Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable 9.11 Concluding Report on T9.3

Insights obtained from the pilot Knowledge Graph in Electoral Studies

Dissemination Level	PU
Due Date of Deliverable	30/11/2021(M35)
Actual Submission Date	17/12/2021
Work Package	WP9 - Data Communities
Task	Task 9.3 Data Community Project: Electoral Studies
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p. 19

Abstract:
This report recapitulates the work of SSHOC Task 9.3 which consisted of the development of a pilot Knowledge Graph (KG) in Electoral Studies. It derives overarching insights that are also relevant for other intended developments of KG in disciplinary sub-domains of social sciences and humanities. The report may be augmented by addenda, as work in SSHOC continues into 2022.

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
0.0	14/10/2021	Structure and Initial content	Cees van der Eijk
0.1	14/11/2021	Draft Sections 1 and 3	Cees van der Eijk
0.2	20/11/2021	Draft Section 2	Cees van der Eijk
0.3	10/12/2021	Draft Sections 4 and 5	Cees van der Eijk
0.4	13/12/2021	Finalisation of content and editing	Cees van der Eijk
0.5	16/12/2021	Revision after WP leader review	Cees van der Eijk
1.0	16/12/2021	Final version for submission	Veronika Heider

Author List

Organisation	Name	Contact Information
UNOTT	Van der Eijk, Cees	cees.vandereijk@nottingham.ac.uk

Executive Summary

This report recapitulates the work done in SSHOC Task T9.3, the Data Community Project for Electoral Studies. Its purpose was to develop a pilot of a Knowledge Graph (KG) in the field of Electoral Studies. The first sections of the report recapitulate the background of and motivation for this purpose, and the manner in which the pilot-KG was developed. Section 4 discusses a number of overarching insights that derive from the work done and the experiences that were gained in the process. These insights are expected to be also relevant for other intended efforts to develop KGs in sub-disciplinary domains of the social sciences and humanities. These insights relate particularly to issues of feasibility and quality of metadata. The most important ones are:

- In spite of the wide variety of invaluable sources of metadata describing scholarly publications, datasets, and other relevant entities, available metadata are not without their problems. The most important of these include:
    - Granularity of many kinds of metadata –particularly those relating to the substantive character of original resources– is often too coarse to be useful when focusing on subdomains of disciplines. This complicates the descriptive characterisation of the content of scholarly publications and datasets, particularly when no widely accepted sub-diciplinary controlled vacabularies exist –as is the case in the field of Electoral Studies.
    - Sources of metadata vary in terms of coverage, richness, timeliness and quality, which leads to the need for KG developers to invest considerable effort in comparison of the strengths and weaknesses of available sources of metadata, taking into account these different aspects.
    - Quality problems are endemic for some types of metadata, most notably in terms of non-completeness, which sometimes reaches for centrally important types of metadata from highly reputed sources levels of 80%.
    - Quality of metadata about data used is generally poor as a result of poor data citation practice in the social sciences and humanities. Such citations lag far behind quality of citations to scholarly and other literature. Often, data citations –if available at all– consist of a free-form text string that is not directly machine actionable. Improvements require sustained effort from data creators and disseminators, but also from publishers, journal editors, and authors of scholarly literature.
- End-users of KGs must be aware that KGs are not and cannot be the definitive answer to all their information needs. Particularly in view of high rates of missingness for some types of metadata, KGs cannot be expected to exhaustively identify instances of the desired kinds of information.

Abbreviations and Acronyms

APA	American Psychological Association – also: citation style developed by APA
AUSSDA	Austrian Social Science Data Archive
CESSDA	Consortium of European Social Science Data Archives
DANS	Data Archiving and Networking Services – Dutch national centre of expertise and repository for research data
doi	Digital object identifier
ERIC	European Research Infrastructure Consortium
ELSST	European Language Social Science Thesaurus
EURHISFIRM	Research infrastructure for long-term company-level data for Europe
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen eV (German Social Science Infrastructure Services) – Leibniz Institute for the Social Sciences
IT	Information Technology
KG	Knowledge Graph
ML	Machine Learning
SSCI	Social Science Citation Index
SSH	Social Sciences and Humanities
SSHOC	Social Sciences & Humanities Open Cloud
SWC	Semantic Web Company
UKDA	UK Data Archive
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WoS	Web of Science

Table of Contents

1. Introduction and purpose 6
2. Context and Content of T9.3 7
3. The Development of the pilot-KG in Electoral Studies 10
4. Insights obtained from the development of the Knowledge Graph 11
   4.1 Feasibility 12
   4.2 Quality of available Metadata 13
5. Conclusion 17
6. References 19

