Linking Study Descriptions to the Linked Open Data Cloud

by Johann Schaible[1], Benjamin Zapilko[2], Thomas Bosch[3], and Wolfgang Zenk-Möltgen[4]

Abstract

The GESIS Data Catalogue contains the study descriptions for all archived studies at GESIS, currently more than 5000 datasets mainly from survey research in the social sciences. These descriptions include information about primary researchers, research topics and objects, used methods, and the resulting dataset, which is mainly used for archiving and retrieval in order to serve secondary researchers. For this purpose the existing metadata can be enriched with further information about the study investigators, involved affiliations, collection dates, content, and more from other sources like DBpedia or the Name Authority File of the German National Library. In recent years the paradigm of Linked Open Data (LOD) encouraged various research organizations to expose their data to the web according to Semantic Web standards. This has increased the number of available data sources and the feasibility of their reuse.

In this paper, we present ways to enrich a study description with various datasets from the LOD cloud. To accomplish this, we expose selected elements of the study description in RDF (Resource Description Framework) by applying commonly used vocabularies. This optimizes the interoperability to other RDF datasets and the discovery of links to them. For link detection we use Silk, a framework for discovering relationships between data items within different LOD sources. Once links are detected, the study description is linked to adequate entities of external datasets and therefore holds additional information for the user, e.g. further metadata on the principal investigator of a study.

Keywords:

Semantic Web, Linked Open Data, Data Transformation, RDF, Link Discovery, Metadata

Introduction

The Linked Open Data (LOD) cloud[5] comprises data from diverse domains. Various best practices and principles (Bizer et al. 2009) guide a data publisher in modeling and publishing data as Linked Data. To use Semantic Web technologies such as RDF[6] and SPARQL[7] and to include links to external data providers are two essential points in the guidelines, as this leads to better discovery of information by Linked Data applications and users (Heath and Bizer 2011). The GESIS Data Catalogue (DBK)[8] comprises study descriptions for all archived studies at GESIS. It contains metadata about each study, such as the primary researchers, research topics and objects, used methods, etc., which is archived to serve as an information pool for secondary researchers. Thus, the visibility of such a dataset is an important aspect. To publish this metadata as Linked Open Data would increase the visibility because external data providers can set links to particular Linked Data sources. This way, secondary users are able to discover the data from multiple points of access. Furthermore, the existing metadata can be enriched with additional information from other external data providers. For example, the GESIS data catalogue can be enriched with additional information about the study investigators, involved organizations, collection dates, content, and more from external sources like

> To publish metadata as Linked Open Data would increase the visibility of the data

DBpedia[9] or the Name Authority File (GND)[10] of the German National Library (also named PND). Note that the publication of the metadata as LOD is intended, not the publication of the quantitative dataset. In terms of computer science both are data and could be published as LOD. But the quantiative datasets can only be ordered or downloaded by agreeing to the usage regulations of the GESIS Data Archive. However, the metadata of the Data Catalogue is freely available and was modeled as LOD in this paper. Please note, that the LOD representation of the Data Catalogue has not been published yet, and the links provided in various examples are as yet hypothetical.

In this article, we describe the modeling and the publishing of a dataset as Linked Open Data and the procedure for how to interlink this resulting Linked Dataset to external data sources. Hereby, we especially focus on the difficulties in producing Linked Open Data. Our dataset is an excerpt from the GESIS data catalogue comprising specific metadata about social science studies. This metadata is stored as XML flat files. The mapping to existing RDF vocabularies is done manually. To transform it into RDF, we use plain XSLT scripts. We use the link discovery tool Silk[11] to detect links from the RDF representation of the GESIS Data Catalogue to external data sources.

We discuss our observations on the benefits of the described approach to publish data. In detail, we inspect whether we gain any efficiency in handling of the data, whether we gain new information from external data providers, and what is possible with such a dataset stored in RDF in contrast to XML. We provide answers to these questions with respect to the effort and difficulty in producing such Linked Open Data.

The article is structured as follows: in Section 2, we describe the GESIS Data Catalogue in detail. Furthermore, we illustrate what metadata it contains and which data elements we used for our excerpt. In Section 3, we demonstrate the transformation of the XML data into RDF. This also includes the choice of the existing vocabularies as well as the mappings to terms from these vocabularies. Section 4 provides an insight into the link discovery framework Silk. We present how Silk can be used to detect links to external datasets containing information on the same resources. We present the results of our work in Section 5 and describe the advantages and the disadvantages of publishing data as Linked Open Data. In Section 6, we conclude our work and give an outlook to future work.

The GESIS Data Catalogue (DBK)

The GESIS Data Catalogue (DBK) comprises the study descriptions from all archived studies and empirical primary data mainly from survey research and historical social research which are published on the GESIS homepage by the application DBKSearch. It is possible to search within the study descriptions by using a simple or advanced search. The simple search is carried out in all or selected fields, whereas the advanced search combines more search terms in different fields. The management of this metadata is implemented by the DBKEdit application that also handles internal metadata and workflows. The GESIS Data Archive uses the Data Catalogue also to publish the metadata in other portals and systems, such as ZACAT[12], the CESSDA data portal[13], Sowiport[14], and the data registration agency da|ra, which again is linked to the metadata store of DataCite[16]. The applications DBKEdit and DBKSearch are also available as an open-source for other providers under the name DBKfree[17].

The list of structured data which describes a dataset of the archive and makes it easier to find is defined by the metadata schema of the Data Catalogue (Zenk-Möltgen and Habbel 2012). Since the establishment of the Central Archive for Empirical Social Research 50 years ago (now part of GESIS), the metadata schema as a system for study description has always been refined in the context of the cooperation of the international archives and is continuously being developed and adapted to new standards (Mochmann 1979, Bauske 1992, and Bauske 2000). The metadata schema contains a number of mandatory core elements which have to exist for the creation of a new study description. Furthermore, optional metadata elements can be used to describe the data more precisely. For some elements other applicable standards are used, e.g., ISO standards for dates or geographic locations.

