Modelling text-fact-integration in digital libraries

Stempfhuber, Maximilian; Zapilko, Benjamin

Modelling Text-Fact-Integration in Digital Libraries

Maximilian Stempfhuber, Benjamin Zapilko

GESIS – Leibniz Institute for the Social Sciences, Germany

max.stempfhuber@gesis.org, benjamin.zapilko@gesis.org

Abstract. Digital Libraries currently face the challenge of integrating many different types of research information (e.g. publications, primary data, expert's profiles, institutional profiles, project information etc.) according to their scientific users' needs. To date no general, integrated model for knowledge organization and retrieval in Digital Libraries exists. This causes the problem of structural and semantic heterogeneity due to the wide range of metadata standards, indexing vocabularies and indexing approaches used for different types of information. The research presented in this paper focuses on areas in which activities are being undertaken in the field of Digital Libraries in order to treat semantic interoperability problems. We present a model for the integrated retrieval of factual and textual data which combines multiple approaches to semantic interoperability und sets them into context. Embedded in the research cycle, traditional content indexing methods for publications meet the newer, but rarely used ontology-based approaches which seem to be better suited for representing complex information like the one contained in survey data. The benefits of our model are (1) easy re-use of available knowledge organisation systems and (2) reduced efforts for domain modelling with ontologies.

1 Introduction

During the last years, Digital Libraries for scientific users are undergoing a huge change according to their role and work (DELOS, 2005b; Gold, 2007). Results from several surveys (Poll, 2004) indicate that harvesting or linking up metadata from different sources and making them available for retrieval by applying only a minimum of standardization techniques on data and retrieval features does not suffice the information needs of users any more. Scientific users are expecting a tight integration of different types of information (full text, bibliographic references, surveys and other primary data, time-series data, project information, researchers' profiles etc.). This reflects their use of these types of information at different stages and in different combinations throughout the research cycle. At an early stage, for example, a scientist might search for publications and project information, whereas at later stages of his research he might be looking for research data to do secondary analysis or for conferences to present his results.

Especially in the social sciences, where at the one hand data archives which document empirical data at a very detailed level are organized at an international level and create dedicated entry points to their holdings, these information and infrastructures are at the other only minimally connected to the holdings of libraries and information centres. This not only challenges information providers in establishing and organizing collaboration with each other to bring together all resources, but also raises research questions on how to integrate research

information at the technical, structural and semantic level. The complexity involved in supporting the full life cycle of data including the accompanying documentation, i.e. different versions of questionnaires, the final data set of a survey, the accompanying codebook, sample frequency distributions and summary statistics for variables, creates domain-specific semantics which currently are not sufficiently matched to the semantic representations produced for e.g. research literature.

But the emerging paradigm of e-Science (Gold, 2007), understood as "enhanced" science, places the focus on creating an holistic infrastructure of hardware, software and (collaboration) networks to support advanced scientific activities which start with data acquisition and laboratory notes, lead to a new level of scientific publishing (e.g. electronic publishing, open access repositories), and at the same point make all research results available for retrieval by fellow researchers. Scientific models and methods are therefore needed to uniformly express the structure and semantics of all types of research information and to define matching and mapping processes to identify and link related information, both for documentation, retrieval, interpretation and re-use. They are also the basis for advanced features, like distributed computation, simulation and visualization of partitioned and heterogeneous data. These tasks can not only be found aside librarians, but also in the community of data scientists (ARL, 2006).

2 Current Research

International efforts are already taking place in several domains and organizations to deal with the current situation and to advance and overcome the described observations. Approaches can be found on the level of information architecture design and on formal organization of Digital Library Systems as well as on the structural, schematic level and the semantic level where dealing with the interoperability of heterogeneous, distributed data and document types is a major issue. Most of these approaches are only loosely connected with each other and stand for their own. In recent years the use of suitable Semantic Web concepts and technologies for Digital Libraries, e.g. the use of ontologies, semantic annotations or inference engines (Sure, Studer, 2005), is discussed in a greater community, especially to improve issues concerning social and knowledge networking and interoperability (Goble et al., 2006) as well as using and representing the semantics of metadata (Svensson, 2007).

