Unofficial Draft

DDI-RDF Discovery Vocabulary

A vocabulary for publishing metadata about data sets (research and survey data) into the Web of Linked Data

Unofficial Draft 02 August 2013

Latest editor's draft:
https://raw.github.com/linked-statistics/disco-spec/master/discovery.html

Editors:
Thomas Bosch, GESIS - Leibniz Institute for the Social Sciences
Richard Cyganiak, DERI, NUI Galway
Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences
Benjamin Zapilko, GESIS - Leibniz Institute for the Social Sciences

Authors:
Thomas Bosch, GESIS - Leibniz Institute for the Social Sciences
Franck Cotton, INSEE - Institut National de la Statistique et des Études Économiques
Richard Cyganiak, DERI, NUI Galway
Arofan Gregory, Open Data Foundation (ODaF)
Benedikt Kämpgen, Karlsruhe Institute of Technology
Olof Olsson, SND - Swedish National Data Service
Heiko Paulheim, University of Mannheim
Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences
Benjamin Zapilko, GESIS - Leibniz Institute for the Social Sciences

---

Abstract

This specification defines the DDI Discovery Vocabulary, an RDF Schema vocabulary that enables discovery of research and survey data on the Web. It is based on DDI (Data Documentation Initiative) XML formats.

Status of This Document

This document is merely a public working draft of a potential specification. It has no official standing of any kind and does not represent the support or consensus of any standards organisation.

This document is a Working Draft produced by the RDF Vocabularies Working Group, a working group at the DDI Alliance.

Development resources:

- Google Group
- Issue tracker
- GitHub repository

Table of Contents

Table of Figures

1. Introduction

Here's the LODE view of the whole thing. And here's the Turtle source. And open issues.

1.1 Scope and Purpose

This specification 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.

The data holdings of data archives are often collected by researchers, and only afterwards disseminated by archives. Other data-producing organizations such as research centers and statistical agencies are also increasingly interested in the DDI standards for documenting their own micro-data. In general terms, most DDI metadata describes data sets for the social, behavioural, and economic sciences. This data is fairly consistent in format, consisting of rectangular data files with columns containing variables for a set of cases, contained in the rows. It is often collected by survey, although in some cases may come from administrative sources, sensors, or registers.

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 document [Scenarios] by Vompras, Gregory, Bosch, Capadisli, and Wackerow describes typical use cases for the applicability of the DDI-RDF Discovery vocabulary. In the Section Example Queries of the Appendix additional discovery use cases are illustrated by several SPARQL queries.

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 and two conference papers [Linked-Statistical-Data] [DDI-RDF-Discovery-Vocabulary]. 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.

1.2 About DDI

The Data Documentation Initiative standards are produced and maintained by a member-based consortium of global scope, the DDI Alliance. Housed currently at the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan, there are currently more than 30 member institutions. The standards have been under development for more than ten years, and are in widespread use among data archives and libraries, producers of research data, secure data centers, and statistical agencies.

There are two major versions of DDI: the “codebook” version, which is an XML format for holding general information about a study, along with its data dictionary; and the “Lifecycle” version of DDI, which allows for the description of more complex multi-wave studies, throughout the data lifecycle, from study conception through data collection and processing.

This vocabulary is not specific to either of these versions, but represents the major types of metadata they contain in a highly simplified form, for the purposes of discovery. The XML Codebook and Lifecycle versions of DDI are very broad: these standards contain hundreds of metadata elements, providing enough information to programmatically work with the data files for such functions as the automatic creation of databases, and transformations between statistical packages. DDI in both versions is generally used to describe data found in ASCII files, whether positional files with fixed-width fields or files using a delimited format such as CSV.

It is difficult to claim that there is a single agreed conceptual model for describing research data in the social, behavioural, and economic sciences—there is a wide range of models and terms. However, the issues faced in this area have been the subject of discussion within the DDI community for many years, and the DDI model represents the best consensus which exists today. As such, it gives us a good basis for creating a vocabulary which will be recognizable to researchers familiar with this type of data.

