Semantic Web Applications for the Social Sciences

Thomas Bosch[1] and Benjamin Zapilko[2]

Abstract

In recent years, semantic technologies have become mature and applicable. They have found their way into various domains, like bio-informatics and eGovernment, where they are used in different applications and provide value-added services for users. In this paper, we present an overview on several representative applications, which use Semantic Web technologies, and show the potential of these technologies and Linked Data in an application context. We then present existing Semantic Web applications for the Social Sciences in detail and highlight their impact on this domain. The takeover of the Semantic Web vision on the Social Sciences results in clear benefits, which we identify and discuss.

Keywords

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

1 Introduction

The 'corporate' landscape is moving. 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 using Semantic Web. Others are using it (or consider using it) as part of their own operations such as Novartis, Pfizer, and Telefónica. Some of the names of active participants in W3C Semantic Web related groups are HP, Agfa, SRI International, Fair Isaac Corp., Oracle, Boeing, IBM, Chevron, Siemens, Nokia, Pfizer, and Eli Lilly. Major communities such as digital libraries, defence, eGovernment, energy sector, financial services, health care, oil and gas industry, and life sciences pick the technology up.

For social science researchers, the Semantic Web and Linked Data hold great promise as Gregory and Vardigan (2010) illustrate in detail. The adaption of semantic technologies enables many possibilities for making the discovery of data and metadata in the Web more efficient, for enhancing the reuse of social science (meta-)data, and for decreasing technical barriers to utilize such data in research. Through Linked Data for the social sciences, users can easily discover the existence of data. They can get details on how the (meta-)data is structured and whether it is suitable for their interest. Therefore, it is necessary that the data is well-documented and that quality and provenance can be exposed. This allows for identifying complex relationships like whether data sets are comparable or whether there are relationships to other versions of the identical data set. Since data collections published as Linked Data can easily be linked with each other, the integration and merging of heterogeneous data sets is facilitated.

In this paper, we provide an overview on tools and applications, which apply Semantic Web technologies and Linked Data. We also present projects and applications which are available or are

planned for the Social Sciences and make use of semantic technologies. Also, we discuss the benefits of the appliance of these technologies for the domain of the Social Sciences.

2 Semantic Web Applications

In this section, we describe applications, which are utilizing Semantic Web technologies and Linked Data and which show the potential of these technologies for their domains and heterogeneous communities. The applications Fedora Repository Project and Virtuoso are popular web repository and server systems, which allow the storage, maintenance and access of data on the Web. PoolParty Thesaurus Server and VocBench are management tools for knowledge organization systems like thesauri, taxonomies, etc. The VIVO Project enables searching and browsing in a network of researchers, their activities and organizations.

2.1 HCLS Demo

Herman (2011) has delivered a presentation about the HCLS demonstration (W3C 2008a). The W3C Health Care and Life Sciences Interest Group (HCLS) have developed demonstrations on the usage of Semantic Web technologies. These demonstrations should show the HCLS community how Semantic Web can be used and 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 part 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 own 'maps', i.e. the URIs of bitmap images. The user then gets the navigation on 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 use SPARQL to combine these into more complex queries that result in the right image URIs, to be then fed into the Google service. Via RDF, one gets a standard query SPARQL interface for free, that gives much power to the end user. And, indeed, 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 allowed you to make it easily without interfering with the original data in any way, and using standard, off-the-shelf tools. The demo shows that, sometimes, RDF and SPARQL helps in a very simple way - Semantic Web applications are not necessarily very complex.

2.2 NASA

Grove and Schain (2008) describe how to find the right experts at NASA. The NASA tool is an expertise locater for nearly 70,000 NASA civil servants. The Semantic Web tool uses RDF integration techniques of over 6 or 7 geographically distributed databases, data sources, and web services.

They 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 a specific

knowledge expertise. The dump results from a faceted browser developed by the company to view result data.

