Use Cases Related to an Ontology of the Data Documentation Initiative

Thomas Bosch[1] and Brigitte Mathiak[2]

Abstract

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

Keywords

Semantic Web, Linked Data, Data Documentation Initiative, DDI, use cases

Introduction

Statistical domain experts worked closely with Linked Data community experts to define an ontology of the DDI data model. This work has 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 (Dagstuhl 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 (European DDI User Conference 2011). The last workshop "Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web" also took place at Schloss Dagstuhl, the Leibniz Center for Informatics in October 2012 (Dagstuhl 2012).

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

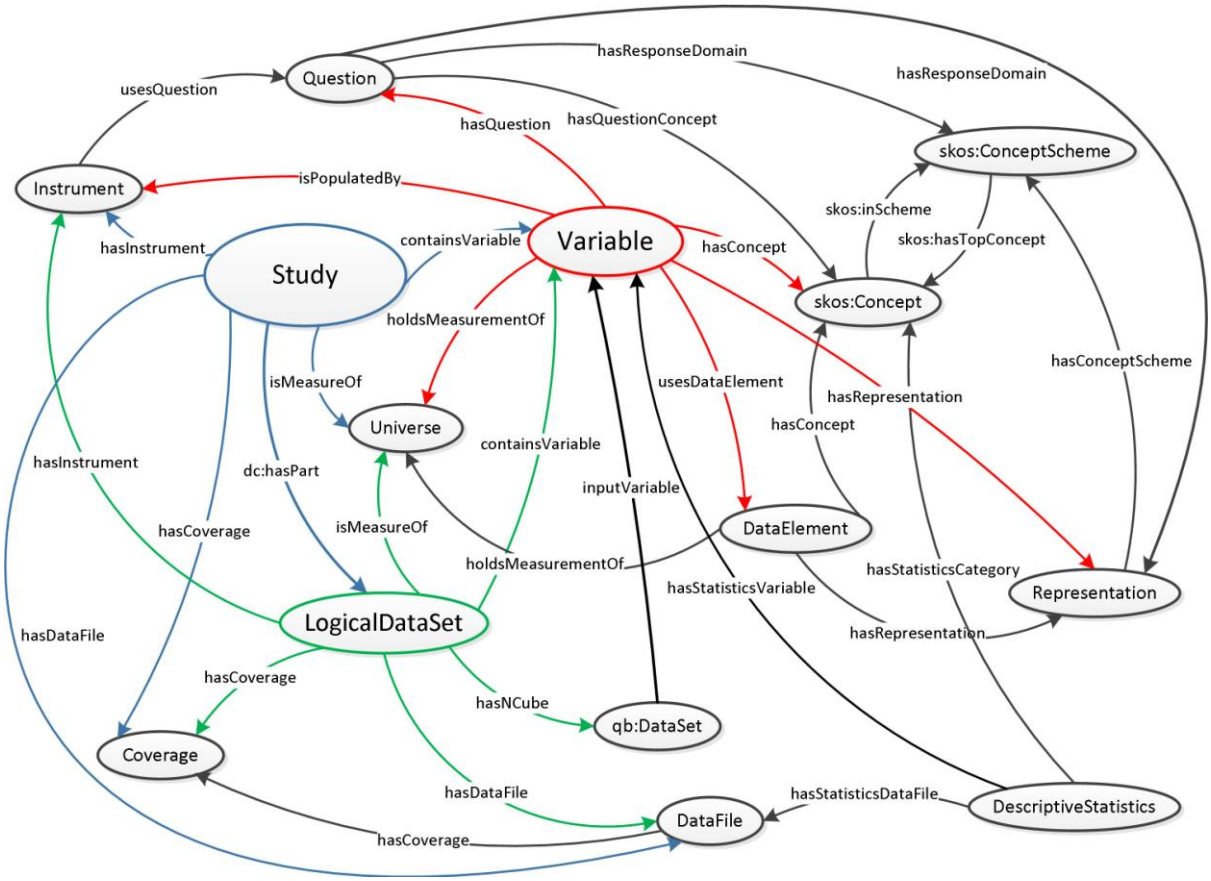

Figure 1. Conceptual Model of the DDI Ontology (Bosch et al. 2012)

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

- the Dublin Core ontology (DCMI Metadata Terms 2012) delineating metadata for citation purposes,
- the Simple Knowledge Organization System (SKOS) (W3C 2009) describing code lists, category schemes, mappings between them, and concepts like topics, and
- the RDF Data Cube Vocabulary which describes aggregated data like multi-dimensional tables (Bosch et al. 2012).

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

In the following chapter, the authors of this paper describe in detail use cases and benefits associated with an ontology of the DDI data model. We want to answer the question why it is

crucially important that an RDF representation of the Data Documentation Initiative has been defined.