The DBK metadata schema is compatible with the Codebook and Lifecycle standards of the Data Documentation Initiative[18] (DDI) and can be exported into the DDI2 and DDI3 XML formats. Moreover, it is compatible with the metadata schema of the GESIS agency for data registration da|ra and DataCite (Hausstein et al. 2011). In addition to the DataCite metadata schema, the DBK metadata contains specific social science information which supports retrieval and especially allows for a methodological comprehensive description of research data. Currently, other social science data archives like the ICPSR[19] in the U.S., DDA[20] in Denmark, NSD[21] in Norway, and the UKDA[22] in the United Kingdom use similar study descriptions for their holdings.

To enrich the study descriptions with additional information using Semantic Web technologies, it is possible to publish the Data Catalogue as Linked Open Data. For this the Data Catalogue XML files have to be transformed into RDF. We used the DDI Codebook XML format and extracted some entities from the DBK which seem to be most promising with respect to finding additional information for the studies. For example, "title," "author," and "abstract" are such important entities, but "caseQnty" (number of variables in the data file) is not. Following is the entire list of the selected important entities for a study description, and Figure 1 displays a pseudo-XML of the structure of the entities.

- **Title statement:** The title statement contains a mandatory element "Title" and an optional list of elements named "Alternative title". Alternative titles can also be of the type project title, original title, or subtitle.
- **Responsibility statement:** The responsibility statement contains the repeatable element "Authoring Entity" with an "Affiliation" of the authoring entity as an attribute. This element contains the principal investigators that should be cited for the creation of the study. Their institution is named in the affiliation attribute. Sometimes institutions are named directly as the principal investigator.
- **Production and Distribution statement:** The production statement comprises the elements "Producer" and "Distributor". The distribution statement currently contains the name of the GESIS Data Archive with its abbreviation and website URL as attributes. The element "Funding Agency" is currently not used by the DBK in the DDI study descriptions.
- **Study Info**: In the entity Study Info there is a list of topic classifications for the study from the ZA-Category System and a detailed thematic description of all the variables in the dataset in the "Abstract" element. Both elements are available in German and English, but for some study descriptions there is still a lack

```
stdyDscr [study description]
      citation
             titlStmt [title statement]
                    titl [title]
                    altTitl [alternative title]
              rspStmt [responsibility statement]
                    AuthEnty [authoring entity] @affiliation
              prodStmt [production statement]
                    producer
                    fundAg [funding agency]
                    distStmt [distribution statement]
                    distrbtr [distributor] @abbr [abbreviation] @URI
      stdyInfo
              subject [language depended]
                    topcClas [category; language depended]
              abstract [language depended]
              sumDscr
                    collDate [collection date]
                    universe [language depended]
      method
              dataColl [data collection]
                    timeMeth [language depended]
                    dataCollector [language depended]
                    sampProc [sampling procedure; language depended]
                    collMode [collection mode; language depended]
      dataAccs
              setAvail [availability statement]
                    accsPlac [access place] @ID @URI
              useStmt [usability statement; how to use the study?]
                    contact
      othrStdyMat [other study material]
             relStdy [related study]
             relPubl [related publication]
             othRefs [other references; further remarks]
```

Figure 1: The extracted entities from the DBK as pseudo-XML

of translations into English for the abstract. In this section there is also the list of "Geographic Coverage," which contains country and region names from the ISO-Format and additional free text, and a description of the "Universe" that the data applies to (both language dependent).

- **Data collection:** In this entity there is the list of collection dates in ISO-Format under the element "Time Method". In addition, there are the elements "Data Collector", "Sampling Procedure", and "Collection Mode" (all language dependent) which describe the methodology of the data collection process.
- **Data access:** Data access comprises a section "Data Set Availability" which contains the element "Access Place" for describing the location of the access place and an URI of the place as attribute, and the "Availability Status" of the study which is described in English and German.
- **Other study material**: In the entity "Other Study Material" there are the elements "Related Material" containing data and document files that may be downloaded with Name and URL, "Related Publications" with the full citation, and "Other References" with further remarks that may contain notes to the study (language dependent).

Converting the Data Catalogue XML into RDF

To convert XML data into RDF two steps have to be passed: the mapping and the technical conversion. While the latter step can be solved by writing and executing scripts like XSL transformations, the mapping of XML elements to RDF properties and classes requires expert knowledge for the domain of the data as well as for Semantic Web vocabularies. That is because on the one hand the data must be converted correctly to RDF without losses or changes in its semantics. On the other hand interoperability with other data expressed in RDF and Semantic Web applications has to be ensured. As described in Bizer et al. (2009) and Heath and Bizer (2011), it has become best practice to reuse properties and classes of existing and popular Semantic Web vocabularies as much as possible. But the search for the most adequate properties and classes for representing the semantics of the source XML data can be a time-consuming task, especially if there are several potential suitable RDF vocabularies or if the data is not fully covered by them. The search is complicated since the number of RDF vocabularies has increased massively during recent years. Hence it requires expert knowledge for deciding which vocabularies should be used for representing the data.