2.3 Semantic MediaWiki

Semantic MediaWiki (2012) is a module to the MediaWiki (2012) software and extends wikis with ideas from the Semantic Web discipline. Semantic MediaWiki enables to make facts available for machines. And this makes it easier for humans to search and reuse information. Articles, like an article about the IASSIST conference in 2012, can be annotated semantically using the RDF format. In this way, RDF triples can be built and stored in the wiki code directly. The IASSIST conference in 2012 would be the subject of the RDF triples stated on the article. Relations between subjects and objects can be defined semantically. 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 programs 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 actualize themselves when the depending sites are actualized. 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 meta-data 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.

2.4 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. Besides 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, which can be applied for various scenarios on storing, maintaining and accessing different kinds of data. Thus, 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 open source edition at http://www.openlinksw.com/wiki/main/.

Considering 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. It is planned to extend the access and storage capabilities of connected databases, which would also enable particular technologies like inferencing on RDF data.

2.5 Thesaurus Management Tools

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

PoolParty Thesaurus Server

PoolParty Thesaurus Server (PPT) (2013) is a software platform, which allows the management and maintenance of complex knowledge models like taxonomies, thesauri, controlled vocabularies and similar data. The metadata of this data is fully organized and modeled using W3C´s Semantic Web standards RDF and SKOS, since SKOS is focusing 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 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 data sets hold at FAO like the AGROVOC thesaurus (2013), the Biotechnology Glossary (2013), and bibliographic metadata used in FAO. For future releases it is planned to include a native interface for SKOS and SKOS-XL and configurable support for hosting of different triple store technologies. Also, it is planned to support generic OWL ontologies.

2.6 Vivo Project

The VIVO project (2013) is an open source semantic web application, which has been 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. By 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 an integrated searching and browsing in 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.

VIVO uses not only semantic technologies. 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.

3 Semantic Web Applications for the Social Sciences

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 of the social sciences community regarding Semantic Web supported functionalities. 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 on the variable and on the study level. NESSTAR enables to publish a huge amount of statistical data and metadata using Semantic Web technologies and Colectica is a fast way to design, document, and publish your survey research using open data standards.

3.1 SofisWiki

Are you performing a social science research project? Are you writing a thesis or are you habilitating about a social science topic? If yes, then 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 10 years of SOFIS (2012), a central database, hosted by GESIS, about social science research activities in the German-speaking countries. The database delivers information about over 50,500 research projects. Using SOFISwiki, research institutions like universities can enter and actualize information about their research projects and their institution 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 in order to support the documentation of research activities, projects, and institutions all over the world. Figure 1 visualizes the graphical user interface of SOFISWiki offered to search for research institutions by institution name, country and location, content orientation, and institution type.

Figure 1. SofisWiki – search for research institutions

Figure 2 shows the result of the above query for the department ‘Monitoring Society and Social Change’ of the research institution ‘GESIS’. For this department metadata like address, homepage, contact person as well as the associated research projects are delivered.