The DELOS, Network of Excellence on Digital Libraries, delivers a state-of-the-art report (DELOS, 2005a) on semantic interoperability in Digital Library Systems which includes a broad range of research activities being undertaken. The following overview describes the current state of diverse activities in the field of semantic interoperability.

2.1 Architectural and Organizational Level

Formal models, e.g. the 5S model (Goncalves et al., 2004), and reference architectures (e.g. Candela et al., 2006) have been developed to organize and structure a Digital Library as a total package in order to bring them to a higher level of effectiveness and to meet users´ claims. These architectures cover nearly every aspect which is relevant for an efficient library in a network connected world, but because of their distance view on the whole package a lot of detailed problems occurring on underlying layers often remain unanalyzed or unsolved. Therefore a lot of activities can be detected which are dealing with specific concerns in different, smaller areas of a system. To recognize all entities which are relevant for documenting research activities and outcomes and to analyze relations between them, models

like the CERIF standard[1], the Common European Research Information Format developed by the European Commission and euroCRIS, or the PolicyGrid ontology (Chorley et al., 2006) have been developed which cover projects, institutes, publications, research facilities, patents etc. These models outline the complete research process.

In a global digital environment another major issue is the unique identification of resources and their long-term availability. Efforts are taking place in organizing long-term access to research information and in standardizing archival formats. With the development of Uniform Resource Identifiers (URI) it is possible to reference information sources like documents, data, persons etc., but URI can also provide a unique identification of individual concepts, terms and relationships of knowledge organization systems. For the approach of URI it is required that the referenced address stays the same for the rest of the life cycle of the referred resource. Digital Object Identifiers (DOI)[2] are an example for creating persistent identifiers which uniquely reference data sets or digital publications and which separate reference to this information from the place of actual storage (Paskin, 2008). For the long-term availability of documents DOI seems to be a effective solution.

2.2 Structural Level

At the structural level, community driven standards for documenting primary data (e.g. the DDI format of the Data Document Initiative[3] or SDMX[4]) are available, as are metadata standards for bibliographic references (e.g. MARC[5] or Dublin Core[6]). But up to now, these different communities are only loosely communicating, and standards focus mostly on (metadata) exchange (e.g. harvesting protocols, like OAI-PMH[7]). Metadata format registries document these formats and mappings between them, and formal methods for schema mapping can be used to at least map similar elements of these formats onto each other. The field of schema matching and metadata exchange is a well researched area; therefore this paper focuses the treatment of heterogeneity on the semantic level.

2.3 Semantic Level

Concerning the semantic level a broad amount of terminology mapping initiatives and projects can be found. There exist a wide range of different approaches on treating semantic heterogeneity (Zeng, Chan, 2004). A semantic network between different information sources, e.g. diverse collections of bibliographic references with contents that are indexed with different vocabularies or thesauri, can be created by mapping the terms of the used vocabularies on each other. This is an important issue when providing a cross-search over distributed and differently indexed information sources. Methods for mapping terminologies onto each other can reach from intellectual methods, e.g. cross-concordances (Mayr, Petras, 2008a), to automatic, statistical or deductive methods.