There are several typical decisions that have to be made when defining a mapping of metadata entries to properties and classes of RDF vocabularies. Some of them depend on the trade-off between a semantically rich expressiveness of the resulted RDF data and an intensive reuse of existing and popular vocabularies. One has to decide consistently for the full mapping and especially for particular data elements whether a correct and full semantic expressiveness of the data or a technical interoperability with other Linked Data sources is of higher relevance. This influences directly the amount of used vocabularies and whether the definition of

an own vocabulary becomes necessary. If the preservation of the semantic meaning of every data element is the highest goal for a conversion, then it is very likely that not all elements can be represented by existing RDF vocabularies and it is necessary to define individual classes and properties in their own vocabulary. The following examples present cases where these considerations are of importance:

- In some cases there is more than one adequate property or class to represent a particular data element. For instance, there are several properties for describing elements of the XML data e.g., title or date. These properties are typically part of different vocabularies like Dublin Core[23] or particular bibliographic vocabularies. One has to decide which property or class of which vocabulary to use for the representation of a particular data element.
- There may be a loss of semantics when mapping a data element to a property of a popular vocabulary instead of mapping it to a property of a less popular vocabulary, which represents the semantics of the element more precisely. For instance, the data element describing a particular time (e.g., the time period observed in a study) is not represented adequately by the general date property from the Dublin Core Elements vocabulary instead of a more precise property of a lesser known vocabulary.
- Two data elements with the same data type, but a slightly different semantic meaning, e.g., starting date of a survey and modification date of a dataset, can lose their meaning if they are represented by the same property (again, e.g., the date property from the Dublin Core Elements vocabulary). Such data elements should be represented in RDF by different properties in order to keep the semantic difference between them.

Additionally, it has to be decided whether data elements should be represented as resources or as properties. A resource is represented with an URI and is in a general sense a "thing". Every resource has properties, which we define as literal values describing the resource. This design decision has to be made carefully, because only resources can be linked to other resources of the Linked Open Data cloud. The instances of properties are commonly expressed as plain literals and cannot be enriched by further information and links. For example, if the principal investigator of a study were modeled as a literal value, it would not be able to interlink this property with an external dataset containing information about persons. On the other side, if the principal investigator were modeled as a resource, it can be interlinked with another resource from an external data source. The structural difference between a resource and a property is defined in the structure of an expression in RDF, as it is a collection of triples, each consisting of a subject, a predicate, and an object. The subject is in most cases an RDF URI that references a resource. The object is usually either an RDF URI that also references a resource or a literal value describing the subject. The predicate is also an RDF URI that links the subject to the object. For example, the resource "study" is the subject. It has the object "principal investigator", which is also a resource, and the object "study title" that is denoted as a literal. The predicate "hasTitle" links the resource "study" to the object "study title" containing a literal value, and the predicate "hasPrincipalInvestigator" links the resource "study" to a resource "principal investigator". These two expressions are considered to be triples.

For the conversion of study descriptions to RDF in order to detect links we decided to reuse existing vocabularies, but as few of them as possible. By choosing popular vocabularies we allow for high interoperability with other datasets of the LOD cloud. This was also the reason we did not define our own properties and classes, although some data elements cannot be covered to the same full semantic extent in RDF as in their original XML representation. The choice of reusable vocabularies that can express the DBK entities in the best possible way was based on the description of the vocabulary and its human-readable documentation. As most appropriate vocabularies, we have identified the DDI-RDF Discovery Vocabulary (DISCO)[24] , the Dublin Core vocabulary (DCTerms), as well as the Semantic Web for Research Communities vocabulary (SWRC)[25] . The DISCO vocabulary covers many DDI2 elements that are used in the Data Catalogue DDI2 XML export. However, all of the terms from the DISCO vocabulary that were considered as appropriate mapping are reused classes and properties from the Dublin Core vocabulary. Thus, it is more convenient to use the classes and property from Dublin Core directly. The SWRC vocabulary is widely used to model entities of research communities such as persons, organizations, and bibliographic metadata on publications, which suits our purpose very well.

As mentioned earlier, the first step to transform the Data Catalogue XML files into RDF is to map the various entities to the classes and properties from the vocabularies we have identified as most appropriate. Table 1 shows the possible mappings of all entities

	DCTerms	SWRC
Title	dcterms:title (*)	swrc:title
Alternative Title	dcterms:alternative (*)	
Authoring Entity	dcterms:creator (*)	swrc:author
Affiliation		swrc:affiliation (*)
Producer	dcterms:Agent (*)	
Distributor	dcterms:publisher (*)	
Category	dcterms:subject (*)	
Abstract	dcterms:abstract (*)	swrc:abstract
Universe	dcterms:coverage (*)	
Time Method	dcterms:date (*)	swrc:startDate swrc:endDate
Data Collector	dcterms:contributor (*)	
Sample Procedure	dcterms: accrualMethod (*)	
Collection Mode	dcterms: accrualMethod (*)	
Access Place	dcterms:Location (*)	
Related Publication	dcterms:relation (*)	
Other References		swrc:note (*)

Table 1: Mapping of the Data Catalogue entities to terms from the different vocabularies. The vocabulary terms marked with a "(*)" are the one that were chosen to be used

from the DBK excerpt to the terms from the different vocabularies. We finally mapped the entities in the left column to the terms that are followed by an asterisk (*). The mapping was done manually. This way it was likely to preserve as much of the semantic richness of the data as possible.

The technical process of the conversion can be conducted by different scripting languages. Since the source data is XML and RDF can also be serialized in XML, it seems likely to use XSL transformations. Hereby, we extracted the entities from the XML we intended to express in RDF and defined an XSLT script, where we specified how the entities should be transformed. Figure 2 provides an example that shows how we have transformed the title entity of an XML file into an RDF representation re-using the Dublin Core property **dcterms:title.**

We can see in Figure 2 that the XML element provides the information about how an entity is encoded. We use this information to make an XSLT script and generate an RDF property. We first identify the entity "title", which is marked purple in the XML. It has a language attribute that is marked orange and a value, which is marked blue. In the XSLT we define a new element with the name "dcterms:title" that has a new attribute with the name "xml:lang" and the value "en". Additionally, the value from the XML element is extracted using XPATH from the path "titleStmt/title/". This results in a new property dcterms:title in RDF that has a language attribute and the value from the XML. This procedure has to be done for every entity in the data catalogue XML. It is very important to note that the example in Figure 2 does not display an entire and valid RDF representation, as it only a single RDF property, without a subject to complete the triple.