1.3 Relationship to Data Cube, DCAT and XKOS

The Discovery Vocabulary is aligned to several other metadata vocabularies used in the RDF community.

The Data Catalog Vocabulary (DCAT) is a W3C standard for describing catalogs of datasets, and we map to it in two places: Our `LogicalDataSet` is a subclass of DCAT's Dataset, and our `DataFile` is a subclass of DCAT's Distribution. 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 the Discovery Vocabulary 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

The Data Cube vocabulary is a W3C standard for representing data cubes, that is, multidimensional aggregate data. Data cubes are often generated by tabulating or aggregating record-level 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 record-level (or "microdata") dataset. The Discovery Vocabulary contains a property "aggregation" that indicates that a Cube dataset was derived by tabulating a record-level 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. This is not the case for the the Discovery Vocabulary, which only describes the structure of a dataset, but is not concerned with representing the actual data in it. The actual data is assumed to sit in a data file (e.g., a CSV file, or in a proprietary stats package file format) that is not represented in RDF.

The interplay of Data Cube and Disco 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. On the one hand aggregate data are often easily available and gives a quick overview. On the other hand microdata enable more detailed analyses.

The use of formal statistical classifications is very common in research data sets—these are treated in our vocabulary 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 publishes XKOS, a vocabulary which extends SKOS to allow for a more complete description of such classifications. While the use of XKOS is not required by this vocabulary, the two are designed to work in complementary fashion.

