DDI-RDF – A Discovery Model for Microdata

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

[2000-7000 words]

[footnotes → endnotes]

Abstract

Ontology engineers and experts from the social, behavioral and economic sciences developed an data discovery ontology of both DDI Codebook and Lifecycle and implemented a rendering of DDI 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) or SKOS and also to define meaningful relationships to the RDF Data Cube vocabulary. Now, organizations have the possibility to publish their DDI data and meta-data in RDF and link it with many other datasets from the Linked Open Data (LOD) cloud. As a consequence, a huge amount of related DDI instances can be discovered, queried, connected, and harmonized. Only the semantic combination of DDI metadata (as well as data) of several organizations will enable derivations of even surprising and not imaginable implicit knowledge out of explicitly stated pieces of information.

- Ontology only a subset of the entire DDI data model

[150-250 words]

Keywords

Semantic Web, Linked Data, Ontology Design, DDI

[3-6 keywords]

Introduction [from DC paper] [Arofan & Achim]

- Why RDF for DDI?
- Motivation
- Technical advantages of RDF / Linked Data
- Goal
  - Business cases
  - Targetet users
  - Ties to datacube & SDMX

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 –

[1] http://www.ddialliance.org/

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. 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, DDI instances can be processed by RDF tools 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 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].

Data Documentation Initiative [from DC paper] [Arofan & Achim]

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 i.e. 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 comprehends 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

[2] http://linkeddata.org/
[3] http://lod-cloud.net/
[4] http://www.dagstuhl.de/11372
[5] http://www.iza.org/eddi11

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 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 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. 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 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.

[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/

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].

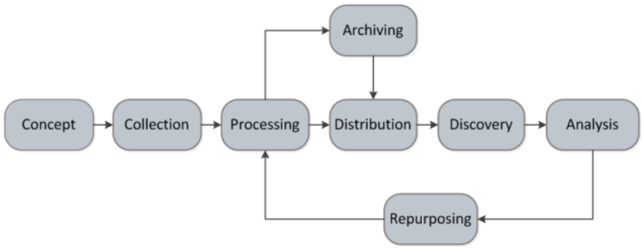

FIG. 1. DDI Data Lifecycle

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.

Related Work [from DC paper]

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 been covered in this extent by other established vocabularies yet. 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

[16] http://www.ddialliance.org/what
[17] http://www.loc.gov/standards/premis/

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 [Thomas]

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 which are seen as most important to solve problems associated with diverse identified use cases. Widely accepted and adopted vocabularies are reused to a large extend. 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[21], and delineating code lists, category schemes, mappings between them, and concepts like topics using SKOS. This section serves as an overview over the conceptual model. More detailed descriptions of all the properties are given in the specification[22] and a conference paper (Bosch et al. 2012).

4.1 Overview

Figure 1 gives an overview over the conceptual model containing a small subset of the DDI-XML specification[23]. 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 data set 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 on-going, data sets may be released as a **StudyGroup**, where

[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-data-cube/
[22] http://rdf-vocabulary.ddialliance.org/discovery
[23] http://www.ddialliance.org/Specification/

each cycle or "wave" of the data collection activity produces one or more data sets. 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.