Publish and Link DDI Data and Metadata

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

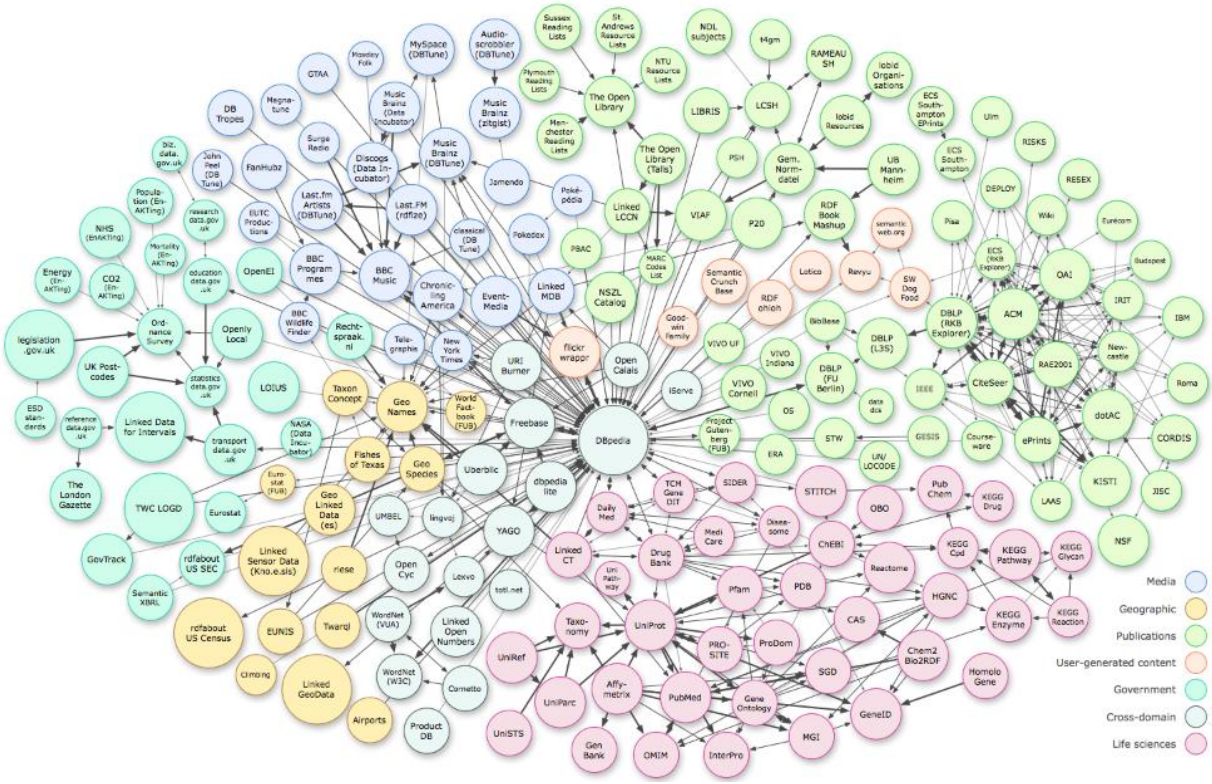

Figure 2. The Linking Open Data Cloud Diagram (Cyganiak 2011)

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

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

specific organization. And secondly, every node in this LOD network can also be processed by individual organizations. Summarized, you can reach a broader audience and a broader audience can reach you.

