DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data

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

Richard Cyganiak
Digital Enterprise Research Institute
richard.cyganiak@deri.org

Arofan Gregory
Open Data Foundation
agregory@opendatafoundation.org

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

ABSTRACT

The Data Documentation Initiative (DDI) is an acknowledged international standard for the documentation and management of data from the social, behavioral, and economic sciences. Statistical domain experts, i.e. representatives of national statistical institutes and national data archives, and Linked Open Data community members have developed the DDI-RDF Discovery Vocabulary – based on a subset of the DDI - in order to support the discovery of statistical data as well as metadata. This vocabulary supports identifying programmatically the relevant data sets for a specific research purpose.

Categories and Subject Descriptors

H.1 **[Information Systems]**: Models and Principles

General Terms

Documentation, Design, Standardization

Keywords

Linked Data, Ontology Design, statistical data, DDI-RDF

1. INTRODUCTION

For data professionals - researchers, data librarians, and statisticians - the term “data” refers to some very specific types of what most people think of as “data”. For those in the Linked Data community, “data” is a very broad term indeed, embracing basically anything accessible on the Web. In developing an RDF vocabulary for describing research data, it is important to understand the narrower “professional” definition of data, since it is this which is to be described using the new vocabulary, not data in the more general sense. Throughout this paper, the narrower term, as used by data specialists, is what is meant unless otherwise stated.

So, what does the term “data” mean? It actually has several distinct meanings: “raw” data; “unit-record” data (also called “micro-data”); and “aggregate” data (also called “tabulated data”). We will attempt to characterize each type here, as they serve different purposes. In order to understand them, it is important to understand how each fits into the data lifecycle, as data is collected and processed to support research.

“Raw” data refers to the set of numbers and other values (often coded values coming from concept schemes r classifications) which are the direct input into the research process. These are often the result of surveys. For instance, each person responding to a survey might be asked: “What is your gender?” The answer would be either “Male” or “Female”, which is a very simple example of a concept scheme for gender. In the data set, the responses might be recorded as “1” (for Male) or “2” (for Female). Raw data can also come from other sources, such as devices performing measurements, or from administrative registers (databases containing information collected for non-research purposes, but which are useful for research, such as registers of births and deaths, clinical systems in hospitals, etc.)

Once collected, “raw” data is processed further, to clean up values which are wrong or likely to distort the research. Some other values are processed into a form which is easy to work with. If there are missing values, these need to be handled (determine why they are missing, etc.) There are several types of processing, and it is not important to understand them all for the purposes of this article. What is important is that the result of this processing in no longer “raw data”, but is instead a useful “unit-record” data set.

The structure of unit-record data is very specific: think of a table, where each column contains a particular type of value (gender, age, response to a particular question in a survey, etc.) and each row represents the responses for a single “unit” (typically an individual, a household, or a business, etc.). By further processing the (usually large) number of cases, a different type of data is produced – aggregate or tabulated data.

Take, for example, the following unit-record data set, recording gender, age, highest education degree attained, and current place of habitation:

Table 1. unit-record data set

Case ID	Gender	Age	Degree	Habitation
1	Male	22	High school	Arizona
2	Female	36	PhD	Wisconsin
3	Male	50	PhD	New Mexico

LDOW2013, May 14, 2013, Rio de Janeiro, Brazil.

[1] http://stats.oecd.org/glossary/

4	Female	65	PhD	Texas
5	Male	23	BA	Vermont

In analyzing this data set, we might decide that older people tend to have higher degrees of education, since no one under the age of 36 has a PhD, and of those under the age of 36, only 50% has a college degree of any type. I could “aggregate” (or “tabulate”) my unit record data into the following aggregate data set:

Table 2. aggregate data set

Age	% with High-School	% with BA	% with PhD
Age < 36 years	50	50	0
Age > 36 years	0	0	100

Now, this is a ridiculous example – we have too few cases. But you can see that by focusing on some of the columns in our unit-record data, we can create a table (in this case, age by educational degree). Such tabulations are used by researchers analyzing data to prove or disprove research hypothesis. Tabulations are also created by government statisticians working to support policy decisions.

