DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2012

Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences

Thomas Bosch
GESIS – Leibniz Institute for the Social Sciences, Germany
thomas.bosch@gesis.org

Richard Cyganiak
Digital Enterprise Research Institute, Ireland
richard@cyganiak.de

Joachim Wackerow
GESIS – Leibniz Institute for the Social Sciences, Germany
joachim.wackerow@gesis.org

Benjamin Zapilko
GESIS – Leibniz Institute for the Social Sciences, Germany
benjamin.zapilko@gesis.org

Abstract

Experts from the statistical domain worked in close collaboration with ontology engineers to develop an ontology of a subset of the Data Documentation Initiative, an established international standard for the documentation and management of data from the social, behavioral, and economic sciences. Experts in the statistics domain formulated use cases which are seen as most significant to solve frequent problems. Various benefits for the Linked Data and the statistics community as well are connected with an RDF representation of the developed ontology. In the main part of the paper, the DDI conceptual model as well as implementations are explained in detail.

Keywords: Semantic Web; ontology design; RDF; DDI; statistical data

1. Introduction

The Data Documentation Initiative (DDI)[1] is an acknowledged international standard for the documentation and management of data from the social, behavioral, and economic sciences. The DDI metadata specification supports the entire research data lifecycle. The focus is on microdata–data collected on an individual object from a survey or administrative source. Aggregated data can also be described. So far, the DDI data model is expressed in XML Schema. We developed DDI-RDF, an OWL ontology for a basic subset of DDI to solve the most frequent and important problems associated with diverse use cases and to open the DDI model to the Linked Open Data[2] community. Possible use cases are mapping search terms to external thesaurus concepts, finding publications and linkage to publications related to specified data, and discovery of data and metadata connected with multiple studies. There are two parallel ways to implement the mapping between DDI-XML document instances and an RDF representation of the DDI data model. A direct mapping on the one side and a generic transformation on the other side can be distinguished. The generic approach can be applied not only within the framework of the DDI, but elsewhere. The benefits for the DDI community are to publish DDI data as well as metadata in the Linked Open Data cloud[3] as RDF data. As a consequence, RDF tools can process DDI instances without supporting the DDI-XML Schemas' data structures. After publishing public available structured data, DDI data and metadata may be linked with other data sources of multiple topical domains. With the possibilities of Semantic Web technologies, requesting

[1] http://www.ddialliance.org/
[2] http://linkeddata.org/
[3] http://lod-cloud.net/

https://doi.org/10.23106/dcmi.952135935

multiple, distributed and merged DDI instances are possible. This work started within the context of a workshop on semantic statistics in Schloss Dagstuhl - Leibniz Center for Informatics, Germany in September 2011[4] and was continued in a working meeting in collocation with the 3rd Annual European DDI Users Group Meeting in Gothenburg, Sweden[5].

2. Data Documentation Initiative (DDI)

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 accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, re-purposing, 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 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. Public accessible metadata of good quality are important for finding the right data. This is especially the case if access to microdata is restricted because a disclosure risk of the observed people exists. DDI is currently specified in XML Schema, organized in multiple modules corresponding to the individual stages of the data lifecycle, and is comprised of 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 means of DDI. This approach supports applications which understand the Dublin Core XML, but which 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, 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 of 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 the transparency. DDI enables the re-use 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., 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

[4] http://www.dagstuhl.de/11372
[5] http://www.iza.org/eddi11
[6] http://sdmx.org/
[7] http://www.ddialliance.org/ddi-at-work/projects
[8] http://www.dataarchive.ac.uk/

https://doi.org/10.23106/dcmi.952135935

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 for international organizations 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. Since 2003, the DDI Alliance develops and promotes 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. Common understanding is that both statistical data and metadata are part of a data lifecycle. 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 should be done when it happens. The metadata could be then re-used along the data lifecycle. Such practices would incorporate documenting 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. Multiple institutions are involved in the data lifecycle which is an interactive process with multiple feedback loops. Figure 1 displays the data lifecycle which is described in more detail on the DDI Alliance website[16].

