Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 9.7 Design and Planning of Knowledge Graph in Electoral Studies

Dissemination Level	PU
Due Date of Deliverable	31/12/19 (M12)
Actual Submission Date	02/04/20
Work Package	WP9 - Data Communities
Task	T9.3 Data Community Project: Electoral Studies
Type	Report
Approval Status	Approved by EC - 03 November 2020
Version	V1.1
Number of Pages	p.1 – p.54

Abstract: This report discusses and defines the fundamental parameters within which a pilot Knowledge Graph in the domain of Electoral Studies will be developed, as well as a planning in terms of sub-tasks and timeline.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.00	02/03/2020	Structure of the document agreed	Cees van der Eijk Albin Ahmeti
0.01	05/03/2020	Added Knowledge Graph Governance and Knowledge Graph basic info	Martin Kaltenböck Albin Ahmeti Sotirios Karampatakis
0.02	09/03/2020	Added Ontology Development	Martin Kaltenböck Albin Ahmeti Sotirios Karampatakis
0.03	09/03/2020	Added sections 1, 2.1, 3	Cees van der Eijk Sylvia Kritzinger
0.04	09/03/2020	Added subsections 3.3, 3.4, 4.3 and 5.1	Julia Partheymüller
0.05	10/03/2020	Added Testing	Martin Kaltenböck Albin Ahmeti
0.06	10/03/2020	Added section 4	Cees van der Eijk Julia Partheymüller
0.07	08/03/2020	Added 5.2, 5.2.2 and 5.2.3	Martin Kaltenböck Albin Ahmeti
0.08	12/03/2020	Added Iterative development section	Martin Kaltenböck Albin Ahmeti
0.09	12/03/2020	Added Technical KG specification and development chapter	Martin Kaltenböck Albin Ahmeti
0.10	18/03/2020	Added final sections 8 and 9	Cees van der Eijk
...	20/03/2020	Peer Review	Tomasz Parkoła
1.0	25/03/2020	Revision of document to accommodate reviewer comments	Cees van der Eijk Martin Kaltenboeck Albin Ahmeti

Author List

Organisation	Name	Contact Information
UNOTT	Cees van der Eijk	cees.vandereijk@nottingham.ac.uk
University of Vienna	Sylvia Kritzinger	sylvia.kritzinger@univie.ac.at
University of Vienna	Julia Partheymüller	julia.partheymueller@univie.ac.at
SWC	Martin Kaltenböck	martin.kaltenboeck@semantic-web.com
SWC	Albin Ahmeti	albin.ahmeti@semantic-web.com
SWC	Sotirios Karampatakis	sotiris.karampatakis@semantic-web.com

Time Schedule before Delivery

Next Action	Deadline	Care of
sent for Peer Review	19/03/2020	Tomasz Parkola (SWC)
received peer review	20/03/2020	Cees van der Eijk (UNOTT)
revisions to produce final document	20/03/2020	(SWC and UNOTT) Cees van der Eijk / Martin Kaltenboeck/Albin Ahmeti
Submit final revised version	19/04/2020	Cees van der Eijk (UNOTT)

Executive Summary

This report constitutes SSHOC Deliverable 9.7 *Design and Planning of Knowledge Graph in Electoral Studies*. It describes the approach as well as the implementation plan for the SSHOC Pilot project in the field of electoral studies, which will constitute Deliverable 9.9.

After a brief introduction that sets out the purpose of this report, Section 2 elaborates the substantive domain to be covered by the Knowledge Graph (KG), and subsequently develops the concept of a KG and reasons for developing such infrastructural tools. The substantive domain is contained within the field of Electoral Studies, which was described in detail in Deliverable 9.6 *Demarcation Report of Electoral Studies User Community*. For reasons of practicality this domain has been further specified in several steps. First it was specified as pertaining to (sub)field of citizen/voter behaviour, and subsequently it was narrowed down further to the field of (studies of) electoral participation. These choices imply that the KG to be developed will not cover the entire field of electoral studies, but that it will be centrally located within the wider field and that its relevance for end-users is not restricted to those who are primarily interested in electoral participation.

The aspired functionality of the pilot-KG is discussed in Section 3. It defines as its intended audience of end-users in first instance scholars engaged in empirical research, but it expects in its later development stages to be also of relevance to journalists, think tanks, government agencies, corporations, political parties and politicians, and individual citizens. Functionalities to be sought relate to (a) focussed searching based on domain-specific criteria not available in other search tools; (b) teaching; and (c) research.

Section 3 also discusses the main datasets and kinds of publications to be covered by the pilot-KG. The overwhelming majority of these datasets are in the public domain, while a considerable part of relevant publications is not (at least not for the first years after publication). This provides challenges that will be addressed in the work program of Deliverable 9.9.

Section 4 discusses how the team developing the pilot-KG plan to involve the user community of electoral studies. Recruitment for such involvement is planned to be done via relevant scientific conferences (which are identified in Section 4) and the authorship of publications in relevant scientific journals; recruited volunteers will be mainly tasked with pre-structured coding and testing. In addition, smaller groups of community members will be personally invited based on their expertise and their willingness to commit somewhat more of their time to assist in the development of an ontology, of coding schemes, and of the design of testing phases.

Section 5 discusses the development of an ontology, which is necessary in view of the *de facto* absence of ontologies, classification schemes or controlled dictionaries that would be able to classify substantive content *within* the field of electoral studies. The section discusses approaches for ontology development and how these will be applied in the context of the development of the intended pilot-KG. In this context, this section also discusses the governance and management of the KG and its ontology.

Section 6 presents technical specifications for processes such as data ingestion, data cleaning, data authoring, data linking, data enrichment, data provisioning and data analysis. It also discusses the technical environment of semantic middleware to be used (for which *PoolParty* was chosen).

Section 7 discusses testing and user-community involvement in that process.

Section 8 discusses post-delivery development issues, and Section 9 presents a planning in terms of tasks and timelines.

Abbreviations and Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
ChEBI	Chemical Entities of Biological Interest
CQ	Competency Question
CQDKGC	Competency Question Driven Knowledge Graph Construction
CSES	Comparative Study of Electoral Studies
CSV	Comma Separated Values
CTCN	Climate Technology Centre and Network
DPU	Data Processing Unit
EES	European Election Study
EOSC	European Open Science Cloud
EPOP	Elections, Public Opinion and Parties
ESCO	European Skills, Competences, qualifications and Occupations ontology
EU	European Union
FIBO	Financial Industry Business Ontology
ISSP	International Social Survey Program
JEPOP	Journal of Election, Public Opinion and Parties
KG	Knowledge Graph
KPI	Key Performance Indicator
LD	Linked Data
LPG	Labelled Property Graph
MEDem	Monitoring Electoral Democracy Consortium
MEDW	Making Electoral Democracy Work
ML	Machine Learning
ORKG	Open Research Knowledge Graph
OWL	Web Ontology Language
PDF	Portable Document Format
RDF	Resource Description Framework
REST	Representational State Transfer
(S)FTP	(Secure) File Transfer Protocol

SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SSHOC	Social Sciences & Humanities Open Cloud
TDKGC	Test Driven Knowledge Graph Construction
TEV	True European Voter Consortium
W3C	World Wide Web Consortium
XLS	Microsoft Excel file

Table of Contents

List of Figures

List of Tables

1. Introduction and purpose of this deliverable

Work package 9 (WP9) of SSHOC focuses on various user communities, including some that are not represented by any of the partners of the Consortium. One of these is the user community working in the field of Electoral Studies, which is an important and highly developed field in Political Science and cognate disciplines. This user community has been defined and demarcated in the report of Deliverable D9.6 (Demarcation Report of Electoral Studies User Community).

The program of activities of WP9 includes piloting the development of infrastructural tools – known as KGs, briefly described later. These tools are expected to be exceptionally valuable in the development of open science; for the re-purposing and re-use of data, scientific publications, and analytical results; for promoting multidisciplinary collaboration; and to increase the potential for societal impact. In short, KGs are seen as one of the ways to realise SSHOC's aspiration of the "Development, realisation and maintenance of user-friendly tools & services, covering all aspects of the full research data cycle …. , taking into account human-centric approach and creating links between people, data, services and training." (Research and innovation action NUMBER – 823782 – SSHOC, page 3).

What a KG is can be defined in a variety of ways (cf. Ehrlinger and Wöß 2016). Fortunately, these definitions overlap considerably. This makes it possible to briefly describe, in first instance, a KG as *a formal and structured representation of a domain of knowledge that offers stakeholders an interface for navigating the accumulated knowledge in the domain, and for discovering relationships between information from otherwise separate and unconnected databases within or relating to that domain*. A further discussion of the nature of KGs is provided in Section 2.2.

Although it is widely expected that KGs will be invaluable as infrastructural tools, their development in the domain of the social sciences and humanities is still in its infancy. Hence the importance of piloting such development and to identify specific problems involved.

WP9 has been tasked to conduct such a pilot, that will substantively be focussed on the field of electoral studies. The experiences and resulting tools from this pilot will not only be of use for the electoral studies community, but also for other user communities (including those that are represented in the Consortium by its partners) that aim to develop their own KGs in due course.

The actual production and delivery of this KG in a user-validated form will constitute another deliverable (D9.9, titled *Delivery of user validated Knowledge Graph, and Election Studies Analytics Dashboard*). The current report defines the fundamental parameters within which this pilot KG will be constructed, as well as a planning in terms of sub-tasks, and timeline; as such it constitutes a deliverable in its own right (D9.7 – *Design and Planning of a Knowledge Graph in Electoral Studies*). This report does not aim to be fully exhaustive in all aspects of specification and planning. Some detailed aspects do not lend themselves well to be included in the narrative report that constitutes D9.7, while decisions on some other (often detailed) aspects of design are path-dependent on experiences and results of earlier stages of the development, and have therefore to be

considered at a later and more relevant moment than at the time of delivery of this report. Instead, this report aims to provide for a non-specialist audience a clear, but in some respects broad-brush perspective on considerations and choices that will guide the further development of the pilot KG.

This report starts, in section 2, with a demarcation of the substantive domain of the KG. Section 3 discusses its aspired functionality. Section 4 deals with user community involvement. Section 5 considers issues relating to ontology development, while Section 6 reviews issues concerning technical specifications. Sections 7 and 8 involve, respectively, testing, and post-delivery development. Section 9 presents planning in terms of tasks, timeline and responsibilities. In view of the character of this report, there is no need for a concluding section.

2. Demarcation of the domain of the pilot-KG

2.1 The substantive domain

The substantive domain of the pilot-KG is defined by electoral studies. However, as indicated in the report of D9.6 (*Demarcation Report of Electoral Studies User Community*), this domain is –in spite of its shared focus on elections– incredibly broad and variegated. Although most of the field is empirically oriented, it also contains important philosophical, normative, and 'positive' (in the sense of non-empirical axiomatic-deductive) work. The (dominant) empirical segment of work in electoral studies is heterogeneous in the specific phenomena under study, which can be the rules and regulations that govern elections, or the behaviour of central actors that are involved (parties and candidates; citizens; media; financiers, etc.), or the 'outcomes' of elections (which range from who gets elected to which office, to policy outcomes, to reactions in the form of trust, and so on). Much of electoral studies has a distinct disciplinary flavour in political science, but smaller segments of this domain are located in other disciplines such as sociology, psychology, law, economics, geography, communication, and so forth. All these, and multiple other, distinctions make the entire field of electoral studies too heterogeneous and too broad for the purpose of developing a pilot-KG within the given resource parameters.

While narrowing down the substantive field to be covered by the pilot-KG, the aim is nevertheless to remain of direct interest and relevance for the largest possible proportion of the overall electoral studies user community. This can be achieved by focussing on the largest subfield within electoral studies (as reflected in scholarly output), which consist of the study of citizen/voter behaviour in elections for public office in democratic societies.[1] Yet, even this restriction of the domain to be covered by the pilot-KG still leaves a very wide and diverse field of scholarly work, as citizen/voter behaviour encompasses such diverse phenomena as

[1] See the report of D9.6 (*Demarcation Report of Electoral Studies User Community*, particularly Section 2.1) for the rationale of the qualifiers 'public office' and 'democratic societies' in this narrowing down. The term *voters* is often used in two slightly different meanings: on the one hand all those who are entitled to participate in a specific election, and, on the other hand, those who actually make use of that right by casting their vote. This report uses the term *voters* in the first of these two meanings.

whether or not one makes use of one's right to vote; one's choice of candidate or party (or, in the case of a referendum: of one of the available options); one's activities in electoral processes as influencer, canvasser, volunteer, donor, etc.; one's openness to communications about an election; and even more. A further narrowing down of the substantive domain of the pilot-KG is therefore necessary, and this was found in a focus on *electoral participation*, which involves whether or not people who have the right to vote do make use of that right in a given election by casting a vote.

The substantive scope of the pilot-KG is thus defined by the phenomenon of electoral participation, which is obviously only a part of the entire field of electoral studies.[2] Yet, in spite of this narrowing down, the pilot-KG to be developed can be expected to remain of direct interest to very large segments of the electoral studies user community. As a phenomenon it defines one of the central questions in very many studies of citizen/voter behaviour; it links up directly to virtually all surveys of citizen/voter behaviour, irrespective of whether these are directed at national elections; and understanding differences in electoral participation between elections or between countries is one of the dominating questions in much comparative electoral research.[3]

2.2 What is a Knowledge Graph

For the past decade or so, KGs have been sneaking into daily life, be it through voice assistants (such as Alexa, Siri or Google Assistant), intuitive search results or even personalized shopping experiences through online store recommenders. Many people are constantly interacting with KGs on a daily basis. However, KGs and underlying graph databases are still a mystery to most and because of its seamless entrance into our lives, most are not even aware of how dependent they are on the technology – or worse, how they have come to expect a certain quality and standard that is now commonplace.

