IASSIST Quarterly

VOLUME 38 – Number 4 – 2014 & VOLUME 39 – Number 1 – 2015

DDI and Semantic Web

SPECIAL ISSUE

SPECIAL ISSUE
DDI and Semantic Web

ON PAGE **6**
Guest Editors

ON PAGE **60**
Membership Information

Online at: *iassistdata.org/iq*

COLOPHON

IASSIST Quarterly

The **IASSIST Quarterly** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors

The Quarterly is normally published four times per year. Authors are encouraged to submit papers as word processing files (for further information see: *http://www.iassistdata.org/iq/instructions*-authors). Manuscripts should be sent to Editor: Karsten Boye Rasmussen.: Email: *kbr@sam.sdu.dk*

.Announcements of conferences, training sessions, or the like are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event

Editor

Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email: *kbr@sam.sdu.dk*

Deputy editor

Walter Piovesan
Simon Fraser University

Website editor

Stuart MacDonald
University of Edinburgh

In this issue

5 Editor's notes
Karsten Boye Rasmussen

6 Guest Editors' Notes
Joachim Wackerow and Mary Vardigan

7 Semantic Web Applications for the Social Sciences
Thomas Bosch and Benjamin Zapilko

17 DDI-RDF Discovery – A Discovery Model for Microdata
Thomas Bosch, Olof Olssoni, Benjamin Zapilko, Arofan Gregory, and Joachim Wackerow

25 Use Cases Related to an Ontology of the Data Documentation Initiative
Thomas Bosch and Brigitte Mathiak

38 Linking Study Descriptions to the Linked Open Data Cloud
Johann Schaible , Benjamin Zapilko, Thomas Bosch and Wolfgang Zenk-Möltgen

47 XKOS - An RDF Vocabulary for Describing Statistical Classifications
Franck Cotton, Daniel W. Gillman, Yves Jaques

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137

Online @ *http://www.iassistdata.org/iq*

Editor's notes

The Making of Meta-Analysis through Metadata of the Data Documentation Initiative for Semantic Web

Welcome to the fourth issue of Volume 38 and the first issue of Volume 39 in this double-issue of the IASSIST Quarterly (IQ 38:4 & 39:1, 2014 & 2015). This special issue is guest edited by Joachim Wackerow of GESIS – Leibniz Institute for the Social Sciences in Germany and Mary Vardigan of ICPSR at the University of Michigan, USA. They have arranged, participated, herded other DDI experts, as well as produced in several workshops and conferences on the issues of the Data Documentation Initiative (DDI). This special issue on DDI addresses the semantic web. I have included the keyword 'meta-analysis' in the header for this DDI-issue as I believe that is where the combination of DDI and the semantic web is going to make one of its big impacts. I expect precise and detailed DDI metadata with strong applications will bring remarkable support for overview and accumulation of results from large amounts of datasets.

Thanks to the guest editors Joachim Wackerow and Mary Vardigan, this issue includes papers from numerous researchers involved in DDI development and use.

In the overview paper on semantic web applications (Thomas Bosch and Benjamin Zapilko) I came across a term like 'machine-understandable' - i.e., the machine (and software) is capable of understanding. In these days of the Turing-movie 'The Imitation Game' you could say that 'understanding' is a huge part of what the Turing-test investigates when looking into intelligence. However, I believe that we still must be content to stick with 'machine-actionable' – i.e., the machine is capable of taking special actions according to the rising conditions. I will frame the difference as the machine is capable of "if then do" while only humans so far are also capable of acting on "if then don't"!

The second paper (Thomas Bosch, Olof Olsson, Benjamin Zapilko, Arofan Gregory, and Joachim Wackerow) shows how the DDI has in twenty years developed into the support of the complete data lifecycle. In several figures the concepts are overviewed graphically and the use cases illustrate several scenarios of support.

Use cases are also at the center of the paper, including the ontology of the DDI (Thomas Bosch and Brigitte Mathiak) and here less database bound graphics are showing representations of the DDI conceptual model with a focus on the Linked Open Data Cloud and the Web of Data. This paper also exemplifies the query language SPARQL.

The Linked Open Data Cloud is continued in the paper on linking study descriptions (Johann Schaible, Benjamin Zapilko, Thomas Bosch, and Wolfgang Zenk-Möltgen) describing how study descriptions can be enriched with datasets from the Linked Open Data Cloud. The trick is to automatically detect that items within different sources can be successfully linked as they carry the same property. The last paper introduces the reader to XKOS, the eXtended Knowledge Organization System (Franck Cotton, Daniel W. Gillman, Yves Jaques). The paper gives some examples of statistical classification and includes data harmonization. This paper applies the term 'machine-understandable' once again. Semantics is about meaning and meaning is understanding. I still don't think that the Turing-test has been passed - although I might have paid insufficient attention – but the DDI and Semantic Web is going in a promising direction.

Articles for the IASSIST Quarterly are always very welcome. They can be papers from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. When you are preparing a presentation, give a thought to turning your one-time presentation into a lasting contribution to continuing development. As an author you are permitted 'deep links' where you link directly to your paper published in the IQ. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at *http://www.iassistdata.org*.

Authors are very welcome to take a look at the instructions and layout:
http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: *kbr@sam.sdu.dk*. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
July 2015
Editor

Guest Editor's notes

DDI and Semantic Web

This issue focuses on the ways in which DDI can play an important role in the Linked Open Data environment and recent accomplishments that move this idea forward into reality.

The first paper, "Semantic Web Applications for the Social Sciences," presents an overview of several representative applications that use Semantic Web technologies, highlighting social science applications and their benefits for the domain.

The second paper, "DDI-RDF Disco – A Discovery Model for Microdata," describes a data discovery ontology based on DDI that enables users to publish their DDI data and metadata in RDF and link them with many other datasets from the Linked Open Data (LOD) cloud.

The third paper, "Use Cases Related to an Ontology of the Data Documentation Initiative," provides additional detail related to the data discovery ontology Disco, offering several use cases to show its value in the world of Linked Open Data.

The fourth paper, "Linking Study Descriptions to the Linked Open Data Cloud," presents ways to enrich a study description with various datasets from the LOD cloud by exposing selected elements of the study description in RDF.

And finally, the fifth paper, "XKOS - An RDF Vocabulary for Describing Statistical Classifications," offers a brief description of the eXtended Knowledge Organization System (XKOS), an extension of SKOS, and a rationale for why it was developed, showing how the semantics of classification systems in the authors' own offices are represented more faithfully by extending SKOS with XKOS.

Joachim Wackerow - Joachim.Wackerow@gesis.org
Mary Vardigan - vardigan@umich.edu

Semantic Web Applications for the Social Sciences

by Thomas Bosch[1] and Benjamin Zapilko[2]

Abstract

In recent years, Semantic Web technologies have matured and have made their way into various domains – for example, bioinformatics and eGovernment -- where they are used in different applications to provide value-added services for users. In this paper, we present an overview of several representative applications that use Semantic Web technologies and show the potential of these technologies in the Linked Data world. We then present existing Semantic Web applications specifically for the social sciences and highlight their impact on this domain. Integrating the Semantic Web vision into the social sciences results in clear benefits, which we identify and discuss.

Keywords

Semantic Web, Linked Data, Semantic Web Applications, Social Sciences

Introduction

The corporate landscape is moving to the Semantic Web in a big way. Major companies like Adobe, Oracle, IBM, HP, Software AG, GE, Northrop Grumman, Altova, and Microsoft offer (or plan to offer) Semantic Web tools or systems. Others like Novartis, Pfizer, and Telefónica are using the Semantic Web (or are considering using it) as part of their own operations. Active participants in W3C Semantic Web related groups include HP, Agfa, SRI International, Fair Isaac Corp., Oracle, Boeing, IBM, Chevron, Siemens, Nokia, Pfizer, and Eli Lilly. In addition to the corporate sector, we see major communities such as digital libraries, defense sector, eGovernment, the energy sector, financial services, health care, the oil and gas industry, and the life sciences adopting the technologies.

For social science researchers, the Semantic Web and Linked Data hold great promise as Gregory and Vardigan (2010) illustrate in detail. The adoption of semantic technologies has the potential to make the discovery of data and metadata in the Web more efficient, to enhance the reuse of social science metadata, and to decrease the technical barriers to employing data in research. Through Linked Data for the social sciences, users can easily discover the existence of data and can determine how the metadata are structured and whether the data are suitable for their interest. Therefore, it is necessary that the data is well-documented and that quality and provenance are explicit. This will enable the identification of complex relationships such as whether datasets are comparable or whether there are relationships to other versions of the same dataset. Since data collections published as Linked Data can easily be linked with each other, the integration and merging of heterogeneous datasets is also facilitated.

In this paper, we provide an overview of tools and applications that apply Semantic Web technologies and Linked Data. We also present social science projects and applications that make use of semantic technologies, and we discuss the benefits of applying these technologies to the domain of the social sciences.

The number of applications utilizing Semantic Web technologies and Linked Data is increasing and illustrates the potential of these technologies for various domains

Semantic Web Applications

The number of applications utilizing Semantic Web technologies and Linked Data is increasing and illustrates the potential of these technologies for

various domains and heterogeneous communities. In this section of the paper we provide illustrative examples.

HCLS Demo

The W3C Health Care and Life Sciences Interest Group (HCLS) have developed demonstrations of the usage of Semantic Web technologies (W3C 2008a), which Herman (2011) has described. The demonstrations are designed to show the HCLS community how the Semantic Web can be used and to show the Semantic Web community how this Semantic Web technology can be useful in this specific application area. The core of the HCLS demo is the access and the integration of public datasets via the Semantic Web. One of the HCLS demonstrations is the Allen Brain Atlas, which is currently available only through an HTML interface. Mouse brains are cut in slices and stained for the presence of gene expression: 20,000 genes, 400,000 images at high resolution. The Allen Brain Atlas is 'mashed up' with Google maps. Google maps allows the user to upload his or her own 'maps', i.e., the URIs of bitmap images. The user then gets the navigation of large bitmaps for free. The goal of the demonstration is to find the right images in the Atlas data to really provide navigable data.

How is it done? What happens is that the query on the Atlas data is based on scraping the HTML structures, extracting the URI, and using SPARQL to combine these into more complex queries that result in the right image URIs being fed into the Google service. Via RDF, one gets a standard query SPARQL interface for free, providing much power to the end user. If the original authors of the brain Atlas had stored the data in a public MySQL database, one could have achieved the same user interface, but the fact is that they did not. RDF and SPARQL allow one to produce the relevant images easily without interfering with the original data in any way, using standard, off-the-shelf tools. The demo shows that sometimes RDF and SPARQL help in a very simple way -- Semantic Web applications are not necessarily very complex.

NASA

Grove and Schain (2008) describe how to find the right experts at NASA using Semantic Web technologies. The NASA tool is an expertise locater for nearly 70,000 NASA civil servants. The tool uses RDF integration techniques across geographically distributed databases, data sources, and web services.

The authors use internal ontologies/vocabularies to describe the knowledge areas, and a combination of the RDF data and these ontologies to search through the (integrated) databases for specific knowledge expertise. The dump results from a faceted browser developed by the company to view the data on experts.

Semantic MediaWiki

Semantic MediaWiki (2012) is a module of the MediaWiki (2012) software and it extends wikis with ideas from the Semantic Web discipline. Semantic MediaWiki enables the user to make facts available for machines, thus making it easier for humans to search and reuse information. Articles, such as an article about the IASSIST conference in 2014, can be annotated semantically using the RDF format. In this way, RDF triples can be built and stored in the wiki code directly. Relations between subjects and objects can be defined semantically. The IASSIST conference in 2014 would be the subject of the RDF triples stated on the article. It could be specified that the IASSIST conference has participants, presentations, publications, key note speeches (e.g., relation 'hasKeyNoteSpeech'). Authors can also write articles for each relation and for each object to explain the semantics. In the context of this example, authors can write articles about the meaning of the relations to key note speeches and also about each key note speech. On the site about a specific key note speech, the key note speaker could write an article about the content of the presentation.

Humans as well as machines can query all the information included in the wiki sites. People can query information about other articles when they edit new articles. Then the query results appear directly in the wiki site. These articles are updated automatically when the dependent sites are updated. As a consequence, overview articles are always up-to-date and consistent with the detail sites. Visitors of Semantic MediaWikis can download semantically enriched information directly in RDF. A large number of general-purpose RDF tools and specialized external programs can now reuse and process the RDF data in an easy and standardized way. The data and metadata exchange in RDF enables the combination of information from different sources like wikis.

Semantic MediaWiki is free software under the GPL license. More than 150 public wikis use the Semantic MediaWiki extension. In particular, semantic annotations in wikis (e.g., LexWiki, Concept-Hub-Wiki) are adopted in medical and biology sciences to create biomedical terminologies and ontologies collaboratively.

Web Repositories and Server Systems

In this section we present two popular systems for accessing, maintaining, and publishing data on the Web that utilize Semantic Web technologies.

Fedora Repository Project

The Fedora (Flexible Extensible Digital Object Repository Architecture) Repository Project (2013) is an open source software system originally developed by researchers at Cornell University. The underlying architecture enables the storage, management, and access of digital content. Fedora allows for expressing digital objects and relationships among them by assigning so-called 'behaviors' (i.e., services) to them. In addition to a core repository service with well-defined APIs, Fedora includes services for searching, OAI-PMH, messaging, and administrative clients, to name a few.

Regarding Semantic Web technologies Fedora supports interaction with RDF data, since the repository software can be connected with RDF triple stores. Data stored in a triple store can be accessed and used by every service of Fedora.

There are various scenarios and domains dealing with digital content where Fedora is applied. It can be used for "digital collections, e-research, digital libraries, archives, digital preservation, institutional repositories, open access publishing, document management, digital asset management, and more" (Fedora, 2013).

Virtuoso

Virtuoso (2013) is a data server that can be applied to various ways of storing, maintaining, and accessing different kinds of data. The core of Virtuoso is an object-relational SQL database. For serving dynamic web pages, Virtuoso provides a flexible built-in web server, which can process pages written in Virtuoso's own web language (VSP) and other standard languages like PHP or ASP. Virtuoso also provides functionalities for maintaining and managing the published web pages like versioning, automatic metadata extraction, and full text searching. It is also available as an open source edition at http://www.openlinksw.com/wiki/main/.

With respect to semantic technologies Virtuoso currently enables the storing and querying of RDF data in its database. Since this is currently a SQL database, a translation of SPARQL queries into SQL is supported in order to query the RDF data and to provide RDF as output format as well. There are plans to extend the access and storage capabilities of connected databases, which would also enable particular technologies like inferencing on RDF data.

Thesaurus Management Tools

Thesauri and classifications are commonly used instruments for describing and annotating metadata about various kinds of documents. They have a long tradition in libraries and archives. In the following paragraphs we describe two thesaurus management tools that use Semantic Web technologies and datasets.

PoolParty Thesaurus Server

PoolParty Thesaurus Server (PPT) (2013) is a software platform that enables the management and maintenance of complex knowledge models such as taxonomies, thesauri, controlled vocabularies, and similar data. The metadata are fully organized and modeled using W3C´s Semantic Web standards RDF and SKOS, since SKOS focuses particularly on knowledge organization systems. Managing this data inside PPT is enhanced by text mining functionalities and linked data mapping technologies. In addition to processing RDF data, the APIs provided by PPT are also based on semantic technologies, the SPARQL standard of W3C. This allows also an integration of the maintained knowledge models in other systems, e.g., CMS, ERP-Systems, or Wikis. By using complex Semantic Web based approaches like text corpus analysis, entity extraction, linked data enrichment, and SKOS thesaurus management, it is possible to build, maintain, and publish large and complex knowledge models based on RDF data.

VocBench

VocBench (2013) is a vocabulary editing and workflow tool developed by the Food and Agriculture Organization (FAO). The web-based application enables the transformation of multilingual knowledge organization systems like thesauri, authority lists, and glossaries into SKOS/RDF concept schemes. Thus, traditionally maintained thesauri can easily be used in Semantic Web applications. Besides the transformation, VocBench also allows for managing, maintaining, and editing the data. This includes collaborative editing as well as validation and quality assurance tasks. VocBench is an open source project and based on Protegé.

Currently, VocBench is used to manage several datasets held at FAO like the AGROVOC thesaurus (2013), the Biotechnology Glossary (2013), and bibliographic metadata used in FAO. For future releases FAO plans to include a native interface for SKOS and SKOS-XL and configurable support for hosting of different triple store technologies. Also, they plan to support generic OWL ontologies.

Vivo Project

The VIVO project (2013) is an open source Semantic Web application, which was originally developed and implemented at Cornell University. The application maintains profiles of researchers and organizations. These profiles can be populated with additional information like activities or interests. Through extensive search and browse capabilities, it is possible to discover information across institutions and disciplines. Although the VIVO software is installed locally, the different installations worldwide are connected with each other in a network, which also enables integrated searching and browsing across the information of all connected installations. The information that can be discovered can be used in different contexts, e.g., in visualizations or in applications like VIVO Searchlight, which allows to search for VIVO profiles based on textual information from any web page. The open source project is available at http://vivo.sourceforge.net.

The structured data in VIVO is represented in RDF using the VIVO ontology (Mitchell et al., 2011). This ontology focuses on describing researchers and networks of researchers across organizations and disciplines. It also covers researchers' teaching activities, their expertise, their research, and which service activities they provide. The ontology has been developed inside the VIVO project.

Semantic Web Applications for the Social Sciences

Figure 1. SOFISwiki – Search for research institutions

So far, there are just a few Semantic Web applications for the social sciences. We present them in this section in detail. For each Semantic Web application for the social sciences, we show individual benefits for users in the social sciences community regarding Semantic Web supported functionalities.

A preview of these social science applications:

- SofisWiki provides research institutions the possibility to publish information about their research institution, their research activities, and their research projects
- The Microdata Information System (MISSY) is an information system to document German and European studies at the variable and the study level.
- NESSTAR enables publishing of a huge amount of statistical data and metadata using Semantic Web technologies.
- Colectica is a fast way to design, document, and publish survey research using open data standards.

SOFISwiki

Are you conducting a social science research project? Are you writing a thesis or doing advanced work on a social science topic? If so, then you may want to create your project in SOFISwiki and make your research work transparent for the scientific community. GESIS has developed SOFISwiki (2012), the first Semantic MediaWiki in the social science community.

SOFISwiki informs researchers about research activities and projects as well as research institutions in the German-speaking social sciences. SOFISwiki contains all entries from the last ten years of SOFIS (2012), a central database hosted by GESIS that delivers information about over 50,500 research projects. Using SOFISwiki, research institutions like universities can enter and represent information about their research projects and their institutions in convenient forms. Other researchers can get an overview of the wiki content and can also search for and find this information. The collected project information is also available using the social science portal Sowiport (2012). As part of future work, SOFISwiki could be extended to support the documentation of research activities, projects, and institutions all over the world. Figure 1 shows the graphical user interface of SOFISwiki to enable a search for research institutions by institution name, country and location, content orientation, and institution type.

Figure 2 shows the result of the above query for the department 'Monitoring Society and Social Change' of the research institution

GESIS - Leibniz-Institut für Sozialwissenschaften Dauerbeobachtung der Gesellschaft (Mannheim)

Institutionsname:	GESIS - Leibniz-Institut für Sozialwissenschaften Dauerbeobachtung der Gesellschaft
Institutions-Nr.:	069258
Inhaltliche Ausrichtung:	Sozialwissenschaften
Organisationstyp:	andere außeruniversitäre Forschungseinrichtung
Homepage:	http://www.gesis.org/das-institut/wissenschaftliche-abteilungen/dauerbeobachtung-der-gesellschaft/
Straße:	B2,1
Ort:	Mannheim
PLZ:	68072
Postfach:	122155
Land:	Bundesrepublik Deutschland
Email:	christof.wolf@gesis.org
Telefon:	0621 1246-0

Projekte

Zur Zeit sind 12 Projekt(e) für diese Institution vorhanden.

Data without Boundaries (DwB)
Erwerbs- und Betreuungspotenziale von Paaren mit Kindern: Realisierungschancen einer gleichmäßigen Arbeitsteilung
Externe Managementunterstützung zur Erleichterung von Ausgründungsvorhaben (Good Practice) mit dem Ausgründungsvorhaben "Bodymonitor" aus dem GESIS Leibniz-Institut für die Sozialwissenschaften
German Longitudinal Election Study (GLES)

Figure 2. SOFISwiki – Research institution and associated research projects

'GESIS'. Department metadata such as address, homepage, and contact person as well as the associated research projects are delivered.

Benefits for the social sciences community

Research institutions like universities can manage and publish information about their research institutions and their research projects using this tool, and researchers can get an overview of research projects, institutions, and activities. Internally, the metadata are represented using RDF to ease the integration of the metadata items. RDF is also used to integrate SOFISwiki with diverse other portals like Sowiport and SOFIS.

The Microdata Information System (MISSY)

General Description of MISSY

The Microdata Information System (MISSY) (GESIS 2012) maintains the largest household survey in Europe – the German microcensus, which provides statistics about the general population in Germany, including the employment market (occupation, professional education, income, legal insurance). MISSY consists of approximately 500 variables and questions and captures data for 25 years, since 1973. Figure 3 shows the graphical user interface of MISSY. In the detailed view of the variable, the user gets details about the variable "gender" including associated question, values, value labels, and absolute and relative frequencies.

MISSY has two parts:

- Missy Web, the end-user front-end
- Missy Editor for the metadata documentation, which is the back-end

Several use cases are covered by MISSY:

- Thematic classification: variables by thematic classification and year
- Variables by year
- Generated variables by year
- Details of variables with statistics
- Variable-Time Matrix: Variables by thematic classification and year (selectable)
- Questionnaire Catalogue

In the third generation of MISSY, additional surveys such as EU-SILC (European Union Statistics on Income and Living Conditions), EU-LFS (European Union Labour Force Survey), and EVS (European Values Study) will be integrated. The MISSY Editor will be implemented as a web application. In future, it should also be possible to browse variables by survey and by country.

MISSY and Linked Data

The MISSY-specific data model is based on the DDI-RDF Discovery Vocabulary – Disco (Bosch et al. 2012) for the following reasons:

- Disco contains the most salient components of both DDI-Codebook and DDI-Lifecycle for data discovery (as a

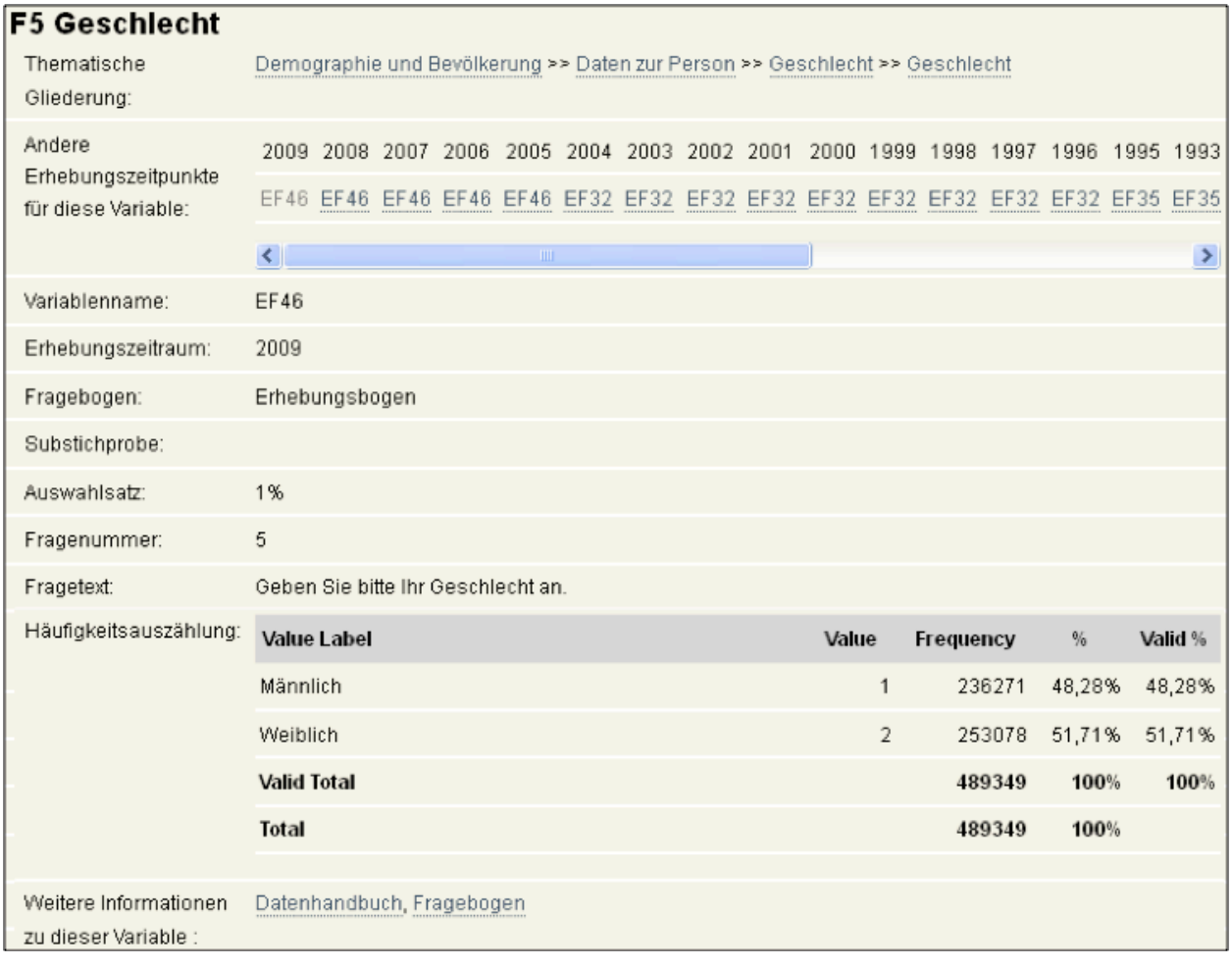


Figure 3. MISSY – Graphical user interface

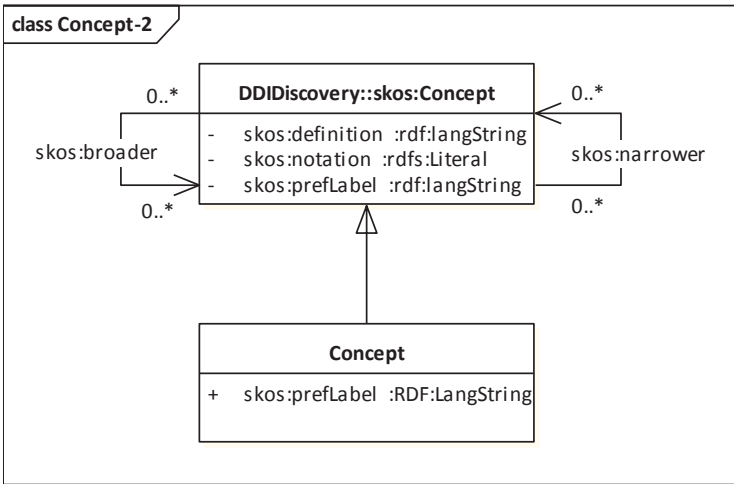


Figure 4. MISSY – Abstract and individual data model

consequence, not all of the over 830 XML elements of DDI 3.1 are covered)

- Disco serves as a first step in developing a model-driven DDI specification to document microdata within the social, behavioral, and economic sciences
- Disco will be officially published and publicly available
- More than 20 experts from the statistics and the Linked Data community of eight different countries have contributed to the development of Disco in three workshops and additional working groups

As not every requirement within the MISSY context is covered by the DDI Ontology, an individual data model is defined on top of the common abstract data model. Other software projects intended to document studies at the study and variable level using DDI-L can also be based on Disco and can reuse existing code, which is made available on a GitHub repository (MISSY 3 2012). To show how this works, an example is provided below. The SKOS (Simple Knowledge Organization System) (W3C 2009), whose purpose is to define hierarchies of concepts, is reused in the Disco ontology to a large extent (see Figure 4).

For instance, codes, categories, DDI concepts, and study subjects are represented in Disco as skos:Concepts. The next figure visualizes skos:Concepts within MISSY. The skos:Concept used in the DDI Discovery Vocabulary is extended by the Concept defined in MISSY. Using the property 'skos:prefLabel', category labels can be stored in an RDF format. The datatype of the category labels is specified as a normal string. One requirement in MISSY is to store category labels in different languages such as English, German, and French. Thus, we have defined the type 'Multilingual' in MISSY. In order to represent category labels, the property 'skos:prefLabel' of the datatype 'Multilingual' is used and therefore the initial common abstract data model is extended.