Figure 2. SofisWiki – research institution and associated research projects

Benefits for the social sciences community

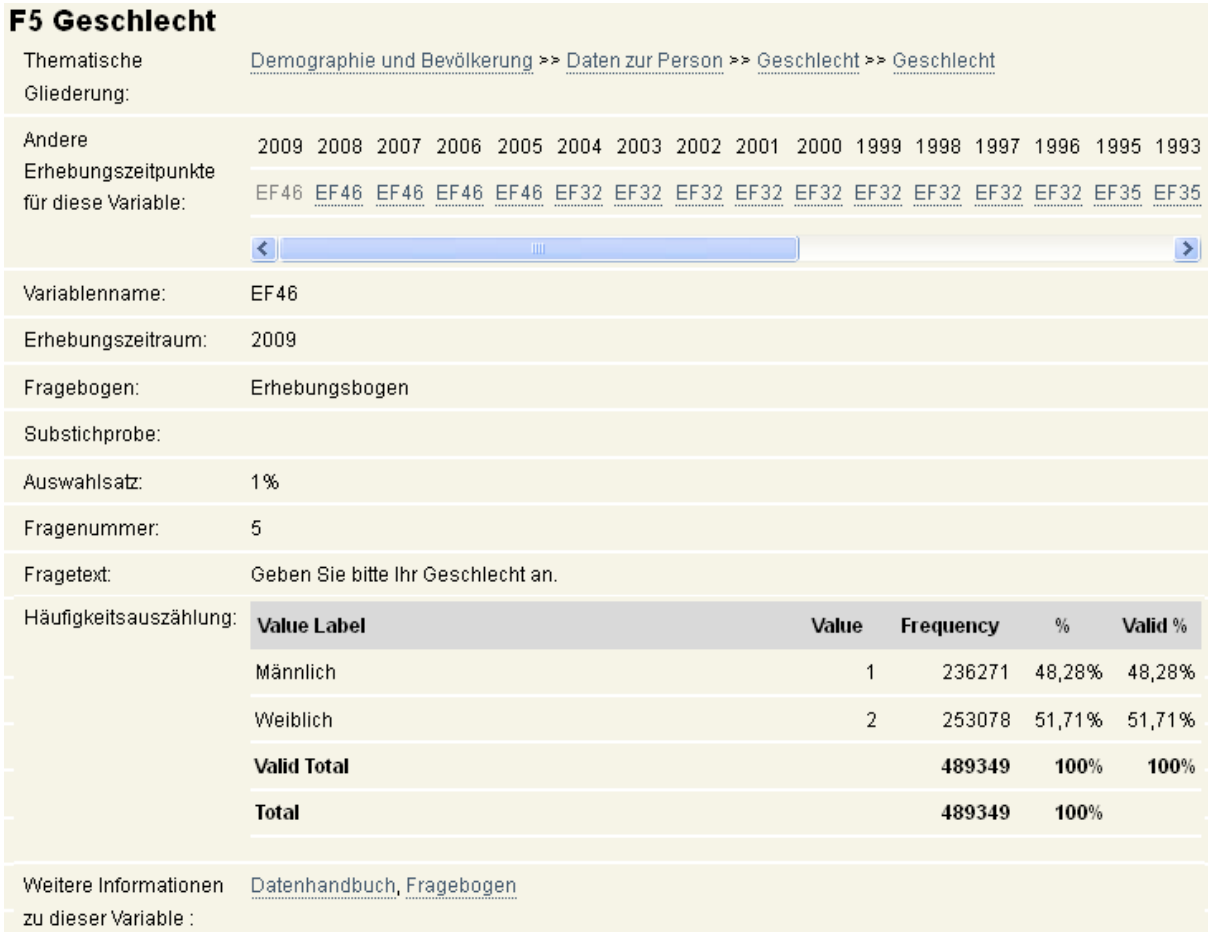
Research institutions like universities can manage and publish information about their research institution and their research projects. Researchers can get an overview of research projects, institutions, and activities. Internally, the metadata is represented using RDF to ease the integration of the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

3.2 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. MISSY provides detailed information about individual data sets. Currently, MISSY consists of the german microcensus survey, which is comprised of statistics about the general population in Germany, the situation about the employment market (occupation, professional education, income, legal insurance). MISSY consists of approx. 500 variables and questions and captures 25 years, since 1973. Figure 3 shows the graphical user interface of MISSY. In

the visualized variable detail view you get details about the variable gender like associated question, values, value labels, absolute, and relative frequencies.



F5 Geschlecht

Thematische Gliederung:	Demographie und Bevölkerung >> Daten zur Person >> Geschlecht >> Geschlecht
Andere Erhebungszeitpunkte für diese Variable:	2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 1993 EF46 EF46 EF46 EF46 EF46 EF32 EF32 EF32 EF32 EF32 EF32 EF32 EF32 EF32 EF35 EF35
Variablenname:	EF46
Erhebungszeitraum:	2009
Fragebogen:	Erhebungsbogen
Substichprobe:	
Auswahlsatz:	1%
Fragenummer:	5
Fragetext:	Geben Sie bitte Ihr Geschlecht an.

Häufigkeitsauszählung:

Value Label	Value	Frequency	%	Valid %
Männlich	1	236271	48,28%	48,28%
Weiblich	2	253078	51,71%	51,71%
Valid Total		**489349**	**100%**	**100%**
Total		**489349**	**100%**	

Weitere Informationen zu dieser Variable : Datenhandbuch, Fragebogen

Figure 3. MISSY – graphical user interface

MISSY may be split in two parts:

- the Missy Web, the end-user front-end part and
- the Missy Editor for the metadata documentation, the back-end part.

