledge Organization System (SKOS) representation was retrieved. For each of the representations, the skos:narrower property was followed by their children in SKOS being extracted, producing 2,513 concepts in total.

The Vocabulary's multilingualism was increased by 1) following LCSH links, 2) following links to Wikidata and extracting labels in various languages, and 3) ingesting existing mappings from national vocabularies (French and Italian National Libraries) into LCSH. The multilingualism was further enhanced by using an automatic translation service to produce missing labels, which were then validated and/or curated by partners. As a result, the coverage increased significantly (e.g. for Greek, from 9.67% to 88.73%; and for Polish, from 13.10% to 99.84%).

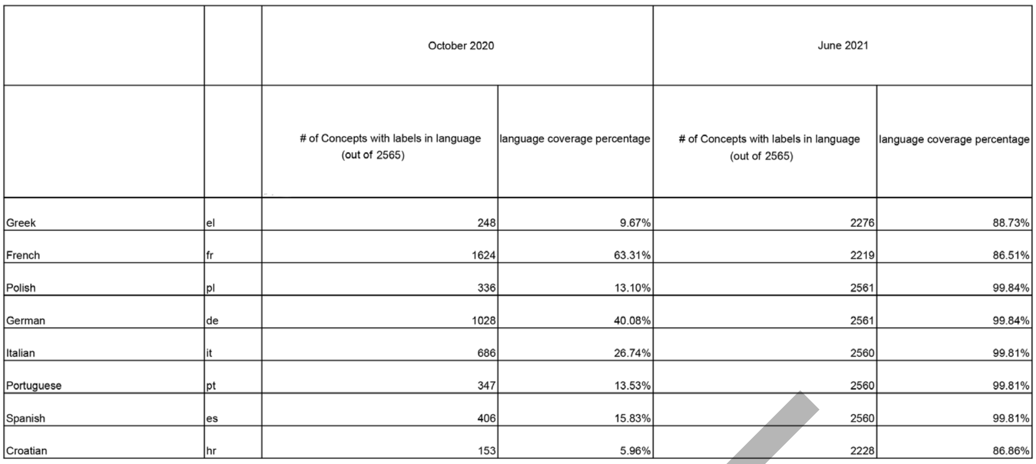
[93] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-6136-7186

[94] semantics.gr/authorities/vocabularies/SSH-LCSH

[95] H. Georgiadis, M. Blaszczynska, and M. Maryl. *TRIPLE Deliverable: D2.4 Report on identification and Creation of New Vocabularies* (Zenodo, 2023), doi.org/10.5281/zenodo.7539922.

[96] www.oecd.org/innovation/frascati-manual-2015-9789264239012-en.htm

[97] APIs for LoC.gov available at loc.gov/apis



		October 2020		June 2021	
		# of Concepts with labels in language (out of 2565)	language coverage percentage	# of Concepts with labels in language (out of 2565)	language coverage percentage
Greek	el	248	9.67%	2276	88.73%
French	fr	1624	63.31%	2219	86.51%
Polish	pl	336	13.10%	2561	99.84%
German	de	1028	40.08%	2561	99.84%
Italian	it	686	26.74%	2560	99.81%
Portuguese	pt	347	13.53%	2560	99.81%
Spanish	es	406	15.83%	2560	99.81%
Croatian	hr	153	5.96%	2228	86.86%

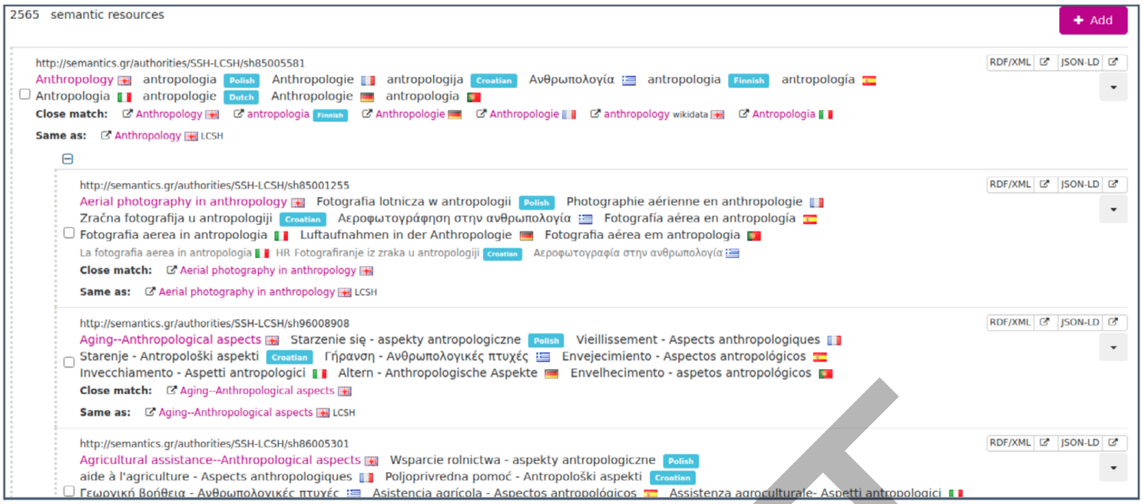
Table 4.3.1. Progress in enhancing multilingualism in the GoTriple Vocabulary

The GoTriple Vocabulary needed to meet certain key requirements in order to be published as a LOD vocabulary in SKOS, which is the standard data model for concept-based vocabularies. This included being accessible via persistent URIs in both HTML and RDF formats following the SKOS data model, and being published under an open licence for anyone to use. Additionally, the Triple consortium was required to continuously update the vocabulary, including adding new labels, links to other vocabularies, and concepts. To ensure these requirements were met, it was decided to host the GoTriple Vocabulary on the dedicated platform Semantics.gr.

Semantics.gr[98] is based on a state-of-the-art infrastructure which supports the development, curation, and interlinking of vocabularies, thesauri, classifications, classification schemes, and authority files (altogether called vocabularies), and their publication as LOD. The infrastructure is being developed in-house by the National Documentation Centre in Greece (EKT, a Greece TRIPLE partner) and has been used for cataloguing and enriching data. It employs semantic knowledge representation technologies.

The novelty of Semantics.gr lies in the fact that, besides SKOS, it can support any data model which can be expressed as an OWL (Web Ontology Language) ontology. The data model chosen in each case generates the template; a vocabulary schema is then created based on this template. In turn, the vocabulary schema stipulates, in detail, the structure of the vocabulary which will be created in the infrastructure. In practice, the vocabulary schema defines, in detail, the entry/update form through which the user adds or updates semantic resources.

[98] www.semantics.gr/authorities/info/semanticsPage?language=en



2565 semantic resources

□ http://semantics.gr/authorities/SSH-LCSH/sh85005581
Anthropology antropologia Polish Anthropologie antropologija Croatian Ανθρωπολογία antropologia Finnish antropología
Antropologia antropologie Dutch Anthropologie antropologia
Close match: Anthropology antropologia Finnish Anthropologie Anthropologie anthropology wikidata Antropologia
Same as: Anthropology LCSH

□ http://semantics.gr/authorities/SSH-LCSH/sh85001255
Aerial photography in anthropology Fotografia lotnicza w antropologii Polish Photographie aérienne en anthropologie
Zračna fotografija u antropologiji Croatian Αεροφωτογράφηση στην ανθρωπολογία Fotografía aérea en antropología
Fotografia aerea in antropologia Luftaufnahmen in der Anthropologie Fotografia aérea em antropologia
La fotografia aerea in antropologia HR Fotografiranje iz zraka u antropologiji Croatian Αεροφωτογραφία στην ανθρωπολογία
Close match: Aerial photography in anthropology
Same as: Aerial photography in anthropology LCSH

□ http://semantics.gr/authorities/SSH-LCSH/sh96008908
Aging--Anthropological aspects Starzenie się - aspekty antropologiczne Polish Vieillissement - Aspects anthropologiques
Starenje - Antropološki aspekti Croatian Γήρανση - Ανθρωπολογικές πτυχές Envejecimiento - Aspectos antropológicos
Invecchiamento - Aspetti antropologici Altern - Anthropologische Aspekte Envelhecimento - aspetos antropológicos
Close match: Aging--Anthropological aspects
Same as: Aging--Anthropological aspects LCSH

□ http://semantics.gr/authorities/SSH-LCSH/sh86005301
Agricultural assistance--Anthropological aspects Wsparcie rolnictwa - aspekty antropologiczne Polish
aide à l'agriculture - Aspects anthropologiques Poljoprivredna pomoć - Antropološki aspekti Croatian
Γεωργική Βοήθεια - Ανθρωπολογικές πτυχές Asistencia agrícola - Aspectos antropológicos Assistenza agroculturale- Aspetti antropologici

Figure 4.3.1. The GoTriple Vocabulary on Semantics.gr

Semantics.gr aspires to serve as a central public platform for publishing trustworthy LOD vocabularies, especially scientific terminology and authority files, which can be further utilised by any third party in order to enhance the quality and interoperability of their digital resources.

The main goal of the GoTriple platform as a discovery system is to aggregate all possible SSH data and provide sufficient query functionalities. Therefore, keywords within GoTriple primarily function as supporting mechanisms for cataloguing information. The service created a rich, controlled vocabulary for SSH, which is stored in a SKOS format using LOD structures. It is possible to enrich the controlled vocabulary both within headings, using links to further services, and within the entire vocabulary, which can be enriched with relevant concepts. However, the GoTriple platform also presents original author-generated keywords alongside each text. Since these keywords exist as strings, they cannot be used in search mechanisms. Moreover, they are not semantically linked to the controlled vocabulary. In the future, it would be worthwhile developing a mechanism for using original keywords and using them to enrich the controlled vocabulary. The use of artificial keywords, on the other hand, can help describe aggregated material more comprehensively, and mechanisms known, for example, from Annif software, could combine the effects of automatic keyword generation with the existing semantic environment of the GoTriple website.

4.4. GUIDELINES FOR KEYWORDS AS HUMANITIES RESEARCH DATA IN THE CONTEXT OF TRIPLE – A SUMMARY

1. Humanities data providers are more prone to content description – e.g. by describing content using keywords in the form of strings – which is not controlled in terms of applied vocabularies (e.g. because of the prevalence of smaller, under-resourced data providers and a more linguistically diverse content). In this case, **a dedicated effort to map existing keywords onto relevant controlled vocabularies** is needed to allow for more interoperability between resources. This effort should take into account **multiple languages, especially those which are less spoken**, which are not present in existing, popular controlled vocabularies.
2. Controlled vocabularies for the humanities need to be multilingual.
3. At least some resources are now being **described using keywords of various origins**, in multiple places (in the humanities, this has become even more systemic because the content of many monographs are being independently indexed through librarian services). In order to create a research dataset, these resources need a certain level of preprocessing in the form of gathering different keywords, then analysing and comparing them. This is a prerequisite for the relevant mapping of these keywords onto existing vocabularies and their enrichment.
4. Due to the growing number of publications which demand content descriptions, there is a **need to develop automated solutions for keyword description**. This automation is especially needed in the humanities where historical content is comparably more relevant. The humanities deals with digitised historical content which cannot be thematically described by humans at a large scale.

5. ABSTRACTS

5.1. ABSTRACTS IN THE SSH

Bianca Kramer[99]

5.1.1. THE ROLE OF ABSTRACTS

Research publication abstracts (including, but not limited to journal articles, book chapters, books, and monographs) fulfil an important role in scholarly communication – they help readers determine the relevance of the publication, communicate key findings, and summarise the content of the publication. They also increase the visibility of the work through being included in bibliographic databases, which assists with discovery. Finally, they are useful in classifying and grouping texts, for instance, in creating subject classifications.

Importantly, these benefits still hold true when the publication itself is open access, as easy access to abstracts facilitates discovery and selection. In addition, for both search and classification, using the full text is not always preferable to using abstracts – for practical as well as conceptual reasons.

While books and monographs may not often include a formal abstract, as is customary with many journal articles and book chapters, they often have a short description available to help readers quickly appraise the topic and content. For the purposes of this chapter, these are considered equivalent to abstracts.

5.1.2. THE IMPORTANCE OF OPEN ABSTRACTS

Abstracts can be considered part of the metadata of publications and, as such, are an important part of open science. To understand and make informed decisions based on research in a transparent way, it is not only important for the data and publications themselves to be openly available, but also for the metadata of these outputs to be openly available and reusable, as well as having open analytical tools and applications to use.

As GoTriple aims to provide the infrastructure to make this possible for social sciences and humanities, it benefits from the availability of open metadata, including abstracts.

[99] Sesame Open Science, Open Abstracts, orcid.org/0000-0002-5965-6560.

5.1.3. WAYS IN WHICH ABSTRACTS ARE (MADE) AVAILABLE

Abstracts are usually available from both publisher websites and bibliographic databases. However, there are several limitations to the availability of abstracts from these sources.

Publishers usually make abstracts free to read on their websites, but while these abstracts can be read by humans who visit an individual publication's landing page, the large-scale processing of the same content by machines is difficult. For search and retrieval purposes, abstracts which are included in bibliographic databases are more useful. These databases, however, often have access restrictions (for instance, only accessible with an institutional licence) and reuse restrictions. In addition, many have a limited disciplinary scope.

5.1.4. INITIATIVE FOR OPEN ABSTRACTS (I4OA)

In 2020, the Initiative for Open Abstracts (I4OA)[100] was started as a collaboration between scholarly publishers, infrastructure organisations, librarians, researchers, and other interested parties to advocate for and promote unrestricted access to the abstracts of the world's scholarly publications which will be held in trusted repositories where they are open and machine-accessible.

While I4OA is not prescriptive in how and where abstracts should be opened, it does ask publishers, where possible, to (also) submit them to Crossref.

5.1.5. CROSSREF AS A CENTRALISED INFRASTRUCTURE FOR METADATA

Crossref provides a centralised infrastructure for making abstracts openly available (together with other publication metadata) in a machine-readable way with minimal restrictions. While obviously limited to publications with a Crossref DOI, for these publications, Crossref is a centralised source of uniformly formatted, authorised metadata. Having abstracts available as part of these metadata opens them up for reuse by many applications and research infrastructures.

Such downstream use can be direct; for example, knowledge extraction for subject classification, screening abstracts for inclusion in systematic reviews, or enriching metadata in institutional repositories (IRs) and research information systems (CRIS). Usage can also be indirect; for example, through being included in bibliographic databases, and being used in bibliographic mapping tools (to create network visualisations around research topics)

[100] i4oa.org/

and in applications which employ machine learning based on abstracts (like Scholarcy[101] and ASReview[102])[103].

Publishers that are members of Crossref can submit abstracts from journal articles, books and book chapters, conference papers, posted content, dissertations, reports, and standards. Abstracts can be deposited in various ways: as part of an XML submitted to Crossref, or by using a web deposit form or a platform-specific plugin like the one developed for OJS. Like other Crossref metadata, they are then available through public Crossref APIs for download and reuse.

WHAT ABOUT COPYRIGHT?

Abstracts have a somewhat unique position – while functionally, they can be considered part of a publication's metadata, they are also in themselves creative textual output and as such, can be under copyright by the authors, their institutions, or the publishers.

