Towards the Discovery of Person-Level Data

Reuse of Vocabularies and Related Use Cases

Thomas Bosch[1], Benjamin Zapilko[1], Joachim Wackerow[1], Arofan Gregory[2]

[1]GESIS – Leibniz Institute for the Social Sciences, Germany
{first name.last name}@gesis.org
[2]Open Data Foundation, USA
agregory@opendatafoundation.org

Abstract. The Linked Data and the Social Science data communities developed the DDI-RDF Discovery Vocabulary, an ontology of the Data Documentation Initiative, in order to support the discovery of person-level data and its metadata. The Data Documentation Initiative (DDI) is an acknowledged international standard for the documentation and management of data from the social, behavioral, and economic sciences. Within the context of DDI-RDF Discovery Vocabulary, we reuse well elaborated and accepted vocabularies to a large extend. Vocabularies like DCMI, FOAF, ORG, ADMS, PROV-O, SKOS and XKOS, DCAT, and Data Cube. This paper focuses on the description of how other vocabularies are reused reasonably and on the description of use cases which are associated with the usage of the DDI-RDF Discovery Vocabulary.

Keywords: Linked Data, Semantic Web, Ontology Design, Statistical Data, DDI-RDF Discovery Vocabulary, Data Documentation Initiative

1 Introduction

For more than a decade, members of the community around research data for the social, behavioural, and economic (SBE) sciences have been developing and using a metadata standard known as the Data Documentation Initiative (DDI) standard [1], an XML format designed for the purposes of supporting the dissemination, management, and re-use of the data collected and archived for research purposes. Recently, this standard has become the basis for the DDI-RDF Discovery Vocabulary, an effort to leverage the mature metadata model found in DDI XML formats for the purposes of exposing these same data holdings as resources within the Web of Linked Data. In designing this vocabulary, every attempt has been made to meet the requirements of the different technologies and needs found in the Linked Data world [2], as opposed to the world of data archives, research institutes, and data libraries [3]. Part of this best practice is the reuse of existing vocabularies wherever possible, and the extension of existing vocabularies where needs are almost, but not completely covered.

It is important to understand what type of data we are describing here, as within the SBE community „data“ have a very specific meaning: the data most often used in

adfa, p. 1, 2011.

research is data collected about individuals (and sometimes also businesses and households) in the form of responses to surveys or taken from administrative registers (such as hospital records, registers of births and deaths, etc.). Common terms for this kind of data are microdata, record-unit data or more specific person-level data. Synonyms for this kind of data are microdata, record-unit data, or more specific person-level data. By its nature, this data are highly confidential, and access is often only permitted for qualified researchers who must apply for access. The range of data is very broad, including census data, all types of social surveys, education data, health, labor force surveys and business surveys. Increasingly, this type of research data is held within data archives or data libraries after it has been collected, so that it may be re-used by future researchers. In performing their research, the detailed person-level data are aggregated into „tables“ or „data cubes“, a process which involves transformation of the individual data into something less confidential, but which answers a particular research question.

The archives and data libraries have no control over the form of the data deposited with them by researchers, and the DDI standard reflects this – it is a standard XML format for the large amount of metadata needed to understand the wide range of data formats used by researchers at a very detailed level. Where a metadata standard such as Dublin Core has dozens of metadata fields, the DDI standard has almost twelve hundred. The metadata is sufficient to support a wide range of uses, including management of data holdings within archives, discovery and dissemination, transformation of the data between different proprietary software formats, and a thorough documentation of the data and how and why it was collected. The key to the re-use and management of data is always metadata, and this has been a major theme within the SBE community for many years.

It should be noted that the typical use of DDI is within controlled environments: because the data are itself so often highly confidential, the metadata are often maintained and used within closed systems, except in those cases where it is exposed for discovery purposes, typically on websites. DDI has been used heavily: three excellent examples are its use within the CESSDA community of European national data archives; its use by the International Household Survey Network (IHSN) community, made up of more than 90 statistical agencies in the developing world; and its use by the largest SBE data archive in the US, ICPSR.; but there are many other examples.

