

# INTEGRATED RETRIEVAL OF RESEARCH DATA AND PUBLICATIONS IN DIGITAL LIBRARIES

*Maximilian Stempfhuber; Benjamin Zopilko*

GESIS – Leibniz Institute for the Social Sciences

Lennéstr. 30, 53113 Bonn, Germany

e-mail: max.stempfhuber@gesis.org; benjamin.zopilko@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.), for which to date no general model for knowledge organization and retrieval 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 focuses on integrating reference data for publications and survey data in the social sciences, but also applies the problems existing in other domains. We present a model for the integrated retrieval of factual and textual data which combines the traditional content indexing methods for publications with the newer, but rarely used ontology-based approaches which seem to be better suited for representing complex information like that 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.

**Keywords:** information retrieval; digital libraries; information integration; research data; ontology; knowledge organization; information architecture.

## **1. Introduction**

During the last few years, Digital Libraries for scientific users have been undergoing a huge change according to their role and work [1]. Results from several surveys [2] 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. Users are more and more 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. Especially in the social sciences, where on one hand data archives which document empirical data at a very detailed level are organized at an international scale and create dedicated entry points to their holdings; this information and infrastructures are on the other hand only minimally connected to the holdings of libraries and information centres. This not only challenges information providers in organizing collaboration 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 [3], understood as "enhanced" science, places the focus on creating a 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 time 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.

## **2. Current Research**

International efforts are already taking place in organizing long-term access to research information and in standardizing archival formats. Digital Object Identifiers (DOI) [4] 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. At the structural level, community driven standards for documenting primary data (e.g. the DDI format of the Data

Document Initiative [5] or SDMX [6]) are available, as are metadata standards for bibliographic references (e.g. MARC [7] or Dublin Core [8]) or entities relevant for documenting research activities and outcomes (e.g. CERIF [9], a model for research information systems which covers projects, institutes, publications, research facilities, patents etc.). 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 [10]). 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.

On the semantic level, treatment of 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 [11]. This information cannot be directly mapped on adequate thesaurus entries, which could be used for retrieving related literature, and vice versa. 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.

### **3. Model for semantic integration of heterogeneous information types**

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 the issues mentioned above. 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. Specifically, the model consists of 3 layers, each layer dealing with a dedicated semantic modelling problem (see figure 1).

In the following paragraphs the three layers are described in detail.

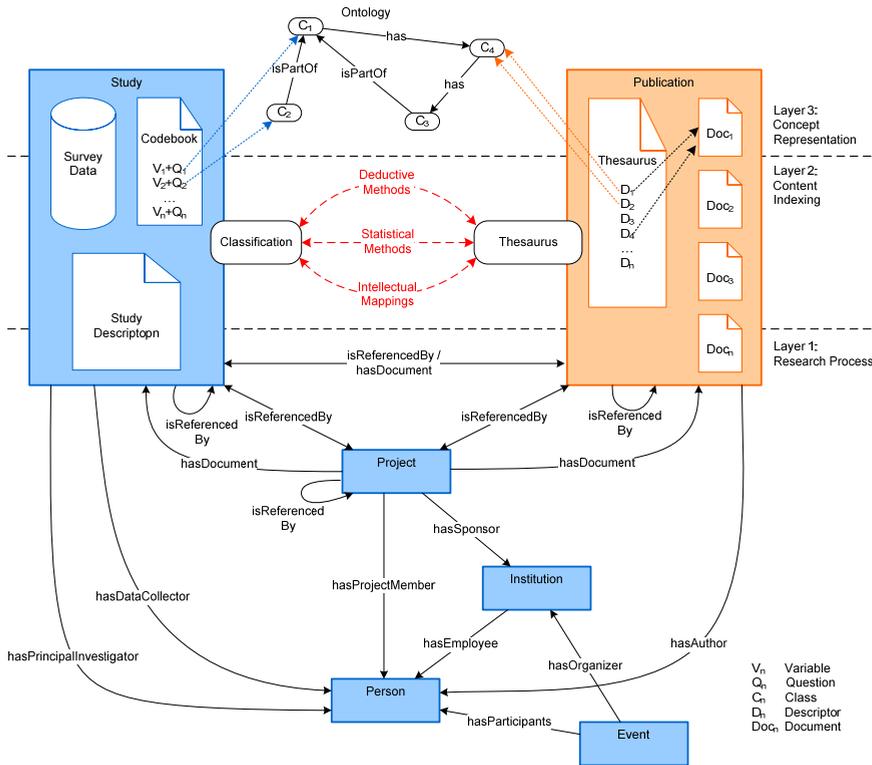


Figure 1: Full Model

### 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, the Common European Research Information Format developed by the European Commission and euroCRIS, or the PolicyGrid ontology [12]. 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.

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.

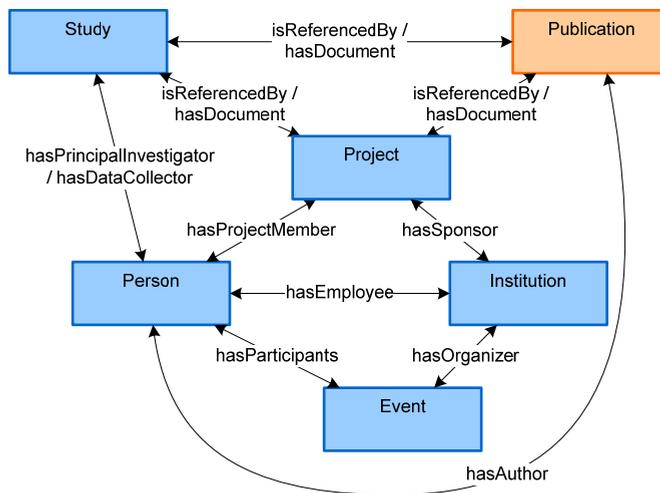


Figure 2: Layer for modelling the research process.