---

1 http://www.eurocris.org/cerif/introduction

2 http://www.doi.org

3 http://www.ddialliance.org

4 http://www.sdmx.org

5 http://www.loc.gov/marc

6 http://dublincore.org

7 http://www.openarchives.org

When dealing with different information types such as textual and factual data, treatment of semantic heterogeneity is much more complex as the different metadata schemas do use non-standard means of representing (semantic) content, and not all of the semantics inherent in the data are fully expressed. For primary data, like surveys or time-series data, different (types of) controlled vocabularies (e.g. nomenclatures and classifications) are used for content indexing, whereas thesauri are mostly used for indexing textual data (e.g. publications). Mapping these different vocabularies is rather difficult due to differences in expressiveness of semantic concepts and the relations used to express different types of linkage between these concepts within each vocabulary (e.g. broader terms, narrower terms, similarity etc.). Both approaches for content indexing are justified in their different usage contexts, but create mapping problems if used within one retrieval system: For primary data, the most relevant information – the scientific intention for phrasing a certain question – is only by occasion encoded in the question itself or the related variable and variable label (Krause, Stempfhuber, 2005). This information can not directly be mapped on adequate thesaurus entries, which could be used for retrieving related literature, and vice verse. In the context of the retrieval of literature, users normally can cope with certain amounts of imprecision and noise by scanning titles and abstracts of results, but in the case of data retrieval, where relevance of a study for re-use has to be judged at the level of a combination of variables, sampling method, size of sample, coding etc., a much higher precision is needed to satisfy the information needs of users.

Ontologies seem to be a suitable solution to index data with richer semantics. As ontologies are used in approaches on the architectural and organizational level it is quite obvious that they can be useful on the semantic level as well. With the development of SKOS[8] (Simple Knowledge Organisation System) the technical feasibility of making classic knowledge organization systems like thesauri accessible to the Semantic Web is available. How to link them effectively to ontologies for an integrated retrieval is still an open issue.

3 Model for Text-Fact-Integration in Digital Libraries

The following model describes the semantic integration of heterogeneous types of information in Digital Libraries. It focuses on treating semantic heterogeneity and is capable of solving some of the issues described above by taking existing approaches and developments and bringing them together in one broader scenario without loosing the focus on detail. It not only covers the semantics contained in different types of data (e.g. survey data or publications), but also includes semantics for linking the data with entities relevant to the overall research process. In specific, the model consists of 3 layers, each layer dealing with a dedicated semantic modelling problem (see figure 1).

8 http://www.w3.org/2004/02/skos

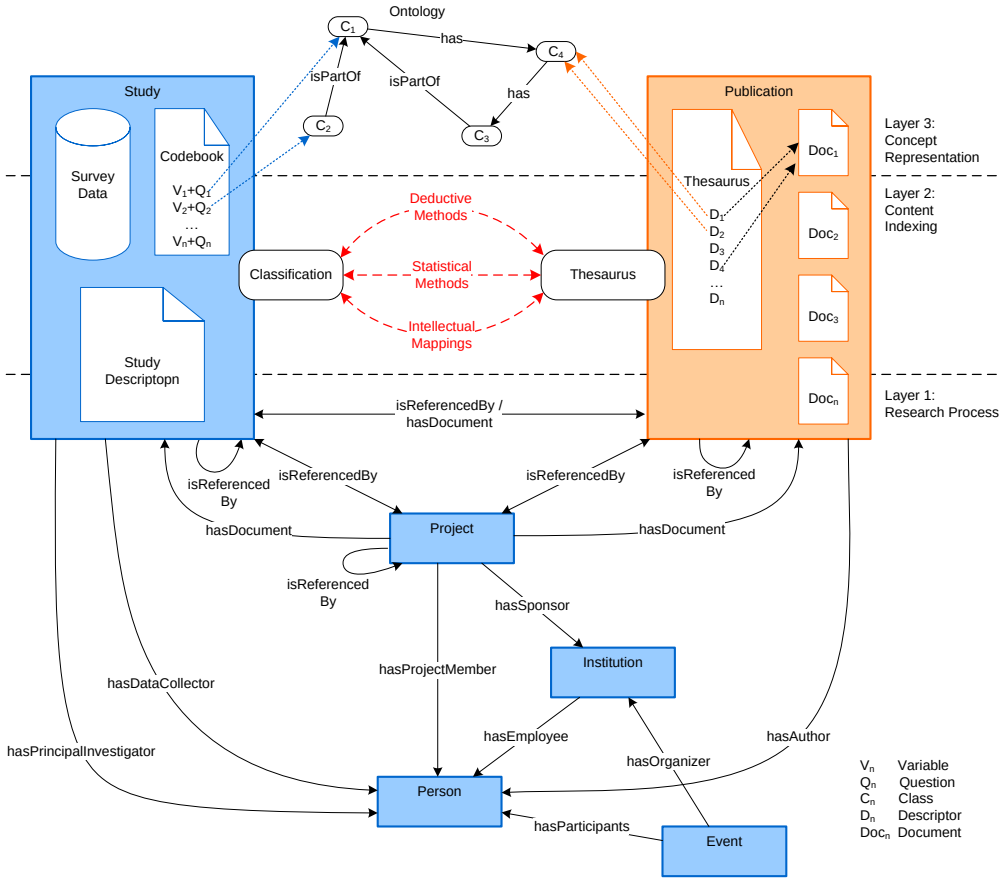

Figure 1. Full Model.