1.Introduction and purpose

Within the SSHOC project, the focus of WP9 is on Data Communities that are not directly represented in the Consortium as partners in the form of an ERIC or other infrastructural organisation. One of these Data Communities is the community of scholars in Electoral Studies, which is at the centre of Task T9.3 "Data Community Project: Electoral Studies & EURHISFIRM". The main aspiration of T9.3, as formulated at the inception of the SSHOC project is to generate a pilot of an Open Research Knowledge Graph in the field of Electoral Studies. Towards this aim, the following deliverables have been produced:

- D9.6 (*Demarcation Report*)[1] which defined the field and data community of Electoral Studies, and the scope of the Knowledge Graph to be developed.
- D9.7 (*Design of Knowledge Graph*)[2] which specified the aspired functionality of the Knowledge Graph and the most important aspects of its overall and technical design.
- D9.8 (*User Community Involvement Plan*)[3] which describes aspired ways of involving the user community in the development, testing and evaluation of the Knowledge Graph.
- D9.9 (*Delivery of Knowledge Graph and Election Studies Analytics Dashboard*)[4] which described the elaboration of the design into an operational prototype of the Knowledge Graph with detailed information about the ingestion of information about materials covered by it.
- D9.10 (*User Community Feedback Report*)[5] which reports on the first stages of testing of the prototype of the Knoweledge Graph.
- D9.3 (*Usability Evaluation Report*)[6] in which Section 3 reports on the usability of the current (alpha) version of the Knowledge Graph, its perceived usefulness for different audiences, and its

[1] See: van der Eijk, C. (2020). SSHOC D 9.6 *Demarcation Report of Electoral Studies User Community*. Zenodo. https://doi.org/10.5281/zenodo.3725823. [accessed 1 December 2021].

[2] See: van der Eijk C., Kritzinger S., Partheymüller J., Kaltenböck M., Ahmeti A., & Karampatakis S. (2020). SSHOC D9.7 Design and Planning of Knowledge Graph in Electoral Studies (v1.0). Zenodo. https://doi.org/10.5281/zenodo.3824266. Zenodo. https://doi.org/10.5281/zenodo.3824266. [accessed 1 December 2021].

[3] See: van der Eijk, C. (2020). D9.8 User-community involvement plan. Zenodo. https://doi.org/10.5281/zenodo.4558312. [accessed December 2021].

[4] See: Karampatakis, S., Ahmeti, A., Kaltenböck, M., Gründl, J., Partheymüller, J., & van der Eijk, C. (2021). SSHOC D9.9 Delivery of user-validated Knowledge Graph, and Election Studies Analytics Dashboard (V1.0). Zenodo. https://doi.org/10.5281/zenodo.4700170. [accessed December 2021].

[5] See: van der Eijk, C. & Partheymüller, J. (2021). D9.10 User community feedback and usage report (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5579217. [accessed December 2021].

[6] See: van der Eijk, C., Saji, A., Morales, L., Degl'Innocenti, E., Di Meo, C., & Coradeschi, F. (2021). D9.3 Usability evaluation report. Zenodo. https://doi.org/10.5281/zenodo.5783356 [accessed 1 December2021].

weaknesses that have to be addressed before a wider roll-out in the form of a stable beta-version can be undertaken.

Further activities on the Knowledge Graph, its tesing and evaluation, and its introduction in the user community will still be undertaken during the remaining months of the SSHOC project. The current report, which constitutes D9.11 (*Concluding Report on T9.3*) differs from all other deliverables and reports listed above by employing a less detailed and more general perspective on the experiences generated by the work done in T9.3. It discusses a set of issues that are likely to be of relevance also when considering the development of Knowledge Graphs in other, cognate fields of the social sciences and humanities.

The remainder of this report contains the following:

- A recapitulation of the context and content of T9.3, its aspirations and their background, and some of the conditions that affected the execution of the planned work (Section 2);
- A concise summary in Section 3 of the development of the pilot Knowledge Graph in Electoral Studies;
- In Section 4 a discussion of major insights obtained from the work with respect to T9.3. These include an assessment of the feasibility of developing a Knowledge Graph in a sub-disciplinary domain of the social sciences, and a discussion of issues relating to available metadata;
- Section 5 concludes this report.