1.4 Limitations

2. Overview

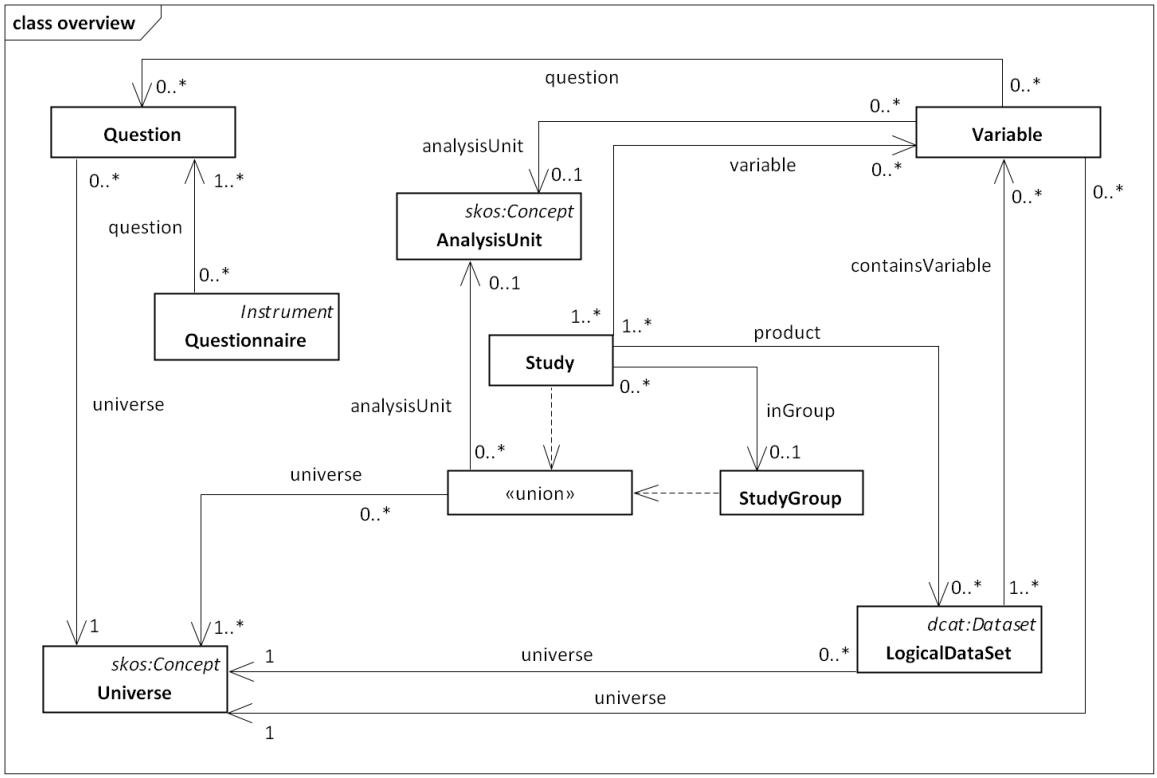


Fig. 1 Vocabulary Overview

To understand the DDI Discovery Vocabulary, there are a few central classes, which can serve as entry points. The first of these is the `Study` class. A **`Study`** in our model 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" (to use the DDI term). In this case, a number of `Study` objects would be collected into a single `StudyGroup`.

Data sets have two representations in our model: 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. In our model the **LogicalDataSet** represents the content of the file (its organization into a set of variables (Variable)). The LogicalDataSet is an extension of the dact:DataSet class. Physical, distributed files are represented by the class **DataFile** (not depicted in the diagram), which is itself an extension of the dcat:Distribution.

When it comes to understanding the contents of the data set, this is done using the **Variable** class. Variables (Variable) 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 (Variable) 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 **Universe** class and the **AnalysisUnit** class. If, for example, the adult population of Finland is being studied, the AnalysisUnit would be individuals or persons. Bosch, Cyganiak, Wackerow, and Zapilko give a detailed overview of the DDI-RDF Discovery Vocabulary in a full paper written for the Dublin Core conference [Linked-Statistical-Data].

3. A Worked Example

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, 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.

To give some idea of what is contained in the metadata set, we will use some screen shots from OpenMetadata Survey Catalog, a portal which indexes the DDI files to facilitate searching, and reflects the contents in a fashion which is easy to view. Follow this link for the information about this DDI file at the OpenMetadata Survey Catalog.