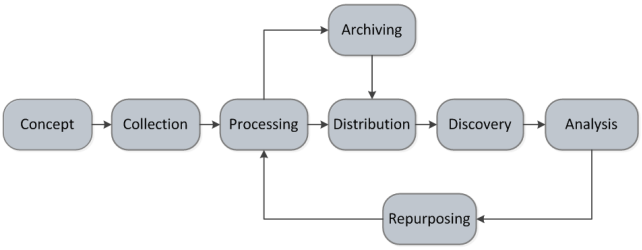

FIG. 1. DDI Data Lifecycle

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.

[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

https://doi.org/10.23106/dcmi.952135935

3. Related Work

Beyond the Semantic Web, there are several relevant metadata standards like SDMX (Statistical Data and Metadata Exchange) for the representation and the exchange of aggregated data, ISO 19115 (ISO 19115-1:2003, 2003) for geographic information, and PREMIS[17] for preservation purposes. The metadata registry ISO 11179 (ISO/IEC, 2004) marks a standard for the modeling of metadata, e.g. reference models, and for their registry. Elements are often used as top-level components, while other standards and concrete implementations are derived. But beside standards in XML for describing and documenting such complex metadata models, there are yet only few adequate RDF-based vocabularies. DDI-RDF has a clearly defined focus on describing microdata, which has not yet been covered to this extent by other established vocabularies. Therefore it applies 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 if suitable for representing basic information about publishing objects on the web as well as for hasPart relationships. For representing concepts, which are organized similar as 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 ownerships and can be used to represent and interchange 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.

An established RDF metadata vocabulary, which seems to be very 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 multi-dimensional tables. Aggregated data are 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. In the Data Cube vocabulary associated metadata is added.

4. DDI as Linked Data

In this section, we present the development process from the DDI-XML metadata standard to the DDI-RDF ontology for exposing DDI data according to Semantic Web standards. The benefits for the Linked Data community lie at hand, as there is currently no such ontology with a comparable level of detail for representing complex entities and relations regarding the complete lifecycle of research data as DDI-RDF provides. The publication of research data in the web of data became popular and important in various domains beside the Social Sciences, so that a valuable contribution can be seen in the introduction of DDI-RDF. The benefits for the DDI community are to publish DDI data as well as metadata in the Linked Open Data cloud as RDF data. As a consequence, DDI instances can be processed by RDF tools without supporting the DDI-XML Schemas' data structures. After publishing public available structured data, DDI data

[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/

https://doi.org/10.23106/dcmi.952135935

and metadata may be linked with other data sources of multiple topical domains. With the possibilities of Semantic Web technologies, requesting multiple, distributed and merged DDI instances will be possible.

Conceptual Model. Figure 2 visualizes in detail the conceptual model including the DDI elements (subset of the whole DDI model) which are seen by various statistical domain experts as most important to resolve problems associated with diverse identified use cases. Experts comprehend core members of the DDI Alliance Technical Implementation Committee, representatives of national statistical institutes, national data archives, and the Linked Open Data community (see Acknowledgements). With the background of the broadness and complexity of DDI, the first version of DDI-RDF focuses on a subset of DDI. The selection relies on use cases which are oriented on the discovery of data in the Linked Data context and possible usage within the web of data. The conceptual model is based on XML Schemas describing the DDI domain data model with extensions that partly borrow from existing vocabularies and partly reside in a new DDI vocabulary. Only relations between exactly two DDI elements and not between one DDI element and an instance of an XML Schema datatype are displayed in the figure, in order to reduce the complexity of the overall conceptual model. Where necessary, individual datatype properties are described in the use cases section. The three components of the DDI conceptual model 'Study', 'Variable', and 'LogicalDataSet' are seen as the most important parts of the data model. Because of this, they are highlighted and outgoing relations are displayed in three different colors.