Finally, before moving to the next sections, it needs to be emphasised that the current version of this report is not necessaily the final one, in view of the possibility that information will be added in the form of addenda during the remaining months of the SSHOC project, depending on the kinds of further insights that will be generated in that period (this also pertains to other reports emanating from T9.3). [7]

2. Context and Content of T9.3

Task 9.3 consists of a Data Community Project that aims to develop a pilot of a Knowledge Graph in the field of Electoral Studies. This aspiration is motivated by existing challenges that researchers in this field encounter when searching for relevant information. Such information is predominantly of two types: scholarly publications, and datasets that are (in principle) available for secondary analysis. While researchers can make use of a variety of invaluable and sophisticated tools that may help them to search for publications and datasets, they nevertheless experience that many of their search needs can only be addressed by a combination of manual effort, perseverance, and inventiveness. This is particularly the

[7] The new version will be published via Zenodo and connected via DOI.

case when searching across different kinds of information. As a case in point, it is relatively straightforward to use widely available tools such as Google Scholar, or the Social Science Citation Index (SSCI) from Web of Science (WoS) to find scholarly publications about, for example, 'political trust'. It is much less straightforward to search for publications that report on the relationship between political trust on the one hand and electoral participation on the other. And yet more problematic is a search for publications about this relationship that are based on data from a particular source (such as the Dutch National Election Study). Although it is quite feasible to formulate the required search string when using Boolean operators, the resulting finds are far from complete while also still needing a considerable amount of manual post-processing to weed out 'false positives'. When addressing this example from the perspective of the data that were used, one can make use of search tools pertaining to data archives. One could use, for example, the CESSDA Data Catalogue[8] to search for datasets that would allow analysing the relationship between political trust and electoral participation. But one quickly finds out that neither topic classifications nor keywords used to describe the contents of datasets are sufficiently specific to determine whether the Dutch national election studies contain items of the looked-for kind.[9] One could, of course, use different catalogues (from various data archives, such as GESIS, DANS, AUSSDA or UKDA) to find the data one is looking for. But each of these tools presents similar problems of insufficient granularity of search terms. Moreover, all of them mainly allow the user to find potentially useful data, but not whether or not these data have already been analysed from the perspective one is interested in, and whether that has been published. The purpose of this little example is not to critique all these tools for searching publications or data, as they are tremendously useful and researchers could not do well without them. Yet, they are no panacea; they usually require extensive pre-existing knowledge of existing publications and datasets, and they often give only partial answers that require extensive further manual refinement. Thus, what this example does highlight is that, in spite of the wide availability of various powerful search tools, searching across different kinds of information (here: scholarly publications and datasets, respectively) is far from straightforward. As a result, dissemination of scholarly work, and reuse of existing work (publications as well as data) is far from optimal.

This is where the potential attraction of a domain-specific Knowledge Graph (KG) comes in. As described in D9.7 (Design and Planning of Knowledge Graph in Electoral Studies)[10]

> "A KG is a model of a knowledge domain created by subject-matter experts with the help of intelligent ML [i.e., Machine Learning – CvdE] algorithms. It provides a structure and common

[8] See: https://datacatalogue.cessda.eu/. [accessed December 2021].

[9] The substantive terms used to describe the datasets in the CESSDA Data Catalogue are derived from the European Language Social Science Thesaurus –ELSST—see https://thesauri.cessda.eu/en/ [accessed 1 December 2021] and from the CESSDA Topic Classification (see: https://vocabularies.cessda.eu/vocabulary/TopicClassification?lang=en) [accessed December 2021]. Neither of these contain 'political trust', but instead mainly broader terms, or occasionally narrower ones. Similarly, 'electoral participation' is not available as a search term, only the wider term 'voting behaviour' which does not distinguish between electoral participation and vote choice.

[10] See: https://doi.org/10.5281/zenodo.3824266. [accessed December 2021].

> 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." (Van der Eijk et al. 2020, p.14)

and

> "domain-specific KGs . . . help . . . 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" (Van der Eijk et al. 2020, p.15)

The benefits of KGs can thus be summarised in terms of their potential to

- Combine and connect disparate data silos
- Bring together structured and unstructured data
- Find things faster
- Future-proof databases when built to be compliant with W3C standards[11]

What is yet unknown is what obstacles and challenges one encounters when attempting to construct such a domain-specific KG, and whether and how these can be overcome. Hence the decision to aim for a pilot-KG to assess the viability of a domain-specific KG rather than for a full implementation and its rollout. This task (known within the SSHOC project as T9.3) was taken up by a group consisting of, on the one hand, experts in KG development from the Semantic Web Company (SWC)[12] and, on the other hand, academic scholars in the field of electoral studies from the University of Nottingham and the University of Vienna.