In a well-defined software architecture, the application itself does not need to know how the data are stored. The application just needs to know the API, i.e., methods that are provided to access and store objects. These methods may be abstracted away from the actual implementation. An actual implementation or strategy can just be a matter of configuration. A strategy is an implementation of the actual type of persistence or physical storage, e.g., DDI-L-XML, DDI-RDF, XML-DB, or Relational-DB. The persistence API defines the persistence functionality for model components regardless of the actual type of physical persistence. Several components implement the persistence functionality defined in the persistence API with respect to the usage of relational DBs, DDI-XML, and DDI-RDF. One concrete implementation of the persistence API is DDI-RDF, the RDF representation of the developed DDI Ontology. MISSY will offer several export formats – one of them will be DDI-RDF.

We will implement two additional concepts providing RDF data. First, MISSY websites will be annotated semantically using RDFa (W3C 2012), which are generic annotations in XHTML documents. Thus, machines can crawl the MISSY websites in order to import exactly the information needed for further processing, since now machines 'know' the meanings of the provided information. RDFa metadata should and will be provided according to the DDI ontology Disco and according to the schema.org vocabulary (Schema.org 2012). Launched in 2011, Schema.org is an initiative from Bing, Google, and Yahoo to provide a vocabulary (a collection of concepts and their properties) to be used by web masters to markup web content in ways recognized by major search providers. Search engines will rely on this markup to improve the display of search results, making it easier for people to find the right pages they search for. The second way of exporting semantic information to the social science community is to build a SPARQL endpoint. RDF triples are stored in a triple store in parallel to the XML documents containing both the data and metadata of multiple studies offered by MISSY.

Benefits for the social sciences community

Other software projects documenting studies on the study and variable level using DDI-L can reuse existing GitHub repository code. MISSY will provide multiple export formats (e.g., DDI-RDF). DDI data as well as metadata can be published in the Linked Open Data cloud. MISSY websites will be annotated using RDFa according to DDI-RDF and schema.org. As a consequence, search engines will improve their query results. By writing SPARQL queries, DDI data and metadata can be accessed from SPARQL endpoints.

NESSTAR

NESSTAR (Norwegian Social Science Data Services 2012) is a Semantic Web application for documenting both statistical data and metadata. NESSTAR can be seen as an extraordinarily easy to understand SW application, as it does not specify sophisticated ontologies, does not use advanced RDF features such as reification, and does not use logical inference.

In 1998, the European Union funded the research and development project called 'Networked Social Science Tools and Resources,' abbreviated as NESSTAR. The EU project with the name 'FASTER' (Faster 2012) followed the goals associated with the NESSTAR project. Assini (2002) gives a rather detailed description of NESSTAR.

The aim of NESSTAR is to make a huge quantity of statistical data and metadata accessible using Semantic Web technologies. Before the implementation of NESSTAR, statistical data as well as metadata was typically only available in a human-readable and understandable form and not additionally in a machine-understandable form that could be further processed by

computer programs. NESSTAR was intended to revolutionize the way people access statistical information, bringing the advantages of instant access to the world of statistical data dissemination. On the Nesstar website (Nesstar 2013), a list of Nesstar catalogues (e.g., surveys, tables) is provided by the Nesstar's Demo Server. Figure 5, for example, shows information such as the associated question, the values and the categories, summary statistics, interviewer instructions, and the total responses about the variable gender of a demo survey.

Conceptual model of NESSTAR

The NESSTAR object model is defined in RDFS. About 15 classes represent the key domain-specific concepts within the statistical domain like studies, data files, variables, indicators, and tables. The conceptual model also includes relationships between domain-specific concepts: studies, for example, may contain cubes having one or more dimensions. Additionally, 10 domain independent support classes are part of the object model of statistical data and metadata. The class Server, for instance, represents the server where the metadata objects are hosted. It provides basic administrative functionality such as file transfer, server reboot, and server shutdown. Starting from a server and by recursively traversing the objects' relationships, applications can reach all the server's objects. The domain independent support class Catalog groups metadata objects. Instances of Catalogs can be browsed and you can get a list of all the metadata objects which are included in the Catalog. There is also the possibility to search for particular objects.

Many research studies contain sensitive information that cannot be made available without restrictions. Within NESSTAR, access control policies can be defined in order to follow a security model. To implement this, classes such as User, Role (e.g., administrator, final user, data publisher), and Agreements (e.g., 'I agree to use this data only for non-commercial research purposes') are specified.

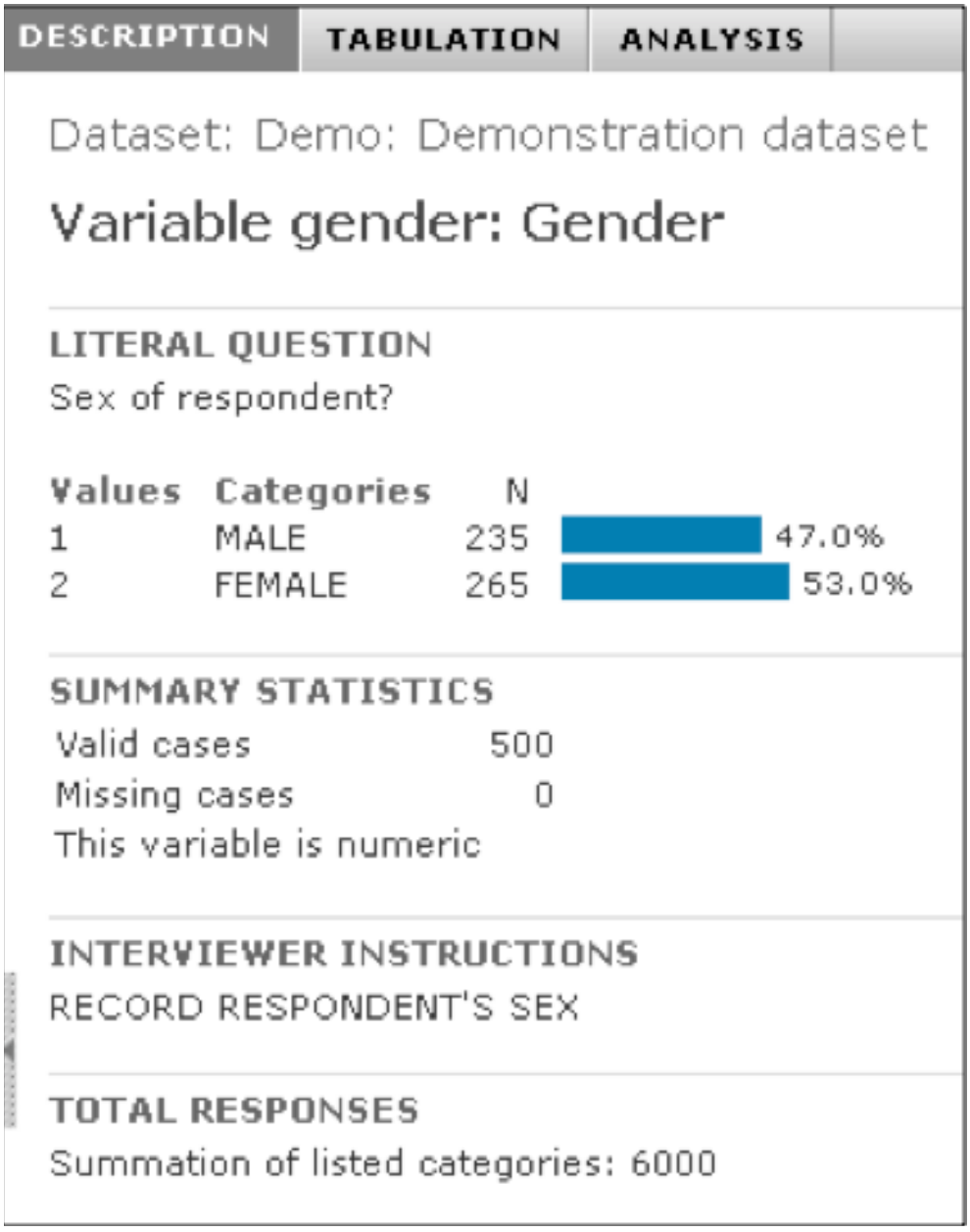


Figure 5. NESSTAR – variable gender

NESSTAR is based on lightweight, object-oriented web middleware named NEOOM - NEsstar Object Oriented Middleware. NEOOM is based on Web and Semantic Web standards like HTML, HTTP, RDF, and RDFS. NEOOM is a set of guidelines on how to use web as well as Semantic Web technologies to build distributed object-oriented systems. The NEOOM guidelines are described extensively in Assini (2001b) and very briefly in Assini (2001a). RDFS does not provide a way to describe the behavior, the operations of statistical objects (e.g., queries, statistical operations, file transfers, tabulate, and frequency). How to specify the operations formally? In the NEOOM Object Model, specific methods (e.g., Login) are defined as sub-classes of the Method class. Concrete method invocations are then instances of the Method class.

Behavioral view on NESSTAR

According to the NESSTAR conceptual model, data publishers make their statistical data and metadata available on the web as objects. These objects are represented by RDF resources according to the NESSTAR object model. Each data publisher runs its own server, which is an instance of the class Server. NESSTAR

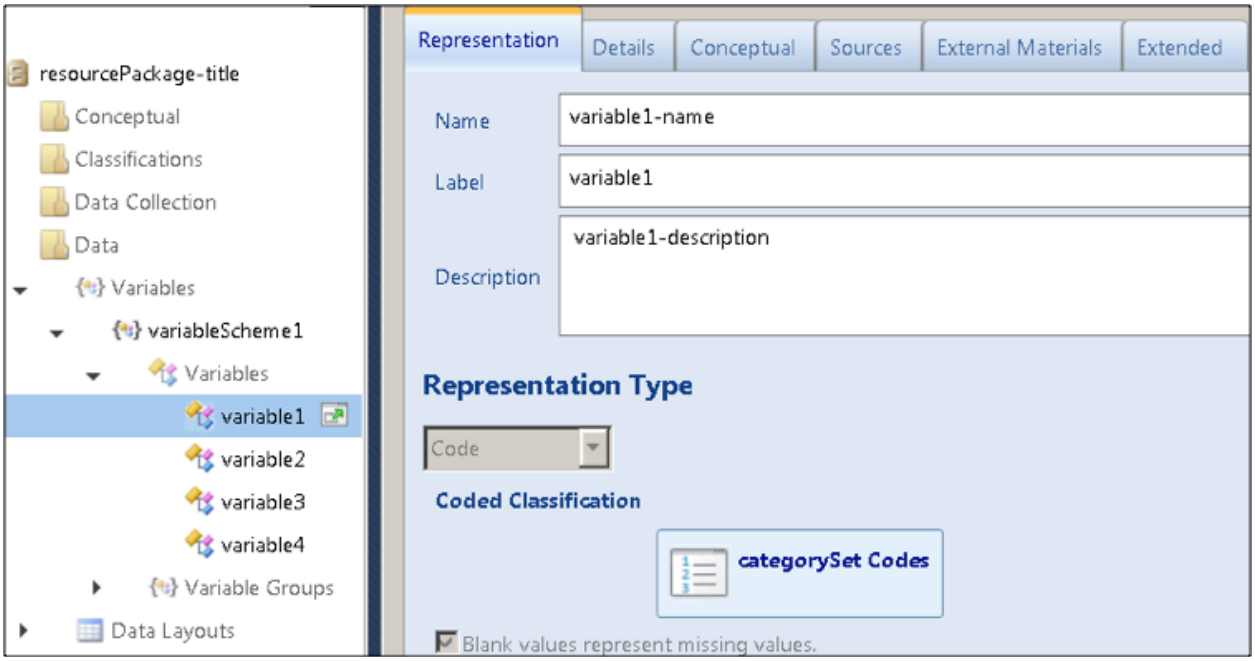


Figure 6. Colectica – Variables

servers host the maintained objects. NESSTAR servers provide WWW resources such as HTML pages and images as well as statistical objects. NESSTAR is fully distributed and each server is totally independent and integrated. Users have the possibility to access statistical objects remotely by simply typing objects' URLs. SOAP (W3C 2007) is used for remote object-oriented calls. Similar to using search engines like Google, users can search for remote statistical objects: they could for example type the search term 'find all variables about political orientation'. In NESSTAR, there are different kinds of user access possibilities: NESSTAR Explorer,

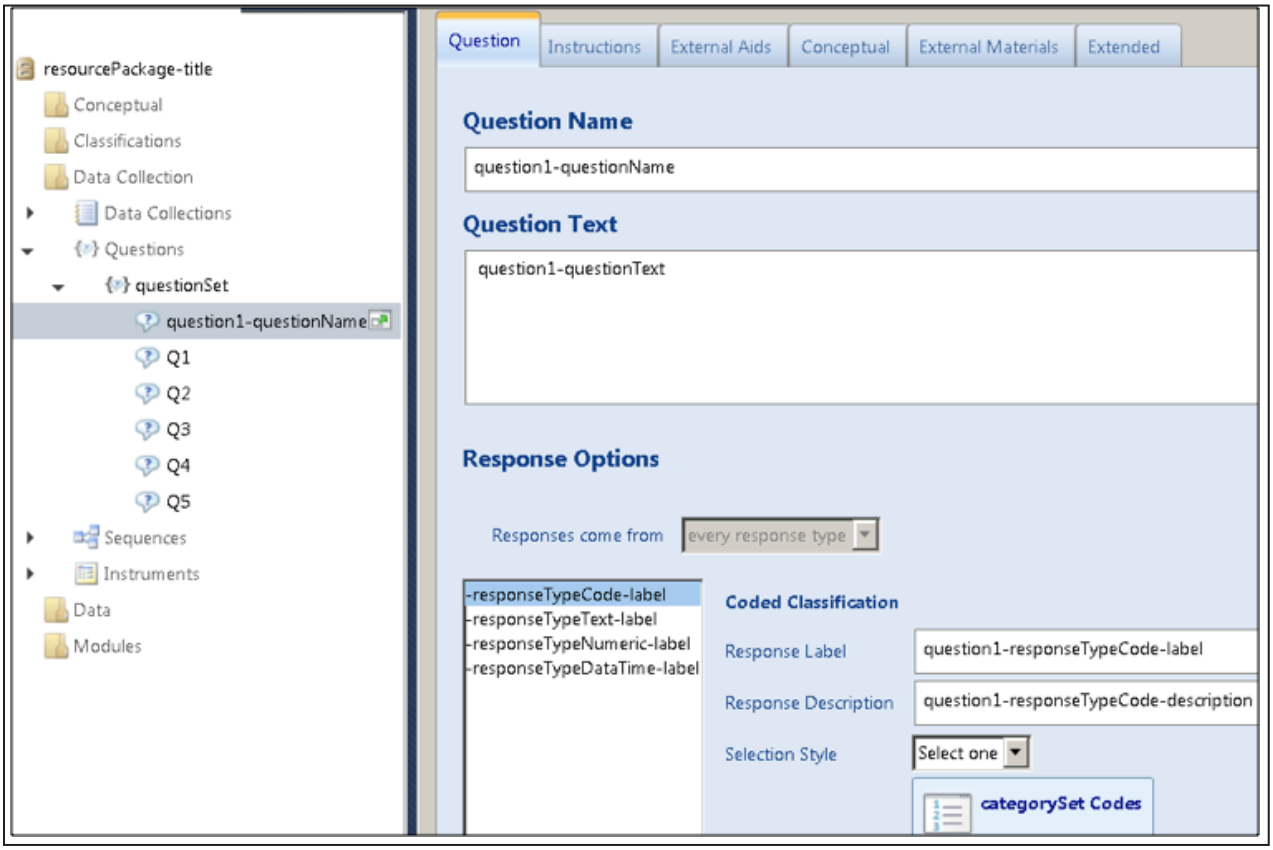


Figure 7. Colectica – Questions

NESSTAR Light Explorer, NESSTAR Publisher, and Object Browser. NESSTAR Explorer is similar to a common web browser. Users can enter objects' URLs and WWW resources are displayed as they would be displayed in a web browser. NESSTAR Publisher is a tool for editing metadata, for validating, and for publishing. The Object Browser's purpose is to test and to administer statistical objects.

Benefits for the social sciences community

NESSTAR enables users to publish a huge amount of statistical data and metadata using Semantic Web technologies. Now statistical data and metadata is not only available in a human-readable and understandable form but also in a machine-understandable form which can be further processed.

Colectica RDF Services

Colectica is a fast way to design, document, and publish survey research using open data standards. The Colectica Platform provides features for statistical agencies, survey research groups, public opinion researchers, data archivists, and other data intensive operations. Colectica can increase the expressiveness and longevity of the data collected through standards-based metadata documentation (Colectica 2012c). DDI-L allows for reuse and harmonization of metadata items through the use of referencing. With the Colectica 4.0 Repository Addin, the relationships between metadata items are indexed (Colectica 2012b), which makes it possible to execute queries on these relationships. Figure 6 displays the documentation of variables using Colectica Designer. You can specify variable names, variables labels, variable descriptions, the response unit, associated concepts and universes, and the representation of the variable (e.g., numeric, textual, or coded representation).
Figure 7 shows how to document questions using Colectica Designer. For questions you can specify question names, question text, the question scheme, and the response domain (e.g., numeric, textual, or coded).

The Colectica RDF model is created by hand based on the Colectica DDI-L model. Each description of a DDI metadata item is stored as a named graph. The RDF Services Architecture can be deployed with Colectica Repository. All DDI metadata items, which are stored and versioned in the Repository, are also stored in RDF (Colectica 2012a). Several external vocabularies such as RDF, RDFS, simple Dublin Core (DC), the DCMI Metadata Terms (DCTERMS), OWL, XSD, and FOAF are reused (Colectica 2012b).

SPARQL (W3C 2008b) is a query language created for searching RDF data and is standardized by the W3C. It allows for searching based on the relationships and literal data stored in an RDF graph or store. SPARQL can be used to construct very precise questions about DDI metadata items referencing multiple metadata items. There are two deployment scenarios which can be distinguished in Colectica: internal RDF stores and external RDF stores. Using Colectica Repository's internal RDF store, SPARQL 1.0 as well as the draft version 1.1 of SPARQL are supported. The SPARQL Update functionality is disabled in order to maintain consistency with the versioned DDI metadata items in the repository. For Colectica Repository, it is also possible to replicate the RDF to external already existing RDF stores, which is the second deployment scenario (Colectica 2012a).

One can query DDI-L as RDF either using a web service from Colectica Repository or using a SPARQL endpoint on Colectica Web. In addition, each DDI-L metadata item, which is stored in the Colectica Repository, can be downloaded as an RDF dump (Colectica 2012a). One example of such a SPARQL query in the statistical domain could be: Which studies has Dan Smith - the software developer of Colectica - authored since the beginning of 2010 (Colectica 2012a)?

```
PREFIX ddi: <urn:ddirdf:>
PREFIX ddit: <urn:ddirdf:type:>
PREFIX dc: <http://purl org/dc/elements/1.1/>
SELECT ?study
WHERE {
        ?study a ddit:StudyUnit;
        dc:date ?creation_date;
        dc:creator <http://dan.smith.name/who#dan>.
        FILTER (xsd:dateTime(?creation_date) > "2010-01-01
        00:00:00"^^xsd:dateTime ) . }
ORDER BY ?study
```

Another example of a SPARQL query would be: How many times has a variable been reused across multiple datasets (Colectica 2012a)?

```
PREFIX ddi: <urn:ddirdf:>
PREFIX ddit: <urn:ddirdf:type:>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?variable COUNT (?parent) AS c
WHERE {
        ?variable a ddit:Variable ;
        ?parent ddi:HasVariable ?variable .
        ?parent a ddit:Dataset. }
GROUP BY ?variable
```

Dan Smith's website (Colectica 2012b) also offers several examples of DDI-RDF serializations. A further SPARQL example from the DDI-L US 2010 Census sample file is also provided. As part of future work, predicates will be updated when the official and community adopted DDI Discovery Vocabulary is available (Colectica 2012a).

Benefits for the social sciences community

DDI-L data and metadata can be queried as RDF using the Colectica Repository web service or the Colectica Web SPARQL endpoint. DDI-L metadata items, stored in the Colectica Repository, can also be downloaded as RDF dumps.

Conclusion and Future Work

We have presented several representative applications that apply Semantic Web technologies to a high degree. While semantic technologies and Linked Data have yet not been widely used in the social sciences, we have identified initial applications exclusively developed for this domain. The impact of Semantic Web and Linked Data are exposed in these applications. Additional potentials and benefits for an adaption of semantic technologies for scientific purposes can easily be identified. We have shown individual benefits for users of the social sciences community regarding Semantic Web functionalities.

References

AGROVOC 2013 AGROVOC Thesaurus, viewed 6 May 2015, < http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus >

Assini, P 2001a 'Objectifying the Web the 'light' way: an RDF-based framework for the description of Web objects', Proceedings of the International World Wide Web Conference, Hong Kong, 01 Mai 2001, tenth International World Wide Web Conference.

Assini, P 2001b NEOOM: A Web and Object Oriented Middleware System, [Online], Available: http://www.nesstar.org/sdk/neoom.pdf [6 May 2015].
Assini, P 2002, 'A Semantic Web Application for Statistical Data and Metadata', Proceedings of the International World Wide Web Conference, Hawaii, 07 May 2002, 11th International World Wide Web Conference.
Biotechnology Glossary 2013 Biotechnology Glossary, viewed 6 May 2015, <http://www.fao.org/biotech/biotech-glossary/en/>
Bosch, T, Cyganiak, R, Wackerow J & Zapilko B 2012 'Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences', Proceedings of the International Conference on Dublin Core and Metadata Applications, Kuching, 03 September 2012, International Conference on Dublin Core and Metadata Applications, pp46 - 55.
Colectica 2012a Accessing DDI 3 as Linked Data: Colectica RDF Services, viewed 6 May 2015,
<http://www.iassistdata.org/conferences/2012/presentation/3326 >.
Colectica 2012b DDI 3 meets RDF and SPARQL with Colectica Repository, viewed 6 May 2015, <http://dan.smith.name/2011/10/ddi-3-meets-rdf-and-sparql-with-colectica-repository/>.
Colectica 2012c Colectica Website, viewed 6 May 2015, <http://www.colectica.com/>.
Faster 2012 Faster, viewed 6 May 2015, <http://fasterproject.eu/>.
Fedora 2013 Fedora Repository Project, viewed 6 May 2015, <http://fedora-commons.org/>.
GESIS 2012 The Microdata Information System (MISSY), viewed 6 May 2015, < http://www.gesis.org/missy/>.
Gregory, A & Vardigan, M 2010 The Web of Linked Data: Realizing the Potential for the Social Sciences, viewed 6 May 2015, < http://odaf.org/papers/201010_Gregory_Arofan_186.pdf>
Grove, M & Schain, A 2008 Case Study: POPS — NASA's Expertise Location Service Powered by Semantic Web Technologies, viewed 6 May 2015, <http://www.w3.org/2001/sw/sweo/public/UseCases/Nasa/>.
Herman, I 2011 Semantic Web Adoption and Applications, viewed 6 May 2015, <http://www.w3.org/People/Ivan/CorePresentations/Applications/>.
MediaWiki 2012 MediaWiki.org, viewed 6 Max 2015, <http://www.mediawiki.org/wiki/MediaWiki>.
Mitchel, S, Chen, S, Ahmed, M, Lowe, B, Marks, P, Rejack, N, Corson-Rikert, J, He, B, Ding, Y 2011 'The VIVO Ontology: Enabling Networking of Scientists' ACM WebScience Conference, Koblenz
MISSY 3 2012 MISSY 3 project, viewed 6 May 2015, <https://github.com/missy-project>.
Nesstar 2013, Welcome to Nesstar's Demo Server, viewed 6 May 2015, < http://nesstar-demo.nsd.uib.no/webview/>.
Norwegian Social Science Data Services 2012 Nesstar, viewed 6 May 2015, <http://www.nesstar.com/>.
PPT 2013 PoolParty Thesaurus Server, viewed 6 May 2015, < http://www.poolparty.biz/>
Schema.org 2012 Schema.org, viewed 6 May 2015, <http://schema.org/>.
Semantic MediaWiki 2012 Semantic MediaWiki, viewed 6 May 2015, <http://semantic-mediawiki.org>.
SOFIS 2012 SOFIS - Social Science Research Information System, viewed 6 May 2015, <http://www.gesis.org/en/services/research/sofis-social-science-research-information-system/>.
SOFISWiki 2012 SOFISWiki, viewed 6 May 2015, <http://www.gesis.org/sofiswiki/Hauptseite>.
Sowiport 2012 Sowiport, viewed 6 May 2015, < http://sowiport.gesis.org/ >.
Virtuoso 2013 Virtuoso Universal Server, viewed 6 May 2015, <http://virtuoso.openlinksw.com/>
VIVO 2013 VIVO Project, viewed 6 May 2015, <http://www.vivoweb.org/>
VocBench 2013 VocBench, viewed 6 May 2015, < http://aims.fao.org/vest-registry/tools/vocbench-2>
W3C 2007 SOAP Version 1.2 Part 0: Primer (Second Edition) - W3C Recommendation 27 April 2007, viewed 6 May 2015, <http://www.w3.org/TR/2007/REC-soap12-part0-20070427/>.
W3C 2008a HCLS/Banff2007Demo, viewed 6 May 2015, <http://www.w3.org/wiki/HCLS/Banff2007Demo>.
W3C 2008b, SPARQL Query Language for RDF, viewed 6 May 2015, <http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/>.
W3C 2009 SKOS Simple Knowledge Organization System Namespace Document - HTML Variant - 18 August 2009 Recommendation Edition, viewed 6 May 2015, <http://www.w3.org/2009/08/skos-reference/skos.html>.
W3C 2012 RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - W3C Working Group Note 07 June 2012, viewed 6 May 2015, <http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/>.

Notes

1.Thomas Bosch GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany E-Mail: thomas.bosch@gesis.org

2. Benjamin Zapilko GESIS - Leibniz Institute for the Social Sciences, Köln, Germany E-Mail: benjamin.zapilko@gesis.org

DDI-RDF Discovery – A Discovery Model for Microdata

by Thomas Bosch[1], Olof Olsson[2], Benjamin Zapilko[3], Arofan Gregory[4], and Joachim Wackerow[5]

Abstract
Ontology engineers and experts from the social, behavioral, and economic sciences developed a data discovery ontology covering a subset of both the DDI Codebook and Lifecycle models, and implemented a rendering of DDI XML instances to RDF (Resource Description Framework). The main goals associated with the design process of the DDI ontology were to reuse widely adopted and accepted ontologies like Dublin Core (DC) and Simple Knowledge Organization System (SKOS) and also to define meaningful relationships to the RDF Data Cube vocabulary. Now, organizations have the possibility to publish their DDI data and metadata in RDF and link it with many other datasets from the Linked Open Data (LOD) cloud. As a consequence, a huge number of related DDI instances can be discovered, queried, connected, and harmonized. The combination of DDI metadata (as well as data) from several organizations, based on this RDF discovery (Disco) vocabulary, will enable powerful derivations of implicit knowledge out of explicitly stated pieces of information.

Keywords: Semantic Web, Linked Data, Ontology Design, DDI

Data Documentation Initiative: Background

Overview

The DDI specification describes social science data, data covering human activity, and other data based on observational methods measuring real-life phenomena. DDI supports the entire research data lifecycle. DDI metadata accompany and enable data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage data (NISO Press, 2004). DDI does not invent a new model for statistical data. It formalizes state of the art concepts and common practice in this domain. DDI focuses on both microdata and aggregated data. It has its strength in microdata -- data on the characteristics of units of a population, such as individuals or households, collected by, for example, a census or a survey. Statistical microdata are not to be confused with microdata in HTML, an approach to nest semantics within web pages. Aggregated data (e.g., multidimensional tables) are likewise covered by DDI. They provide summarized versions of the microdata in the form of statistics like means or frequencies. Publicly accessible metadata of good quality are important for finding the right data. This is especially the case if access to microdata is restricted due to potential risk of disclosure of respondent identities. DDI is currently specified in XML Schema, organized in multiple modules corresponding to the individual stages of the data lifecycle, and includes over 800 elements (DDI Lifecycle).