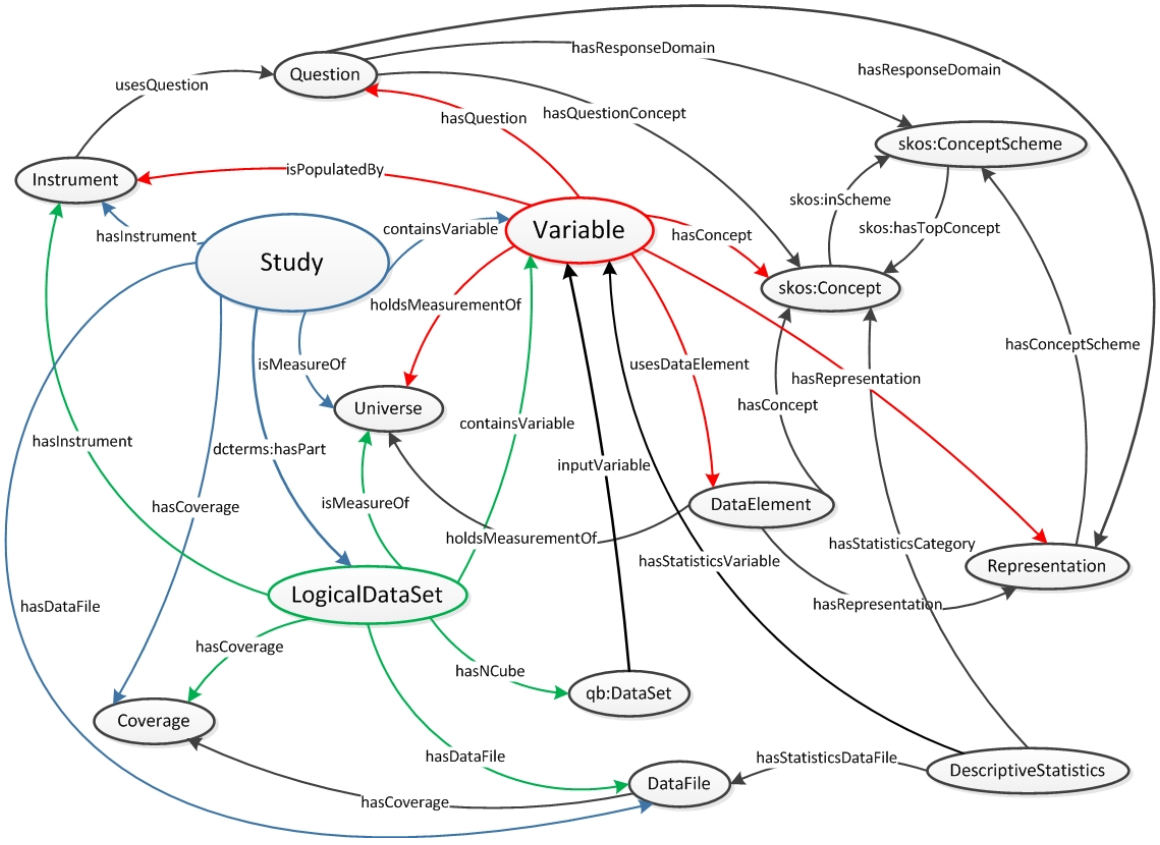

FIG. 2. Conceptual Model

There are features of DDI which can be addressed through other vocabularies, such as: describing metadata for citation purposes using Dublin Core, describing aggregated data like multi-dimensional tables using the RDF Data Cube Vocabulary, and delineating code lists,

https://doi.org/10.23106/dcmi.952135935

category schemes, mappings between them, and concepts like topics using SKOS. Widely adopted and accepted vocabularies are reused to a large extend. Diverse relations between DDI elements (e.g. dcterms:hasPart), between classes defined in other namespaces (e.g. skos:inScheme or skos:hasTopConcept) and between DDI elements and XML Schema datatypes (e.g. dcterms:identifier or skos:definition) have been implemented using object and datatype properties from the Dublin Core and the SKOS ontologies. Overall, two object properties and 13 datatype properties are re-used from the dcterms namespace.