In this Concluding Report of T9.3 it is necessary to acknowledge that the development of T9.3 has, in several ways, not followed the paths planned and described in D9.7 (*Design and Planning of Knowledge Graph in Electoral Studies*) and D9.8 (*User-community involvement plan*). To some extent this is because of evolving insights, as occurs in virtually every kind of research enterprise. But it was (and is continuing to be) the Covid-pandemic that has played havoc on how the work could be executed and on how it progressed. The pandemic severely undermined most of the anticipated opportunities to involve the user community via variety of academic and professional meetings, conferences, seminars, and similar events that normally occur regularly across the year, as these events were cancelled, postponed, or because they restricted opportunities for 'tagging on' the kind of activities we originally had hoped to conduct. Moreover, the pandemic prevented the regularly planned face-to-face meetings of the

[11] See: W3C standards website: https://www.w3.org/standards/. [accessed December 2021].

[12] See: https://semantic-web.com/. [accessed December 2021].

members of the KG-development group while the functionality and efficiency of these intended events cannot be fully compensated by the various possibilities of online video-conferencing. And finally, the pandemic generated extra complications for other professional activities of members of the group (such as teaching, research supervision and administrative duties) which impacted in turn on opportunities for their efficient interaction. All of this did not so much the affect the planned SSHOC-related workload as it did negatively impact the productivity of the time invested. As a result, this Concluding Report on T9.3 is not as complete as it otherwise would have been. The pilot-KG will still be adapted in various ways in the remaining months of the SSHOC project, and additional testing and user-evaluation will still take place as well. Insights to be derived from these activities in 2022 will in due course be linked to this report in the form of addenda.

3. The Development of the pilot-KG in Electoral Studies

The substantive domain of the pilot-KG was originally specified as 'Electoral Studies'. However, for the purpose of the development of a pilot, this domain is too broad and heterogeneous to be practical –see for a more detailed discussion of what all is contained in the field of Electoral Studies the report of D9.6 (*Demarcation Report of Electoral Studies User Community*), and Section 2 of the report of D9.7 (*Design and Planning of Knowledge Graph in Electoral Studies*).[13] Therefore, the domain to be covered by the pilot-KG was specified narrower as 'empirically oriented studies of citizen/voter electoral participation in elections for public office in democratic societies'. However, this narrowing down does not render the pilot-KG irrelevant to the user community involved, on the contrary. It 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. In short, the narrowed down substantive domain of this pilot-KG will, in and of itself, be highly relevant for very large segments of the total community of scholars, students and professionals in the field of electoral studies.

For the pilot-KG an ontology was developed that has been documented in D9.9 (*Delivery of user-validated Knowledge Graph, and Election Studies Analytics Dashboard);* Figure 2 (p.10) and Appendix A (pp.31-37).[14] It consists of 23 classes, 12 relations and 14 attributes. The ontology uses partly existing taxonomies and a

[13] See: https://doi.org/10.5281/zenodo.3725823. [accessed December 2021] and https://doi.org/10.5281/zenodo.3824266. [accessed December 2021].
[14] See: https://doi.org/10.5281/zenodo.4700170. [accessed December 2021]

few newly constructed ones. The existing taxonomies include the CESSDA Topic Classification Scheme and the European Language Social Science Thesaurus (ELSST), as well as an existing taxonomy of countries. Newly constructed taxonomies relate to Research Methods and to Concepts and Theories. It is foreseen that these new taxonomies will be evaluated in greater detail and developed further in the remaining period of the SSHOC project (as well as thereafter). Datasets to be covered by the pilot-KG were ingested using the Datacite source of metadata.[15] Scholarly publications to be covered were derived in a first step from a set of extensive meta-analyses, followed by utilizing Scopus, and, in a further step, Microsoft Academic. In a final stage additional use was made of Crossref. Details of the procedure and of the specification of the respective queries are provided in Sections 3.2 and 3.3 of D9.9 (see footnote 13).

The implementation of the pilot-KG was constructed by the SWC team on the PoolParty platform, which provides a intuitively easy end-user interface.

The alpha-version of the pilot-KG, as delivered in the spring of 2021 covers 3865 scholarly publications, and 1081 datasets. It is envisaged that further ingestion of relevant materials will occur in the spring of 2022. On the basis of recommendations by a group of domain experts who performed a first set of end-user functionality tests,[16] the pliot-KG has been adapted before having been subjected to further testing.[17] As emphasised earlier in this report, further testing will take place in the remaining period of the SSHOC project, with –as much as possible– subsequent adaptations of the pilot-KG to accommodate experiences and feedback.