Discovering Links to External Data Sources

The rationale for publishing data as Linked Open Data is to increase its visibility and make it easier for secondary users to consume the data, but also to gather information from other data providers who published their data as Linked Open Data. To achieve the latter, we have to identify external data sources that might hold noteworthy data; second, we have to discover links to equivalent resources; and third, we have to include the links in our RDF representation.

The search for external data sources containing further information for the Data Catalogue's study descriptions was performed manually, since currently there is no satisfactory way of searching LOD instances automatically. The data hub[26] Linked Open Data group provided an appropriate set of data sources for this, as it contains all datasets included in the LOD cloud. The first candidate is the Integrated Name Authority File (GND). It originates from the German library community and contains a broad range of elements to describe authorities in detail. This way it aims to solve the name ambiguity problem. Another candidate that might comprise data for enriching the study descriptions is DBpedia. It contains structured information that was extracted from Wikipedia, i.e., the information boxes on the top right corner of many Wikipedia pages. The data comprises information on persons, places, organizations and more.

To discover links to instances from these two external data sources, there are so-called "Link Discovery Tools". One of these tools is Silk – A Link Discovery Framework. It detects relationships between items within different Linked Open Data sources based on various comparison methods that are applied on literal properties of all items. The included comparison methods cover typical similarity measures like Levenshtein distance, Jaccard similarity coefficient, or even geographical distance. Figure 3 displays the general workflow of this procedure, where the relationship is defined as **owl:sameAs** and the comparison method is an absolute string equality measure. If the value of "Property 1" in the initial dataset is equal to the value of "Property 1" in the external dataset, the value of "Property 2" in the initial dataset is equal to the value of "Property 2" in the external dataset, and the value of "Property 3" in the initial dataset is equal to the value of "Property 3" in the external dataset, the both resources are considered to be related to each other in the meaning of **owl:sameAs (**Note http://www.w3.org/TR/owl-ref/#sameAs-def**)**. This relatedness is expressed by a value, which is computed out of the applied similarity measures. As a benefit the

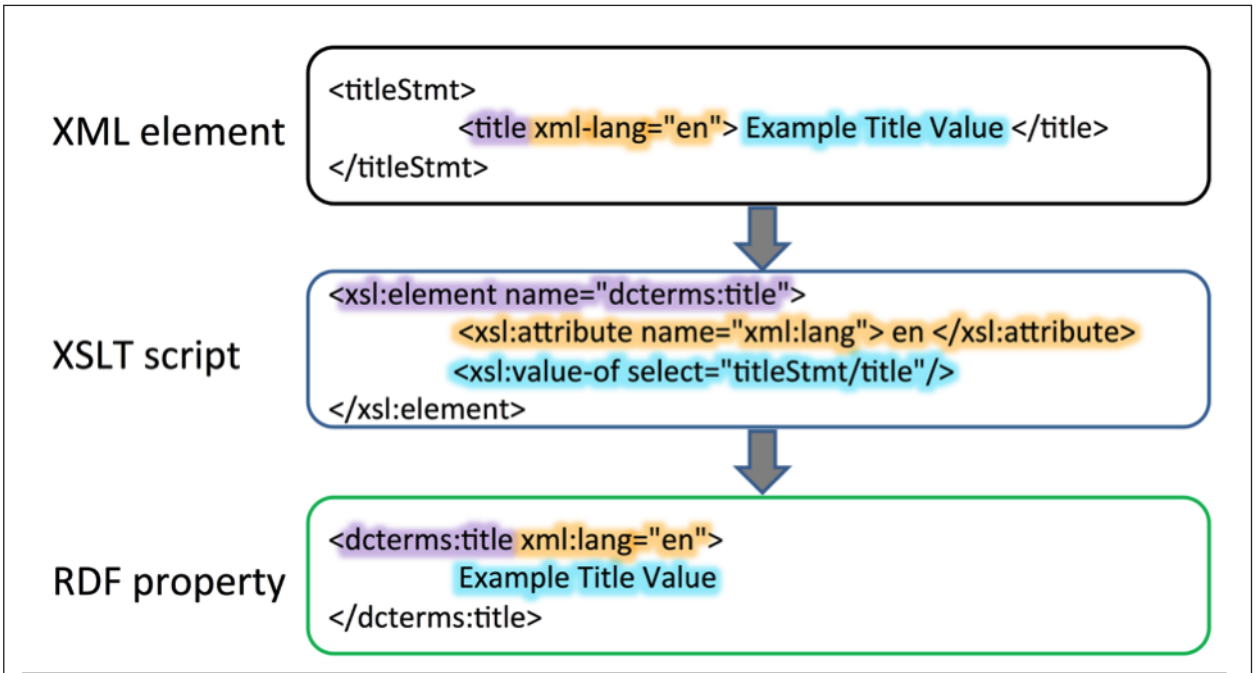

Figure 2: The XSL transformation of the entity "title"