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, further surveys like 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 Discovery Vocabulary – the DDI Ontology (Bosch et al. 2012). The reasons why the MISSY data model is built on top of this ontology are the followings:

- The DDI Ontology contains the most important components of both DDI-Codebook as well as DDI-Lifecycle (as a consequence, not all of the over 830 XML elements of DDI 3.1 are covered),
- the DDI Ontology serves as a first step to a model-driven further development of the DDI metadata standard to document microdata within the social, behavioral, and economic sciences,
- the DDI Ontology will be officially published at the end of the year 2013, and
- more than 20 experts from the statistics and the Linked Data community of 8 different countries have contributed to the development of the DDI ontology within 3 workshops and further 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 – the DDI Ontology. Other software projects with the purpose to document studies on study and variable level using DDI-L can also be based on the DDI Ontology and can reuse existing code which is made available on a GitHub repository (MISSY 3 2012). To show how this works, an example is now given. The SKOS (Simple Knowledge Organization System) (W3C 2009), whose purpose is to define hierarchies of concepts, is reused in the DDI Ontology to a large extend (see figure 4).

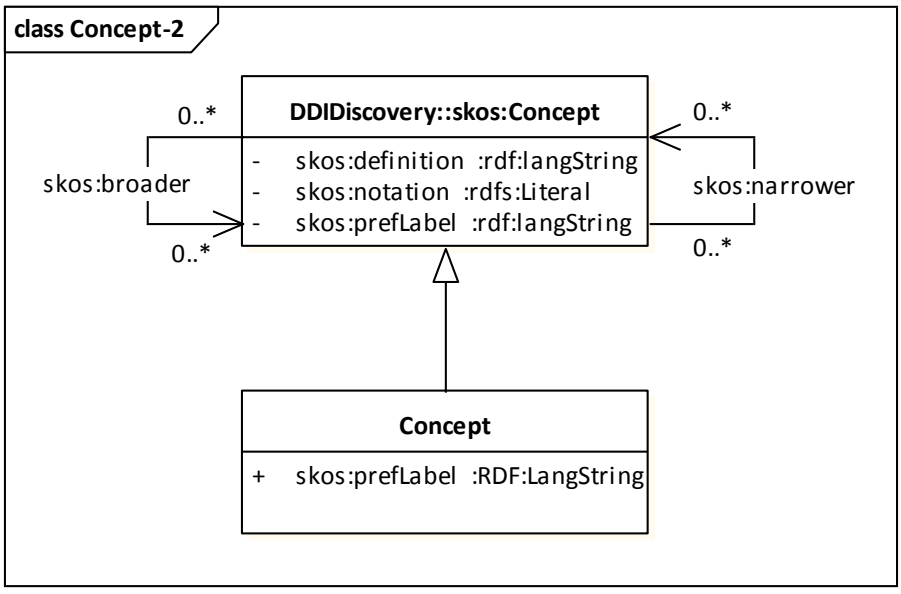


Figure 4. MISSY – abstract and individual data model

For instance, codes, categories, DDI concepts, and study subjects are represented in the DDI Ontology 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 like 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 is 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 persistency 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), that are generic annotations in XHTML documents. Thus, machines can crawl the MISSY websites in order to import exactly the information needed for further processings, since now machines 'know' the meanings of the provided information. RDFa metadata should and will be provided according to the DDI Ontology 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 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 both containing the data and metadata of multiple studies offered by MISSY.

Benefits for the social sciences community

Other software projects documenting studies on 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 LOD 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.

3.3 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 extraordinary easy and simple to understand SW application, as it does not specify sophisticated ontologies, it does not use advanced RDF features such as reification, and it 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 is only available in a human readable and understandable form and not additionally in a machine understandable form which can be further processed by computer programs. NESSTAR should revolutionize the way people access statistical information, as it should bring 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.

