gesis
Leibniz-Institut
für Sozialwissenschaften

TECHNICAL *Reports* 2009|07

Converting the TheSoz to SKOS

Benjamin Zapilko, York Sure

GESIS-Technical Reports 2009|07

Converting the TheSoz to SKOS

Benjamin Zapilko, York Sure

GESIS – Leibniz-Institut für Sozialwissenschaften 2009

GESIS-Technical Reports

GESIS – Leibniz-Institut für Sozialwissenschaften
Informationelle Prozesse in den Sozialwissenschaften
Lennéstraße 30
53113 Bonn
Telefon: (0228) 22 81 - 175
Telefax: (0228) 22 81 - 4175
E-Mail: benjamin.zapilko@gesis.org

ISSN: 1868-9051 (Online)
Herausgeber,
Druck und Vertrieb: GESIS - Leibniz-Institut für Sozialwissenschaften
Lennéstraße 30, 53113 Bonn

1 Introduction

The Thesaurus for the Social Sciences (TheSoz) [1] serves as a crucial instrument for indexing the content of documents in the databases SOLIS (Social Science Literature Information System) and SOFIS (Social Science Research Information System), both are owned and maintained by GESIS - Leibniz Institute for the Social Sciences [2]. It contains overall about 11,600 keywords and covers all topics and sub-disciplines of the social sciences. Additionally terms from associated and related disciplines are included in order to support an accurate and adequate indexing process of interdisciplinary, practical-oriented and multi-cultural documents.

Paying attention to recent developments in the fields of e-Science and the web of Linked Open Data (LOD) it seems to be quite obvious that the TheSoz has to be made available in the web in a compatible and machine-readable format for providing and sharing its relevant information with a greater community. With the development of SKOS (Simple Knowledge Organization System) [3] there is a "standard way to represent knowledge organization systems using the Resource Description Framework (RDF)"[1]. Due to the use of RDF [4] information can be passed and re-used in a very interoperable way. First attempts for modeling the TheSoz with SKOS have been made since the beginning of 2009 due to the fact that a lot of organizations and libraries were bringing their thesauri and vocabularies to the web in SKOS format. These developments can also be observed in Germany where recently the Thesaurus for Economics (STW)[2] of the ZBW has been published as linked data [5]. After SKOS has been announced as a standard in August 2009 by the W3C a stable SKOS version of the TheSoz has been worked on. The following article describes the transformation process of the TheSoz in detail and takes an outlook on possible future developments.

2 Converting the TheSoz to SKOS

The transformation process was split up into three steps and the method followed the established method described in [6]. It began with a conceptual analysis of the thesaurus. Based on this all elements of the thesaurus have been mapped to adequate properties of SKOS. Special use cases in this mapping process are described in detail. The technical conversion process itself is automated and done via XSLT methods.

2.1 Thesaurus Analysis

The Thesaurus for the Social Sciences contains about 11,600 keywords, of which more than 7,750 are descriptors (authorized keywords) and about 3,850 are non-descriptors. Relationships between these keywords are expressed as broader, narrower or related terms as well as there are also "use instead" and "use combination" relations and their counterparts ("used for" and "used for combination"). Additionally a classification hierarchy is provided and each thesaurus term is dedicated to one or more classification terms. The TheSoz contains a special type of non-descriptor called "AD" (for alternative descriptor) which differs from the international standard norms for thesauri[3] and holds more than one "use instead" and/or "use combination" relation at the same time. There are about 200 of such "AD" terms in the TheSoz.

---

[1] http://www.w3.org/2004/02/skos/intro

[2] http://zbw.eu/stw/

[3] ISO 2788 (Guidelines for the establishment and development of monolingual thesauri) and ISO 5964 (Guidelines for the establishment and development of multilingual thesauri)

- Example: The term "AD alternative energy" contains the relations "USE INSTEAD renewable energy" as well as "USE COMBINATION alternative AND energy". Both of them are (a combination of) descriptors with the same relevance. In most cases terms of the type "AD" describe generic terms which have different concrete meanings in specified sub-contexts. This is expressed through the use of more than one "use instead" and/or "use combination" relations for only one term.

2.2 Mapping to SKOS

Most of the thesaurus items could easily be mapped to adequate SKOS properties and classes (see table 1) due to the broad consistency of the TheSoz to the standard norms for thesauri. Problems could be observed when mapping special data items and/or relations which are not conform to thesauri standards to SKOS. This problem is an already well-known issue in research context [5], but due to the fact that SKOS is based on RDF it is possible to define own relations without greater effort.