4. Insights obtained from the development of the Knowledge Graph

The development of the pilot of a Knowledge Graph in Electoral Studies has resulted not only in a set of submitted deliverables, one of which is the alpha version of the pilot Knowledge Graph (see https://sshoc.poolparty.biz/GraphSearch/), but also in some more general insights that are of relevance for assessing this data community pilot. They are potentially also to be taken into account when one

[15] Datacite: https://datacite.org/. [accessed 1 December 2021]

[16] Reported in D9.10 (*User community feedback and usage report*, v1.0). Zenodo. https://doi.org/10.5281/zenodo.5579217. [accessed 1 December 2021]

[17] Reported in Section 3 of D9.3 (*Usability Evaluation Report*); Zenodo. https://doi.org/10.5281/zenodo.5783356 [accessed 1 December2021].

would consider developing a knowledge graph in other substantive domains of the social sciences and humanities. These insights are discussed below and pertain to

- The feasibility of developing a knowledge graph for a sub-disciplinary domain;
- Issues with respect to available metadata: their domain-specific relevance, their quality, and the consequences of data citation practices for what can be derived from available metadata.

4.1 Feasibility

The experiences of T9.3 have demonstrated that developing a useful knowledge graph in a substantive (sub)domain of the social sciences and humanities is certainly feasible. The fact that an alpha version of the pilot-KG has been produced is in itself not relevant for assessing feasibility, because the technical expertise to do so is sufficiently widely available. Of much greater importance are two other experiences. Firstly, it could be shown that that this KG is able to address realistic substantive queries of a kind that can currently not be addressed by existing tools and instruments for identifying and locating relevant web-based information –as documented in Section 5, p.26 of D9.9 (*Delivery of User-Validated Knowledge Graph and Election Studies Analytics Dashboard*),[18] and experienced in a testing workshop of end-users. Secondly, the participants of this same workshop evaluated the usefulness and usability of the alpha-version of the pilot-KG in very promising terms –as documented in Section 3 of the report of D9.3 (*Usability Evaluation Report*).[19] Thus, the answer to the question whether a KG in a substantive domain such as Electoral Studies is not only technically possible, but also substantively useful, is certainly affirmative.

This general conclusion requires two accompanying remarks. First, a successful development of a KG requires (at least) three kinds of expertise to be combined in the project group. In no particular order these are (a) deep- and broad-ranging expertise of the substantive field in question; (b) extensive expertise in KG-methodology, relevant IT platforms, and computer science; and (c) extensive knowledge of existing and domain-relevant taxonomies, ontologies, and of relevant web-based meta-data. These different competencies do rarely coexist in single persons, and it is necessary that specialists in these several areas do interact closely, intensively and frequently in order for these different forms of expertise to be productively combined.

Second, the conclusion that a useful domain-specific KG is feasible in a specialised field of social sciences and humanities should acknowledge some of the unavoidable limitations of such a KG. These limitations have at least two distinct, yet related origins. The first set of limitations of any KG derives from choices made by its developers with respect to design and implementation. In the case of the pilot-KG in Electoral

[18] See https://doi.org/10.5281/zenodo.4700170. [accessed 1 December 2021].

[19] Reported in Section 3 of D9.3 (*Usability Evaluation Report*); enodo. https://doi.org/10.5281/zenodo.5783356 [accessed 1 December2021].

Studies, for example, the design choice was made to narrow the scope of the domain to be covered from 'Electoral Studies' to 'empirical studies of electoral participation by citizens in elections for public office in democratic societies'. In the implementation stage of the project choices had to be made about sources of metadata. Sources differ between them in terms of coverage of relevant material, as well as qua richness of metadata, two aspects that are often inversely related. Thus, the choice made in the pilot-KG of Electoral Studies for using Datacite (rather than the CESSDA Data Catalogue)[20] for populating the KG with information about datasets prioritised coverage of national election studies programmes over richness and detail of metadata. Such choices of design and implementation are unavoidable in any other substantive domain and result in limitations of the eventual KG.[21] In short: choices that have to be made in design and implementation imply limitations of the resulting KG in terms of breadth, depth, and emphasis.

The second set of limitations of any KG does not originate in choices by developers, but from the web-based environment on which it rests. Particularly in the case of KGs of substantive academic domains this environment is not under the control of developers. Limitations deriving from this constitute an important insight that needs to be discussed separately, in the next subsection.

4.2 Quality of available Metadata

Knowledge Graphs require metadata about materials of interest in a specified domain. Metadata can, in principle, be developed from scratch, but it is obviously more efficient to make use of existing metadata. Scholarly publications and professionally curated datasets have been subject to multiple efforts to document them with structured metadata and it stands therefore to reason that any attempt to develop a domain-specific KG that intends to cover such publications and datasets builds upon available sources of metadata. This brings three potential problems:

- The domain-specific relevance of available metadata
- The quality of available metadata
- Data Citation practice

Domain-specific Relevance

The first potential problem inherent in the use of externally harvested metadata is their relevance for the specific substantive domain to be covered by the KG in question. This is not a concern for technical and administrative elements of metadata, such as the doi, authors, date of publication or an URL. But it is a

[20] For Datacite see: https://datacite.org/ [accessed December 2021]; for CESSDA Data Catalogue see: https://datacatalogue.cessda.eu/. [accessed December 2021].