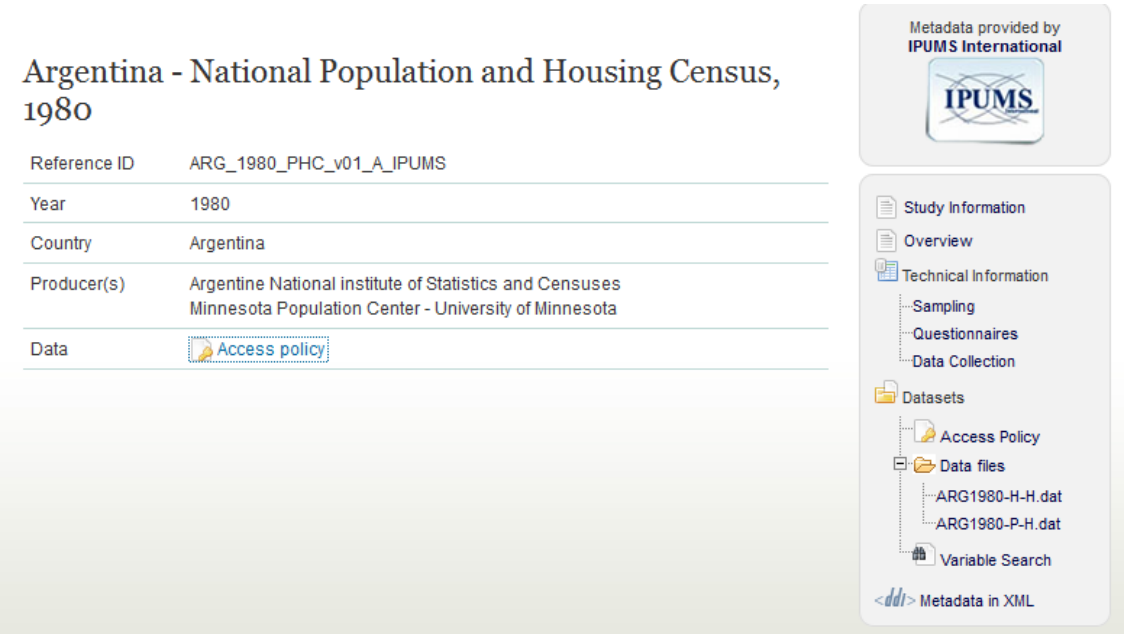


Fig. 2 Overview

Figure 2 shows us the overview page for this study, giving us some basic information - title, an identifier for the study, data producers, year, country, and a link to the access policies. If we look at the right-hand panel, we see an outline of the metadata contents of the file, including information about the questionnaire used, sampling methodology, and data collection activities, as well as detailed information about the variables contained in the two data files.

Not all of this information is useful in a data discovery scenario—sampling and data collection methodologies are not typically indexed for searches. Information about the questionnaire is, as is detailed information about the variables contained in the files. We will look more closely at the metadata of primary interest for our discovery scenario.

Using RDF and the DDI Discovery Vocabulary, the study can also be described in triples: An instance of type of 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. A study is further described using additional information which is described further below.

EXAMPLE 1

```
<#Study> a disco:Study;
    dcterms:title "National Population and Housing Census, 1980"@en;
    dcterms:identifier "ARG_1980_PHC_v01_A_IPUMS".
    dcterms:creator [
        rdfs:label "Minnesota Population Center"@en;
        skos:notation "MPC";
        org:memberOf [
```

```
            rdfs:label "University of Minnesota"@en;
        ];
    ];
    dcterms:creator [
        rdfs:label "Argentine National institute of Statistics and Censuses"@en;
    ]
        dcterms:temporal [
                a dcterms:PeriodOfTime ;
                disco:startDate "1980-10-22"^^xsd:date;
                disco:endDate "1980-10-22"^^xsd:date;
                rdfs:comment "The interviews take place on the expected census day. In
                        some areas the enumeration took place the following day because of
                        access problems due to heavy rains.";
    ];
    dcterms:spatial [
        # This is the DC-strictly compatible way to do it
        a dcterms:Location;
        rdfs:label "Argentina, national coverage"@en;
    ];
    # Only a subset of subjects mentioned in the original file
    dcterms:subject [
                skos:definition "Technical Variables -- HOUSEHOLD"@en ;
    ] ;
    dcterms:subject [
                skos:definition "Group Quarters Variables -- HOUSEHOLD"@en ;
    ] ;
        dcterms:abstract "IPUMS-International is an effort to inventory, preserve,
                harmonize, and disseminate census microdata from around the world. The
                project has collected the world's largest archive of publicly available
                census samples. The data are coded and documented consistently across
                countries and over time to facilitate comparative research. IPUMS-
                International makes these data available to qualified researchers free
                of charge through a web dissemination system. The IPUMS project is a
                collaboration of the Minnesota Population Center, National Statistical
                Offices, and international data archives. Major funding is provided by
                the U.S. National Science Foundation and the Demographic and Behavioral
                Sciences Branch of the National Institute of Child Health and Human
                Development. Additional support is provided by the University of
                Minnesota Office of the Vice President for Research, the Minnesota
                Population Center, and Sun Microsystems.";

    owl:versionInfo "Version 1.0. This version contains selected variables from
                the original census micro data plus harmonized variables from the IPUMS
                International data base."@en;

    disco:universe <#Universe>;
    disco:instrument <#Questionnaire>;
    disco:product <#Dataset>;

    disco:analysisUnit <#AnalysisUnit>;
    disco:kindOfData <#KindOfData>;

    # stdyInfo/notes currently not represented.
    disco:variable <#AR80A401>, <#AR80A402>, <#AR80A404>, <#AR80A407>, <#AR80A411>.
```

While the sampling methodology may not be of great interest for those searching for data, one field within this section is: the “universe”, that is, the population being studied. Figure 3 gives us an example of this information.