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

Discovery

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

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

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

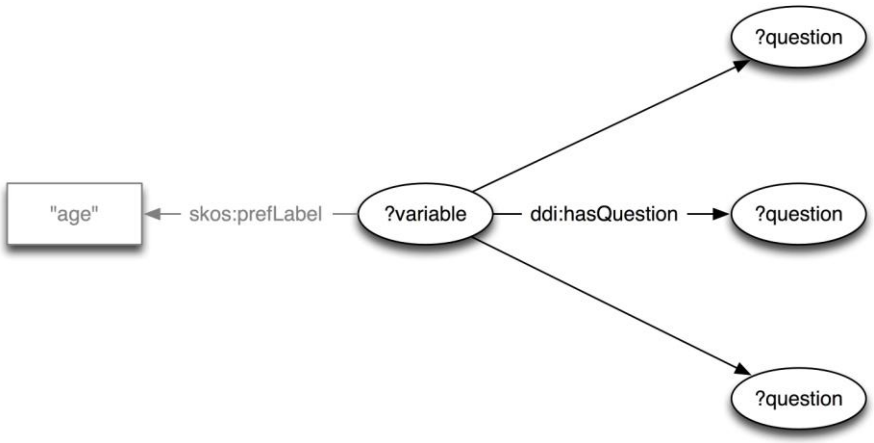

Figure 3. Requesting all questions of a given variable

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

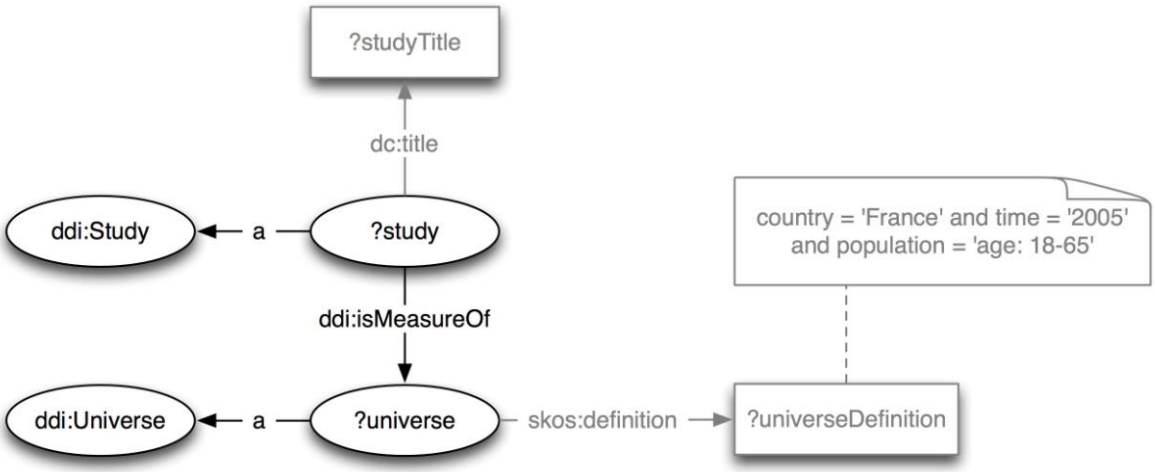

Figure 4. Discovery – Study, Universe

```
SELECT ?studyTitle
WHERE

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

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

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

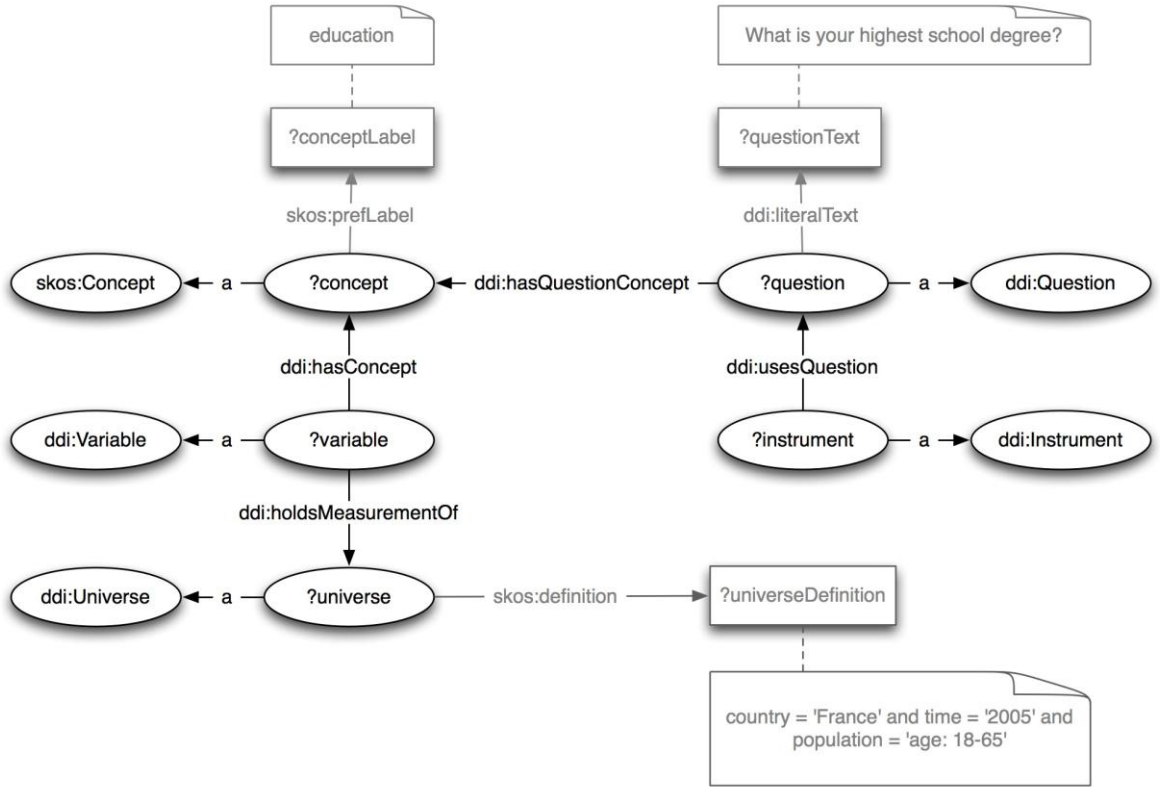

Figure 5. Discovery – Universe, Variable, Concept, Question, Instrument