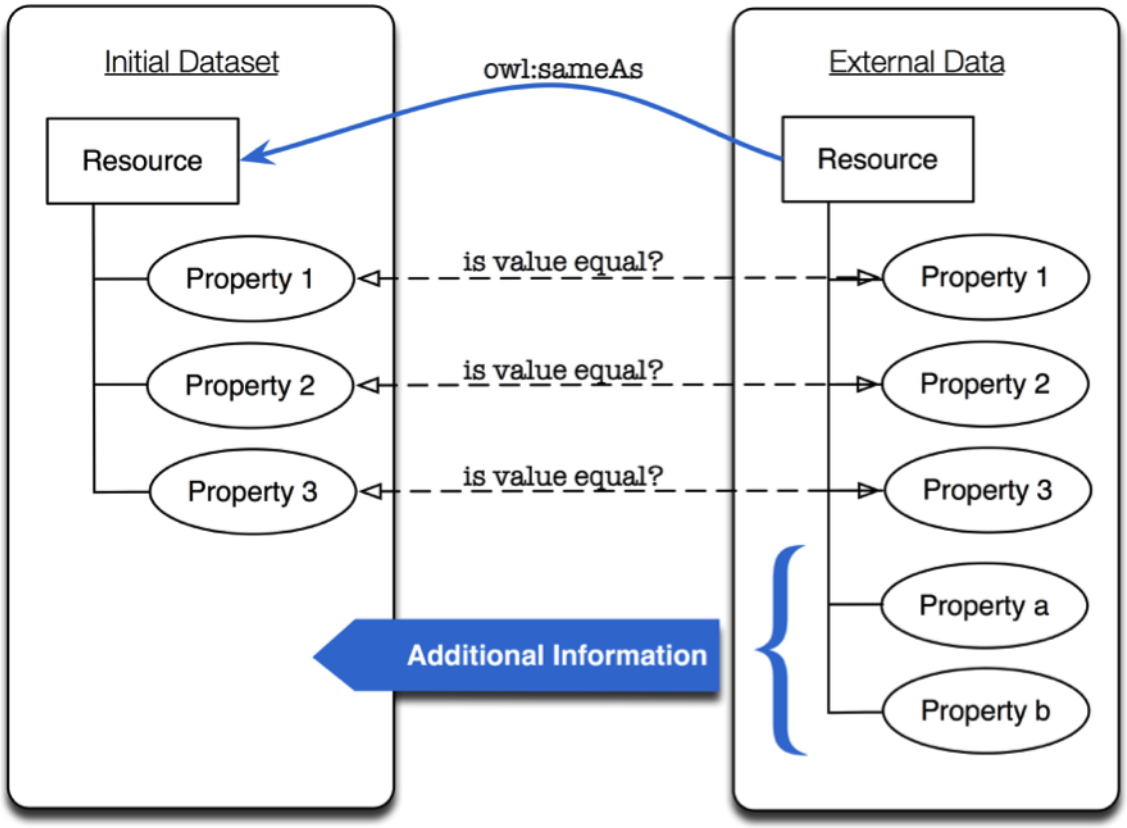

Figure 3: General link discovery procedure with owl:sameAs as defined relationship

properties "Property a" and "Property b" in the external dataset can now be gathered as additional information.

To guide the user through the process of creating link specification for such relationships, Silk provides the "Silk Workbench". The user has to go through three basic steps: (1) specify the data sources and the linking tasks, (2) define explicit linkage rules, and (3) evaluate the correctness of the discovered links. In the following, we will describe the link discovery procedure along an example study description from the Data Catalogue.

For the first step, Silk allows the user to specify several data sources by either providing the SPARQL Endpoint of the data source or its RDF dump that has to be downloaded and stored on the local machine. Figure 4 shows the Data Catalogue and the GND data sources (named PND) that are specified as RDF dumps.

After defining the data sources, it is essential to specify a linking task. The user can also denote an output file, where all results can be saved, but this has to be done for every linking task. A linking task describes what kind of relationship shall be found between two data sources. Therefore, the user has to declare the source dataset, the target dataset, and the link type. Figure 5 illustrates that for our work we have chosen the Data Catalogue as the source dataset and the GND Name Authority file as the target dataset. The link type is set to **owl:sameAs**, as we intend to find equal resources. This is the most common approach to find the resource within an external data source. Data represented in RDF is structured as a graph. The user can add source and target restrictions that specify the node in the RDF graph from which Silk starts to compare the property values. This can be very helpful if the data is very big or if specific concepts should not be part of the comparison. If no restrictions are provided, Silk starts at the root node.

Having defined the linking task along with two data sources, the user comes to the second step and has to define linkage rules that specify how two literal values have to be compared. Hereby, Silk displays a set of all properties used in both data sources the user has specified in the previous step. The user chooses the properties he intends to compare. To accomplish this task, Silk provides an intuitive drag and drop mechanism. Every literal value can also be transformed, e.g., by capitalizing or extracting all numerical values, in order to avoid miss matches due to different encoding schemes. Also, it is possible to select different comparators. For example, the user can choose the comparator that utilizes the Levenshtein distance. This way it is possible to deal with spelling mistakes. Figure 6 illustrates the creation of such a linkage rule. It is shown that for our purpose we selected the property swrc:name from the Data Catalogue and the gnd:preferredNameForThePerson from the GND Name Authority File. Each value of these properties is transformed to lower case. Then each value of swrc:name is compared to each value of gnd:preferredNameForThePerson by applying the Levenshtein distance. The user can also specify other options for the comparator to make the comparison even more precise. Furthermore it is possible to compare several properties with each other. For example, the user could also compare the values of the properties foaf:birthday and gnd:dateOfBirth. This allows the user to define linkage rules such as "Only if the names are the same AND the birthday dates are the same, then the resources should have an owl:sameAs relationship".

The third step comprises the evaluation of the links that Silk has detected between the two specified data sources. As an output, Silk shows the compared values and to which percentage it considers the resources to be related. Figure 7 displays such an output. It is displayed that the comparison is a Levenshtein distance transformed on the input properties **swrc:name** and **gnd:preferredNameForThePerson**. The values "tomka, miklós" and "tomka, miklós" are considered to be a 100% match. Therefore the resources containing these properties are considered to be related in the meaning of **owl:sameAs**.