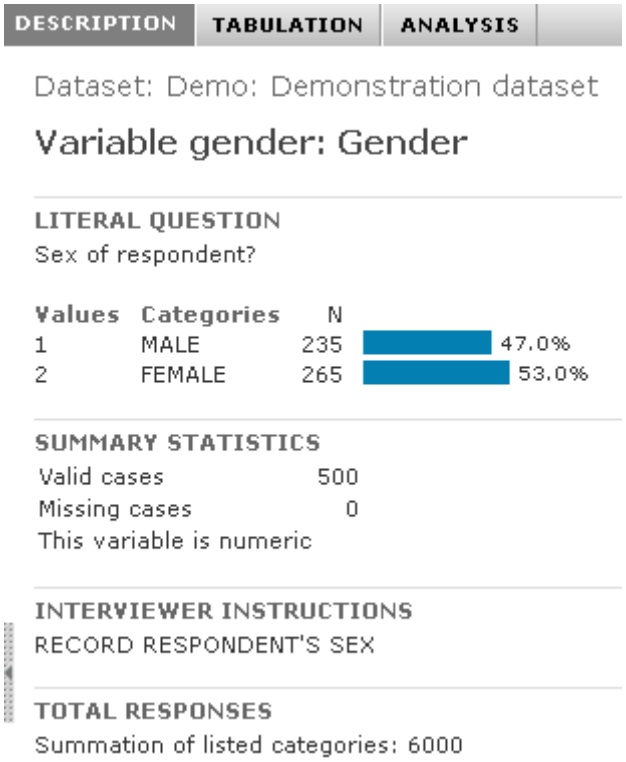


Figure 5. NESSTAR – variable gender

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 statistical 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.

NESSTAR is based on a lightweight object-oriented and web middleware which is named NEOOM - NEsstar Object Oriented Middleware. NEOOM is based on Web and Semantic Web standards like HTML, HTTP, RDF, and RDFS. NEOOM is considered as a set of guidelines of 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 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: the NESSTAR Explorer, the NESSTAR Light Explorer, the NESSTAR Publisher, and the Object Browser. The NESSTAR Explorer is very 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 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.

3.4 Colectica RDF Services

Colectica is a fast way to design, document, and publish your 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 state 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).