### 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.

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 [13] and support users by automatically transforming queries for a specific type of in-

formation (e.g. publications) to other types of information (e.g. statistical data), therefore eliminating the need to learn new indexing vocabularies or reformulating 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.

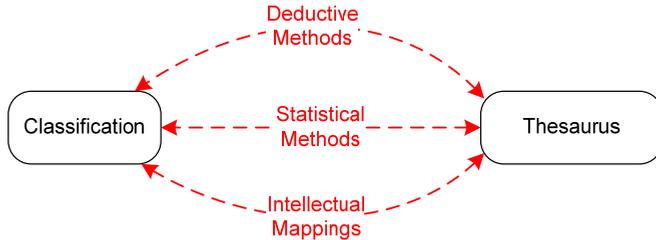


Figure 3: Layer for integrating the content indexing

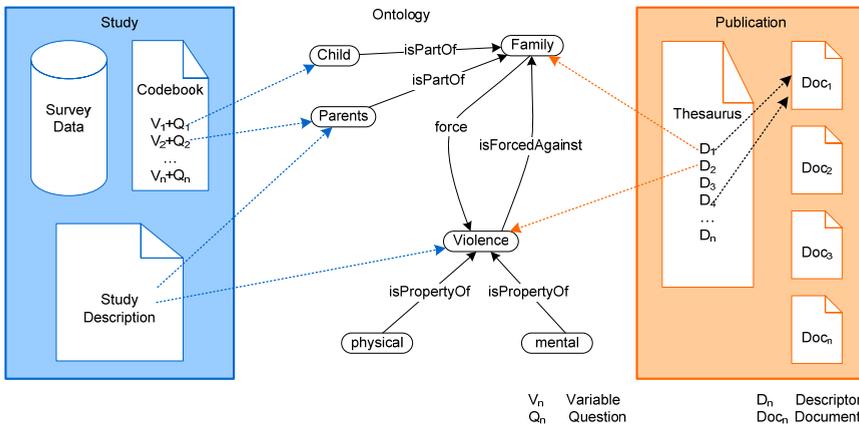


Figure 4: Layer for concept representation.

### 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.

#### 4. Conclusion and further research

The model presented here tries to combine 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.

The application area and test bed of this semantic integration model is the GESIS Data Catalogue [14] and its integration into the social science portal [sowiport.de](http://sowiport.de) [15] which contains over 2.5 million records of publications, projects, institutional profiles etc. While the first results on the effectiveness of Layer 2 (Content Indexing) show that recall of relevant information can be improved [16], semantic relationships like these on Layer 1 and Layer 2 currently have not been evaluated to the same extent. This will be the focus of our future work.

#### Notes and References

- [1] DELOS. The DELOS Network of Excellence on Digital Libraries: Recommendations and Observations for a European Digital Library (EDL). 4<sup>th</sup> DELOS Brainstorming Workshop on Digital Libraries, December 2005.
- [2] POLL, R. Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft. In *ZfBB* 51, 2004, p. 59-75.
- [3] GOLD, A. Cyberinfrastructure, Data and Libraries. Part 1 & 2. In *D-Lib Magazine*, 2007, Volume 13 Number 9/10.
- [4] DOI <http://www.doi.org>.
- [5] DDI <http://www.ddialliance.org>.

- [6] SDMX <http://www.sdmx.org>.
- [7] MARC <http://www.loc.gov/marc>.
- [8] DUBLIN CORE <http://www.dublincore.org>.
- [9] CERIF <http://www.eurocris.org/cerif/introduction>.
- [10] OAI-PMH <http://www.openarchives.org>.
- [11] KRAUSE, J. and STEMPFHUBER, M. Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen. In KÖNIG, C. et al. Datenfusion und Datenintegration: 6. Wissenschaftliche Tagung. Bonn, 2005, p. 141-158.
- [12] CHORLEY, A., EDWARDS, P., HIELKEMA, F., PHILIP, L. and FARRINGTON, J. Developing Ontologies to Support eSocial Science: The PolicyGrid Experience. In Proceedings of the 4<sup>th</sup> International Conference on e-Social Science, Manchester, 2008.
- [13] KRAUSE, J. Standardization, Heterogeneity and the Quality of Content Analysis: a key conflict of digital libraries and its solution. IFLA Journal: Official Journal of the International Federation of Library Associations and Institutions 30, 2004, No. 4, S. 310 - 318.
- [14] GESIS DATA CATALOGUE <http://www.gesis.org/en/services/data/retrieval-data-access/data-catalogue>.
- [15] SOWIPORT.DE <http://www.sowiport.de>.
- [16] MAYR, P. and PETRAS, V Cross-concordances: terminology mapping and its effectiveness for information retrieval. IFLA World Library and Information Congress, 2008.

June 2009  
Printed on demand  
by "*Nuova Cultura*"  
[www.nuovacultura.it](http://www.nuovacultura.it)

Book orders: [ordini@nuovacultura.it](mailto:ordini@nuovacultura.it)