Towards Integrating the Research Data Life Cycle of the Social Sciences based on Semantic Technology

Dennis Wegener, Erdal Baran, Wolfgang Zenk-Möltgen, Benjamin Zapilko

GESIS - Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne
{dennis.wegener, erdal.baran, wolfgang.zenk-moeltgen, benjamin.zapilko}@gesis.org

Abstract: Social science today comprises collaborative research, which is more and more carried out on the basis of comprehensive digital infrastructures. These infrastructures aim at integrating all relevant resources for the research and providing tools to use these resources. In this paper, we present the research data life cycle in the social sciences and its challenges. We discuss prior work on integrating the parts of the research data life cycle and identify open issues regarding infrastructure, linking, and privacy. Based on that, we present an outlook on how these problems could be addressed by using semantic technology.

1 Introduction

Research in the social sciences today can be illustrated based on research data life cycles comprising different phases performed by researchers. It is more and more carried out on the basis of comprehensive digital infrastructures supporting the different phases of research data life cycles (RDLC). Lots of models have been produced to capture the nature of RDLCs. [WGISS11] describes 44 models from different domains and with different aims, with some of them also covering the social sciences. Nearly all models show the iterative nature of research, either by using the term “cycle” or stating that the stages of the model may be repeated. Most of the RDLC models contain stages that are generic to any empirical science, e.g. data collection, data management, or data archiving. Some models cover specific aims like supporting data preservation or helping in the discovery of data. Most, but not all, also cover activities like data analysis or publication. In general, life cycle models are used to identify the processes in conducting research at an institution, and to further define the tasks that have to be done at each of the stages. Similar to the field of e-Science, infrastructures supporting RDLCs aim at integrating all relevant resources for the research and providing tools to use these resources. There exist a lot of repositories, libraries, tools, and services that support different phases, tasks or processes in the RDLC. However, current infrastructures and processes are not fully integrated, which hinders the discovery, reuse and publication of data. Although there is a lot of information and data available, digital libraries and repositories are integrated or cross-linked only in few cases. Semantic technology could potentially play a role in resolving these issues. The potential of using semantic technology for the social sciences has been already identified in [Gr10].

In the following, we present the GESIS research data life cycle in Section 2. In Section 3 we present and discuss prior work in the context of semantic technology. Afterwards, Section 4 presents concrete issues regarding the research data life cycle and semantic technology that remain open. Finally, Section 5 concludes.

2 The GESIS research data life cycle

Different to some of the sophisticated models of the data life cycle, GESIS has created a research data life cycle that tries to organise the products and services of GESIS around five simple stages [GESIS2013]. The phases of the GESIS model are seen from the researchers perspective and describe activities which deal with research data (Figure 1). Each individual phase of this cycle requires specific know-how to obtain significant results. The GESIS research data cycle provides a holistic view of the production, preparation, acquisition and usage of research data and the special information associated to each of the stages. The model should provide social science researchers with intuitive, comprehensible access to the products and services of the institute.

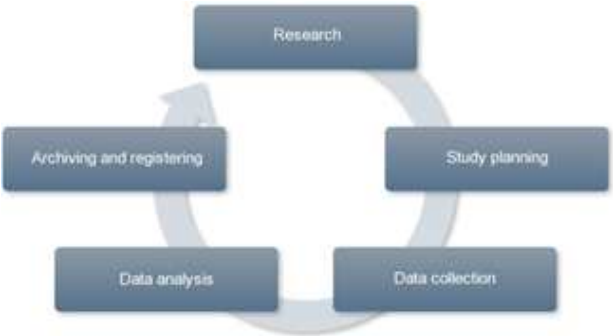

Figure 1: The GESIS research data life cycle [GESIS2013]