[21] These kind of choices between, e.g., sources of metadata are not always necessary in the long run, as metadata from sources that were not used initially can used in later stages to enrich the KG. Whether such opportunities will actually be used depends on resources and priorities. But in first instance, at the development stage of a KG, such choices are necessary to reduce complexity and enhance practicability.

concern for metadata that aim to characterise the substantive nature of a publication or a dataset. For such aspects of metadata, the issue is either one of granularity, or one of absence of standardisation. Many sources provide metadata at too coarse a level of granularity, as is the case for 'subject_areas' in Crossref and Scopus, or in the CESSDA Topic Classification used to describe datasets. These classifications are invaluable for distinguishing between broad disciplinary categories, such as demography, economics or politics. They may even allow, by way of subcategories to identify publications or datasets that are of likely relevance for the field of Electoral Studies. But they are not suitable when one is interested in concepts and terms that are indispensable to distinguish between studies *within* the field of electoral studies; concepts and terms such as 'party identification' (also known as 'partisanship'), or 'valence issue', 'perceptions of parties ideological or issue positions' or 'left-right'. This is not a shortcoming of these sources of metadata that were not intended to serve such detailed sub-disciplinary interests. But when trying to develop a KG in a sub-disciplinary domain such as Electoral Studies, it becomes abundantly clear that much of metadata intended to characterise the substantive character of publications or datasets is of limited relevance.

This problem of too coarse a level of granularity does not apply to metadata that reflect a 'title', an 'abstract' and 'keywords'. These usually replicate what has been submitted by the original author of a publication or a dataset and can, therefore, be accepted as being of pertinent relevance to characterise the content of the material in question. But here one encounters lack of standardisation: authors have all their own unique way for describing their material in title and abstract, and for formulating and choosing keywords.[22] This means that the presence of a particular keyword (e.g., 'socialisation') is more informative than its absence: when the keyword is used the material in question has some relevance for end-users interested in socialisation, but when that keyword is not used it cannot be inferred that the publication ort dataset has no relevance for such end-users. In the long run this problem can be ameliorated by ML approaches and corpus analysis (preferably of full texts),[23] but this is not a realistic option in the development stage of a KG.

Quality of Metadata

A second set of potential problems of externally harvested metadata pertains to their quality, which is used here as an overarching term for such matters as coverage, completeness, accuracy, consistency, timeliness, and so forth.

[22] An additional problem with keywords is that many scholarly journals allow only a very limited set of keywords (often 5 to 7) to be specified, which is often too little to adequately and exhaustively characterise a publication. The work on the pilot-KG has revealed as an unintended by-product, the absence of any kind of common practice in authors' selection and formulation of keywords for scholarly publications. The development of a KG would be supported immensely by the equivalent of the widely used subject descriptors for economic literature (see https://www.aeaweb.org/econlit/jelCodes.php. [accessed 1 December 2021].

[23] See the discussion about such possibilities in Section 5.2.3 (pp. 31-33) of D9.7 (*Design and Planning of Knowledge Graph in Electoral Studies*), https://doi.org/10.5281/zenodo.3824266. [accessed 1 December 2021].

The previous section commented that it is often necessary to make a choice –at least in first instance-- between different sources of metadata that overlap with respect to the kinds of information they provide. Such choices often involve a trade-off between coverage and richness.[24] In a longer time perspective, the additional available metadata from sources with the more limited coverage can be used to augment a KG, but this can only be achieved in a next round of development, and only if resources allow so at later stages. Such augmentation or enrichment comes, obviously, with the additional disadvantage of leading to incompleteness of metadata (see also below, for a further discussion of completeness).

Other aspects of quality of metadata that have to be considered are their accuracy and completeness, neither of which can be taken for granted. Following up on multiple similar studies, Van Eck and Waltman (2019), for example, documented that citation data in several highly regarded sources of metadata are not free from errors. Obviously, there is no reason to assume that such errors are limited to citation data; rather they can be expected to occur in all kinds of metadata.

Completeness (or rather lack of completeness) is yet another kind of problem one encounters when using metadata in the development of a KG. Lack of completeness is the equivalent of the 'item-wise missing data' data problem in virtually all empirical research projects.[25] Such missingness can present itself on a massive scale in many sources of metadata. As an example, when considering using Crossref –one of the largest and most reputed sources of scholarly metadata– one might be pleased to know that it contains, among other kinds of metadata, also the abstract of the publication or dataset in question. However, one might subsequently be somewhat disappointed to find that only in about 20 percent of instances this information is actually available (for 2019, marginally higher than for earlier years), while it is missing in all other instances.[26] Missingness varies across metadata types, but levels in excess of 50% are far from exceptional. It is likely to vary also across sources of metadata, but a cursory check of studies into completeness/missingness suggests that Crossref does not really stand out in a negative way when compared to others.