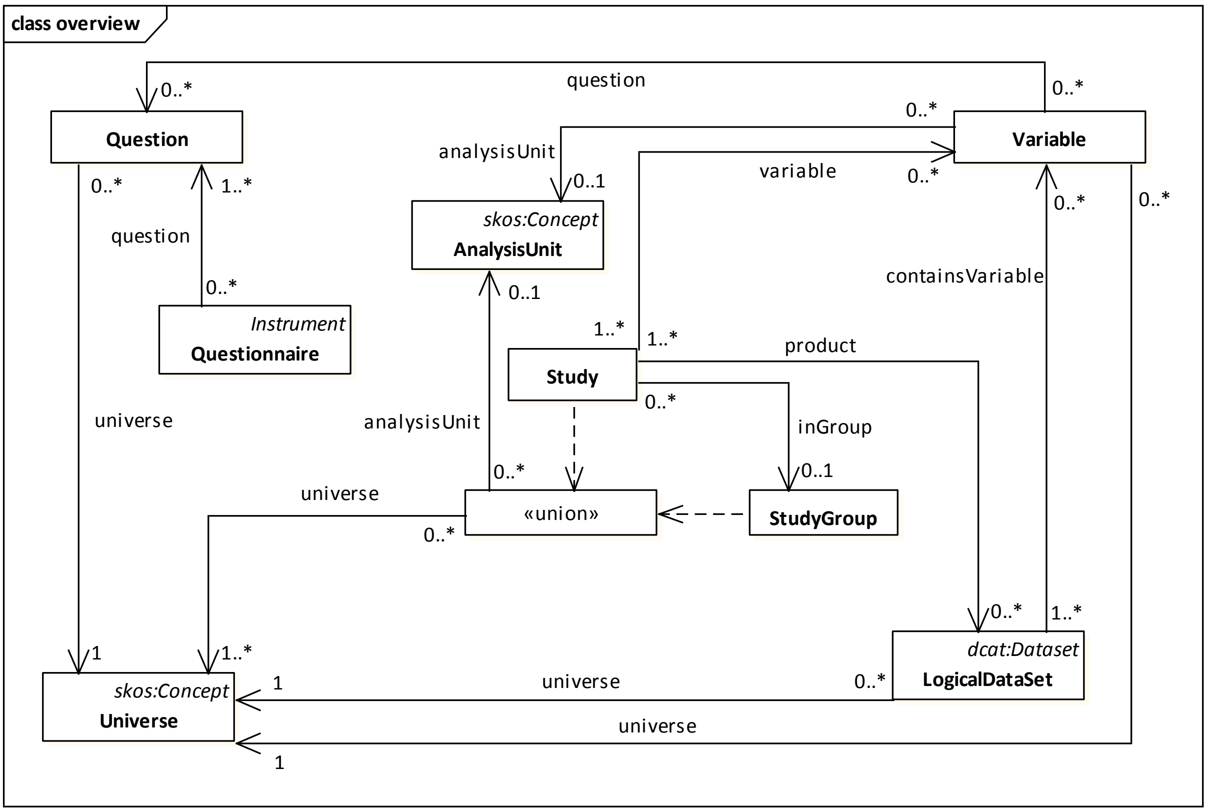

Figure 1. DDI-RDF Discovery Vocabulary

Data sets have two representations: a logical representation, which describes the contents of the data set, 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 the dact: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 data set, 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 is 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 life cycle 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 a spatial (dcterms:spatial), temporal (dcterms:temporal), and topical (dcterms:subject) coverage.

4.2 Studies and StudyGroups

A simple **Study** supports the stages of the full data lifecycle in a modular manner. A Study represents the process by which a data set 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 on-going, data sets may be released as a **StudyGroup**, where each cycle or "wave" of the data collection activity produces one or more data sets. This is typical for longitudinal and panel studies. In this case, a number of Study objects would 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 the time since when the Study is 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 defined as follows: The process collecting data is focusing on the analysis of a particular type of subject. If, for example, the adult population of Finland is being studied, the AnalysisUnit would be individuals or persons. Studies and groups of Studies must have 1 to n Universes which are sub-classes of skos:Concepts. For Universes you 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.

4.3 Logical Data Sets, Data Files, Descriptive Statistics, and Aggregated Data

Data sets have a logical representation, which describes the contents of the data set, and a physical representation, which is a distribuited 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 over 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[24], an approach to map the SDMX information model to an ontology. A DataSet represents aggregated data such as multi-dimensional tables. Aggregated data are derived from microdata by statistics on groups, or aggregates such as counts, means, or frequencies. **SummaryStatistics** pointing to variables and **CategoryStatistics** pointing to categories and codes are both descriptive statistics.