As a result, the detected link can be included in the initial dataset of our study description and thus enriches it with additional information. According to Figure 7 this would be the link to http://d-nb.info/gnd/134232240 (the person Tomka, Miklós). The entire procedure including all the three steps that were explained in this section has to be done for every concept which is intended to be enriched with additional information. For example, the Data Catalogue comprises descriptions of topical categories of the studies. These are mostly very general terms such as "political attitude" that can be linked to similar terms from DBpedia or various thesauri like the GESIS TheSoz (Zapilko et al. 2012).

data sources we intended to link to, Silk was able to detect links to enrich the data on various entities. The Name Authority File of the German National Library provided a lot of additional information on persons who contributed to a study. Unfortunately, we were not able to gather further information from DBpedia on the topic category of a study, as Silk did not return any links. The same applies to the specification of the data of a study. We intended to link it to an extraction of time events from Wikipedia that is published as LOD (Hienert and Luciano 2012). However, no links were detected, as the dates in the DBK data are encoded as a timespan ("January 2003 to December 2003"), whereas the dates from the extracted Time events are encoded as a point of time ("2003").

Another challenging task was the disambiguation of a person, as the set of the first name, the last name, and the affiliation is simply not unique to certainly identify a person. We did not set the Levenshtein distance very low in order to link resources despite spelling mistakes. Hence, the evaluation of the discovered links took longer than intended to ensure the disambiguation of the persons, and sometimes it was simply impossible.

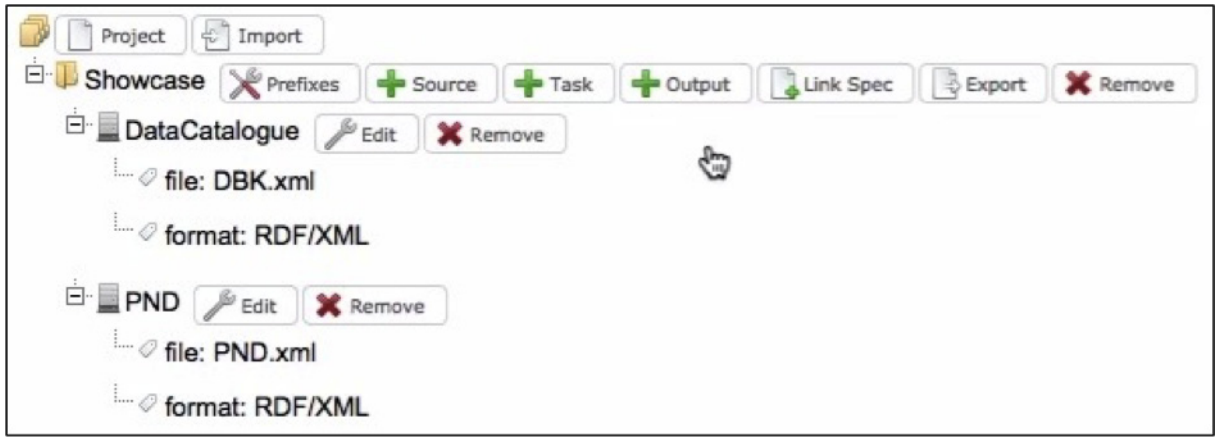

Figure 4: The definition of the Data Catalogue and the GND data sources as RDF dumps in Silk

Results

Based on the entities that we have extracted from the Data Catalogue, the RDF modeling decisions, and the chosen external

The topic category of a study in the Data Catalogue is described with terms from a controlled vocabulary[27]. In RDF the category was first described as a resource that had the terms from the controlled vocabulary as a property. We designed a linkage rule in Silk, which compared the term from the controlled vocabulary

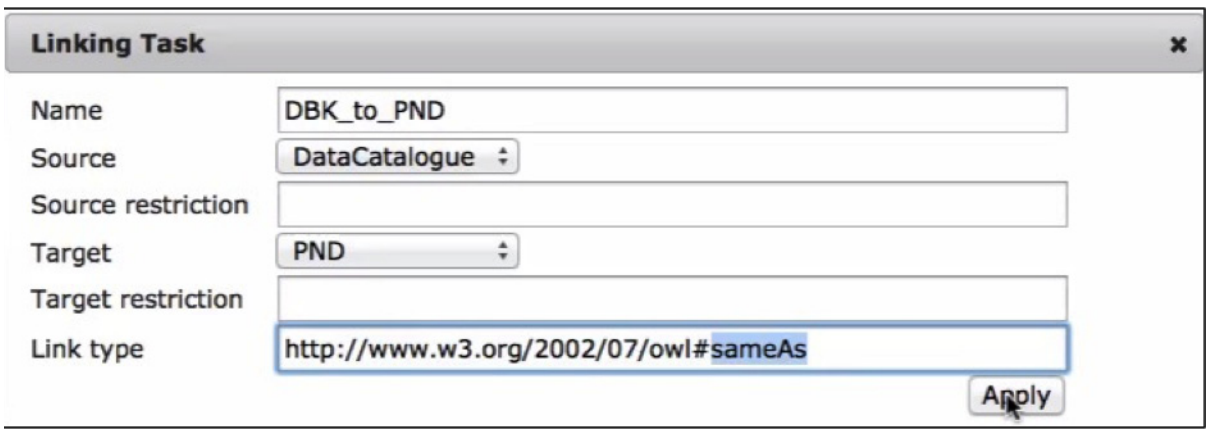

Figure 5: Defining an owl:sameAs link type between the data sources Data Catalogue and the GND