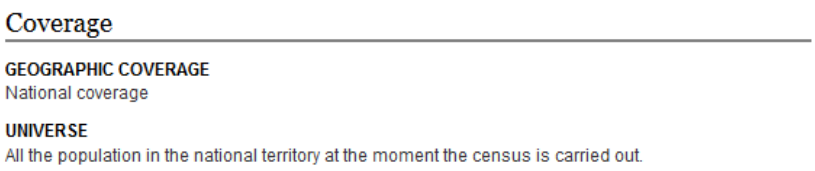


Fig. 3 Coverage and Universe

Thus, the study refers to a specific universe.

EXAMPLE 2

```
<#Universe> a disco:Universe;
    skos:definition "All the population in the national territory at the moment the census is carried out."@en .
```

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

EXAMPLE 3

```
<#Dataset> a disco:LogicalDataset;
    disco:instrument <#Questionnaire>;
    dcterms:accessRights <#AccessRights>;
    disco:dataFile <#Datafile>;
    disco:containsVariable <#AR80A401>, <#AR80A402>, <#AR80A404>, <#AR80A407>, <#AR80A411>.

<#AccessRights> dcterms:description "IPUMS-International distributes
        integrated microdata of individuals and households only by agreement ...
        designed to extend this record.";
    rdfs:seeAlso <http://microdata.worldbank.org/index.php/catalog/442/accesspolicy>.
```

Figure 4 shows us the information about access policies, which typically is of interest to those searching for data.

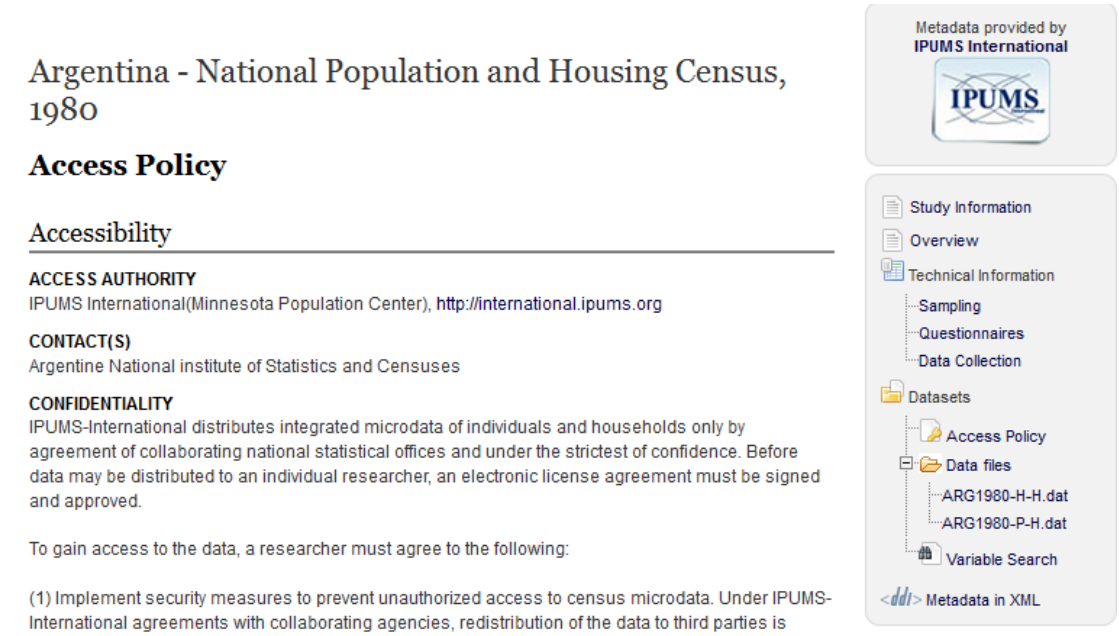


Fig. 4 Access Policy

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

EXAMPLE 4

```
<#AnalysisUnit> a disco:AnalysisUnit ;
    skos:definition "Dwelling, quarter dwelling, census household, and population"@en .

<#KindOfData> a skos:Concept ;
        rdfs:label "Census/enumeration data [cen]"@en .
```

In some cases we may have a lot of information about the questionnaires used, and it is very common to search for data by the text of the question used to collect it. Sometimes there will be a PDF of a questionnaire, and sometimes question text may be linked to individual variables within a file. In this case, we have only a textual description of the set of forms used in the census (Figure 5).