When we consider the types of data which exist on the Web, and which could be represented on the Web as the result of open data initiatives, we can see that at least the second categories (unit-record data and aggregate data) would be very useful, and in some cases even raw data might be useful, as in the case of government expenditures, for example.

It is very important to understand the distinctions between these various types of data, because they are useful for different purposes. Aggregate data can be used to draw many useful types of charts and visualizations, but this cannot be done usefully with unit-record data or raw data, to make a simple point. For most of us, the aggregate data is very useful and easy to understand – unit-record data requires a higher degree of statistical literacy to make sense of.

Both, unit-record and aggregate data is understood as research data. This means any data which is used for research not just data which is collected for research purposes.

When working with data of any type, it is not the data alone which is needed – also required to understand and analyze data is a very detailed level of metadata. How is a column in my unit-record data defined? Where do the values come from? What was the population being surveyed, and how was it selected? This type of metadata includes, but goes well beyond the metadata found in something like Dublin Core, for example. It is highly specialized, and is specific to the type of data being described.

Within the world of data professionals, two metadata standards have emerged which are becoming widely adopted. For raw data and unit-record data, the metadata standard is called the “Data Documentation Initiative” (DDI). For aggregate data, the standard is known as the “Statistical Data and Metadata Exchange” (SDMX, now ISO-TS 17369). Both, DDI and SDMX can describe the structure of multi-dimensional data cubes, and then provide the data formats for these. SDMX focuses on processing and exchanging the data, DDI on documenting the aggregation processes, in case they are of interest to researchers. The overlap and the difference are described in detail by Gregory and Heus [5].

An RDF vocabulary has already been created for describing aggregates – the “Data Cube Vocabulary”, which was based on the SDMX metadata model. Until now, there has not been an RDF vocabulary for describing raw data or unit-record data.

Today, we are seeing individuals from the Linked Data community and the data professional community come together to produce such a vocabulary, on the basis of the DDI model.

2. MOTIVATION

We will look at the motivations of individuals from the two different communities separately, because while they are complementary, they are very different.

2.1 Data Professionals and the DDI Community

For data professionals, the use of the data they produce or disseminate is often a primary goal. For those working in data archives and data libraries, the service they offer is access to research data for secondary use, and an increase in the use of their data is perceived as a positive sign. For government statisticians, it is the same – they produce the “official” data to support policy, and they perceive the use of their data as a contribution to society, and a fulfillment of their mission. For researchers, the re-use of the data they collect is something which enhances their reputation and career, through an emerging system of “data citation” which is very similar to the traditional citation of research papers. Thus, the various members of the DDI community are very interested in having their data be discovered and used.

This is somewhat problematic, however – it is not enough simply to publish data to the Web, which is very often illegal for reasons of privacy and confidentiality. Instead, a researcher looking for unit-record data is often required to apply for access, and to make commitments about how the data will be used and released. These issues are taken very seriously by data professionals for a variety of reasons: first, if people asked to fill out a survey do not trust the person administering the survey, they will refuse to respond, making the collection of good raw data with surveys difficult or impossible. Thus, researchers want to be trusted by the people they study. Additionally, the release of confidential information is illegal and potentially very destructive, and can result in prosecution.

The degree of “statistical literacy” among users is always a major concern with those who work at data libraries and archives, supporting researchers. When using raw data and unit-record data, there is a significant set of skills which are required to produce valid research. These skills require access to the best possible metadata about the data. This is especially true when working with data coming from different sources, something which researchers are often very keen to do.

Thus, the DDI community is torn in two directions: on one hand, they very much want people to use their data, and thus are very interested in advertising their data through the Web of Linked

Data; on the other hand – and especially after seeing the metadata-free nature of many data sets published at open data sites such as data.gov in the US – they are concerned at the lack of standard, highly-detailed metadata which is required for the correct analysis and use of unit-record data.

Ultimately, the DDI-based RDF vocabulary being developed here is done as a way of making sure that when unit-record (or raw) data is published into the Web of Linked Data, this will be done in a way which allows for correct and responsible use of that data. The basic idea here is to reap the benefits of broader use of existing data resources, while benefitting from the knowledge and experience of working with data which is the hallmark of the DDI community and its members.