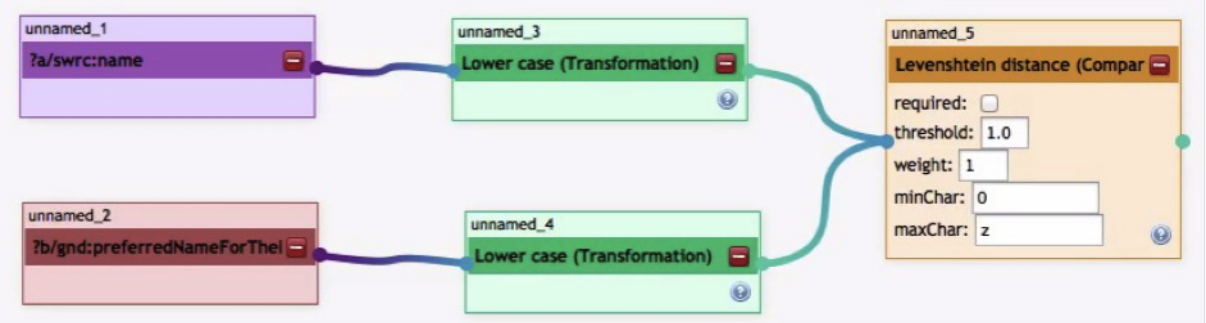

Figure 6: Definition of the linkage rule to compare two property values with the Levenshtein distance

with the labels of articles in DBpedia. For example the category "Income" was supposed to be linked to the DBpedia data of the Wikipedia article about "Income". Silk did not find any links, though. This was due to several reasons. First, there are a lot of articles in Wikipedia that do not have a structured information box. Therefore, there is no DBpedia entry for such articles. Second, some categories are described with multiple terms, like "Legal system, Legislation, Law". DBpedia on the other hand does not describe entities with multiple terms. Therefore, Silk will not find any links between resources that probably describe the same thing, but the comparison of their properties fails due to syntactical difference. To bypass these problems, we have mapped the categories of the study descriptions to concepts of the Thesaurus for the Social Sciences (TheSoz). TheSoz has been already published as Linked Data (Zapilko et al. 2012). This way we were able to gather additional information from the TheSoz such as the translations of the categories in German and French language as well as the hierarchical structure of the categories. For further information we specified the linking rules in Silk to detect links between the concepts of the TheSoz and other thesauri such as EuroVoc[28] . Silk detected these links without any problem providing further information about the categories.

The properties "abstract" and "other references" of a study description were not as helpful for discovering links as we had intended. In order to use the information within these entities, some Natural Language Processing (NLP) algorithms have to be applied to extract keywords and perform a link discover using those keywords. However, this is not part of this work, but can be strongly considered as future work.

Besides discovering links for the authority entities, topic categories, and the date of a study, the properties "title", "alternative title", "producer", and "publisher" were modeled to help to link the study another instance of itself from an external data source. Unfortunately, such a data source was not found on the Linked Open Data cloud. The remaining properties "universe", "data collector", "sample procedure", "collection mode", "access place", "related publication", and "other references" have not been used yet for link discovery and remain as future work.

Conclusion and Discussion

In this work, we have demonstrated how Semantic Web technologies can be used to link study descriptions to external data sources and enrich them with additional information on various entities such as contributors and the categories of the study. We first extracted the entities, which we intended to enrich with further information and several other entities, which seemed to be most promising to help the link discovery process. We transformed the representation of the study description from XML to RDF using XSLT scripts. Hereby, we provided detailed information on the difficulties of such a transformation, especially the mapping of entities to classes and properties from existing vocabularies. We illustrated the workflow of the linking process with the link discovery framework Silk along with an example and provided the results of our work.

For publishing data as Linked Open Data, one has to have good knowledge of RDF as well as the principles and best practices of the modeling and publishing process. "It is especially important to understand whether information should be published as a resource or as a literal, if the intention is to interlink the data with external data sources. In Linked Open Data only resources can be linked together via link types like the **owl:sameAs** statement. Therefore, if the intention is to gather additional information on a specific entity such as the principal investigator, it has to be modeled as a resource containing properties that describe the resource such as "first name" and "last name". For disambiguation purposes, it is strongly advised to use unique identification characteristics such as an ISBN number for books, or ORCID[29] for researchers. Another possibility to disambiguate entities is to use several identification characteristics such as "first name", "last name", "birthplace", and "date of birth". If the aim is to reuse existing vocabularies to express the data, it is important to know which

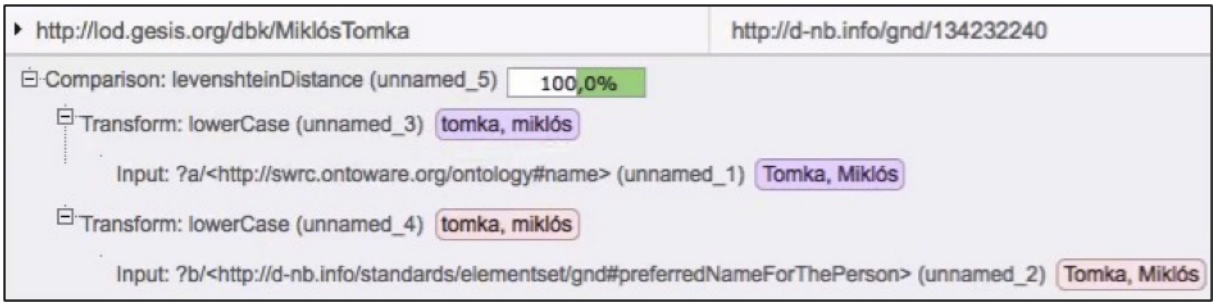

Figure 7: The result from the comparison

vocabularies will fit the best. Resources are modeled as classes, so it is important to investigate several vocabularies to determine if they provide classes that can represent entities as resources in a semantically correct way. The same applies to entities which are intended to be modeled as properties. If the data publisher does not know such vocabularies, the search for them might result in a lot of effort. To help the data publisher to find appropriate terms from existing vocabularies, there are vocabulary search engines like LOV[30] or Swoogle[31] , or novel concepts that recommend classes and properties during the modeling process (Schaible et al. 2013).