In practice, (at least in the European Union) abstracts can be freely used for text- and data mining under the Text and Data Mining (TDM) exception for academic use in the *European Copyright Directive*. However, they cannot be republished without permission, unless they are covered by the same licence as the publication itself, for example, the Creative Commons licences used by many open access publications. It is important to note that the above is meant to provide some guidance and clarification regarding the reuse of abstracts, but does not constitute legal advice.

The restrictions mentioned here also apply to abstracts made available through Crossref as part of a publication's metadata. Crossref itself states that it 'generally provides metadata without restriction; however, some abstracts contained in the metadata may be subject to copyright by publishers or authors'[104].

CURRENT AVAILABILITY OF OPENSSH ABSTRACTS IN CROSSREF

The amount of coverage for the abstracts in Crossref varies greatly across publication types. As shown in Figure 5.1.1., abstracts are available for close to 40% of recent journal articles (published between 2020 and 2022), and preprints are covered to an extremely high level of 80%. For recent books and monographs, though, abstract coverage is lower than 20%, while for book chapters it is even lower than 10%.

[101] www.scholarcy.com/

[102] asreview.nl/

[103] See: A. Tay, B. Kramer, and L. Waltman, *Why Openly Available Abstracts are Important – Overview of the Current State of Affairs*, (Medium, 2020), medium.com/a-academic-librarians-thoughts-on-open-access/why-openly-available-abstracts--are-important-overview-of-the-current-state-of-affairs-bb7bde1ed751

[104] www.crossref.org/documentation/retrieve-metadata/

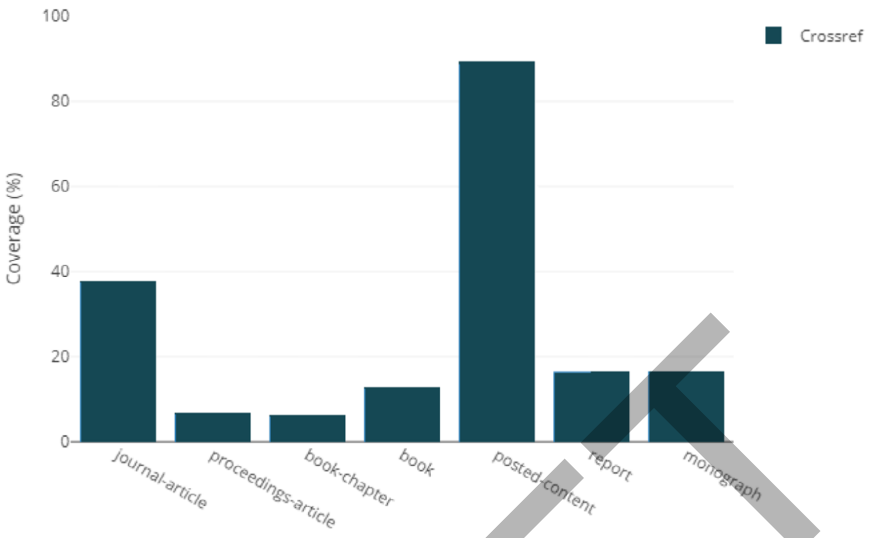


Figure 5.1.1. Abstract coverage in Crossref for publication type (publication years 2020–2022).[105]

The reason for abstracts not being included in Crossref can either be due to a function of publishers not providing them as part of publication metadata (either for all their content, or for a subset of journals), or the publication not having an abstract in the first place. For journal articles, we can see a clear distinction between providers who choose to submit abstracts to Crossref (and often formally support I4OA) and those that don't (see Figure 5.1.2.). For SSH, the latter notably include CAIRN and Project Muse. However, even those publishers that do provide abstracts to Crossref, often don't reach 100% abstract coverage. Especially for journal publishers in SSH (like Brill and Erudit in Figure 5.1.2.), this could be affected by the proportion of journal publications which do not contain abstracts, including, for instance, book reviews (in addition to editorials, letters to the editor, and other non-journal content which can be found in journals from all disciplines).

[105] Data sources here and for the figures below: Crossref metadata (author's own analysis).

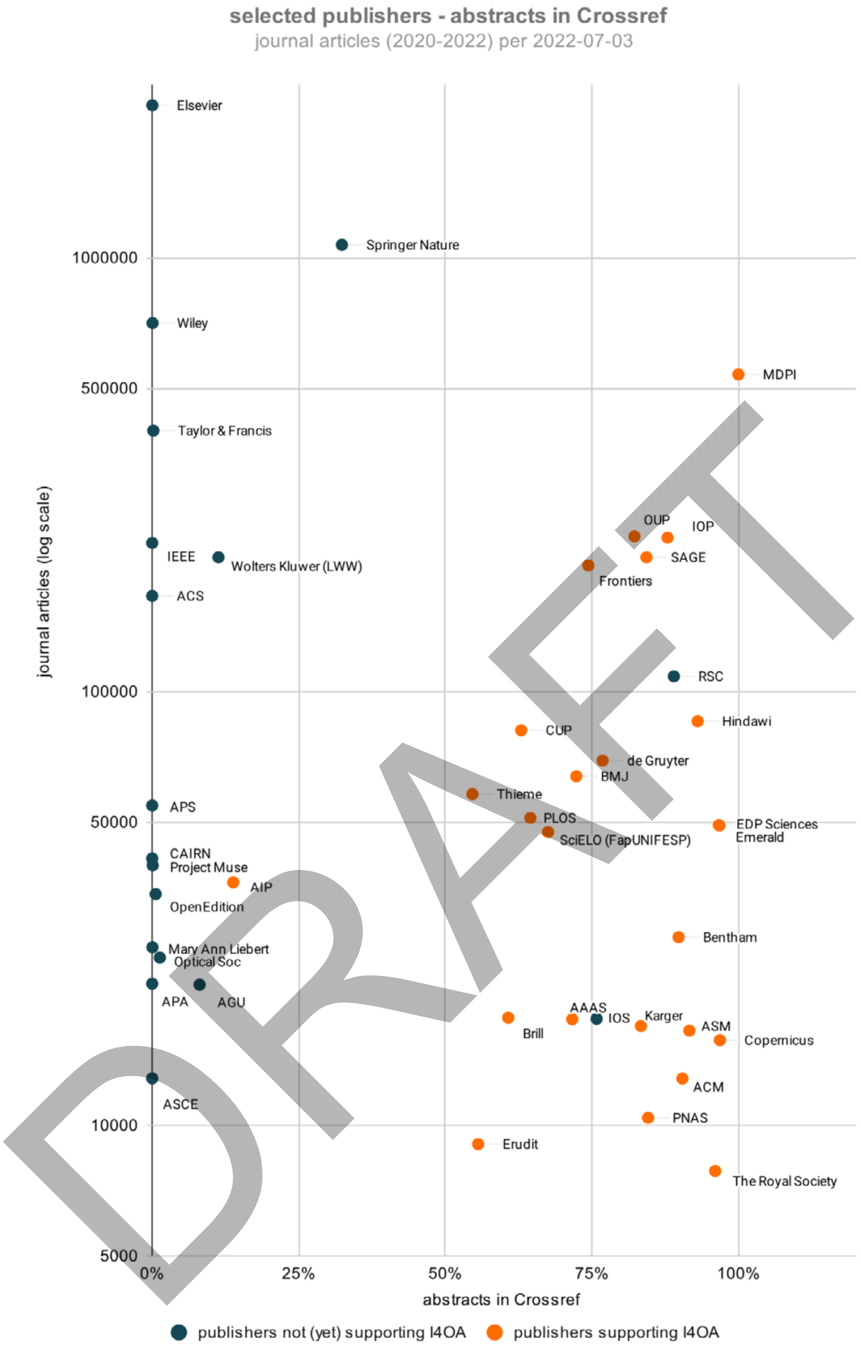


Figure 5.1.2. Abstract coverage in Crossref for journal articles – selected publishers (publication years 2020–2022)

For books, book chapters, and monographs, the picture which emerges is even more binary. Most larger publishers, even those that support I4OA and supply it with article abstracts, do not supply abstracts for long-form publications. Only very few (e.g. Oxford University Press (OUP), IGI Global) do it for almost all long-form content (Figure 5.1.3.). In addition to the lack of availability of abstracts for these publication types and/or the technical barriers in publisher workflows, there might also be a limited awareness among book publishers of the possibility and importance of supplying abstracts as part of Crossref metadata.

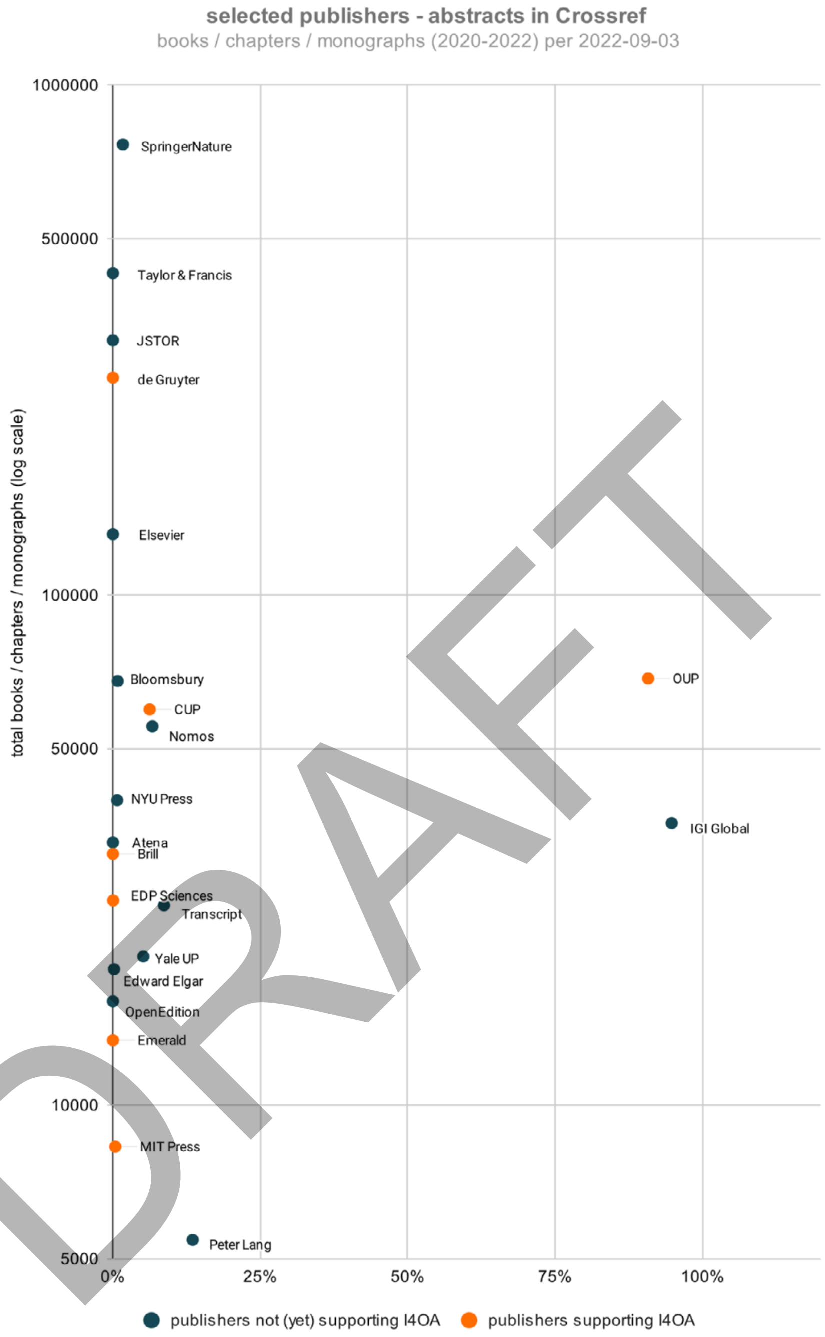
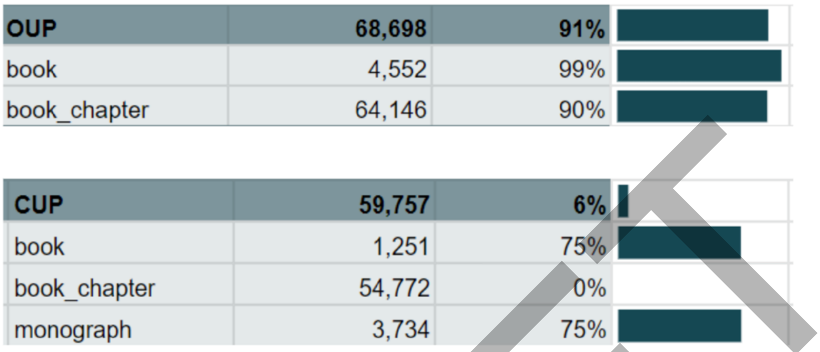


Figure 5.1.3. Abstract coverage in Crossref for books, book chapters, and monographs – selected publishers (publication years 2020–2022)

Even though books, book chapters, and monographs have, so far, been taken together, there are, between publishers, interesting differences in abstract coverage for these publication types which cannot be explained by the overall larger availability of abstracts for one publication type over another. For instance, while Oxford University Press (OUP) supplies abstracts for close to 100% of both their books and book chapters, Cambridge University Press (CUP) only supplies abstracts for books and monographs, not book chapters (Figure 5.1.4.). Incidentally, these differences in coverage for a given publisher also influence the publisher’s overall percentage of abstract coverage for long-form materials as depicted in Figure 5.1.3. For instance, because Cambridge University Press publishes many book chapters, for which they do not supply abstracts, their overall abstract coverage for long--form materials is only 6%, masking the fact that they do, in fact, provide abstracts for the majority of their books and monographs.

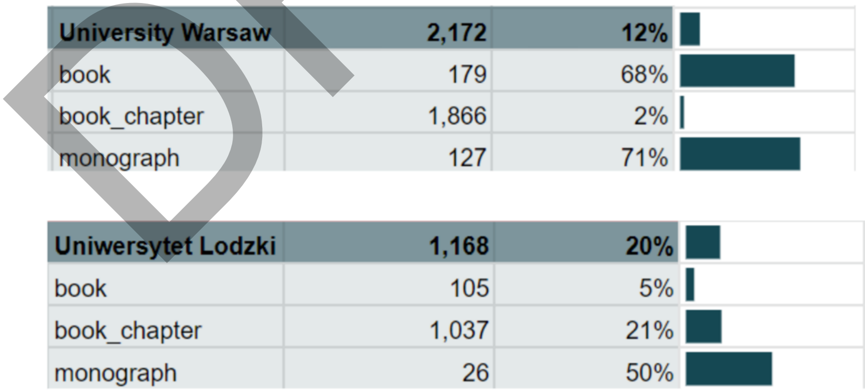


OUP	**68,698**	**91%**
book	4,552	99%
book_chapter	64,146	90%

CUP	**59,757**	**6%**
book	1,251	75%
book_chapter	54,772	0%
monograph	3,734	75%

Figure 5.1.4. Abstract coverage in Crossref for books, book chapters, and monographs separately – Oxford University Press (OUP) and Cambridge University Press (CUP) (publication years 2020–2022)

It is important not only to look at the larger (book) publishers, but also take into account the situation of mid-size and smaller publishers. One reason is the diversity in abstract coverage for long-form publications among smaller publishers. Two examples for Polish institutions are given in Figure 5.1.5.