The GESIS model starts with the phase "Research". Here researchers are supported in finding information, in particular by offering extensive databases with information about publications and research data. This helps to identify important information related to the intended research question. The next phase is the "Study planning", in which researchers can use services for the planning and methods for the survey design of their studies, e.g. sampling methods. After that, researchers may find information about collecting and preparing research data according to state-of-the-art methodologies in the phase "Data collection". This includes consultation services about item wording as well as about survey mode and using demographic standard indicators. Then, the phase "Data analysis" provides tools and information to analyse research data together with methods for that. Finally, the phase "Archiving and registering" covers the registration of research data with persistent identifiers and the archiving and distribution of them, including long-term preservation. Even if these five phases are simpler than most of the models, they map to the DDI lifecyle model, which is currently the most widely known model in the social science archival community [DDI11]. For each of the phases, different tools and services are offered to social scientists. Among others, GESIS provides the portals

Sowiport and da|ra for the research phase, consultation for sampling or for conduction of online surveys for the planning phase, the electronic handbooks ZIS and EHES with survey instruments for the data collection phase, the system for computer-aided quantitative content analysis TEXTPACK for the data analysis phase, and the metadata management system DBKfree for the archiving and registering phase [GESIS2013].

Social science researchers would, e.g., follow the RDLC as follows: First, information about prior research is needed and the relevant literature may be searched at sowiport and the datasets at da|ra. Second, the question arises which study design may be suited, and the study descriptions and method reports from the Data Catalogue DBK may be consulted to evaluate possible datasets [Ze12]. Third, data collection methods for new or of existing data have to be investigated. Fourth, different available data analysis routines and rules have to be evaluated. For both phases, GESIS does offer consultation services. Fifth, after performing the study, the results must be published, and the data and publications should be archived and registered with a persistent identifier, which can be done by the GESIS services SSOAR, the data archiving into DBK, and da|ra.

In the past, there were no problems when following the RDLC when the research was mainly carried out by single researchers who used only their own collected data. But already when data archiving in the social sciences began, the need of good documentation was well known [Bi67]. Today, research in the social science is highly collaborative, which means that several people from different organizations have to cooperate within a project. This results in demands for collaborative research environments and sharing of data, processes and results. In addition, the variety of tools, services and processes for the different phases of the research data cycle makes it hard to reuse and integrate these. Metadata standards like DDI that cover the whole RDLC are not yet implemented widely enough. The end user is mainly interested in the following things in order to enhance his research process: 1) Inclusion of more and broad data for research, which could be addressed by enhancing the findability and interlinking of existing data (research phase); 2) Faster generation of data due to data reuse, e.g. supported by phase/process/tool integration and transition (data collection phase); 3) Generation of high-quality data, which could be addressed by metadata enhancement (like the Enhanced Publications [Surf13]) and by ensuring that all important concepts for future research are collected (study planning phase); 4) Easy archiving and registration, including unique identification and citability to increase visibility and reuse, which could be supported by standard processes (archiving and registering phase).

3 Prior work in the context of semantic technology

[Go11] presents a study on requirements for building a semantic data library for the social sciences. In the study, a set of open issues were identified, which hinder a reuse of available and valuable data resources. These issues include lacks in data citation, a variety of data formats, the absence of RDF-based vocabularies for data description, missing integration of existing decentralized data repositories, lack of metadata annotation, and the incapability to gain deep insights to both data and documentation for judging the relevance and quality of data.

Recently, several approaches have been proposed that address some of these issues. E.g., the da|ra system [DBW13] allows for the registration of datasets to get a global unique persistent identifier (DOI). A large set of metadata can be documented with the datasets, and these can be searched in the da|ra system [HQJSZ13]. Via the DOIs the original data at the providing organizations that registered the data can be retrieved. The da|ra system offers a web-portal for form-based registration of datasets and a web service for the registration, so that institutions can integrate their archiving with da|ra, like done with the GESIS Data Catalogue DBK [Ze12]. In addition, efforts on the development of the DDI-RDF Discovery Vocabulary have been started, which aim at supporting the discovery of microdatasets and related metadata using semantic technologies [BCGW13]. The vocabulary is a subset of the DDI standard data format for data documentation in the social, behavioural and economic sciences and allows the representation of DDI metadata as Linked Data for dissemination on the Web. Thus, it covers the major concepts of the DDI model like Study, StudyGroup, Variable, Question, Questionnaire, AnalysisUnit, Universe, and LogicalDataSet. By simplifying the complex DDI model it is intended to increase the applicability of the vocabulary.