To link to external data sources, one has to discover such data sources in the first place, for example, by searching a repository like the data hub. The next step is to understand the structure of the external datasets and locate the concepts of interest for linking and their properties for comparison. In the beginning, this might be time consuming, as datasets are generally modeled differently. However, this is a crucial step because it is necessary to specify the linkage rules in link detection tools like Silk. The setup of datasets, linking task, and linkage rules in Silk is straightforward. Nevertheless, several problems did occur, due to the complexity of the specifications of comparison methods and the not very detailed documentation.

Data from the domain of the social sciences are not very widespread in the Linked Open Data cloud. To find additional information on such type of data is very hard. Once the LOD cloud gets populated with datasets covering social science studies with detail about their contributors, it will be a lot easier to link the GESIS Data Catalogue to these data sources and thereby enrich its study descriptions with additional information. One example for such a domain would be the publications of scientific papers in the area of the Semantic Web, as was discussed by Schaible and Mayr (2012).

References

Bauske, F. (1992), 'Europäische Informationsbasis über Datensätze in CESSDA-Archiven', ZA-Information, vol. 31, pp. 109-111.

Bauske, F. (2000) 'Das Studienbeschreibungsschema des Zentralarchivs', ZA-Information, vol. 47, pp. 73-80.

Bizer, C., Heath T. & Berners-Lee, T. (2009), 'Linked Data-the Story so Far', International Journal on Semantic Web and Information Systems, vol. 4, no. 2, pp. 1–22.

Brank, J., Grobelnik, M. & Mladenić, D. (2005), 'A survey of ontology evaluation techniques', Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD).

Hausstein, B., Zenk-Möltgen, W., Wilde, A. & Schleinstein, N. (2011), 'da|ra Metadatenschema Version 1.0.', GESIS Working Papers 2011/14, doi:10.4232/10.mdsdoc.1.0.

Heath, T. & Bizer, C. (2011), 'Linked Data: Evolving the Web into a Global Data Space', Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 1, no. 1, pp. 1–136.

Hienert, D., Luciano, F. (2012), 'Extraction of historical events from Wikipedia', Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (KNOW@LOD 2012).

Mochmann, E. (1979), 'Bericht über die IASSIST Konferenz in Ottawa', ZA-Information, vol. 4, pp. 24-27.

Schaible, J., Gottron T., Scheglmann S. & Scherp A. (2013), "LOVER: support for modeling data using linked open vocabularies", Proceedings of the Joint EDBT/ICDT 2013 Workshops (EDBT '13), ACM, New York, NY, USA, 89-92, DOI=10.1145/2457317.2457332, http://doi.acm.org/10.1145/2457317.2457332.

Schaible, J., Mayr, P. (2012): "Discovering links for metadata enrichment on computer science papers", GESIS-Technical Reports, 2012/10, Köln: GESIS.

Volz, J., Bizer, C., Gaedke, M. & Kobilarov, G. (2009), 'Discovering and Maintaining Links on the Web of Data', Proceedings of the International Semantic Web Conference (ISWC), pp. 650-665.

Zapilko, B., Schaible, J., Mayr, P. & Mathiak, B. (2012), "TheSoz: a SKOS representation of the thesaurus for the social sciences", Semantic Web: interoperabilty, usability,applicability, DOI: 10.3233/SW-2012-0081.

Zenk-Möltgen, W. & Habbel, N. (2012), 'Der GESIS Datenbestandskatalog und sein Metadatenschema', Version 1.8, GESIS Technical Reports 2012/01.

Notes

1. Johann Schaible Research associate and Ph.D. student at GESIS. Unter Sachsenhausen 6-8, 50667 Köln, Germany. Email: johann.schaible@gesis.org
2. Benjamin Zapilko Research associate and Ph.D. student at GESIS. Unter Sachsenhausen 6-8, 50667 Köln, Germany. Email: benjamin.zapilko@gesis.org
3. Thomas Bosch Research associate and Ph.D. student at GESIS. B2, 1, 68159 Mannheim, Germany. Email: thomas.bosch@gesis.org
4. Wolfgang Zenk-Möltgen Team leader and project manager at GESIS. Unter Sachsenhausen 6-8, 50667 Köln, Germany. Email: wolfgang.zenk-moeltgen@gesis.org

5. http://lod-cloud.net/
6. http://www.w3.org/RDF/
7. http://www.w3.org/TR/rdf-sparql-query/
8. https://dbk.gesis.org/dbksearch/
9. http://dbpedia.org/About
10. https://wiki.d-nb.de/display/LDS
11. http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
12. http://zacat.gesis.org
13. http://cessda.net/Data-Catalogue
14. http://www.sowiport.de
15. http://www.da-ra.de/
16. http://www.datacite.org/
17. https://dbk.gesis.org/dbkfree2.0/
18. http://www.ddialliance.org/
19. http://www.icpsr.umich.edu
20. http://samfund.dda.dk/dda/default-en.asp
21. http://www.nsd.uib.no/nsd/english/index.html
22. http://data-archive.ac.uk/
23. http://dublincore.org/documents/dcmi-terms/
24. http://rdf-vocabulary.ddialliance.org/discovery
25. http://ontoware.org/swrc/
26. http://datahub.io/group/lodcloud
27. https://dbk.gesis.org/dbksearch/Categories.htm
28. http://eurovoc.europa.eu/drupal/?q=node
29. http://orcid.org/
30. http://lov.okfn.org/dataset/lov/
31. http://swoogle.umbc.edu/