2.2 The Linked Data Community

From the perspective of the linked Data community, the benefit is a simple one – to put all of the data holdings of data archives and statistical organizations into the Web of Linked Data. The vocabulary being developed is one which will encourage the holders of the data to be confident that sufficient metadata is being published to permit discovery and use of the data. RDF-based tools will be able to take advantage of this publication, without requiring the use of the complicated XML schemas which most DDI implementations require. Additionally, data sets described using this vocabulary can be easily linked with other data sets, and can be more easily connected to related Web-based descriptions, making the data and the results of research more closely connected. Further, the possibility exists of making explicit the metadata around published, but under-documented data sets from open government initiatives, in a standard and understood form, by organizations other than those which published the data sets themselves.

3. DATA DOCUMENTATION INITIATIVE

The DDI specification[2] describes social science data, data covering human activity, and other data based on observational methods measuring real-life phenomena [7]. DDI[3] 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 e.g. a census or a survey. Statistical microdata is 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 when access to microdata is restricted as 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 which understand the Dublin Core XML, but which do not understand DDI. DDI is aligned with other metadata standards as well, with SDMX[4] (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[5]. Academic users include the UK Data Archive at the University of Essex, the DataVerse Network at the Harvard-MIT Data Center, and the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. Official data producers in more than 50 countries include the Australian Bureau of Statistics (ABS) and many national statistical institutes of the Accelerated Data Program for developing countries. Examples for international organizations are UNICEF, the Multiple Indicator Cluster Surveys (MICS), The World Bank, and The Global Fund to Fight AIDS, Tuberculosis and Malaria.

Data Lifecycle. Common understanding is that both statistical data and metadata is 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 [6]. 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.

[2] http://www.ddialliance.org/Specification/

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

[4] http://sdmx.org/

[5] http://www.ddialliance.org/ddi-at-work/projects

Multiple institutions are involved in the data lifecycle which is an interactive process with multiple feedback loops.

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.

4. 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 chosen the DDI elements which are seen as most important to solve problems associated with diverse identified use cases in the area of data discovery. 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[6], 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[7] and a conference paper [2]. The DDI-RDF Discovery Vocabulary is intended to provide means to describe data 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 DDI-RDF components and reused components of other RDF vocabularies which make only sense in the Linked Data field.

4.1 Overview

Figure 1 gives an overview over the conceptual model containing a small subset of the DDI-XML specification[8]. To understand the DDI-RDF 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 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`.

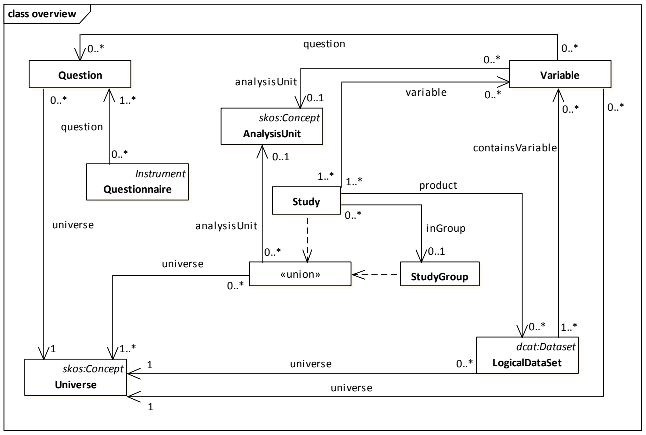

Figure 1. Overview

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.

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

[7] http://rdf-vocabulary.ddialliance.org/discovery

[8] http://www.ddialliance.org/Specification/

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 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 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[9], an approach to map the SDMX information model to an ontology. A DataSet represents aggregated data such as multi-dimensional tables. Aggregated data is 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.