```
SELECT ?question
WHERE

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

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

So Far, the researcher gets the questions joined with the question text ‘What is your highest school degree?’, the concept ’education’, and the universe with the three dimensions country, time, and population. The same researchers are now interested in the representation as wording and as code of the returned questions. Variables are interconnected with their representations which are typed as ‘Representation’ as well as ‘skos:ConceptScheme’, since the wording (the category) and the code

are both represented as instances of the class ‘skos:Concept’. Two datatype properties are defined for this class: ‘skos:notation’ and ‘skos_prefLabel’. The datatype properties ‘skos:notation’ points to the code and ‘skos_prefLabel’ to the wording representation. Figure 6 shows the class axioms needed to formulate the SPARQL query below in order to implement the stated discovery sub use case. The where clause of the previous SPARQL query has to be included.

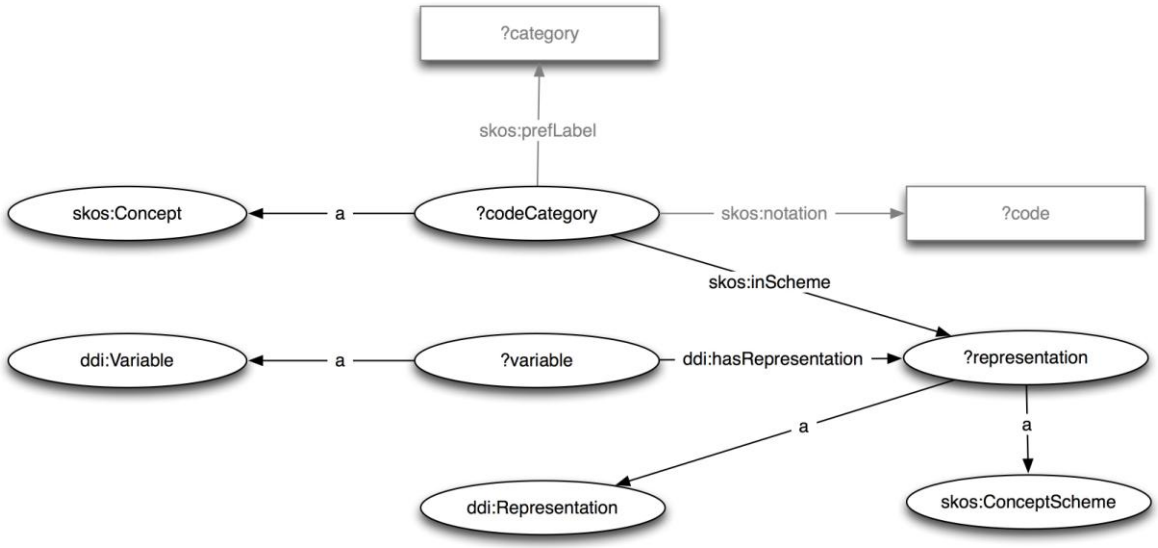

Figure 6. Discovery – Representation

```
SELECT ?code ?category
WHERE

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

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

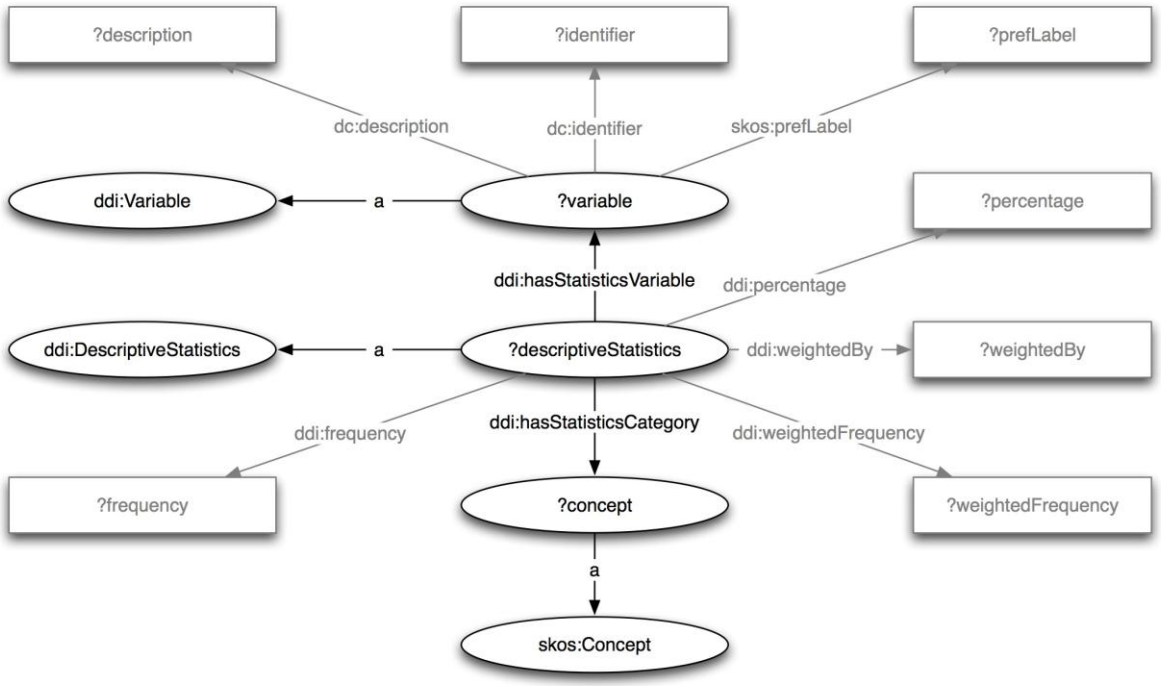

Figure 7. Discovery – Descriptive Statistics

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

If access to microdata is limited or to get an overview over the entire microdata, researchers could request the aggregated data (e.g. a two-dimensional table with the dimensions ‘age’ and ‘highest school degree’) for particular studies, variables, universes and concepts. Logical datasets build the link between studies and associated aggregated data which is represented by the RDF Data Cube Vocabulary’s class ‘DataSet’. In a similar way, microdata for a specific study, variable, universe, concept may be queried for own analyses. The study is interconnected with an instance of the ’DataFile’ class across the logical dataset. The classes ’LogicalDataSet’, ’DataSet’, and ’DataFile’, as well as their datatype an object properties are visualized in the next figure.

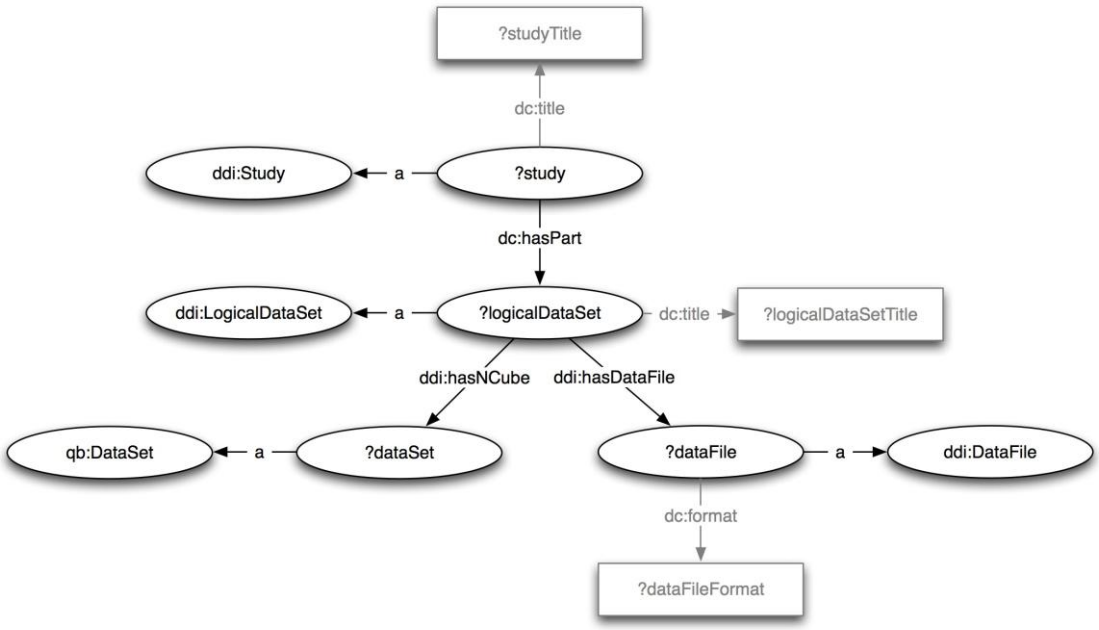

Figure 8. Discovery – LogicalDataSet, DataSet, DataFile

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

Integration of Other Ontologies

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