In the following paragraphs the three layers are described in detail and their linkage to the current research context is shown.

3.1 Layer 1: Research Process

This layer (see figure 2) reflects the complete research process and expresses relationships between all entities (e.g. persons, institutes, research programmes, projects, results, facilities, patents). Moreover, this layer represents the context in which research is carried out and in which research outcomes are produced. It is based on established models like the CERIF standard or the PolicyGrid ontology (Chorley et al., 2006). The relationships within this layer allow deductive processes within the realm of research, e.g. about authorship of results, linkage of results to projects, linkage of complementary projects to research programmes etc. They can be used for browsing related information and outline the core of a research information system on which the other layers (see below) are based.

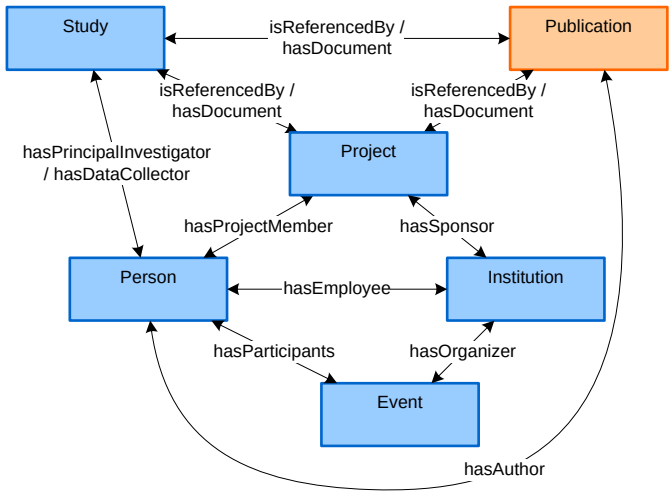

Figure 2. Layer for modeling the research process.

The semantics encoded on this layer help to reduce vagueness in the retrieval process as they provide background information to the first order objects normally retrieved by users in the context of digital libraries (e.g. literature or primary data). Not only do they provide unique and persistent identifiers for persons in different roles (e.g. author, researcher, project manager etc.), they also provide complementary information (e.g. about the strategic goals of funding programmes) which normally is not expressed at the level of single information entities and which can be used to support end user search strategies.

3.2 Layer 2: Content Indexing

This layer deals with the semantics expressed in the data itself or in the document surrogates (the accompanying metadata including content indexing with key words, notations from classifications etc.). It handles the heterogeneity between the indexing vocabularies used in different collections and for different types of information, e.g. classifications and nomenclatures for primary data and thesauri for publications (see figure 3) together with means of mapping these vocabularies onto each other.

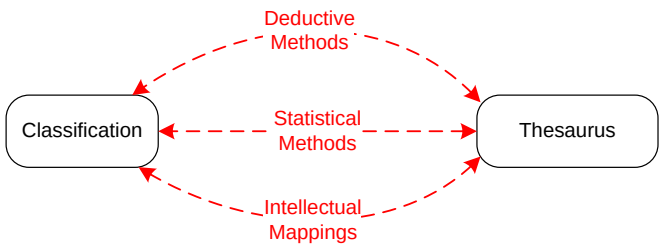

Figure 3. Layer for integrating the content indexing.