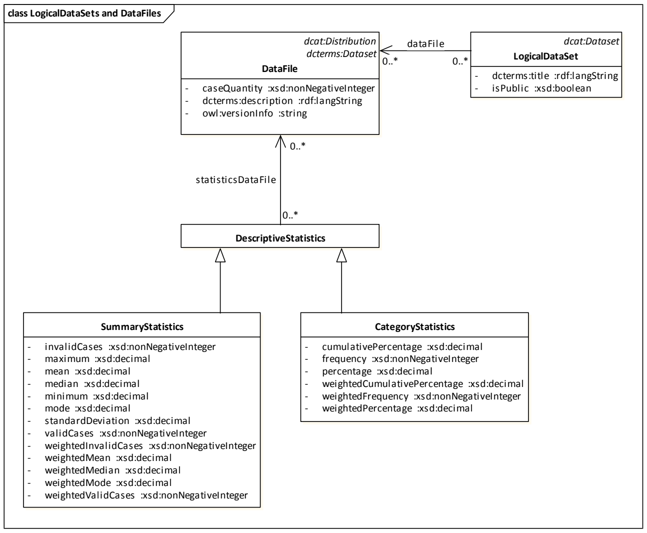
class LogicalDataSets and DataFiles
dcat:Distribution
dcterms:Dataset
DataFile
- caseQuantity :xsd:nonNegativeInteger
- dcterms:description :rdf:langString
- owl:versionInfo :string
dataFile
0..*
0..*
dcat:Dataset
LogicalDataSet
- dcterms:title :rdf:langString
- isPublic :xsd:boolean
0..*
statisticsDataFile
0..*
DescriptiveStatistics
SummaryStatistics
- invalidCases :xsd:nonNegativeInteger
- maximum :xsd:decimal
- mean :xsd:decimal
- median :xsd:decimal
- minimum :xsd:decimal
- mode :xsd:decimal
- standardDeviation :xsd:decimal
- validCases :xsd:nonNegativeInteger
- weightedInvalidCases :xsd:nonNegativeInteger
- weightedMean :xsd:decimal
- weightedMedian :xsd:decimal
- weightedMode :xsd:decimal
- weightedValidCases :xsd:nonNegativeInteger
CategoryStatistics
- cumulativePercentage :xsd:decimal
- frequency :xsd:nonNegativeInteger
- percentage :xsd:decimal
- weightedCumulativePercentage :xsd:decimal
- weightedFrequency :xsd:nonNegativeInteger
- weightedPercentage :xsd:decimal

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. **Variable**s 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 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").

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

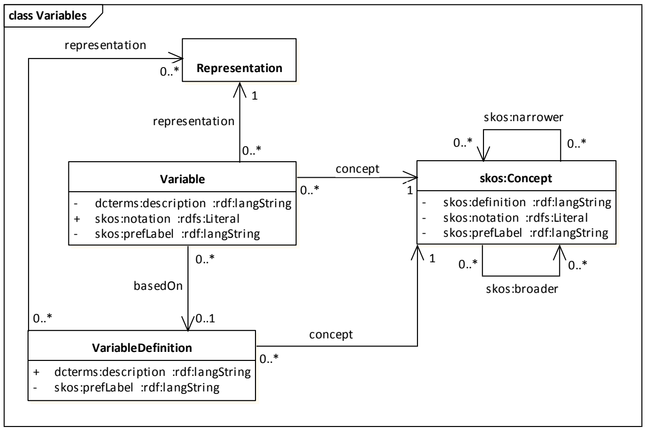
class Variables
representation
0..*
Representation
1
representation
0..*
Variable
- dcterms:description :rdf:langString
+ skos:notation :rdfs:Literal
- skos:prefLabel :rdf:langString
concept
0..*
1
skos:Concept
- skos:definition :rdf:langString
- skos:notation :rdfs:Literal
- skos:prefLabel :rdf:langString
skos:narrower
0..*
0..*
skos:broader
0..*
0..*
1
0..*
basedOn
0..1
0..*
VariableDefinition
+ dcterms:description :rdf:langString
- skos:prefLabel :rdf:langString
concept
0..*

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. We use the general **`skos:ConceptScheme`** class to represent 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`.

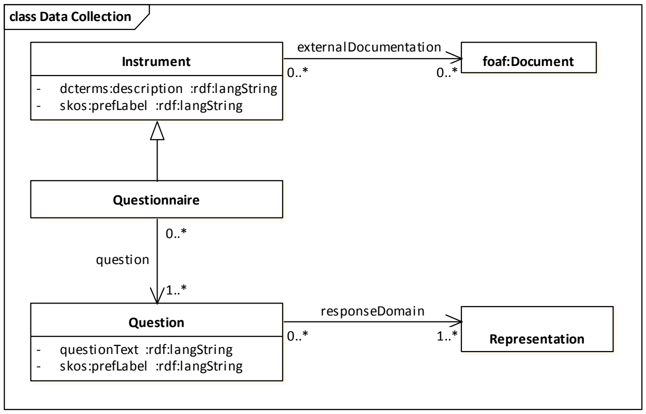

Figure 4. Data Collection

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`).