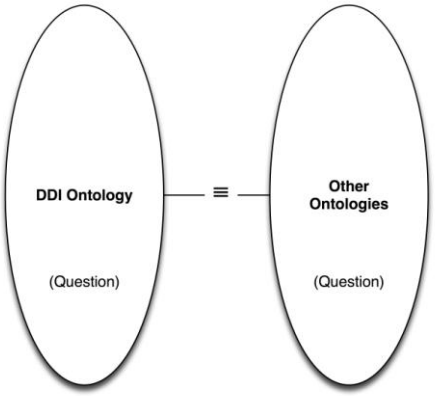

Figure 9. Integration of other ontologies

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

An ontology of the SPSS data model and respective tools transferring metadata between DDI and SPSS may be built. Now, class axioms (i.e. classes, datatype and object properties) of the SPSS ontology and the DDI ontology can be stated as equivalent. As a consequence, there is no need to write transformation code translating RDF instances of the SPSS ontologies to RDF representations of the DDI ontology, as an SPSS class ”Variable” could be defined as equivalent to a DDI class ”Variable”. So far, there is no ontology of the SPSS data model available, but as this statics program is commonly used, this task should be executed in the future.

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

Expressiveness of Ontologies

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

schema languages like XML Schema concentrate on structuring documents instead of structuring data.

Consistency Check of the DDI Data Model

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

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

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

Facilitation of DDI Elements' Comparability

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

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

these two variables ‘Grade_USA’ and ‘Grade_Germany’ are equivalent, too.

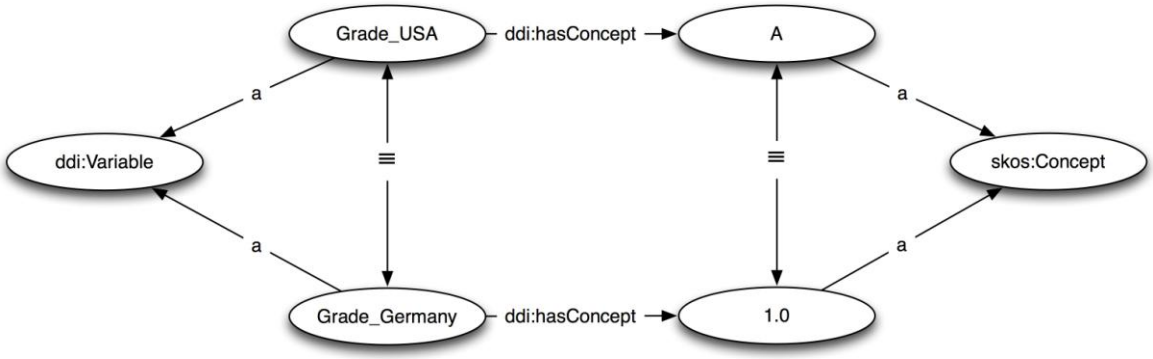

Figure 10. Equivalence of variables

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

Finding and Linking External Resources like Publications Related to Data

Publications, which describe ongoing research or its output based on research data, are typically held in bibliographical databases or information systems. Adding unique, persistent identifiers established in scholarly publishing to DDI-based metadata for datasets, these datasets become citable in research publications and thereby linkable and discoverable for users. 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 in DDI Lifecycle. In version 3.1 of the DDI metadata standard, the element ‘OtherMaterial’ is used to reference resources such as publications that are related to the content of the relevant module. This element includes a description, a bibliographic