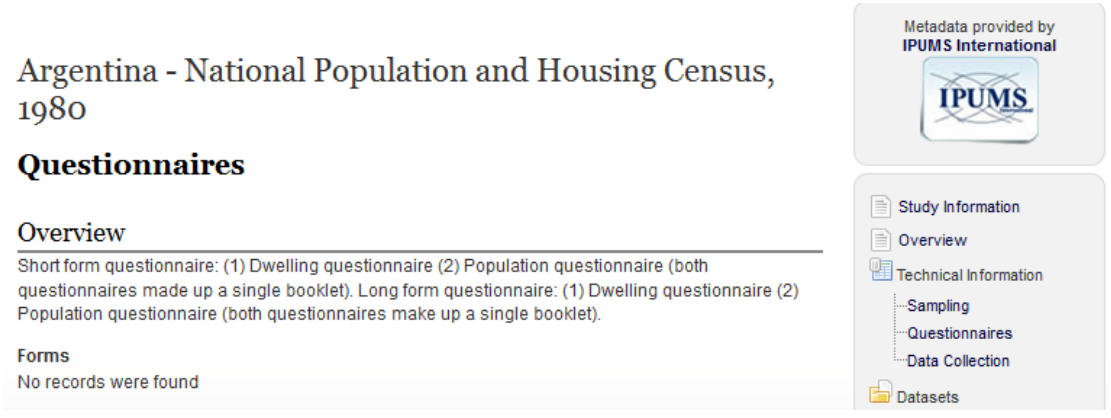


Fig. 5 Questionnaires

The following example illustrates three questions. Each question does have a text.

EXAMPLE 5

```
<#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"@en.

<#QuestionAge> a disco:Question;
```

```
        disco:questionText "3. What is his or her age? ___ Mark the age in completed
                    years at the date of the census for those younger than one year old mark
                    00. For those younger than 10 years old, mark 01, 02, 03, etc. For those
                    older than 99 years old, mark 99."@en.

    <#QuestionCitizenship> a disco:Question;
        disco:questionText "6. [Immigration status] Only for persons who have usual
                    residence in Argentina and were born in another country. [Questions 6A
                    and 6B asked only of persons born outside Argentina and who currently
                    reside in Argentina.] B. Are you a naturalized citizen of Argentina?
                    [] Yes [] No [] Unanswered"@en.
```

In Figure 6 we see the list of variables contained in the data file. For each of these we will also have a detailed view, showing the codes and categories used to encode the actual responses in the variables (Figure 7).

Variables

ID	NAME	LABEL	QUESTION
RECTYPE	RECTYPE	Record type	
CNTRY	CNTRY	Country	
YEAR	YEAR	Year	
SAMPLE	SAMPLE	IPUMS sample identifier	
SERIAL	SERIAL	Household serial number	
PERSONS	PERSONS	Number of person records in the household	
WTHH	WTHH	Household weight	
SUBSAMP	SUBSAMP	Subsample number	
GQ	GQ	Group quarters status	
UNREL	UNREL	Number of unrelated persons	
URBAN	URBAN	Urban-rural status	
REGIONW	REGIONW	Continent and region of country	

Fig. 6 Variables List

ROOF ROOF Roof material

Roof material(ROOF)

File: ARG1980-H-H.dat

Overview

Type: Discrete
Format: numeric
Width: 2
Decimals: 0

DEFINITION

This variable indicates the dwelling's predominant roofing material.