4.6 Implementation

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 OWL ABox corresponding to the ontology. The direct mappings are realized by XSLT stylesheets[10]. Bosch and Mathiak [3] 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 [4]. 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 [1].

5. USE CASES

The use cases are oriented on the discovery of data in the Linked Data context and possible usage within the web of data.

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.

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

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.

Vompras et al. describe further possible use cases in detail [8]. 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.

6. CONCLUSIONS AND FUTURE WORK

We introduced the DDI-RDF Discovery Vocabulary which is designed to support the discovery of microdata sets and related metadata using RDF technologies in the Web of Linked Data. Many archives and other organizations have large amounts of data, sometimes publically available, but often confidential in nature, requiring applications for access. Many such organizations use the Data Documentation Initiative standard, which is a proven and highly detailed XML metadata format for describing rectangular data sets of this type. This vocabulary leverages the DDI specification to create a simplified version of this model for the discovery of data files. This vocabulary is intended not only for use by the research data community, but also by any others needing an RDF vocabulary for describing this type of rectangular data. This vocabulary will provide a useful model for describing some of the data sets now being published by open government initiatives, by providing a rich metadata structure for them. 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. This vocabulary would help to overcome this difficulty by allowing for the creation of standard queries to programmatically identify data sets, whether made available by government or held within a data archive.

The DDI-RDF Discovery Vocabulary is planned as a specification of the DDI Alliance a global consortium which drives the development of standards in the area of the social, behavioral, and economic sciences. The DDI Alliance is a self-sustaining membership organization whose members have a voice in the development of the DDI standards. The ongoing work is continued in a DDI Alliance working group. A public review of the vocabulary is planned while 2013.

7. ACKNOWLEDGEMENTS

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"[11] 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)[12] in Gothenburg, Sweden in December 2011, second workshop on "Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web"[13] 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

[11] http://www.dagstuhl.de/11372

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

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

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).

8. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Gregory, A., Heus, P. 2007. DDI and SDMX: Complementary, Not Competing, Standards", Open Data Foundation.

[6] Jacobs, J.A. and Humphrey, C. 2004. Preserving Research Data. Communications of the ACM 47, 9.

[7] Vardigan, M., Heus, P., and Thomas, W. Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation 3, 1 (2008), 107–113.

[8] Vompras, J., Gregory, A., Bosch, T., and Wackerow, J. 2013. Scenarios for the DDI-RDF Discovery Vocabulary. DDI Working Paper Series – Semantic Web 2 (May 2013). DOI= http://dx.doi.org/10.3886/DDISemanticWeb02.

Appendix – Working Example

In this section, we describe a full working example which is explained in detail in the DDI-RDF Discovery Vocabulary specification[14]. You can download the full working example from a GitHub repository[15]. We have a sample of a survey which has been documented using DDI-XML - the 1980 Argentine National Population and Housing Census. The version of this data we are using as our example is the one disseminated by IPUMS[16], which provides internationally harmonized census data, to make it more useful for cross-border research. Thus, this data set is produced by two organizations: The Argentine National Institute of Statistics and Censuses, and the Minnesota Population Center housed in the University of Minnesota.

Using the DDI-RDF Discovery Vocabulary, the study can also be described in triples: an instance of type `Study` is given the title and the identifier; also, the two data producers are linked and further described. The year and country are described in the form of a temporal and spatial coverage of the study. Also, the topics of the study are represented. The study instance further contains an abstract. Since a study is a versionable object in DDI, we attach a version to it.

```
<#study> a disco:Study;
  dc:title "National Population and Housing
Census, 1980";
  dc:identifier "ARG_1980_PHC_v01_A_IPUMS";
  dc:creator [
    rdfs:label "Minnesota Population…";
    skos:notation "MPC".];
  dc:temporal [
    a dcterms:PeriodOfTime;
    disco:startDate "1980-10-22"^^xsd:date;
    disco:endDate "1980-10-22"^^xsd:date.];
  dc:spatial [
    a dc:Location;
    rdfs:label "Argentina…".];
  dcterms:subject [
    skos:definition "Technical…".];
  dcterms:abstract "…";
  owl:versionInfo "Version 1.0…";
  disco:universe <#universe>;
  disco:instrument <#questionnaire>;
  disco:product <#dataset>;
  disco:analysisUnit <#analysisUnit>;
  disco:kindOfData <#kindOfData> ;
  disco:variable <#AR80A401>.
```