Many organizations are already using KG technology to help themselves stay ahead of the game. And KGs and graph databases have been in use for all types of industries, ranging from banking, the auto industry, oil and gas to pharmaceutical and health, retail, publishing, the media and more. Although these organisations use KGs for different use cases, the purpose is the same: taking large amounts of data from various data silos and adding value to it so that it can be used (and ultimately re-used) in a meaningful and more intelligent way.

The rising role of content and context for delivering insights with AI technologies, as well as recent KG offerings for AI applications have pulled KGs to the surface Gartner (2018): "Hype Cycle for Artificial Intelligence"[4]

[2] See also the report of Deliverable 9.6 (*Demarcation Report of Electoral Studies User Community)* for a further discussion of how the study of electoral participation fits in the wider field of electoral studies.

[3] It must be emphasised that, although closely linked, electoral participation is to be distinguished from its aggregate manifestation, which is the *rate* of electoral participation, also known as the *turnout* of an election. Turnout is an aggregate-level phenomenon, while electoral participation is individual-level. Although the two terms are often used somewhat interchangeably, they refer to different units of observation and analysis, and therefore relate also to different research questions, explanatory approaches and empirical insights.

[4] https://www.gartner.com/en/documents/3883863

A similar approach is taken by the ORKG[5], that is an open project driven by TIB Hannover, Germany[6], where a KG is being used to make scientific (library) content better accessible and explorable as well as comparable.

A KG is a model of a knowledge domain created by subject-matter experts with the help of intelligent ML algorithms. It provides a structure and common interface for all data and enables the creation of smart multilateral relations throughout databases. Structured as an additional virtual data layer, the KG lies on top of existing databases or data sets to link all data together at scale – irrespective of whether these data are structured or unstructured.

2.2.2 What is the difference between a Knowledge Graph and a Graph Database?

KGs are data. They have to be stored, managed, extended, quality-assured and can be queried. This requires databases and components on top, which are usually implemented in the Semantic Middleware Layer. This 'sits' on the database and at the same time offers service endpoints for integration with third-party systems.

Thus, graph databases form the foundation of every KG. Typically, these are technologies based either on the Resource Description Framework (RDF)[7], a W3C standard, or on LPG.

In order to roll out KGs in organisations and/or for a certain domain, however, more than a database is required. Only with the help of components such as taxonomy and ontology editors, entity extractors, graph mappers, validation, visualization and search tools, etc. can it be ensured that a KG can be sustainably developed and managed. While graph databases are typically maintained by highly qualified data engineers or Semantic Web experts, the interfaces of the Semantic Middleware also allow people to interact with the KG who can contribute less technical domain-specific knowledge to the graphs.

2.2.3 How is a Knowledge Graph different from Google's Knowledge Graph?

KGs are all around; Facebook, Microsoft, Google, all operate their own KGs as part of their infrastructure. Google introduced in May 2012 its own version and interpretation of a KG. Since then the notion of a 'Knowledge Graph' has become more and more popular. On the surface, the information from the Google KG is used to augment search results, as illustrated in Figure 1, which was originally used by Gligoric and Kaltenböck (2018, reproduced with permission).

[5] https://www.orkg.org/
[6] https://projects.tib.eu/orkg/
[7] https://www.w3.org/RDF/

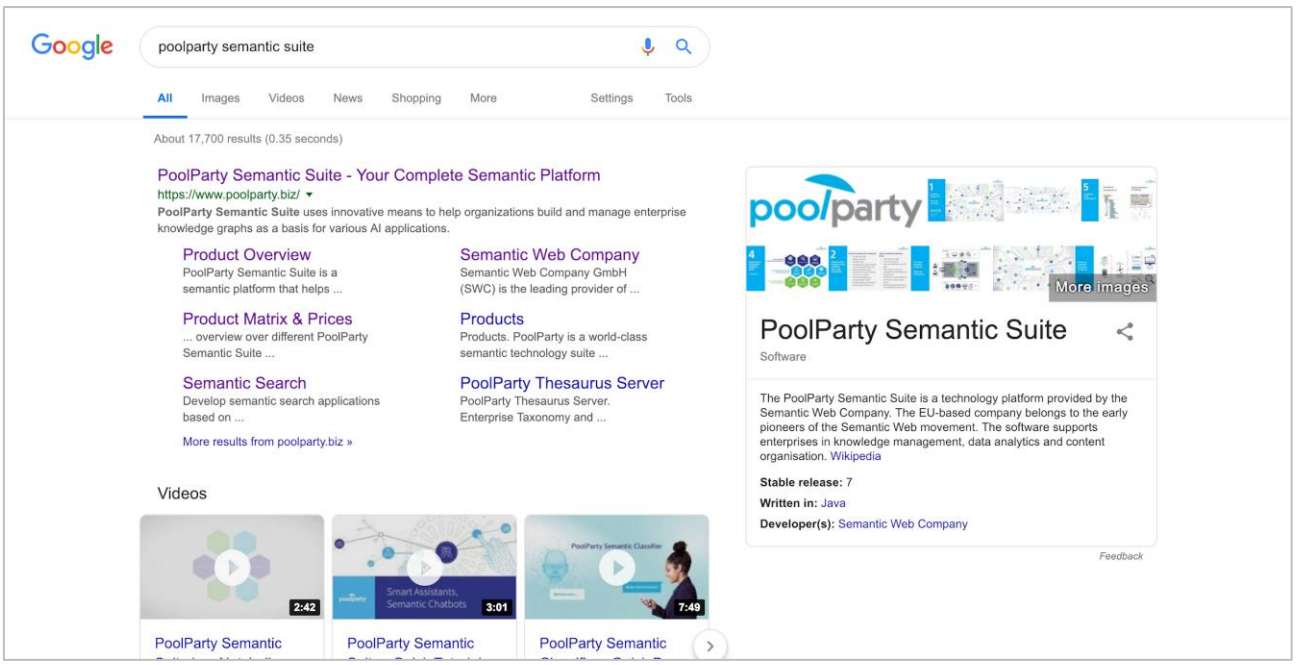

Figure1: KG data about PoolParty Semantic Suite displayed on Google Search (February 2019)

On top of that, the Google KG also enhances its AI when answering direct spoken questions in Google Assistant and Google Home voice queries. Behind the scenes and in return, Google uses its KG to improve its ML algorithms.

But Google's KG is quite limited in how users and software agents can interact with it. It only covers a few areas of industry-specific or domain-specific knowledge, and it doesn't cover the internal knowledge of organisations or specialised domains. Its API returns only individual matching entities, rather than graphs of interconnected objects.

This is where domain-specific KGs come into place. KGs help organisations or domain-specific communities (like the area of electoral studies) create their specific web of knowledge representing their very own domain. As a result, they can seamlessly break down data silos to use information assets in an agile way. Furthermore, it is a cost-efficient solution that does not replace but boosts existing IT systems. KGs fulfil today's requirements to process real-time sources of information and retrieve knowledge from data stored in disparate systems (see Figure 2 for an illustration in the domain of the Climate Technology Centre and Network -- CTCN). The content from CTCN website such as requests, response plan, technology dossiers, are tagged via the module of Climate Tagger[8] that is using Reegle taxonomy. The same taxonomy is used to drive the search, where one can search

[8]Climate Tagger website: https://www.drupal.org/project/climate_tagger; [20 March 2020]

- in a faceted search mode - through the different content, which is a-priori tagged with concepts from that taxonomy.

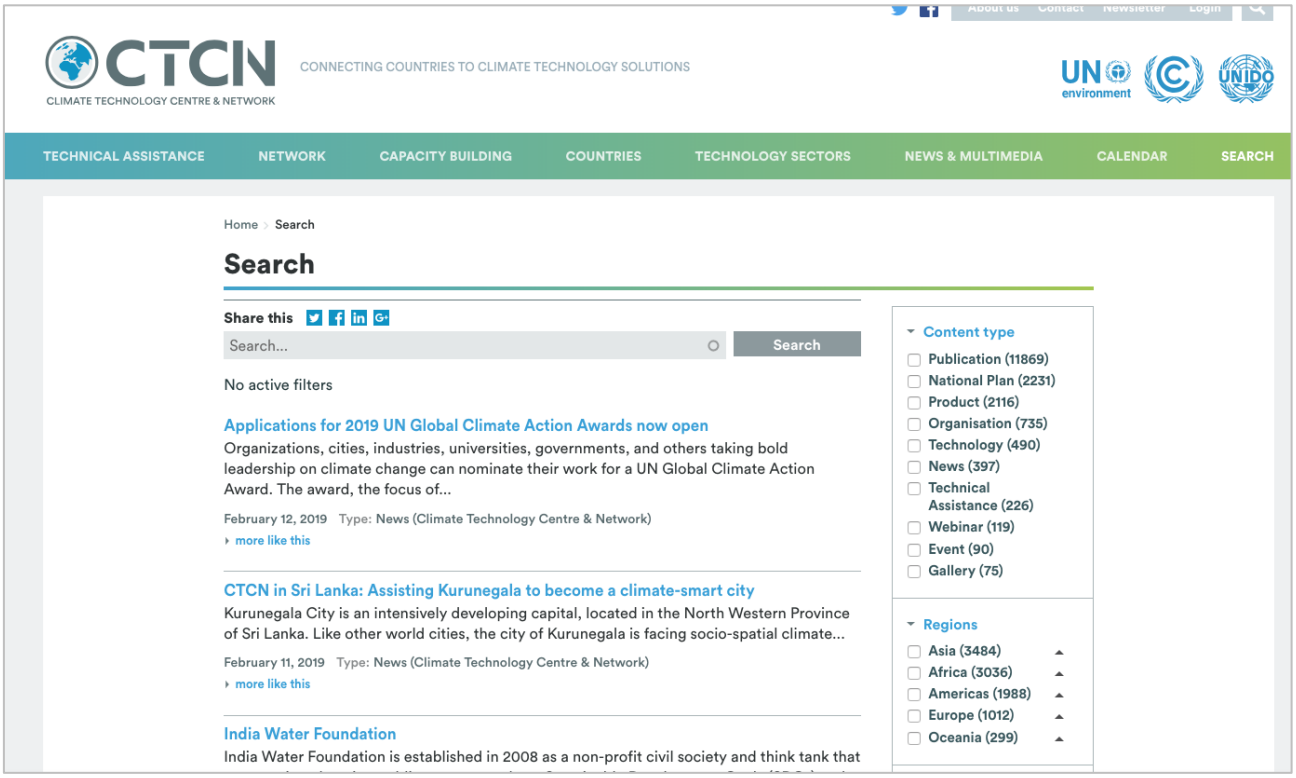

Figure 2: CTCN Search Portal enriched by an KG

2.2.4 Four reasons a Knowledge Graph can help

1. Combine Disparate Data Silos

Ever wonder how there could be a complete overlap of work from two separate departments or groups of people working in a specific domain, and neither one bothered to communicate with one another? This happens more often than not and requires organisations to reassess what they are spending their money and resources on: wasted effort on the knowledge they already have, or having employees re-learn things on a continual basis. KGs help to combine disparate silos of data, giving an overview of all of available knowledge – not only departmentally but also across departments and global organisations, as well across different groups of people working in the same area, for instance research groups in a specific domain like electoral studies.

2. Bring Together Structured and Unstructured Data

Apart from their possibility to combine information from different 'silos', KGs also allow the combination of information of different data formats. Accumulating data doesn't mean just assembling documents and excel sheets. KG technology means being able to connect different types of data in meaningful ways and supporting

richer data services than most knowledge management systems. Organisations will then use this technology to extract and discover deeper and more subtle patterns with the help of AI and ML technology. So can do groups of people working on the same topic, for instance a research domain.

3. Make Better Decisions by Finding Things Faster

Even prior to computers, looking for information meant digging through piles of documents to find a particular sentence, or number, etc. that is valuable to your train of thought. Using KG technology mitigates this by allowing more enriched and in-depth search results, helping to provide relevant facts and contextualized answers to specific questions, rather than a broad search result with many (ir)relevant documents and messages. KGs are able to do this because of their networks of 'things' and facts that belong to these 'things'. 'Things' could be any objects in a domain as well as their attributes or facets. Any graph can be linked to other graphs as well as to relational databases. On top of that, one can use KGs to derive new implicit information from the explicit one by leveraging reasoning capabilities that are possible in KG. With all of these linkages in place, a fully-fledged KG can provide domain-specific communities with a solid infrastructure and foundation for any smart application.

4. Future Proof Databases with Standards

Without quality data, it is impossible to get quality knowledge. With a KG in place, domain-specific communities benefit from higher reusability of their data because their KGs are compliant with W3C standards[9]. This triggers, not only internal network effects, but it also allows for the re-use of publicly available graphs, models and ontologies (e.g., FIBO[10], ChEBI[11], ESCO[12], etc.). This also allows that user communities are fully in control of their own KG.

3. Aspired functionality of the pilot-KG

The *domain-specific* information to be covered by the pilot-KG includes the following:

- professionally curated ***empirical data***. Given the specification of the domain (see Section 2.1) such data are overwhelmingly in the form of structured survey datasets, although they may, occasionally, also be in less structured form such as (transcripts of) unstructured interviews, other unstructured textual data, etc. Much of such data are relatively easily discoverable (see also Section 3.3).
- scholarly ***publications*** on electoral participation. These consist mainly of books and book chapters, published articles, and publications in the so-called 'grey' literature (conference papers, reports that

[9] W3C standards website: https://www.w3.org/standards/; [20 March 2020]
[10] EDMCouncil website: https://spec.edmcouncil.org/fibo/; [20 March 2020]
[11] ChEBI website: https://www.ebi.ac.uk/chebi/; [20 March 2020]
[12] European Skills, Competences, qualifications and Occupations ontology website: https://ec.europa.eu/esco/resources/data/static/model/html/model.xhtml; [20 March 2020]

are not published in book or article form). Here too, much of this information is relatively easily discoverable.