When we consider how such a standard could be used as the basis for an RDF vocabulary, we realize that the requirements are very different. The most obvious use case is that of discovery, given that much of the data is highly confidential, and that access to the data must be applied for in most cases. Further, the challenges of searching the Web of Linked Data are enormous – the sheer range of information is incredibly broad. Thus, the almost twelve hundred of metadata fields within DDI is itself a problem. The DDI model must be significantly reduced in complexity to be meaningful to cover these requirements. The fact that DDI is not specific to any particular research domain or type of research data is a positive feature, however, as the range of data to be exposed into the Web of Linked Data is also very broad.

The DDI-RDF Discovery Vocabulary (*disco* is the namespace abbreviation) has emerged as a massive simplification of the DDI XML standard, optimized for query-

ing using technologies such as SPARQL. Because the use cases of data management and other applications are not supported, many of the fields found within the base model have been ignored. For some functions – such as the description of tabulated data „cubes“ the Data Cube Vocabulary has been directly used. Further, there is a heavy use of SKOS, or the extended version of SKOS – XKOS – which is also being developed by the DDI Alliance – to add the additional information needed to describe formal statistical classifications. Several other common vocabularies are also used, where it makes sense.

It is worth noting that the SDMX standard – used as the basis for the Data Cube Vocabulary – and DDI have traditionally made efforts to align their content [4]. Similarly, some of the developers of the DDI vocabularies were also involved in the development of Data Cube, allowing the RDF versions of these standard models to retain that alignment.

The DDI-RDF Discovery Vocabulary presents a good model for vocabulary development: it was the joint product of collaboration between members of the SBE community, DDI experts and implementers, and members of the Linked Data Community. It re-uses other popular vocabularies wherever possible, and can be applied to the research data from many different domains, rather than being specific to a single set of domain data (i.e. census). And it is based on a proven and widely implemented metadata model, sufficient for the demanding requirements of discovering and describing person-level research data.

2 DDI as Linked Data

Statistical domain experts (core members of the DDI Alliance Technical Implementation Committee, representatives of national statistical institutes, national data archives) and Linked Open Data community members have selected the DDI elements which are seen as most important to solve problems associated with use cases in the area of data discovery. This section gives an overview of the conceptual model. More detailed descriptions of all the properties are given in the specification[1] and two conference papers [5, 6]. Disco is intended to provide means to describe microdata by essential metadata for the discovery purpose. Existing DDI-XML instances can be transformed into this RDF format and therefore exposed in the Web of Linked Data. The vice-versa process is not intended, as we have defined Disco components and reused components of other RDF vocabularies which make only sense in the Linked Data field.

Figure 1 gives an overview of the conceptual model containing a small subset of the DDI-XML specification[2]. To understand Disco, there are a few central classes, which can serve as entry points. A **`Study`** represents the process by which a dataset was generated or collected and supports the stages of the full data lifecycle in a modular manner. Literal properties include high-level information about the funding, or-

[1] http://rdf-vocabulary.ddialliance.org/discovery
[2] http://www.ddialliance.org/Specification/

ganizational affiliation, abstract, title, and version. In some cases, where data collection is cyclic or on-going, datasets may be released as a **`StudyGroup`**, where each cycle or "wave" of the data collection activity produces one or more datasets. This is typical for longitudinal studies, panel studies, and other types of "series". In this case, a number of `Study` objects would be collected into a single `StudyGroup`.