The study refers to a specific universe.

```
<#universe> a disco:Universe;
  skos:definition "All the population…".
```

Using a questionnaire, the study produced a dataset. The dataset has access rights. The dataset has a concrete data file that will populate certain variables.

```
<#dataset> a disco:LogicalDataset;
  disco:instrument <#questionnaire>;
  dcterms:accessRights;
  disco:dataFile <#datafile>;
  disco:containsVariable <#AR80A401>.
```

The Units of Analysis and Kind of Data further describe the study.

```
<#analysisUnit> a disco:AnalysisUnit;
  skos:definition "Dwelling…".
<#kindOfData> a skos:Concept;
  rdfs:label "Census/enumeration data".
```

The questionnaire contains several questions having a text.

```
<#questionnaire> a disco:Questionnaire;
  disco:question <#questionGender>;
  disco:question <#questionAge>;
  disco:question <#questionCitizenship>.
<#questionGender> a disco:Question;
  disco:questionText "2. Is the person a man or a
woman? [] Man, [] Woman".
```

Any variable has a text and is based on a variable definition.

```
<#AR80A401> a disco:Variable;
  dc:identifier "AR80A401";
  skos:prefLabel "Sex";
  dc:description "This variable…";
  disco:basedOn <#sexVD>;
  disco:hasQuestion <#questionGender>.
```

Any variable definition has a representation defining the possible values of a variable. Also, a variable definition has its own universe and DDI concepts further describing the variable.

```
<#sexVD> a disco:VariableDefinition;
```

[14] http://rdf-vocabulary.ddialliance.org/discovery

[15] https://github.com/linked-statistics/disco-spec

[16] http://www.ipums.org/

```
  disco:universe <#universePerson>;
  disco:representation <#sexRepr>;
  disco:concept <#ipumsC1>;
  skos:prefLabel "Sex";
  dc:description "Sex data element".
<#sexRepr> a skos:ConceptScheme,
disco:Representation;
  skos:hasTopConcept <#sexM>, <#sexF>.
<#sexM> a skos:Concept;
  skos:notation "1";
  skos:prefLabel "Male";
  skos:inScheme <#sexRepr>.
<#sexF> a skos:Concept;
  skos:notation "2";
  skos:prefLabel "Female";
    skos:inScheme <#sexRepr>.
```

Any universe of a variable definition is a subset of the universe of the entire study.

```
<#universePerson> a disco:Universe;
  skos:definition "All persons." ;
  skos:narrower <#universe>.
```

DDI concepts can be hierarchically structured.

```
<#ipumsCS> a skos:ConceptScheme;
    skos:hasTopConcept <#ipumsC1>.
<#ipumsC1> a skos:Concept;
  skos:prefLabel "Demographic Variables…";
  skos:inScheme <#ipumsCS>.
```

The usage of a variable definition within a data file can be described using statistics.

```
<#dstat1> a disco:DescriptiveStatistic;
  disco:frequency 13314444;
  disco:percentage 49.97;
  disco:hasStatisticsVariable <#AR80A401>;
  disco:hasStatisticsCategory <#sexM>;
  disco:hasStatisticsDatafile <#datafile>.
```

Finally, the data file more concretely describes the actual physical file.

```
<#datafile> a disco:Datafile;
  dc:identifier "ARG1900-P-H.dat";
  dc:description "Person records";
  disco:caseQuantity 2667714;
  dc:format "ascii";
  dc:provenance "Minnesota Population…";
  owl:versionInfo "Version 1.0…";
  dc:spatial[
    a dc:Location;
    rdfs:label "Argentina…".];
  dc:temporal "PeriodOfTime";
  dc:subject "To be defined".
```