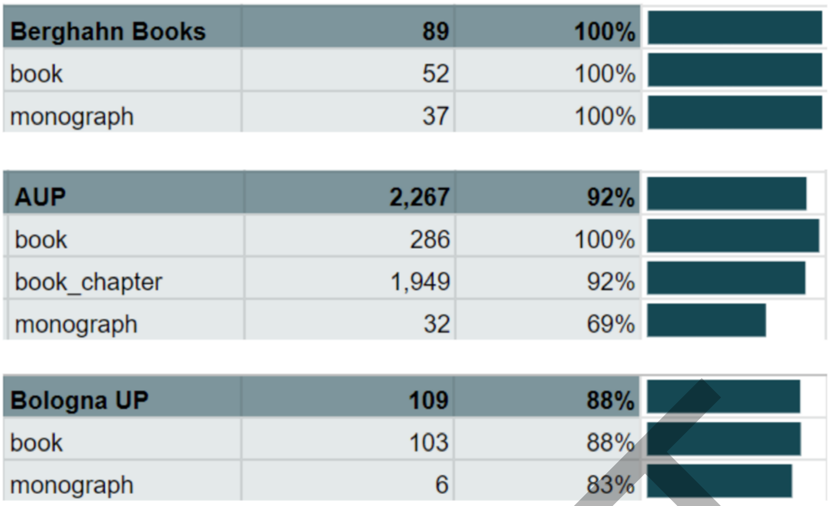


University Warsaw	**2,172**	**12%**
book	179	68%
book_chapter	1,866	2%
monograph	127	71%

Uniwersytet Lodzki	**1,168**	**20%**
book	105	5%
book_chapter	1,037	21%
monograph	26	50%

Figure 5.1.5. Abstract coverage in Crossref for books, book chapters, and monographs separately – University of Warsaw and Uniwersytet Lodzki (publication years 2020–2022)

There are also some excellent examples of good practice among mid-size and smaller publishers, for instance Berghahn Books (who also formally supports I4OA), Amsterdam University Press (AUP), and Bologna University Press, which all provide abstracts for the large majority of their long-form publications, irrespective of publication type (Figure 5.1.6).



Berghahn Books	**89**	**100%**
book	52	100%
monograph	37	100%

AUP	**2,267**	**92%**
book	286	100%
book_chapter	1,949	92%
monograph	32	69%

Bologna UP	**109**	**88%**
book	103	88%
monograph	6	83%

Figure 5.1.6. Abstract coverage in Crossref for books, book chapters, and monographs separately – Berghahn Books, Amsterdam University Press (AUP), and Bologna University Press (publication years 2020–2022)

The final reason to pay attention to abstract coverage for smaller publishers (and include them in awareness-raising and outreach actions) is the potential knock-on effect of the availability of abstracts in downstream applications. If attention is solely focused on increasing abstract availability for larger publishers, then those are the abstracts which will be available in downstream services as well, potentially leading to a serious bias in the visibility and use of abstracts.

For any Crossref member, a visual representation of metadata coverage (including abstract coverage) can be found in the *Crossref Participation Reports*[106] for various publication types, including journal articles and book chapters. Unfortunately, book abstracts are currently not included in the metadata types shown.

5.1.6. SOME ISSUES AROUND (OPEN) ABSTRACTS IN SSH

PERSISTENT IDENTIFIERS

The previous paragraphs discussed the workflow through which abstracts from articles, books, book chapters, and monographs, among others, could be made available for reuse via Crossref. As mentioned, this is an approach suitable for publications which have a Crossref DOI, which, however, is still less common for long-form publications than it is for journal articles. Even though the value of persistent identifiers has been well-established (both for identification and linking purposes, as well as for use of associated metadata), DOIs, and among them Crossref DOIs, are not the only persistent identifiers available,

[106] www.crossref.org/members/prep/

and solely focusing on Crossref DOIs therefore excludes many publications. Especially for smaller and non-Western publishers, the cost and required technical expertise are sometimes felt to be prohibitive for participation in Crossref (notwithstanding Crossref's efforts to remedy this through its sponsorship program and community engagement activities).

Another source of publication metadata (including abstracts) is available through harvesting information via OAI-PMH endpoints – both from publishers which enable this, and from institutional repositories. For instance, OpenEdition makes all its abstracts available through OAI-PMH, even if they do not provide them through Crossref. Many aggregators make use of OAI-PMH to harvest metadata, including OpenAIRE, BASE (Bielefeld Academic Search Engine), CORE (Connecting Repositories), and GoTriple itself. It would be interesting to compare the coverage and quality of metadata (including abstracts) retrieved via OAI--PMH and those retrieved via Crossref, both across the landscape of SSH output as well as for the specific publishers which supply metadata using both routes. In practice, infrastructures like GoTriple might consider enriching the metadata harvested through OAI-PMH with metadata from DOI registrars like Crossref for those publications which do have DOIs.

THE PRESENCE OF ABSTRACTS

As mentioned earlier, books, and monographs may often not include a formal abstract, as is customary for journal articles and book chapters. However, they often do have a short description which is available to help readers quickly appraise the topic and content; this, then, can function as an abstract. It will depend on the publisher, as well as the metadata schema being used, whether this text is included as an abstract in the metadata. For instance, Crosssref has a dedicated metadata field for abstracts, whereas the Dublin Core standard (often used for OAI-PMH and also mapped to the metadata scheme for DataCite DOIs) contains a more general metadata field labelled 'description' which can be used for abstracts, but also for other information (either with or without labels using controlled vocabulary).

The above illustrates the confluence of factors which affect the availability and the retrieval potential of abstracts in metadata for long-form publications – running the gamut from disciplinary norms to technical implementation.

LANGUAGE DIVERSITY

To properly account for language diversity, abstracts for publication outputs should be captured in all the languages in which they are made available. This is important for all areas of research, not just SSH, although language diversity is, arguably, most prominent in SSH. The availability of abstracts in multiple languages depends both on the inclusion criteria of research infrastructures and on the publishers' practices for including abstracts in metadata. In this respect, outreach and awareness efforts should purposely include non-English providers.

In the case of multilingual abstracts for one publication, it is important that all language versions are included in the publication's metadata. While metadata schemes do allow for this, in practice this is not always done correctly. One example, discussed during the booksprint for this report, concerned metadata for the same publication provided by the publisher in both Dublin Core[107] and JATS XML[108] formats, both of which contained an English and Polish abstract; while the Crossref metadata[109] for the same publication contained only the English abstract. This was not due to limitations in Crossref metadata, as Crossref encourages members to include titles and abstracts in multiple languages in their metadata[110] (with confirmation provided via personal communication that this is true for all content types which accept abstracts). Here again, awareness, and where needed, support to build providers' capacity, seems crucial to ensure the optimal availability of metadata.

OTHER SOURCES

As discussed earlier, publication metadata, including abstracts, can be sourced in multiple ways, including through metadata associated with DOIs and metadata harvested via OAI-PMH. Additionally, aggregators might have agreements to receive metadata from publishers directly and/or employ web scraping to collect and/or enrich metadata. These methods will vary in transparency and provenance, as well as in coverage and completeness for various aggregators, both as an end-user and as an aggregator or provider (like GoTriple) who is looking for sources to enrich existing metadata. For research output metadata in SSH, some potential sources include WorldCat, BASE, OpenAIRE, DataCite (which includes metadata from Zenodo), and OpenAlex.

As an example of the potential added value which multiple metadata sources represent, Figure 7 shows the abstract coverage for various publication types in OpenAlex, both for Crossref DOIs (Figure 7A) and for research output without DOIs in OpenAlex (Figure 7B). As can be seen, OpenAlex does have abstracts for a considerable proportion of journal articles and book chapters and, to a somewhat lesser extent, also books and monographs, for which Crossref does not have abstracts. In addition, OpenAlex can, to some extent, also be a source of abstracts for long-form publications without DOIs.

[107] See: bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:bibliotekanauki.pl:1968869.
[108] See: bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=jats&identifier=oai:bibliotekanauki.pl:1968869.
[109] See: api.crossref.org/works/10.17651/SOCJOLING.35.2.
[110] www.crossref.org/documentation/principles-practices/best-practices/multi-language/

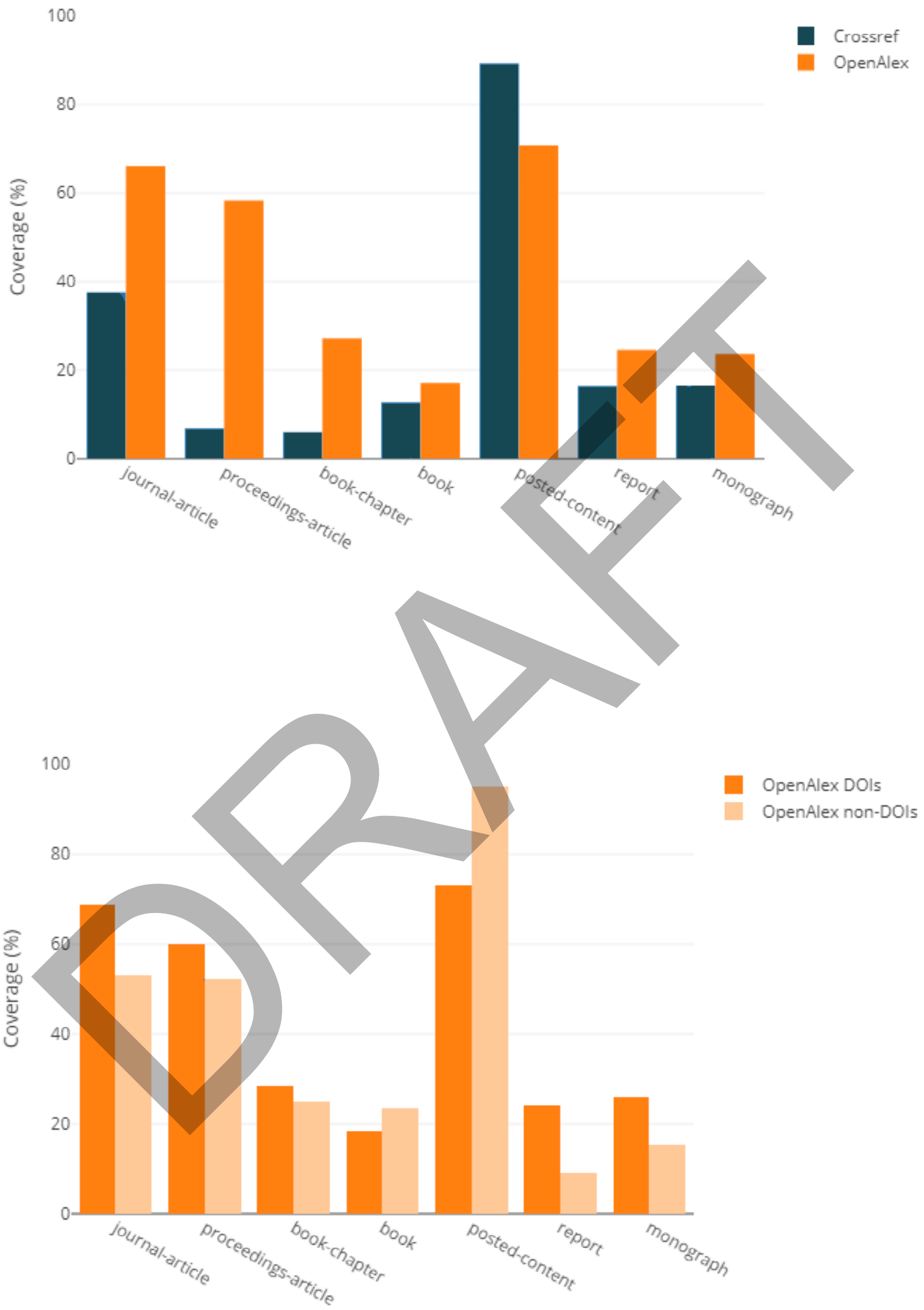


Figure 5.1.7A, 5.1.7B. Abstract coverage for Crossref DOIs in Crossref and OpenAlex (A); and for publications with or without DOIs in OpenAlex (B), for publication type (publication years 2020–2022, sampled June 2022).

5.1.7. CONCLUSIONS, AND IMPLICATION FOR GOTRIPLE

The open availability of abstracts and other metadata is, in principle, independent of the platform or infrastructure. Ideally, publishers and other providers would make abstracts available in machine-readable format as part of metadata for all their publications at all their endpoints in order to allow maximal availability and re-use.

In practice, both disciplinary norms (predominant publication types and the presence of abstracts for these publication types) and social, financial, and technical barriers limit the availability of abstracts as metadata.

For aggregators like GoTriple, the coverage of abstracts will therefore depend on how and where metadata are harvested (e.g. directly via OAI-PMH endpoints or via Crossref). A combined approach, where metadata are enriched by harvesting from additional sources, might provide added value.

To prevent, as much as possible, any bias in the metadata which are collected and made available, careful attention should be paid to the inclusion of the long tail of publishers as well as to the accommodation of language diversity, including multilingual abstracts.

5.2. THE HIGHWAY TO ABSTRACT. THE PRESENT AND FUTURE OF AUTOMATICALLY GENERATED ABSTRACTS

Agnieszka Karlińska[111], Cezary Rosiński[112], Nikodem Wołczuk[113]

5.2.1. MISSING ABSTRACTS

As stated in the previous subchapter, 'Abstracts in SSH', one of the challenges within the current SSH data ecosystem is the lack of abstracts in the form of structured metadata fields. This can happen when an abstract has not been created at all for a particular document, or when the information is present in the document (such as a PDF file), but has not been inserted into the metadata schema for the document's description. This is also the case for GoTriple providers – this will be closely investigated below, followed by a discussion concerning the possible technological challenges and solutions which could be applied to improve the platform's abstract coverage.

In the case of one of GoTriple's providers – the Library of Science (Bibliotekanauki.pl) – this happens in 20% of cases. So, around 90,000 documents (out of more than 490,000) are missing an abstract from their metadata description.

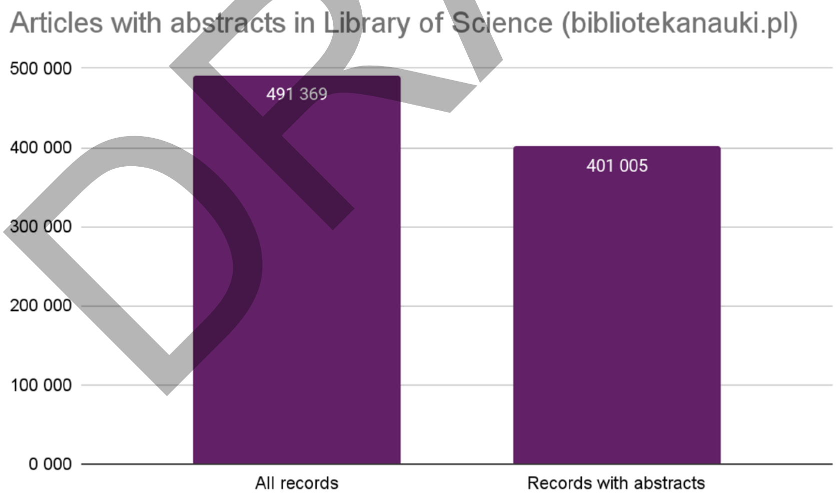


Figure 5.2.1. Articles with abstracts in the Library of Science (bibliotekanauki.pl, September 2022).

