Thesaurus-Based Indexing of Research Data in the Social Sciences:

Opportunities and Difficulties of Internationalization Efforts

by Katrin Baum[1] and Andreas Oskar Kempf[2]

ELSST

Abstract
Efforts towards internationalization have become increasingly important in scientific environments. As for content-based indexing of scientific research data, however, standards leading to internationally coherent indexing which is vital for retrieval purposes are not yet sufficiently developed. Even concerning the concrete use of indexing instruments, launched by initiatives on an international scale, there are still no binding policies and guidelines.
Against this backdrop, essential criteria which internationally applicable indexing systems should meet will be outlined. These will be illustrated through the multilingual European Language Social Science Thesaurus (ELSST), originally based on the UK Data Archive's (UKDA) Humanities and Social Science Electronic Thesaurus (HASSET) and ultimately developed by the Council of European Social Science Data Archives (CESSDA). Additionally, the general pros and cons of using international versus national indexing languages will be weighed using the ELSST and the Thesaurus for the Social Sciences (TSS) developed by GESIS – Leibniz-Institute for the Social Sciences. In this light, the benefit of vocabulary crosswalks for supporting a combined use of international and national indexing systems will be discussed.

Keywords: research data, cataloguing, thesaurus, internationalization, social sciences.

Introduction

Over the past several years, multiple efforts pertaining to the standardization of workflows, working instruments and working methods have been undertaken in various scientific domains. These efforts have been at national levels and, increasingly, on an international scale because standardization both supports and facilitates interoperability between cooperating institutions.

In the field of the social sciences the Data Documentation Initiative (DDI) metadata specification has been developed to serve as an international standard for describing data from the social sciences and related disciplines. By using this standard, coherent documentation across institutions in different countries is ensured and data exchange facilitated.

Internationalization and standardization efforts can also be observed in the context of subject indexing. The use of commonly applied indexing systems or the mapping of dispersed terminological resources is an attempt to support subject retrieval across distributed collections.

Subject Indexing of Research Data in the Social Sciences in Europe

The Council of European Social Science Data Archives (CESSDA), founded in the 1970s, is an umbrella organization for European social science data archives. Membership is comprised of data archives and other organizations which archive and provide social science data for secondary use. CESSDA currently provides access to 25,000 datasets with the collection growing

by approximately 1,000 datasets annually. Among other functions CESSDA is responsible for the development and maintenance of the European Language Social Science Thesaurus (ELSST) and the Topic Classification, both of which are used for subject indexing of research data by the member organizations. The CESSDA Catalogue enables retrieval of data stored at CESSDA archives throughout Europe and provides besides free-text search options for searches by topic of studies indexed with the Topic Classification or searches by keyword for studies indexed with ***ELSST.***

The European Language Social Science Thesaurus (ELSST) is used for subject indexing of research data by the CESSDA member organizations. It is based on the subject thesaurus HASSET which is hosted by the UK Data Archive and is being further developed by the CESSDA thesaurus management team.

It is a multilingual thesaurus for the social sciences and has been translated from English into Danish, Finnish, French, German, Greek, Norwegian, Spanish and Swedish. It consists of approximately 3,300 concepts extracted from HASSET. These concepts aim to be culturally neutral thereby reflecting a European perspective instead of one that is country-specific, thus allowing international applicability of terms. Furthermore, ELSST allows for the addition of local extensions, which means that concepts of local importance can be added to meet institutional needs.

Currently, indexing practices using ELSST vary widely across the participating institutions due to the lack of binding indexing guidelines. For example, indexing specificity ranges from the description on a very general level with only some descriptors for one study to a very precise and deep indexing with more than a hundred descriptors for one study. This can lead to the result that the same issue is very differently described. This again has implications on retrieval as the same issue can only be retrieved by using different search terms – a fact that users will not be aware of.

Additionally, due to the requirement that concepts be internationally applicable, fine-grained local issues as well as historical, juridical, religious, political and other country-specific aspects cannot be displayed if using solely ELSST. Consequently, retrieval is limited to internationally valid concepts.