These comments about metadata quality are not meant to be unduly critical of the producers of metadata, but lead to two important conclusions. The first is that developers of KGs cannot naively count on the availability of various kinds of metadata, but they have to include in their planning a phase of systematic comparison of the strengths and weaknesses of available sources of metadata, taking into account such aspects as coverage of the desired domain, richness, accuracy, and completeness. The second conclusion that can be drawn from this concise discussion about aspects of metadata quality is about the expectations that end-users may validly entertain when employing Knowledge Graphs. Knowledge Graphs of scientific (sub)domains are not and cannot be a complete solution to the information needs of researchers, professionals and others. KGs can be regarded sensibly as valuable

[24] Many studies of the quality of metadata illustrate the presence of such trade-offs between coverage and richness –e.g., García-Pérez (2010).

[25] Coverage (or lack of coverage) also relates to missing information, but then on a unit-basis. Completeness refers to missing metadata information about some, but not all, aspects of a publication or dataset.

[26] Information derived from Hendricks et al. (2020).

additions to a wider repertoire of tools. They can, potentially, illuminate relations between various kinds of information (such as, for example, publications and datasets, respectively). But particularly in view of high rates of missingness for some types of metadata, they cannot be expected to exhaustively reveal such relationships.

Data Citation Practice

The third problem that was experienced in the development of the Knowledge Graph for Electoral Studies relates to data citation. One of the most important types of metadata consists of citations. Citations to publications are well-established and existing metadata sources can handle the wide array of citation styles quite well (e.g., Harvard, Chicago, APA, etc.). Matters are somewhat less well-established when it comes to data citation, i.e., the reference to data that have been used in a study. Partly this reflects different traditions of data citation and data sharing in the social sciences and humanities, as documented exhaustively in the report of SSHOC D3.2 (*Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning*),[27] and partly it reflects absence of established routines and procedures for data citation among the diverse actors involved (data creators, data users, publishers, journal editors, etc.). As a consequence, "citations tend to be more free-form text strings and thus not readily machine-actionable" (Larousse et al., 2019, p.11). The disadvantages of this practice of 'free-form text strings' complicated the development of a Knowledge Graph in Electoral Studies, where one of the aspirations involved connecting publications (scholarly journal articles, books, etc.) with the particular datasets used in the analyses reported in these publications. Where such connections could be made an additional complication was that any available citations to datasets tend to include a 'date of publication', which for a dataset is usually the date on which it is deposited or included in a data archive or repository. However, in the field of electoral studies another date is usually of at least of equal, if not of greater relevance, namely the date of the election to which the data pertain. For example, for a 'national election study [in country x]' it is more important to know that it relates to the election of, for example, 2017, than that it was published in 2019, particularly if 2019 would have seen a subsequent election in country x. It would be helpful if in the development of machine actionable metadata standards for data citation not only the 'year of publication' were included, but also the year of origin of the data (in the example just given this would be the year of the election in question).

Within WP3 of SSHOC innovative work is done to improve data citation practice by providing practical guidelines, particularly directed at data creators and disseminators.[28] It would be important to extend such guidelines to the end-users of research data, who are often the authors of publications in the form of scholarly articles and books. Additionally, it would be of importance that publishers and journal editors would actively require (and enforce) adherence to such data citation guidelines.[29] Better practice of data

[27] Available at: https://zenodo.org/record/3595965#.YbdkAZHP3Vi. [accessed December 2021].

[28] For example, in the form of the Webinar on 'FAIR SSH Data citation: a practical guide' on 3 December 2021. [materials of this webinar not yet web-based but will be so in the near future]. See also Larrousse and Gray (2021).

[29] Many scholarly journals in the social sciences and humanities are rather vague in their requirements for data citation and, judging from their recent articles, they do not systematically enforce even their own vague guidelines.

citation would be of immeasurable importance for the development of Knowledge Graphs for sub-disciplinary domains, of which the T9.3 example of a pilot-KG in Electoral Studies is just one example.

5. Conclusion

This report presented an overview of the work done in Task 9.3 of SSHOC, the remit of which is to construct a pilot Knowledge Graph in Electoral Studies as a Data Community Project for the scholarly community in the field of electoral studies.

In sections 2 and 3 this report recapitulated the aspirations and context of the project, and the form of its implementation.