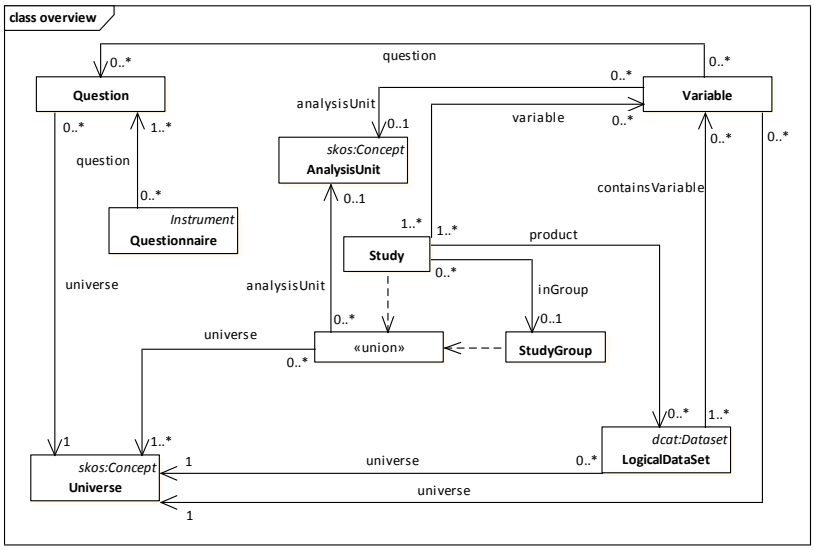

Fig. 1. Overview

Datasets have two representations: a logical representation, which describes the contents of the dataset, and a physical representation, which is a distributed file holding that data. It is possible to format data files in many different ways, even if the logical content is the same. **`LogicalDataSet`** represents the content of the file (its organization into a set of `Variables`). The `LogicalDataSet` is an extension of the `dcat:DataSet`. Physical, distributed files are represented by the **`DataFile`**, which is itself an extension of `dcat:Distribution`. An overview of the microdata can be given either by descriptive statistics or aggregate data (**`qb:DataSet`** originates from the RDF Data Cube Vocabulary). **`DescriptiveStatistics`** may be minimal, maximal, mean values, and absolute and relative frequencies. **`SummaryStatistics`** pointing to variables and **`CategoryStatistics`** pointing to categories and codes are both descriptive statistics.

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

`Definitions` may be related to **`Representation`**`s` 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. **`skos:Concept`** 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. We use **`skos:ConceptScheme`** to represent a set of metadata describing statistical concepts.

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. Data are collected about a specific phenomenon, typically involving some target population of a defined class of people, objects or events. (**`Universe`**), and focusing on the analysis of a particular type of subject (**`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 Disco metadata and the original DDI-XML files. Every element can be related to any `foaf:Document` (DDI-XML files) using `dcterms:relation`. Any entity (especially metadata, studies, data files) can have version information (`owl:versionInfo`). 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. A complete overview of all disco classes and properties can be found in the Disco specification.

3 Use of External Vocabularies

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: representing detailed provenance information of Web data and metadata using the PROV Ontology (PROV-O)[3], describing catalogues of datasets using the Data Catalog Vocabulary (DCAT)[4], describing aggregate data like multi-dimensional tables using the RDF Data Cube Vocabulary[5], describing formal statistical classifications using the SKOS Extension for Statistics (XKOS)[6], delineating code lists, category schemes,

[3] http://www.w3.org/TR/prov-o/
[4] http://www.w3.org/TR/vocab-dcat/
[5] http://www.w3.org/TR/vocab-data-cube/
[6] http://htmlpreview.github.io/?https://github.com/linked-statistics/xkos/blob/master/xkos.html

mappings between them, and concepts like topics using the Simple Knowledge Organization System (SKOS)[7], and the Asset Description Metadata Schema (ADMS)[8] for representing persistent identifiers. Furthermore, we reuse the external vocabularies Friend of a Friend (FOAF)[9] to describe person-level data, the Organization Ontology (ORG)[10] to model organization related information, and the DCMI Metadata Terms (DCMI)[11] to describe general metadata of Disco constructs.