- ***scholars*** actively involved in the domain of the pilot-KG. When using authorship as a criterion for active involvement, relevant scholars are discoverable from scholarly publications.
- ***elections*** that are described or analysed by publications, or about which relevant data is available in empirical data. Most of these elections can be uniquely described by their spatio-temporal locators (which relate to the 'where' and 'when'), although occasionally additional attributes may be required to specify them, as in instances of different elections that are conducted concurrently in the same place.
- ***theories; concepts*** and their operationalisations in the form of variables; and analytical ***approaches*** and ***methods*** that are used in publications. These are plausibly among the most difficult kinds of information to be covered. They must be derived from publications, in first instance by expert-coding that can serve as a basis for ML. How successful classifications based on ML are will have to be established in practice. Moreover, the absence of generally accepted, widely used and uncontested descriptors in the field of electoral studies[13] may make these aspects of the pilot-KG potentially problematic.

In addition to the sorts of *domain-specific* information mentioned above, the pilot-KG should be able to link to other KGs that relate to these phenomena. One can think of other KGs that contain information about the spatial-temporal locations of elections (i.e., about countries, or about the weather on election day, etc.).

The remainder of this section discusses a set of issues that specify some of the parameters within which the production of the pilot-KG will take place. These include the specification of the kind of end-users for whom the KG will be produced, and how these end-users can be identified and approached for purposes of feedback, beta-testing, and, hopefully, contributing to the development of the KG in a variety of ways (section 3.1). Subsequently, section 3.2 specifies what the KG should allow end-users to do. Section 3.3 then specifies the main resources to be covered by the pilot-KG, such as empirical data and scholarly publications. Finally, section 3.4 discusses some of the attributes in terms of which these resources are to be described.

It must be emphasised that all the aspects discussed here can, and probably will be further elaborated and refined over the course of the actual development of the pilot-KG. Thus, this section (including its subsections) should be interpreted as an opening gambit rather than as an unalterable definition of the pilot-KG at the time of its delivery.

[13] See also the discussion about this in the following report: Van der Eijk, Cees. (2020). SSHOC D9.6 Demarcation Report of Electoral Studies User Community. DOI: 10.5281/zenodo.3725823

3.1 Intended audience

The pilot-KG is directed towards different scholars in the field of electoral studies. It is envisaged that the KG will, in due course, also be of relevance to other end-users, such as journalists, think tanks, government agencies (including electoral commissions and election regulators), corporations, political parties and politicians, and individual citizens. Yet, during the development phase and the evaluation of the resulting KG the focus will be primarily on academic scholars. This audience of academic scholars can be identified and approached in a variety of ways, as follows:

a) Research programs

There are large-scale research programs that focus in their data collection efforts and methodological developments on electoral studies in particular (including electoral participation). These are, *inter alia*, the (1) CSES that surveys (in collaboration with national election studies) democratic elections in more than 30 countries, using an identical questionnaire across all countries and elections covered; (2) the EES that surveys voter behaviour (including electoral participation) at European Parliament elections since 1989 in all EU member states; (3) various *national election studies* (e.g., ANES, AUTNES, BES, GLES, NES, etc.) which survey in many democratic countries voter behaviour at the national level over a time span up to 60 years; (4) the TEV Consortium that builds upon European national election study data and integrates these into a post-harmonized database; and (5) the Consortium MEDem that aims to link data on electoral participation to contextual data. The pilot-KG will provide important information for the future development of all these research programs.

b) Conferences and Scholarly organizations

Conferences focusing in particular on electoral studies, and thus being of particular relevance for this pilot-KG include (a) *EPOP*, the organized section of the Political Studies Association of the UK, defined by its focus on Elections, Public Opinion and Parties; it has annual conferences at end of Summer/early Fall; (2) *DACH*, the Austrian-German-Swiss-electoral workshop, that convenes regularly; and (3) the *AK Wahlen*-the organized section of the German Political Science Association with a particular focus on electoral research; it conducts annual conferences the May/June.

c) Individual Researchers / Scholars as identified via scholarly journals

The pilot-KG will target all researchers in the area of electoral studies. In order to reach this important group of scholars a survey will be conducted in which the benefits of such a KG will be conveyed (for the various benefits see below) and that simultaneously aims to recruit individual scholars as suppliers of feedback and beta-testers. An important way to identify such individual scholars is from their role as editors, editorial board members or authors of specialized academic journals of electoral research. The following journals fulfil this profile: *Electoral Studies* and *JEPOP*.

d) Students

Students who specialize in electoral studies are an important audience of the pilot-KG. Graduate and under-graduate students can be identified via universities that offer opportunities for specialization in this field, and via Summer Schools (or their equivalent in other seasons) that focus in particular on electoral studies.

As already indicated, the intended KG will, in a longer time perspective (i.e., beyond the time span planned for the production of this pilot KG) also be relevant for and would benefit from feedback and input from others than academic scholars.

A major challenge will be to obtain long-term involvement and commitment of a sufficiently large and sufficiently diverse group of scholars to develop the KG and subsequently to keep it up to date and to develop it further, both in terms of refinement and by extending its scope.

3.2 Intended end-user functionality

For the "success" of a KG, end-users need to acknowledge its usability and its utility, which thus have to be specified during its development process. The pilot-KG in electoral studies (which focuses in the first instance on electoral participation – see also Section 2.1) has to provide important end-user functionalities to be accepted and used by the research community. Three main areas of end-user functionality can be identified: first, the pilot-KG should serve as a general search tool; second, it should be an important teaching tool; third, it should provide relevant information for future research on electoral participation.

a) General searching tool

Researchers, pundits, students and the interested society at large can browse the pilot-KG in highly focused as well as undirected ways to obtain information on research about electoral participation in an easy and quick way. Although the pilot will in first instance be English language based, these queries should, in due course, be possible in various languages. This is an important facet of the pilot-KG as the hegemony of the English language in research at large can be interrupted and additional, less-known research outputs can be taken into account.

Furthermore, research is a dynamic field with new research outputs published on a regular basis. Keeping track of new research outputs is a challenging endeavour consuming many resources. The pilot-KG can reduce this workload substantially by providing alerts for new published research papers (including conference papers).

b) Teaching tool

The pilot-KG will serve as a teaching tool in two ways.

First, it will facilitate the teaching preparations of lectures on electoral research in general and on electoral participation in particular. Lecturers can navigate the pilot-KG to create their syllabi for lectures, seminars and exercises. It will not only allow them to prepare for course work in a fast way, but most importantly, it will

enable them to prepare and update their syllabi with the latest research papers in the area. Teaching will thus be able to include the newest research in an easier fashion than is currently the case.

Second, the pilot-KG will also facilitate the work of students of social sciences in general, and students of electoral studies in particular. Reading materials for seminars and lectures can be found in an easier way, and in particular, it will facilitate the literature search for undergraduate and postgraduate theses allowing students to move forward in a swifter way in their research endeavours.

c) Research tool

The inherent goal of any research is to move the field forward by developing new theoretical and empirical knowledge. The pilot-KG will facilitate identifying theoretical and empirical lacunas in existing research in an easy manner and thus will substantially contribute to new research areas and outputs.

Furthermore, the pilot-KG will provide the necessary information to conduct comprehensive meta-analyses of research in its substantive domain. The goal of a meta-analysis is to leverage research findings from multiple research outputs. However, it is increasingly difficult to keep track of the various research output in a certain research field. The pilot-KG can provide this necessary information by capturing research findings in the daily growing research field of electoral participation.

3.3 Resources covered: Datasets and publications

The raw material that the KG will be generated from are datasets and publications. In the field of electoral studies, the core of relevant datasets are the voter surveys produced by national election studies. There is a lot of variation across countries and over time, though, in how and how consistently such surveys have been carried out. While, in some countries, elections have been consistently monitored over many decades, in other countries voter data is rather scarce or non-existent. At the same time, the design of voter surveys can vary within and across countries. In general, the datasets are often archived by national social science data archives, but modes of data access can be quite heterogeneous. Overall, the heterogeneity of existing studies and data access can make it difficult for researchers to gain an overview of the available resources.

Table 1 gives an overview over the national election study programs that will serve as one of the major resources covered by the KG. Each election study program has produced multiple datasets that will need to be indexed, tagged with relevant meta-data, and linked to other relevant elements in the KG. Based on that, the KG should facilitate gaining an overview over the existing data resources within and across time and political systems.

Table 1: National election study programs

National election study program	Since	Data provider
Australian Election Study (AES)	1987	Australian Data Archive (ADA)
Austrian National Election Study (AUTNES)	2008	Austrian Social Science Data Archive (AUSSDA)
Belgian National Election Study (BNES)	1991	Data Archiving and Networked Services (DANS)
Canadian Election Study (CES)	1965	Canadian Opinion Research Archive (CORA)
Danish National Election Study (DNES)	1971	Centre for survey and Survey/Register data (CSSR)
Estonian National Election Study (ENES)	2003	Estonian National Election Study (ENES)
Finnish National Election Study (FNES)	2003	Finish Social Science Data Archive (FSD)
French Election Study (FES)	1958	French Data Archive For Social Sciences (CDSP)
German Longitudinal Election Study (GLES)	1949	GESIS - Leibniz-Institut für Sozialwissenschaften
British Election Study (BES)	1963	UK Data Service
Hellenic National Election Studies (ELNES)	2009	Inter-university Consortium for Political and Social Research (ICPSR)
Hungarian Election Study	1996	TÁRKI Social Research Institute (TARKI)
Icelandic National Election Study (ICENES)	1983	Social Science Research Institute (SSRI)
Irish National Election Study (INES)	2002	Irish Social Science Data Archive (ISSDA)
Israel National Election Studies (INES)	1969	Israel National Election Studies (INES)
Italian National Election Study (ITANES)	1968	Italian National Election Study (ITANES)
Japanese Election Study (JES)	1983	Social Science Japan Data Archive (SSJDA)
Lithuanian National Election Study (LNES)	2012	Lietuvos HSM duomenų archyvas (LiDA)
Dutch Parliamentary Election Studies (DPES)	1971	Data Archiving and Networked Services (DANS)
New Zealand Election Study (NZES)	1990	New Zealand Election Study (NZES)
Norwegian National Election Studies (NNES)	1957	Norwegian Centre for Research Data (NSD)
Polish National Election Study (PGSW)	1997	GESIS - Leibniz-Institut für Sozialwissenschaften
Portuguese Election Study (CEP)	2002	Production and Archive of Social Science Data (PASSDA)
Spanish Election Studies (CIS)	1977	Centro de Investigacione Sociológicas (CIS)
Swedish National Election Studies (SNES)	1960	Swedish National Data Archive (SND)
Swiss Election Study (Selects)	1971	FORS
American National Election Study (ANES)	1948	American National Election Study (ANES)

Besides national election studies, there are further relevant datasets that serve as an important resource for electoral researchers (see Table 2). These include most notably cross-national study programs. Some cross-national datasets such as the CSES and TEV integrate data produced by various national election study programs whereas other programs such as the EES and the MEDW cover voting in other arenas such as at the European and subnational level. Finally, surveys of the general resident population sometimes do include questions on voting behaviour. This, most notably, includes most notably the ESS and the ISSP. Again, these survey programs include multiple independent surveys that will need to be mapped by the KG to enable researchers to gain an overview of the additional resources that are available beyond national election studies.

Table 2: Other relevant survey programs

Study Program	Since	Data Provider
Comparative Study of Electoral Systems (CSES)	1996	GESIS - Leibniz-Institut für Sozialwissenschaften
The European Voter (TEV)	1956-1998	GESIS - Leibniz-Institut für Sozialwissenschaften
European Election Study (EES)	1979	GESIS - Leibniz-Institut für Sozialwissenschaften
Making Electoral Democracy Work (MEDW)	2010-2015	Harvard Dataverse
European Social Survey (ESS)	2002	European Social Survey (ESS)
International Social Survey Program (ISSP)	1985	GESIS - Leibniz-Institut für Sozialwissenschaften

Relevant types of publications can include books, book chapters, journal articles, as well as conference paper and pre-prints. To identify relevant publications, relevant databases and catalogues can be searched using a set of relevant keywords and synonyms that will need to be identified beforehand (e.g. voter turnout, electoral participation). In particular, the main journals in the field of electoral research (such as Electoral Studies, the JPOP, and Political Behaviour) deserve a close examination as they are likely to include relevant material with a very high propensity. Table 3 gives an overview of different types of sources that can be used to identify relevant publication resources. The KG will organize and classify the content of those publications and link them to relevant datasets, theories, concepts, research methods, and scholars, so that researchers will be able to gain an overview of the current state of the art very rapidly.

Table 3: Sources of relevant publications

Source	Type of source	Type of publications to be retrieved
Specialist Journals: - Electoral Studies - Journal of Elections, Public Opinion and Parties - Political Behaviour	Specialized journals focusing on electoral behaviour	Journal articles
Social Sciences Citation Index (SSCI)	Commercial citation index that covers 3000 of the world's leading academic journals in the social sciences	Journal articles
Google Scholar	Web search engine that indexes a wide range of different types of publications	Books, book chapters, journal articles, conference papers, pre-prints
WorldCat.org	Meta-library catalogue that covers the collections of 17900 libraries in 123 countries	Books
Social Science Research Network (SSRN)	Biggest open access repository for the rapid dissemination of scholarly research	Conference papers, pre-prints

3.4 Elements of the KG: Theories, concepts, research methods and scholars

From the datasets and publications, further elements of the KG can be derived (see also Section 3). These will include

- theories,
- concepts,
- research methods,
- and relevant scholars in the field.