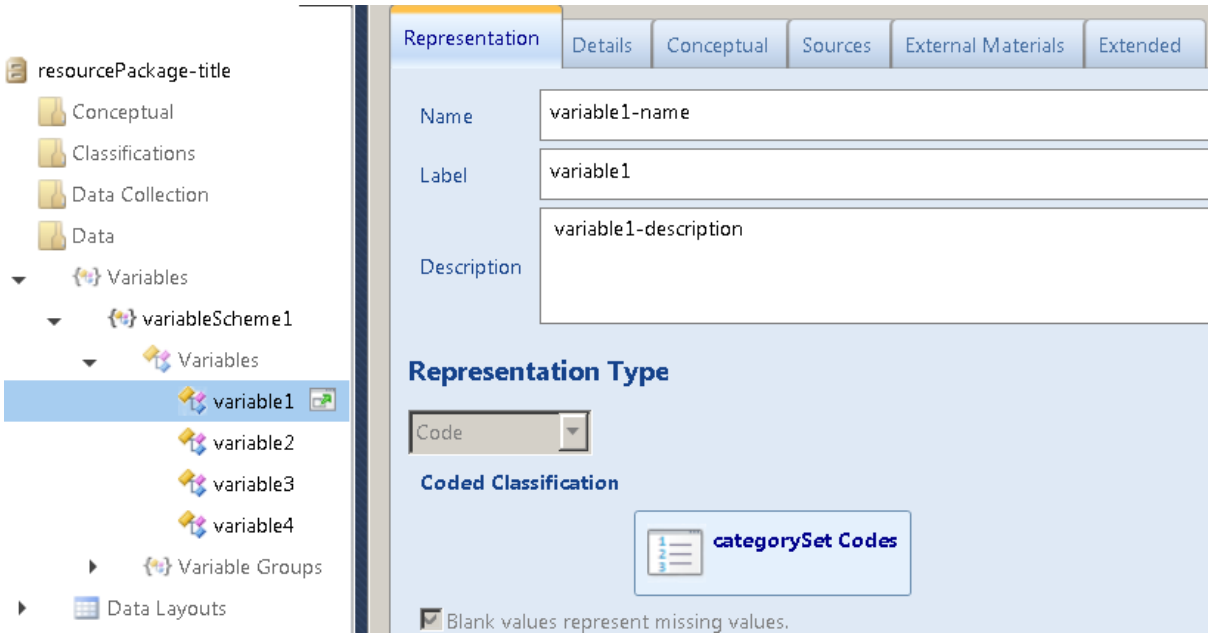


Figure 6. Colectica – variables

Figure 7 presents how to document questions using Colectica Designer. For questions you can state question names, question texts, the question scheme, and the response domain (e.g. numeric, textual, or coded).

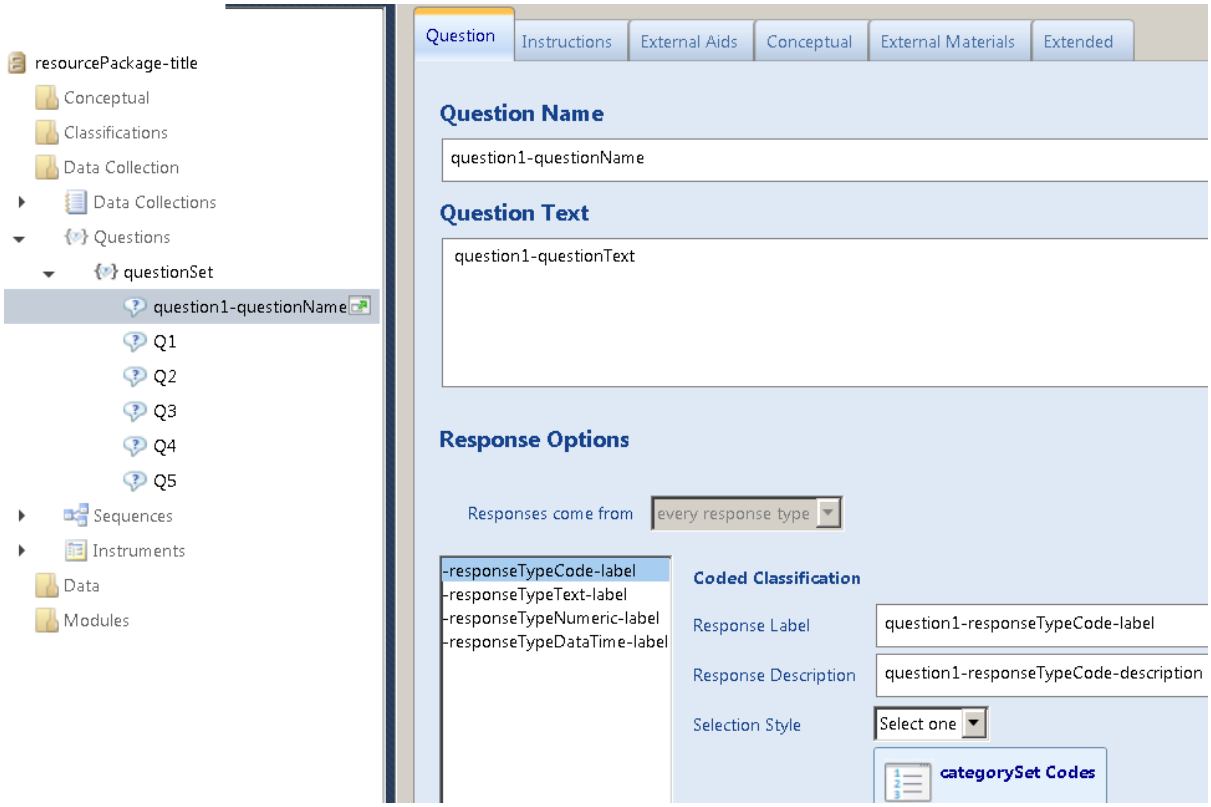


Figure 7. Colectica – questions

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).

You 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 data sets (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.