A specific DDI module (using the simple Dublin Core namespace) allows for the capture and expression of native Dublin Core elements, used either as references or as descriptions of a particular set of metadata. This is used for citation of the data, parts of the data documentation, and external material in addition to the richer, native DDI. This approach supports applications that understand the Dublin Core XML, but do not understand DDI. DDI is aligned with other metadata standards as well, with SDMX[6] (time-series data) for exchanging aggregate data, ISO/IEC 11179 (metadata registry) for building data registries such as question, variable, and concept banks (ISO/IEC,

> DDI has its strength in the domain of social, economic, and behavioral data

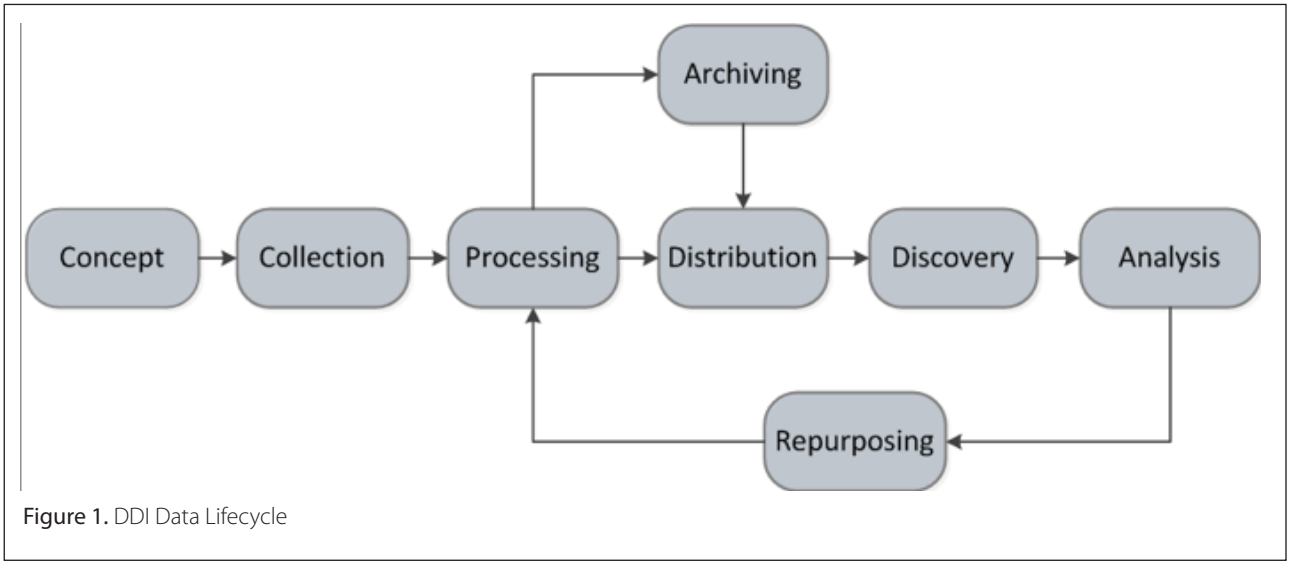


Figure 1. DDI Data Lifecycle

2004), and ISO 19115 (geographic standard) for supporting GIS (geographic information system) users (ISO 19115-1:2003, 2003).

Goals

DDI supports technological and semantic interoperability in enabling and promoting international and interdisciplinary access to and use of research data. Structured metadata with high quality enable secondary analysis without the need to contact the primary researcher who collected the data. Comprehensive metadata (potentially along the whole data lifecycle) are crucial for the replication of analysis results in order to enhance research transparency. DDI also enables the reuse of metadata of existing studies (e.g., questions, variables) for designing new studies, an important ability for repeated surveys and for comparison purposes. DDI supports researchers who follow the above mentioned goals.

DDI Users

A large community of data professionals, including data producers (e.g., of large, academic international surveys), data archivists, data managers in national statistical agencies and other official data producing agencies, and international organizations use the DDI metadata standard. The DDI Alliance hosts a comprehensive list of projects using the DDI[7]. Academic users include the UK Data Archive at the University of Essex[8], the Dataverse Network at the Harvard-MIT Data Center[9], and the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan[10]. Official data producers in more than 50 countries include the Australian Bureau of Statistics (ABS)[11] and many national statistical institutes of the Accelerated Data Program for developing countries[12]. Examples of international organizations using DDI are UNICEF, the Multiple Indicator Cluster Surveys (MICS)[13], The World Bank[14], and The Global Fund to Fight AIDS, Tuberculosis and Malaria[15].

DDI History and Versions

The DDI project, which started in 1995, has steadily gained momentum and evolved to meet the needs of the social science research community. In 2003, the DDI Alliance was established to develop and promote the DDI specification and associated tools, education, and outreach program. The DDI Alliance is a self-sustaining membership organization whose institutional members have a voice in the development of the DDI specification. To ensure continued support and ongoing development of the standard, DDI has been branched into two separate development lines. DDI-Codebook (formerly DDI2) is a more light-weight version of the standard, intended primarily to document simple survey data for archival purposes. Encompassing all of the DDI-Codebook specification and extending it, DDI-Lifecycle (formerly DDI3, first version published in 2008) is designed to document and manage data across the entire data lifecycle, from conceptualization to data publication and analysis and beyond.

Data Lifecycle

The common understanding is that both statistical data and metadata are part of a data lifecycle (Figure 1 displays this lifecycle -- it is described in more detail on the DDI Alliance website[16]). Multiple institutions are involved in the data lifecycle, which is an interactive process with multiple feedback loops. Data documentation is a process, not an end condition where a final status of the data is documented. Rather, metadata production should begin early in a project, and metadata should continue to be captured at the source as data come into being. The metadata can then ideally be reused along the data lifecycle. Such practice would incorporate documentation as part of the research method (Jacobs et al., 2004). A paradigm change would be enabled: on the basis of the metadata, it becomes possible to drive processes and generate items like questionnaires, statistical command files, and web documentation, if metadata creation is started at the design stage of a study (e.g., survey) in a well-defined and structured way.

Limitations

DDI has its strength in the domain of social, economic, and behavioral data. Ongoing work focuses on the early phases of survey design and data collection as well as on other data sources like register data. The next major version of DDI will incorporate the results of this work. It will be opened to other data sources and to data of other disciplines.

Related Work

With respect to documenting data, there are several relevant metadata standards like SDMX (Statistical Data and Metadata Exchange) for the representation and exchange of aggregated data, ISO 19115 (ISO 19115-1:2003, 2003) for geographic information, and PREMIS[17] for preservation purposes. The metadata registry standard ISO 11179 (ISO/IEC, 2004) addresses the modeling

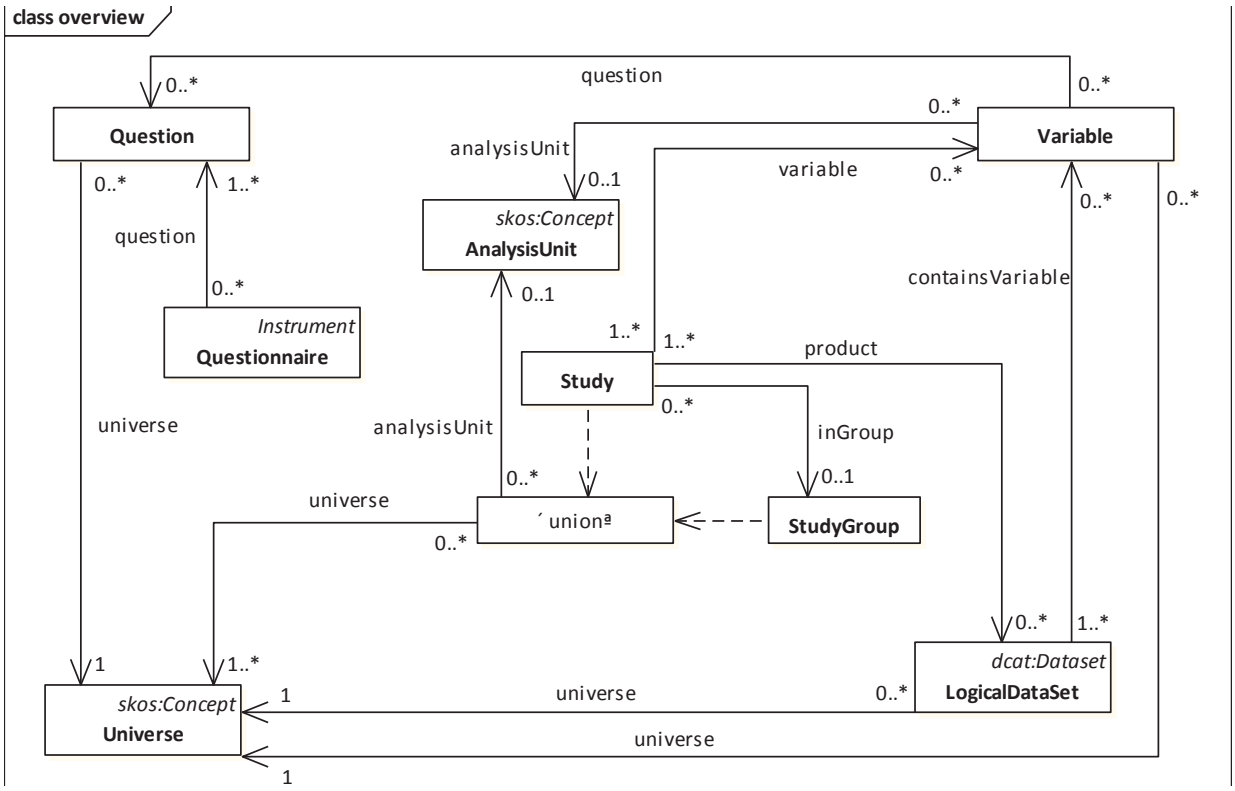


Figure 2. DDI-RDF Discovery Vocabulary (Disco)

of metadata, e.g., reference models, and registries. However, there are as yet few adequate RDF-based vocabularies for documenting data. DDI-RDF for discovery, or Disco, has a clearly defined focus on describing microdata, which has not been covered to this extent by other established vocabularies yet. Therefore it fits well alongside other metadata standards on the web and can clearly be distinguished. Connection points to classes or properties of other vocabularies ensure equivalent or more detailed possibilities for describing entities or relationships.

An RDF expression of the Simple Dublin Core specification exists which could be used for citation purposes (DCMI, 2008). Furthermore, the DCMI Metadata Terms (DCMI, 2010) have been applied when suitable for representing basic information about publishing objects on the web as well as for hasPart relationships. For representing concepts that are organized in ways similar to thesauri and classification systems, classes and properties of Simple Knowledge Organization System (SKOS)[18] have been used. Some aspects of DDI-RDF are already similarly represented in other metadata vocabularies, e.g., data management and documentation. The vocabulary of interlinked datasets (VoID)[19] represents relationships between multiple datasets, while the Provenance Vocabulary[20] provides the possibility to describe information on ownership and can be used to represent and exchange provenance information generated in different systems and under different contexts. In this context, a study can be seen as a data-producing process and a logical dataset as its output artifact. Data Catalog Vocabulary (DCAT)[21] is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs.

An established RDF metadata vocabulary, which seems similar to DDI-RDF at first glance, is the RDF Data Cube vocabulary (Cyganiak et al., 2010). This model maps the SDMX information model to an ontology and is therefore compatible with the cube model that underlies SDMX. It can be used for representing aggregated data (also known as macrodata) such as multidimensional tables. Aggregate data are data derived from microdata by statistics on groups or aggregates, such as counts, means, or frequencies. A dataset presented with the Data Cube vocabulary consists of a set of values organized along a group of dimensions, which is comparable to the representation of data in an Online Analytical Processing system. In the Data Cube vocabulary associated metadata are added.

DDI as Linked Data

Statistical domain experts (core members of the DDI Alliance Technical Implementation Committee, representatives of national statistical institutes, national data archives) and Linked Open Data community members have chosen the DDI elements that are seen as most important to solve problems associated with diverse identified use cases around data discovery. Widely accepted and adopted vocabularies are reused to a large extent. There are features of DDI that can be addressed through other vocabularies, such as: describing metadata for citation purposes using Dublin Core, describing aggregated data like multidimensional tables using the RDF Data Cube Vocabulary[22], and delineating code lists, category schemes, mappings between them, and concepts like topics using SKOS. This section serves as an overview of the conceptual model for the Disco vocabulary. More detailed descriptions of all the properties are given in the specification[23] and a conference paper (Bosch et al. 2012).

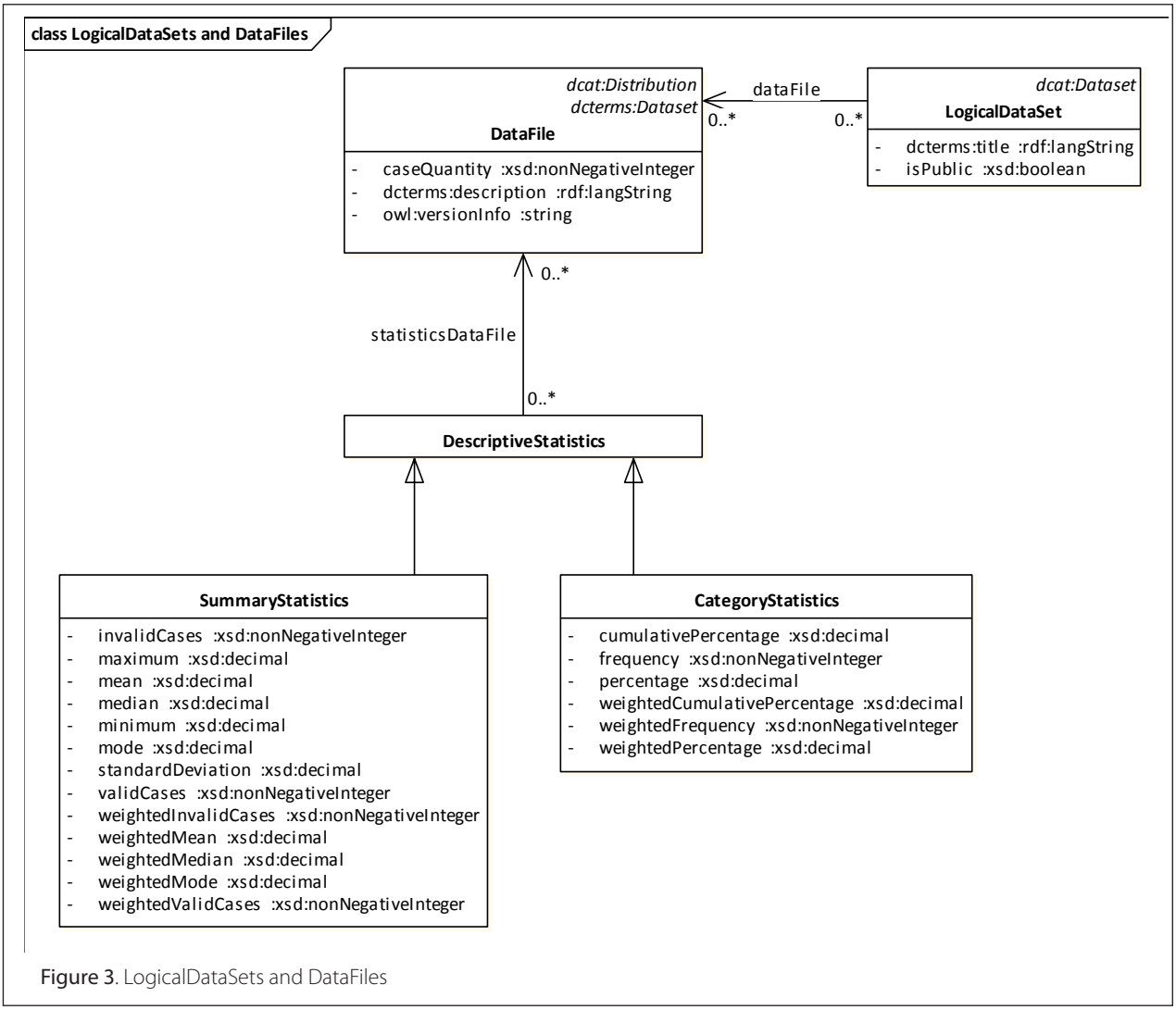


Figure 3. LogicalDataSets and DataFiles

Overview

Figure 2 provides a diagram of the conceptual model containing a small subset of the DDI-XML specification[24]. To understand the DDI Discovery Vocabulary, there are a few central classes, which can serve as entry points. The first of these is Study. A Study represents the process by which a dataset was generated or collected. Literal properties include information about the funding, organizational affiliation, abstract, title, version, and other such high-level information. In some cases, where data collection is cyclic or ongoing, datasets may be released as a StudyGroup, where each cycle or "wave" of the data collection activity produces one or more datasets. This is typical for longitudinal studies, panel studies, and other types of "series". In this case, a number of Study objects would be collected into a single StudyGroup.

Datasets have two representations: a logical representation, which describes the contents of the dataset, and a physical representation, which is a distributed file holding that data. It is possible to format data files in many different ways, even if the logical content is the same. LogicalDataSet represents the content of the file (it is organized into a set of Variables). The LogicalDataSet is an extension of the dcat:DataSet. Physical, distributed files are represented by the DataFile, which is itself an extension of dcat:Distribution.

When it comes to understanding the contents of the dataset, this is done using the Variable class. Variables provide a definition of the column in a rectangular data file, and can associate it with a Concept and a Question (the Question in the Questionnaire which was used to collect the data). Variables are related to a Representation of some form, which may be a set of codes and categories (a "codelist") or may be one of other normal data types (dateTime, numeric, textual, etc.). Codes and Categories are represented using SKOS concepts and concept schemes.

Data are collected about a specific phenomenon, typically involving some target population, and focusing on the analysis of a particular type of subject. These are respectively represented by the classes **Universe** and **AnalysisUnit.** If, for example, the adult population of Finland is being studied, the AnalysisUnit would be individuals or persons.

Unique identifiers for specific DDI versions are used for easing the linkage between DDI-RDF metadata and the original DDI-XML files. Every element can be related to any foaf:Document (DDI-XML files) using dcterms:relation. Any entity can have version information (owl:versionInfo). However, the most typical cases are the versioning of the metadata (the DDI or the RDF file), the versioning of the study (as a study goes through the lifecycle from conception through data collection), and the versioning of the data files. Every LogicalDataSet may have access rights statements (dcterms:accessRights) and

licensing information (dcterms:license) attached to it. Studies, logical datasets, and data files may have spatial (dcterms:spatial), temporal (dcterms:temporal), and topical (dcterms:subject) coverage.

Studies and StudyGroups

A simple **Study** supports the stages of the full data lifecycle in a modular manner. As noted above, a Study represents the process by which a dataset was generated or collected, and a number of Study objects can be collected into a single StudyGroup.

Studies may have multiple disco:instrument relationships to Instruments and may have disco:dataFile connections with 0 to n DataFiles. Studies are associated with 0 to n Variables using the object property disco:variable. Studies may have multiple LogicalDataSets (disco:product). Studies or StudyGroups (the **union of Study and StudyGroup)** may have an abstract (dcterms:abstract), a title (dcterms:title), a subtitle (disco:subtitle), an alternative title (dcterms:alternative), a purpose (disco:purpose), and information about the date and time the Study was made publicly available (dcterms:available). Disco:kindOfData describes the kind of data documented in the logical product(s) of a Study (e.g., survey data or administrative data). Disco:ddiFile leads to foaf:Documents which are the DDI-XML files containing further descriptions of the Study or the StudyGroup. Creators (dcterms:creator), contributors (dcterms:contributor), and publishers (dcterms:publisher) of Studies and StudyGroups are foaf:Agents which are either foaf:Persons or org:Organizations whose members are foaf:Persons. Studies and StudyGroups may be funded by (disco:fundedBy) foaf:Agents. The object property disco:fundedBy is defined as sub-property of dcterms:contributor.

Universe is the total membership or population of a defined class of people, objects, or events. **AnalysisUnit** is the particular type of subject being analyzed, for example, individuals or persons. Studies and groups of Studies must have 1 to n Universes which are sub-classes of skos:Concepts. For Universes one can state definitions using skos:definition. The union of Study and StudyGroup may have 0 or 1 AnalysisUnit reached by the object property disco:analysisUnit. AnalysisUnit is specified as a sub-class of skos:Concept.

Logical Datasets, Data Files, Descriptive Statistics, and Aggregated Data

As noted, datasets have a logical representation, which describes the contents of the dataset, and a physical representation, which is a distributed file holding that data. It is possible to format data files in many different ways, even if the logical content is the same. **LogicalDataSet** represents the content of the file (its organization into a set of Variables). The LogicalDataSet is an extension of dcat:DataSet. Physical, distributed files containing the microdata datasets are represented by **DataFile**, which are sub-classes of dcterms:Datasets and dcat:Distribution.

An overview of the microdata can be given either by descriptive statistics or aggregated data. **DescriptiveStatistics** may be minimal, maximal, mean values, and absolute and relative frequencies. qb:**DataSet** originates from the RDF Data Cube Vocabulary[25], an approach to map the SDMX information model to an ontology. A DataSet represents aggregated data such as multidimensional tables. **SummaryStatistics** pointing to variables and **CategoryStatistics** pointing to categories and codes are both descriptive statistics.

Variables, Variable Definitions, Representations, and Concepts

When it comes to understanding the contents of the dataset, this is done using the Variable class. **Variables** provide a definition of the column in a rectangular data file, and can associate it with a **Concept**, and a **Question.** Variable is a characteristic of a unit being observed. A Variable might be the answer to a question, have an administrative source, or be derived from other Variables. **VariableDefinitions** encompass study-independent, reusable parts of Variables like occupation classification.

Questions, Variables, and VariableDefinitions may have Representations. Representation is defined as a sub-class of the union of **rdfs:Datatype** (e.g., numeric or textual values) and **skos:ConceptScheme,** as for example questions may have as their response domain a mixture of a numeric response domain containing numeric values (rdfs:Datatype) and a code response domain (skos:ConceptScheme) -- a set of codes and categories (a "codelist").

Codes and Categories are represented using **SKOS Concepts** and concept schemes. SKOS defines the term skos:Concept, which is a unit of knowledge created by a unique combination of

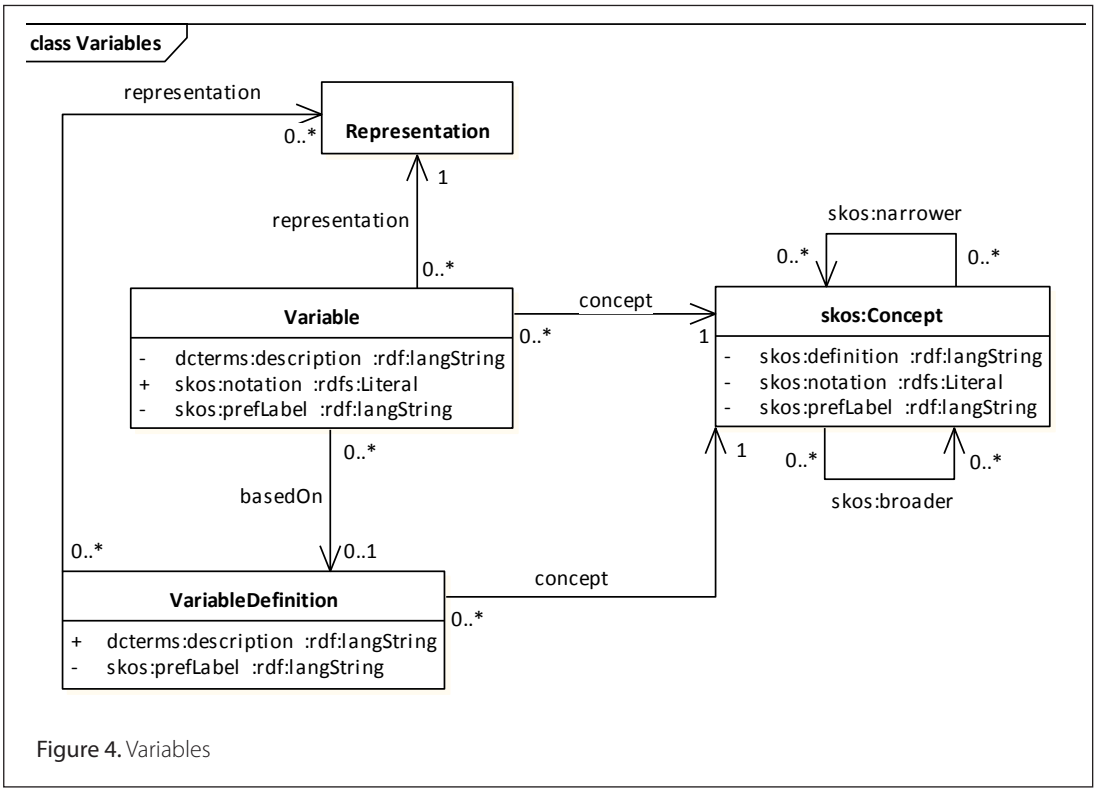


Figure 4. Variables

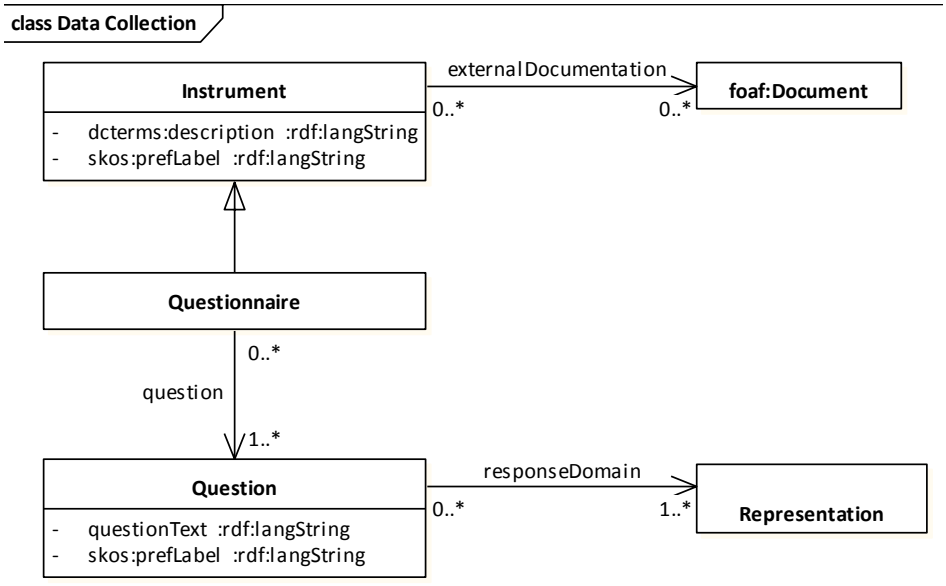


Figure 5. Data Collection

characteristics. In the context of statistical (meta)data, concepts are abstract summaries, general notions, or knowledge of a whole set of behaviors, attitudes, or characteristics which are seen as having something in common. Concepts may be associated with variables and questions. A skos:ConceptScheme is a set of metadata describing statistical concepts. Skos:Concept is reused to a large extent to represent DDI concepts, codes, and categories.

Data Collection

The data for the study are collected by an Instrument. The purpose of an Instrument, e.g., an interview, a questionnaire, or another entity used as a means of data collection, is in the case of a survey to record the flow of a questionnaire, its use of questions, and additional component parts. A Questionnaire contains a flow of questions. A Question is designed to elicit information on a subject, or sequence of subjects, from a respondent. The next figure visualizes the datatype and object properties of Instrument and Question.

One can describe (dcterms:description) Instruments and associate labels (skos:prefLabel) to Instruments. Instruments may have multiple external documentation files of the type foaf:Document. Questionnaires are special instruments having at least one collection mode (disco:collectionMode) which is a skos:Concept. Questionnaires must contain at least one Question. Questions have a question text (disco:questionText), a label (skos:prefLabel), exactly one universe (disco:universe), multiple concepts (disco:concept), and at least one response domain (disco:responseDomain).

Use Cases

This section describes the scenarios that the DDI-RDF Discovery Vocabulary was designed to support. These are not formal UML use cases -- instead, they are scenarios for the possible use of the vocabulary, based on an analysis of existing search interfaces and known behaviors for those looking for research data. The process around these discovery scenarios is to posit the thinking of the researcher/user seeking to find data, to identify needed classes and properties in the vocabulary, and then to render the search as it might be implemented.

Enhancing Discovery of Data by Providing Related Metadata

Many archives and government organizations have large amounts of data, sometimes publicly available, but often confidential in nature, requiring applications for access. While the datasets may be available (typically as CSV files), the metadata which accompanies them is not necessarily coherent, making the discovery of these datasets difficult. A prospective user has to read related documents to determine if the data are useful for his/her research purposes. The data provider could enhance discovery of data by providing key metadata in a standardized form. This would allow the creation of standard queries to programmatically identify datasets. The DDI-RDF Discovery Vocabulary would support this approach.

Link Publications to Datasets

Publications, which describe ongoing research or its output based on research data, are typically held in bibliographical databases or information systems. By adding unique, persistent identifiers established in scholarly publishing to DDI-based metadata for datasets, these datasets become citable in research publications and thereby linkable and discoverable for users. And in addition the extension of research data with links to relevant publications is possible by adding citations and links. Such publications can directly describe study results in general or further information about specific details of a study, e.g., publications of methods or design of the study or about theories behind the study. Exposing and connecting additional material related to data described in DDI is already covered in DDI. In DDI-RDF, every element can be related to any foaf:Document using dcterms:relation. Researchers may also want to search for publications where specific questions are discussed.

Discovering Studies Using Free Text Search in Study Descriptions

The most natural way of searching for data is to formulate the information need by using free text terms and to match them against the most common metadata, like title, description, abstract, or unit of analysis. A researcher might search for relevant studies that have a particular title or keywords assigned to them in order to further explore the datasets. The definition of an analysis unit might help to directly determine which datasets the researcher wants to download afterwards. A typical query could be 'Find all studies with questions about commuting to work'.