In the field of electoral studies, there are currently no authoritative indexes of the relevant theories, concepts, and research methods. To generate a widely accepted ontology, the community of electoral researchers will be involved (see section 4.) and existing taxonomies and proto-ontologies will need to be reviewed (see section 5.1). By collaboratively generating a common ontology and classifying the underlying material based on it, researchers will be able to answer very specific questions based on the state of the art and identify potential research gaps very quickly.

There is currently also no database that can help to identify relevant experts on specific subtopics within the field. The names of relevant scholars can be indexed based on the publications and datasets. They can be linked to theories, concepts, and research methods, which will allow for the identification of experts on very specific questions. This database of relevant scholars will help to identify relevant colleagues, reviewers, project partners, speakers, lecturers and (co-)supervisors as well as experts for external media queries.

4. User community involvement

4.1 General perspectives on user community involvement

In accordance with the overall aim of SSHOC to maximise the involvement of relevant user communities, the development of the KG in Electoral Studies will include as much as possible such involvement. This is expected to be relevant in a number of ways, including

- Avoiding the development of a KG the functionality of which does not match needs and expectations of end-users, which would then lead to suboptimal usage in and by the user community and to obstruct further development of the KG beyond the timeline of the deliverables involved;
- Providing feedback at various stages of development
- Contributing to ontology development
- Crowd-contributing to classification of materials as a basis for training of algorithms
- Beta testing

Additionally, all forms of involvement of relevant user communities will help promote awareness of and familiarity with EOSC/SSHOC and its products, tools and services.

Forms of planned Involvement of user communities will be discussed in subsections 4.2 and 4.3. In case additional opportunities for user community involvement will present themselves during the development of the KG, efforts will be made to incorporate them into this further development.

4.2 Expert encounters

Particularly in the development of the necessary ontology –see also Section 5– involvement of leading and expert members of the user community of Electoral Studies is expected to be of crucial importance. In some instances, such experts can be involved in a decentralised and on-line fashion, while in other instances face-to-face consultations with several of such experts are expected to be of particular value. Such face-to-face encounters will be organised in the context of relevant meetings of other organisations in the field of electoral studies. Three of these have been identified as being of unique value in this respect, and contacts with their organisers have already resulted in their willingness to facilitate and support the work of WP9.9 (*Delivery of user validated KG, and Election Studies Analytics Dashboard*). If any similarly relevant expert meetings were to occur during the development of D9.9, efforts will be made to include those also in the upcoming work. The three meetings identified so far, which agreed to accommodating face-to-face sessions with expert members of the user community are:

- **DACH**: meeting of German (D), Austrian (A) and Swiss (CH) national election study investigator teams. This meeting was originally planned for the end of March 2020 in Frankfurt. Owing to the Covid19 pandemic, it has been rescheduled for October 2020.
- **AK Wahlen**: the annual meeting of the organised section on Elections and Political Attitudes of the German Political Science Association (*Arbeitskreis Wahlen und Politische Einstellungen*). This meeting was originally planned for May 2020, and has also been rescheduled because of the Covid19 pandemic, to October 2020.
- **EPOP**: the annual meeting of the specialised section of the British Political Studies Association on *Elections, Public Opinion and Parties*. Its annual conference is scheduled to be held in London from 4-5 September 2020.

In each of these meetings, volunteers will be recruited in advance for dedicated sessions contributing to evaluation and further development of the then current state of the ontology. At the same time, additional events at these meetings will disseminate information about EOSC/SSHOC, and recruit interested contributors for crowd-based coding and classification tasks.

4.3 Recruitment for and tasks for crowd-involvement

Organizing the user community requires the recruitment of relevant scholars and the delegation of meaningful tasks to them. As the KG identifies relevant scholars through their publications (see section 3.4), the KG itself

can be used as a tool for user recruitment by inviting those who have published on the topic to sign up as a member of the community. This mechanism of recruitment ensures that members of the community have relevant expertise on the subject and the mechanism itself can easily be automated.

The user community can then get involved at different stages of the development of the KG. In a first step, the user community can be surveyed to identify the relevant demands, questions and challenges that researchers face when trying to access information in the field. An initial online survey can be launched to measure the frequency and relevance of potential questions that the KG can answer. This will guarantee that the developed product will be effectively addressing researchers' genuine demands.

Apart from that, users can get involved in ontology development (see Section 5.). This may take several forms. Again, through online surveys or in-person encounters available proto-ontologies can be crowd-tested and consolidated. Crowd-involvement in that context can, for instance, also be used to generate synonyms, to transform keywords into taxonomies, and create translations of the ontology into multiple languages. This will ensure that the ontology is familiar to users and reflects their demands and use of language.

Once user demands have been identified and the ontology has been developed, users assist in classifying publications and datasets. Most notably, the content of publications will need to be coded with regard to theories, concepts, and research methods. The manually coded documents can then be used as training material to automatically classify new publications along those dimensions. Here, the expertise of the members could, for example, be leveraged by assigning documents to be classified to those who have cited these works in their own publications. Likewise, the national teams of election studies can be asked to classify their datasets to the extent that current meta-data (e.g. keywords) does not yet provide sufficient information on the concepts measured by the surveys.

Finally, user feedback will be enormously valuable when testing the KG. Testing may involve assessing the appropriateness of the ontology, the precision of classifications, as well as the usefulness of the user interface. Ideally, after initial development users should become continuously involved in curating and updating the content of the KG. This will make sure that the KG can serve as a focal point for knowledge exchange in the long run.

5. Ontology Development

5.1 Review of existing taxonomies and proto-ontologies

The first step in the process of ontology development is the review of existing taxonomies and proto-ontologies. Data providers and libraries have developed classification schemes that have been applied to classify diverse materials in the social sciences. Table 5.1-1 gives an overview of some of these classifications.

Table 4: Existing classification schemes in the social sciences

Classification scheme	Description
CESSDA Controlled Vocabulary for CESSDA Topic Classification	Typology of main themes of datasets
Data Documentation Initiative (DDI)	Schema to describe and document datasets
European Language Social Science Thesaurus	Multilingual thesaurus for the social sciences
Thesaurus Sozialwissenschaften (TheSoz)	German language thesaurus for the social sciences
EuroVoC	Multilingual, multidisciplinary thesaurus covering the activities of the EU, the European Parliament in particular

These classification schemes can mainly distinguish between electoral studies and other types of studies. In that sense, they can be useful to select relevant studies and describe datasets at a technical level. They are in general, though, too broad and unspecific to classify substantive content *within* the field of electoral studies. Given that, one of the largest challenges will be to develop novel substantive classification schemes for the theories, concepts, and research methods within the field of electoral behaviour.

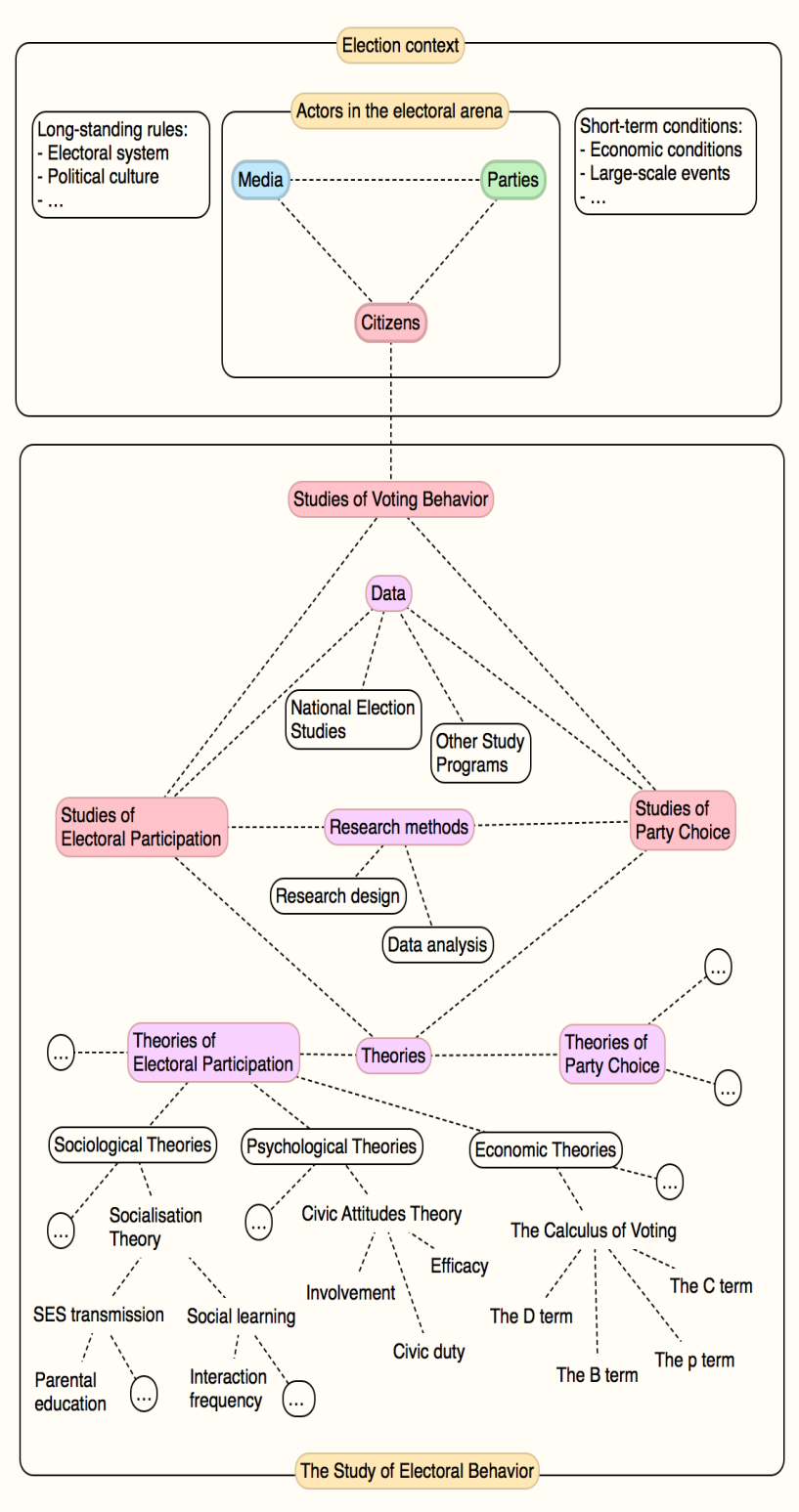

Figure 3: The study of elections and electoral behaviour

How the ontology for a KG on electoral participation might fit in a larger field electoral studies is sketched in Figure 3 that situates the study of electoral behaviour in the broader field of the study of actors in the electoral arena.[14] The figure locates citizens as an actor in the electoral arena together with the media and the political parties. All actors are situated within the electoral context that is shaped by long-standing institutional or quasi-institutional rules as well as by short-term conditions. The study of the behaviour of citizens as electoral actors itself branches out into studies of electoral participation (which is the domain of the pilot-KG to be developed) and studies of their party choice. Both types of studies often rely on the same data sources and share common research methods. The fundamental theoretical approaches are connected but differ with regard to their main outcome variable – turnout and party choice. Although this is a matter of discussion, theories can broadly be grouped into sociological, psychological, and economic theories. Each kind of theory introduces several concepts and sub concepts that translate into specific measurements in voter surveys. Theories are typically not mutually exclusive but rather complementary. They can share concepts and measures. The structure of these lower branches, though, is required to be developed in collaboration with the experts and broader user community.

5.2 Approaches to ontology development

There are a number of techniques that can be applied when collaborating with domain experts to create taxonomies. The best tool to use will depend on the circumstances. Typically, one can start with building a skeleton, i.e., main entities of the domain which is so-called the "top-bottom" approach (cf. card sorting, see section 5.2.1), or alternatively, one can start with the corpora of the domain in order to deduce concepts in a "bottom-top" approach (cf. corpus analysis in section 5.2.1, manual tagging in section 5.2.3). In an ideal world, one can use a mix of these two approaches in a "hybrid" mode.

The following subsections describe various approaches to ontology development that will be used in the development of the pilot-KG for Electoral Studies.

5.2.1 Expert-based first specification

To make optimal use of the planned expert encounters (see Section 4.2) a proto-ontology will be developed by a relatively small number of people consisting of scholars involved in SSHOC-WP9 (mainly from the University of Vienna, AUSSDA and the University of Nottingham) with possibly a handful of others based on invitation. This will be done iteratively in a series of video-conference meetings, alternating with individual contributions. It is envisaged that three or four of such iterations will suffice to arrive at a 'square 1 ontology' that is sufficiently indicative of its eventual aspirations to generate creative contributions - either on an individual basis, or in group sessions - from experts in the user community.

[14] Figure 3 does not sketch the entire field of electoral studies, but only of studies of the central actors in the electoral arena. This, in turn, is a subfield of the wider field of electoral studies, as indicated and discussed in Deliverable D9.6, *Demarcation Report of Electoral Studies User Community*.