Approaches for dealing with the semantic heterogeneity between indexing vocabularies include intellectual mappings (bilateral concordances), statistical and deductive methods, which generally increase recall during information retrieval (Krause, 2004) and support users by automatically transforming queries for a specific type of information (e.g. publications) to other types of information (e.g. statistical data), therefore eliminating the need to learn new indexing vocabularies or re-formulating an information need several times and by using different vocabularies to find proper search terms. This automatic transfer can be provided as an automatic and transparent background service during query processing; the mappings

actually used during retrieval can be presented to the user for explanatory purposes and for further exploration of the result set.

3.3 Layer 3: Concept Representation

The topmost layer (see figure 4) handles specific differences in semantic expressiveness between thesauri, classifications, codebooks etc. by mapping the hidden semantics underlying e.g. survey data (i.e. the scientific intention for phrasing a certain question) onto the less expressive keywords e.g. used for indexing publications. Typical problems arising from this gap in expressiveness are situations where many surveys are considered relevant because of simple key word searches in question phrases or variable labels, but an in-depth analysis of study descriptions shows that the survey as a whole is not relevant to the user's information needs. Ontologies here could be used to model certain aspects in the realm of social sciences and act as a linkage between the simpler semantics of thesauri for literature databases (e.g. narrower and broader term relationships) and the complex aspects embedded in survey questions and code books. The development of SKOS could provide a technical feasibility for this approach.

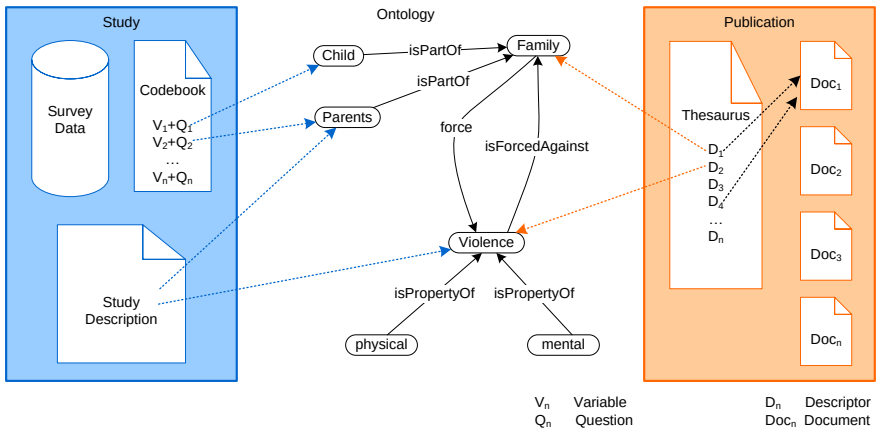

Figure 4. Layer for concept representation.

4 Conclusion and further research

The model presented here tries to combine parallel developed and complementary approaches to knowledge organization and information retrieval in the context of Digital Libraries with all their heterogeneity involved at the structural and semantic level. It seeks to overcome the shortcomings of the individual approaches with an integrated viewpoint that builds on the vast amount of traditional knowledge organisation systems available (thesauri) and, by combining them with ontologies, reduces the amount of work necessary there from modelling whole domains to modelling – as a first step - only these areas of a domain where thesauri are not expressive enough to yield satisfying retrieval quality. This approach is embedded in the full research process, which can not be ignored because of its important role for analyzing semantic relations between different entities on which the other two layers are built on. Although the view on these different approaches is expanded by combining them with each other, the focus on detail level will not get lost.

The application area and test bed of this semantic integration model is the GESIS Data Catalogue[9] and its integration into the social science portal sowiport.de[10] which contains over 2.5 million records on publications, projects, institutional profiles etc. While first results on the effectiveness of Layer 2 (Content Indexing) show that recall of relevant information can be improved (Mayr, Petras, 2008b), semantic relationships like these on Layer 1 and Layer 3 currently have not been evaluated to the same extent. In specific a combined and expanded retrieval as intended on Layer 3 is far away from first results as technical feasibilities still have to be researched and information retrieval on the Semantic Web is in general still an open research issue (Scheir et al., 2007; Finin et al., 2005). Our future work will focus on implementing and evaluating Layer 1 and Layer 3.