Searching for Studies by Publishing Agency

Researchers are often aware of the organizations that disseminate the kind of data they want to use. This scenario shows how a researcher might wish to see the studies disseminated by a particular organization, so that the datasets that comprise them can be further explored and accessed. "Show me all the studies for the period 2000 to 2010 disseminated by the ESDS service of the UK Data Archive" is an example of a typical query.

Searching for Datasets by Accessibility

This scenario describes how to retrieve datasets that fulfill particular access conditions. Many research datasets are not freely available, and access conditions may restrict some users from accessing some datasets. It is common to want to search only for those datasets that are either publicly available, or that have specific types of licensing/access conditions. Access conditions vary by country and institution. Users may be familiar with the specific licenses that apply in their own context. It is expected that the researcher looking for data might wish to see the datasets that meet specific access conditions or license terms. Here, a researcher is using a tool that will generate a SPARQL query that returns the titles of datasets that are, for example, publicly available under the Canadian Data Liberation Initiative Community policy. Optionally it would also be possible to provide links to the rights statement and the license.

There is a paper[26] describing further possible use cases in detail. Researchers can search for studies by producer, contributor, coverage, universe (i.e., study population), and data source (e.g., study questionnaire). Social science researchers can search for datasets using variables, related questions, and classifications. Furthermore, one can search for reusable questions using related concepts, variables, universe, and coverage, or by text.

RDF from Codebook and Lifecycle

We have implemented a direct and a generic mapping between DDI-XML and DDI-RDF. DDI-Codebook and DDI-Lifecycle XML documents can be transformed automatically into an RDF representation corresponding to the ontology. The direct mappings are realized through XSLT stylesheets[27] . Bosch and Mathiak (2011) have developed a generic approach for designing domain ontologies. XML Schemas are converted to ontologies automatically using XSLT transformations, which are described in detail by Bosch and Mathiak (2012). After the transformation process, all the information located in the underlying XML Schemas of a specific domain is also stored in the generated ontologies. Domain ontologies can be inferred automatically out of the generated ontologies in a subsequent step (Bosch 2012). In this section, only the direct approach is described in detail.

The structure of DDI-Codebook differs substantially from DDI-Lifecycle. DDI-C is designed to describe metadata for archival purposes, and the structure is very predictable and focused on describing variables with the option to add annotations for used question texts, etc. DDI-L on the other hand is designed to capture metadata from the early stages in the research process. A lot of the metadata can be described in modules, and references are used between, for example, questions and variables. DDI-L enables capturing and reuse of metadata through referencing.

The Disco vocabulary is developed with this in mind -- the discovery of studies, questions, and variables should be the same regardless of which version of DDI was used to document the study. DDI-L has more elements and is able to describe studies, variables, and questions in greater detail than DDI-C. However, the core metadata for the discovery purpose is available in both DDI-C and DDI-L. The transformation can be automated and standardized for both. That means that regardless of the input -- DDI-C or DDI-L -- the resulting RDF is the same. This enables an easy and equal search in RDF resulting from DDI-C and DDI-L. Also, interoperability between both is increased.

Creating Triples from DDI XML via XSLT

There is a huge ecosystem of tools exporting DDI-XML. This makes it possible to act on the output in a standardized way via XSLT. XSLT is implemented in a wide variety of environments and is a good method for making the transformation from DDI-XML to Disco. The flexibility of XSLT allows us to generate one conversion process for both DDI-C and DDI-L, which can be detected automatically inside the XSLT by paths and nodes of the input files. This corresponds to the goal to generate a consistent and equal Disco output independently of the DDI input.

The goal of making this implementation is to provide a simple way to start publishing DDI as RDF. XSLT is also easy to customize and extend so users can take the base and add output to other vocabularies if they have specialized requirements. It can also be adjusted if special requirements to the input are given. Keeping the XSLT as general as possible, we provide the basis for a broad reusability of the conversion process.

The implementation can also be used as a reference to show how elements in DDI-C and DDI-L map to Disco. The current version of the XSLT can be found at <https://github.com/linked-statistics/DDI-RDF-tools>.

Future Work on the Mapping and DDI-RDF XSLT

Currently, we have created two separate XSLT files for the conversion of DDI-C and DDI-L. According to the flexibility of XSLT we aim to merge them into one generic conversion XSLT that automatically detects which DDI input is given. Also, we plan on including parameters into the conversion process in order to select and define particular languages and URI prefixes.

Since the work on the conceptual model of Disco is currently not finished, the finalized mappings of DDI to Disco have to be included into the XSLT.

Future Work on Integrated Use of Disco and Related RDF Vocabularies

The description of the relationship of aggregated data to the original microdata by Disco, RDF Data Cube, and Prov will be further explored. Another focus will be how data portals can benefit of the combined use of Disco with DCAT, and the new RDF vocabulary on Physical Data Description (PHDD)[28].

Conclusions

In this paper, we introduced the DDI-RDF model, an approach for applying a non-RDF standard to the web of data. We developed an RDFS/OWL ontology for a basic subset of DDI to solve the most frequent and important problems associated with diverse use cases (especially for discovery purposes) and to open the DDI model to the Linked Open Data community. There are two implementations of mappings between DDI-XML and DDI-RDF: a direct mapping and a generic one, which can be applied within various contexts. The most important use cases associated with an ontology of the DDI data model are to find and link to publications related with particular data, to map terms to concepts of external thesauri, and to discover data and metadata that are interlinked with more than one study.

Diverse benefits are connected with the publication of DDI data and metadata in the form of RDF. Users of the DDI social science metadata standard can query multiple, distributed, and merged DDI instances using established Semantic Web technologies.

Members of the DDI community can publish DDI data as well as metadata in the Linked Open Data cloud. Therefore, DDI instances can be processed by RDF tools without supporting and knowing the DDI-XML Schemas' data structures. After publishing public available structured data, DDI data and metadata can be connected with other data sources of multiple topical domains.

Acknowledgements

The work described in this paper was started at the first workshop on "Semantic Statistics for Social, Behavioral, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web"[29] at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in September 2011. This work was continued at three meetings: a follow-up working meeting in the course of the 3rd Annual European DDI Users Group Meeting (EDDI11)[30] in Gothenburg, Sweden, in December 2011; a second workshop on "Semantic Statistics for Social, Behavioral, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web"[31] at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in October 2012; and a follow-up working meeting at GESIS - Leibniz Institute for the Social Sciences in Mannheim, Germany, in February 2013. This work has been supported by contributions of the participants of the events mentioned above: Archana Bidargaddi (NSD - Norwegian Social Science Data Services), Thomas Bosch (GESIS - Leibniz Institute for the Social Sciences, Germany), Sarven Capadisli (Bern University of Applied Sciences, Switzerland), Franck Cotton (INSEE - Institut National de la Statistique et des Études Économiques, France), Richard Cyganiak (DERI, Digital Enterprise Research Institute, Ireland), Daniel Gillman (BLS - Bureau of Labor Statistics, USA), Arofan Gregory (ODaF - Open Data Foundation, USA and DDI Alliance Technical Implementation Committee), Rob Grim (Tilburg University, Netherlands), Marcel Hebing (SOEP - German Socio-Economic Panel Study), Larry Hoyle (University of Kansas, USA), Yves Jaques (FAO of the UN), Jannik Jensen (DDA - Danish Data Archive), Benedikt Kämpgen (Karlsruhe Institute of Technology, Germany), Stefan Kramer (CISER - Cornell Institute for Social and Economic Research, USA), Amber Leahey (Scholars Portal Project - University of Toronto, Canada), Olof Olsson (SND - Swedish National Data Service), Heiko Paulheim (University of Mannheim, Germany), Abdul Rahim (Metadata Technologies Inc., USA), John Shepherdson (UK Data Archive), Dan Smith (Algenta Technologies Inc., USA), Humphrey Southall (Department of Geography, UK Portsmouth University), Wendy Thomas (MPC - Minnesota Population Center, USA and DDI Alliance Technical Implementation Committee), Johanna Vompras (University Bielefeld Library, Germany), Joachim Wackerow (GESIS - Leibniz Institute for the Social Sciences, Germany and DDI Alliance Technical Implementation Committee), Benjamin Zapilko (GESIS - Leibniz Institute for the Social Sciences, Germany), Matthäus Zloch (GESIS - Leibniz Institute for the Social Sciences, Germany).

References

Bosch, T., Cyganiak, R., Wackerow, J., and Zapilko, B. 2012. Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences. International Conference on Dublin Core and Metadata Applications, 46–55.

Bosch, T. 2012. Reusing XML schemas' information as a foundation for designing domain ontologies. Proceedings of the 11th International Semantic Web Conference, Part II (Berlin, Heidelberg, 2012), 437–440.

Bosch, T. and Mathiak, B. 2011. Generic Multilevel Approach Designing Domain Ontologies based on XML Schemas. Proceedings of the ISWC 2011 Workshop Ontologies Come of Age in the Semantic Web (OCAS) (Bonn, Germany, 2011), 1–12.

Bosch, T. and Mathiak, B. 2012. XSLT transformation generating OWL ontologies automatically based on XML Schemas. 6th International Conference for Internet Technology and Secured Transactions (ICITST) (Abu Dhabi, United Arab Emirates, 2012), 660 –667.

Notes

1. Thomas Bosch GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany E-Mail: thomas.bosch@gesis.org
2. Olof Olsson SND - Swedish National Data Service, Gothenburg, Sweden E-Mail: olof.olsson@snd.gu.se
3. Benjamin Zapilko GESIS - Leibniz Institute for the Social Sciences, Köln, Germany E-Mail: benjamin.zapilko@gesis.org
4. Arofan Gregory Open Data Foundation, Tucson, USA E-Mail: agregory@opendatafoundation.org
5. Joachim Wackerow GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany E-Mail: joachim.wackerow@gesis.org
6. http://sdmx.org/
7. http://www.ddialliance.org/ddi-at-work/projects
8. http://www.dataarchive.ac.uk/
9. http://thedata.org/
10. http://www.icpsr.umich.edu
11. http://www.abs.gov.au/
12. http://www.ihsn.org/adp
13. http://www.childinfo.org/mics3_surveys.html
14. http://data.worldbank.org/
15. http://www.theglobalfund.org/
16. http://www.ddialliance.org/what
17. http://www.loc.gov/standards/premis/
18. http://www.w3.org/2004/02/skos/
19. http://www.w3.org/TR/void/
20. http://www.w3.org/TR/prov-o/
21. http://www.w3.org/TR/vocab-dcat/
22. http://www.w3.org/TR/vocab-data-cube/
23. http://rdf-vocabulary.ddialliance.org/discovery
24. http://www.ddialliance.org/Specification/
25. http://www.w3.org/TR/vocab-data-cube/
26. http://www.ddialliance.org/resources/publications
27. https://github.com/linked-statistics/DDI-RDF-tools
28. http://rdf-vocabulary.ddialliance.org/phdd.html
29. http://www.dagstuhl.de/11372
30. http://www.iza.org/eddi11
31. http://www.dagstuhl.de/12422

Use Cases Related to an Ontology of the Data Documentation Initiative

by Thomas Bosch[1] and Brigitte Mathiak[2]

Abstract
Ontology engineers worked in close collaboration with experts from the statistical domain in order to develop an ontology of a subset of the Data Documentation Initiative. In this paper, we give a brief overview of the DDI ontology's current status and discuss in detail the most significant use cases associated with the DDI data model's ontology and therefore various benefits for the statistics community. By means of this ontology, DDI data as well as metadata can be published in the Linked Open Data Cloud and as a consequence be combined with an extensive number of datasets from diverse heterogeneous data sources. Researchers will have the opportunity to discover both data and metadata related to multiple studies which are interlinked in the Web of Data. In case a user searches for a specific study and does not know which terms to state, it is necessary to link DDI concepts to external thesaurus concepts. As a result, users' search tasks are facilitated in a significant manner. Semantic Web technologies enable the ability to check the consistency of the overall DDI data model and ease the comparison of DDI elements among multiple DDI instances. Furthermore, external resources like publications related to specific data can be found and linked, if they are semantically specified.

Keywords: Semantic Web, Linked Data, Data Documentation Initiative, DDI, use cases.

Introduction

Statistical domain experts worked closely with Linked Data community experts to define an ontology of the DDI data model. This work has begun at the workshop "Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web" at Schloss Dagstuhl - Leibniz Center for Informatics, Germany, in September 2011 (Dagstuhl 2011) and continued at the follow-up workshop in the course of the 3rd Annual European DDI Users Group Meeting (EDDI11) in Gothenburg, Sweden (European DDI User Conference 2011). A final Dagstuhl workshop on Semantic Statistics took place in October 2012 (Dagstuhl 2012).

Figure 1 depicts the DDI ontology's conceptual model containing the DDI elements that are seen by diverse experts of the statistical domain as the most important ones to solve problems connected with various use cases the authors of this paper identified. XML Schemas, which describe the DDI data model, build the basis of the visualized DDI ontology's conceptual model. Extensions partly borrow from existing vocabularies and partly lead to a new DDI vocabulary. The most important parts of the data model are the three components of the DDI conceptual model "Study", "Variable", and "LogicalDataSet". Thus, they are highlighted and outgoing relations are displayed in three different colors (Bosch et al. 2012).

Widely adopted and accepted ontologies are heavily reused as they can also address some DDI features. Some of the reused vocabularies are:

- Dublin Core ontology (DCMI Metadata Terms 2012) delineating metadata for citation purposes

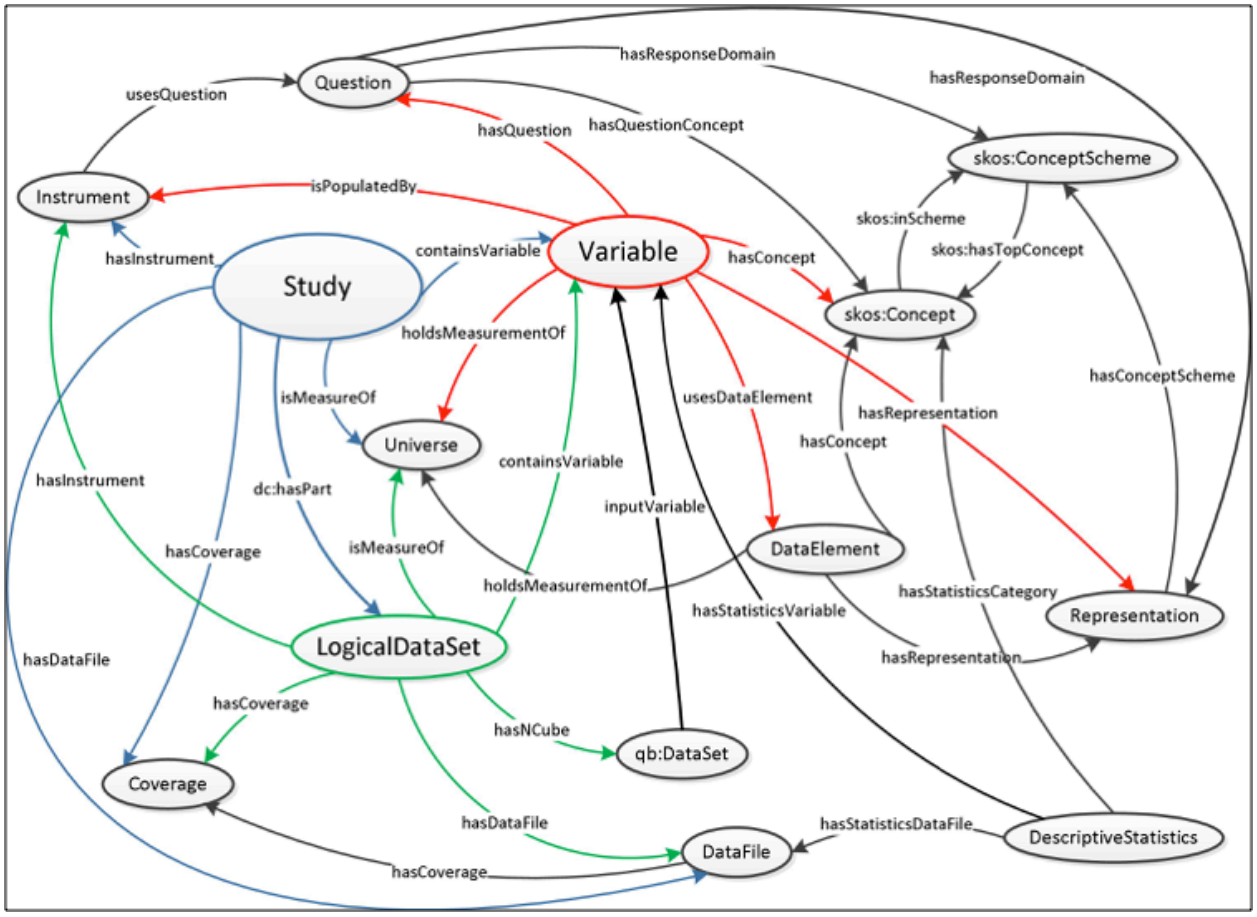


Figure 1. Conceptual Model of the DDI Ontology (Bosch et al. 2012)

- Simple Knowledge Organization System (SKOS) (W3C 2009) describing code lists, category schemes, mappings between them, and concepts like topics
- RDF Data Cube Vocabulary, which describes aggregated data like multi-dimensional tables (Bosch et al. 2012)

We defined a direct and a generic mapping between DDI-XML and DDI-RDF. Both DDI-Codebook and DDI-Lifecycle XML documents can be transformed automatically to an RDF representation, as the syntactic structure is described using XML Schemas. Bosch et al. (2011) have developed a generic multi-level approach for designing domain ontologies based on XML Schemas. XML Schemas are converted to OWL generated ontologies automatically using XSLT transformations which are described in detail by Bosch et al. (2012). After the transformation process, all the information located in the underlying XML Schemas of a specific domain is also stored in the generated ontologies. OWL domain ontologies can be inferred completely automatically out of the generated ontologies using SWRL rules.

In the following, the authors of this paper describe in detail use cases and benefits associated with an ontology of the DDI data model. We want to answer the question why it is crucially important that an RDF representation of the Data Documentation Initiative has been defined.

Publish and Link DDI Data and Metadata

Using an ontology of the DDI data model, DDI data as well as metadata can be published in the Linked Open Data cloud in the form of the standard-based exchange format RDF. The LOD cloud comprises approximately 29 billion RDF triples. The number of RDF links of nearly 400 million refers to out-going links that are set from data sources within a topical domain to data sources of other thematic areas (Bizer, Jentzsch & Cyganiak 2012). Figure 2 visualizes the current state of the entire Web of Data with its RDF triples, links between them, and diverse topical sections depicted using different colors.

One has to fulfill different conditions before datasets can be published in the LOD network. Data must be published in accordance with the Linked Data principles. Another precondition is offering RDF data through a SPARQL endpoint (W3C 2008). DDI instances can be processed by RDF tools without supporting the complex DDI XML Schemas' data structures and can be displayed using mature Linked Data browsers like Tabular (The Tabulator 2005), Marbles (Marbles 2012,) or LinkSailor (LinkSailor 2012). After publishing publicly available structured data, DDI data and metadata may be linked with other data sources of multiple topical domains. Organizations offering RDF representations of their DDI instances will be additional nodes in this LOD network.

Two major advantages are connected with the publication of DDI data and metadata in the LOD cloud and with the relation to other RDF datasets. First, each organization which is part of this continuously growing cloud can search for, find, and operate with the published DDI instances of a specific organization. And

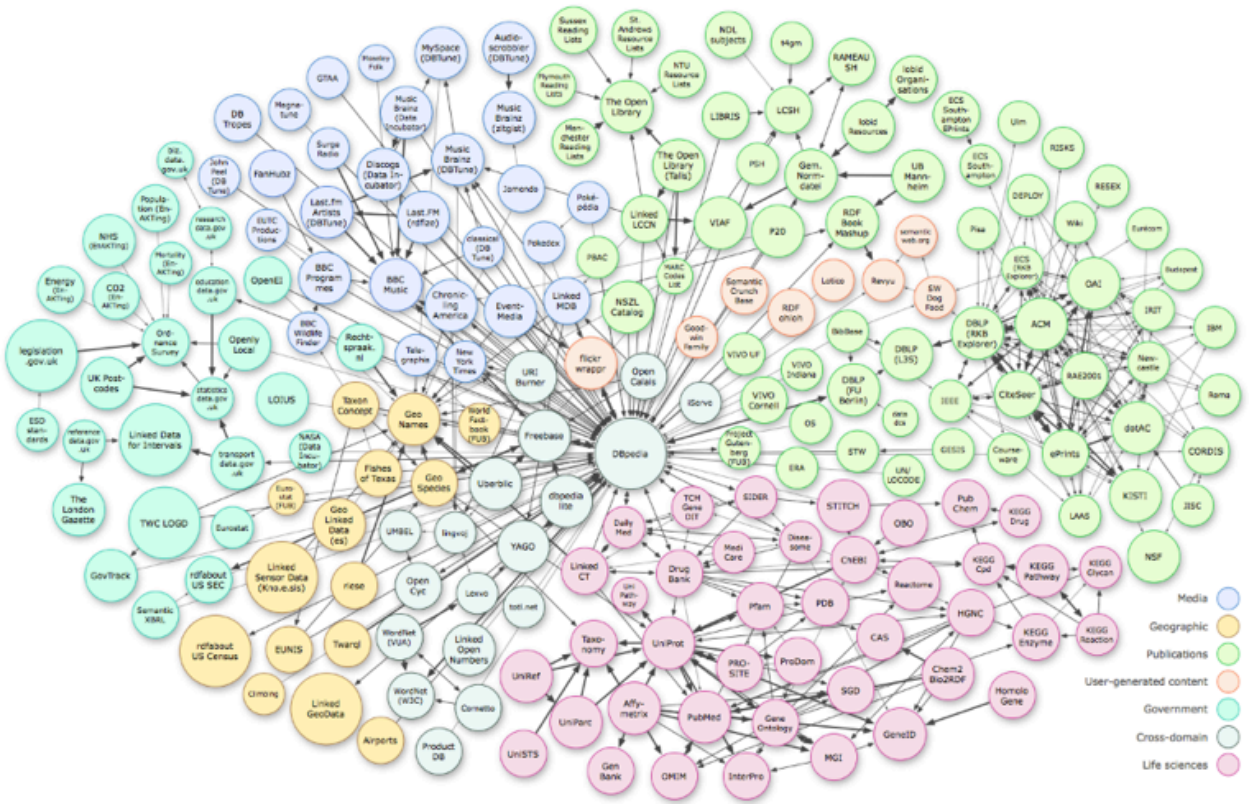


Figure 2. The Linking Open Data Cloud Diagram (Cyganiak 2011)

secondly, every node in this LOD network can also be processed by individual organizations. In summary, you can reach a broader audience and a broader audience can reach you.

Linked Data Search engines like Sig.ma (SIG.MA Semantic Information Mashup 2012), Falcons (Falcons 2011), or SWSE (Semantic Web Search Engine 2012) can search for DDI instances which can be found in the directory of all known sources of linked data with an open license (Linking Open Data Project) (LinkedData 2012). Linked Data Crawlers such as the publicly available LDSpider use RDF links between various data sources to provide extensive search functionalities (Isele et al. 1996). Even semantic mashups utilize linked RDF data from several data sources. Furthermore, the publication of Linked Data in the LOD cloud is the prerequisite of the development of Linked Data driven web applications.

Discovery

What kinds of problems can't be solved without an ontology of the DDI data model, what types of problems can be solved in a better way using such an ontology, and what is the associated additional value? Requesting multiple, distributed, and merged DDI instances will be possible. The Semantic Web query language SPARQL is applied to traverse the RDF graph (W3C 2008). The SPARQL Protocol and Query Language is similar to SQL, the Structured Query Language, within the framework of requesting relational databases. But before executing SPARQL queries, you have to generate a SPARQL endpoint (W3C 2008). Semantic queries are formulated using simple and intuitive DDI domain concepts without knowledge of complex DDI XML Schemas' structures. In the following program listing, all the questions belonging to a given variable with the variable label "age" are requested.

```
SELECT ?question
WHERE
{
  ?variable rdf:type Variable;
       skos:prefLabel ?variableLabel;
       hasQuestion ?question.
?question rdf:type Question.
Filter
   (
        ?variableLabel = 'age'
   )
}
```

The next figure visualizes the RDF representation of this specific SPARQL query. Other examples would be querying all the studies in which variables with a specific variable label exist or to request all the publications belonging to a given topic.

Bosch et al. (2012) provide a detailed description of the discovery use case summarized in this sub-section. By means of the DDI ontology, researchers can discover both data and metadata belonging to more than one particular study. Researchers often wish to know which studies are connected with a specific universe consisting of the three dimensions: time (e.g., 2005), country (e.g., France), and population (e.g., age between 18 and 65). Figure 4 depicts the SPARQL query shown below its visualization. The SPARQL query's results are the titles of the studies related to the defined universe. These individual studies are of the type 'Study' and are connected with the mentioned universe via the object property 'isMeasureOf'. This particular study is related to its title using the datatype property 'title' borrowed from the Dublin Core namespace. The universe consisting of the three dimensions time, country, and population is defined as from the type 'Universe'

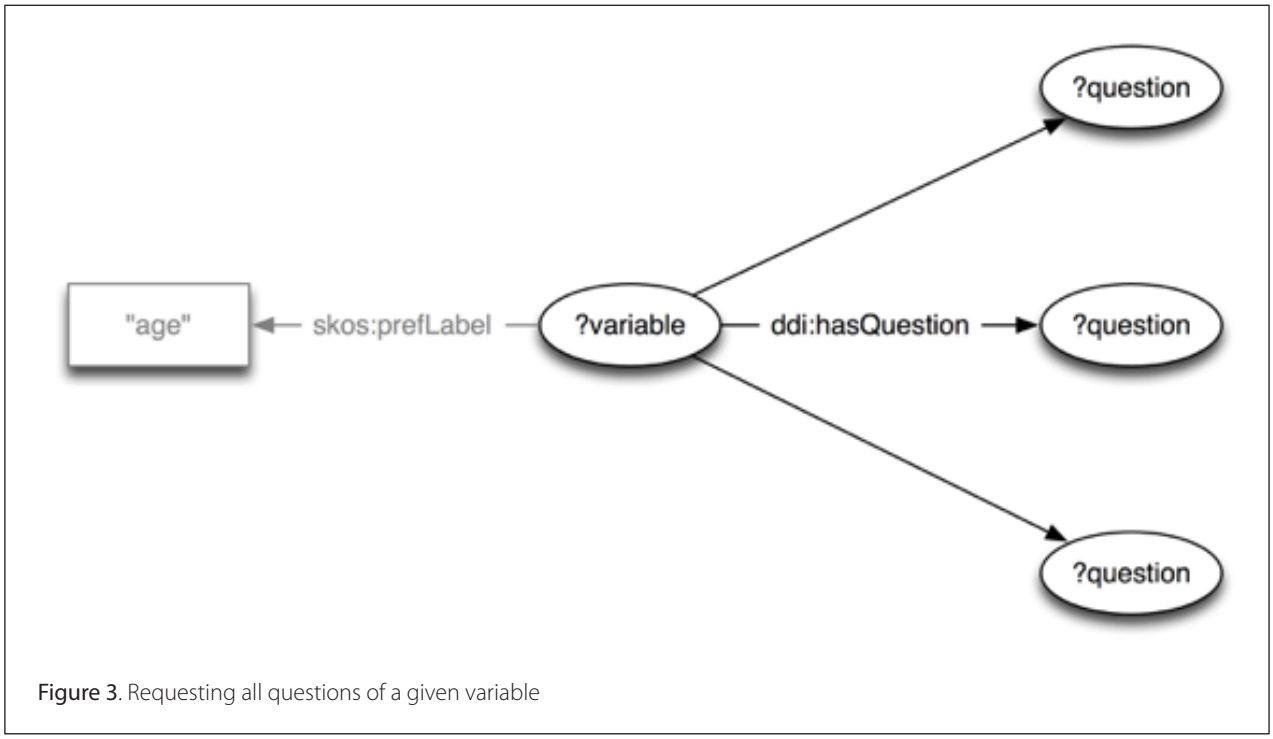


Figure 3. Requesting all questions of a given variable

and is combined with its definition via the datatype property 'definition'. The individual namespaces where the class axioms are specified are shown in the figure in the form of namespace prefixes such as 'ddi', 'dc', and 'skos'.

```
SELECT ?studyTitle
WHERE
{
   ?study rdf:type ddi:Study;
          dc:title ?studyTitle;
          ddi:isMeasureOf ?universe.
   ?universe rdf:type ddi:Universe;
          skos:definition ?universeDefinition.
    FILTER
    (
          ?universeDefinition = "country ='France' and time =
'2005' and population = 'age: 18-65'"
    )
}
```

The result of the SPARQL query is a table including all the titles of the studies which are associated with the given universe. The next step could be to request exactly those studies returned from the first query in which a particular concept (e.g., education) exist. In this case, variables associated with the three-dimensional universe and the returned studies are linked to the DDI element 'Concept' via the object property 'hasConcept'. The concept label is realized using the datatype property 'prefLabel' borrowed from SKOS.