5.2.2 User-involvement via sorting tasks

Card sorting is a technique popular in the discipline of usability. There are many existing resources online, for instance the description in Usability.gov[15] or the examples described in the Encyclopedia of Human-Computer Interaction[16]. A typical use of card sorting is to get stakeholders to group functionalities in menus or content in websites. The application described in the Encyclopedia of Human-Computer Interaction is on how to display fruit and vegetables in supermarket scales, so that patrons can easily find the product that they wish to weigh and print the cost. This is a clear use of the technique, but it is quite obvious that one could also apply card sorting to help stakeholders to come up with a taxonomy structure. There are variations, but the technique is applied as a workshop exercise where, ideally, a group of domain experts are involved that share a deep understanding of the domain.

PREPARATION PHASE

Although the card sorting exercise provides an opportunity for stakeholders to suggest other concepts that might be missing, the start is defined by a set of candidate concepts that derived from the expert-based first specification (Section 5.2.1). Determine the scope of the exercise

Having too many concepts to organise will introduce a significant cognitive overload in the participants. The number of concepts will be limited to 50 or less. The goal is to determine an agreed upon skeleton for the taxonomy. In case of doubt, it might be useful to run the exercise with all relevant concepts split into two or more subsets (thus, iterations). With a follow-up session to consolidate results of each iteration into one single taxonomy skeleton.

PREPARE THE CARDS

Have each concept within scope of the exercise written in a different card. There should also be blank cards for stakeholders to add previously unidentified concepts. There should also be blank cards of colour different from the concept cards. These coloured cards will be used to create categories under which the concepts will be organised.

Participants in workshops will be working on the same set of concepts to collaboratively build an agreed taxonomy. Each workshop will develop its own version of the taxonomy, which will be compared to assess whether there are different interpretations of the domain. Any such differences will be accommodated in subsequent workshops, and by the experts referred to in Section 5.2.1.

[15] https://www.usability.gov/how-to-and-tools/methods/card-sorting.html

[16] https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/card-sorting

EXECUTION PHASE

Participants will be encouraged to think aloud. This will happen naturally if they are working on teams, but even if they are working alone, they should do this. Where possible, this 'thinking aloud' will be recorded.

Participants will be requested to sort the cards in groups that make sense for them. After they have done this each of the resulting groups has to be labelled as a higher-order term. Finally, participants are requested to establish relations between these higher-order terms. Participants will be encouraged to add new cards if they feel that important concepts are missing or discard concepts that are not relevant.

AFTER THE SESSION

From the results of the workshops a skeleton of the taxonomy will be created in PoolParty Thesaurus. Categories of concepts with relations between them other than SKOS broader/narrower are to be modelled as different concept schemes. Narrowers of these are added in the taxonomy. Other relations are created as part of a taxonomy.

5.2.3 Corpus analyses

Document tagging is based on one or more taxonomies. This means that the richness of the taxonomy is a determining factor in providing a good experience in doing e.g. semantic search. Concepts in text will not be found if they are not modelled in the taxonomy through a preferred, alternative or hidden label. The question of what should be added to the taxonomy is highly dependent on the domain and nature of documents to be searched for. Experts will be asked to apply the taxonomy to texts, and what leads them to say that a text is about a certain topic. The question is then how to enrich the taxonomy. This is clearly a task for the domain experts, individually (see Section 4.3) and in workshops (Section 4.2). The starting point will be a taxonomy skeleton produced during previous activities (see section 5.2.1) and a set of very relevant documents for the area that needs enrichment.

Usually a good starting point is to have around 100 domain specific articles / documents in place. For the SSHOC pilot the team will make use of articles on electoral participation that will be identified and provided by the domain experts of the pilot team.

The process would be as follows:

- Build a document corpus with the identified documents. If there is a very large set of documents in place, subsets of documents need to be prepared. As a rule of thumb, an order of magnitude of 100 starts to bring useful results but around 1000 (relevant) documents should provide really good results in the corpus analysis.
- Run the corpus analysis.

    - PoolParty here by leveraging the statistical model generated from the taxonomy concepts and relationships, as well as the corpus data, creates a statistical analysis of the terms, as well as concepts, their co-occurrences:
        - one can sort by different scores such as frequency (total number of occurrences of the term in the corpora), relevancy, "multi-token" terms and so on.
        - one can also find so-called "shadow concepts", i.e. concepts that do not appear in the corpora, but due to co-occurrences between terms and concepts are derived
        - one can also see if the taxonomy is a good fit for annotation based on the results, i.e., if it has a good coverage
- For the purpose of this activity, the domain experts will be instructed to focus on the extracted terms. They will look at the list of extracted terms and for each of them decide if:
    - the term is relevant as a concept and added as a 'candidate concept' before it gets integrated in the taxonomy hierarchy as described next
    - It should become a new concept (and thereby a part of the KG), and where should it be placed in the taxonomy hierarchy
    - It should become an altLabel or hiddenLabel of an existing concept
    - It is not relevant for the taxonomy and it can safely be ignored
- The expert goes through as many terms as needed until they run out of resources or until they consider that the taxonomy is sufficiently rich to support the document tagging.

Note that:

- The candidate concept mechanism of PoolParty Semantic Suite (the software tool used for this work) can help adding the new knowledge into the taxonomy.
- Once a term has been added as a concept to the taxonomy, it will not appear as a term in subsequent corpus analyses, as it will be found as a concept and the process continues iteratively

MANUAL TAGGING

The case of the SSHOC Pilot involves tagging documents, thereby it is an effective technique to organise a workshop with domain experts to create the skeleton of a taxonomy. This approach is a variation of the card sorting technique described in the Card sorting exercise section (see section 5.2.1), with the difference that one needs a corpus. Depending on the scope of the workshop and the available resources (e.g. an initial taxonomy) the workshop will provide the following advantages:

- The skeleton of a taxonomy is agreed upon
- The domain experts get a better understanding on the tagging process and its limitations
- The domain experts get an initial hands-on experience using the (PoolParty) Thesaurus

This technique is based on extensive experience at the SWC. In a nutshell the following steps are necessary and will be carried out in the SSHOC Pilot project:

- Texts in the domain of interest (electoral participation) of the domain experts are identified. Ideally, these texts are part of the sample of texts provided by the domain experts as part of the target set of texts to be tagged by the use case. If we would not have access to these texts, or we do not have the sufficient domain understanding to select appropriate samples, we need to ask the domain experts liaison to pick such texts for us. In the SSHOC Pilot project the pilot team consists of domain experts that will support this. The choice of texts must be appropriate:
    - The text should be highly relevant for the use case, as it needs to be representative for the texts that will be targeted by the real / production system.
    - The text should fall within the expertise domain of the experts attending the workshop. This can be non-trivial in large corporates (or domain-specific communities), as areas of expertise are well-delimited, and people tend to have a deep understanding of relatively narrow topics.
    - The text should be relatively short. Although key guideline documents that describe certain areas in detail for 50 or 60 pages will be useful at larger stages, for this exercise paragraphs of 10-20 lines have much more appropriate length.
- During the session, the KGraph facilitator explains the purpose of the exercise (as per the description in this section) and kicks off the exercise. Participants are shown the text on a screen. The experts are asked to identify the key concepts in a passage relevant for their domain of expertise. These are the concepts that an expert would highlight if asked "what is this text about?". These concepts will later become part of the taxonomy.
- The resulting concepts will be sorted into concepts that are part of the same category.
- Experts will be queried about the relation between the categories in terms of things that are subtypes of others, or examples of others. This leads to a hierarchy of concepts that will materialise in the (PoolParty) Thesaurus.

With this the basic skeleton of the taxonomy will be obtained. From this point on, the experts can go and enrich each concept scheme. One possible approach to enrich the taxonomy is described in 5.2.1 the Corpus Analysis section.

5.2.4 Iterative further development

The creation of KGs is not a linear process, but rather an iterative one. On the one hand, the creation of the KG can be reviewed after a certain point in order to see how well it performs. For instance, if the KG is used to drive the search, then the search results are evaluated and provide an indication on how well the KG is constructed and serves the use case. On the other hand, the iterative process can be continued by ingesting different (structured) data sources in order to make the KG better describe the domain, and by that better serve its intended purpose.

As indicated in the previous sections, the building of a KG starts with building a domain-centred taxonomy. Then, ontological elements are added (iteratively) to the taxonomy. As an illustrative example, given two concepts in the taxonomy, typically described by SKOS[17], can be described to be 'skos:related' between each

[17] https://www.w3.org/2004/02/skos/

other, which might not be sufficient in practice. What if the relationship is "qualified", meaning that the relationship between those two concepts is specific like 'isLocated', 'belongsTo' and so on? In that case dedicated relations between those concepts will be added in order to better describe the intended domain. For that such relations can be defined either as properties (relations between concepts) or as attributes (concepts that have data type as range). Those relations are typically described by OWL standard, and on top of that classes can be assigned to the concepts, in order to determine their class type. The techniques described in the previous approaches can be re-used repeatedly to further refine the KG. This process can be also done by ingesting data in an automatic way, as indicated by the following.

INGESTING FROM THE STRUCTURED DATA

A KG can be serialized in RDF format, which can be consumed by other applications. This can be achieved either via exports or bespoke APIs. By default, each KG can serve the clients and applications via so-called SPARQL endpoints. RDF is a structured format that enables to easily map and transform various heterogeneous (semi-) structured data to this format.

If structured data is located in a legacy relational database, then one can map the data from tables to either put in the taxonomy as SKOS or convert it simply in RDF instance data by using any OWL properties or attributes.

For instance, via different structured queries as in SQL, one can deduce if certain columns have repeating values, if that is the case then typically, they have to be encoded as a controlled vocabulary and have to go to taxonomy as concepts. Then, one can check other columns and see if relations can also be derived from such columns between concepts, if that's the case then they have to be populated in the taxonomy accordingly.

Once the taxonomy data are specified, the rest can be ingested as simply RDF instance data by providing the mapping using OWL ontology one has defined, for example if there is a column that describe where an employee works, then the "worksFor" property can be mapped in the ontology. The ingested data can go to a separate graph in the KG. In this way the KG can be further enriched.

The governance workflows to ingest new data, check for changes in the existing data can be set up to automatically refine the KG in an iterative process. Such workflows can be achieved by defining an orchestrator component, e.g., PoolParty UnifiedViews[18].

5.3 Knowledge Graph Governance

"Poor data quality is the unintended consequence of data silos and poor data & analytics governance." (Gartner, Inc: 'Think Big, Start Small, Be Prepared — Master Data Management', Sally Parker and Simon Walker, October 2019)

[18] https://www.poolparty.biz/agile-data-integration

This statement shows the importance of policies and governance models being in place as soon as any work is done with data, as high quality of data is crucial for the quality benefits and value of any data-driven end user application.

So, if one starts to develop a KG in the form of taxonomy and ontology for a well-defined prototype, knowing very well what data will be used, the first thing to do is to see what is already there! There are already a lot of great taxonomies and ontologies out there for different domains, commercial and non-commercial. However, as discussed in Section 5.1, most of what already exists is useful for identifying whether data or publications belong to the domain of electoral studies (rather than to other, unrelated domains), but not very useful for identifying, classifying or connecting materials within the domain of electoral studies.

So, the project will set up a governance process around the ontology and taxonomy management process. Especially the design of the Pilot's ontology affects the retrieval of data in smart applications and it also will affect the use of external data, as other people may use existing ontologies in a different way.

5.3.1 Taxonomy Governance

In contrast to the rather static, often monolithic and usually hardly collaborative process of maintaining classification schemes (e.g. Dewey decimal system[19]), the development of taxonomies and thesauri, especially when they are later used as part of a larger (domain-specific) KG, is highly collaborative, networked and agile.

The purpose of taxonomies is mainly to tag and retrieve content later, but not to model a domain of knowledge or establish a strict regime for the later classification of digital assets.

In many cases there are several taxonomies for a domain. These are managed by different groups of people according to their custom governance model, and in many cases these taxonomies are linked together to form the backbone of a larger (domain-specific) KG. This can be achieved through a central vocabulary hub or through a more decentralized approach (similar to peer-to-peer networks) and requires a different way of thinking than that often developed by traditional librarians or cataloguers.

Managing taxonomies also means establishing a continuous process to ensure that new developments in the domain are well reflected and incorporated. This requirement deserves a balanced process in which automatic and manual work support each other.

In short, a domain-specific taxonomy governance model[20] for the sustainable management, linking and provision of domain-specific taxonomies throughout a community must be well thought through and strictly aligned with the objectives of a larger KG initiative. A set of roles, responsibilities and processes must be defined

[19] https://en.wikipedia.org/wiki/Dewey_Decimal_Classification

[20] Taxonomy Governance Best Practices (Zach Wahl, 2017), https://enterprise-knowledge.com/taxonomy-governance-best-practices/ [22 March 2020]

to manage the development and application of a taxonomy so that it remains consistent and coherent over time (See also Section 8).

PROCESS MODEL

Developing taxonomies for a community means bringing together different stakeholders to agree on the scope, structure and content of a knowledge area. In addition, a common understanding of the objectives to be pursued through the development of taxonomies and the higher-level KG must be established.

This agreement process rarely starts on a greenfield site, but in this case the method of card sorting can often provide a good starting point. See further information in Section '5.2.2 User-involvement via sorting tasks'.

Furthermore, in many cases, thanks to the open standards of the Semantic Web, it is possible to fall back on already well-developed taxonomies, which often provide a solid starting point for further steps in various industries or domains. In the field of electoral studies, and more specifically of citizens behaviour in electoral studies, this is not the case, as was discussed in Section 5.1. Therefore, the Pilot team needs to develop most of it from scratch, while reusing, where possible, parts of existing taxonomies in other domains (e.g. from social science or from EuroVoc, the Thesaurus of the European Parliament[21]).