The case of the Library of Science, however, is an exception. The high number of released abstracts is due to the service's focus on aggregating current scientific materials, which are usually better equipped with metadata than older resources. Services which collect

[111] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-4846-7086
[112] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-6136-7186
[113] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-4303-2016

materials from a wider time range do not have such a rich representation of abstracts in their data. Old materials requiring digitisation often suffer from a lack of such metadata. An example of such a service is the Digital Repository of Scientific Institutes (RCIN[114]), which, among its 105,287 records whose designation type is 'text', has only 8,984 indexed abstracts – that is a mere 8.5%.

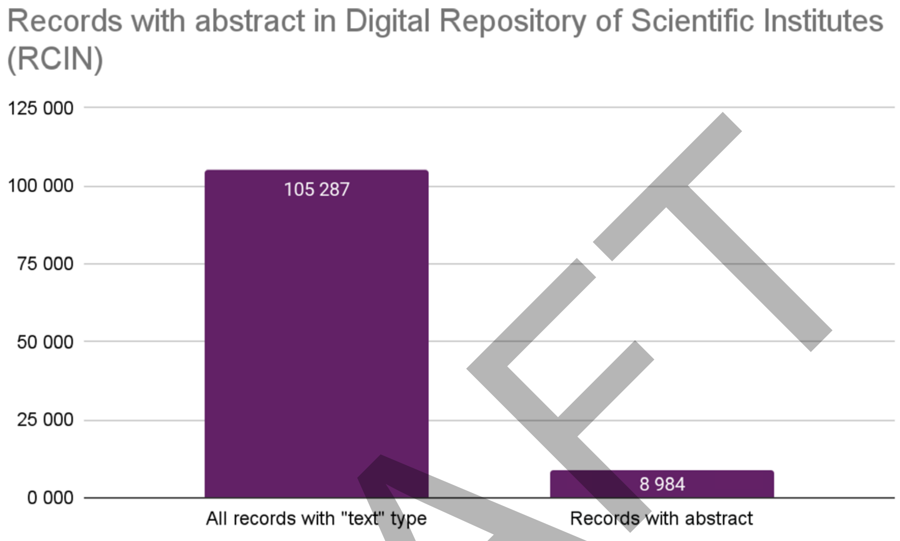


Figure 5.2.2. Records with abstracts in the Digital Repository of Scientific Institutes (RCIN) (rcin.org.pl, September 2022).

The GoTriple dataset demands the inclusion of not only contemporary scientific output, but also older materials originating from the GLAM sector, where abstract coverage is limited.

5.2.2. MAKING UP FOR MISSING ABSTRACTS

There are two complementary approaches to acquiring missing abstract data: metadata extraction and metadata generation[115] (Figure 5.2.3.).

The first strategy, when applied to a digitised item without abstract metadata, involves optical character recognition, optical layout recognition, and semantic segmentation procedures to create a textual layer for the digitised document, identify the abstract, and extract its content, followed by adding the extracted abstract as a new metadata type.

The second strategy leads to a different set of difficulties. Where an abstract is not available – and this kind of situation may be common for older documents – an abstract needs to be generated from scratch. The traditional method, which involves authorial factoring

[114] rcin.org.pl/

[115] The manual creation of missing abstracts has been left out here due to how time- and cost-consuming this solution would be.

or even having professional bibliographers or librarians adding the new piece of metadata, is rather unlikely to happen due to financial constraints. Therefore the only possible solution is to use natural-language programming (NLP) and machine learning to do so.

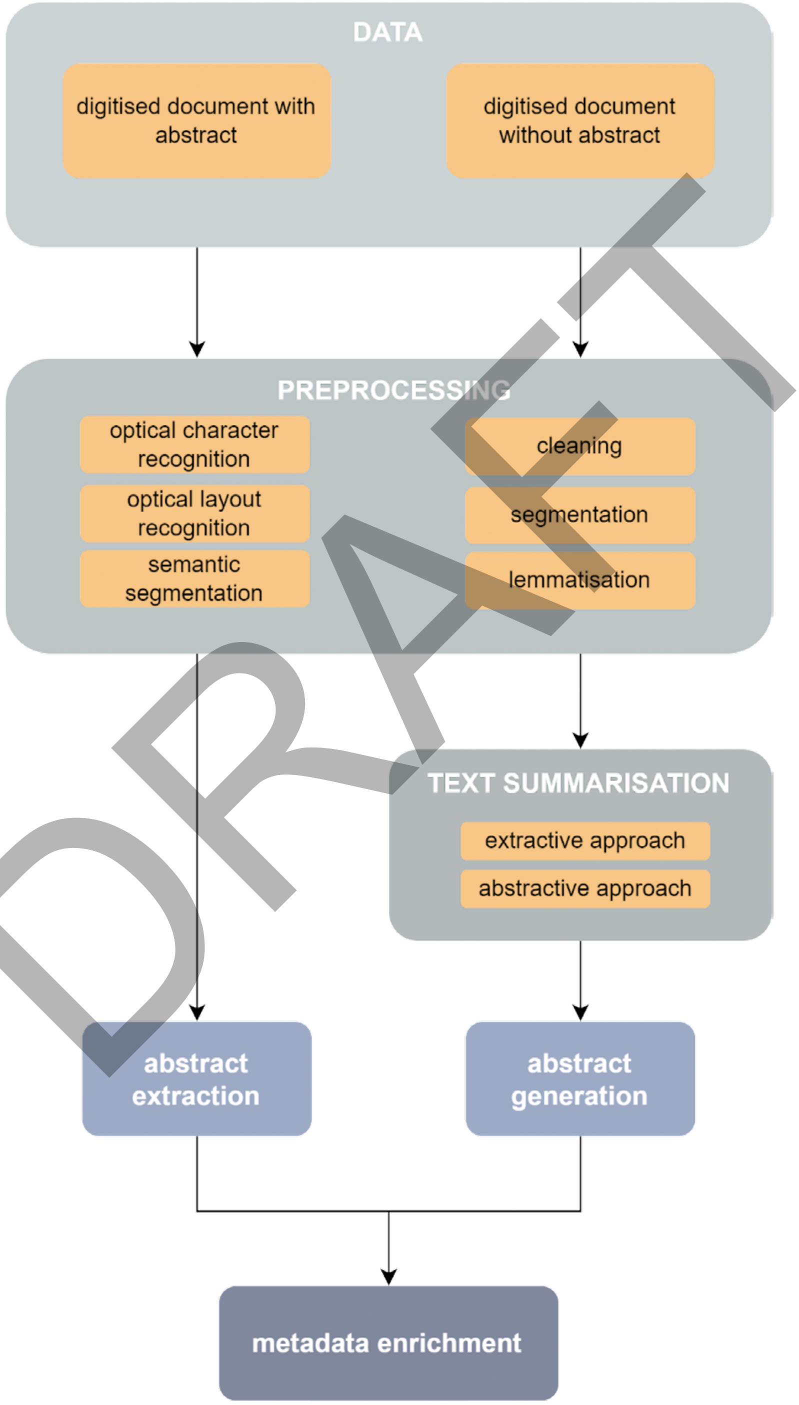


Figure 5.2.3. Workflow for acquiring missing abstracts

In NLP, the task of creating a short text which is the best possible semantic representation of the original document – i.e. includes the most important and relevant information – is referred to as 'automatic text summarisation'. Research on text summarisation has been carried out since the 1950s. The goal is to develop a system capable of creating machine-generated summaries which are equivalent to human-generated ones[116].

As with keyword generation systems, text summarisation systems are classified into two types: extractive and abstractive[117]. In extractive approaches, paragraphs, sentences, and phrases are obtained from the source document and then put together to produce the summary[118]. In abstractive techniques, text is paraphrased and the summary contains sentences and phrases which are different to those in the source document. Abstractive methods are usually highly complex as they require full-text comprehension and extensive natural language processing[119]. Therefore, the research community has long focused primarily on improving extractive techniques[120].

As such, extractive text summarisation has become a popular field within NLP, resulting in a whole range of approaches which implement several machine learning and optimisation techniques. They perform different types of clustering, which are usually aimed at extracting the most diverse topics from the original document[121]. For this reason, extractive summarisation is said to rely solely on sentence scoring to maximise topical coverage and minimise redundancy, while coherence (i.e. the extent to which ideas are related to each other) and cohesion (i.e. textual fluency) are considered only in cases of abstractive summaries[122]. This is not always the case, as there are extractive methods which allow for both sentence scoring and text fluency[123].

Since extractive text summarisation is a relatively well-developed field, current research is increasingly shifting toward abstractive and hybrid text summarisation methods[124]. A promising area is the recent advances in neural methods, which provide a feasible framework for obtaining an abstract representation of the meaning of the original text[125]. At the core

[116] M. Gambhir and V. Gupta. 'Recent Automatic Text Summarization Techniques: A Survey', *Artif Intell Rev* 47, (2017): 1–66.

[117] G. Sharma and D. Sharma, 'Automatic Text Summarization Methods', *A Comprehensive Review. SN COMPUT. SCI* 4 (2022): 33.

[118] A. Khan and N. Salim, 'A review on abstractive summarization methods', *Journal of Theoretical and Applied Information Technology* 59.1 (2014): 64–72.

[119] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and I. M. De Rosal Setiadi, 'Review of Automatic Text Summarization Techniques & Methods', *Journal of King Saud University – Computer and Information Sciences* 34 (4) (2020): 1029–1046.

[120] Gambhir and Gupta, 'Recent automatic text...'

[121] I. Okulska, 'Team Up! Cohesive Text Summarization Scoring Sentence Coalitions'. In *Artificial Intelligence and Soft Computing. ICAISC. Lecture Notes in Computer Science*, vol 12416, eds L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J. M. Zurada (Springer, Cham, 2020).

[122] C. C. Aggarwal, *Machine Learning for Text* (Springer, Cham, 2018).

[123] Okulska, 'Team Up! Cohesive Text Summarization ...'

[124] Widyassari, Rustad, Shidik, Noersasongko, Syukur, Affandy, De Rosal Setiadi, 'Review of automatic text...'

[125] H. Lin, V. Ng, 'Abstractive Summarization: A Survey of the State of the Art', *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01), (2019): 9815–9822.

of abstractive summarisation are three pipelined tasks: information extraction, which obtains useful information from text using noun or verb phrases; content selection, which works by selecting a subset of important phrases from the extracted text; and surface realisation, which combines selected words or phrases in a sequence by using grammatical rules and lexicons[126]. Abstractive text summarisation methods are divided into structure-based approaches, semantic-based approaches, and deep learning-based[127] and linguistic approaches; for example, for linguistic approaches, information-based or tree-based methods, and semantic approaches such as template-based methods or ontology-based methods[128]. Neural network-based text summarisation is considered to be state-of-the-art[129].

All summarisation methods and models, both extractive and abstractive, share the common goal of generating summaries which are informative, non-redundant, and coherent. Despite several improvements, they still have some issues and challenges. One of the key challenges for abstractive text summarisation is ensuring factual consistency, i.e., including in the summary only those statements which can be derived directly from the original document[130]. Recent studies have shown that about 30% of abstractive summaries which are generated by neural network sequence-to-sequence models involve fact fabrication, such as, including names which never appeared in the original text[131]. The main challenge from the perspective of generating abstracts from scientific texts is the strong dependence on the domain (in this case the scientific discipline), which makes it impossible to develop a fully universal approach. Multilingual text summarisation is also a challenge. New developments are emerging in this field, but there are still a limited number of datasets for low/ mid-resource languages[132].

[126] Lin and Ng, 'Abstractive Summarization...'

[127] S. Gupta and S. K Gupta, 'Abstractive Summarization: An Overview of the State of the Art', *Expert Systems with Applications* 121 (2019): 49–65.

[128] Widyassari, Rustad, Shidik, Noersasongko, Syukur, Affandy, and De Rosal Setiadi 'Review of automatic text...'

[129] Gupta and Gupta, 'Abstractive summarization...'

[130] F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. Xiang, 'Entity-level Factual Consistency of Abstractive Text Summarization', *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* Main Volume (2021): 2727–2733. [Online. Association for Computational Linguistics.]

[131] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, 'Evaluating the Factual Consistency of Abstractive Text Summarization', Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), (2020): 9332–9346.

[132] T. Hasan, A. Bhattacharjee, M. Saiful Islam, K. Mubasshir, Y-F. Li, Y-B. Kang, M. Sohel Rahman, and R. Shahriyar, 'XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages', *Findings of the Association for Computational Linguistics*: ACL-IJCNLP, (2021): 4693–4703. [Online. Association for Computational Linguistics.]

5.2.3. WHO NEEDS THIS DATA AND FOR WHAT

The GoTriple dataset must have the best possible abstract coverage, as it must facilitate both the needs of the end users of its public interface, and the value they bring to understanding society, culture, and science. The reasons for having abstracts have already been discussed in detail in previous sections. Abstracts are critical for enriching the workflow which GoTriple employs, especially keyword attribution and generating data visualisations such as Streamgraphs and Knowledge Maps.

From this point of view it is critical to tackle both abstract extraction and generation, especially when we take into account the fact that many types of documents lack them – such as books, book chapters, and non-article documents within journals – as was discussed by Bianca Kramer in the section entitled 'Abstracts in SSH'. The question remains – who should tackle this challenge and at what stage of the document's lifecycle? GoTriple is a high-level aggregator of the content created by its providers and other aggregators. The need for larger abstract coverage is also well-understood further down the pipeline, by scientific journals and publishers.

In a survey[133] conducted within the Dariah.lab[134] project, respondents (Polish scientific journals and publishers) answered questions about their knowledge of automatic metadata extraction tools and their willingness to use them. Although 70% of publishers and 60% of scientific editors responded that they added or enriched metadata in their publications, none of the publishers, and only two scientific journals, used automatic techniques. What is more, a large group of the respondents was not even aware of the existence of automatic metadata extraction tools. The vast majority of editors expressed interest in using new, automatic metadata extraction software, provided, however, that it was free. Interest in paid software was very low in this group (Figure 5.2.4.).

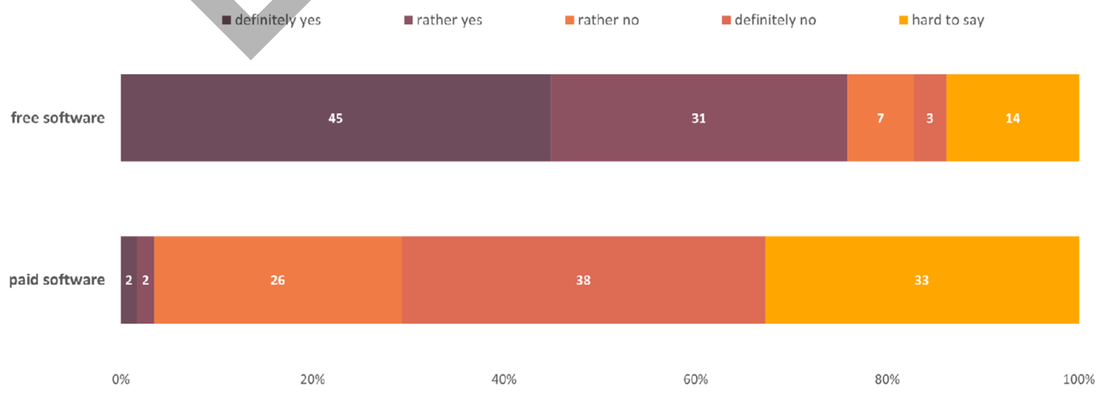


Figure 5.2.4. Interest among editors of Polish scientific journals in using new automatic metadata extraction software

[133] operas.pl/2022/04/22/pytamy-redakcje-czasopism-i-wydawnictw-naukowych-o-technologie/
[134] lab.dariah.pl/en/

More interest was shown by scientific publishers (Figure 5.2.5). Over 80% of these respondents said they would use free metadata extraction software and 24% would be willing to pay for such a solution.