Figure 2 visualizes the terms described in the following and explained in examples in the use case section. A simple '**Study**' supports the stages of the full data lifecycle in a modular manner. This does not comprehend groups of studies (like repeated annual surveys). The key criteria for a study are: a single conceptual model (e.g. survey research concept), a single instrument (e.g. questionnaire) made up of one or more parts (ex. employer survey, worker survey), and a single logical data structure of the initial raw data (multiple data files can be created from this such as a public use microdata file or aggregate data files) (DDI Alliance, Technical Specification, Part I, 2009). The DC datatype properties 'dcterms:abstract', 'dcterms:title', and 'dcterms:identifier' are used to describe studies.

'Concept', the 'Universe', and the 'Coverage' define a study. SKOS defines the term '**Concept**', which is a unit of knowledge created by a unique combination of characteristics (ISO 1087- 1:2000, 2000). In context of statistical (meta)data, concepts are abstract summaries, general notions, knowledge of a whole set of behaviours, attitudes or characteristics which are seen as having something in common. Concepts may be associated with variables and questions. A '**ConceptScheme**', also defined within the SKOS namespace, is a set of metadata describing statistical concepts. '**Universe**' is the total membership or population of a defined class of people, objects or events. There are two types of population, target population and survey population. A target population is the population outlined in the survey objects about which information is to be sought. A survey population (also known as the coverage of the survey) is the population from which information can be obtained in the survey[13]. '**Coverage**' comprehends the key features of the scope of the data (e.g. geographic product occupation). The 'Coverage' has the datatype property 'dcterms:subject' and the object properties 'dcterms:temporal' and 'dcterms:spatial' pointing to 'dcterms:PeriodOfTime' and 'dcterms:Location'. As in the DC reference namespace'dcterms:spatial' is defined as a sub-property of 'dcterms:coverage', it could be derived that resources of type 'Coverage' have also spatial or temporal coverages to individuals of the class 'dcterms:Location'. Within the DDI context, these range resources are only understood as locations.

The data for the study are collected by an instrument. The purpose of an '**Instrument**', i.e. 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 (DDI Alliance, Technical Specification, Part II, 2009). A questionnaire contains a flow of questions. A '**Question**' is designed to get information upon a subject, or sequence of subjects, from a respondent. '**Variable**' is a characteristic of a unit being observed. A variable might be the answer of a question, have an administrative source, or be derived from other variables. Two DC datatype properties 'dcterms:identifier' and 'dcterms:description' are used to describe resources of type 'Variable'. The '**Representation**' of a variable is the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set (ISO/IEC, 2004). 'Representation' is one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, and sex: male coded as 1). '**DataElement**' encompasses study-independent, re-usable parts of variables like occupation classification. Data elements can be further described using the datatype property 'dcterms:description'.

Each study has a set of logical metadata ('**LogicalDataSet**') associated with the processing of data, at the time of collection or later during cleaning, and re-coding. This includes the definition of variables (paired code and category schemes). 'dcterms:title' specifies the title of a logical dataset. The collected data result in the microdata represented by the '**DataFile**'. Four DC

https://doi.org/10.23106/dcmi.952135935

datatype properties share the same domain 'dcterms:identifier', 'dcterms:description', 'dcterms:format', and 'dcterms:provenance'. An overview over the microdata can be given either by the descriptive statistics or the aggregated data. '**DescriptiveStatistics**' may be minimal, maximal, mean values, and absolute and relative frequencies. '**DataSet**' originates from the RDF Data Cube Vocabulary, an approach to map the SDMX information model to an ontology. A dataset represents aggregated data (also known as macrodata) such as multi-dimensional tables (Cyganiak et al., 2010). Aggregated data are derived from microdata by statistics on groups, or aggregates such as counts, means, or frequencies.