Suitable software tools can be used to extract subgraphs from larger KGs such as DBpedia[22], which can then serve as base taxonomies[23]. Furthermore, with the help of a reference text corpus and the corresponding corpus analyses, it can be determined which topics within the defined area should be represented in a taxonomy in any case. Corpus analyses (see also Section: 5.2.3 Corpus analyses) can also play an important role later on in the ongoing development and enhancement process of the taxonomy (as illustrated in Figure 4).

[21] EUR-Lex official website: https://eur-lex.europa.eu/browse/eurovoc.html?locale=en [22 March 2020]
[22] DBpedia website: https://wiki.dbpedia.org/ [20 March 2020]
[23] Harvest Linked Data to Generate a Seed Thesaurus (PoolParty Manual, 2020), https://help.poolparty.biz/pages?pageId=35921550; [20 March 2020]

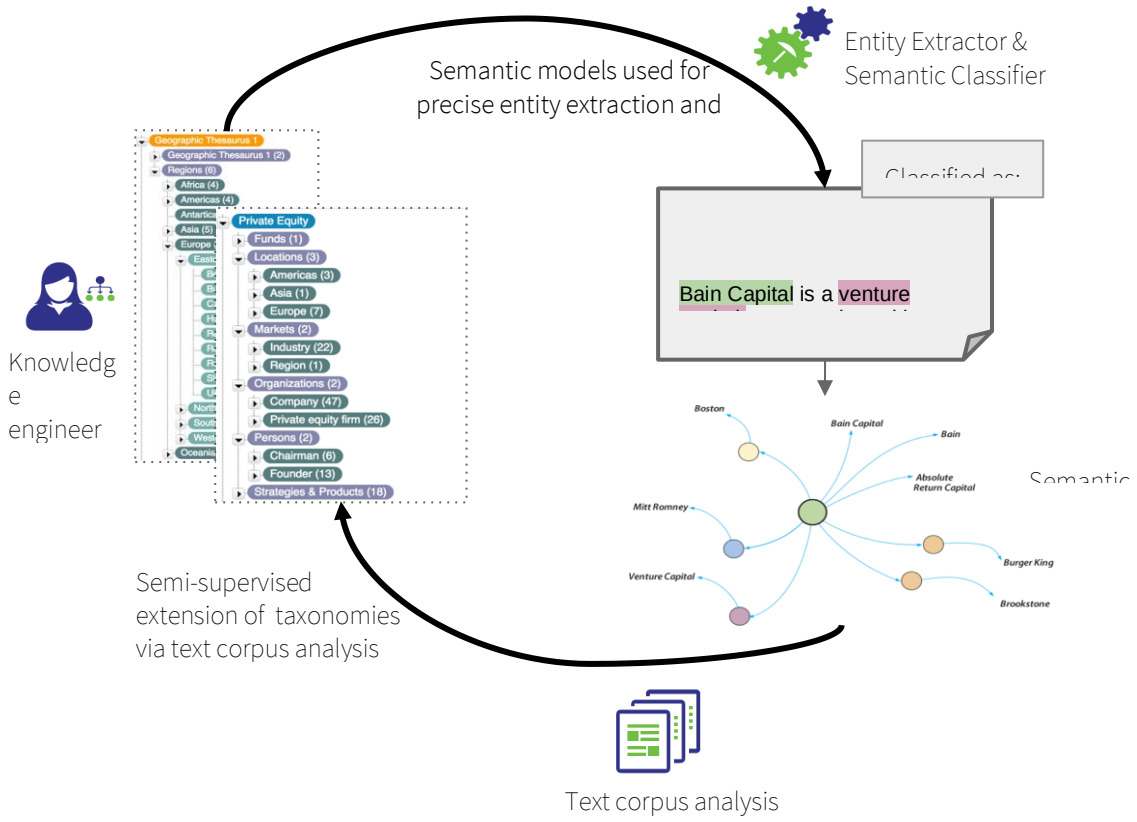

Figure 4: Interplay of Corpus Analysis and KGraph modelling

The process model will also make greater use of crowdsourcing methods (see also Section 4.3). For example, if a suitable user interface is provided that allows each user to suggest missing concepts or labels (for example, embedded in a tagging or search dialog), or if the search behaviour of users is simply analysed using search log analysis, then, in conjunction with a suitable approval workflow, this can lead to a taxonomy that grows with user requirements and can quickly identify missing components.

5.3.2 Ontology Management

Available methods for ontology management differ more than the approaches for taxonomy management. There are several reasons for this:

- The range of semantic expressivity and complexity of ontologies is much wider than is usually the case with taxonomies. In many cases, ontologies as well as taxonomies are concentrated on hierarchical or is-a relations.

- In some cases, however, the development of ontologies also has a strong focus on axioms, which goes far beyond the expressiveness of SKOS taxonomies.
- Some ontology management approaches combine all building blocks of the semantic knowledge model into one ontology, i.e. classes, instances, mappings. Other approaches are focussed only on the creation of the schema (called the 'TBox'[24]), but not on the facts or instances ('ABox').
- Some ontology engineers still stick to the idea of building expert systems in the classical sense instead of supporting the Semantic AI approach (aka 'Knowledge-based AI'[25]), which has a fundamental impact on the design process since basic concepts of the semantic web like the open world assumption are not applied in this case.
- Many ontologies do not have any project goals in mind or requirements of applications that should be based on them. In order to develop universally valid ontologies (sometimes also called 'upper ontologies'), different design principles and management methods must of course be applied than for specific ontologies that are often only relevant for a single subdomain. This leads to confusion, and some people believe that *the ontology is already the KG*.
- It also implies that the pilot-KG will remain realistic and agile, avoiding the temptation to try to come up with the ultimate ontology that will be the single source of all truth. Ontology management, and the development of KGs in particular will remain an ongoing process iterating and evolving along the learning curves of involved stakeholders.

The SSHOC Pilot project will make use of all the above listed and described methodologies and tools to create and publish a useful KG on Electoral Studies, that can potentially become a de-facto standard in the field.

5.3.3 The Knowledge Graph Governance Model

KGs have been recognized as an efficient approach for data governance, semantic enrichment, and as a data integration technology that brings unstructured and structured data together. Use cases for graph technologies make use of automatically generated unified views of heterogeneous and disconnected data sources. These 'virtual graphs' provide richer data sets to feed analytics platforms or to train AI algorithms. Subsequently, advanced tools for knowledge discovery and data analytics can be built based upon a semantic layer.

KGs are not 'just another database'

Appropriate methodologies to build, maintain and govern the KG itself in a sustainable way and on a large scale are not obvious for many potential users and stakeholders. KGs are not 'just another database', they rather serve as a vehicle to rethink and rework the existing data governance model, while, at the same time, a governance model for the KG management itself has to be developed.

[24] The Fundamental Importance of Keeping an ABox and TBox Split (Michael K. Bergman, 2009), http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/

[25] Knowledge-based Artificial Intelligence (Michael K. Bergman, 2014), http://www.mkbergman.com/1816/knowledge-based-artificial-intelligence/

How to mitigate the risk of generating a 'not-invented-here syndrome'

Additionally, various stakeholders have to be positioned well as integral parts of KG initiatives to mitigate the risk of generating a 'not-invented-here syndrome'. This will include roles like enterprise architects, data scientists and analysts, data warehouse specialists, knowledge managers, and of course all the business lines that ultimately will benefit from KGs.

Involving business and data stewards as soon as possible is essential, since users will become an integral part of the continuous KG development process nurturing the graph with change requests and suggestions for improvement.

Laying the foundation of a KG governance model

Some foundational decisions have to be made for the SSHOC Pilot project:

- Which parts of the graph will be governed merely centrally, which are driven rather by collaborative and decentralized processes?
- How can diverse requirements be fulfilled, also with regards to differing ideas of what a good-quality KG actually is.
- Which parts of the KG can be generated automatically without harming the overall quality criteria, which elements have to be curated by human beings?
- Which data elements, e.g. structured datasets or already existing taxonomies can be incorporated potentially into the evolving KG?

During the development of the Pilot-KG in electoral studies these issues are still quite manageable, as the data-stewardship, evaluation, quality assessment and choices about incorporation of existing datasets and taxonomies are all in one (collective) hand: the team working on the production and delivery of the pilot-KG. But these issues are of utmost importance for the period after the initial development and the production of the pilot-KG. Therefore, this developmental process (the delivery of D9.9) will include sustained efforts to resolve these issues (see also the discussion in Section 8 about post-delivery development).

Introducing quality metrics and KPIs at the right point in time

Introducing quality metrics and KPIs at the right point in time of a KG project is a key success factor. Any KG project shouldn't remain a stand-alone initiative but should be embedded in the organisation's or domain-specific community's overall data governance framework as early as possible.

A closer look at the anatomy of a KG, its layers, and its structural elements, allows a better understanding of the methodology, as well as a more specific definition of the parameters that potentially will have an impact on the quality metrics of a corresponding KG governance framework.

The SSHOC Pilot project will identify and specify 'key questions to be answered by the KG' as most important KPIs. The details will be developed in the implementation phase, but the following key questions could already be identified:

- How can end users (target groups, see section 3.1 Intended audience) find relevant information and data on citizens behaviour (in the field of electoral studies), by using clearly specified filter mechanisms and, where possible, question-answering functionality (asking questions in natural language)
- Where and how can end users find classified and annotated information, documents and data in the field.
- How can end users find relevant other stakeholders (e.g. scientists) in a specific field of electoral studies.

KGs as an extension of data governance in place

Ultimately, KGs as an agile approach for data management based on the linked data life cycle implies the need for an extension of the existing data governance framework. Any graph project triggers changes on various layers of an organisation and its information and data architecture:

- New roles, their interplay and their responsibilities have to be defined.
- Content and data authoring / curation processes will be extended and partially automated.
- Diversification of access points to data and knowledge have a direct impact on the existing data governance model.
- New ways to gain insights to enterprise data will be developed, e.g. automated generation of links between data points which initially were not connected yet.
- These new insights, in return, trigger new questions related to GDPR compliance.
- Algorithms that automatically generate personalized views on content and data enhance customer experience.
- New ways to filter and contextualize data objects will be available 'as a service'.
- New and diversified perspectives on data quality make the necessity to establish a Data Governance Board even more obvious.
- New ways to make use of data standards and harmonized metadata boost the value of existing data sources.

In its inner core a sustainable methodology to create and maintain KGs will be implemented, which will change the use of AI technologies substantially: while ML components continuously learn from raw data, advanced AI systems combine this with existing knowledge from the KG, and produces new knowledge, answers and explanations at the same time.

Conclusion: KG Governance

To ensure that a KG is of high quality and useful over time a dedicated KG Governance Model needs to be developed and will be developed for the SSHOC Pilot project in the course of implementation.

This means the pilot project team will develop a model that clearly specifies roles and processes to maintain the KG in regard to:

- add / expansion (create)
- changes (edit)

- deletion (remove)
- publishing
- linking (e g with other KGs)
- continuous corpus management
- license strategy (which license do parts of / the whole KG have)
- workflows (if applicable) for certain tasks (e g adding a new concept)
- the overall decision-making processes

The governance model shall include clear processes for all listed areas but should still be lightweight to avoid complex and complicated decision-making processes.

6. Technical KG specifications and development

During the development process of a KG the Linked Data Life Cycle[26] approach will be followed. Developing a KG along the whole Linked Data Life Cycle is not a linear process. In all phases, several activities take place in parallel, and are highly dependent on each other, see the next figure with the detailed description about each phase.

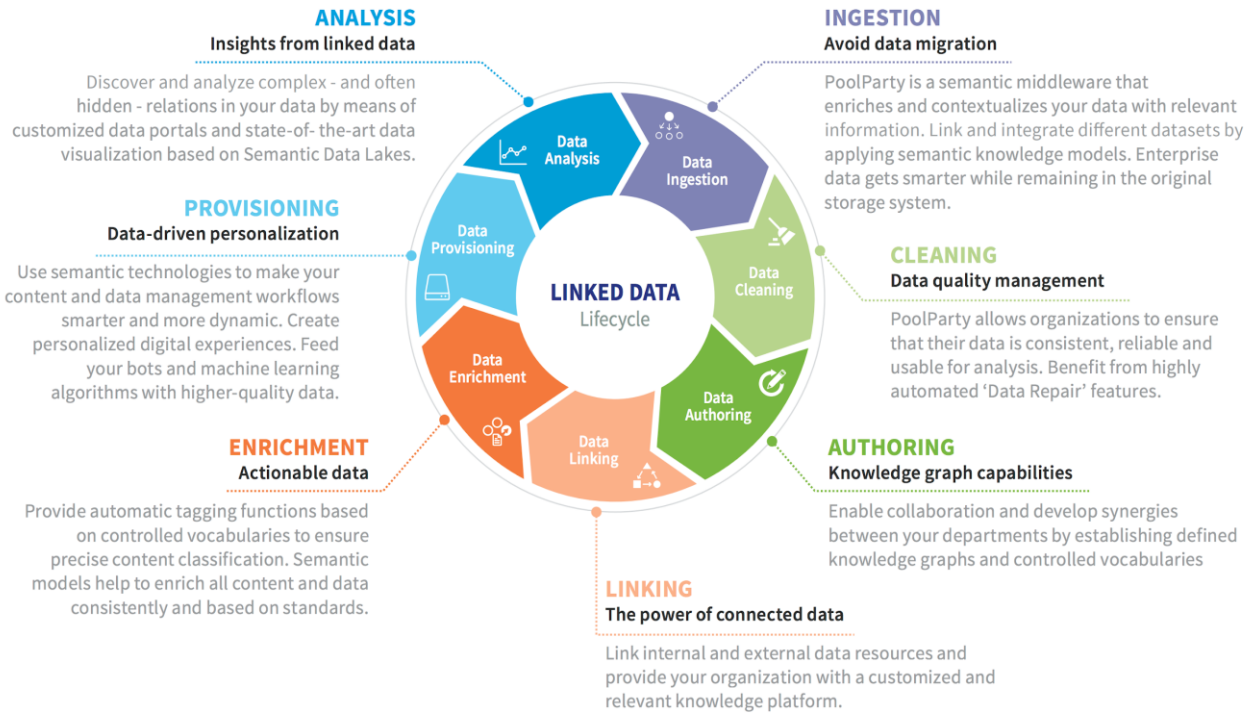

Figure 5: Linked Data Lifecycle activities and their description

As an initial step, the team will define more specifically which insights users would like to gain from a resulting "semantic data lake". The generation of a list of key questions can serve as a guideline: which types of questions should be answered by our digital systems if a perfect database would exist for our user community? In addition to this rather user-centric approach, also resulting data services for automatic bots should be defined more specifically.