In 2011, a Linked Open Data (LOD) pilot application for the Social Sciences has been developed [Go11]. The motivation was to implement an initial use case relevant to Social Science research (data analysis phase) and to investigate further areas of research utilising state-of-the-art technologies. Hence, the prototype aims to expose, integrate, aggregate and visualize data as Linked Data from two sources: 1) survey data of individuals rating personal and national economic situation from the ALLBUS and 2) election statistics from the German federal state North Rhine-Westphalia from IT.NRW. While the prototype provides central services for processing the data sources, their physical storage location remains distributed. Wrappers are used to generate RDF data on-the-fly from the original data sources. By using OpenRDF's Sesame library, all data is loaded into a memory store and can be accessed integrated via a SPARQL endpoint.

4 Open issues regarding the research data life cycle and semantic technology

Some challenges and issues were already addressed by the approaches described in the previous section. In the following, we present some issues that still remain open, but could be addressed by semantic technologies. We focus on issues related to infrastructure, linking, and privacy, as it is important to integrate the different and partially privacy-sensitive data collections at GESIS inside the research data life cycle.

In [HB11] six recipes for publishing Linked Data are defined, which consider different infrastructure situations like using custom server-side scripts, defining mappings from relational databases or wrapping applications or web APIs. Considering the historically grown distributed infrastructure for publishing data at GESIS none of the recipes fits to all currently existing situations. Data collections are not only distributed, but also published technically different. Since it is intended to integrate and link data between different data collections, the challenge for a suitable infrastructure increases. Data sets of different data collections may complement each other, e.g. publications may analyze

and discuss results from a survey. Also, there are entities, e.g. persons, organizations, which are part in several data collections. They can be unified as central entities, to which particular data sets from different data collections can link to. In addition, external data sources from the Web can provide relevant information regarding to users' research interests. Current (semi-)automatic approaches can be used for this task. Most of them are applying standard similarity-based or machine-learning techniques. In order to reduce manual efforts it is necessary to investigate candiate entities first, which can serve as linking entities for meaningful and relevant links for users. To enable an integrated search over the distributed, but interlinked data collections it is necessary to bring the Linked Data resources to standard frontends, e.g. to Solr search. There exist first approaches in this regard, e.g. in the VIVO project [VIVO2013], where RDF documents are converted into Solr documents for building an index.

Because LOD offers researchers a lot of related information about studies, authors, individuals, data collections etc., it can raise a privacy problem due to this data enrichment. The privacy of social science data has been already identified as open issue in [Go11]. During the archiving phase of the RDLC, several steps are conducted: acquisition, ingest, data processing, archival storage, and access [Ma12]. The ingest step includes checks of the data on technical correctness, documentation, and persistence. Since the archived studies at the GESIS Data Archive are primarily micro data, they are being investigated in terms of data protection and privacy aspects, too. Personal information in studies, such as zip code, profession, salary data etc., has to be defacto anonymized. Our goal is to optimize this manual and time-consuming process based on semantic web technologies by finding the combinations of attributes in the data that violate the privacy. Based on the identified bad combinations we also want to provide the experts possible solutions to prevent this privacy injury. Instead of focussing on new anonymization algorithms, we want to focus on the development of a concept for a semi-automated recommendation system for professional experts that makes the anonymization process easier and clearer. In the access step, data is provided for the future reuse by researchers. Before that, the conditions under which the data can be made accessible for research have been decided upon with the data provider and are regulated by access categories[1]. However, all data that is provided via the standard distribution mechanisms must be defacto anonymized. Data that are privacy sensitive can only be provided under highly controlled regulations like in the recently established secure data center (SDC) [Ma12], so that there is a need for a privacy concept for these different data sets. Hence, a next step would be to analyze these data sets for appropriate solutions which could, e.g. be based on an authentication and authorization model supporting RDF based access control lists [HP09] or a policy-aware data access model for Linked Data [WSRH10].