Implementation. We defined a direct[21] as well as 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. On the instance level, XML document instances can be translated automatically into the RDF representation of the generated ontologies by means of Java code. Individuals of domain ontologies can relate to resources of generated ontologies using equivalence relationships. In comparison with tools converting XML Schemas to OWL ontologies, the novelty of the evolved method is that the translation of XML Schemas into generated ontologies is based on the meta-model of XML Schema.

5. Use cases

Diverse problems can be solved with an RDF representation of the DDI data model. We describe three exemplary and representative use cases in detail to show the benefits associated with the developed ontology. First, we will depict the task of social science researchers to discover data and metadata which are connected with more than one study. The second use case deals with the linkage to publications related to specified data. It is necessary to link to external thesaurus concepts if a user searches for a specific study and does not know which terms have to be stated. This is treated in the third use case.

Discovery. The first use case deals with the discovery of both data and metadata associated with multiple studies. Researchers often want to know which studies exist for a specific country (e.g. France), time (e.g. 2005), and subject (e.g. election). The coverage in this example is separated by the three dimensions: country, time, and subject and the studies are connected to this coverage via the object property 'hasCoverage'. In order to get the titles of each of those studies, a SPARQL query like the following can be executed:

```
SELECT ?studyTitle
WHERE{
 ?study dcterms:title ?studyTitle.
 ?study ddi:hasCoverage ?coverage.
 ?coverage dcterms:subject 'election'.
 ?coverage dcterms:temporal ?periodOfTime.
 ?coverage dcterms:spatial ?location.}
```

As a result, we have now the titles of all the studies linked to the coverage consisting of the dimensions country, time, and subject. The next step could be to request exactly these studies returned form the first query in which a particular concept (e.g. education) exists. In this case, variables associated with the three-dimensional coverage and the returned studies are linked to the DDI element 'Concept' via the object property 'hasConcept'. The concept label is described

[21] XSLTs available: http://ddixslt.googlecode.com/svn/trunk/ddi-rdf/

https://doi.org/10.23106/dcmi.952135935

by the datatype property 'prefLabel' borrowed from SKOS. Another frequent and study comprehensive information retrieval task would be to ask for all the questions (e.g. 'What is your highest school degree?') which are linked to specific concepts (e.g. operationalizing education) and which are raised within the context of the example coverage. The question text and the concept label are described by means of datatype properties pointing to the primitive datatype string. Questions and concepts are directly related to each other and concepts are connected with the coverage indirectly via variables and the study.

Almost the same SPARQL query should be performed in order to get each of the variables (e.g. highestSchoolDegree) which are linked to particular concepts (e.g. measuring education) and which are linked with a specific coverage. At this time, we received the question with the question text 'What is your highest school degree?' connected with the concept 'education' and the coverage with the three dimensions country, time, and subject. The next query could be: How is this question represented both as wording (e.g. 'high school') and as code (e.g. 4)? Variables are interconnected with their representations. These representations are of the two types 'Representation' and 'skos:ConceptScheme', as concept schemes may include multiple skos:Concepts. The wording (the category) and the code are both represented as instances of the class 'skos:Concept'. This class has the two datatype properties 'skos:notation' pointing to the code and 'skos:prefLabel' pointing to the wording representation.