This has an immediate effect on *Data Ingestion*: in which existing data sources are at least chunks of information available, which could be used together with other sources to get 'unified views' on our data during *Data Provisioning* and *Data Analysis*?

[26] The term "Linked Data Life Cycle" was coined for the first time in 2012 by the LOD2 project consortium. https://link.springer.com/content/pdf/10.1007/978-3-642-35173-0_1.pdf

During the *Data Provisioning* phase information needs are further analysed down to the level of specific services. Contexts are introduced, which could be described by virtual 'personas'. This specific view on the need for data will give, in return, also a clearer picture on which aspects of data quality should be focussed upon.

As a result, missing pieces of the envisioned KG like, for example, domain-specific taxonomies can be identified. Which concrete steps during *Data Authoring* have to be made will depend on the requirements of *Data Linking* and *Data Enrichment*. Efficiency of data linking on entity level and of schema mapping is dependent from the quality of available ontologies and taxonomies, which are essential building blocks of any KG.

Iterating across the various steps involved in a Linked Data Life Cycle results in a higher degree of linking, which in return helps to improve quality of automatic *Data Enrichment*, and semi-automatic support for *Data Authoring*. What seems to be an endless loop at first sight is very similar to processes that can be seen in software engineering. KGs become increasingly complete, serve as means for more and more accurate semantic enrichment services, or will begin to learn from user behaviour or even crowd-sourced efforts.

In the following, particular steps in the LD life cycle will be discussed in somewhat more detail, focusing on technical specificities.

Data Ingestion

The software component that will be used for the Pilot project is ***PoolParty Semantic Suite***. PoolParty is a semantic middleware that enriches and contextualizes data with relevant information. The platform links and integrates different datasets by applying semantic knowledge models. Enterprise data gets smarter while remaining in the original storage system.

Different connectors in PoolParty Semantic Suite are capable of ingesting data, and afterwards use that data in conjunction with the existing one. For instance, in PoolParty one can ingest RDF data via means of Excel import, RDF import, or APIs, hence creating a new or refining the existing taxonomy. In PoolParty's UnifiedViews[27] component, one can download CSV, XLS or RDF data from an (S)FTP location on a server, ingest data from a SPARQL endpoint[28] etc. Each of these functionalities is possible due to availability of DPUs capable of doing the respective tasks by encapsulating the bespoke logic.

[27] Pool Party website: https://www.poolparty.biz/unifiedviews/; [22 March 2020]

[28] Medium Open Link Virtuoso Blog: https://medium.com/virtuoso-blog/what-is-a-sparql-endpoint-and-why-is-it-important-b3c9e6a20a8b [22 March 2020]

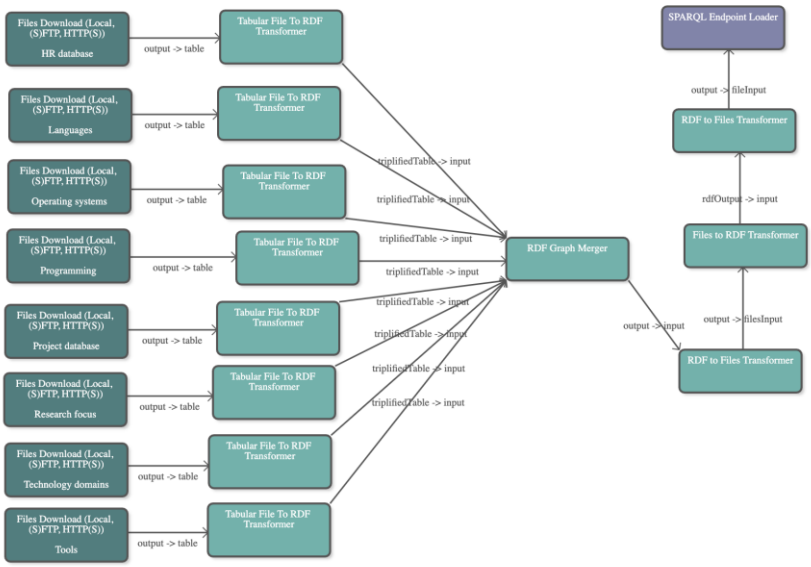

Figure 6: PoolParty UnifiedViews pipeline consisted of building blocks - so-called DPUs - and the data flow depicted with arrows

Data Cleaning

PoolParty allows organisations and/or domain-specific communities to ensure that their data is consistent, reliable and usable for analysis. Users benefit from highly automated 'Data Repair' features.

The ingested data (that in the SSHOC Pilot is harvested from different data sources) is typically far from being clean. Thus, in PoolParty if one imports any taxonomy data, then SKOS quality checks are run to check for its quality. In case there are violations, if data are prone to be inconsistent or incomplete, then the user has to intervene and choose between the options such as deleting or adding data. In UnifiedViews, one can always use ASK queries or SHACL rules[29] in order to check if RDF data respect the model that the application is expecting.

Data Authoring

KGs enable collaboration and develop synergies between departments by establishing defined controlled vocabularies.

Typically, the data is spread across different named graphs, each serving its purpose. In UnifiedViews it is very common to change data resulting from a given named graph of a remote store. Running SPARQL Update queries on top of (union of) named graphs is a very common task as well. PoolParty GraphEditor[30] is capable

[29] https://www.w3.org/TR/shacl/

[30] https://www.poolparty.biz/poolparty-grapheditor/

of also doing that, namely access and process all required data in their graph context and process them in one go.

Data Linking

Data resources are automatically linked on entity level to build the basis for a knowledge platform of an organization.

Once entities are defined in different named graphs, they will be related to each other. First, entities are studied, and then decisions are made how to align them, for instance, based on label matching, or even more sophisticated patterns (regular expressions etc.). With specific DPUs in place, even more sophisticated algorithms can be used like spreading activation[31].

Data Enrichment

Provide automatic tagging functions based on controlled vocabularies to ensure precise content classification. Semantic models help to enrich all content and data consistently and based on standards.

Data does not always come in a structured form. Data can be obtained in PDF, Word format and the task is to extract meaning out of them. Or, structured and unstructured data can be linked together (see 'Data Linking'). Based on a taxonomy, PoolParty Extractor is able to extract tags of such a data file, or PoolParty Semantic Classifier[32] is able to classify the file in one of a specified category.

Data Provisioning

Use of semantic technologies makes content and data management workflows smarter and more dynamic. Personalized digital experiences can be created. ML algorithms feed bots with higher-quality data.

Data processing in RDF is typically done via UnifiedViews. Once this process is finished, then the data is stored in one of the triple stores by specifying a named graph. Different applications can get the data using the SPARQL endpoint or the REST APIs.

Data Analysis

Discover and analyse complex - and often hidden - relations in large datasets by means of customized data portals and state-of- the-art data visualization based on Semantic Data Lakes.

After data is stored in the format and based on the data model needed, one can do further analysis. PoolParty GraphSearch is the component tailored for this task. One can drill-down information based on facets until one

[31] https://en.wikipedia.org/wiki/Spreading_activation
[32] https://www.poolparty.biz/semantic-classifier/

gets the information satisfying the criteria. Also, pie and bar charts are in disposal to get graphical insights on the data (see Figure 7).

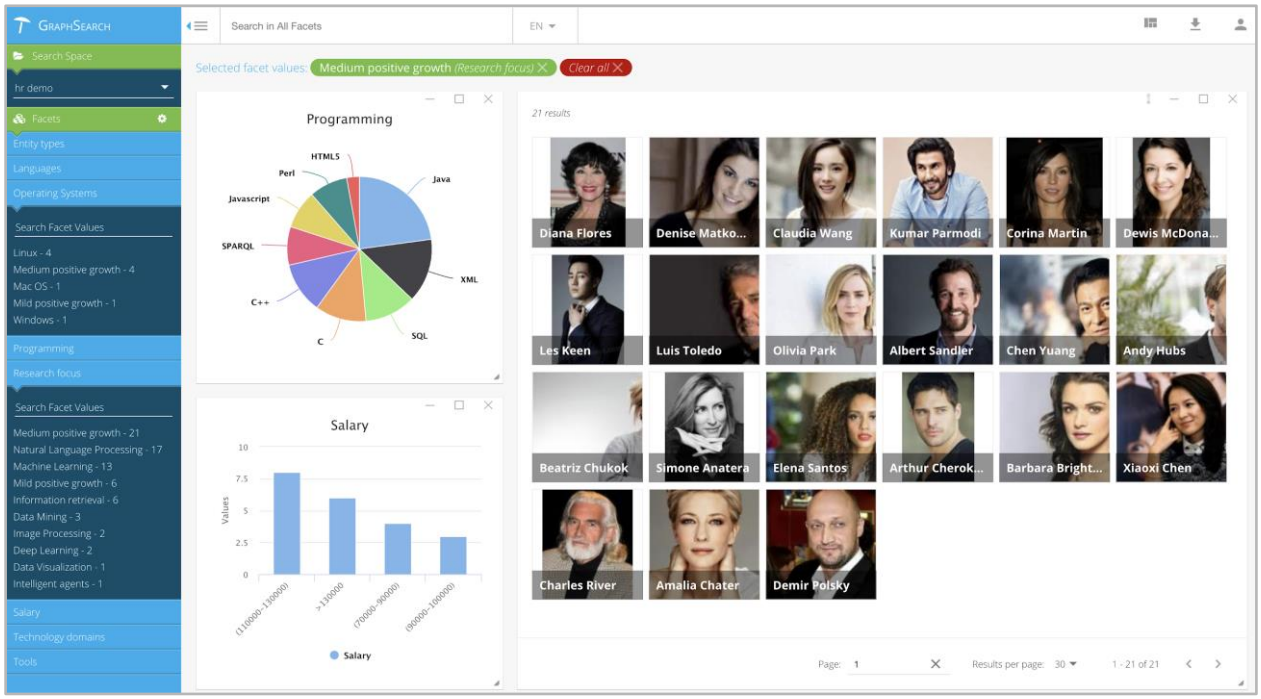

Figure 7: PoolParty GraphSearch as a semantic search application that can reside on top of unified data

In the end, Figure 8 illustrates how PoolParty components are mapped to Linked Data Lifecycle that is the basis for the (technical) implementation of the SSHOC Pilot project.

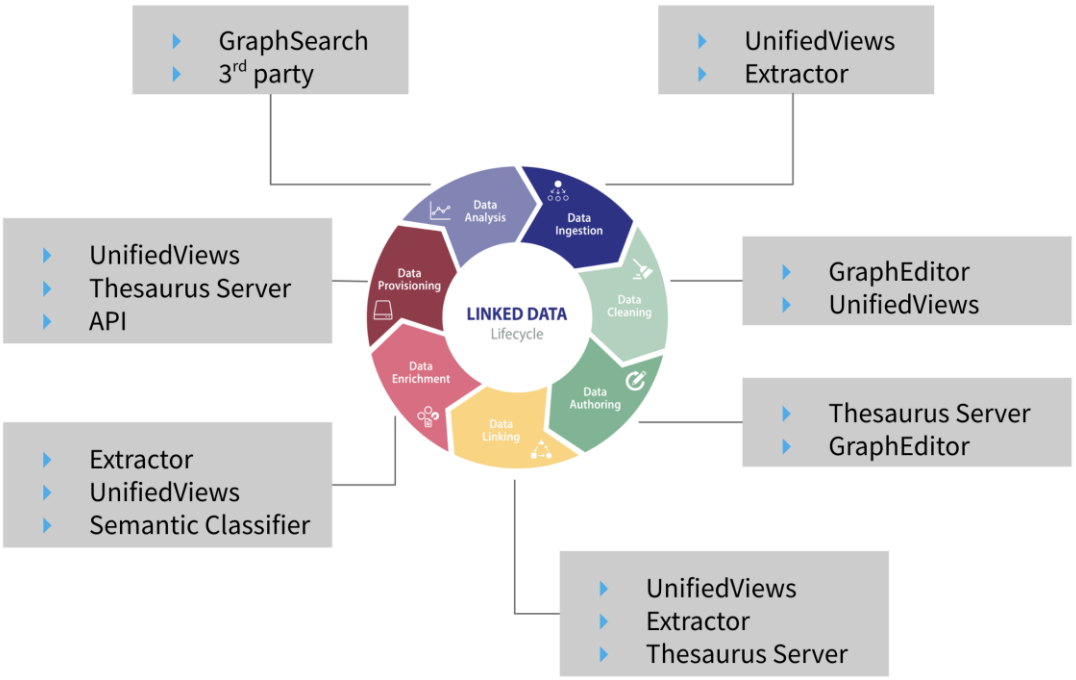

Figure 8: PoolParty components mapped to Linked Data Lifecycle, each serving different activities

7. Testing

Given the heterogeneity of the sources used to create a KG, it is a common fact that the data extracted for the seed KG will be incomplete, and will probably contain duplicate, contradictory or even incorrect statements[33]. Additionally, the resulting KG should be able to serve the requirements that drive its construction. Therefore, assessing the quality of the KG is an important step after the initial creation and enrichment of a KG.