Although developed in the context of social sciences the model is not restricted to this domain. It should be applicable on any other domain or discipline and provides the reusing of existing knowledge organization systems.

References

ARL (2006): 'To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering', ARL/NSF Workshop on Long-Term Stewardship of Digital Data Collections, 2006.

Candela, L; Castelli, D.; Pagano, P. (2006): 'A Reference Architecture for Digital Library Systems', in: *ERCIM News, Special theme: European Digital Libraries*, 2006, No. 66.

Chorley, A.; Edwards, P.; Hielkema, F.; Philip, L.; Farrington, J. (2008): 'Developing Ontologies to Support eSocial Science: The PolicyGrid Experience', in *Proceedings of the 4th International Conference on e-Social Science*, Manchester, 2008.

DELOS (2005a): 'D5.3.1: Semantic Interoperability in Digital Library Systems', The DELOS Network of Excellence on Digital Libraries, 2005.

DELOS (2005b): 'The DELOS Network of Excellence on Digital Libraries: Recommendations and Observations for a European Digital Library (EDL)', 4th DELOS Brainstorming Workshop on Digital Libraries, December 2005.

Finin, T.; Mayfield, J.; Joshi, A.; Cost, R.; Fink, C. (2005): 'Information Retrieval and the Semantic Web', 38th Annual Hawaii Internation Conference on System Sciences. Waikoloa, Hawaii.

Goble, C., Corcho, O., Alper, P. and De Roure, D. (2006): 'e-science and the semantic web: A symbiotic relationship', in: *Discovery Science 2006*, Barcelona, Spain.

Gold, A. (2007): 'Cyberinfrastructure, Data and Libraries. Part 1 & 2', *D-Lib Magazine*, 2007, Volume 13, Number 9/10.

Gonçalves, M.; Fox, E.; Watson, L.; Kipp, N. (2004): 'Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries', *ACM Trans. Inf. Syst*, Volume 22, pp. 270—312.

Krause, J. (2004): 'Standardization, Heterogeneity and the Quality of Content Analysis: a key conflict of digital libraries and its solution', *IFLA Journal: Official Journal of the*

9 http://www.gesis.org/en/services/data/retrieval-data-access/data-catalogue

10 http://www.sowiport.de

International Federation of Library Associations and Institutions, 2004, No. 4, pp. 310-318.

Krause, J.; Stempfhuber, M. (2005): 'Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen', in König, C.; et al. (eds.): *Datenfusion und Datenintegration: 6. Wissenschaftliche Tagung*, Bonn, 2005, pp. 141-158.

Mayr, P.; Petras, V. (2008a): 'Building a terminology network for search: the KoMoHe project', in: Greenberg, J.; Klas, W. (eds.): *International Conference on Dublin Core and Metadata Applications (DC 2008)*. Berlin. pp. 177-182.

Mayr, P.; Petras, V. (2008b): 'Cross-concordances: terminology mapping and its effectiveness for information retrieval', IFLA World Library and Information Congress, 2008.

Paskin, N. (2008): 'Digital Object Identifier (DOI) System', in: *Encyclopedia of Library and Information Sciences*, 3rd Edition (to appear).

Poll, R. (2004): 'Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft', in: *ZfBB*, 2004, No. 51, pp. 59-75.

Scheir, P.; Pammer, V.; Lindstaedt, S. (2007): 'Information Retrieval on the Semantic Web – Does it exist?', in: *Proceedings of Lernen-Wissen-Adaption*, Germany, pp. 252-257.

Sure, Y.; Studer, R. (2005): 'Semantic Web Technologies for Digital Libraries', *Library Management, Special Issue: Semantic Web*, 2005, 26 (4/5), pp. 190-195.

Svensson, L. (2007): 'National Libraries and the Semantic Web: Requirements and Applications', in Prasad, A.R.D.; Madalli, D. (eds.): *International Conference on Semantic Web and Digital Libraries*, Bangalore, 2007.

Zeng, M.; Chan, L. (2004): 'Trends and issues in establishing interoperability among knowledge organization systems', *Journal of the American Society for Information Science and Technology*, *55*(3), pp. 377-395.