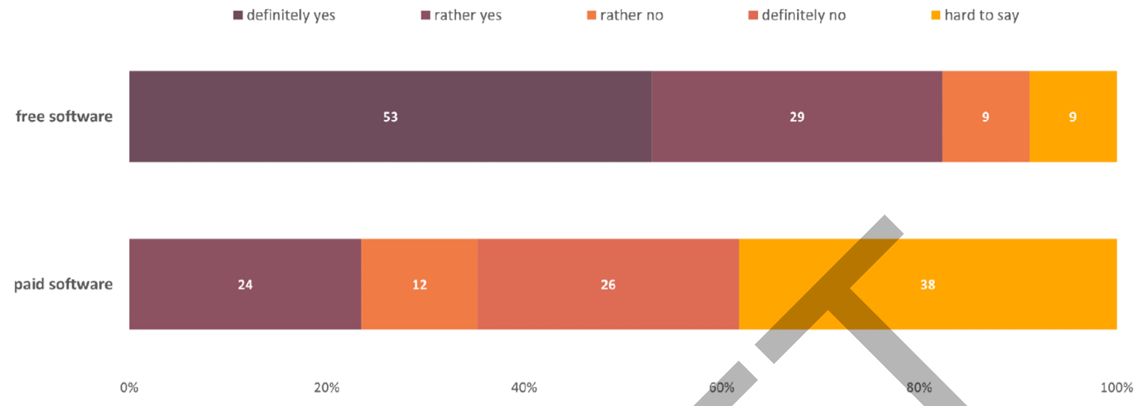


Figure 5.2.5. Interest among Polish scientific publishers in using new automatic metadata extraction software

None of the respondents used abstract generation tools. As in the case of metadata extraction, a large group of the respondents was not even aware of the existence of such solutions.

For scientific journal editors, there was slightly less interest in using a new service to automatically generate abstracts than there was for metadata extraction tools, but more than half of respondents were still willing to use such a service – provided it was free (Figure 5.2.6.).

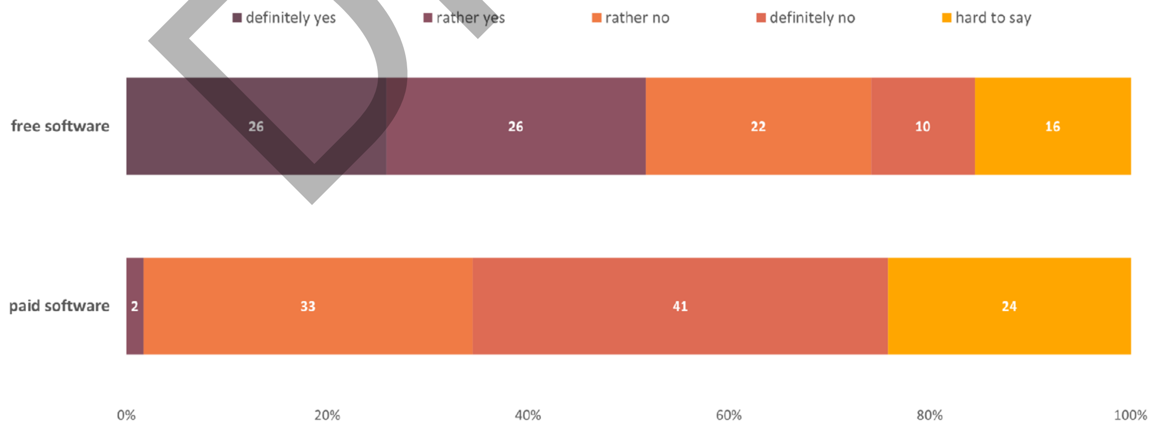


Figure 5.2.6. Interest among editors of Polish scientific journals in using new abstract generation tools

Scientific publishers again showed a little more interest – 70% of respondents would use a free service and 12% a paid one (Figure 5.2.7).

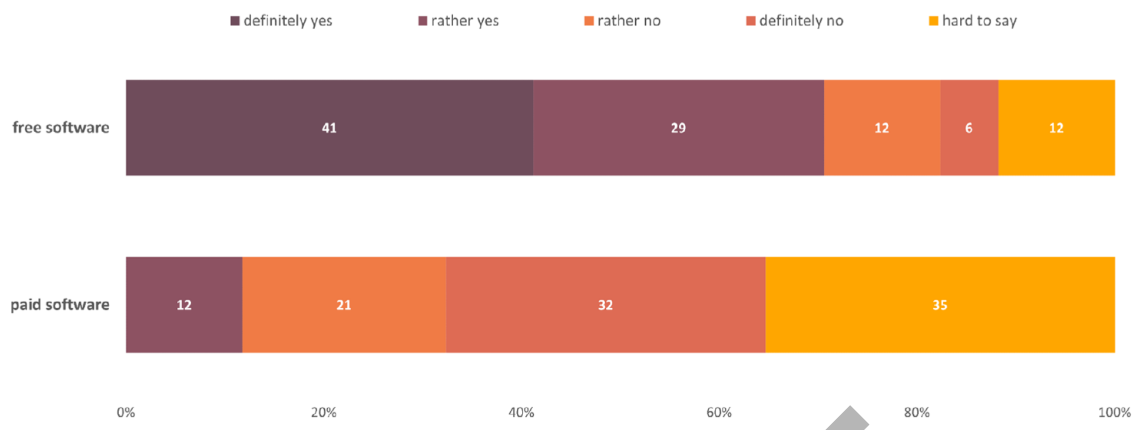


Figure 5.2.7. Interest among Polish scientific publishers in using new abstract generation tools

5.3. GUIDELINES FOR ABSTRACTS AS HUMANITIES RESEARCH DATA IN THE CONTEXT OF TRIPLE – A SUMMARY

1. The humanities need to **take full advantage of the existing infrastructure for open metadata** – such as Open Abstracts – which includes, for example, registering abstracts through Crossref services. Through this, **humanities data becomes more accessible and easier to link to different data spaces**.

2. As different engines and algorithms rely on abstracts – for example, for automated keyword extraction – **there is a need to add abstracts if they do not exist**. This might also apply to non-scientific documents such as cultural documents, which are important for the humanities. Hence the relevance of automated approaches to text summarisation and the possibilities of applying these methods to the creation of large research datasets.

3. With the development of NLP methods and resources such as large corpora, **text summarisation is able to provide insight into those documents which cannot be accessed in their entirety** (provided copyright considerations and regulations regarding data mining allow for it).

4. The multilingual aspects of abstracts are especially important for the humanities; this means both capturing all language versions of abstracts, but also leveraging the potential for the automated translation of abstracts into all necessary languages (current developments in NLP and AI might be of great help here).

6. CITATIONS

6.1. TRANSPARENCY MEETS OPEN CITATIONS

Silvio Peroni[135]

Within the scholarly ecosystem, a bibliographic citation is a conceptual, directional link from a citing entity to a cited entity which is used to acknowledge or ascribe credit for the contribution made by the author(s) of the cited entity. Citations are one of the core elements in scholarly communication. They enable our independent research endeavours to be integrated into a global graph of relationships which can be used, for instance, to analyse how scholarly knowledge develops over time, assess scholars' influences, and make wise decisions about research investment.

However, as citation data (i.e. pieces of factual information aimed at identifying entities and the relationships between them) are of great value to the scholarly community, it is a 'scandal'[136] that they have not been recognised as part of the commons. Indeed, only recently have we seen some efforts – such as those already discussed in the previous chapter, 'Initiative for Open Citations' (I4OC[137]) – which have tried to change the behind-the-paywall status quo enforced by the companies controlling the major citation indexes used worldwide, by convincing scholarly publishers to support unrestricted availability of scholarly citation data by publishing them in suitable open infrastructures such as Crossref[138] and DataCite[139].

Of course, as with many other kinds of data, putting bibliographic and citation data behind a paywall is a threat to enabling the full reproducibility of the research studies based upon them (e.g. in bibliometrics, scientometrics, and science of science domains), even when such studies are published in open access articles. For instance, the results of a recent open access article by Spinaci et al. (2022)[140], published in *Digital Scholarship in the Humanities*, which aimed to analyse the citation behaviour of digital humanities (DH) research across different proprietary and open citation databases, are not fully reproducible, since the majority of the databases used – namely Scopus, Web of Science, and Dimensions – do not make their bibliographic and citation data openly available.

In addition, the coverage of publications and their related citations in specific disciplines, in particular those within the social sciences and humanities (SSH), is not adequate when

[135] University of Bologna, OpenCitations, orcid.org/0000-0003-0530-4305

[136] D. Shotton, 'Open citations', *Nature* 502 (7471) (2013): 295–297, dx.doi.org/10.1038/502295a

[137] i4oc.org

[138] crossref.org

[139] datacite.org

[140] G. Spinaci, G. Colavizza, and S. Peroni, 'A Map of Digital Humanities Research Across Bibliographic Data Sources', *Digital Scholarship in the Humanities*, (2022), doi.org/10.1093/llc/fqac016

compared to other fields[141]. Usually, this is due to the limited availability of born-digital publications, as well as by the wide variety of publication languages, publication types (e.g. monographs), and the complex referencing practices which may limit automatic processing and citation extraction. As a side effect, such partial coverage may result in considerable bias when analysing SSH disciplines compared to STEM disciplines, which usually have better coverage in existing citation databases.

All these scenarios have at least one other negative effect on the area which is strictly concerned with research assessment and often uses quantitative metrics based on citation data to evaluate articles, people, and institutions. Indeed, the unavailability and partial coverage of bibliographic and citation data create an *artificial* barrier to the transparency of the processes used to decide the careers of scholars in terms of research, funding, and promotions.

In the past few years, several initiatives from around the world have highlighted the importance of reforming research assessment, such as those summarised in Figure 6.1.1: the French *National Plan for Open Science*[142]; the San Francisco Declaration on Research Assessment[143]; the *Leiden Manifesto for Research Metrics*[144]; and the recent proposal for reforming the research assessment system by the European Commission (2021)[145], which is being formalised under the Coalition for Advancing Research Assessment (COARA)[146]. All these initiatives agree on some essential characteristics which are necessary to have a trustworthy assessment system:

- The system must be open and transparent by providing *machine-readable*, *unrestricted*, and *reusable* data, as well as methods for calculating the metrics used in research assessment exercises;
- The control and ownership of the crucial infrastructures and tools used to retrieve, use, and analyse such data within research assessment systems must be left to the *research community* instead of commercial players.

[141] G. Colavizza, S. Peroni, and M. Romanello, 'The Case for the Humanities Citation Index (HuCI), A Citation Index by the Humanities, for the Humanities', *International Journal on Digital Libraries*, (2022), doi.org/10.1007/s00799-022-00327-0; A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E.Delgado López-Cózar, 'Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A Multidisciplinary Comparison of Coverage via Citations', *Scientometrics* 126(1), (2020): 871–906, doi.org/10.1007/s11192-020-03690-4; V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr, 'The journal coverage of Web of Science, Scopus and Dimensions: A Comparative Analysis', *Scientometrics* 126(6) (2021): 5113–5142, doi.org/10.1007/s11192-021-03948-5; M. Visser, N. J. van Eck, L. Waltman, 'Large-scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic', *Quantitative Science Studies* 2(1) (2021): 20–41, doi.org/10.1162/qss_a_00112

[142] www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018/

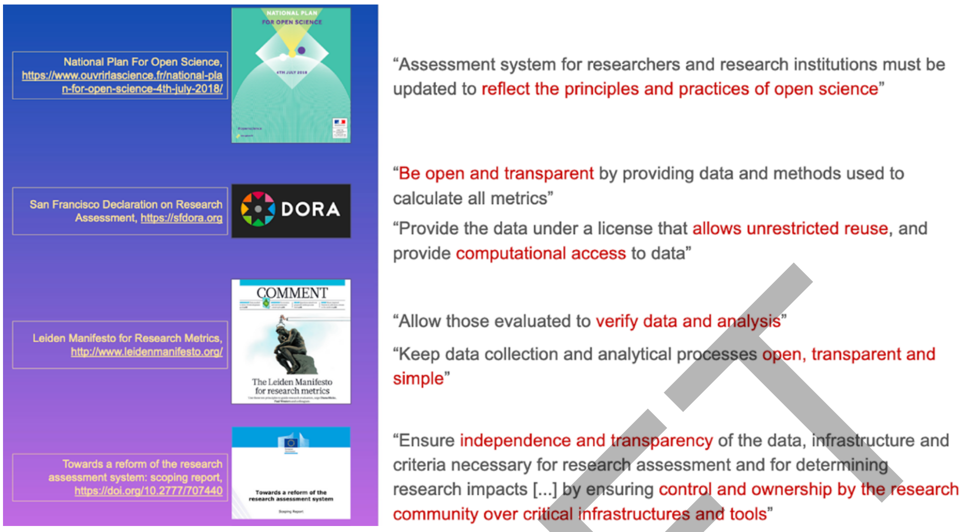
[143] sfdora.org

[144] www.leidenmanifesto.org/

[145] European Commission, *Towards a Reform of the Research Assessment System. Scoping report*. (KI-09-21-484-EN-N), Publications Office, (2021), doi.org/10.2777/707440.

[146] coara.eu/

Thus, the leading guideline which can be drawn from this is to follow open science practices, and not only when performing research.