Thesauri in Subject Indexing

Thesauri are being used for verbal subject indexing in documentation. Consistently applying the same, controlled descriptor for specific issues results in consistent documentation and facilitates retrieval.

Indexing systems are usually based on specific collections, meaning that content and structure of systems even in the same domain can differ considerably. As well, levels of abstraction and hence specificity can vary among different thesauri depending on local conditions and needs. Moreover, different classification aspects following from a variety of perspectives on a topic can lead to different semantic relations between concepts.

Requirements of an internationally applicable thesaurus

One of the most important requirements for an internationally applicable thesaurus is that it be free of bias. The concepts it contains need to exist in every participating culture and have to be displayed in a hierarchical and semantic structure that fits all cultures and languages. Terms for concepts have to be multilingual to allow access in all of the languages in use.

However, the characteristics of an internationally usable system such as this include numerous limitations and constraints. Fine-grained issues at both the institutional and country-, respectively language-specific level cannot be displayed; thus retrieval is limited to internationally applicable concepts.

Local indexing systems

Local indexing systems are able to reflect the scope of the local collection very accurately and with respect to cultural characteristics. This allows for more precise indexing. Beyond that, they are easier to maintain as there are no cross-institutional agreements to follow.

On the other hand, the exclusive use of a local indexing system has its own deficiencies. Since it remains a solely locally applied system, without further measures there can be no access points for unified subject retrieval across dispersed collections that have been indexed using different terminological resources.

Recommended Indexing Model

One possible way to offer uniform subject access to heterogeneously indexed collections in a dispersed environment is the mapping of institutionally used indexing systems (Doerr 2001). Applied on retrieval, mappings aid the user in finding documents indexed no matter which indexing system was used by being able to only employ search terms in the system he or she is familiar with.

Though, mappings often carry a certain amount of intrinsic vagueness due to incomplete congruity between concepts of different indexing systems. For this reason we propose an aggregate of local thesauri with

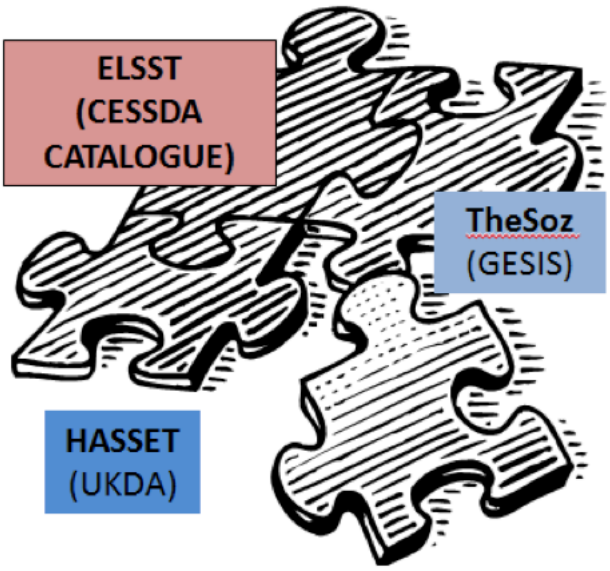

Figure 1: Interplay between local and international indexing system.

common, internationally applicable core concepts (Boteram & Hubrich 2008; Gödert 2008). It should contain concepts that exist in any language and its hierarchical structure should fit all languages as close as possible ensuring that it is free of bias. Concepts already existing in local systems should be mapped to concepts of the core system. As well, concepts missing in the local systems should be added.

To reiterate, we would argue for a direct interplay between international and local indexing system in a way that will permit internationally applicable concepts standing for the core vocabulary to be fully integrated, i.e., represented, in all of the different local indexing systems attached to it. Consequently, the whole content of the core vocabulary is simultaneously part of every single local indexing system. This is the reason that the universal core system, i.e., the ELSST, must contain all central concepts which exist in all of the included languages. And, vice versa, these central concepts must be integrated into local indexing systems. For instance, the key indexing tool for German-language social sciences, the Thesaurus for the Social Sciences (TSS), translated into English and French, contains more than 8,000 concepts and, like the ELSST, includes a wide range of subdisciplines of the social sciences.