One could also be interested in descriptive statistics like minimal, mean, or maximal values, standard deviations, and absolute or relative frequencies to get a first impression of the microdata of datasets. Variables are directly connected with descriptive statistics. These descriptive statistics are of the type 'Descriptive Statistics' and may have datatype properties like 'percentage' to state relative frequencies. To get an overview over the overall microdata (makes especially sense in the case the accessibility of the microdata is limited), the aggregated data (e.g. a two-dimensional table with the dimensions 'age' and 'highest school degree') for a specific study, variable, coverage, and/or concept could be requested. A study and the aggregated data, instantiated using the class 'DataSet' which is specified within the RDF Data Cube Vocabulary namespace, are joined by the logical dataset. In a similar way, microdata for a specific study, variable, coverage, concept may be queried for own analyses. The study is interconnected with an instance of the DataFile class across the logical dataset.

Finding and Linking 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. But, 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, 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 using a 'Relationship'/ 'RelatedToReference' element in DDI Lifecycle. Transferring the connection between publications and data to DDI-RDF, possible link predicates can be 'ddilink:backgroundPublication' for a theoretical background of the study, 'ddilink:methodologyPublication' for a methodical background of the study and 'ddilink:resultsPublication' for the representation of main results, e.g. a publication based on study. Kauppinen et al. (2012) also talk about linking publications and data together.

Links to External Thesauri. In DDI, concepts can be connected with e.g. questions, variables, data elements or descriptive statistics in order to provide information about their topic. Such concepts are typically organized in DDI in concept schemes, which are often similar to traditional

https://doi.org/10.23106/dcmi.952135935

thesauri or classification systems regarding their structure and content. When assigning concepts to questions, etc. either an existing concept scheme has to be used or a new one has to be defined. A precise annotation of such entities is relevant for users when searching for studies, which e.g. cover specific concepts or contain questions regarding a very specific theme. But in a lot of cases the user does not know, which terms or classification systems have been used to provide these concepts. In such a case mappings from concepts to terms of other established thesauri or dictionaries like EUROVOC[22], Wordnet[23] or LCSH[24] and more specific thesauri such as the STW Thesaurus for Economics[25] or the Thesaurus for the Social Sciences TheSoz[26] can be helpful in order to recommend users suitable terms for search, which are used in the DDI study as concepts. Such mappings between thesauri are a typical instrument for information retrieval.

Therefore, it is quite reasonable to connect concepts of DDI to terms of existing knowledge systems for (a) using existing knowledge systems for the description of DDI entities with concepts and (b) providing information retrieval related services for users like search term recommendation during search. The inclusion of external thesauri, which often provide an established and mature term corpora in their specific discipline, does not only disseminate the use of such vocabularies, but also the potentially reuse of the DDI concepts in other Linked Data applications. DDI-RDF can technically be connected with Linked Data thesauri very easily. The latter ones are typically represented in SKOS format as well as concepts in DDI-RDF. Conceptually there are two possibilities to establish a connection between DDI-RDF and Linked Data thesauri. Concepts in DDI-RDF can be aligned to SKOS concepts of other thesauri. This can be achieved with the use of the SKOS mapping properties like 'skos:exactMatch', 'skos:relatedMatch', etc. The result is a network of related concepts over different thesauri and classification systems, which can be used for information retrieval methods. Another approach is the direct use of concepts of external thesauri instead of own concept schemes in DDI. Therefore all questions, variables, etc. in a study would reference directly via the DDI-RDF object properties to concepts from external data sources as their concepts.

6. Conclusions and future work

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 which are interlinked with more than one study.

Divers benefits are connected with the publication of DDI data and metadata in 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.

DDI-RDF for discovery purposes as well as the SKOS extension on concepts are planned as DDI Alliance specifications and therefore appropriate instances expressed by the DDI-RDF

[22] http://eurovoc.europa.eu/
[23] http://wordnet.princeton.edu/wordnet/
[24] http://www.loc.gov/aba/cataloging/subject/
[25] http://zbw.eu/stw/versions/latest/about.en.html
[26] http://lod.gesis.org/

https://doi.org/10.23106/dcmi.952135935

vocabulary can be published in the LOD cloud. This ongoing work is continued in core working groups. A review of the current work, an exploration of usage possibilities, and first evaluation attempts are planned at the second workshop on semantic statistics at Schloss Dagstuhl - Leibniz Center for Informatics in October 2012.