The next figure delineates another frequent research discovery process. Researchers want to know which questions -- such as 'What is your highest school degree?' -- are linked to specific concepts like 'education' and a certain universe, as was shown in the three-dimensional universe in our previous example. Questions have a connection with their texts using the datatype property 'literalText' and are related to concepts via the object property 'hasQuestionConcept'. These concepts can have a label which has to be stated in the form of the datatype property 'prefLabel' from the SKOS namespace. Resulting questions are indirectly interlinked to the three-dimensional universe via

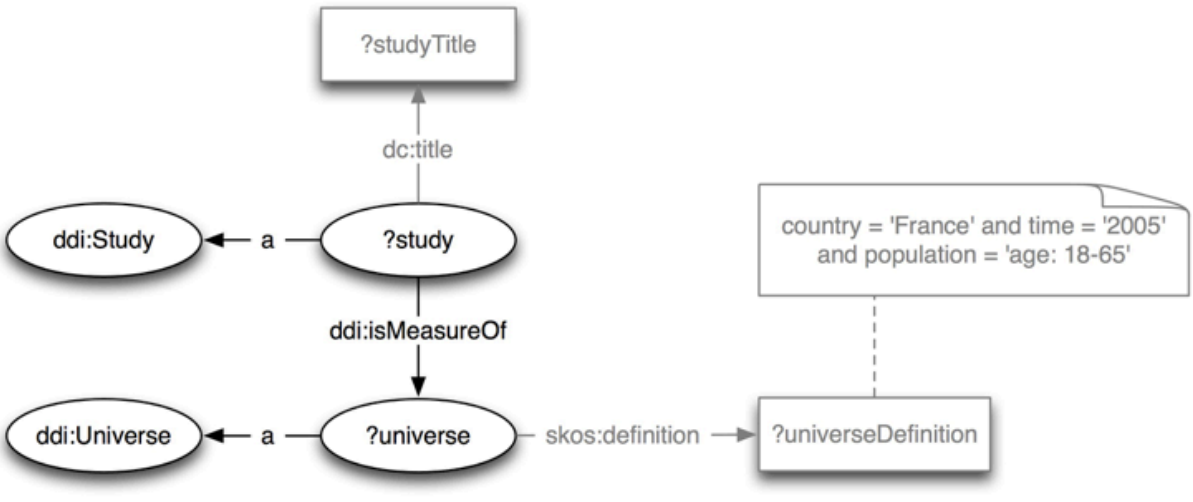


Figure 4. Discovery – Study, Universe

relationships from the concepts to variables and from variables to the universes. The SPARQL query following the figure illustrates in detail the navigation from the queried questions to the concepts and the universes.

Almost the same SPARQL query should be performed in order to get each of the variables (e.g., highestSchoolDegree) which are assigned to particular concepts (e.g., education) and which are linked to a specific universe.

and the universe with the three dimensions country, time, and population. The same researchers are now interested in the representation as wording and as code of the returned questions. Variables are interconnected with their representations which are typed as 'Representation' as well as 'skos:ConceptScheme', since the wording (the category) and the code are both represented as instances of the class 'skos:Concept'. Two datatype properties are defined for this class: 'skos:notation' and 'skos:prefLabel'. The datatype properties 'skos:notation' points to the code and 'skos:prefLabel' to the wording representation. Figure 6 shows the

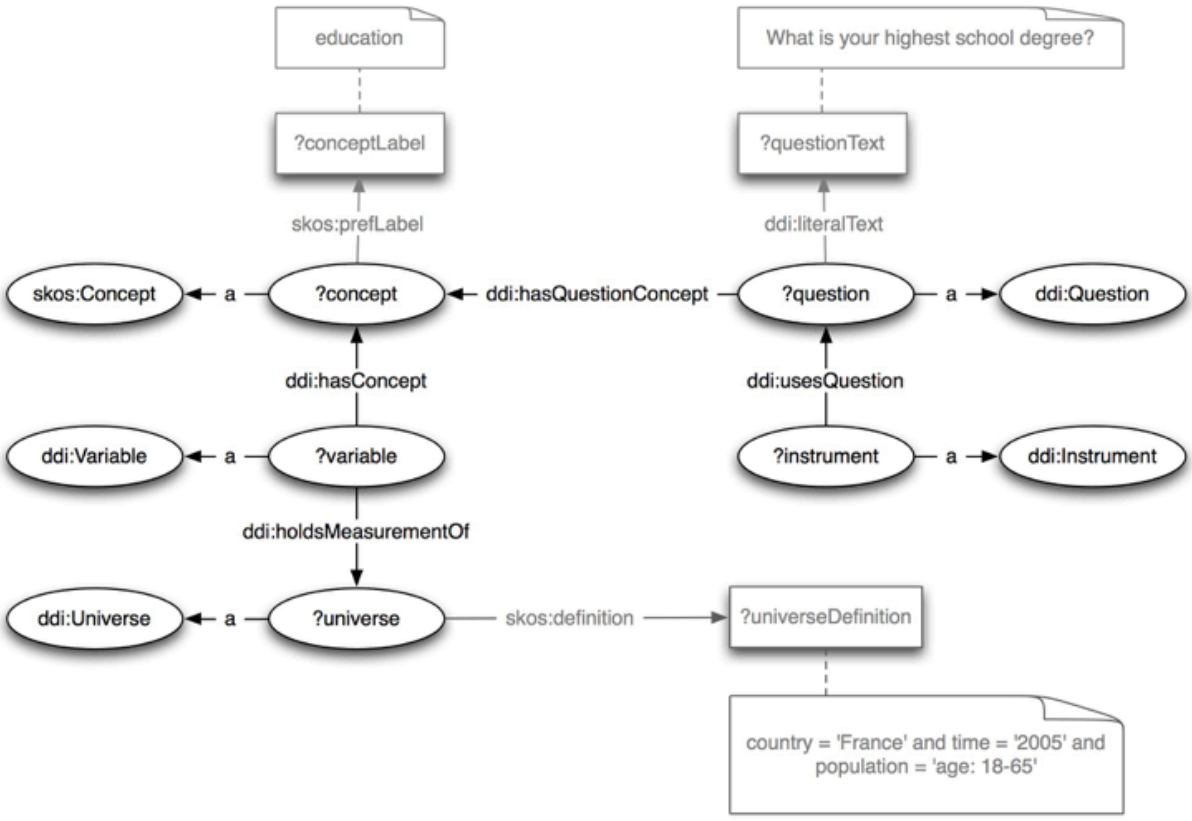


Figure 5. Discovery – Universe, Variable, Concept, Question, Instrument

```
SELECT ?question
WHERE
{
?universe rdf:type ddi:Universe;
   skos:definition ?universeDefinition.
  FILTER(?universeDefinition = "country = 'France' and time  ='2005'
    and  population = 'age:18-65'")
?variable rdf:type Varble;
  ddi:holdsMeasurementOf ?universe;
  ddi:hasConcept ?concept.
?concept rdf:type skos:Concept;
    skos:prefLabel ?conceptLabel.
    FILTER(?conceptLabel = "education")
?question rdf:type Question;
   ddi:hasQuestionConcept ?concept;
   ddi:literalText ? questionText.
   FILTER(?questionText = "What is your highest schooldegree?")
}
```

So far, the researcher gets the questions joined with the question text 'What is your highest school degree?', the concept 'education',

class axioms needed to formulate the SPARQL query below in order to implement the stated discovery sub use case. The 'where' clause of the previous SPARQL query has to be included.

```
SELECT ?question
WHERE
{
<WHERE clause of previous SPARQL query>
?variable rdf:type Variable;
  ddi:hasRepresentation ?representation.
?representation rdf:type skos:ConceptScheme
  rdf:type ddi:Representation.
?codeCategory rdf:tyoe skos:Concept;
   skos:inScheme ?representation;
   skos:prefLabel ?category.
   skos:notation ?code;
}
```

To get a first impression of the datasets' microdata, researchers are interested in descriptive statistics such as standard deviations, absolute or relative frequencies, and minimal, mean, or maximal

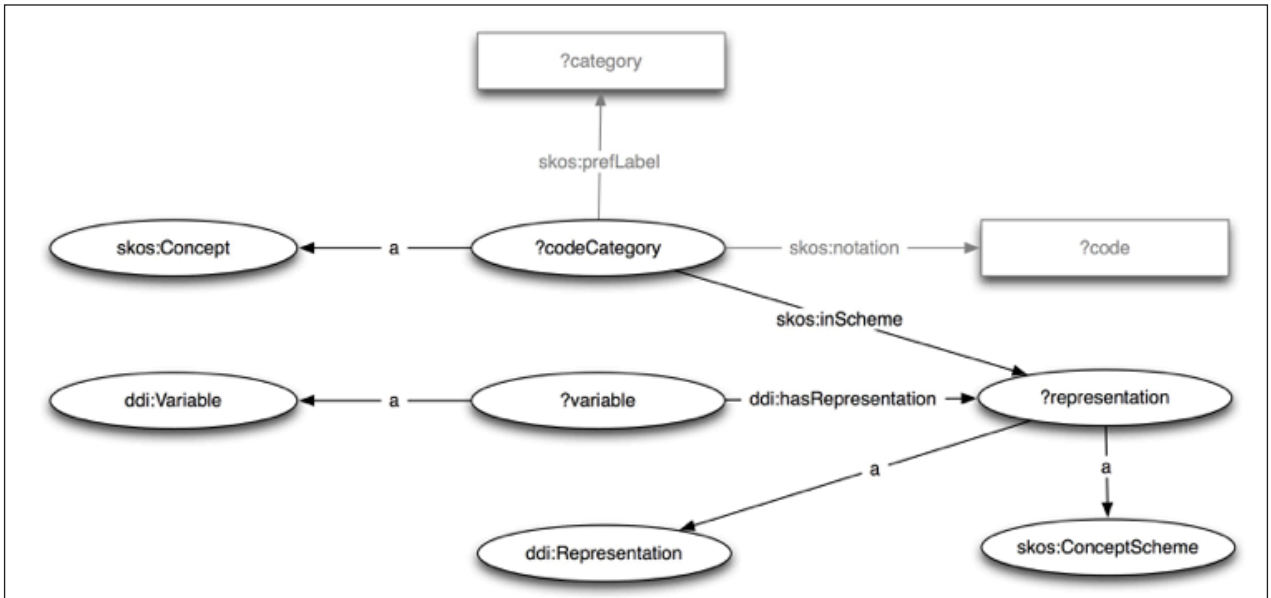


Figure 6. Discovery – Representation

values. Variables and values are directly connected with descriptive statistics which are of the type 'DescriptiveStatistics' and may have datatype properties like 'percentage' to state relative frequencies, as can be seen in the succeeding figure. If summary statistics (e.g., minimal, maximal, mean values, or standard deviations) have to be stated, instances of the class 'DescriptiveStatistics' point to variables using the object property 'hasStatisticsVariable'. If the purpose is to define category statistics like absolute and relative frequencies, descriptive statistics point to skos:Concepts representing values as well as categories via the object property 'hasStatisticsCategory'.

Which questions, connected with more than one study and the three-dimensional universe, include particular keywords (e.g., "school") in the question text? This would be another useful query, if no concepts are defined. To implement this, supplementary filters have to be set in SPARQL queries like FILTER regex(?questionText, "school", "i").

If access to microdata is limited or to get an overview over the entire microdata, researchers could request the aggregated data (e.g., a two-dimensional table with the dimensions 'age' and 'highest school degree') for particular studies, variables, universes, and concepts. Logical datasets build the link between studies and associated aggregated data which is represented by the RDF Data Cube Vocabulary's class 'DataSet'. In a similar way, microdata for a specific study, variable, universe, and concept

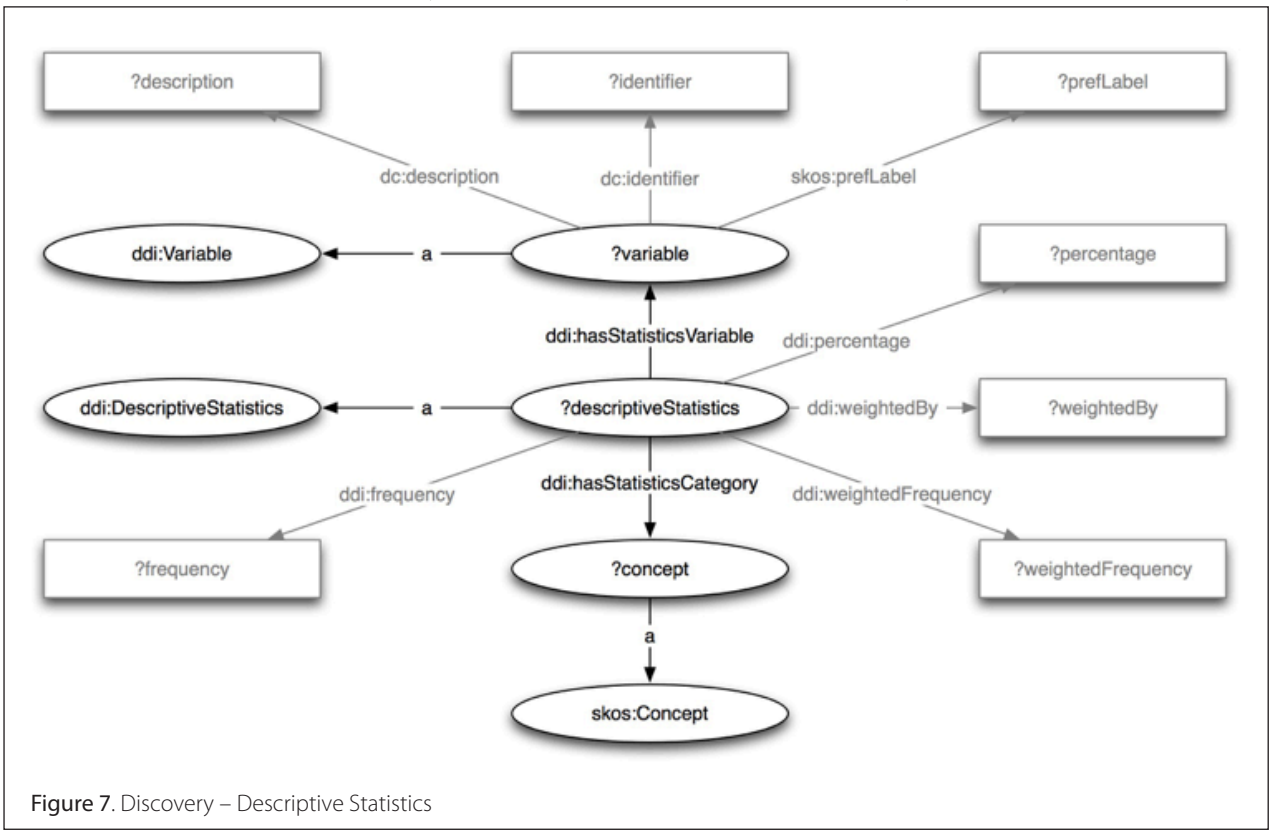


Figure 7. Discovery – Descriptive Statistics

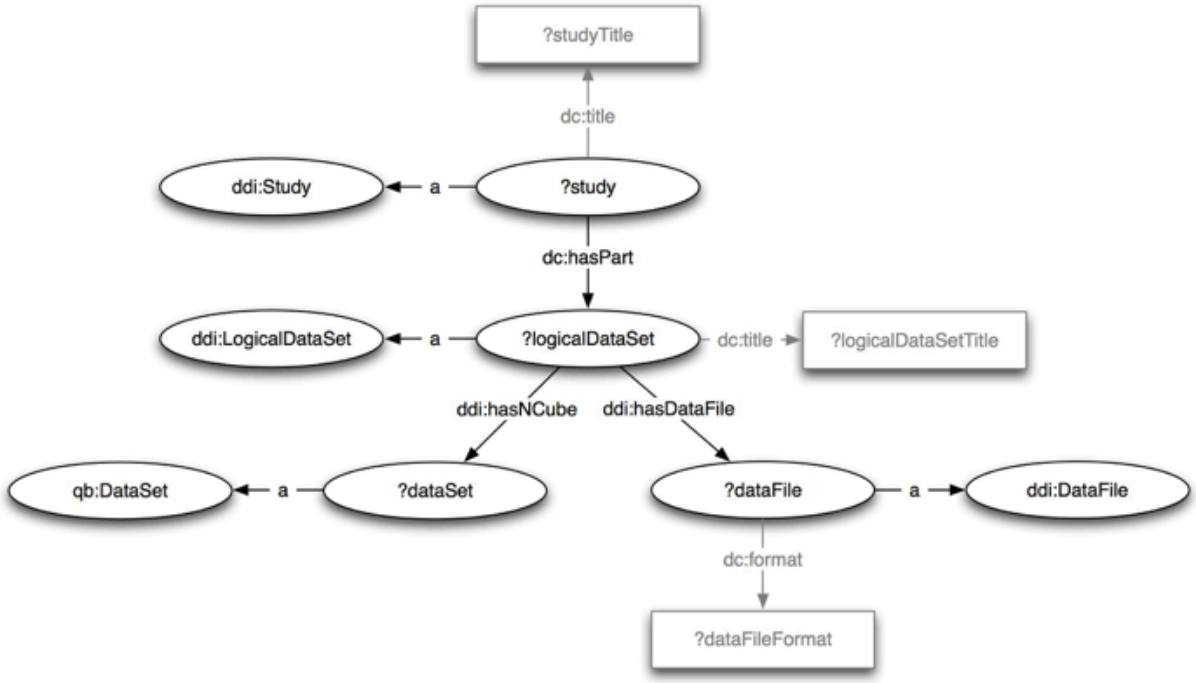


Figure 8. Discovery – LogicalDataSet, DataSet, DataFile

may be queried for analysis. The study is interconnected with an instance of the 'DataFile' class across the logical dataset. The classes 'LogicalDataSet', 'DataSet', and 'DataFile', as well as their datatype and object properties, are visualized in Figure 8.

Further use cases would be retrieving studies in which specific variables are contained or variables which are included in a specific study. Using an RDF representation of the DDI data model, comparable data could be found following the same approach describing the characteristic of a given study. Parts of studies could be compared, if the same 'DataElement' (study-independent re-usable units of information) is used. Many more use cases are conceivable like finding source data related to published aggregates (tables) and finding data related to an organization or person.

Integration of Other Ontologies

Classes, datatype, and object properties of the DDI domain ontology can relate to existing similar classes, object, and datatype properties of other external accepted and widely adopted ontologies. Conjunctions of multiple ontologies can be realized using the OWL constructs owl:equivalentClass and owl:equivalentProperty (W3C 2004). If, for instance, a concept like 'Question' is defined in the DDI domain ontology, information about possible answers and respective codes may be provided by other ontologies (see Figure 9).

Figure 9. Integration of other ontologies

The study title could be represented using a datatype property called "title". This datatype property could be newly defined in the DDI ontology or reused from already available knowledge representation systems such as Dublin Core (Dublin Core Metadata Initiative 2008) or the Semantic Web Research Communities ontology (Ontoware.org 2012). URIs are used referring to remote resources and reasoners may use additional semantic information defined in other ontologies for deductions (Kupfer et al. 2007). As external ontologies may change over time, the referred concepts might not exist anymore. Therefore it will be necessary to jump to past versions of respective ontologies (Kupfer et al. 2007) using the OWL language construct owl:versionInfo.

An ontology of the SPSS data model and respective tools transferring metadata between DDI and SPSS may be built. Now, class axioms (i.e., classes, datatype, and object properties) of the SPSS ontology and the DDI ontology can be stated as equivalent. As a consequence, there is no need to write transformation code translating RDF instances of the SPSS ontologies to RDF representations of the DDI

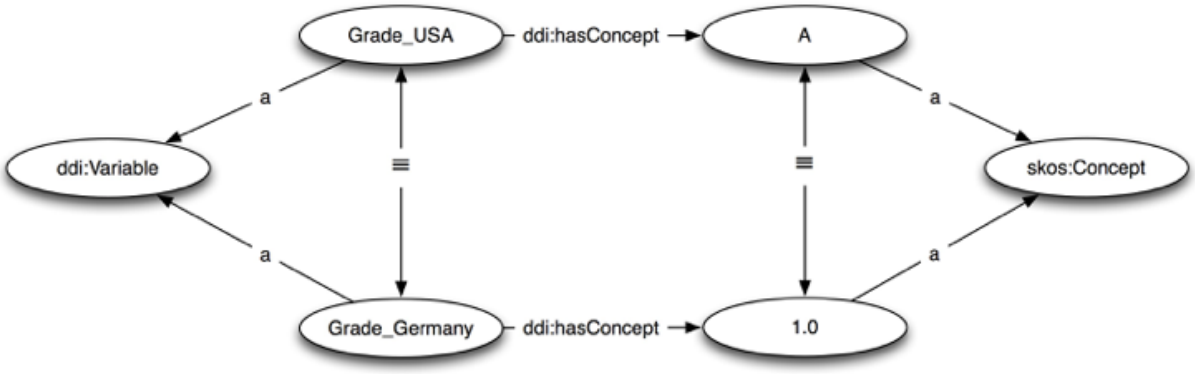


Figure 10. Equivalence of variables

ontology, as an SPSS class "Variable" could be defined as equivalent to a DDI class "Variable". So far, there is no ontology of the SPSS data model available, but as this statistics program is commonly used, this task should be executed in the future.

The members of the data.gov.uk project's working group developed the Data Cube vocabulary which is based on the SDMX (Statistical Data and Metadata eXchange 2012) data model (Cyganiak, Reynolds & Tennison 2010). SDMX (Statistical Data and Metadata Exchange), a metadata standard describing aggregate data and focusing on quantitative data, is increasingly adopted across the globe (Gregory 2011). As both microdata and aggregated data are part of study descriptions, there has to be a link between DDI and SDMX. In the current version of the DDI ontology, two appropriate relations are specified. The object property "inputVariable" points from Data Cube datasets to DDI variables and the object property "hasNCube" has the domain class "LogicalDataSet" from the DDI namespace and the range class "DataSet" specified in the Data Cube vocabulary. Moreover, top-level components of the DDI conceptual model could be defined as elements of the ISO/IEC 11179 metadata standard (ISO/IEC JTC1 SC32 WG2 2012). In order to realize this, ISO/IEC 11179 elements may be mapped to an ontology formalizing parts of the ISO/IEC 11179 metadata standard.

Expressiveness of Ontologies

Ontologies based on formal logic are more expressive than XML Schemas. On that score the DDI data model can be depicted more precisely and additional complexity to describe concepts can be formalized as well. You cannot use XML Schemas, for instance, to express that two complex types or classes are disjoint. Ontologies can describe data models in greater detail than XML Schemas, because they not only describe the syntax but semantics as well. XML Schema and OWL follow different modeling goals. On the one hand, the XML data model describes the terminology and the syntactic structure of XML documents, a node labeled tree. OWL, on the other hand, is based on formal logic and on the subject-predicate-object triples from RDF. OWL specifies semantic information about specific domains of interest, describes relations between domain classes, and thus allows the sharing of conceptualizations. More effective and efficient collaborations between individuals and organizations are possible if they agree on a common syntax (specified by XML Schemas) and have a common understanding of the domain classes (defined by OWL ontologies). XML is intended to structure and exchange documents (document-oriented), but is used to structure and exchange data (data-oriented), a purpose for which it has not been developed. Also, XML schema languages like XML Schema concentrate on structuring documents instead of structuring data.

Consistency Check of the DDI Data Model

OWL reasoning techniques, terminological and assertional OWL queries, are executed in order to determine if domain data models are consistent. Terminological OWL queries can be divided into checks for global consistency, class consistency, class equivalence, class disjointness, subsumption testing, and ontology classification. A class is inconsistent if it is equivalent to owl:Nothing, an OWL language construct. In general, this indicates a modeling error. Are there any objects satisfying the concept definition (Stuckenschmidt 2009)? If this question cannot be answered with 'yes', the respective concept is not consistent. An ontology is globally consistent if it is devoid of inconsistencies. Unsatisfiability is often an indication of errors in concept definitions and for this reason you can test the quality of ontologies using global consistency checks (Stuckenschmidt 2009). By means of classification, the ontology's concept hierarchy can be calculated on the basis of concept definitions (Stuckenschmidt 2009).

Instance checks, class extensions, property checks, and property extensions can be classified to assertional OWL queries. Instance checks are used to test if a specific individual can be assigned to a particular class (Stuckenschmidt 2009). The search for all individuals contained in a given class may be performed in terms of class extensions (Stuckenschmidt 2009). Role checks and extensions can be defined similarly with regard to pairs of individuals.

Verifications of class and global consistencies provide means to check the overall consistency of the DDI-L data model and corresponding XML Schemas by association of XML Schema declaration and definitions with OWL domain concepts. If it can be verified that the DDI ontology is consistent, meaning that the ontology does not have any contradictions, it may be derived that the DDI data model is consistent as well.

Facilitation of DDI Elements' Comparability

An extension of the actual DDI ontology and therefore its RDF representation will ease the comparability of diverse DDI elements among different DDI instances. In order to realize this, sufficient conditions specifying equality, inequality, and similarity of DDI elements have to be delineated. These conditions may be defined as immutable or recommended by an information system to researchers. Thereon, scientists will only choose conditions which are relevant for their individual research questions.

There is a limited number of DDI-L elements like variables, questions, concepts, codes (values), categories (value labels), and study descriptions which may be compared. Variables with different scales can be compared by mapping between these scales or by generating derived variables. One possible application example would be the comparison of the general qualification for university entrance in diverse countries. In figure 10, the two variables 'Grade_USA' and 'Grade_Germany' are compared. In this example, it is defined that variables are equivalent, if all their values are defined as equivalent. It is also defined that all the value pairs of the two variables ('A' and '1.0', for instance) are equivalent. As a consequence, OWL reasoners can now derive that these two variables 'Grade_USA' and 'Grade_Germany' are equivalent, too.

Both necessary and sufficient conditions may be defined. If an individual is a member of a specific class then it must satisfy the necessary conditions. If, on the other hand, some individual satisfies the sufficient conditions then this individual must be a member of a specific class (Horridge 2009). In the example above, the necessary condition could be: if an individual, a particular variable, is a member of the class called 'GradeUSA_Equivalent_GradeGermany', for instance, then it must satisfy the condition that all the values of the variables 'Grade_USA' and 'Grade_Germany' are equivalent. The sufficient condition, however, could be defined as follows: if some individual satisfies the condition that all the values of the individual 'Grade_Germany' are equivalent to values of the individual 'Grade_USA' then this individual must be a member of the class 'GradeUSA_Equivalent_GradeGermany'. In this case, the class 'GradeUSA_Equivalent_GradeGermany' is consistent, and this means that this class has at least one assigned individual and therefore these two variables can be seen as equivalent. Another application example would be to define necessary conditions determining when two variables with a different number of age classes as values are similar and when they are not.

Finding and Linking External Resources like Publications Related to Data

Publications, which describe ongoing research or its output based on research data, are typically held in bibliographical databases or information systems. Adding unique, persistent identifiers established in scholarly publishing to DDI-based metadata for datasets, these datasets become citable in research publications and thereby linkable and discoverable for users. Also, the extension of research data with links to relevant publications is possible by adding citations and links. Such publications can directly describe study results in general or further information about specific details of a study, e.g., publications of methods or design of the study or about theories behind the study.

Exposing and connecting additional material related to data described in DDI is already covered in DDI Codebook as well as in DDI Lifecycle. Because related material can vary from e.g., appendices to related sampling methods or instruments to related or outcome publications, the way to represent such information in DDI can vary from elements like 'RelatedMaterials' or 'OtherStudyMaterials' in DDI Codebook to the 'OtherMaterial' element in DDI Lifecycle 3.*. In version 3.1 of the DDI metadata standard, the element 'OtherMaterial' is used to reference resources such as publications that are related to the content of the relevant module. This element includes a description, a bibliographic citation (containing 15 Dublin Core elements like identifier, title, creator, or date), an external reference using a URL or a URN, and a reference to the item within the module to which the external resource is related (DDI Alliance 2009). Thus, all the necessary information characterizing the referenced resources can be stated. Figure 11 depicts the XML tree of the 'OtherMaterial' element.

Various drawbacks are associated with this approach modeling references to external resources. The attribute 'type' classifies external resources. To state that the resource is a publication, the value 'publication' can be assigned to the attribute. This specific value, however, is not a part of a controlled vocabulary, and the possible values of the attribute 'type' are not explicitly defined. For this reason, applications cannot understand and process the type of the external resource. As can be seen in the XML tree of the 'OtherMaterial' element, this section of the overall data model is very complex. The semantics of the 'OtherMaterial' element are not intuitive, so you have to read the documentation to get the semantics. The number of elements for bibliographic citation is limited. In DDI 3.1 you can cite works using only 15 unqualified Dublin Core elements. With an extension to qualified Dublin Core you could realize more detailed bibliographic citations (DDI Alliance 2009). References are backwards from OtherMaterial and Note in DDI 3.* to the elements using these reusable elements. This seems to be a weighty disadvantage from a modeling perspective. Ensuring reusability, it is important to store references to reusable elements in the elements using these reusable elements.