Looking at this direct interplay between local and international indexing system in practice, we explicitly advocate for further use of the local system as key indexing system. Taking "secondary school" as an example from the education sector, this concept, referring to the local indexing system introduced so far, needs to be, if not already the case, incorporated into the TSS. To summarize, missing concepts which are part of the internationally used core vocabulary must be created in the local indexing systems.

In addition to these internationally applicable concepts, the local indexing system must also contain any locally distinctive specificities. For example, the German concept "Gymnasium", stands for a certain type of secondary school, one with a strong emphasis on academic learning. It is comparable to the British grammar school system or preparatory schools in the United States. Moreover, the local indexing system contains collection-specific concepts, e.g., the geographic subject heading "Nordrhein-Westfalen", a federal state in Germany, indicating, for instance, the provenance of a dataset.

It becomes clear that a connection between internationally used core concepts and local indexing systems is necessary. In our judgment, this linkage could be best achieved by terminology mapping between international and local indexing system. Referring again to the two thesauri mentioned above, we like to hint at the major terminology mapping initiative conducted by GESIS - Leibniz-Institute for the Social Sciences as part of the project Competence Center Modeling and Treatment of Semantic Heterogeneity (KoMoHe) (Mayr & Petras 2008). Carried out shortly after the ELSST extension in the framework of the European Commission project Multilingual Access to Data Infrastrctures of the European Research Area (MADIERA) its main objective was to create crosswalks between various controlled vocabularies and also between the ELSST and the TSS. Approximately 2,300 equivalent relations were built up in each direction which could be reused when translating ELSST vocabulary for inclusion into the TSS. A search using the ELSST-concept "secondary schools", would directly create a link to datasets indexed with the German term "weiterführende Schule", and respectively into their English and French translations.

Additionally with the help of these crosswalks, the extent to which the international core system is already part of local indexing systems becomes apparent. Moreover, looking at these crosswalks from the opposite direction, in our case from the TSS to the ELSST, and looking at non-equivalent relations of this mapping, which had also been built up in the past, gives a hint at local specificities being part of the local indexing system. For example, the German concepts "Gymnasium", "Realschule" and "Hauptschule" are narrower terms of the above mentioned concept "weiterführende Schule".

Retrieval aspects

We suggest there are significant information retrieval benefits to be obtained as a result of the direct interplay that occurs between internationally applicable concepts and local indexing systems. First of all, using an integrated retrieval system, e.g., the CESSDA Catalogue, the researcher is able to use proper terminology for core concepts included in the multilingual ELSST. Due to the hierarchical semantic structure of the thesaurus, the researcher is aided in the search for narrower subject terms. At this point the researcher has access to all the data collections of the CESSDA member organizations indexed with those commonly shared international core concepts.
For fine-grained regional datasets with existing vocabulary mappings between international and local indexing system, a linkage between both vocabularies could be established. Similar to the CESSDA catalogue's tree-like hierarchical search structure, additional local specific terms could be connected to the broader terms in the international indexing system. Thus, datasets indexed with the specific subject headings become searchable and accessible. It will prevent locally embedded information from being buried under broad general indexing terms.

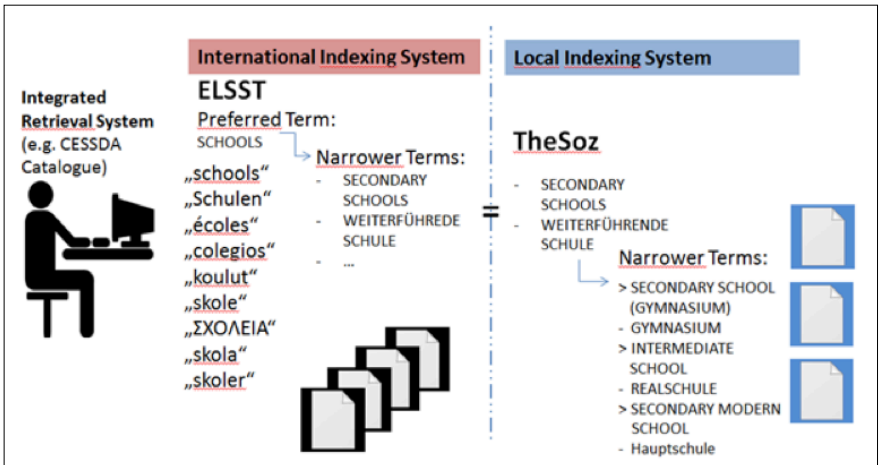

Figure 2: Application of the indexing model for retrieval.