In order to represent detailed provenance information of Web data and metadata, classes and properties of PROV-O can be used. Thus, it can be used as a natural vocabulary to attach provenance information to Disco metadata. Terms of PROV-O are organized among three main classes: `prov:Entity`, `prov:Activity` and `prov:Agent`. While classes of Disco can be represented either as entities or agents, particular processes for, e.g. creating, maintaining and accessing data can be modeled as activities. Properties like `prov:wasGeneratedBy`, `prov:hadPrimarySource`, `prov:wasInvalidatedBy`, or `prov:wasDerivedFrom` describe the relationship between classes for the generation of data in more detail. In order to link from a `disco:Study` to its original DDI XML file, the property `prov:wasDerivedFrom` can be used. Moreover, PROV-O allows for representing versioning information by e.g., using the terms `prov:Revision`, `prov:hadGeneration` and `prov:hadUsage`. PROV-O can also be used to model information and relationships that are relevant for determining accuracy, quality and comparability of a data set with others. By utilizing the properties `prov:qualifiedInfluence` or `prov:wasInformedBy`, qualified statements can be made about a relationship between entities and activities, e.g. that and how a particular method influenced a particular data collection or data preparation process.

DCAT is a W3C working draft for describing catalogs of datasets. DCAT makes few assumptions about the kind of datasets being described, and focuses on general metadata about the datasets (mostly using Dublin Core), and on different ways of distributing and accessing the dataset, including availability of the dataset in multiple formats. Combining terms from both DCAT and Disco can be useful for a number of reasons:

- Describing collections (catalogs) of research datasets
- Providing additional information about physical aspects (file size, file formats) of research data files
- Providing information about the data collection that produced the datasets in a data catalog
- Providing information about the logical structure (variables, concepts, etc.) of tabular datasets in a data catalog

---

7 http://www.w3.org/2004/02/skos/
8 http://www.w3.org/TR/vocab-adms/
9 http://www.foaf-project.org/
10 http://www.w3.org/TR/vocab-org/
11 http://dublincore.org/documents/dcmi-terms/

The LogicalDataSet is an extension of the dcat:DataSet. Physical, distributed files are represented by the DataFile, which is itself an extension of dcat:Distribution.

The RDF Data Cube Vocabulary is a W3C candidate recommendation for representing data cubes, that is, multidimensional aggregate data. A DataSet represents aggregate data such as multi-dimensional tables. Aggregate data is derived from microdata by statistics on groups, or aggregates such as counts, means, or frequencies. Data cubes are often generated by tabulating or aggregating unit-record datasets. For example, if an observation in a census data cube indicates the population of a certain age group in a certain region is 12345, then this fact was obtained by aggregating that number of individual records from a unit-record dataset. Disco contains a property "aggregation" that indicates that a Cube dataset was derived by tabulating a unit-record dataset. Data Cube provides for the description of the structure of such cubes, but also for the representation of the cube data itself, that is, the observations that make up the cube dataset [7]. This is not the case for Disco, which only describes the structure of a dataset, but is not concerned with representing the actual data in it. The actual data are assumed to sit in a data file (e.g. a CSV[12] file, or in a proprietary statistical package file format) that is not represented in RDF.

Skos:Concept is reused to a large extent to represent DDI concepts, codes, and categories. 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. Skos:Concepts may be associated with variables, variable definitions, and questions and are reused to a large extent to represent DDI concepts (skos:prefLabel), codes (skos:notation), and category labels (skos:prefLabel). Skos:Concepts may be organized in skos:ConceptSchemes (skos:inScheme), sets of metadata describing statistical concepts. Hierarchies of DDI concepts can be built using the object peoperties skos:broader and skos:narrower. Topical coverage can be expressed using dcterms:subject. Disco foresees the use of skos:Concept for the description of topical coverage. Spatial, temporal, and topical coverage are directly attached to studies, logical datasets, and datafiles. Universes and AnalysisUnits are also skos:Concepts. Therefore the properties defined for skos:Concept can be reused. KindOfData, pointing to a skos:Concept , describes, with a string or a term from a controlled vocabulary, the kind of data documented in the logical product(s) of a Study. Using dcterms:format, DataFiles formats can be defined.