citation (containing 15 Dublin Core elements like identifier, title, creator, or date), an external reference using a URL or a URN, and a reference to the item within the module to which the external resource is related (DDI Alliance 2009). Thus, all the necessary information characterizing the referenced resources can be stated. Figure 11 depicts the XML tree of the 'OtherMaterial' element.

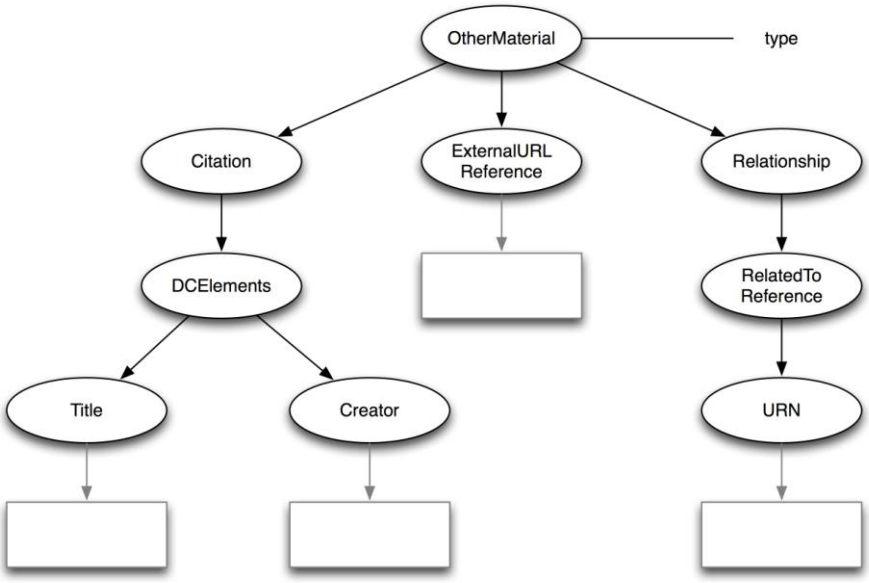

Figure 11. References to external resources in DDI 3.1

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

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

study, ‘methodologyPublication’ for a methodical background of the study and ‘resultsPublication’ for the representation of main results, e.g. a publication based on study.

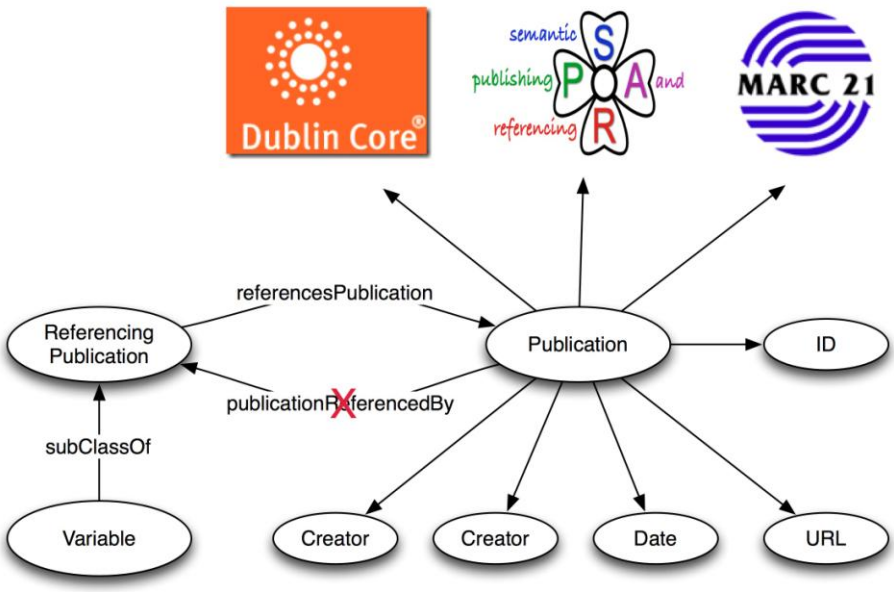

Figure 12. Semantic references to external resources using Semantic Web technologies

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

As you can see in figure 12, the recommended data model is very simple, intuitive, and generic, that implies that this data model can be applied in multiple contexts. To ensure reusability, variables only reference publications and not the other way round. Using this data model, both the reference and the referenced resources are defined semantically. This model can be expanded if the classes ‘ReferencingResource’ and ‘Resources’ are specified as super-classes of ‘ReferencingPublication’ and ‘Publication’. In this manner, each possible type of resource can be stated as referenceable by each possible kind of things which can have a reference to specific resources. A further example, similar to semantic references to external resources such as publications, would be to define references to notes in a semantic way.

Concept Relationships

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

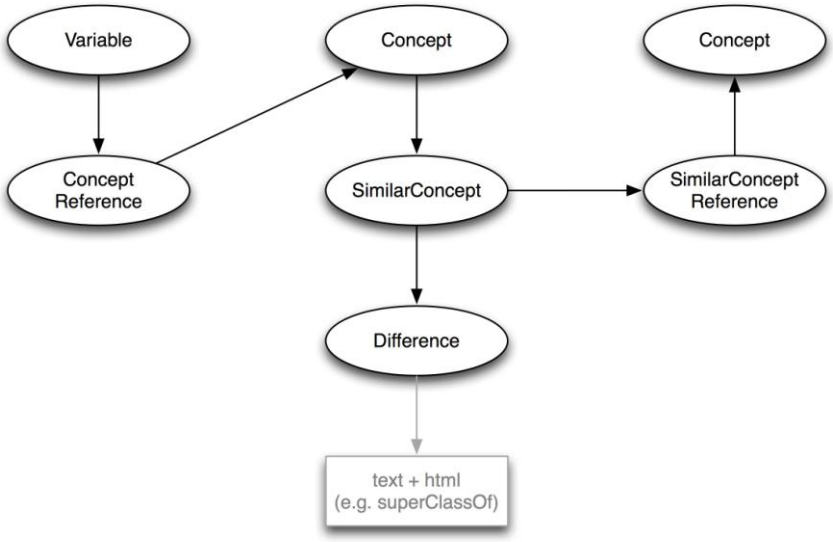

Figure 13. Concept relationships in DDI 3.1

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

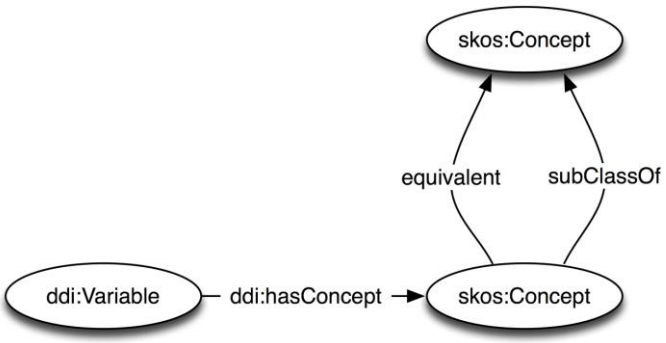

Figure 14. Semantic concept relationships

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

Links to External Thesauri

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

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

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

- DDI concepts can be aligned to SKOS concepts of other external thesauri using SKOS properties like skos:exactMatch, skos:relatedMatch. This mapping serves a network of related concepts over different thesauri and classification systems, which can be used to identify equivalent or related concepts.