National Plan For Open Science, https://www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018/

"Assessment system for researchers and research institutions must be updated to reflect the principles and practices of open science"

San Francisco Declaration on Research Assessment, https://sfdora.org

"Be open and transparent by providing data and methods used to calculate all metrics"

"Provide the data under a license that allows unrestricted reuse, and provide computational access to data"

Leiden Manifesto for Research Metrics, http://www.leidenmanifesto.org/

"Allow those evaluated to verify data and analysis"

"Keep data collection and analytical processes open, transparent and simple"

Towards a reform of the research assessment system: scoping report, https://doi.org/10.2777/707440

"Ensure independence and transparency of the data, infrastructure and criteria necessary for research assessment and for determining research impacts [...] by ensuring control and ownership by the research community over critical infrastructures and tools"

NATIONAL PLAN
FOR OPEN SCIENCE
4TH JULY 2018
DORA
COMMENT
The Leiden Manifesto for research metrics
Towards a reform of the research assessment system
Scoping Report

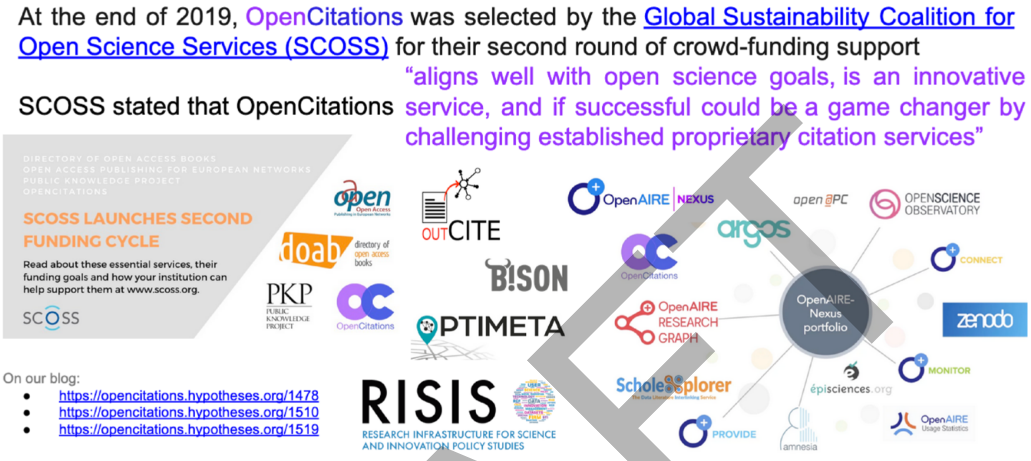
Figure 6.1.1. Some of the initiatives pushing for reform of the principles behind research assessment systems.

Within this context, OpenCitations[147] play an important role, acting as a key infrastructure component for global open science, and pushing to actively involve universities, scholarly libraries and publishers, infrastructures, governments and international organisations, research funders, developers, academic policy-makers, independent scholars, and ordinary citizens. The mission of OpenCitations is to harvest and openly publish accurate and comprehensive metadata which describes the world's academic publications and the scholarly citations which link them, with the greatest possible global coverage and subject scope, encompassing both traditional and non-traditional publications, and with a breadth and depth which surpasses existing sources of such metadata, while maintaining the highest standards of accuracy, and accompanying all records with a rich provenance of information, while providing this information – both human-readable form and in interoperable machine-readable linked open data formats – under open licences, at zero cost, and without restriction for third-party analysis and re-use.

For OpenCitations, 'open' is the crucial value and the final purpose. The distinctive mark and founding principle of OpenCitations is that everything it provides – data, services and software – is open and free, and will always remain so. OpenCitations fully espouses the aims and vision of the UNESCO Recommendations on Open Science, complies with the FAIR data principles, and promotes and practises the Initiative for Open Citations' recommendation that citation data, in particular, should be *structured*, *separable*, and *open*.

[147] opencitations.net

The most important collection of such open citation data is COCI, OpenCitations' Index of Crossref open DOI-to-DOI citations[148]. The latest release, dated August 2022, contains more than 1.36 billion citation links between more than 75 million bibliographic entities, which can be accessed programmatically using its REST API, queried via the related SPARQL endpoint, and downloaded in full as dumps in different formats (CSV, JSON, and RDF).



At the end of 2019, OpenCitations was selected by the Global Sustainability Coalition for Open Science Services (SCOSS) for their second round of crowd-funding support

SCOSS stated that OpenCitations "aligns well with open science goals, is an innovative service, and if successful could be a game changer by challenging established proprietary citation services"

DIRECTORY OF OPEN ACCESS BOOKS
OPEN ACCESS PUBLISHING FOR EUROPEAN NETWORKS
PUBLIC KNOWLEDGE PROJECT
OPENCITATIONS

SCOSS LAUNCHES SECOND FUNDING CYCLE

Read about these essential services, their funding goals and how your institution can help support them at www.scoss.org.

On our blog:

- https://opencitations.hypotheses.org/1478
- https://opencitations.hypotheses.org/1510
- https://opencitations.hypotheses.org/1519

Figure 6.1.2. Collaborations between OpenCitations and other open science infrastructures and services.

In addition to the publication of citation data, considerable effort has been dedicated to collaborating with other open science infrastructures working in the scholarly ecosystem, as summarised in Figure 6.1.2. Since 2020, OpenCitations has benefited significantly from the scholarly community, as, in 2019, the Global Sustainability Coalition for Open Science Services (SCOSS[149]) identified OpenCitations as being a scholarly infrastructure worthy of financial support. This community funding allowed OpenCitations to appoint people who were dedicated to the administration, communication, community development, and maintenance and improvement of the OpenCitations software and the computational infrastructure on which it runs. In addition, OpenCitations started its involvement with OpenAIRE[150] and the European Open Science Cloud (EOSC[151]), and is now collaborating with other funded projects such as RISIS 2[152], OUTCITE[153], OPTIMETA[154] and B!SON[155].

While OpenCitations currently provides a good set of citation data which is already approaching parity with other commercial citation databases[156] and which has already been used in a few studies for research purposes, there is still room for improvement. Currently, the citations included in the OpenCitations Indexes come mainly from Crossref data, one

[148] w3id.org/oc/index/coci
[149] scoss.org
[150] www.openaire.eu
[151] eosc-portal.eu
[152] www.risis2.eu
[153] excite.informatik.uni-stuttgart.de
[154] projects.tib.eu/optimeta/en
[155] projects.tib.eu/bison/en/project
[156] A. Martín-Martín, 'Coverage of open citation data approaches parity with Web of Science and Scopus', *OpenCitations* (blog), 27/10/2021, opencitations.hypotheses.org/1420.

of the biggest open reference providers. However, Crossref does not cover all publishers of DOI-based resources. Indeed, other DOI providers, in some cases, expose citation relations in their metadata, such as DataCite[157]. In addition, DOI-based publications represent only a limited set of all the bibliographic entities published in the scholarly ecosystem. Other identifier schemas have been used to identify bibliographic entities, and, for some publications, no identifiers exist at all.

Thus, to address these two issues, OpenCitations is working to expand its coverage in two different directions. On the one hand, OpenCitations is developing two new citation indexes for open references based on the holdings of DataCite and the National Institute of Health Open Citation Collection[158], which, together with COCI, will be cross-searchable through the Unifying OpenCitations REST API[159].

On the other hand, OpenCitations has started working to create a new database entitled OpenCitations Meta, which will provide three major benefits. First, it will store in-house bibliographic metadata for both the citing and cited entities involved in all OpenCitations indexes, including author identifiers using ORCID and VIAF identifier schemes where available. Second, it will provide better query performance than the present API system, which obtains bibliographic metadata on-the-fly by live API calls to external services such as Crossref and DataCite's APIs. Finally, it will permit the indexing of citations involving entities which lack DOIs by providing them with OpenCitations Meta Identifiers.

This last collection, combined with automatic tools for extracting citations from digital formats, is crucial for increasing the coverage of underrepresented disciplines and fields in bibliographic databases, such as SSH publications. One of OpenCitations' goals is to reduce this gap in citation coverage by setting up crowdsourcing workflows for ingesting missing citation data from the scholarly community (e.g. libraries and publishers). In the future, another contribution will be to set up tools for the automatic extraction of citations which can also support small and local publishers – assets which are crucial for SSH research – who may have difficulties in carrying out citation extraction tasks on their own, as using and maintaining a tool for this (or paying a company to address those tasks on behalf of the publisher) requires extra costs beyond the publishers' finances.

To conclude – OpenCitations is one piece in a puzzle which is working to change existing scholarly practices in order to create an open and inclusive future for science and research, in which the scholarly community owns and is responsible for its own data.

[157] datacite.org

[158] B. I. Hutchins, K. L. Baker, M. T. Davis, M. A. Diwersy, E. Haque, R. M. Harriman, T. A. Hoppe, S. A. Leicht, P. Meyer, and G. M. Santangelo, 'The NIH Open Citation Collection, A Public Access, Broad Coverage Resource', *PLOS Biology* 17(10) (2019), doi.org/10.1371/journal.pbio.3000385

[159] w3id.org/oc/index/api/v1/

6.2. CITATION DATA AND GOTRIPLE'S DATA PROVIDERS

Cezary Rosiński[160], Tomasz Umerle[161], Nikodem Wołczuk[162]

6.2.1 MISSING OR LOST CITATIONS?

The availability of citations is dependent on the data format offered by the provider. In the case of one of GoTriple's providers, the Library of Science[163], Dublin Core data does not expose citation data. On the other hand, the citations are present in JATS format in almost 60% of the articles:

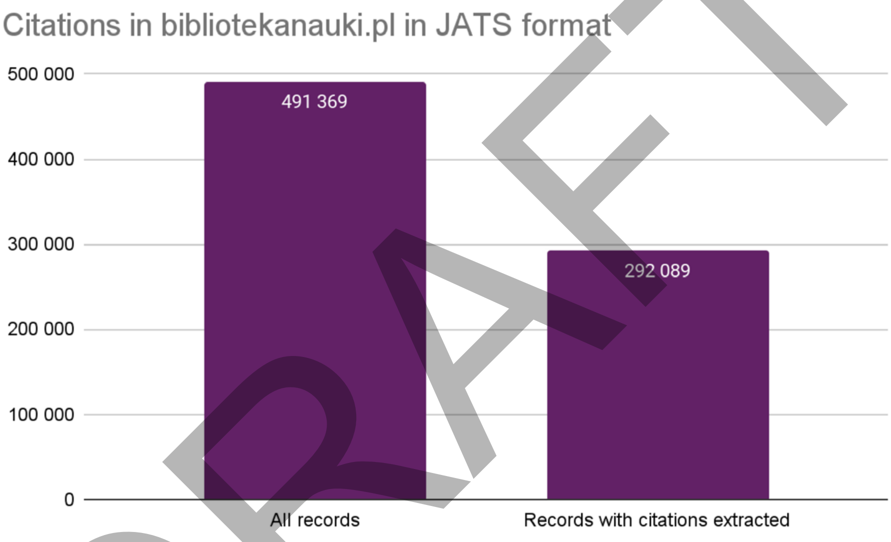


Figure 6.2.1. Citations in the Library of Science in JATS format (bibliotekanauki.pl, September 2022).

The presence of citations in local services does not guarantee, however, that this information is carried to other services, for example Crossref. The analysis of the Library of Science's data which was prepared for the purposes of this report confirms that even though a certain metadata entity (record of a document) exists in Crossref, not every piece of original information is preserved there. For instance, metadata in JATS format[164] in the Library of Science may be equipped with extracted citations, but the record presented in Crossref[165] will not contain such information. It is worth mentioning that the exact same record in the Library of Science database, in JATS format, has extracted citations, but it does not contain

[160] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-6136-7186.
[161] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-7335-0568.
[162] Institute of Literary Research, Polish Academy of Sciences (IBL PAN), orcid.org/0000-0002-4303-2016.
[163] bibliotekanauki.pl
[164] See: bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=jats&identifier=oai:bibliotekanauki.pl:1968869.
[165] See: api.crossref.org/v1/works/10.17651%2FSOCJOLING.35.2.

that part of metadata which is in DC format[166], even though it is technically possible. In the end, we are dealing with an open access aggregator of scientific output which already exposes citation data, but only locally.

From the perspective of the GoTriple dataset, this means that, even if GoTriple aims to aggregate and expose citation data, 1) reliance on Dublin Core data is not a viable solution, 2) analysing the different local formats offered by multiple providers so as to identify which one of them stores citation data is a complex task, and 3) relying on dedicated services which aggregate citation data (such as Crossref) does not guarantee that the local metadata will be exposed by the aggregator.

6.2.2. MISSING CITATIONS: EXTRACTION AND ENRICHMENT

When structured citation data is truly missing from the whole citation data lifecycle, we are left with two solutions, which depend on the document and metadata quality.

The first case is a document without any citation metadata, and this data needs to be extracted from the body of the text. The second case is when the document's metadata contains a bibliography or reference list provided either as plain text or a set of divided citations – this data needs to be parsed and enriched.

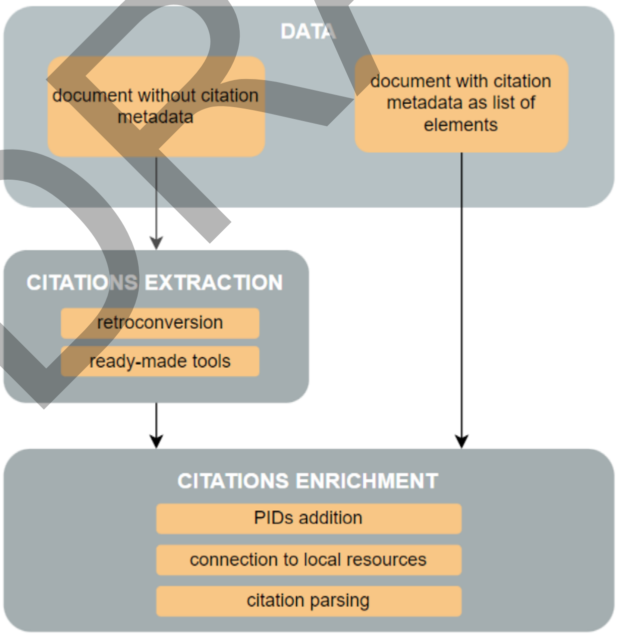

DATA
document without citation metadata
document with citation metadata as list of elements
CITATIONS EXTRACTION
retroconversion
ready-made tools
CITATIONS ENRICHMENT
PIDs addition
connection to local resources
citation parsing

Figure 6.2.2. Missing citations – possible solutions.

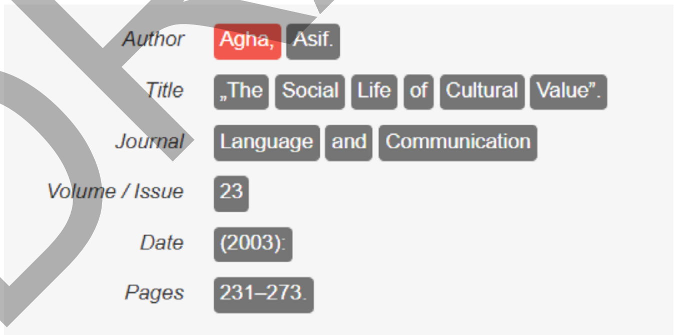
Citation extraction relies on software which uses OCR, OLR, and semantic segmentation to identify parts of the document as references and enhance metadata with this information.

[166] See: bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:bibliotekanauki.pl:1968869.

Citation enrichment is applied to a segmented set of extracted references. The further structuralisation of each reference may be provided by three strategies. The first one consists of adding persistent identifiers to bibliographic information, such as DOIs. For instance, this is done with Crossref plug-ins for the Open Journal System[167]. This solution is limited, as it depends on the scope of Crossref, Worldcat, and similar resources; but it remains especially useful for contemporary research output from journals, as DOI attribution is mostly focused on these.

The second solution requires connection to external resources such as national catalogues, bibliographies, or scientific repositories. The scope of these resources might be more suited for local needs – and they are not limited to records with full texts. These authority services could provide knowledge at the local level via URIs attached to the references. This type of linking may better shape the regional landscape of scientific research and prepare the information for international exchange. However, this solution would require effort from local communities to provide enrichment tools.

The third way to enrich citations requires automatic parsing and may be provided, for instance, with Open Journal Systems (OJS) plugins such as ParsCit Citation Extractor or ParaCite Citation Extractor. However, the most promising application is AnyStyle.io[168] – a free parser for references, which works regardless of citation style, and structures bibliographic data using machine learning heuristics based on conditional random fields.