The use of formal statistical classifications is very common in research datasets - these are treated in Disco as SKOS concepts, but in some cases those working with formal statistical classifications may desire more expressive capability than SKOS provides. To support such users, the DDI Alliance also develops XKOS, a vocabulary which extends SKOS to allow for a more complete description of such classifications [8]. While the use of XKOS is not required by this vocabulary, the two are designed

[12] comma-separated values

to work in complementary fashion. SKOS properties may be substituted by additional XKOS properties.

Especially persons and organizations may hold one or more persistent identifiers of particular schemes and agencies (e.g. ORCID[13], FundRef[14]) that are not considered by the specific IDs of Disco. In order to include those identifiers and for distinguishing between multiple identifiers for the same class, ADMS is utilized. As a profile of DCAT, ADMS aims to describe semantic assets, i.e. reusable metadata and reference data. The class `adms:Identifier` can be added to a `rdfs:Resource` by using the property `adms:identifier`. That identifier class can contain properties that define the particular identifier itself, but also its scheme, version and managing agency. However, although utilized primarily for describing identifiers of persons and organizations, it is allowed to attach an `adms:Identifier` class to all classes in Disco.

4 Use Cases

In this section, we introduce three real world use cases that show the connection and interaction between Disco and other vocabularies. Moreover, all three use cases have in common that they represent real information needs from researchers. First, researches could be interested in the question: which persons and organizations are associated with particular datasets? The second use case describes how to find datasets with a specific statistical classification. The third use case shows the search of data in a data catalog. Additional real world use cases are described in Vompras et al. [9]. 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 datasets using variables, related questions, and classifications. Furthermore, you can search for reusable questions using related concepts, variables, universe, coverage, or by text. The Disco specification contains example data of real world use cases which can be consulted in order to get details of how to construct real world instance data and in order to get a feeling of the full potential of Disco to represent metadata on statistical data.

4.1 Which Persons and Organizations Are Associated with Specific Datasets?

Within the context of Disco, we reuse other well elaborated and accepted vocabularies as often as possible and reasonable. DCMI, FOAF, ORG, ADMS, and PROV-O build one block of complementary vocabularies. Their use is shown in one combined use case. DCMI is used in order to describe general metadata, FOAF and ORG are used to describe persons and organizations, we use ADMS for the persistent identification of objects like persons and organizations, and PROV-O is used to provide provenance

[13] http://orcid.org/
[14] http://www.crossref.org/fundref/

information. A typical scenario within the social sciences community could be the following one:

- John (foaf:person) aggregates (disco:aggregation) microdata datasets (disco:LogicalDataSet) which are associated with (disco:product) the European study EU-SILC (disco:Study). The aggregate dataset is represented using qb:DataSet. The prov:Agent :john was associated with (prov:wasAssociatedWith) the prov:Activity :aggregationActivity. The :aggregationActivity used (prov:used) the prov:Entity :europeanDataSet (a European dataset), and generated (prov:wasGeneratedBy) a new prov:Entity :aggregatedEuropeanDataSet that aggregates the microdata in :europeanDataSet. The prov:Agent :john acted on behalf of (prov:actedOnBehalfOf) the organization :deri (prov:Agent, org:Organization). The European study (disco:Study) was funded by (disco:fundedBy) the research institution GESIS (org:Organization) for which John is working for (org:memberOf). In order to identify foaf:Persons and org:Organizations permanently, the object property adms:identifier is used pointing to adms:Identifiers. Further possible example queries using the vocabularies TERMS, FOAF, ORG, ADMS, and PROV-O would be: Which persons (foaf:Person), working for (org:memberOf) the research institute GESIS (org:Organization), created (dcterms:creator) the survey ALLBUS (Germany General Social Survey), which is a particular group of studies (disco:StudyGroup) in Germany?
- Which organizations (org:Organization) and which persons (foaf:Person) contributed (dcterms:contributor) to the creation of the European study EU-SILC (disco:Study)?
- Which persistent identifier (adms:identifier) are assigned to persons and organizations (foaf:Agent) publishing (dcterms:publisher) the European study EU-LFS (disco:Study)?