5 Conclusion and outlook

In this paper, we have presented the GESIS research data life cycle and prior work on integrating this cycle utilizing semantic technologies. We have identified and discussed

[1] http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/#3_Access_categories

open issues that can be addressed by semantic technologies in the future. Altogether, a semantic technology driven infrastructure can enable a tight integration of the GESIS research data cycle. But, since existing semantic technologies can only serve as reference and best practices, the adaptation of these solutions have to be investigated. For GESIS this means that a general concept for LOD is needed, which focuses on data standardization based on DDI and improves the linking of social science research datasets to other Linked Data. This will enhance data discovery and standardization of metadata. In addition, we have to identify whether besides anonymization and access control additional mechanisms are needed for assuring the privacy of the data involved.

References

[BCGW13] Bosch, T.; Cyganiak, R.; Gregory, A.; Wackerow, A. (2013): DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. Accepted for publication at Linked Data on the Web Workshop@WWW2013.

[Bi67] Bisco, Ralph L. "Social science data archives: progress and prospects." Social Science Information 6.1 (1967): 39-74.

[DDI11] DDI Alliance Web Site: Online: http://www.ddialliance.org/ [accessed on 2013-03-25]

[DBW13] Dimitrov, D.; Baran, E.; Wegener, D.: Making Data Citable - A Web-based System for the Registration of Social and Economics Science Data. Accepted for publication at WEBIST 2013.

[GESIS2013] GESIS: Services for the Social Sciences. Online: http://www.gesis.org/en/services/ [accessed on 2013-03-25].

[Go11] Gottron, T.; Hachenberg, C.; Harth, A.; Zapilko, B. (2011): Towards a Semantic Data Library for the Social Sciences. In: Proc. of the Int. Workshop on Semantic Digital Archives, Berlin, Germany. CEUR Workshop Proceedings, vol. 801, Berlin, pp. 48-59

[Gr10] Gregory, A., Vardigan, M.: The Web of Linked Data. Realizing the Potential for the Social Sciences (2010), http://odaf.org/papers/201010_Gregory_Arofan_186.pdf

[HQJSZ13] Hausstein, B.; Quitzsch, N.; Jeude, K.; Schleinstein, N.; Zenk-Möltgen, W. (2013): da|ra Metadata Schema. Version 2.2.1. GESIS Technical Reports, 2013/03.

[HB11] Heath, T.; Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool.

[HP09] Hollenbach, J.; Presbrey, J. (2009): Using RDF Metadata to Enable Access Control on the Social Semantic Web Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, CK'09.

[Ma12] Mauer, R. (2012): Das GESIS Datenarchiv für Sozialwissenschaften. In: Altenhöner, Reinhard; Oellers, Claudia (Hrsg.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen, Berlin: Scivero, S. 197-215

[Surf13] Enhanced publications. [accessed on 2013-04-11] Online: http://www.surf.nl/en/themas/openonderzoek/verrijktepublicaties/pages/default.aspx

[VIVO2013] VIVO Project. Online: http://vivoweb.org/ [accessed on 2013-04-12]

[WSRH10] Wagner, A.; Speiser, S.; Raabe, O.; Harth, A. (2010): Linked Data for a Privacy-aware Smart Grid GI Jahrestagung (1), pp. 449-454

[WGISS11] Committee on Earth Observation Satellites - Working Group on Information Systems and Services (WGISS) (2011): Data Life Cycle Models and Concepts. Version 1.0.

[Ze12] Zenk-Möltgen, W. (2012). The metadata in the data catalogue DBK at the GESIS data archive. RatSWD Workshop "Metadata and Persistent Identifiers for Social and Economic Data“. Berlin, 07-05-2012 – 08-05-2012.