Author	Agha, Asif.
Title	„The Social Life of Cultural Value”.
Journal	Language and Communication
Volume / Issue	23
Date	(2003):
Pages	231–273.

Figure 6.2.3. Example of automatic parsing using Annif software.

As can be observed, reference enrichment largely depends on the availability and quality of external resources. SSH publications often cite older resources which have not yet been digitised, and in this case a particular dataset would be needed to enrich such citations, for example, printed bibliographies which may not have been digitised. These resources need to be subjected to retrospective conversion to obtain properly formatted bibliographic information.

[167] See: www.crossref.org/documentation/content-registration/ojs-plugin/, docs.pkp.sfu.ca/crossref-ojs-manual/en/references
[168] arxiv.org/abs/2205.14677

6.3. GUIDELINES FOR CITATIONS AS HUMANITIES RESEARCH DATA IN THE CONTEXT OF TRIPLE – A SUMMARY

1. The humanities need to **take full advantage of existing infrastructure for open citations** – for instance, OpenCitations – which includes, for example, registering abstracts through Crossref services. Through this, **humanities data becomes more accessible and easier to link to different data spaces**.

2. In the humanities, **reliance on citation recognition and linking based on DOIs, yields limited results**. What is far more promising are approaches which allow for **citation recognition through more varied reference datasets** (such as national library databases or local repositories). Citations are relevant, even without global PIDs present in the bibliographic description.

3. In the humanities, citation extraction is more challenging, as bibliographic styles and reference formatting are more varied. **Dedicated approaches are needed to fully achieve a breakthrough in the humanities' citation extraction**. This, however, might be challenging, as many services (such as metadata aggregators or citation indexes) would still rely on DOI-based solutions for citation extraction. A solution tailored to the needs of the humanities[169] would, for example, call for the reuse of cultural heritage collections, and other local data collections, to ensure citation identification and linking. A novel approach is needed to tackle the systemic challenges in this regard. For example, most current solutions for citation extraction and identification rely on using resources like Crossref to recognize references and provide structured expressions for them. For the humanities, besides resources like Crossref, it makes sense to also use national library catalogues or national databases for research output.

[169] Colavizza, Peroni, Romanello, 'The case for the Humanities...'

7. CONCLUSIONS

GoTriple is an important dataset for all those who are interested in understanding and analysing European (and wider) societies and cultures, and/or who would like to enrich their services or datasets through normalised and carefully selected SSH data.

As a research data collection which is available via open protocols, it can be reused by a number of stakeholders, such as metadata aggregators (knowledge graphs, semantic search engines etc.) and indexes and information services (metrics providers, data analysis companies, and current research information systems). Its future development is aimed at making it more scalable and interoperable in order to facilitate this kind of interoperability. To leverage this potential, it is crucial to understand the key components of the GoTriple data model and the pathways leading to their improvement. This report has offered a discussion on the components and guidelines for shaping research data collections such as GoTriple.

The main finding and recommendation of this report is that in order to secure the organic development of a service like GoTriple, it is **necessary to work on intrinsic modifications** of the current data quality and data model; but it is **equally important to be involved in the larger processes and workflows** which pertain to key elements of the data model and its development (such as PIDs, keywords, persistent identifiers, and citations). This report identifies initiatives related to research data quality which are worth engaging in (i.e. Open Abstracts, OpenCitations, and Dublin Core-related working groups), and technologies which demand closer and continued inspection (NLP, ML). This also helps to contextualise a heavily discussed issue concerning the specificity of humanities data – **although we have identified numerous dimensions to this specificity, many crucial developments are discipline agnostic and can help the humanities community face their own challenges – if they are properly understood and applied**. One example is the implementation of linked open data and Semantic Web technologies in vocabulary creation (which helps in the semantic linking of dispersed and heterogeneous ontologies or thesauri). Another is NLP and ML, which could drastically impact the ability to produce multilingual content.

In other words, the key to the development of services like GoTriple – services which process humanities research data and aim to provide the output of this processing for further reuse, including scientific and analytical – **is the creation of environments where the humanities (SSH) domain experts**, computer scientists, and research data experts can work together in a sustained manner. This deliverable is an example of this kind of effort to bring together various actors in search of solutions relevant to the challenges which contemporary humanities are facing.

BIBLIOGRAPHY

Aggarwal, C. C., *Machine Learning for Text* (Springer, Cham, 2018).

Angelaki, G., K. Badzmierowska, D. Brown, V. Chiquet, J. Colla, J. Finlay-McAlester, K. Grabowska, et al., *How to Facilitate Cooperation between Humanities Researchers and Cultural Heritage Institutions. Guidelines*, (Warsaw, Poland: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 10 March 2019), doi.org/10.5281/zenodo.2587481

Balula, A., L. Caliman, S. Fiorini, S. Jarmelo, D. Leão, P. Mounier, J-F. Nomine, et al., *Innovative Models of Bibliodiversity in Scholarly Publications: OPERAS Special Interest Group Multilingualism* (White Paper, 8 November 2021), doi.org/10.5281/zenodo.5653084

Bauer, F., M. Kaltenböck, *Linked Open Data: The Essentials*, (Vienna: edition mono/monochrom, 2012), www.reeep.org/linked-open-data-essentials

Boudin, F. 'Unsupervised Keyphrase Extraction with Multipartite Graphs', *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Volume 2 (Short Papers), (New Orleans, 2018), arxiv.org/abs/1803.08721

Bougouin, A., F. Boudin, and B. Daille, 'TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction', *International Joint Conference on Natural Language Processing* (IJCNLP), (Nagoya, 2013), hal.science/hal-00917969/

Britt Holbrook, J., B. Penders, and S. de Rijcke, 'The Humanities do not Need a Replication Drive', *CWTS Blog* (archive), (21 January 2019),
www.cwts.nl/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive

Car, N., P. Golodoniuc, and J. Klump, 'The Challenge of Ensuring Persistency of Identifier Systems in the World of Ever-Changing Technology', *Data Science Journal* 16 (2017), doi.org/10.5334/dsj-2017-013

Chang, W-C., H-F. Yu, K. Zhong, Y. Yang, and I. D. Dhillon, 'Taming Pretrained Transformers for Extreme Multi-label Text Classification', (KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, 2020),
dl.acm.org/doi/abs/10.1145/3394486.3403368?casa_token=-E8vT-FHdnwAAAAA:PmQdjkC9gsvvg-sZok4T3MptmbP NOzh6DyGC1MkElwOwDnwaV1S0OF5lXKsRVVol7tqqBOjG3Xmc

Cioffi, A., Peroni, S. (2022). Structured References from PDF Articles: Assessing the Tools for Bibliographic Reference Extraction and Parsing, doi.org/10.48550/arXiv.2205.14677

Colavizza, G., S. Peroni, and M. Romanello, 'The Case for the Humanities Citation Index (HuCI), A Citation Index by the Humanities, for the Humanities', *International Journal on Digital Libraries*, (2022), doi.org/10.1007/s00799-022-00327-0

De Santis, L. *TRIPLE Deliverable: D6.6 API's Development -RP3* (Draft), (Zenodo, 2022),
doi.org/10.5281/zenodo.7371832

Devlin, J., M-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *NAACL-HLT* (1), (2019).

Edmond, J., N. Horsley, R. Kalnins, J. Lehman, M. Priddy, and T. Stodulka, *Big Data & Complex Knowledge. Observations and Recommendations for Research from the Knowledge Complexity Project*, 8–9, (K-PLEX. University College Dublin, 2018),
kplexproject.files.wordpress.com/2018/04/trinity-big-da ta-report-jklr_04.pdf

Edmond, J., and Erzsébet Tóth-Czifra, *Open Data for Humanists, a Pragmatic Guide*. (Zenodo, 2018), doi.org/10.5281/zenodo.2657248

El-Beltagy, S. R. and A. Rafea, ‘KP-Miner: Participation in SemEval-2’, *Proceedings of the 5th International Workshop on Semantic Evaluation*, (Uppsala 2010), aclanthology.org/S10-1041.pdf

European Commission, *PID Architecture for the European Open Science Cloud. Report from the EOSC Executive Board Working Group (WG) Architecture PID Task Force (TF)*, (2020), doi.org/10.2777/525581

European Commission, *Towards a Reform of the Research Assessment System. Scoping report*. (KI-09-21-484-EN-N), Publications Office, (2021), doi.org/10.2777/707440

Fellbaum, Ch. ed., WordNet – *An Electronic Lexical Database*, (Cambridge, MA: MIT Press, 1998).

Ferwerda, E., Frances Pinter, and Niels Stern, *A Landscape Study on Open Access and Monographs: Policies, Funding and Publishing in Eight European Countries*, (Zenodo, 2017),
doi.org/10.5281/ze nodo.815932

Florescu, C. and C. Caragea, ‘PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents’, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* Volume 1 (Long Papers), (Vancouver, 2017), aclanthology.org/P17-1102/

Frantzi, K., S. Ananiadou, and H. Mima, ‘Automatic Recognition of Multi-Word Terms: The Cvalue/NC-value Method’, *Int. Journal on Digital Libraries* (3) (2000).

Gambhir, M. and V. Gupta. ‘Recent Automatic Text Summarization Techniques: A Survey’, Artif Intell Rev 47, (2017).

Georgiadis H., M. Blaszczynska, and M. Maryl. TRIPLE Deliverable: D2.4 *Report on Identification and Creation of New Vocabularies* (Zenodo, 2023), doi.org/10.5281/zenodo.7539922

Golub, K., ‘Evaluating Automatic Subject Indexing: A Framework’, *keynote speech at the 7th ISKO Italy Meeting Bologna*, 20 April 2015, www.iskoi.org/doc/bologna15/golub.htm

Gould, M. (2022). People, places, and things: Persistent identifiers in the scholarly communication landscape. *College & Research Libraries News*, 83(9),
crln.acrl.org/index.php/crlnews/article/view/25638

Grootendorst, M., Keybert: *Minimal Keyword Extraction with Bert*, (2020),
doi.org/10.5281/zeno do.4461265

Gualandi, B., L. Pareschi, S. Peroni. 'What Do We Mean by "Data"? A Proposed Classification of Data Types in the Arts and Humanities', (arXiv, 15 July 2022), doi.org/10.48550/arXiv.2205.06764

Gupta, S. and S. K Gupta, 'Abstractive Summarization: An Overview of the State of the Art', *Expert Systems with Applications* 121 (2019).

Haak, L. L., 'Persistent Identifiers Can Improve Provenance and Attribution and Encourage Sharing of Research Results', (1 Jan. 2014), doi.org/10.3233/ISU-140736

Hakala, J., 'Persistent Identifiers: An Overview', *KIM Technology Watch Report* (2010), www.persid.org/downloads/PI-intro-2010-09-22.pdf

Harrower, N., M. Maryl, T. Biro, B. Immenhauser, (ALLEA Working Group E-Humanities), *Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities*, (Berlin: ALLEA - All European Academies, February 2020), Digital Repository of Ireland, repository.dri.ie/catalog/tq582c863

Hasan, T., A. Bhattacharjee, M. Saiful Islam, K. Mubasshir, Y-F. Li, Y-B. Kang, M. Sohel Rahman, and R. Shahriyar, 'XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages', *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, (2021). [Online. Association for Computational Linguistics.] aclanthology.org/2021.findings-acl.413/

Hutchins, B. I., K. L. Baker, M. T. Davis, M. A. Diwersy, E. Haque, R. M. Harriman, T. A. Hoppe, S. A. Leicht, P. Meyer, and G. M. Santangelo, 'The NIH Open Citation Collection, A Public Access, Broad Coverage Resource', *PLOS Biology* 17(10) (2019), doi.org/10.1371/journal.pbio.3000385

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov, 'Bag of Tricks for Efficient Text Classification', *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 2, Short Papers, (Association for Computational Linguistics, Valencia, Spain, 2017).

Khan, A. and N. Salim, 'A review on abstractive summarization methods', Journal of Theoretical and Applied Information Technology 59.1 (2014).

Klein, M. and L. Balakireva, 'On the Persistence of Persistent Identifiers of the Scholarly Web', (arXiv, 2020), arxiv.org/abs/2004.03011

Kotarski, R., et al., *Developing Identifiers for Heritage Collections*, (Zenodo, 2021), doi.org/10.5281/zenodo.5205757

Kryscinski, W., B. McCann, C. Xiong, and R. Socher, 'Evaluating the Factual Consistency of Abstractive Text Summarization', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), (2020).

Kunze, J., Calvert, S., DeBarry, J.D., Hanlon, M., Jenee, G. and Sweat, S. (2017). Persistence Statements: Describing Digital Stickiness. *Data Science Journal*, 16 1–11, DOI: doi.org/10.5334/dsj-2017-039

Leão, D., M.Angelaki, A. Bertino, S. Dumouchel, and F. Vidal, OPERAS *Multilingualism White Paper*, (Zenodo, 2018), doi.org/10.5281/zenodo.1324026

Lin, H., V. Ng, 'Abstractive Summarization: A Survey of the State of the Art', *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01), (2019).

Liu, Y., J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, 'Multilingual Denoising Pre-training for Neural Machine Translation', *Transactions of the Association for Computational Linguistics* 8, (2020).

Madden, Frances, 'Why Use Persistent Identifiers?', *The PID Forum* (2019),
pidforum.org/t/why-use -persistent-identifiers/714

Marciniak, M., A. Mykowiecka, and P.Rychlik, 'TermoPL – A Flexible Tool for Terminology Extraction'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, eds N.Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, (Portorož, Slovenia: LREC, 2016). [European Language Resources Association (ELRA)].

Martín-Martín, A., 'Coverage of open citation data approaches parity with Web of Science and Scopus', *OpenCitations* (blog), 27/10/2021, opencitations.hypotheses.org/1420

Martín-Martín, A., M. Thelwall, E. Orduna-Malea, and E.Delgado López-Cózar, 'Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A Multi- disciplinary Comparison of Coverage via Citations', *Scientometrics* 126(1), (2020), doi.org/10.1007/s11192-020-03690-4

Maryl, M., M. Błaszczyńska, B. Szleszyński, and T. Umerle, 'Dane badawcze w literaturoznawstwie',
Teksty Drugie. Teoria literatury, krytyka, interpretacja, 2 (1 March 2021),
journals.openedition.org/td/14190

Mihalcea, R. and P. Tarau, 'TextRank. Bringing Order into Texts', *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, 2004), aclanthology.org/W04-3252.pdf

Nan, F., R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. Xiang, 'Entity-level Factual Consistency of Abstractive Text Summarization', *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* Main Volume (2021). [Online. Association for Computational Linguistics.]

O'Sullivan, J., 'The Humanities have a "Reproducibility" Problem', *Talking Humanities*, (9 July 2019), talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/.

Okulska, I., 'Team Up! Cohesive Text Summarization Scoring Sentence Coalitions'. In *Artificial Intelligence and Soft Computing. ICAISC. Lecture Notes in Computer Science*, vol 12416, eds L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J. M. Zurada (Springer, Cham, 2020).