Using Semantic Web technologies, you can specify references to external resources semantically. One possible application example would be the definition of semantic references to publications as can be seen in the following figure. The class 'ReferencingPublication' is specified as the class of all the things which can have a reference to a publication via the object property 'referencesPublication'. The class 'Variable' is a sub-class of 'ReferencingPublication'. So it can be derived that every variable can

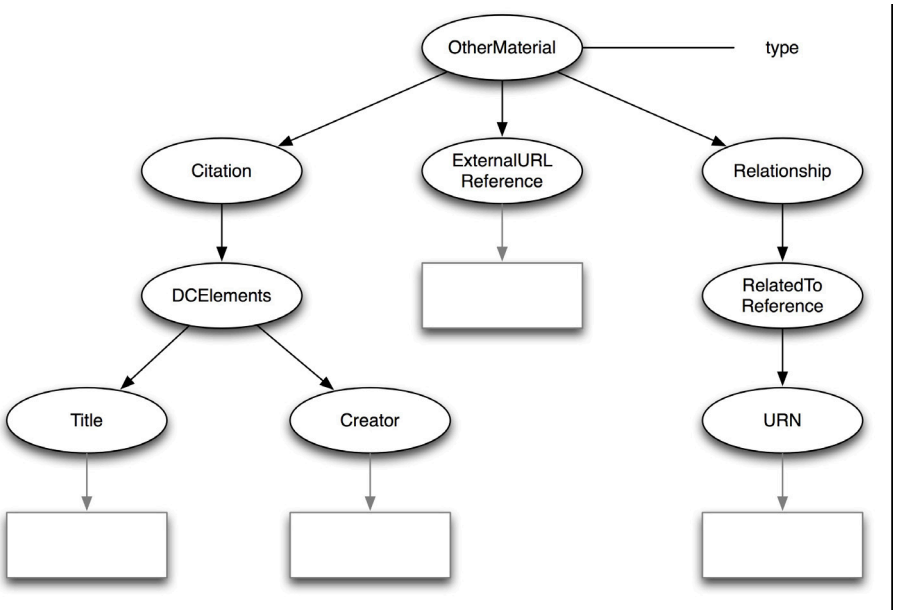


Figure 11. References to external resources in DDI 3.1

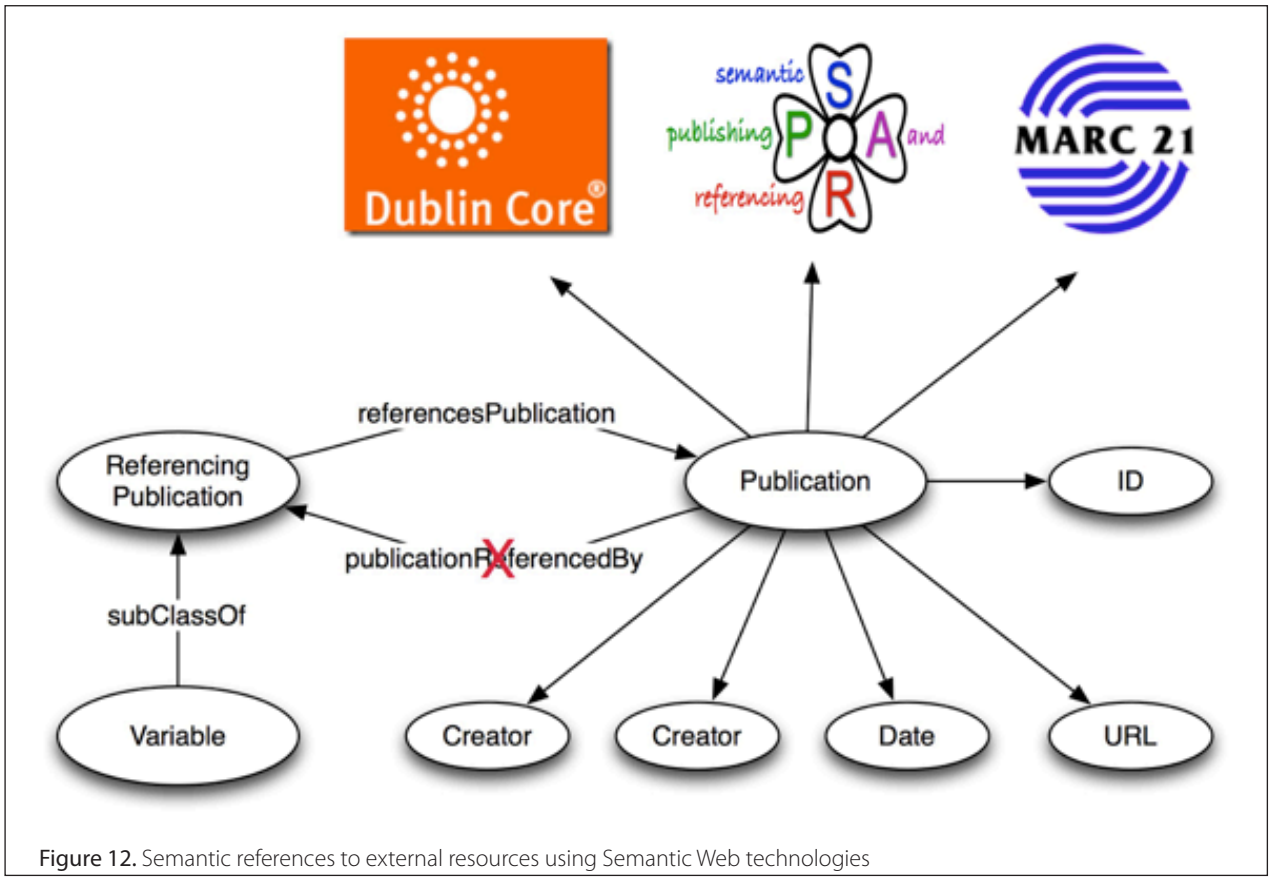


Figure 12. Semantic references to external resources using Semantic Web technologies

also have a reference to a publication. As related publications can vary, possible link predicates can also be 'backgroundPublication' for a theoretical background of the study, 'methodologyPublication' for a methodical background of the study, and 'resultsPublication' for the representation of main results, e.g., a publication based on study.

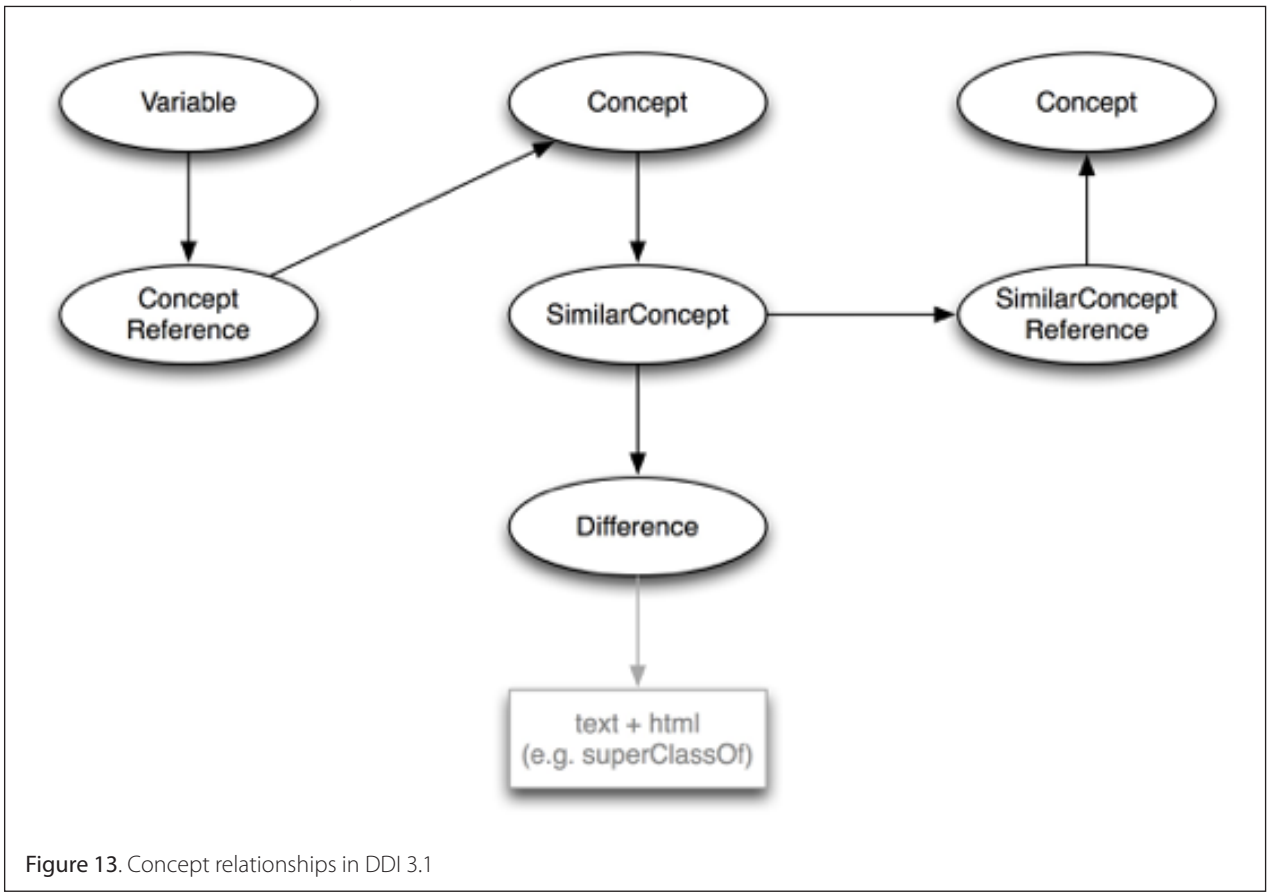


Figure 13. Concept relationships in DDI 3.1

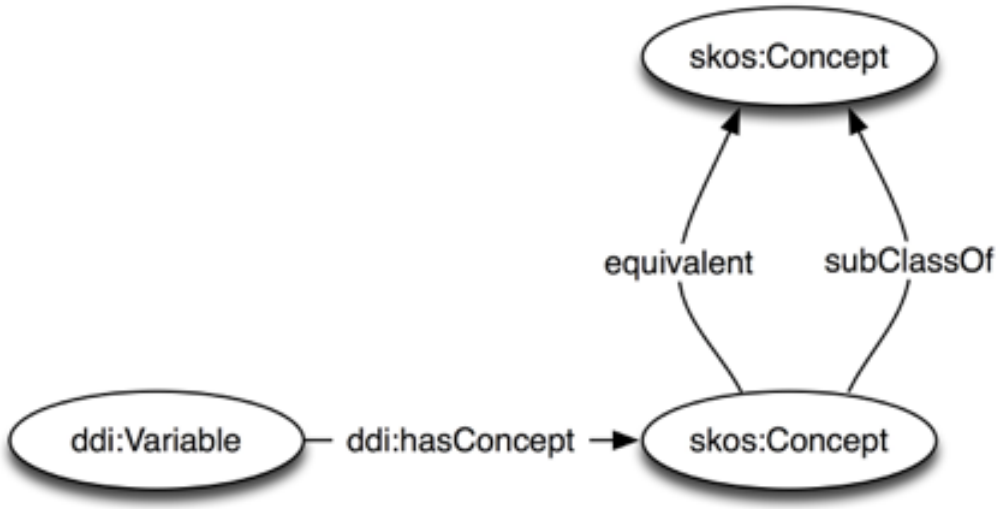


Figure 14. Semantic concept relationships

You are able to describe publications further using classes of different external ontologies such as DC, Marc 21, and SPAR. By this means, information about publications like identifier, title, creator, date, or URL can be stated. MARC (Machine-Readable Cataloging), for example, is the standard for the representation and communication of bibliographic and related information in machine-readable form (Library of Congress - MARC STANDARDS 2012). Bibliographic citations using 15 unqualified Dublin Core elements can be expanded to qualified Dublin Core elements. Dublin Core represents a very primitive way of bibliographic citing. As a consequence, Dublin Core has to be connected with other metadata standards. Nevertheless, Dublin Core is a well adopted metadata standard supported by many tools (Dublin Core Metadata Initiative 2008). SPAR (Semantic Publishing and Referencing Ontologies) is a suite of complementary ontology modules for creating machine-readable RDF metadata for all aspects of semantic publishing and referencing. SPAR consists of eight ontologies, encoded in OWL 2.0, which can be used either individually or in conjunction. The ontologies are revised, checked, stable, and ready for use (Peroni & Shotton 2011).

As you can see in Figure 12, the recommended data model is very simple, intuitive, and generic, implying that this data model can be applied in multiple contexts. To ensure reusability, variables only reference publications and not the other way around. Using this data model, both the reference and the referenced resources are defined semantically. This model can be expanded if the classes 'ReferencingResource' and 'Resources' are specified as super-classes of 'ReferencingPublication' and 'Publication.' A further example, similar to semantic references to external resources such as publications, would be to define references to notes in a semantic way.

Concept Relationships

In DDI-L, users are able to state different types of relationships between concepts. The element 'Variable' may include the element 'ConceptReference', a reference to the concept measured by this variable. The element 'Concept' can contain multiple elements called 'SimilarConcept'. The content of this element is the element 'SimilarConceptReference', a reference to another concept that is similar to the one included in the 'Concept' element description. The 'SimilarConcept' element may incorporate diverse elements called 'Difference' describing the difference and the type of relationship between the concept referenced in 'ConceptReference' and the concept referenced by the 'SimilarConceptReference' element (DDI Alliance 2009). Figure 13 demonstrates an excerpt of the DDI 3.1 XML Schemas used to describe concept relations. The 'Difference' element can only contain text and a fraction of html markup components. No controlled vocabulary and no semantics are defined. As a consequence, the content of this element is neither machine-readable nor machine-understandable, and applications cannot know how to handle this kind of content.

The DDI data model, defined using Semantic Web technologies, would be very simple, as you can see in figure 14, and reusable in other contexts.

According to this modeling approach, variables can have concepts measured by the appropriate variables. Now, you will be able to define all possible types of relations between concepts such as sub-, super-concept, and equivalence relations. As a result, the variety of connections between concepts is both readable and understandable by software components which can process the semantic information from now on in a controlled way.

Links to External Thesauri

In the current version of the DDI data model, questions, variables, data elements, descriptive statistics, and other DDI elements can relate to concepts in order to provide information about topics. DDI concepts are organized in so-called concept schemes which are similar to thesauri or classification systems regarding structure and content. When assigning concepts to DDI elements, either already available concept schemes can be reused or new concept scheme can be defined. Bosch et al. (2012) describe the thesaurus linkage use case in more detail.

The connection between DDI concepts and thesauri as well as knowledge systems' terms is relevant for two reasons: When DDI entities' concepts are defined and described, terms included in existing classification systems can be reused and provide search terms recommendation services for users. If researchers are searching for specific entities in studies, they have to state one of the concepts these entities are linked to. Therefore, a precise annotation of studies' content is significant. In many cases, researchers do not know which terms to use in the search process. To solve this problem, information systems can recommend to users suitable search terms from established thesauri or dictionaries like EuroVoc (EuroVoc 2012), Wordnet (Princeton University 2012), or LCSH (Library of Congress – Library of Congress Subject Headings 2012), when DDI concepts are mapped to these terms. Via such mappings paths from entered search terms to actually used concepts can be detected. The advantage of the reuse of different external thesauri and knowledge organization systems is that these have often been maintained for over decades and consist of well-known and established term corpora

in their specific disciplines. The inclusion of external thesauri not only disseminates the use of such vocabularies, but also promotes the potential reuse of the DDI concepts in other Linked Data applications.

As external thesauri are published in the LOD cloud, DDI as Linked Data can technically be connected with thesauri. Concepts in the DDI-RDF data format, Linked Data thesauri, and other Linked Data classification systems are typically represented on the web in the SKOS format. Conceptually there are two possibilities to establish a connection between Linked Data thesauri and DDI-RDF:

- DDI concepts can be aligned to SKOS concepts of other external thesauri using SKOS properties like skos:exactMatch, skos:relatedMatch. This mapping serves a network of related concepts over different thesauri and classification systems, which can be used to identify equivalent or related concepts.
- Often DDI metadata doesn't include concepts because they are not captured. In these cases, after-the-fact relations to external thesauri could be done by means of the semantic web. Therefore all questions, variables, data elements, or descriptive statistics in a study would reference directly via the DDI-RDF object properties to concepts from external data sources as their concepts.

Conclusions

Several use cases are associated with the development and the usage of an ontology of the Data Documentation Initiative. OWL reasoning enables the classification of DDI components such as studies. In a previous step, necessary conditions for the classification of studies have to be defined. DDI 3.1 can be used to depict quantitative data. Dealing with qualitative data will be implemented within the scope of the next DDI subversions. One additional goal of the ontology creation is to describe both quantitative and qualitative datasets. Examples of qualitative data are pictures, texts and open answers (e.g., 'Others' as a possible response to the question 'For what party did you vote?'). Metadata of pictures, structure models of texts, and relations from qualitative to quantitative data may be formulated.

Researchers often do not know which terms to use if they want to search for specific topics. DDI concepts can be annotated as equivalent to concepts defined in thesauri or classification systems. As a consequence, information systems may recommend appropriate search terms in order to build more sophisticated search processes. Researchers also want to discover microdata as well as aggregated data using graphical user interfaces on the internet. They can investigate, for example, which variables are connected with a specific question with a particular question text. By means of an RDF representation of the DDI Ontology, both DDI data and metadata can be published in the Linked Open Data cloud and be linked to other RDF datasets within the LOD cloud. A plethora of tools can be used to process RDF data without knowing the complex DDI XML Schemas' structures. Another benefit of an ontology of the DDI would be to define hierarchies and other types of relationships between DDI concepts in a semantic manner. Using Semantic Web technologies, you can specify references to external resources like publications semantically and the comparability of DDI elements is facilitated. Other external ontologies can be reused to a large extent, the DDI data model can be defined more precisely, additional more complex classes can be formalized, and OWL reasoning techniques can be used to check the consistency of the overall DDI data model.

References

Bizer, C, Jentzsch, A, Cyganiak, R 2012. State of the LOD Cloud. Available from: < http://lod-cloud.net/state/ [6 May 2015].

Bosch, T, Mathiak, B 2011. Generic Multilevel Approach Designing Domain Ontologies based on XML Schemas. Paper presented at the Workshop Ontologies Come of Age in the Semantic Web, Bonn, Germany.

Bosch, T, Mathiak, B 2012. XSLT Transformation Generating OWL Ontologies Automatically Based on XML Schemas. Paper presented at The 6th International Conference for Internet Technology and Secured Transactions, Abu Dhabi.

Bosch, T, Cyganiak, R, Wackerow J, Zapilko B 2012. Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences. Paper presented at the International Conference on Dublin Core and Metadata Applications, Malaysia.

Cyganiak, R 2011. The Linking Open Data cloud diagram. Available from: http://lod-cloud.net/. [6 May 2015].

Cyganiak, R, Reynolds, D & Tennison, J 2010. The RDF Data Cube vocabulary. Available from: http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html. [14 July 2010].

Dagstuhl 2011. Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Web. Available from: http://www.dagstuhl.de/11372. [6 May 2015].

Dagstuhl 2012. Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web. Available from: http://www.dagstuhl.de/12422. [6 May 2015].

Dublin Core Metadata Initiative 2012. DCMI Metadata Terms. Available from: http://dublincore.org/documents/dcmi-terms/. [6 May 2015].

DDI Alliance 2009. DDI 3.1 XML Schema Documentation. Available from: http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/XMLSchema/FieldLevelDocumentation/. [6 May 2015].

Dublin Core Metadata Initiative 2008. Expressing Dublin Core metadata using the Resource Description Framework (RDF). Available from: http://dublincore.org/documents/dc-rdf/. [6 May 2015].

European DDI User Conference 2011. European DDI User Conference. Available from: http://www.iza.org/conference_files/EDDI2011/call_for_papers. [6 May 2015].

EuroVoc, 2012. Available from: <http://eurovoc.europa.eu/>. [6 May 2015].

Falcons, 2011. Available from: <http://ws.nju.edu.cn/falcons/objectsearch/index.jsp>. [6 May 2015].

Gregory, A 2011. Open data and metadata standards: Should we be satisfied with "good enough"? Technical report, Open Data Foundation. Available from: http://odaf.org/papers/Open%20Data%20and%20Metadata%20Standards.pdf [6 May 2015]

Horridge, M 2009. A practical guide to building OWL ontologies using Protégé 4 and CO-ODE tools edition 1.2. University of Manchester.

Isele, R, Harth, A, Umbrich J & Bizer, C 2010. Ldspider: An open-source crawling framework for the web of linked data. ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, Vol-658.

ISO/IEC JTC1 SC32 WG2 2012, ISO/IEC 11179, Information Technology -- Metadata registries (MDR). Available from: <http://metadata-stds.org/11179/>. [6 May 2015].

Library of Congress 2012. MARC STANDARDS. Available from: http://www.loc.gov/marc/. [6 May 2015].

Library of Congress 2012. Library of Congress Subject Headings. Available from: <http://www.loc.gov/aba/cataloging/subject/>. [6 May 2015].

Linked Data, 2012. Available from: <http://linkeddata.org/>. [6 May 2015].

LinkSailor, 2012. Available from: < http://www.w3.org/2001/sw/wiki/LinkSailor>. [6 May 2015].
Kupfer, A, Eckstein, S, Störmann B, Neumann K, Mathiak B 2007. 'Methods for a synchronised evolution of databases and associated ontologies.' In Proceeding of the 2007 conference on databases and information systems IV.
Marbles, 2012. Available from: < http://mes.github.io/marbles/>. [6 May 2015].
Ontoware.org 2012, SWRC Ontology. Available from: http://ontoware.org/swrc/. [6 May 2015].
Peroni, S & Shotton, D 2011. Semantic Publishing and Referencing Ontologies (SPAR) Available from: http://sempublishing.sourceforge.net/. [6 May 2015].
Princeton University 2012. WordNet – A lexical database for English. Available from: <http://wordnet.princeton.edu/wordnet/>. [6 May 2015].
The Tabulator, 2005. Available from: <http://www.w3.org/2005/ajar/tab>. [6 May 2015].
Semantic Web Search Engine 2012. Available from: <http://www.swse.org/index.php> [1 May 2012].
SIG.MA Semantic Information Mashup 2012. Available from: < https://www.w3.org/2001/sw/wiki/Sig.ma> [6 May 2015].
Statistical Data and Metadata eXchange 2012. Available from: <http://sdmx.org/> [6 May 2015].
Stuckenschmidt, H (2009). Ontologien: Konzepte, Technologien und Anwendungen, Springer-Verlag, Berlin Heidelberg.
W3C 2004, OWL Web Ontology Language Overview. Available from: http://www.w3.org/TR/2004/REC-owl-features-20040210/. [6 May 2015].
W3C 2008, SPARQL Query Language for RDF. Available from: http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/. [6 May 2015].
W3C 2009, SKOS Simple Knowledge Organization System Namespace Document - HTML Variant - 18 August 2009 Recommendation Edition. Available from: http://www.w3.org/2009/08/skos-reference/skos.html. [6 May 2015].

Notes

1. Thomas Bosch | GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany |E-mail: thomas.bosch@gesis.org
2. Brigitte Mathiak | GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany |E-mail: brigitte.mathiak@gesis.org

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

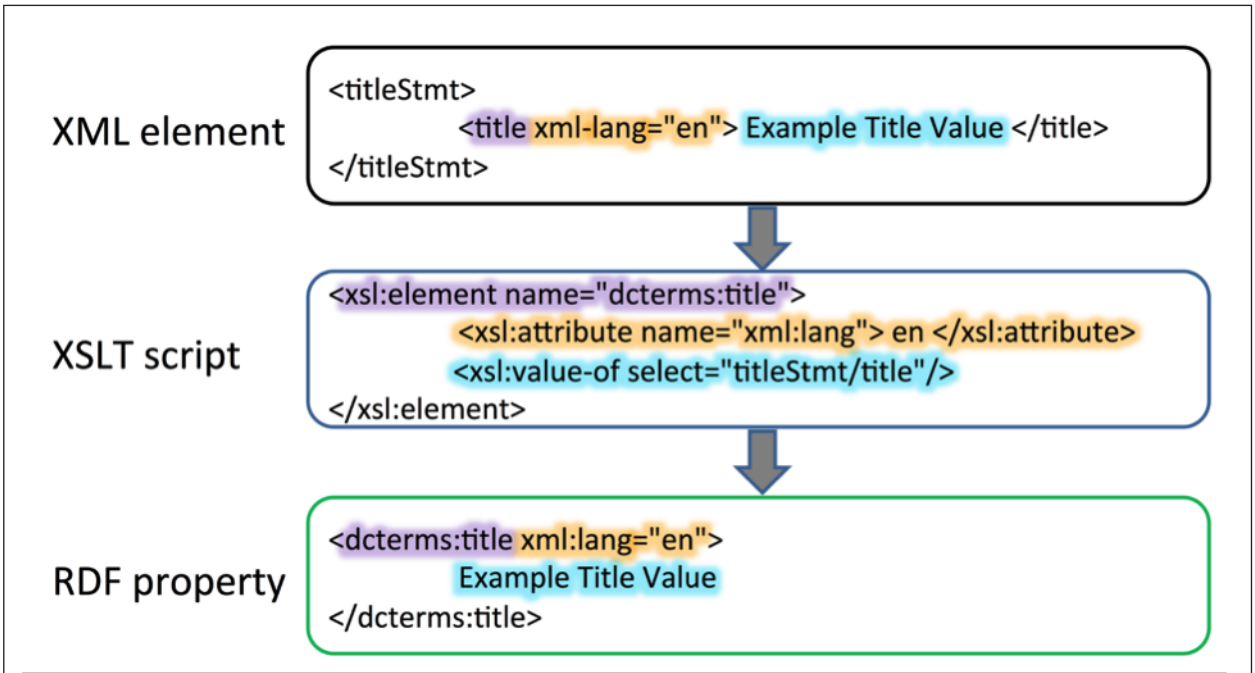


Figure 2: The XSL transformation of the entity "title"

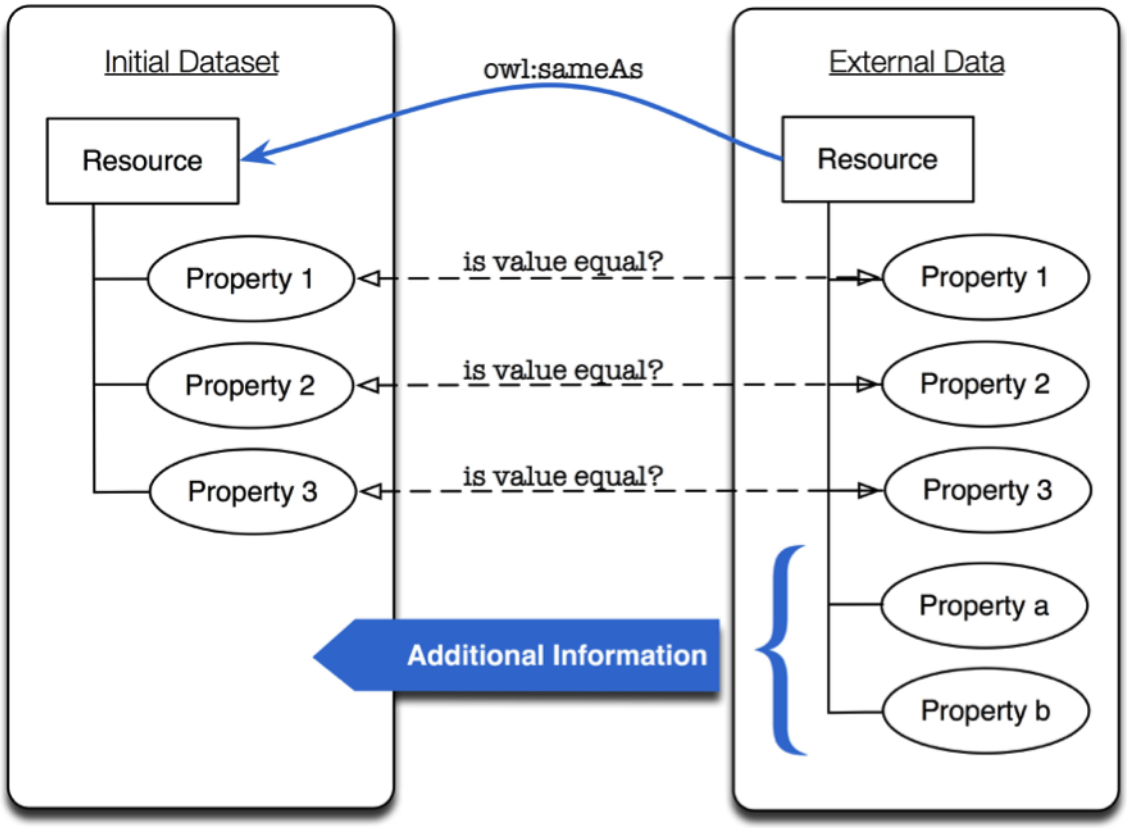


Figure 3: General link discovery procedure with owl:sameAs as defined relationship

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.