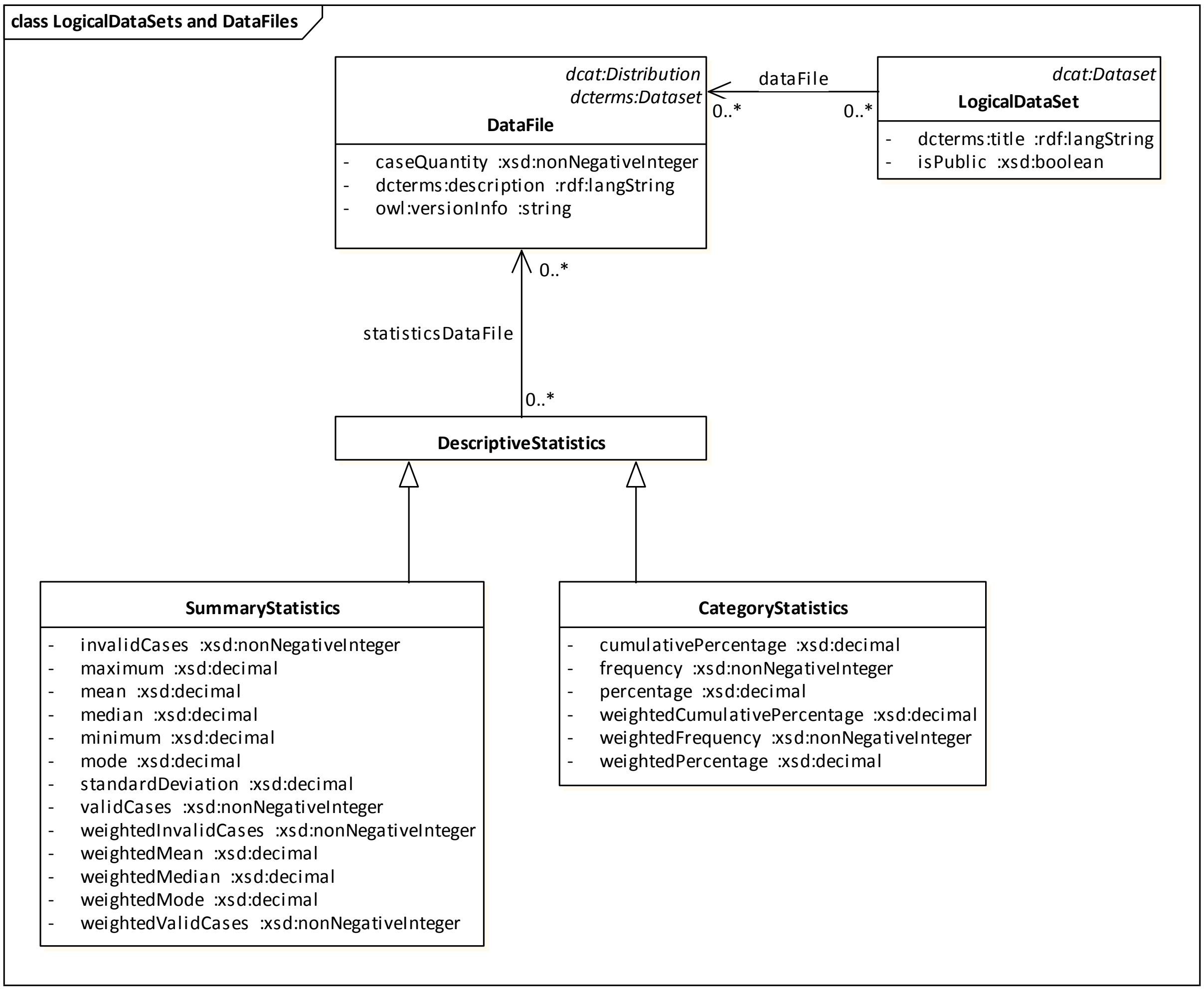

Figure 2. LogicalDataSets and DataFiles

4.4 Variables, Variable Definitions, Representations, and Concepts

When it comes to understanding the contents of the data set, 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 of a question, have an administrative source, or be derived from other Variables.

[24] http://www.w3.org/TR/vocab-data-cube/

VariableDefinitions encompasse study-independent, re-usable parts of Variables like occupation classification.

Questions, Variables, and VariableDefinitions may have Representations. Representation is defined as sub-class of the union of **rdfs:Datatype** (e.g. numeric or textual values) and **skos:ConceptScheme**, as for example questions may have as 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").