Section 4 addressed the question of wider insights that were engendered by the project, and that could be of potential relevance in other contexts where the development of a knowledge graph for a sub-siciplinary domain is contemplated. The most important of these insights can be summarised as follows:

- The development of a useful knowledge graph for a scholarly sub-domain of the social sciences or humanities is feasible as demonstrated by (a) the experience that the pilot-KG in Electoral Studies was shown to be capable to address queries that cannot be addressed by existing tools and instruments for identifying and locating relevant web-based information, and (b) evaluations of usefulness and usability by end-users in a testing workshop.

- Successful development of a KG requires (at least) three kinds of expertise: (a) deep- and broad-ranging expertise of the substantive field of the KG; (b) extensive expertise in KG-methodology, relevant IT platforms, and computer science; and (c) extensive knowledge of existing and domain-relevant taxonomies, ontologies, and of relevant sources of web-based meta-data. Frequent and intensive interaction between these different kinds of experts is essential.

- Available metadata for the wider SSH domain are often of a granularity that is too coarse for sub-disciplinary domains. This problem is not much ameliorated by metadata on author-assigned keywords, because of the usually too small number of keywords allowed by scholarly journals and because of the absence of agreed-upon vocabularies for the sub-disciplinary domain of electoral studies (condition that are, in all likelihood, the same for many other sub-disciplinary domains in SSH). Data communities could improve these conditions by focussing on the development of sub-disciplinary controlled vocabularies.

- Given endemic quality issues of metadata, developers of KGs cannot naively count on the availability of various kinds of metadata, but they have to include in their planning a phase of

systematic comparison of the strengths and weaknesses of available sources of metadata, taking into account such aspects as coverage of the desired domain, richness, accuracy, and completeness.

- End-users of KGs must be aware that KGs are not and cannot be the definitive answer to all their information needs. Particularly in view of high rates of missingness for some types of metadata, KGs cannot be expected to exhaustively identify instances of the desired kinds of information.

- Poor data citation practice in the SSH domain complicates the realisation of the potential of KGs to connect information from scholarly publications on the one hand, and datasets on the other. Improvements in this respect can be actively driven by creators/disseminators of datasets, as well as by publishers and journal editors who are in a position to require authors of scholarly publications that are based on datasets, to adhere to adequate data citation standards.

6. References

García-Pérez, M.A., (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. Journal of the American society for information science and technology, 61(10), pp.2070-2085.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022. [accessed 1 December2021].

Karampatakis, S., Ahmeti, A., Kaltenböck, M., Gründl, J., Partheymüller, J., & van der Eijk, C. (2021). SSHOC D9.9 *Delivery of user-validated Knowledge Graph, and Election Studies Analytics Dashboard* (V1.0). Zenodo. https://doi.org/10.5281/zenodo.4700170. [accessed 1 December2021].

Larrousse, N., Broeder, D., Brase, J., Concordia, C., & Kalaitzi, V. (2019). SSHOC D3.2 *Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning* (v1.0). Zenodo. https://doi.org/10.5281/zenodo.3595965. [accessed 1 December2021].

Larrousse, N., Gray, E.J. (2021). *Recommendations for FAIR Data Citations in the Social Sciences and Humanities*. Zenodo. https://doi.org/10.5281/zenodo.5361718. [accessed 1 December2021].

van der Eijk, C. (2020). SSHOC D 9.6 *Demarcation Report of Electoral Studies User Community*. Zenodo. https://doi.org/10.5281/zenodo.3725823. [accessed 1 December2021].

van der Eijk, C. (2020). D9.8 *User-community involvement plan*. Zenodo. https://doi.org/10.5281/zenodo.4558312. [accessed 1 December2021].

van der Eijk, C., Saji, A., Morales, L., Degl'Innocenti, E., Di Meo, C., & Coradeschi, F. (2021). D9.3 *Usability evaluation report*. Zenodo. https://doi.org/10.5281/zenodo.5783356 [accessed 17 December2021].

van der Eijk C., Kritzinger S., Partheymüller J., Kaltenböck M., Ahmeti A., & Karampatakis S. (2020). SSHOC D9.7 *Design and Planning of Knowledge Graph in Electoral Studies* (v1.0). Zenodo. https://doi.org/10.5281/zenodo.3824266. [accessed 1 December2021].

van der Eijk, C. & Partheymüller, J. (2021). D9.10 *User community feedback and usage report* (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5579217. [accessed 1 December2021].

van Eck, N.J. and Waltman, L. (2019). Accuracy of citation data in Web of Science and Scopus. *arXiv preprint* arXiv:1906.07011. [accessed 1 December2021].