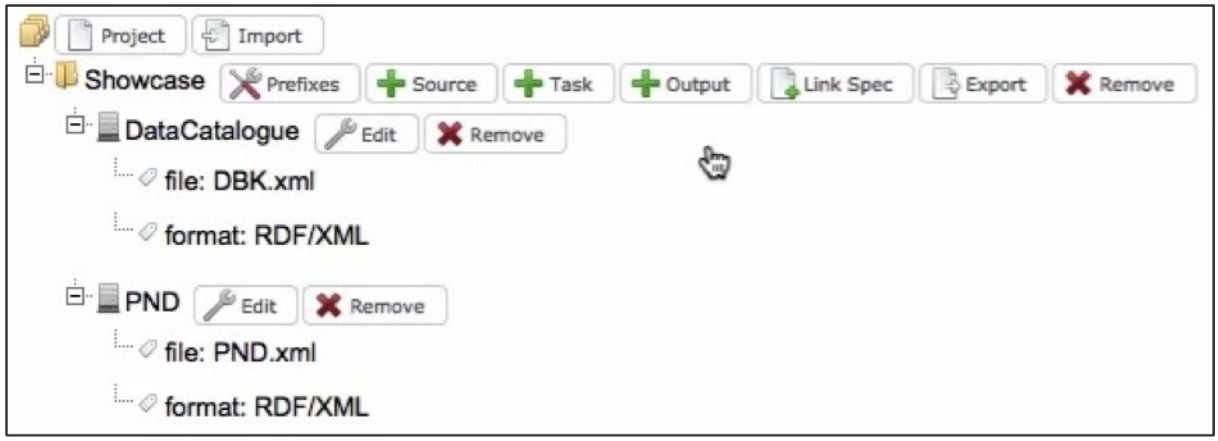


Figure 4: The definition of the Data Catalogue and the GND data sources as RDF dumps in Silk

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

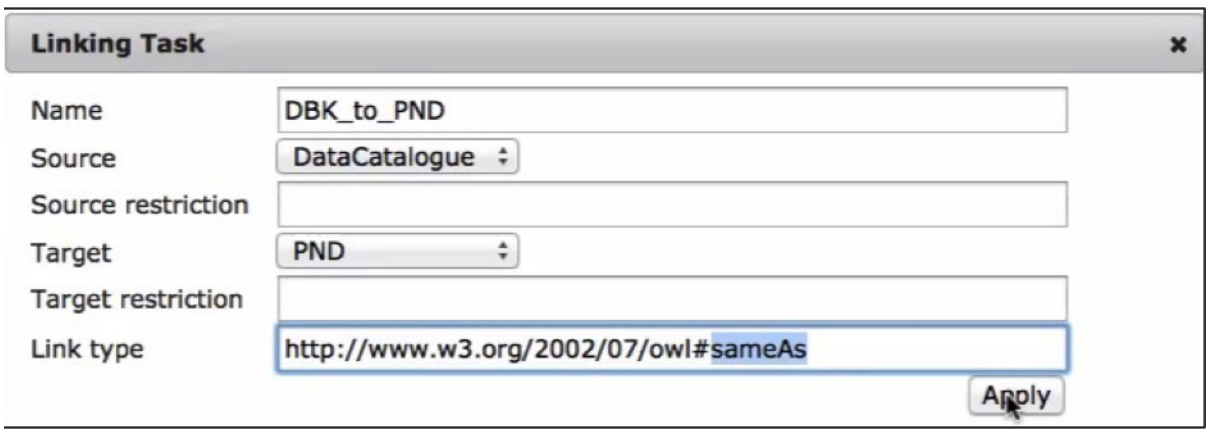


Figure 5: Defining an owl:sameAs link type between the data sources Data Catalogue and the GND

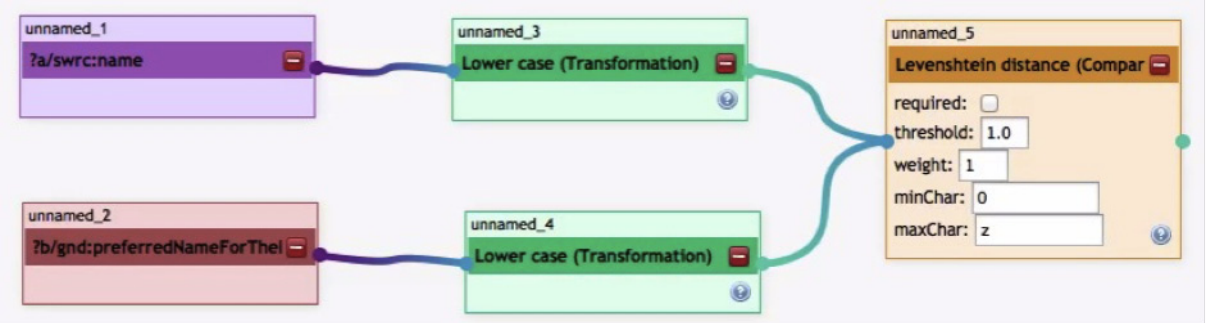


Figure 6: Definition of the linkage rule to compare two property values with the Levenshtein distance

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

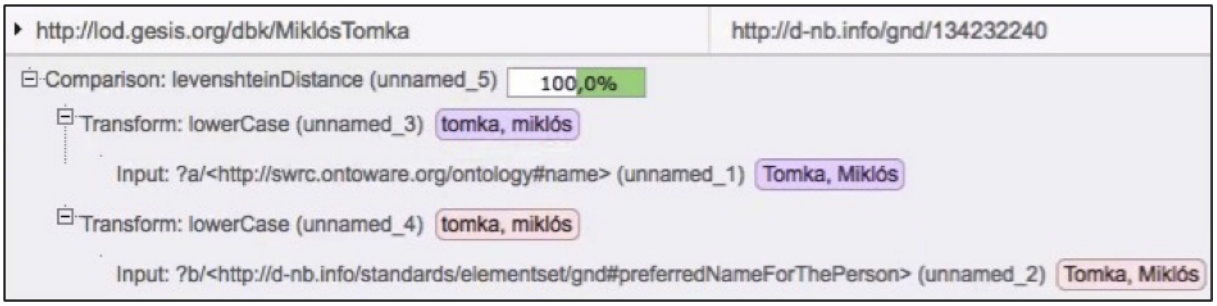


Figure 7: The result from the comparison

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/

XKOS - An RDF Vocabulary for Describing Statistical Classifications

by Franck Cotton, Daniel W. Gillman, Yves Jaques[1]

Introduction

This paper contains a brief description of the eXtended Knowledge Organization System (XKOS) and a rationale for why it was developed. In particular, there is a focus on describing statistical classifications with XKOS. For statistical data, statistical classifications are essential for categorizing complex domains, such as industries or occupations; presenting dimensions on which to aggregate data, such as in tables or time series; providing the means to stratify populations; and supplying survey respondents with standard response choices.

XKOS is an extension of the Simple Knowledge Organization System (SKOS)[2] applicable to the needs of statistical offices and social science data users. As we show in this paper, some limitations in SKOS leave it inadequate to the task of describing statistical classifications. XKOS is designed to fill these gaps.

SKOS was published in 2009 as a World Wide Web Consortium (W3C)[3] recommendation, and in the same year was extended in another vocabulary named SKOS-XL. This was to better meet the needs of multi-lingual thesauri. The purpose of SKOS is to provide a representation for knowledge organization systems, of which statistical classifications and thesauri are examples, in a machine-understandable way within the framework of the Semantic Web[4]. Therefore, SKOS-encoded statistical classifications are appropriate for use within the Linked Open Data (LOD)[5] community.

LOD is a set of recommendations for building the Semantic Web, described by Tim Berners-Lee in 2006,[6] and has been taken up by a wide variety of communities including biodiversity, environment, statistics, GIS, libraries, archives, and museums. Its promise to provide crosswalks across domains and types of data is especially attractive to the growing "open access" and "open data" movements that in the social science data community are beginning to force change to the business-as-usual practice of considering each dataset part of its own closed world.

The purpose of SKOS is to provide a representation for knowledge organization systems

Implementing the LOD recommendations provides new abilities to find, understand, and combine data on similar or otherwise related domains by organizing and linking data and metadata. LOD adds value to disparate, difficult to link datasets by employing frameworks such as the Resource Description Framework (RDF).[7] RDF, described further in the

Resource Description Framework section, is a W3C standard used for organizing and linking data. Links are used to navigate and find related data and metadata; therefore the technique, among other features, provides an easy to leverage mechanism for building mash-ups (data from multiple sources).

Implications of using LOD for data harmonization were initially explored in a paper by Gillman (2010[8]), which includes references to work to mash-up crime, traffic, workplace safety, and natural disaster risk data to create a livability index for US cities. Even though the cited work did not employ LOD *per se*, the ideas are very similar to LOD recommendations, and the reader is encouraged to understand the example. Moreover, the example shows that to do LOD right in the statistical framework is not at all straightforward. However, as a growing collection of new tools and many applications have been built with LOD, there is an expanding community of interest in employing the technology, and many benefits are promised.[9] The statistical data community needs to be paying attention to these developments.

Along with XKOS, other RDF developments that affect the statistical data community have taken place. The Data Cube vocabulary built through cooperation between LOD experts and SDMX technical experts has produced a rendition of SDMX for LOD[10] which is already in wide use by major initiatives such as data.gov.uk.[11] Similar work is planned for DDI, and the workshops held at Schloβ Dagstuhl[12] in Germany on *Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web* in September 2011[13] and October 2012[14] were devoted to the topic. In particular, this is where XKOS was first developed.

The original SKOS is used widely in LOD applications, as seen in the SKOS Implementation Report.[15] As a result, a group was formed at the Dagstuhl Workshops (in 2011 and 2012) to look at the suitability of using SKOS in the statistical data community for LOD work. As will be described in the SKOS / What is Missing section, SKOS was found to have shortcomings, so the group looked to address the issues of how to extend SKOS to meet the needs of the statistical data community. Several extensions were deemed important enough for inclusion under a new initiative, XKOS, with the intention of submitting this as a W3C Editor's Draft. Fortunately, the entire design and culture of RDF is based on a spirit of re-use and extension, so extending SKOS is technically easy. The results of the workshops and subsequent output are reported here.

In this paper, we provide introductory remarks to set the stage for discussion, provide a short primer on RDF, describe SKOS in general and the limitations to statistical classifications embedded in the design in some detail, and lay out the extensions to SKOS that form the XKOS specification. In particular, we show how the semantics of classification systems in our own offices are represented more faithfully by extending SKOS with XKOS.

Resource Description Framework

This section gives a brief primer on RDF, a W3C standard that facilitates the exchange of structured data on the Internet. Based on a simple subject-predicate-object model commonly referred to as "triples," it allows for a generic, standardized structuring of resources that can be used to model and disseminate everything from taxonomies to statistical observations to metadata records. The model used by RDF is also commonly referred to as a "graph model" consisting of "nodes" (which are vertices) and "edges" or "arcs." See the Figure 1 below for an example.

The RDF model, which by itself contains only the barest set of classes (subjects and objects) and properties (predicates), is extended using RDF Schema,[16] another fairly limited set of classes and properties that together with RDF form the foundation of the framework which can then be endlessly extended and specialized as needed. Each extension is known as a *vocabulary*, which is bounded by a *namespace*. Namespaces allow implementers to specify the *set* of classes and properties that belong to a vocabulary and give a strong assurance of uniqueness even in the open waters of the World Wide Web (WWW). This is a concept that will be familiar to those who know XML schemas.

The other very important aspect of RDF is that as with its namespaces, all of its classes and properties are also uniquely identified using the underpinning naming mechanism of the Internet, the URI[17] (Uniform Resource Identifier). In the same way that all web pages are uniquely identified by a URI (web pages actually use the URL,[18] a subset of the URI specification), all RDF classes and properties are uniquely identified by a URI. In practice this enables a powerful, standardized method for uniquely identifying information of all kinds with great certainty that the information will remain unique not only within the closed context of an internal database, but also across the WWW.

As mentioned before, each vocabulary uses a namespace to scope its set of classes and properties. This namespace is known by a URI, and by common convention the unique identifiers for the classes and properties are appended to this common namespace URI with an intervening hash or forward slash. For example, the commonly used Friend of a Friend (FOAF)[19] vocabulary, designed to link instances of people and information, uses the common namespace *http://xmlns.com/foaf/0.1/*. All of the FOAF classes and properties are then appended to this namespace, e.g., the FOAF class Person is uniquely identified by its URI as *http://xmlns.com/foaf/0.1/Person*.

Just as in an XML schema, one can define a namespace prefix to act as a shortcut for the entire namespace. Thus in a group of FOAF statements (written in XML syntax) one will commonly find a statement such as *xmlns:foaf=http://xmlns.com/foaf/0.1/*. This simply means that once this foaf shortcut has been defined, one can now refer to the URI that uniquely identifies the class FOAF Person more compactly as *foaf:Person*.

One of the other important aspects of RDF is that it does not rely on a particular syntax for its expression. Thus, there are a handful of interchangeable syntaxes that can and are used depending on a variety of requirements that one may have such as brevity or readability. This paper uses the popular Turtle (Terse RDF Triple Language[20]) syntax, prized for its readability.

Returning to the FOAF example, here is how one might make the simple triple statement that one of the authors of this paper is a thing known as a person (with a web page to provide an identifier for the actual person):

< http://aims.fao.org/community/profiles/yjaques>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person>.

So to recap, we have a subject "Yves Jaques", a "type" predicate (defined in RDFS), and an object *foaf:Person*.
To put it in another way, "Yves Jaques" is an instance of the class "Person". In RDF "type" gets used so often that Turtle lets you simply use "a" for convenience:

< http://aims.fao.org/community/profiles/Yves-Jaques>
a
<http://xmlns.com/foaf/0.1/Person>.

Let's say we want to make our statement a little shorter. We can define namespace prefixes one time and then use the shortcut for all the other triples in our graph:

@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix aims: <http://aims.fao.org/community/profiles/> .

So with those shortcuts defined, we can now write the same statement as (putting the triple on a single line this time):

aims:Yves-Jaques rdf:type foaf:Person .

Or using the Turtle shortcut for rdf:type:

aims:Yves-Jaques a foaf:Person .

Let's say we want to put a few triples together so we can say a little bit more:

aims:Yves-Jaques
a foaf:Person ;
foaf:name "Yves Jaques" .

So here we are seeing the short-hand Turtle notation for two sets of triples. In words, these triples are

"The Yves-Jaques AIMS profile web page is a person."
"The person is named Yves Jaques."

This illustrates another feature of RDF. The triples may be linked together to tell a story. The object in the first triple is then used as the subject in the next (possibly many) triple(s).

To think about what RDF looks like graphically, here is a nice diagram courtesy of Marek Obitko.[21] The round-cornered boxes are classes or instances of classes (subjects/objects), the arrows are properties (predicates), and the square boxes are *literals*. Literals are typically used to represent simple numeric values, dates, or labels. Literals can also have a datatype, a powerful mechanism to enforce restrictions on permissible values:

And here is the corresponding Turtle (note the use of the empty namespace shortcut):

@prefix : <http://www.example.org/~joe/contact.rdf#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

:joesmith a foaf:Person ;
foaf:givenname "Joe" ;
foaf:family_name "Smith" ;
foaf:homepage <http://www.example.org/~joe/> ;
foaf:mbox <mailto:joe.smith@example.org> .

To briefly recap, RDF is a framework that is designed to organize structured data about resources and their relationships over the Internet in a standard way. It is designed from the ground-up to be endlessly extensible and able to maintain the uniqueness of the things it represents even in the radically decentralized WWW.

SKOS

What Is Missing

SKOS provides a means for representing knowledge organization systems using RDF, and this makes the use of SKOS immediately applicable to LOD and the Semantic Web. So, SKOS is important for

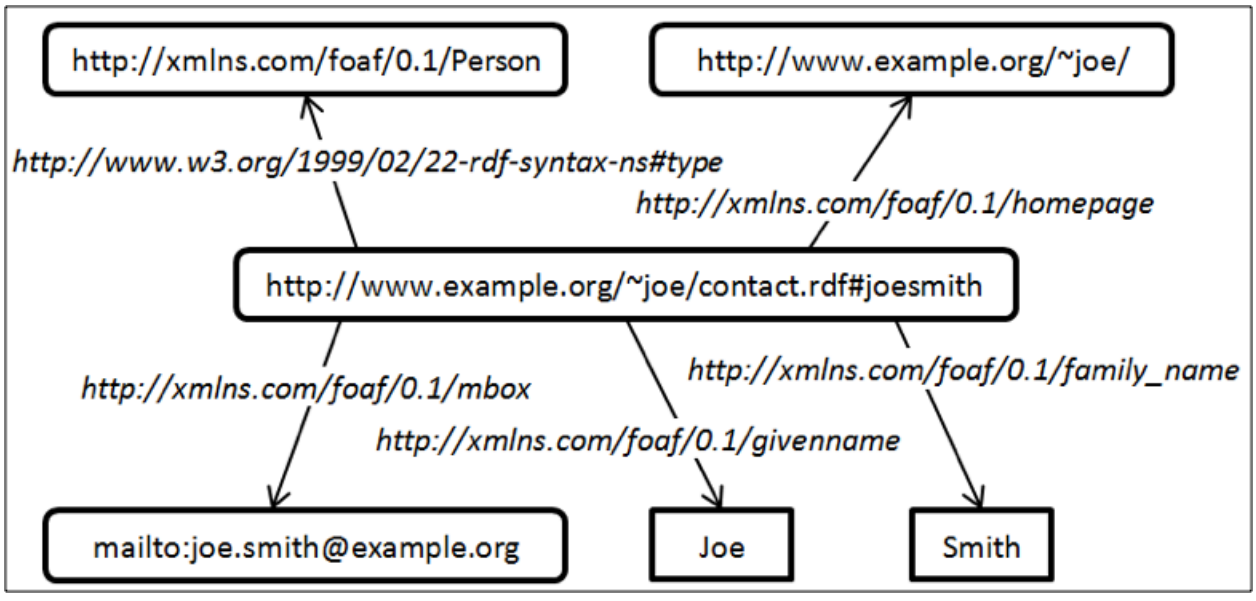


Figure 1: RDF Graph

organizations that wish to use LOD and employ classifications and code sets.

It is beyond the scope of this paper to provide a detailed description of SKOS. We direct the interested reader to the SKOS website (see End Note 2). However, SKOS contains the following basic ideas, whose definitions we paraphrase here:

- Concept Scheme – any knowledge organization system (including statistical classifications and code sets)
- Concept – any abstract idea or unit of thought
- Definition – formal statement conveying the meaning of a concept
- Label – lexical representation for a concept, may be preferred or alternate; provides means to communicate the concept
- Notation – a symbolic notation for the concept (such as a code) that is typically data-typed
- Semantic Relation – broad category for relations between concepts, such as broader than, narrower than, and related to (these relations can include relations to concepts found in other concept schemes)

The basic ideas listed above are the minimum required to describe a classification scheme. We can account for the scheme itself (*concept scheme*), all its underlying concepts (or categories as they are often called in statistics) with *concept*, what each concept means (*definition*), the labels and codes associated with a category (*label / notation*), and relationships between a concept and its parent and between and concept and all of its children (*semanticRelation*). So, is anything missing that is needed for statistics?

SKOS is based on the now withdrawn standard ISO 2788 - *Guidelines for the establishment and development of monolingual thesauri.* This standard describes three basic kinds of relations between concepts: *generic, partitive,* and *instantiation*. The *generic* relation refers to a generic / specific situation, such as between family and genus/species in the biological classification of living things. For instance, all *Homo sapiens* are mammals. The *partitive* relation refers to a part / whole situation, such as between an automobile and a steering wheel. *Instantiation* is the relation between a kind and an instance, such as each of the authors of this paper are instances of the class of people. Both the *generic* and *partitive* relations are used in statistical classifications, but *instantiation* is not.

Interestingly, the *generic* and *partitive* relations are not provided in SKOS, only the more generic *broader than* and *narrower than*, which are often referred to in more technical settings as *super-ordinate* and *sub-ordinate*, respectively. Both the *generic* and *partitive* relations are specializations of *broader than / narrower than*. In the SKOS Primer,[22] this simplification is acknowledged by the following:

> "Not covered in basic SKOS is the distinction between types of hierarchical relations: for example, instance-class and part-whole relationships. The interested reader is referred to Section 4.7, which describes how to create specializations of semantic relations to deal with this issue."

These more specialized relations were included in the past in SKOS, but they are now deprecated. XKOS, in part, is the effort to put them back.

SKOS also specifies the possibility of an *association* relation between concepts, but this is not made any more detailed. It is possible to specialize *associations* somewhat, and that is done in XKOS through *sequential, temporal*, and *causal* relations, none of which are in SKOS. The *sequential* relation refers to ideas where one is the antecedent of the other, either temporally or spatially. An example is the relationship between production and consumption. The specialized *temporal* relation is based on time. An example is the relationship between spring and summer. Finally, the *causal* relation relates cause and effect, such as the detonation of a hydrogen bomb and nuclear fall-out. Upon inspection of some classification schemes in the statistical offices of the authors, some of these relations are needed.

There is also a structural deficiency in SKOS; there is no satisfactory way to represent the idea of levels in concept schemes. Levels in statistical classifications are used to identify aggregation levels in reported statistics, which provide producers a consistent way to report their data or provide a way to reduce the threat of disclosures. Therefore, XKOS also needs to account for levels in concept schemes.

Examples

Below are some examples that illustrate the need for the extensions we have identified above:

1. The US Standard Occupational Classification System (SOC – 2012).

Take, for example

27-2000 – Entertainers and Performers, Sports and Related Workers
27-2040 – Musicians, Singers, and Related Workers
27-2042 – Musicians and Singers

The appropriate relation between 27-2000 and 27-2040 is generic, i.e. Musicians, Singers and Related Workers is a specialization of Entertainers and Performers, Sports and Related Workers. The same relation is found between 27-2040 and 27-2042, i.e., Musicians and Singers is a specialization of Musicians, Singers and Related Workers. So, the *generic* relation is needed to specify the semantics of the US SOC.

2. The US Occupational Injury and Illness Classification[24] (OIICS – 2012).

Occupational injury and illness is a four-facet classification: nature, body part, source, and event. In the body part facet, for example

3 – Trunk
31 – Chest
313 – Heart
315 – Lungs
32 – Back, including spine, spinal cord
321 – Thoracic
322 – Lumbar

Going from broad to lower detail in this snippet of the body part classification illustrates the *partitive* relation. The chest and back are parts of the trunk. The heart and lungs are part of the chest. Finally, the thoracic and lumbar regions are part of the back and spine. Note that it would not be proper to use the *generic* relation here. Therefore, the partitive relation is needed to specify the semantics of the US OIICS.

3. The US American Time Use Survey — Activity Coding Lexicons,[25] last updated in 2011. The classification is a hierarchy, but some activity categories depend on what has occurred before. For instance,
04 – Caring For & Helping non-Household Members
0402 – Caring For & Helping non-Household Children
040204 – Arts & Crafts with non-Household Children
040212 – Dropping Off/Picking Up non-Household Children

Dropping off non-household children is a sequential activity related to having supervised arts-and-crafts activities (or some other activity in the 04 group) previously. So, there are associations between some pairs of activities within this classification. In this case, the sequential or possibly the temporal relation is needed to convey the additional semantics that some activities depend on the triggering of other prior activities.

XKOS

We move now to a description of the XKOS vocabulary. As already mentioned, just as SKOS-XL extends SKOS for the needs of multi-lingual thesauri, XKOS extends SKOS for the needs of statistical classifications. It does so in two main directions. First, it defines a number of terms that allow the representation of statistical classifications with their structure and textual properties, as well as the relations between classifications. Second, it refines SKOS semantic properties to allow the use of more specific relations between concepts. Those specific relations can be used for the representation of classifications or for any other case where SKOS is employed.

Classifications

For the representation of statistical classifications, XKOS borrows from the Neuchâtel Model,[26] which is a *de facto* standard created by a group of statistical institutes and maintained in the United Nations Economic Commission for Europe's Common Metadata Framework.[27] XKOS is not a complete translation of the model, though. In particular, the notion of a classification index is not supported. There are other areas where minor differences exist between the XKOS and Neuchâtel Model approaches: these will be described below.

To begin with the classification itself, we distinguish within XKOS the notion of classification and that of classification scheme. A classification is a set of classification schemes that share a well-known name, for example, the European Statistical Classification of Economic Activities (NACE) or the International Standard Industrial Classification (ISIC). Typically, a classification scheme will be a *major version* of a given classification. For example, NACE is a *classification*, and each version of NACE (the original 1970 version, the 1990 NACE Rev. 1, the 2003 NACE Rev. 1.1, and the 2008 NACE Rev. 2) are *classification schemes* belonging to this classification.

The Neuchâtel Model also defines the Classification Variant, which is an adaptation of a classification version to a certain context or usage. In a variant, items can be split, aggregated, added, or suppressed relative to the standard structure of the base version. A variant can also be represented as an XKOS Classification Scheme, albeit of a particular type.

XKOS does not create its own object classes to represent classifications, classification schemes, and classification items, but directly uses classes already defined in SKOS. Classification items will be represented as instances of *skos:Concept*, with normal SKOS properties for codes, labels, etc. A classification scheme will simply be a *skos:ConceptScheme*, which is defined as an aggregation of concepts and semantic relationships between those concepts. A classification itself will also be a *skos:Concept*, which can in turn be included in concept schemes representing classification families (e.g., "Occupational classifications", "Activities classifications", etc.).

However, XKOS defines a set of properties that can be used to link classifications and classification schemes. For example, *xkos:belongsTo* allows one to attach a classification scheme to its classification, and *xkos:follows* or its sub-property *xkos:supersedes* can link classification schemes representing successive versions of a classification. XKOS also provides a set of properties that indicate how a classification covers its field (e.g., exhaustively, without overlap, both). The field itself would be a SKOS concept that can be taken from a well-known thesaurus such as Eurovoc[28] or the Library of Congress Subject Headings.[29]

Of course, existing standard RDF properties are available to capture versioning information, textual documentation, etc. Examples of these are the Dublin Core[30] *dcterms:valid* property, or the RADion[31] *radion:version* property. Also, *skos:note* can be used to record documentation or other descriptive resources relative to classifications and schemes. In keeping with the RDF spirit of re-use, the existing classes and properties of broadly supported vocabularies are used wherever possible.

The main purpose of a classification is to classify the entities that belong to or operate in the field that it covers. In linked data terms, classification results in the creation of an RDF triple where the subject is the resource representing the entity and the object is the concept representing the classification item. XKOS defines a generic property, *xkos:classifiedUnder,* that can be used in such statements, but classification criteria are often quite complex: for example, the same enterprise could be classified in different items of a classification of activities, depending on the rules that are used to measure its main economic activity. Thus, it is expected that *xkos:classifiedUnder* will be specialized for use in specific contexts.

Another important notion in the classifications terminology is the notion of level. Many statistical classifications, especially those that are international standards, are organized in embedded levels. For example, the ISIC Rev. 4 has four levels: the top is composed of 21 sections that cover broad economic sectors, and there are three more levels that go into greater and greater detail: divisions, groups, and classes.