Figure 2:

Conclusion

In a period when efforts towards standardization and internationalization of subject indexing of research data have become increasingly important, there is a pressing need to determine ways to integrate local indexing systems into widely launched internationally applicable vocabularies. Even though work on this began in the 1970s, there are still no binding and coherent indexing guidelines. Vast differences in cross-country indexing of research data remain.

With this as our backdrop, our aim was to present a concrete proposal on how to create an interconnection between local and international indexing system. Hence, we argued for an aggregate of local thesauri with a common core vocabulary of internationally applicable

key concepts. These concepts would exist in all of the participating languages represented in CESSDA. The vocabulary would be kept free of bias as much as possible. Concepts already integrated in local systems would be mapped to the core system and any missing concepts in the local systems would be added. Doing this would achieve a coherent and unified subject indexing of dispersed collections of research data. Information retrieval would be significantly improved. Fine-grained institutional and locally distinctive datasets that have been indexed with collection-specific subject headings will become searchable via crosswalks built up to the international indexing system. The local indexing system will remain the key tool as it is much easier to maintain.

Concrete efforts to move forward will include as a first step the review of ELSST to ensure no bias in the concepts and the interconcept-relations. Following this, the second step is to adapt the local systems. Subsequently, universally accepted indexing guidelines for the core system need to be developed.

References

Boteram, F, Hubrich, J 2008, Towards a comprehensive international Knowledge Organization System, Available from: < http://linux2.fbi.fh-koeln.de/crisscross/vortraege.html >. [19 July 2013].

Council of European Social Science Data Archives 2013, Available from: <http://www.cessda.org/>. [19 July 2013].

CESSDA Catalogue 2013, Available from: <http://www.cessda.org/accessing/catalogue/>. [19 July 2013].

Doerr, M 2001, 'Semantic Problems of Thesaurus Mapping', Journal for Digital Information, vol. 1, no. 8. Available from: <http://journals.tdl.org/jodi/index.php/jodi/article/view/31/32>. [19 July 2013].

European Language Social Science Thesaurus 2013, Available from: <http://elsst.esds.ac.uk/login.aspx>. [19 July 2013].

Gödert, W 2008, Ontological Spine, Localization and Multilingual Access: Some Reflections and a Proposal, Available from: <http://linux2.fbi.fh-koeln.de/crisscross/publikationen/goedert2008.pdf>. [19 July 2013].

Mayr, P, Petras, V 2008, Cross-concordances: terminology mapping and its effectiveness for information retrieval, Available from: <www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf>. [19 July 2013].

Thesaurus for the Social Sciences 2013, Available from: <http://www.gesis.org/en/services/research/thesauri-und -klassifikationen/social-science-thesaurus/>. [19 July 2013].

NOTES

1. Katrin Baum is a librarian at GESIS – Leibniz-Institute for the Social Sciences in Cologne, Germany, where she works in the Data Archive department in the area of study descriptions. Her professional focus is on subject indexing and information retrieval. Katrin can be reached by email: mailto:katrin.baum@gesis.org.

2. Andreas Oskar Kempf is a research associate at GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany. He holds a PhD in sociology from Goethe University, Frankfurt am Main, and received a Master´s degree in Library and Information Science from Humboldt-University, Berlin, Germany. Andreas conducts applied research on GESIS authority data for content cataloguing (i.e. Thesaurus for the Social Sciences) of social science research literature, projects and data. He can be reached by email: mailto:andreas.kempf@gesis.org.