- Another approach can be the absence of own concept schemes in DDI and the use of existing, external thesauri for all suitable concepts. Therefore all questions, variables, data elements, or descriptive statistics in a study would reference directly via the DDI-RDF object properties to concepts from external data sources as their concepts.

Conclusions

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

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

References

Bizer, C, Jentzsch, A, Cyganiak, R 2012, *State of the LOD Cloud*. Available from: <http://www4.wiwiss.fu-berlin.de/lodcloud/state/>. [1 Mai 2012].

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

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

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

Cyganiak, R 2011, *The Linking Open Data cloud diagram.* Available from: http://richard.cyganiak.de/2007/10/lod/. [1 Mai 2012].

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

Dagstuhl 2011, *Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Web*. Available from: http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=11372. [15 December 2012].

Dagstuhl 2012, *Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web*. Available from: http://www.dagstuhl.de/de/programm/kalender/evhp/?semnr=12422. [15 December 2012].

DCMI Metadata Terms 2012, *DCMI Metadata Terms*. Available from: http://dublincore.org/documents/dcmi-terms/. [15 December 2012].

DDI Alliance 2009, *DDI 3.1 XML Schema Documentation*. Available from: http://www.ddialliance.org/sites/default/files/documentation/ddi3.1/index.html. [18 October 2009].

Dublin Core Matadata Initiative 2008, *Expressing Dublin Core metadata using the Resource Description Framework (RDF)*. Available from: http://dublincore.org/documents/dc-rdf/. [14 January 2009].

European DDI User Conference 2011, *EuropeanDDI User Conference*. Available from: http://www.iza.org/conference_files/EDDI2011/call_for_papers. [15 December 2012].

EuroVoc, 2012. Available from: <http://eurovoc.europa.eu/>. [2 Mai 2012].

Falcons, 2011. Available from: <http://ws.nju.edu.cn/falcons/objectsearch/index.jsp>. [1 Mai 2012].

Gregory, A 2011, *Open data and metadata standards: should we be satisfied with "good enough"?*, Technical report, Open Data Foundation.

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

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