In SKOS terms, classification levels are just *collections* or at most *ordered collections* of concepts, but their hierarchical organization within a classification scheme gives them extra characteristics not covered by SKOS. Thus, XKOS defines a dedicated subclass of *skos:Collection* to represent them, which is the *xkos:ClassificationLevel.* The levels or instances of *xkos:ClassificationLevel*, are structured as an RDF List, starting with the most aggregated, and the list is attached to the classification scheme by the *xkos:levels* property. An *xkos:depth* property can be used to express the distance of a given level from the (abstract) root node of the level hierarchy, and an *xkos:organizedBy* property

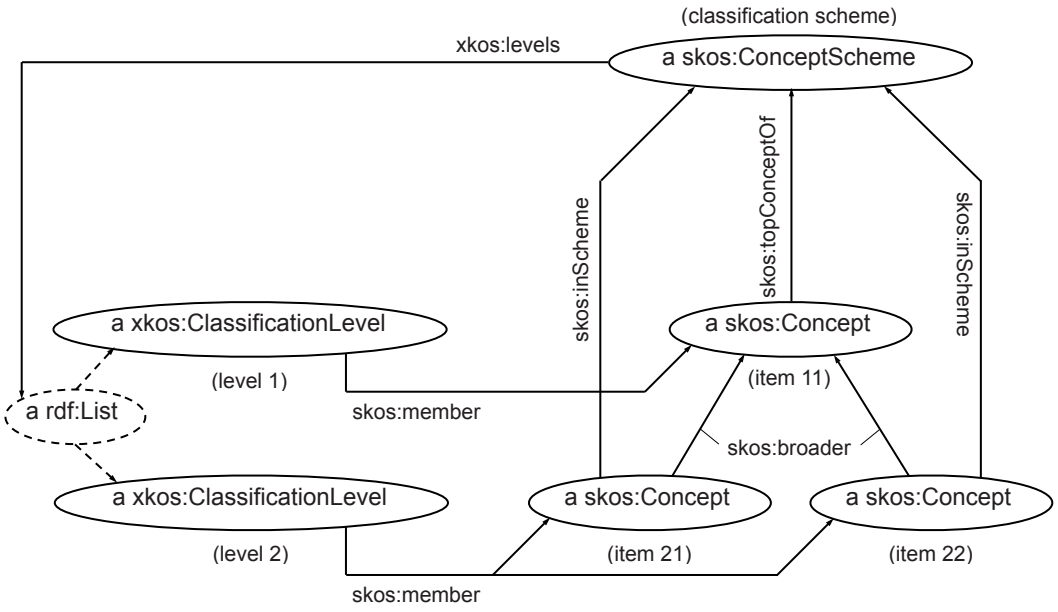


Figure 2: Structure of a Classification Scheme

can be used to record the generic name of the items of a given level (e.g., "section", "division", etc.).

The structure of a classification scheme can be described using the usual SKOS properties. More precisely:

- *skos:inScheme* (or the more specific sub-property *skos:topConceptOf* if the items belong to the most aggregated level) links the classification items to the classification scheme
- *skos:member* connects the classification level to the items that it contains
- *skos:broader* and *skos:narrower* represent the hierarchical relations between the classification items

In this last case, the more precise sub-properties defined by XKOS to express partitive or generic relations between concepts (see below) may be used instead of *skos:narrower* or *skos:broader*.

Figure 2 illustrates a simple abstract case of the usage of SKOS properties to represent the structure of a classification scheme.

Textual properties

Good classifications usually come with a fair amount of textual material, generally organized as notes attached to the classification items or to the scheme itself. These notes typically explain the content of a given classification item by describing what should be classified under this item and what should go elsewhere.

For example, here is an excerpt from the official publication of NACE:[33]

46.34 Wholesale of beverages

This class includes:

- wholesale of alcoholic beverages
- wholesale of non-alcoholic beverages

This class also includes:

- buying of wine in bulk and bottling without transformation

This class excludes:

- blending of wine or distilled spirits, see 11.01, 11.02

We see that the explanatory notes have a defined structure: they first describe what is included in the item, then what is excluded. For the inclusions, a distinction is made between what is evidently included (sometimes called "central content" or "core content"), and what is "also" included, by convention or experts' decisions, even if it does not result obviously from the item's label. For the exclusions, the note often refers explicitly to the item(s) where the content should in fact be classified.

It is perfectly satisfactory to represent explanatory notes with SKOS generic notes (*skos:note*) or scope notes (*skos:scopeNote*), but it can be useful to be able to easily distinguish between the different types of note. For this purpose, XKOS introduces four sub-properties of *skos:scopeNote*, which are represented in the Figure 3 below.

In the case of the NACE class 46.34 cited before, we would have three RDF triples to represent the explanatory notes with predicates, respectively, *xkos:coreContentNote*, *xkos:additionalContentNote* and *xkos:exclusionNote*. SKOS does not specify which type the objects of these triples should be, nor does XKOS. As a side note, Eurovoc uses an interesting mechanism that allows the representation of the notes as XHTML fragments, thereby opening the possibility of rendering the references to other items as HTML links.

Correspondences between classifications

Different classification schemes can cover the same classification, the same field, or even fields that are different but semantically related. This induces semantic relations between the classification items that belong to these schemes. A simple example of this is given by two successive major versions of a classification: some items may remain unchanged in the new version, but others will disappear, merge, be created, etc. More complicated n to m correspondences between items of the two versions are frequent.

A much more complex example of relations between classifications or classification schemes is given by the international system of economic classifications maintained by

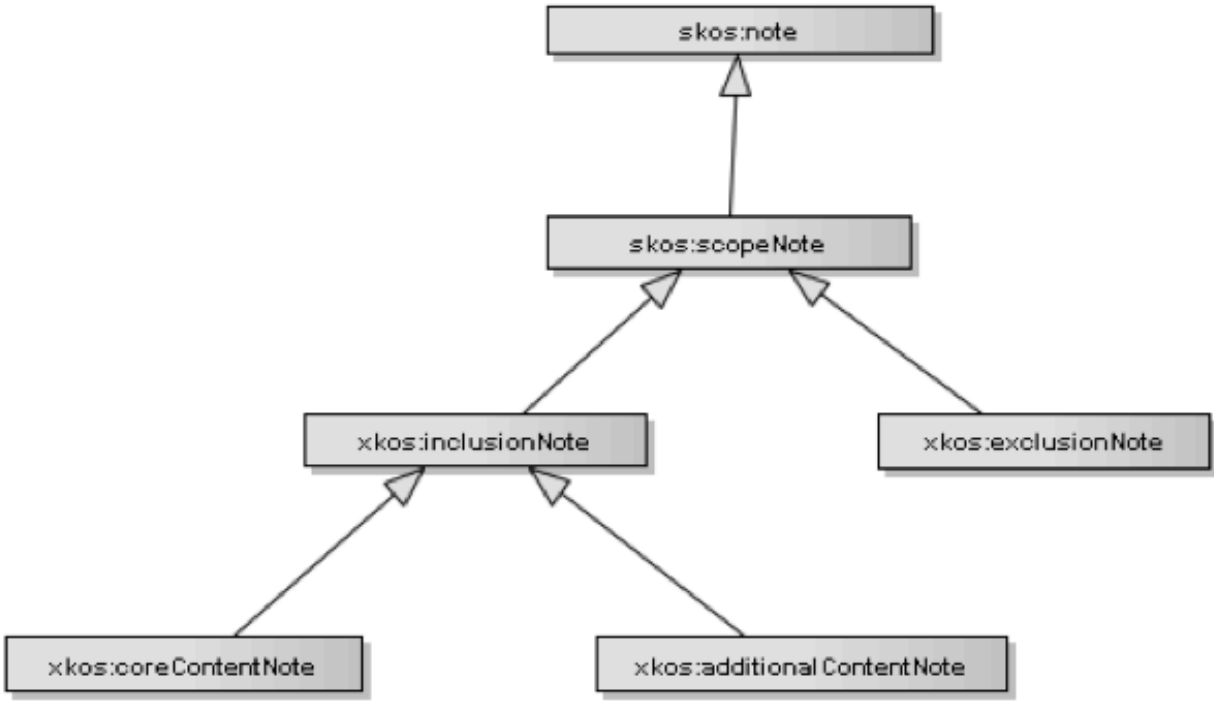


Figure 3: XKOS Note Properties

the United Nations Statistical Division. The European view of this system is well described in the online publication of the NACE Rev. 2 (*op. cit.*, chapter 1.1). The economic classifications forming this system are linked either by a common structure which gets more detailed as one goes from the international to the European to the national levels, or by semantic correspondences between the economic fields covered: activities, products, and goods (e.g., activities create products). Here again, the high-level links established between classifications result in more fine-grained correspondences between items: a given activity will create one or more specific products.

Thus, there are different types of correspondences between classifications, schemes, or items:

- Between classifications on the same field, for example, North American and European activities classifications
- Between different linked fields, for example, classifications of activities and products
- Historical correspondences, for example, SIC to NAICS
- Versioning of items over time within a given classification scheme

Since classification items are represented as SKOS Concepts, we could use the usual SKOS associative properties to represent

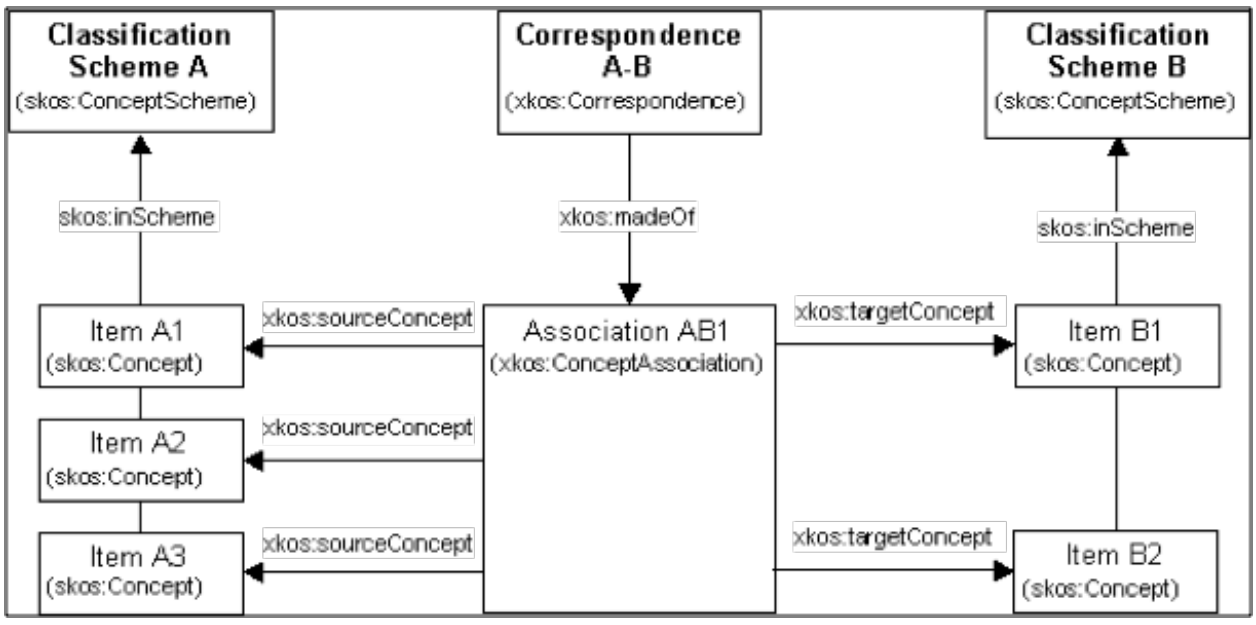


Figure 4: Concept Association Example

correspondences between them. However, this simple approach has some limitations:

- As mentioned above, relations between items in correspondences are often n to m, whereas SKOS properties relate one unique concept to another unique concept. It is always possible to decompose an n to m relation into several *1 to 1* relations, but it is better to have a global vision of a given correspondence. We also want to be able to represent *0* to *n* relations, for example, when an item is created or disappears in a new version of a classification.
- More globally, we want to be able to group all the fine-grained item associations that compose a given high-level relation between two classification schemes, such as the ones that exist in the international system of economic classifications. Such a collection of item associations is called a correspondence table, conversion table, or concordance.
- Lastly, it is often useful to be able to attach additional information (for example, notes) to item associations, for example, to describe what proportion of the different items are linked in the association.

For these reasons, XKOS defines the *xkos:ConceptAssociation* class that can be used to represent correspondences between classification items where the SKOS properties are not sufficient. Each *xkos:ConceptAssociation* may have input or source *skos:Concept*(s) and output or target *skos:Concept*(s). The complete collection of such associations for all the concepts in two SKOS Concept Schemes forms a correspondence and is expressed as an instance of the *xkos:Correspondence* class. The *xkos:madeOf* property is used to link the *xkos:Correspondence* to its *xkos:ConceptAssociation* components. To those familiar with entity-relationship diagrams, what XKOS does is to take the *skos:related* relationship (property) and "decompose" it into its own entity (class) to solve the n to m relationship problem as well as to be able to add additional properties to the relationship.

Figure 4 illustrates a simple example of a concept association: three classification items are re-combined into two.

The *xkos:ConceptAssociation* is similar to the Correspondence Item in the Neuchâtel model, but it can describe in a single instance the relationship of any number of source concepts to any number of target concepts rather than expressing the association through a set of pair-wise relations. The XKOS concept association can also represent the *Item Change* class of the Neuchâtel model. However, in this version, XKOS does not define any properties or sub-classes for *xkos:Correspondence* and *xkos:ConceptAssociation* for modeling the different types of correspondences that we described above, nor can XKOS describe the typology of item changes detailed in the Neuchâtel model (Annex 3). These may be added in a future version.

Semantic properties

Semantic properties constitute the second direction in which XKOS extends SKOS. Concept schemes are not just lists of concepts: as the SKOS Primer puts it (section 2.3), "The meaning of a concept is defined not just by the natural-language words in its labels but also by links to other concepts in the vocabulary."

SKOS intentionally defines few properties, but introduces the fundamental distinction between hierarchical and associative relations. In both these categories, XKOS creates more precise properties which are described below. The reader can refer to the figure provided in Annex 1 to find a panoptic view of SKOS and XKOS properties.

Hierarchical properties

SKOS defines several hierarchical properties, but the most used are *skos:broader* and *skos:narrower*, which are each other's inverse. These are the two properties that are refined in XKOS. A concept is broader than another one if it encompasses a wider portion of the field covered by the concept scheme, and thus includes the scope of the narrower concept. Note that the *skos:broader* property has the narrower concept for the subject and the broader one for the object, for example, "Car""broader""Vehicle"; and the *skos:narrower* property has the broader concept for the subject and the narrower one for the object, for example, "Green""narrower""Olive".

As we made clear in the previous sections, it is important, at least for statistical purposes, to represent generic and partitive relations between concepts. XKOS therefore defines two couples of inverse properties: *xkos:specializes* and *xkos:generalizes* on the one hand, *xkos:isPartOf* and *xkos:hasPart* on the other. All are sub-properties of *skos:broader* and *skos:narrower*, but the terminology is a bit tricky here: *xkos:specializes* goes from the more specific concept to the more generic one, and thus is a sub-property of *skos:broader*. Similarly, *xkos:hasPart* is a sub-property of *skos:narrower*. For example, head *isPartOf* body and chest *hasPart* heart.

Associative properties

In terms of associative properties, SKOS defines the very general *skos:related,* and a set of mapping properties (*skos:closeMatch, skos:exactMatch*, etc.) intended for establishing links between concepts of different schemes. XKOS proposes a hierarchy of *skos:related* sub-properties that convey more precise semantics. This hierarchy is organized in three branches.

The *xkos:disjoint* property forms a branch of its own. In some circumstances, it is useful to explicitly state that two given concepts do not overlap (for example, *private company* and *non-profit organization* in the *Class-of-Work* classification of the *US Current Population Survey*), especially when it has not been specified that the scheme covered its field without overlap (see A.1 in figure 4 above).

The second line of XKOS associative properties is dedicated to causal relationships. This class of link between concepts is frequently encountered (physics, biology, history, law, etc.). The generic *xkos:causal* is further subdivided into *xkos:causes* and *xkos:causedBy*, so that the direction of the causality can be expressed.

The last branch of properties is the most populated and deals with sequential relationships; it is represented on Figure 5 below. The top node of this branch is *xkos:sequential*, a refinement of *skos:related* that just indicates that two concepts in a scheme are in a sequential relationship, for example, notes in a musical scale. Below are *xkos:succeeds* and *xkos:precedes* that can be used when the sequence has a known order between the concepts. A third sub-property of *xkos:sequential* is *xkos:temporal*, which can be used when the sequence is of a temporal nature (i.e., events in time.). *xkos:temporal* itself is the parent of *xkos:before* and *xkos:after*.

It was found useful to add two more precise sub-properties of *xkos:precedes* and *xkos:succeeds*, namely *xkos:previous* and *xkos:next*. Previous and next imply that there is no intermediary concept

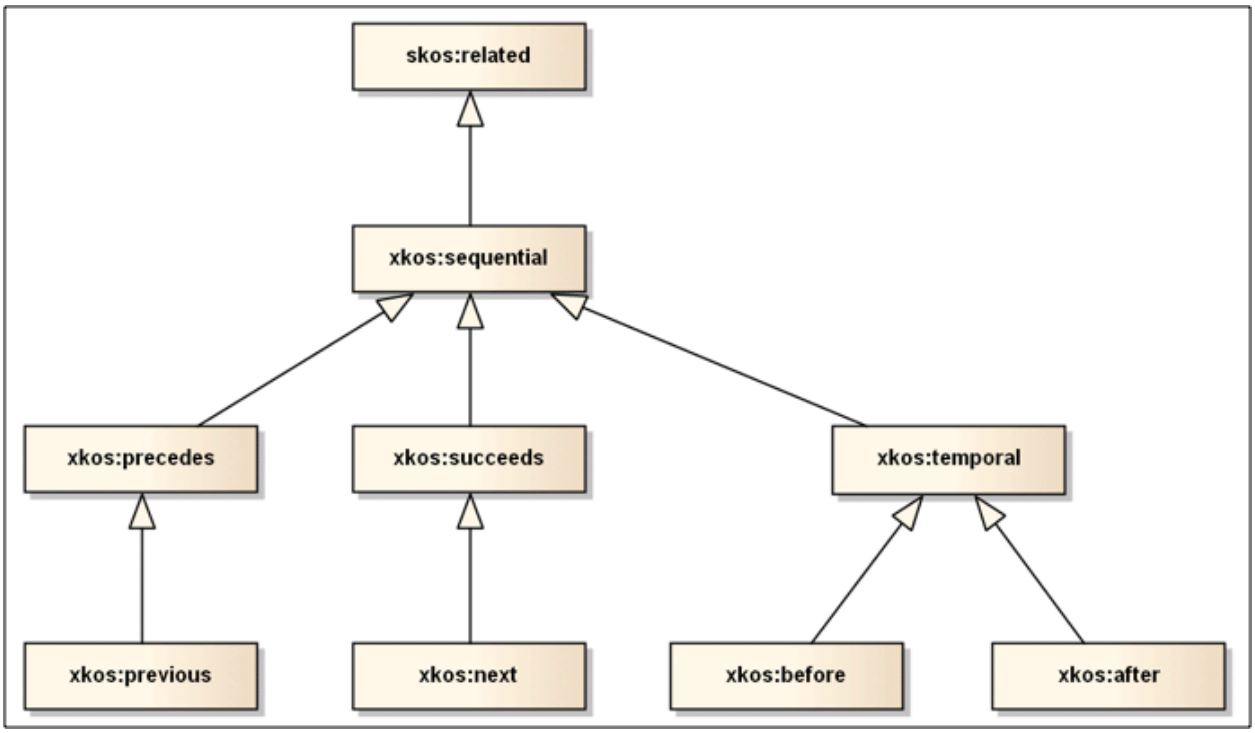


Figure 5: XKOS Sequential Properties

between two sequentially linked concepts. These two properties are of course not transitive, although their parents are.

Conclusion

In this paper, we laid out the general rationale and purpose for why XKOS was developed. We explained the basic extensions to SKOS that were identified as needed to describe statistical classifications in the LOD domain, and we gave examples from the statistical community to justify our choices. Some unresolved issues were also discussed. Finally, we gave a rationale for the importance of SKOS and XKOS, appealing to the burgeoning LOD community of practice, the use of RDF, and the growth of the Semantic Web in general.

It is interesting that some of the extensions (*generic* and *partitive* relations) were originally included in SKOS. Given the amount of discussion in the LOD and Semantic Web communities about semantics and precision, it is even more remarkable that these specific relations were left out. On the other hand, there was a clear desire by the SKOS designers to make building Semantic Web applications as simple as possible. Since SKOS is the Simple Knowledge Organization System, this design choice begins to make sense.

Yet, we have also seen that the worlds of thesauri and classifications are often too complex to model in SKOS. Thus, the vocabulary was quickly extended with SKOS-XL to handle the need to treat labels, not as literals, but as actual class instances (a process sometimes referred to as *reification*) that could participate in relationships with other instances and have properties of their own. While SKOS-XL extends SKOS for the particular needs of the multi-lingual thesaurus community, XKOS adds the extensions that are desirable to meet the requirements of the statistical community.

SKOS is a very popular specification, and we hope the XKOS extensions will simply serve to increase its adoption. The proof of whether XKOS is useful will be found when statistical offices implement it. This work is already underway. However, XKOS is still a work in progress, and unresolved issues remain. We hope the users of XKOS will offer help with these issues, provide comments to the authors on the effectiveness of XKOS, and give guidance as to what other areas should be extended as we prepare to submit the standard as a W3C Editor's Draft.

Acknowledgements

The authors wish to thank the organizers of the Dagstuhl workshops – Richard Cyganiak, Arofan Gregory, Wendy Thomas, and Joachim Wackerow – for their support and encouragement in developing the XKOS ideas. The authors also wish to thank the participants not already mentioned in the XKOS development group: Thomas Bosch, Rob Grim, and Jannik Jensen.

Notes

1. Franck Cotton, Institut National de la Statistique et des Études Économiques. Daniel W. Gillman, US Bureau of Labor Statistics, Yves Jaques, Food and Agriculture Organization of the United Nations.

The opinions in this paper are those of the authors only and do not necessarily reflect the policies and programs of the Institut National de la Statistique et des Études Économiques, the US Bureau of Labor Statistics, or the Food and Agriculture Organization of the United Nations.

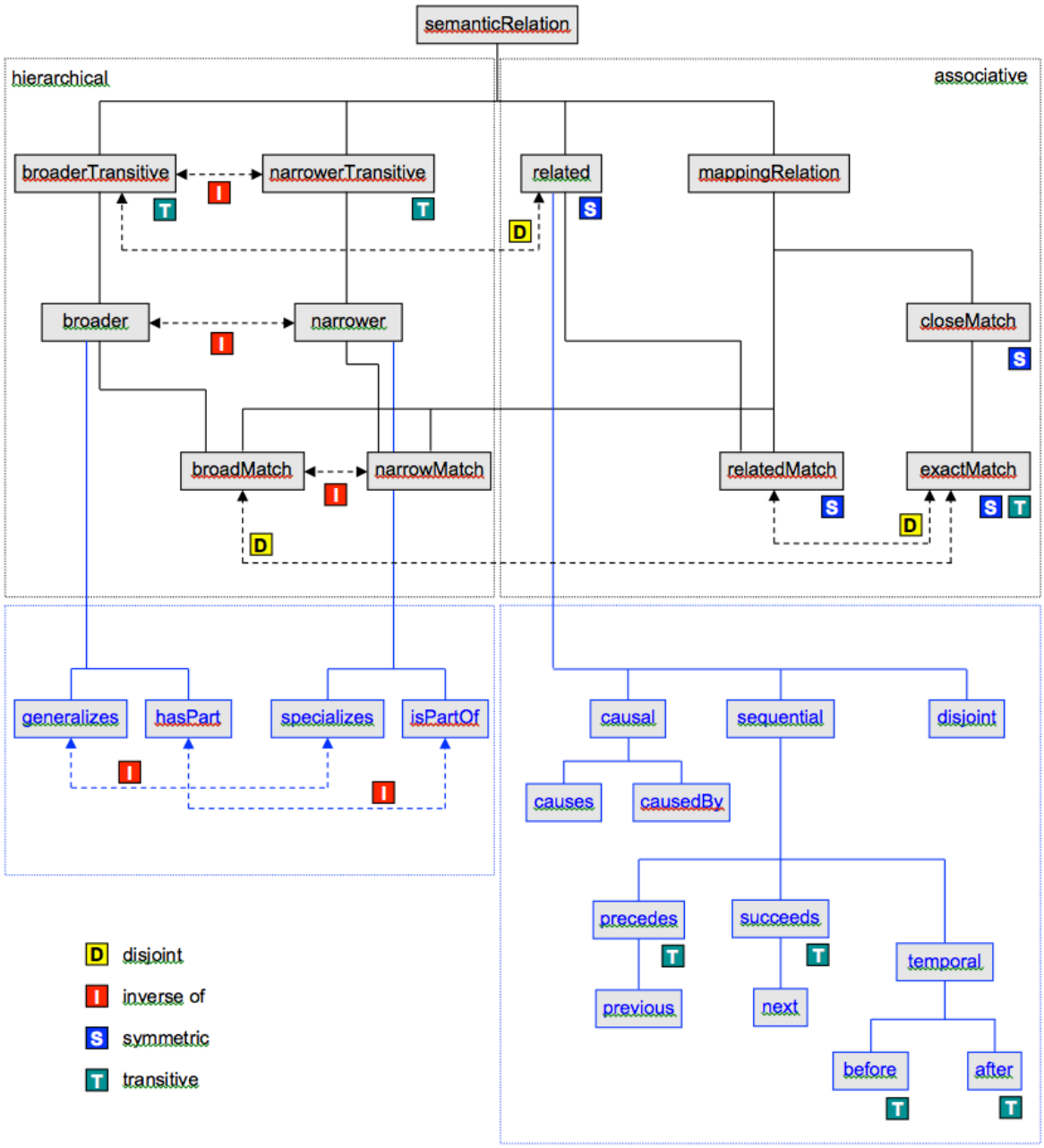
2. http://www.w3.org/2004/02/skos
3. http://www.w3.org
4. http://en.wikipedia.org/wiki/Semantic_Web and http://semanticweb.com/tag/tetherless-world-constellation
5. http://linkeddata.org
6. http://www.w3.org/DesignIssues/LinkedData.html

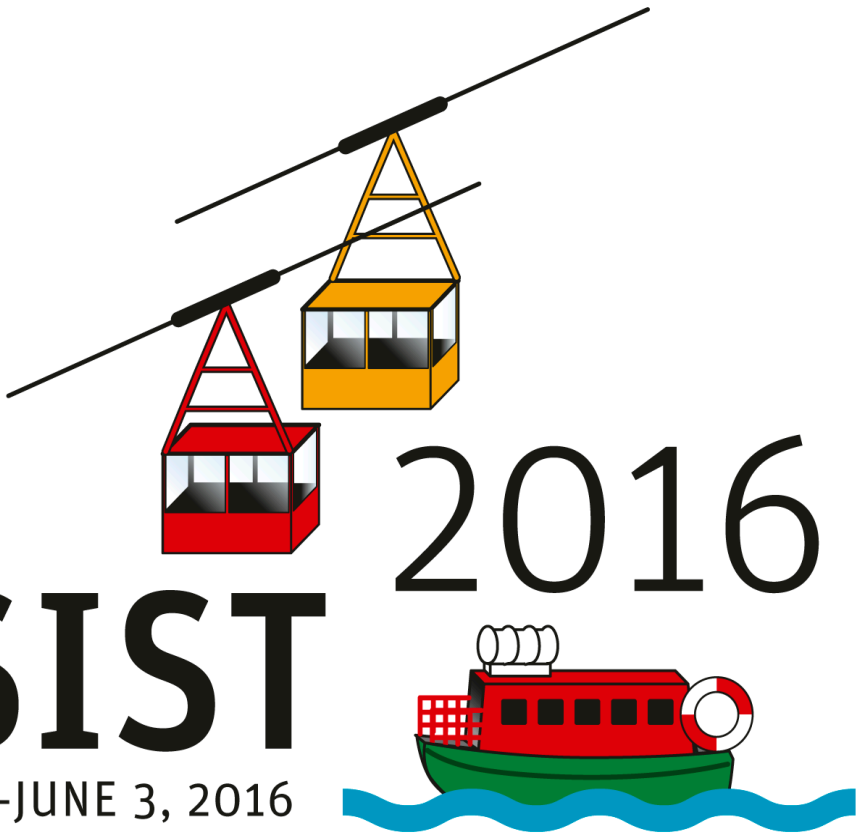
7. http://www.w3.org/RDF
8. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2010/wp.4.e.pdf
9. http://linkeddata.org
10. http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
11. http://data.gov.uk
12. http://www.dagstuhl.de
13. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=11372
14. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=12422
15. http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/implementation.html
16. http://www.w3.org/TR/rdf-schema/
17. http://tools.ietf.org/html/rfc3986
18. http://www.ietf.org/rfc/rfc1738.txt
19. http://xmlns.com/foaf/spec/
20. http://www.w3.org/TeamSubmission/turtle/
21. http://www.obitko.com/tutorials/ontologies-semantic-web/rdf-graph-and-syntax.html
22. http://www.w3.org/TR/skos-primer/
23. http://www.bls.gov/soc/
24. http://www.bls.gov/iif/oshoiics.htm
25. http://www.bls.gov/tus/lexicons.htm
26. http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=14319930
27. http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework
28. http://eurovoc.europa.eu/
29. http://id.loc.gov/authorities/subjects.html
30. http://dublincore.org/documents/dcmi-terms/
31. http://www.w3.org/ns/radion
32. http://unstats.un.org/unsd/cr/registry/isic-4.asp
33. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-015/EN/KS-RA-07-015-EN.PDF

Annex 1

SKOS and XKOS properties relating concepts

Note: SKOS properties are in the two upper boxes, XKOS in the two lower.





IASSIST 2016 will take place in Bergen, Norway, hosted by the Norwegian Social Science Data Services.

For any questions - please contact:
heidi.tvedt@nsd.uib.no

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Service and Technology** (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: *1 year, on: July 1st*
Automatic renewal: *no*

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page