Table 1: Mapping of TheSoz data items to SKOS

Data Item	Feature / Function	Property / Class
DD	Descriptor / Preferred Term	skos:prefLabel
ND	Non-Descriptor / Non-Preferred Term	skos:altLabel
AD	Alternative Descriptor	skos:altLabel
NT	Narrower Term	skos:narrower
BT	Broader Term	skos:broader
RT	Related Term	skos:related
USE	Use X instead of Y	skos:prefLabel
UF	Opposite of USE: X used for Y	skos:altLabel
USK	Use Combination X AND Y	skos:prefLabel; skos:Collection; skos:member
UFK	Opposite of USK: Used for X with Y	skos:altLabel
scope	Scope Notes for specific terms	skos:scopeNote
notationcode	Numerical code of one or more items of the classification hierarchy to which the term is dedicated to	skos:notation
-	Editorial Notes are used to include relations which would get lost due to the non-conform modeling of some relations of the AD descriptor	skos:editorialNote

In case of the TheSoz the "use combination" relation has been modeled via grouping the affected terms as multiple "skos:member" in a "skos:Collection" which again is included in one "skos:prefLabel". But as mentioned above the TheSoz also contains a special type of non-descriptor called "AD" which holds more than one "use instead" and/or "use combination" relation at the same time. Modeling such a term to SKOS would invoke more than one "skos:prefLabel" in one single concept. Therefore these relations were modeled backwards via their "used for" and/or "used for combination" relations in the associated descriptors as it is suggested in various discussions in the web, but a small loss of information could not be avoided with this solution. To avoid a complete loss of this relevant information these relations were

included in additional "skos:editorialNotes" until there is a satisfying way to model them correctly with SKOS.

The classification hierarchy of the TheSoz (originally stored in a separate file) could be mapped to SKOS without problems (see table 2). The numerical code of each classification term which appears in "rdf:about" as well as in "skos:notation" is the same code which is included in "skos:notation" of each concept containing descriptors (see table 1). By referencing these notation codes via URIs a connection between descriptors and their according classification terms is established.

Table 2: Mapping of TheSoz classification hierarchy to SKOS

Data Item	Feature / Function	Property / Class
code	Numerical code of the classification item	skos:notation
description	Name of the classification item	skos:prefLabel
child	Child nodes of the current classification item	skos:narrower
parent	Parent node of the current classification item	skos:broader

2.3 Conversion Program

Based on the mappings in the paragraph above the conversion program was developed. The technical conversion process is done by XSL transformations. The original digital format of the TheSoz which was already encoded in XML was converted to SKOS RDF. Additionally to the mapping of each data item each concept got its own URI which provides a persistent and unique identification. This is a very important aspect for re-use, availability and linkage in the web and possible collaborations. The SKOS version of the thesaurus contains two types of URIs, one for the descriptors and non-descriptors of the thesaurus and one for the terms of the classification hierarchy. This allows easy distinguishing between descriptors and classification terms by only knowing the concept URI. After the transformation process the resulting SKOS version of the TheSoz was tested and validated by various established validation services for RDF[4] and SKOS[5].

3 Future Developments

Although the current SKOS version of the Thesaurus for the Social Sciences represents a stable and valid data set which contains all of the original content of the thesaurus there is room for improvements and extensions coming up in the near future. Beside mandatory technical enhancements and an upcoming update of the content there are currently a few more issues being discussed and planned at GESIS:

- Thesaurus Browser
  According to the Thesaurus Browser which is already available at the social science portal sowiport.de [7], a similar web representation of the SKOS version of the TheSoz is discussed. Exporting the data as RDF should be possible and provided as well as annotating the html pages with RDFa [8] elements.

[4] www.w3.org/RDF/Validator/

[5] www.w3.org/2004/02/skos/validation

- SPARQL Endpoint
  A SPARQL [9] endpoint on top of sowiport.de is planned, not only for querying concepts and relations of the TheSoz, but also for retrieving all data provided in sowiport.de.
- Multi-Thesauri Setting
  GESIS holds a large network of crosswalks between over 20 vocabularies which was resulting from the project KoMoHe[6]. The relevant information represented in these crosswalks could (at a first step partially) be modelled with the mapping properties of SKOS. This could establish a multi-thesauri setting in SKOS which uses already existing crosswalks between the involved thesauri. The participated thesauri should also be available in SKOS.

These further plans indicate major steps towards the web of Linked Open Data (LOD) which seems to be a very relevant research field for GESIS in the next years.

The current SKOS version of the TheSoz can be found and downloaded at the GESIS website[7].

References

1. Thesaurus for the Social Sciences: http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/

2. GESIS - Leibniz Institute for the Social Sciences: http://www.gesis.org/

3. SKOS - Simple Knowledge Organization System: http://www.w3.org/2004/02/skos/

4. RDF - Resource Description Framework: http://www.w3.org/RDF/

5. Neubert, J.: Bringing the "Thesaurus for Economics" on to the Web of Linked Data. Linked Data on the Web (LDOW) 2009. Workshop at the WWW 2009 Conference, April 20th, Madrid, Spain (2009)

6. Assem, M. van, Malaisé, V., Miles, A., Schreiber, G.: A Method to Convert Thesauri to SKOS. In: The Semantic Web: Research and Applications. pp. 95-106 (2006)

7. SOWIPORT: http://www.sowiport.de/

8. RDFa: http://www.w3.org/TR/xhtml-rdfa-primer/

9. SPARQL - Query Language for RDF: http://www.w3.org/TR/rdf-sparql-query/

[6] Competence Center Modeling and Treatment of Semantic Heterogeneity"; funded by the German Federal Ministry for Education and Research; http://www.gesis.org/en/research/programs-and-projects/information-science/project-overview/komohe/

[7] http://www.gesis.org/dienstleistungen/tools-standards/thesaurus-sozialwissenschaften/