Overton, J.A. et al. (2020). String of PURLs – Frugal Migration and Maintenance of Persistent Identifiers. *Data Science*, 3(1), doi.org/10.3233/DS-190022

Papagiannopoulou E. and G. Tsoumakas, 'A Review of Keyphrase Extraction', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2), arxiv.org/pdf/1905.05044.pdf

Paskin, N., ‘Digital Object Identifier (DOI®) System’, *Encyclopedia of Library and Information Sciences* 3 (2010), 0-www.doi.org.oasis.unisa.ac.za/topics/020210_CSTI.pdf

Peels, R., ‘Replicability and Replication in the Humanities’, *Research Integrity and Peer Review* 4, 2 (2019), doi.org/10.1186/s41073-018-0060-4

Peels, R., L. Bouter, and R. van Woudenberg, ‘Do the Humanities Need a Replication Drive? A Debate Rages on’, *Retraction Watch*, (13 February 2019), retractionwatch.com/2019/02/13/do-the-humanities-need-a-replication-drive-a-debate-rages-on/

Plomp, E., ‘Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments’, *Data Science Journal* 19(1) (2020), doi.org/10.5334/dsj-2020-046

Pęzik, P,. A. Mikołajczyk, A. Wawrzyński, B. Nitoń, M. Ogrodniczuk, ‘Keyword Extraction from Short Texts with a Text-to-Text Transfer Transformer’. In ‘Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2022’, eds E. Szczerbicki, K. Wojtkiewicz, S.V. Nguyen, M. Pietranik, M. Krótkiewicz, *Communications in Computer and Information Science*, vol 1716. (Springer, Singapore, 2022), doi.org/10.1007/978-981-19-8234-7_41

Pęzik P., A.Mikołajczyk, A.Wawrzyński, B. Nitoń, and M. Ogrodniczuk, *Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer*, (arXiv, 2022), ACIIDS2022, arxiv.org/abs/2209.14008.

Sharma, G. and D. Sharma, ‘Automatic Text Summarization Methods’, *A Comprehensive Review. SN COMPUT. SCI* 4 (2022).

Shearer, K., L. Chan, I. Kuchma, and P. Mounier, *Fostering Bibliodiversity in Scholarly Communications: A Call for Action*, (Zenodo. 2020), doi.org/10.5281/zenodo.3752923

Shillum, Chris, Julie Anne Petro, Tom Demeranville, Ivo Wijnbergen, Sarah Hershberger, Will Simpson, From Vision to Value: ORCID's 2022–2025 Strategic Plan, (ORCID, 2021), Online resource, doi.org/10.23640/07243.16687207.v1

Shotton, D., ‘Open citations’, *Nature*, 502 (7471), (2013), dx.doi.org/10.1038/502295a

Singh, V. K., P. Singh, M. Karmakar, J. Leta, and P. Mayr, ‘The journal coverage of Web of Science, Scopus and Dimensions: A Comparative Analysis’, *Scientometrics* 126(6) (2021), doi.org/10.1007/s11192-021-03948-5

Spinaci, G., G. Colavizza, and S. Peroni, ‘A Map of Digital Humanities Research Across Bibliographic Data Sources’, *Digital Scholarship in the Humanities*, (2022), doi.org/10.1093/llc/fqac016

Strader, C. Rockelle, ‘Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations’, *Library Resources & Technical Services* 53, 4 (2009), doi.org/10.5860/lrts.53n4.243

Suominen, O., J. Inkinen, and M. Lehtinen. ‘Annif and Finto AI: Developing and Implementing Automated Subject Indexing’, *JLIS*.It 13, 1 (2022), www.jlis.it/index.php/jlis/article/view/437

Tay, A., B. Kramer, and L. Waltman, *Why Openly Available Abstracts are Important – Overview of the Current State of Affairs*, (Medium, 2020), medium.com/a-academic-librarians-thoughts-on-open-access/why-openly-available-abstracts-are--important-overview-of-the-current-state-of-affairs-bb7bde1ed751

Tóth-Czifra, E., and N. Truan, 'Creating and Analyzing Multilingual Parliamentary Corpora', *Research Data Management Workflows* Volume 1, (2021), halshs.archives-ouvertes.fr/halshs-03366486

Visser, M., N. J. van Eck, L. Waltman, 'Large-scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic', *Quantitative Science Studies* 2(1) (2021), doi.org/10.1162/qss_a_00112

Vogel, D., 'Qualified Dublin Core and the Scholarly Works Application Profile: A Practical Comparison', *Library Philosophy and Practice (e-journal)*, 1085 (2014), digitalcommons.unl.edu/libphilprac/1085

Walk, P., *PIDs in Dublin Core*, (Zenodo, January 28, 2019), doi.org/10.5281/ZENODO.2551181

Wittenburg, P., 'From Persistent Identifiers to Digital Objects to Make Data Science More Efficient', *Data Intelligence* 1 (2019), doi.org/10.1162/dint_a_00004

Widyassari, A. P., S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and I. M. De Rosal Setiadi, 'Review of Automatic Text Summarization Techniques & Methods', *Journal of King Saud University – Computer and Information Sciences* 34 (4) (2020).

Wydmuch, M., K. Jasinska, M. Kuznetsov, R. Busa-Fekete, and K. Dembczyński, 'A No-Regret Generalization of Hierarchical Softmax to Extreme Multi-label Classification', *Advances in Neural Information Processing Systems* 31, (2018).

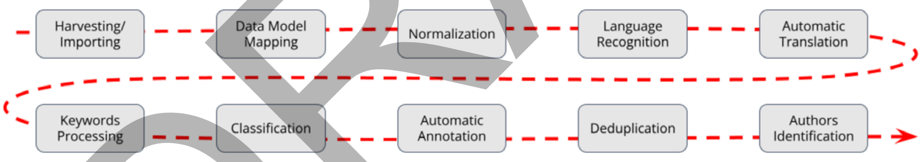
APPENDIX. DATA ENRICHMENT IN TRIPLE: THE CURRENT STATE OF WORK

Luca De Santis[170]

This section presents the data enrichment strategies which were devised in the TRIPLE project, focusing on the main topics of the booksprint – abstract and keyword management. Also, a small digression about permanent identifiers (PIDs) in GoTriple is presented at the end.

All the platform’s data are automatically harvested from various sources using GoTriple's metadata ingestion and curation service, named SCRE.

SCRE processes data using a pipeline approach. This, in fact, consists of several specialised components developed using Apache Camel[171] technology, each dedicated to implementing a particular feature of the ‘data flow’ – starting from the retrieval of a single piece of information (a publication, or also a project description), the curation and enrichment of its metadata, and finally the memorisation of the final result by the platform indexes, which is implemented through the Elasticsearch search engine[172].



Appendix Figure 1. Publication data flow

At present, GoTriple’s publication metadata are acquired in three possible ways: by harvesting OAI-PMH endpoints, OAI-PMH with data formatted either in Dublin Core or in the Europeana Data Model, and by importing data dumps (archives of metadata files in JSON or XML formats) from OpenAIRE and Isidore.

At the time of writing, more than 4.3 million publications and datasets from aggregators and providers have been integrated, including Isidore, OpenAIRE, EKT, DOAJ, Biblioteka Nauki, ZRC Sazu, Cessda, and Coimbra University. Soon, publication metadata from Open Edition, BASE, and Europeana will also be available.

[170] Net7, orcid.org/0000-0003-0527-840X
[171] camel.apache.org/
[172] www.elastic.co/elasticsearch/

Harvesting data from different sources proved to be a challenge for TRIPLE. Several issues were encountered, including

- difficulties in mapping metadata onto the TRIPLE data model, especially if they were in a format which had limited expressivity, such as Dublin Core;
- difficulties in trying to manage the 'specific extensions/interpretations' of the data providers' standards;
- difficulties in dealing with multilingualism for specific attributes (in particular, authors' names when they are mentioned in multiple languages with different alphabets);
- issues with data quality, including obvious mistakes (errors in dates or in language attribution);
- issues with the use of multiple codifications of dates or languages (e.g. 'en', 'en-us', 'eng', etc.);
- issues with the frequent use of textual strings instead of standard vocabularies for many attributes.

The general rules which have been followed in the TRIPLE project for data normalisation are

- removing duplicates when they appear;
- cleaning textual strings by trimming leading and trailing spaces and removing all the HTML codes;
- defining a controlled vocabulary for some elements;
- for normalised attributes, always maintaining the original metadata received, which are then stored in separate elements of the final GoTriple publications' index on Elasticsearch.

Processing **abstracts** is, in general, a straightforward task. Normally, abstracts are found in a dedicated element along with its language attribute. Some typical erroneous situations include

- sometimes, an abstract in its original form is split into multiple elements, possibly due to the data provider's incorrect interpretation of paragraphs;
- the language attributes might be missing or wrong;
- the text contains HTML elements.

In the first case, elements of the same language are taken and merged into a single abstract.

After doing tests on a sample from the first GoTriple dataset, the decision was made to always detect the language using the Apache Tika[173] library and store its code in ISO-639-1 format.

[173] tika.apache.org/

Text cleaning includes the removal of HTML and the leading or trailing blank characters.

If there is no English translation, the abstract, together with the title of the article, is translated using the eTranslation service[174]. In this way, it is guaranteed that GoTriple will practically always have an English version of textual descriptions.

Finally, titles and abstracts are used for automatic classification (TRIPLE/Moress categories) and annotation (TRIPLE vocabulary) enrichment.

As far as **keywords** are concerned, there are two types in GoTriple:

- free text keywords from the original sources,
- TRIPLE vocabulary entities, detected using the annotation service developed by the French company Foxcub.

This discussion only involves the first type, whose normalisation proved quite tricky and gave way to many discussions within the TRIPLE consortium. Basic curation is always applied to keywords, including the removal of duplicates and trimming the blank spaces before and after every string. A decision was also made to normalise the language attribute associated with the keywords, if present, (the 'lang' attribute) by using a controlled vocabulary (see TRIPLE deliverable D2.5[175]). At the same time, in contrast to what had been implemented for titles and abstracts, the keywords' language code is maintained, instead of assessing it with Apache Tika.

Another decision which was taken is to remove any keywords which could possibly refer to the codes or labels of taxonomies used by data providers. This is necessary to present data in a cleaner way and to improve search facet filtering in GoTriple. The rule for identifying whether a keyword is a taxonomic element or not has been adapted to the various data sources and presented in TRIPLE deliverable D2.5 Report on Data Enrichment. The keywords which are 'removed' are not lost but are stored in a special attribute (discarded_keywords) of the Elasticsearch index.

Finally, the decision was made to accept strings containing commas or points as single keywords – no assumption is therefore made as to whether a case like this refers to a single element or a set of keywords.

[174] ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

[175] L. De Santis. *TRIPLE Deliverable: D2.5 – Report on Data Enrichment* (Draft). (Zenodo, 2022), doi.org/10.5281/zenodo.7359654

Other curation and enrichment carried out in TRIPLE for publication data include the following.

- Publication date normalisation: we accept only dates after 1700 AD; all dates are formatted according to the ISO 8601 format (yyyy, yyyy-mm, yyyy-mm-dd).
- Controlled vocabularies have been introduced for
  - language codes, with 24 ISO-639-1 language codes, plus 'other' and 'undefined' elements;
  - document types, with 18 elements, mapped onto the corresponding COAR[176] resource types plus 'other' and 'undefined';
  - licences, with 11 licences, plus 'other' and 'undefined';
  - access rights, with 7 elements, plus 'other' and 'undefined'.
- Publication deduplication.
- Author disambiguation, that is, trying to match a person to author names which have been spelled in different ways (e.g. 'Suzanne Dumouchel', 'Dumouchel, Suzanne', 'Dumouchel, S.' etc.).

It is paramount for TRIPLE to guarantee the **wide reuse of its data**. At present, this is possible either by using the public search REST APIs or the OAI-PMH endpoint. It is also very desirable to embrace a linked open data approach in which GoTriple entities are returned in a semantic open data format (e.g. JSON-LD), as described by an official TRIPLE ontology which takes into account not only the TRIPLE data model but also the controlled vocabularies introduced in the normalisation phase cited above. Experiments in this area are ongoing and their results will be presented at the TRIPLE 2023 conference in Bonn.

As far as **persistent identifiers** are concerned, at present this concept isn't supported by GoTriple. There isn't even an internal TRIPLE ID, since documents are identified by the original data source's main identifier.

During an informal discussion on this topic two possibilities were considered:

- using the GoTriple URL/URI as a PID,
- integrating with some established service, in particular, handle.net.

The first option comes from the Semantic Web experience and is based on improving the current URL logic within GoTriple by considering three components:

- the access protocol and the gotriple.eu domain,
- the element's class in singular form (e.g. 'document', instead of 'documents' as it is now),
- a unique ID, generated using a solid random mechanism, which can guarantee that there is negligible risk of generating duplicates, for example, Nano ID[177].

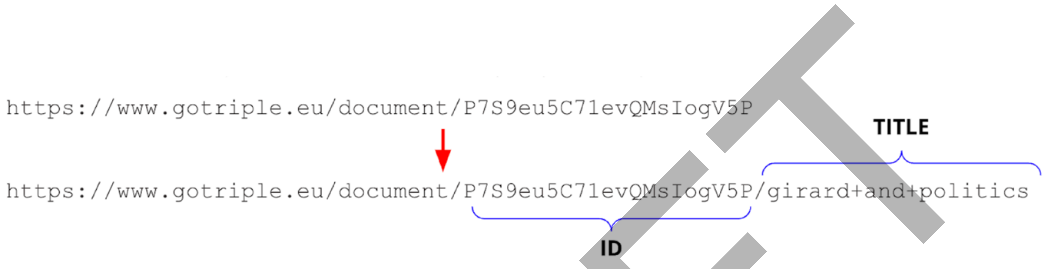
[176] vocabularies.coar-repositories.org/resource_types/

[177] zelark.github.io/nano-id-cc/

The result would be in the form of both an ID and an 'actionable' URL which can be accessed via the web. The content negotiation mechanism would also be able to distinguish

- requests for linked data, for example, generated by an application, which return all the document's information in JSON-LD, formatted accordingly to the (yet to be developed) TRIPLE ontology;
- whether access comes from a browser (or a search engine bot) in which case there should be a 302 HTTP redirection to a SEO-friendly URL, with the added title at the end.

This is shown in the image which follows.


https://www.gotriple.eu/document/P7S9eu5C71evQMsIogV5P
TITLE
https://www.gotriple.eu/document/P7S9eu5C71evQMsIogV5P/girard+and+politics
ID

Appendix Figure 2. PID URL redirection to a SEO and user friendly URL

On the other hand, the Handle.net integration might exploit a (yet to be established) OPERAS registry entry. Specific IDs, produced according to Handle.net formatting rules, would be created for every document and sent to the Handle.net registry by a new SRCEE (University of Zagreb University Computing Centre) service.

As we are approaching the final part of the project, **introducing changes in metadata management** can become quite complicated, especially regarding PIDs. Every structural change in metadata translates into all the existing data being reprocessed, which is very time consuming and can take several weeks to finish. The introduction of PIDs in particular might also have an impact on the data collected so far by the Recommender, which are based on GoTriple's current method of identifying documents.

At the same time it is important to keep a 'long-term' perspective on GoTriple as an OPERAS service, that is, as an infrastructure built to last. As a consequence, all possible improvements must be considered and evaluated, either for being implemented in these few remaining months of the TRIPLE project or in light of a future evolution of the platform.