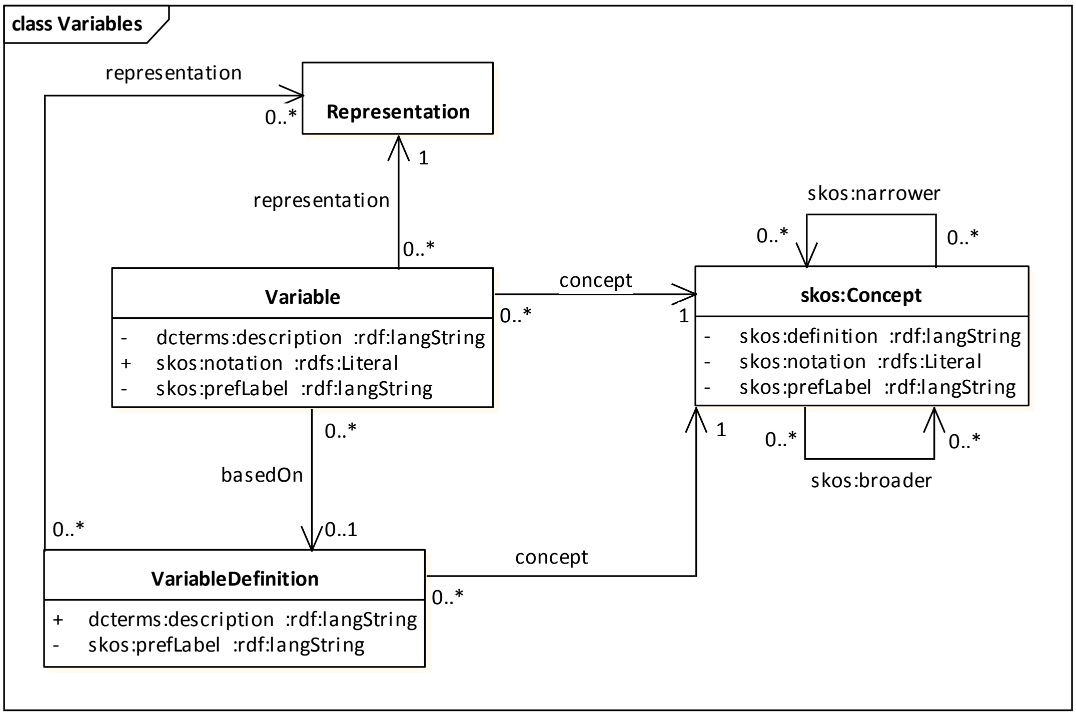

Figure 3. Variables

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 characteristics. 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 **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.

4.5 Data Collection

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. A **questionnaire** contains a flow of questions. A **Question** is designed to get information upon a subject, or sequence of subjects, from a respondent. The next figure visualizes the datatype and object properties of Instrument and Question.