Here a twofold approach will be used in order to test the quality of the KG, which will eventually contribute to the construction of the KG as well: on the one hand a 'test driven' approach (see Section 7.1), and on the other hand a 'competency question' approach (see Section 7.2). .

Crowdsourcing mechanisms will be used in these approaches (see Section 4.3 Recruitment for and tasks for crowd-involvement) to continuously (i) check the KG development and (ii) ensure involvement of relevant stakeholders (here scientists in the field of electoral studies). This approach enables an additional testing cycle to ensure the KG is developed in a way that is useful for the end users.

[33] https://arxiv.org/abs/2003.02320

7.1 Test Driven Knowledge Graph Construction

The notion of Test Driven Software Development is a methodology followed to develop software components. Tests are specified before the actual software component is implemented, so the target of the implementation is to develop a component able to pass the tests. The same idea is the basis for TDKGC. Tests are specified before the target KG is constructed, leading to a guaranteed quality of the KG by making sure the tests are passed. This method forces the authors of the KG to think about the requirements before the construction of the KG and will help to detect and resolve errors sooner. Though, designing a complete test suite for a KG is hardly easier than designing the KG itself.

The TDKGC method will be used to test the KG in terms of consistency and input validity.

In particular, constraints will be provided as a set of conditions, formulating a "shapes graph" using SHACL[34]. This will allow the systematic testing of the KG during development, as well as ensuring long term consistency.

7.2 Competency Question Driven Knowledge Graph Construction

A CQ is a sentence in natural language, representing a pattern for a type of question one would want to be able to answer by using a KG. For instance, in a KG regarding research papers, in order to answer a question "Which paper uses x method?", the KG should contain concepts like Papers and Methods. Additionally, their instances should be able to have a relation called Uses.

The first set of CQs or key questions are listed in section 5.3.3 The KG Governance Model. A more comprehensive set of CQs will be developed in the course of Pilot implementation phase.

Elicitation of the questions will be performed on the requirements input as described in Section 3, as well as the targeted Community Involvement described in Section 4. Gathering these types of questions will allow us to form test queries against the KG to test its fitness for the purpose.

[34] W3C Recommendation website: https://www.w3.org/TR/shacl/ [20 March 2020]

8. Post-delivery development

During the development of the Pilot-KG in electoral studies, the entire process of further detailing of design choices, implementation, and quality assessment is concentrated in the hands of the team that is designated to deliver this (this team is composed of members from the Semantic Web Company (SWC), the University of Vienna (UNIVIE), the Austrian Social Science Data Archive (AUSSDA), and the University of Nottingham (UNOTT). This team will involve as much as possible, the user community of electoral studies, as discussed earlier in Section 4.

A yet unresolved question is how the pilot-KG to be developed will be governed, curated, and further developed after the delivery of the pilot-KG, and particularly after the end of the SSHOC project.

The simple answer is that the KG will be made publicly available via SSHOC's Marketplace. But this answer is also overly simplistic, because it ignores inevitable issues of maintenance, possibilities and needs for further development, for refinement, aspirations for widening the domain (see the discussion in Section 2.1) and so on. Moreover, the KG is intended to become a part of the infrastructure of the user community in electoral studies, and to fulfil that function it needs regular updating, in addition to a gradual expansion of its currently intended domain of electoral participation, to include also other aspects of (the study) of voter behaviour, and other elements of the field of electoral studies (see Figure 3 in this report, and also the report of D9.6 *Demarcation Report of Electoral Studies User Community*). It is therefore necessary that some kind of organisation assumes responsibility for all these aspects of post-delivery maintenance, upgrading, refinement and expansion, including roles of data-stewardship and the organisers of user community involvement. Such an organisation would need to have a relatively stable organisational basis and would need to be a legitimate and authoritative body representing the user community of electoral studies. As discussed in the report of D9.6, this user community currently lacks such organisations. However, the MEDem initiative is well placed to remedy this. MEDem is an emerging European research infrastructure that connects many existing comparative and national projects in the field of electoral studies. It aims to gain a position on the ESFRI roadmap as a central and strategically important hub for electoral research. Its foundational meeting took place in February 2020. Once firmly established, MEDem could be an obvious body to assume responsibilities for the post-delivery roles sketched in this section.

The pilot-KG development team will liaise with MEDem, and possibly other organisations as well, to clarify by the time of delivery of D9.9 (the delivery of the pilot-KG) how post-delivery maintenance and further development of the pilot-KG can be assured for the future.

9. Planning

This section reports the planning of the development of the pilot-KG in terms of tasks and of a timeline. It does so with the usual caveat that where necessary this planning will be revised and adapted according to changing circumstances. This proviso is of particular relevance since the outbreak of the Covid19 pandemic, and the uncertainties this generates for holding meetings, for conducting planned workshops at the time of academic conferences, and for the health of the people involved in the development of the KG.

The period covered by this planning starts mid-March 2020 and ends ultimo February 2021, the due date of D9.9 (titled *Delivery of user validated Knowledge Graph, and Election Studies Analytics Dashboard*). This planning is the product of a 2-day workshop (26-27 February 2020) in Vienna of the entire team that will be involved in D9.9. Although the planning period reported here starts in March 2020, some of the work has already started earlier and has also contributed to the current report (which constitutes D9.7).

The following major tasks and subtasks are foreseen and planned as follows:

1. ***Ontology development*** (March 2020 - ultimo December 2020)
   As discussed at several places in this document, ontology development is a continuous process, starting with a relatively simple skeleton that is to be updated and refined after every new round of usage and feedback. For this reason, this task is planned as continuous until shortly before the deliverable is due. This 'ending' is, of course, artificial, as further usage and feedback, as well as developments in the domain of electoral studies will prompt the need for continuing development, however, that is beyond the planning of this pilot-KG (but, see the discussion in Section 8 about Post Delivery development).
   Although ontology development is a continuous stream of work in the development of the KG, specific parts of it can be distinguished and planned separately:
   a) *Construction of expert-based first specification (March 2020 – ultimo April 2020)*
      This part of ontology development was discussed in Section 5.2.1; it serves as the 'kick start' for further development
   b) *User involvement via sorting tasks (May 2020 – ultimo October 2020)*
      This was discussed in section 5.2.2. This will be done in separate meetings of, respectively, DACH, AK Wahlen and EPOP (see section 4.2). Originally these three meetings would take place in March 2020, May 2020 and September 2020, but the Covid19 pandemic has led to the postponement of the March and May meetings to October 2020 (on the -yet untested- assumption that the pandemic is brought under control by then). As a result, much of these planned activities will take place in the early fall, but their preparations will start earlier. Moreover, the team will assess what possibilities exist to organise some of this user community involvement on an individual online basis, starting at an earlier moment than the fall of 2020.

c) *Construction of a set of relevant publications for corpus analysis (April 2020 – ultimo May 2020)*
This relates to the discussion in Section 5.2.3 of this report and provides the basis for manual tagging and subsequent ML.

d) *Expansion of the skeleton ontology (June 2020 – ultimo December 2020)*
This involves the iterative expansion and updating of the skeleton ontology on the basis of experiences from 1b and 1c (above). Because of the rescheduling of the DACH and AK Wahlen meetings, it is expected that major efforts in this respect will occur at the end of October and in December 2020.

2. ***Specification of resources to be covered by the pilot-KG (March 2020 – ultimo October 2020)***
This part of the work relates to the discussion in Section 3.3 of this report, which also reports on some of the work that has already been done in this respect. However, additional work needs to be done on this basis, much of which will have to be done in the period until July 2020, with a minor extension to ultimo October to allow for the effective use of feedback from conferences, meetings and workshops discussed in Section 4.2 (see also point 1b, above).
This part of the work includes the specification of a designated subset of publications to be used in corpus analysis (see point 1c, above).

3. ***User involvement (June 2020 – ultimo November 2020)***
This relates to Section 4 of this document. User involvement that has to be channelled vie meetings or workshops has been included in point 1b, above. Additional work components include:

a) *Recruitment of participants in crowd-based user involvement (May 2020)*
This involves communicating with the user community of electoral studies in a variety of ways in order to recruit as large a pool of volunteers who can contribute to, e.g., expertise-based tagging and classification of resources.

b) *Construction of crowd-based tasks (May 2020 – July 2020)*
This involves the definition of the specific tasks that can be distributed to crowd-volunteers, as well as setting up a system of oversight and quality control of the returns from crowd-involvement.

c) *Implementing and monitoring crowd-based user involvement (July 2020 – February 2021)*
This involves the actual operation of crowd-based user involvement, the updating of crowd-tasks in line with ontology development, collating and systematising crowd-based contributions and feeding these back into ontology development. In principle, the possibilities for this form of user involvement do not end with the delivery date of the pilot-KG, or even with the end of the SSHOC project. Assignment of stewardship for such ongoing user community involvement must be agreed with other organisations, as discussed in Section 8 of this report, and as reflected in point 7, below.

d) *Organising and implementing crowd involvement in testing (October 2020 – January 2021)*
This relates to the discussion in Section 7 of this document and includes setting up ways in which crowd-volunteers can take part in testing, collating and systematising the results of such

testing, and implementing conclusions in the KG and its interface under development. As with other elements of this planning, the end date does not reflect an absolute end to the (continuing need for) testing, but merely a reflection of a planned delivery date. Here too, Section 8 of this report and point 7 below indicate that the team is aware of the need for post-delivery development, but how this will take its form cannot, at the time of writing, be specified.

4. ***Integrating data sources (July 2020 – ultimo October 2020)***
   This pertains to the implementation of what was discussed in Sections 5 and 6 of this report. This means that the specified data and information sources will be harvested by developing several harvesters (Data Processing Units, DPUs) in the UnifiedViews component of PoolParty Semantic Suite that allows fully automated data acquisition from specified and harvested sources afterwards. Data and information will be harvested continuously and will be mapped and linked to the KG to enrich the KG with all that domain-specific data and information. Where required data will be converted (into RDF) and harmonised (e.g. names or date formats if applicable).

5. ***User Interfaces development and implementation (October 2020 – mid-December 2020)***
   This pertains specifically to Section 6 of this report. The user interfaces will be developed on top of PoolParty Semantic Suite APIs, means the interfaces for the end users that will make use of the application on top of the knowledge graph. This will be search based applications like faceted semantic search interfaces, questions-answering mechanisms and recommender interfaces. The work includes the development of MockUps[35] that are non-functional and not designed screen prototypes that show all relevant elements of the application and explains its functionality, screen design prototypes that are designed mockups, and finally the full implementation that will be realised with JavaScript technology (or similar). Furthermore, this includes documentation and unit tests[36] of the software components and user interfaces.

6. ***Testing (January 2020- mid-February 2021)***
   As described in Section 7 testing will follow two approaches, 'test driving' and the evaluation of the KG performance when queried by CQs. Parts of testing are planned to be done via crowd-involvement (see point 3d, above).

7. ***Charting post-delivery development (May 2020 – February 2021)***
   As indicated at various places in this report (particularly Section 8), the development team considers the mapping of possibilities for post-delivery stewardship of the KG as part of remit. It is expected that this will require a continuous stream of contacts with relevant organisation of and in the user community of electoral studies. If possible, at all, the team will contribute to the actual organisation such stewardship and its responsibilities.

[35] https://en.wikipedia.org/wiki/Mockup
[36] https://en.wikipedia.org/wiki/Unit_testing

8. ***Delivery D9.9 (pilot-KG on electoral studies with dashboard interface) (ultimo February 2021)***
   The delivery will include the pilot-KG and its associated dashboard interface, a user guide, and a report detailing the experiences of the developmental process, 'lessons to be learned' for the further application of KGs in and for academic user communities, and prospects for further development.

References

Bergman, Michael K. 2009. "The Fundamental Importance of Keeping an ABox and TBox Split", AI3, online at http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/

Bergman, Michael K. 2014. "Knowledge-based Artificial Intelligence", AI3, online at http://www.mkbergman.com/1816/knowledge-based-artificial-intelligence/

Brant, Kenneth and Svetlana Sicular. 2018. "Hype Cycle for Artificial Intelligence", Gartner Research, online at https://www.gartner.com/en/documents/3883863.

Ehrlinger, Lisa and Wolfram Wöß. 2016. "Towards a Definition of Knowledge Graphs." *SEMANTiCS* (2016), online at http://ceur-ws.org/Vol-1695/paper4.pdf

Gligoric, Nenad and Martin Kaltenböck. 2018. "Enhanced data management techniques for real time logistics planning and scheduling". Online at https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5c03e0e8b&appId=PPGMS

Pan Jeff Z., Matentzoglu Nico, Jay Caroline, Vigo Markel, Zhao Yuting (2017). Understanding Author Intentions: Test Driven Knowledge Graph Construction. In: Pan Jeff et al. (eds) Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering. Reasoning Web 2016. Lecture Notes in Computer Science, vol 9885. Springer, Cham. Online at https://doi.org/10.1007/978-3-319-49493-7_1

Van der Eijk, Cees. (2020). SSHOC D9.6 Demarcation Report of Electoral Studies User Community. DOI: 10.5281/zenodo.3725823, online at https://zenodo.org/record/3725823#.XoSbSdMzajg