4 Conclusion and Future Work

We have presented several representative applications that apply Semantic Web technologies at a high degree. While semantic technologies and Linked Data have yet not been widely used in the Social Sciences, we could identify initial applications exclusively developed for this domain. The impact of Semantic Web and Linked Data could be 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 11 April 2013, <http://aims.fao.org/standards/agrovoc/about>

Assini, P 2001a 'Objectifying the Web the 'light' way: an RDFbased 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 [15 December 2012].

Assini, P 2002, 'A Semantic Web Application for Statistical Data and Metadata', Proceedings of the International World Wide Web Conference, Hawaii, 07 Mai 2002, eleventh International World Wide Web Conference.

Biotechnology Glossary 2013 Biotechnology Glossary, viewed 11 April 2013, <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 16 December 2012, <http://www.iassist2012.org/indexfolder/program/files/41_IASSIST2012_Colectica_RDF_Services.pdf>.

Colectica 2012b DDI 3 meets RDF and SPARQL with Colectica Repository, viewed 16 December 2012, <http://dan.smith.name/2011/10/ddi-3-meets-rdf-and-sparql-with-colectica-repository/>.

Colectica 2012c Colectica Website, viewed 16 December 2012, <http://www.colectica.com/>.

Faster 2012 Faster, viewed 16 December 2012, <http://fasterproject.eu/>.

Fedora 2013 Fedora Repository Project, viewed 11 April 2013, <http://fedora-commons.org/>.

GESIS 2012 The Microdata Information System (MISSY), viewed 15 December 2012, <http://www.gesis.org/missy>.

Gregory, A & Vardigan, M 2010 The Web of Linked Data: Realizing the Potential for the Social Sciences.

Grove, M & Schain, A 2008 Case Study: POPS — NASA's Expertise Location Service Powered by Semantic Web Technologies, viewed 15 December 2012, <http://www.w3.org/2001/sw/sweo/public/UseCases/Nasa/>.

Herman, I 2011 Semantic Web Adoption and Applications, viewed 15 December 2012, <http://www.w3.org/People/Ivan/CorePresentations/Applications/>.

MediaWiki 2012 MediaWiki.org, viewed 14 December 2012, <http://www.mediawiki.org/wiki/MediaWiki/de>.

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 15 December 2012, <https://github.com/missy-project>.

Nesstar 2013, Welcome to Nesstar's Demo Server, viewed 01 April 2013, < http://nesstar-demo.nsd.uib.no/webview/>.

Norwegian Social Science Data Services 2012 Nesstar, viewed 15 December 2012, <http://www.nesstar.com/>.

PPT 2013 PoolParty Thesaurus Server, viewed 11 April 2013, <http://poolparty.biz/products/poolparty-thesaurus-manager/>

Schema.org 2012 Schema.org, viewed 15 December 2012, <http://schema.org/>.

Semantic MediaWiki 2012 Semantic MediaWiki, viewed 14 December 2012, <http://semantic-mediawiki.org>.

SOFIS 2012 SOFIS - Sozialwissenschaftliches Forschungsinformationssystem, viewed 14 December 2012, <http://www.gesis.org/unser-angebot/recherchieren/sofis/>.

SOFISWiki 2012 SOFISWiki, viewed 14 December 2012, <http://www.gesis.org/sofiswiki/Hauptseite>.

Sowiport 2012 Sowiport, viewed 14 December 2012, <http://www.gesis.org/sowiport>.

Virtuoso 2013 Virtuoso Universal Server, viewed 11 April 2013, <http://virtuoso.openlinksw.com/>

VIVO 2013 VIVO Project, viewed 11 April 2013, <http://www.vivoweb.org/>

VocBench 2013 VocBench, viewed 11 April 2013, <http://aims.fao.org/tools/vocbench-2>

W3C 2007 SOAP Version 1.2 Part 0: Primer (Second Edition) - W3C Recommendation 27 April 2007, viewed 16 December 2012, <http://www.w3.org/TR/2007/REC-soap12-part0-20070427/>.

W3C 2008a HCLS/Banff2007Demo, viewed 14 December 2012, <http://www.w3.org/wiki/HCLS/Banff2007Demo>.

W3C 2008b, SPARQL Query Language for RDF, viewed 16 December 2012, <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 15 December 2012, <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 16 December 2012, <http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/>.

[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