4.2 Which Datasets Have A Specific Statistical Classification and What Are Its Semantic Relations?

XKOS extends SKOS with two main objectives: the first one is to allow the description of statistical classifications, the second one is to introduce refinements of the semantic properties defined in SKOS. The semantic properties extend the possible relations that can be applied between pairs of skos:Concepts. SKOS allows the following relations: skos:broader than, skos:narrower than, and skos:related to. The first two are hierarchical relations, one in each direction. In Disco, these SKOS properties may be substituted by additional XKOS properties like xkos:generalizes, xkos:hasPart, xkos:caused, xkos:previous, and xkos:next.

One question, typically asked by social science researchers, could be to query all the datasets (disco:LogicalDataSet) which have a specific statistical classification (skos:ConceptScheme) like ISCO (International Standard Classification of Occupations) or ANZSIC (Australian and New Zealand Industry Classification). It is also possible to query on the semantic relationships which are defined for statistical classifications using XKOS properties. By means of these properties not only hierarchical relations can be queried but also for example part of relationships (xkos:hasPart), more general (xkos:generalizes) and more specific (xkos:specializes) concepts, and positions of concepts in lists (xkos:previous, xkos:next).

4.3 Searching For Data in a Data Collection

While Disco and Data Cube provide terms for the description of datasets, both on a different level of aggregation, DCAT enables the representation of these datasets inside of data collections like repositories, catalogs or archives. The relationship between data collections and their contained datasets is useful, since such collections are a typical entry point when searching for data.

A search for data may consist of two phases. In a first phase, the user searches for different records described by dcat:CatalogRecord inside a data catalog. This search can differ according to the users' information need. While it is possible to search for metadata provided inside such a record like dcterms:title, dcterms:description, etc., the user can also formulate a query to search for more detailed information about the dataset (represented as dcat:Dataset) or its distribution (dcat:Distribution), which are part of the record. For example, a user may want to search for datasets covering a particular topic (dcat:keyword), particular temporal and spatial coverages (dcterms:temporal and dcterms:spatial), or particular formats in which a distribution of the data is available (dcterms:format). Instances of dcat:DataSet are also described by specific themes they cover (dcat:theme). Since these themes are organized in a theme taxonomy (implemented by a skos:ConceptScheme and classes of skos:Concept), these themes can also be used for an overall search in all datasets of the data catalog.

Nevertheless, the search of the first phase will result in one or presumably multiple hits of datasets. Hence, another search has to be executed in a second phase in order to find out which datasets are relevant for the user, e.g. particular universes or samples. The search regarding particular criteria in multiple Disco datasets materializes as those described in the previous two use case sections and those presented in [9]. However, the user may find data sets which are published in Data Cube. In order to discover the original microdata source of a qb:DataSet, the property prov:wasDerivedFrom can hold the link the particular DDI data set disco:Study.

5 Implementation

We have implemented a direct and in parallel a generic mapping between DDI-XML and Disco. In the direct mapping, different versions of DDI XML documents (as defined in the DDI Specification[15]) can be transformed automatically into an OWL ABoxes corresponding to the DISCO vocabulary. The mappings are implemented as XSLT stylesheets[16]. This transformation is useful for existing DDI XML data and enables an easy publication of this data as RDF. Moreover, regardless of different input formats, i.e. different DDI versions, the same Disco output is generated.