ISO/IEC JTC1 SC32 WG2 2012, ISO/IEC 11179, Information Technology -- Metadata registries (MDR). Available from: <http://metadata-stds.org/11179/>. [1 Mai 2012].

Library of Congress 2012, *MARC STANDARDS*. Available from: http://www.loc.gov/marc/. [10 April 2012].

Library of Congress 2012, *Library of Congress Subject Headings.* Available from: <http://www.loc.gov/aba/cataloging/subject/>. [2 Mai 2012].

Linked Data, 2012. Available from: <http://linkeddata.org/>. [1 Mai 2012].

LinkSailor, 2012. Available from: <http://linksailor.com/nav>. [1 Mai 2012].

Kupfer, A, Eckstein, S, Störmann B, Neumann K, Mathiak B 2007, 'Methods for a synchronised evolution of databases and associated ontologies', in Proceeding of the 2007 conference on databases and information systems IV.

Marbles, 2012. Available from: <http://www5.wiwiss.fu-berlin.de/marbles/>. [1 Mai 2012].

Ontoware.org 2012, *SWRC Ontology.* Available from: http://ontoware.org/swrc/. [15 December 2012].

Peroni, S & Shotton, D 2011, *Semantic Publishing and Referencing Ontologies (SPAR)* Available from: http://sempublishing.svn.sourceforge.net/viewvc/sempublishing/SPAR/index.html. [1 Mai 2012].

Princeton University 2012, *WordNet – A lexical database for English.* Available from: <http://wordnet.princeton.edu/wordnet/>. [2 Mai 2012].

The Tabulator, 2005. Available from: <http://www.w3.org/2005/ajar/tab>. [1 Mai 2012].

Semantic Web Search Engine 2012. Available from: <http://www.swse.org/index.php> [1 Mai 2012].

SIG.MA Semantic Information Mashup 2012. Available from: <http://sig.ma/> [1 Mai 2012].

Statistical Data and Metadata eXchange 2012. Available from: <http://sdmx.org/> [1 Mai 2012].

Stuckenschmidt, H 2009, *Ontologien: Konzepte, Technologien und Anwendungen*, Springer-Verlag, Berlin Heidelberg.

W3C 2004, *OWL Web Ontology Language Overview.* Available from: http://www.w3.org/TR/2004/REC-owl-features-20040210/. [10 February 2004].

W3C 2008, *SPARQL Query Language for RDF.* Available from: http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/. [15 January 2008].

W3C 2009, *SKOS Simple Knowledge Organization System Namespace Document - HTML Variant - 18 August 2009 Recommendation Edition.* Available from: http://www.w3.org/2009/08/skos-reference/skos.html. [15 December 2012].

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

[2] Brigitte Mathiak | GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany |E-Mail: brigitte.mathiak@gesis.org