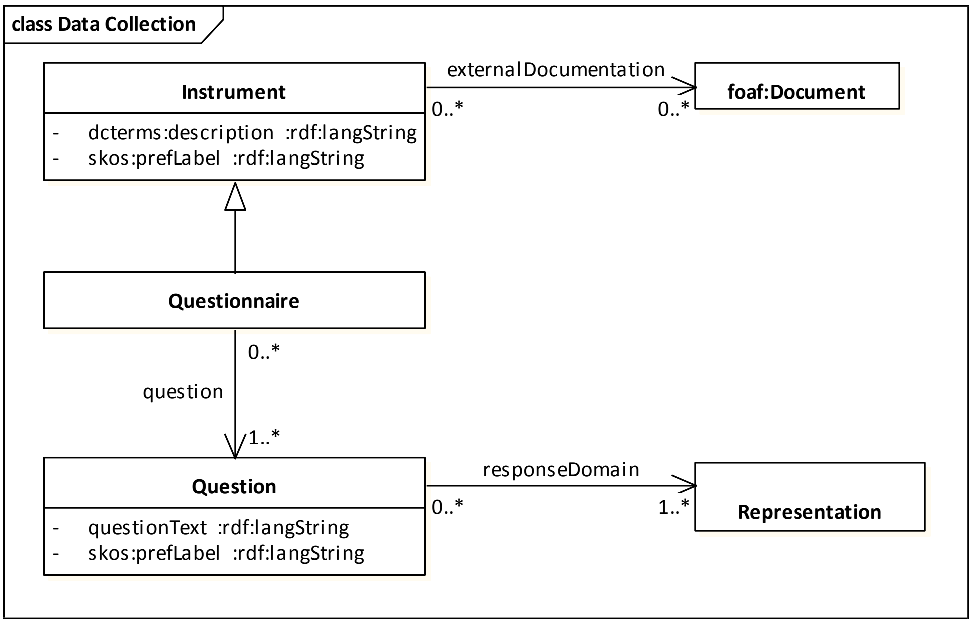

Figure 4. Data Collection

You can describe (dcterms:description) Instruments and associate labels (skos:prefLabel) to Instruments. Instruments may have multiple external documentations of the type foaf:Document. Questionnaires are special instruments having at least 1 collection mode (disco:collectionMode) which is a skos:Concept. Questionnaires must contain at least 1 Question. Questions have a question text (disco:questionText), a label (skos:prefLabel), exactly 1 universe (disco:universe), multiple concepts (disco:concept), and at least 1 response domain (disco:responseDomain).

Use Cases [Thomas]

This section describes the scenarios which 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 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 publically available, but often confidential in nature, requiring applications for access. While the data sets may be available (typically as CSV files) the metadata which accompanies them is not necessarily coherent, making the discovery of these data sets difficult. A possible user has to read related documents to determine if the data is 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 data sets. The DDI-RDF Discovery Vocabulary would support this approach.

Link publications to data sets. 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. 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.

Searching for studies by 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 which have a particular title or keywords assigned to in order to further explore the data sets attached to them. The definition of an analysis unit might help to directly determine which data sets 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 which disseminate the kind of data they want to use. This scenario shows how a researcher might wish to see the studies which are disseminated by a particular organization, so that the data sets which comprise them can be further explored and accessed. "Show me all the studies for the period 2000 to 2010 which are disseminated by the ESDS service of the UK Data Archive" is an example of a typical query.

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

There is a paper[25] describing further possible use cases in detail. Researchers can search for studies by producer, contributor, coverage, universe (i.e. study population), data source (e.g. study questionnaire). Social science researchers can search for data sets using variables, related questions, and classifications. Furthermore, you can search for reusable questions using related concepts, variables, universe, coverage, or by text.

[25] http://www.ddialliance.org/resources/publications

RDF from Codebook and Lifecycle [Olof & Ben]

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 by XSLT stylesheets[26]. 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-C compared to DDI-L is a lot different. DDI-C is designed to describe metadata for archival purposes, 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 be able to capture metadata from the early stage in the research process. A lot of the metadata can be described in modules and references are used between e.g. questions and variables. DDI-L enables capturing and reuse of metadata trough 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 have a lot of new elements and is able to describe studies, variables and questions in greater details. 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 the input of DDI-C or DDI-L the resulting RDF is the same. This enables an easy and equal search in RDF resulted from DDI-C and DDI-L. Also, the interoperability between both is increased.

Create 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 personalize and extend so the 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. As 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 on how elements in DDI-C and DDI-L map to the 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

[26] https://github.com/linked-statistics/DDI-RDF-tools

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, which automatically detects which DDI input is given. Also, we plan on including parameters into the conversion process in order 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. We aim to finalize the XSLT conversion process up to the IASSIST 2013.

Conclusions and Future Work [from DC paper]

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 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 [Achim]

The work described in this paper has been started at the first workshop on "Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web"[27] at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in September 2011. This work has been continued at these three meetings: follow-up working meeting in the course of the 3rd Annual European DDI Users Group Meeting (EDDI11)[28] in Gothenburg, Sweden in December 2011, second workshop on "Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI

[27] http://www.dagstuhl.de/11372
[28] http://www.iza.org/eddi11

Model for the Linked Data Web"[29] at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in October 2012, and 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 Gilman (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.

[29] http://www.dagstuhl.de/12422

[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