The current DDI-XML specification is described using multiple XML Schemas. Bosch and Mathiak [10] have developed a generic approach for designing domain ontologies based on the XML Schema metamodel. XML Schemas are converted to OWL ontologies automatically using XSLT transformations which are described in detail by Bosch and Mathiak [11]. 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' TBoxes and ABoxes can be inferred automatically out of the generated ontologies using SWRL rules [12]. The benefit of the general approach is that the entire DDI model and not only a small subset can be transformed into Disco. Other vocabularies have to be included using intellectual decisions in a second step.

6 Conclusions and Future Work

In this paper, we presented how Disco is connected with other vocabularies and illustrated the interplay between them based on real world use cases. The introduced use cases also show that there is a benefit for searching for data when it is being published using Disco. They motivate an implementation of Disco in information systems (like MISSY[17]) at organizations that hold DDI data. The official review of Disco is planned for late 2013. Additionally, a publication of XKOS is planned.

The interplay of Data Cube, Disco, and PROV-O needs further exploration regarding the relationship of aggregate data, aggregation methods, and the underlying microdata. The goal would be to drill down to the related microdata based on a search resulting in aggregate data. A researcher could then analyze the microdata - often only with constraints of access restrictions to the data (i.e. access only in the closed shop of research data centers or anonymizing methods to assure confidentiality). On the one hand aggregate data are often easily available and gives a quick overview. On the other hand microdata enable more detailed analyses.

Work is going on to add features to Disco which can describe simple data structures like rectangular data (record units by variables). Then Disco could be used to describe for example CSV files on the physical level, the contained variables and

[15] http://www.ddialliance.org/Specification/
[16] https://github.com/linked-statistics/DDI-RDF-tools
[17] https://github.com/missy-project

categories on the logical level, and complementary summary and category statistics as aggregate data.

7 Acknowledgements

The authors would like to thank the participants[18] of several workshops and meetings which were the profound basis of the development of Disco.

8 References

1. Vardigan, M., Heus, P., Thomas, W. Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation 3, 1 (2008), 107–113.
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 2011
3. Jacobs, J.A., Humphrey, C.: Preserving Research Data. Vol. 47, Communications of the ACM (2004)
4. Gregory, A., Heus, P. 2007. DDI and SDMX: Complementary, Not Competing, Standards", Open Data Foundation.
5. Bosch, T., Cyganiak, R., Wackerow, J., Zapilko, B.: Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences. In: International Conference on Dublin Core and Metadata Applications, pp. 46–55. (2012)
6. Bosch, T., Cyganiak, R., Gregory, A., Wackerow, J.: DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. In: Proceedings of the WWW2013 Workshop on Linked Data on the Web, vol. 996, CEUR Workshop Proceedings, Aachen (2013)
7. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic Statistics: Bringing Together SDMX and SCOVO. In: Proceedings of the WWW2010 Workshop on Linked Data on the Web, vol. 628, CEUR Workshop Proceedings, Aachen (2010)
8. Dan Gillman, Franck Cotton, and Yves Jaques. eXtended Knowledge Organization System (XKOS). In: METIS, Work Session on Statistical Metadata, Geneva, Switzerland (2013)
9. Vompras, J., Gregory, A., Bosch, T., Wackerow, J.: Scenarios for the DDI-RDF Discovery Vocabulary. In: DDI Working Paper Series – Semantic Web 2 (2013)
10. Bosch, T., Mathiak, B.: Generic Multilevel Approach Designing Domain Ontologies based on XML Schemas. In: Proceedings of the ISWC 2011 Workshop Ontologies Come of Age in the Semantic Web (OCAS), vol. 809, CEUR Workshop Proceedings, Aachen (2011) 1-12
11. Bosch, T., Mathiak, B.: XSLT transformation generating OWL ontologies automatically based on XML Schemas. In:6th International Conference for Internet Technology and Secured Transactions (ICITST), pp. 660 –667 (2012)
12. 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

[18] http://rdf-vocabulary.ddialliance.org/discovery.html#acknowledgements