Acknowledgements

The work described in this paper has been started 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 and has been continued at the follow-up workshop in the course of the 3rd Annual European DDI Users Group Meeting (EDDI11) in Gothenburg, Sweden[27]. This work has been supported by contributions of the participants of both events, Archana Bidargaddi (NSD - Norwegian Social Science Data Services), Thoms Bosch (GESIS – Leibniz Institute for the Social Sciences, Germany and LOD community), Franck Cotton (INSEE - Institut National de la Statistique et des Études Économiques, France and LOD), Richard Cyganiak (DERI, Digital Enterprise Research Institute, Ireland and LOD), Daniel Gilman (BLS - Bureau of Labor Statistics, USA), Arofan Gregory (ODaF - Open Data Foundation, USA and DDI Alliance Technical Implementation Committee (TIC)), Marcel Hebing (SOEP - German Socio-Economic Panel Study), Larry Hoyle (University of Kansas, USA), Jannik Jensen (DDA - Danish Data Archive), 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), 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 TIC), Johanna Vompras (University Bielefeld Library, Germany), Joachim Wackerow (GESIS – Leibniz Institute for the Social Sciences, Germany and DDI Alliance TIC), and Benjamin Zapilko (GESIS – Leibniz Institute for the Social Sciences, Germany and LOD).

References

Bosch, Thomas, and Brigitte Mathiak. (2011). Generic Multilevel Approach Designing Domain Ontologies based on XML Schemas. Workshop Ontologies Come of Age in the Semantic Web, 2011, 1-12.

Bosch, Thomas, and Brigitte Mathiak. (2012). XSLT Transformation Generating OWL Ontologies Automatically Based on XML Schemas. The 6th International Conference for Internet Technology and Secured Transactions, 2012, 660-667.

Cyganiak, Richard, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. (2010). Semantic Statistics: Bringing Together SDMX and SCOVO. Proceedings of the Linked Data on the Web Workshop, 2010.

DCMI. (2008). Expressing Dublin Core Metadata using the Resource Description Framework (RDF). Retrieved March 22, 2012, from http://dublincore.org/documents/dc-rdf/.

DCMI. (2010). DCMI Metadata Terms. Retrieved March 22, 2012, from http://dublincore.org/documents/dcmi-terms/.

DDI Alliance. (2009). Data Documentation Initiative (DDI) Technical Specification, Part I, Overview, Version 3.1. Retrieved March 22, 2012, from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/.

DDI Alliance. (2009). Data Documentation Initiative (DDI) Technical Specification, Part II, User Guide, Version 3.1. Retrieved March 22, 2012, from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/.

ISO. (2000). ISO 1087- 1:2000, Theory and Application. Retrieved March 22, 2012, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=20057

ISO. (2003). ISO 19115-1:2003, Geographic Information – Metadata. Retrieved March 22, 2012, from http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020.

ISO/IEC. (2004). ISO/IEC 11179-1:2004, Information Technology – Metadata registries (MDR) - Framework. Retrieved March 22, 2012, from http://metadata-standards.org/11179/.

Jacobs, James. A., and Charles Humphrey. (2004). Preserving Research Data. Communications of the ACM 47, 9, 2004.

Kauppinen, Tomi, Baglatzi, Alkyoni, and Keßler, Carsten. (2012). Linked Science: Interconnecting Scientific Assets. Terence Critchlow and Kerstin Kleese-Van Dam (Eds.): Data Intensive Science. USA: CRC Press.

NISO Press. (2004). Understanding Metadata, NISO Press. Retrieved June 21, 2012, from http://www.niso.org/publications/press/ UnderstandingMetadata.pdf.

[27] http://www.iza.org/eddi11

https://doi.org/10.23106/dcmi.952135935