Categories

Value	Category	Cases	
00	NIU (not in universe)	24383	3.6%
10	Masonry, concrete, clay tile, or tiles of unspecified type	0	0.0%
11	Concrete or cement	0	0.0%
12	Reinforced concrete or brick	0	0.0%
13	Cement or sheet metal	0	0.0%
14	Tile, unspecified	34885	5.2%
15	Clay tile	0	0.0%
16	Tile or cement	0	0.0%
17	Modern tiles, industrial	0	0.0%
18	Traditional tiles, locally made	0	0.0%
19	Tile or flat stone	0	0.0%
20	Fibercement or plastic	41567	6.2%

Fig. 7 Variable Details

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

EXAMPLE 6

```
<#AR80A401> a disco:Variable;
    dcterms:identifier "AR80A401";
    skos:prefLabel "Sex"@en, "Sexe"@fr;
    dcterms:description "This variable indicates the person's gender."@en;
    disco:basedOn <#SexVD>;
    disco:question <#QuestionGender>.
```

```
<#AR80A402> a disco:Variable;
    dcterms:identifier "AR80A402";
    dcterms:description "This variable indicates the person's age in years."@en;
    skos:prefLabel "Age"@en, "Âge"@fr.
    disco:basedOn <#AgeVD>;
    disco:question <#QuestionAge>.

<#AR80A407> a disco:Variable;
    dcterms:identifier "AR80A407";
    dcterms:description "This variable indicates whether or not the person is
                a naturalized citizen of Argentina."@en;
    skos:prefLabel "Citizenship"@en, "Citoyenneté"@fr;
    disco:basedOn <#CitizenshipVD>;
    disco:question <#QuestionCitizenship>.
```

Any variable definition has a representation defining the possible values of a variable. Also, a variable definition has its own universe (possibly the same as the study, possibly a narrower one) and (DDI) concepts further describing the variable.

EXAMPLE 7

```
<#SexVD> a disco:VariableDefinition;
    disco:universe <#UniversePerson>;
    disco:representation <#SexRepr>;
    disco:concept <#IpumsC1>;
    skos:prefLabel "Sex"@en, "Sexe"@fr;
    dcterms:description "Sex data element"@en.

<#SexRepr> a skos:ConceptScheme, disco:Representation;
    skos:hasTopConcept <#SexM>, <#SexF>.

<#SexM> a skos:Concept;
    skos:notation "1";
    skos:prefLabel "Male"@en, "Homme"@fr;
    skos:inScheme <#SexRepr>.

<#SexF> a skos:Concept;
    skos:notation "2";
    skos:prefLabel "Female"@en, "Femme"@fr;
    skos:inScheme <#SexRepr>.

<#ageVD> a disco:VariableDefinition;
    disco:universe <#UniversePerson>;
    disco:representation <#AgeRepr>;
    disco:concept <#IpumsC1>;
    skos:prefLabel "Age"@en, "Âge"@fr;
    dcterms:description "Age data element"@en.

<#AgeRepr> a skos:ConceptScheme, disco:Representation;
    skos:hasTopConcept <#Age0>, <#Age1>, <#Age99>.

<#Age0> a skos:Concept;
    skos:notation "0";
    skos:prefLabel "0";
    skos:inScheme <#AgeRepr>.

<#Age1> a skos:Concept;
    skos:notation "1";
    skos:prefLabel "1";
    skos:inScheme <#AgeRepr>.

# ...

<#Age99> a skos:Concept;
    skos:notation "99";
    skos:prefLabel "99";
    skos:inScheme <#AgeRepr>.

<#CitizenshipVD> a disco:VariableDefinition;
    disco:universe <#UniverseNonArgentines>;
    disco:representation <#CitizenshipRepr>;
    disco:concept <#IpumsC2>;
    skos:prefLabel "Citizenship"@en;
    dcterms:description "Citizenship data element"@en.

<#CitizenshipRepr> a skos:ConceptScheme, disco:Representation;
    skos:hasTopConcept <#CYes>, <#CNo>, <#CUnknown>, <#CNIU>.

<#CYes> a skos:Concept;
    skos:notation "1";
    skos:prefLabel "Yes";
    skos:inScheme <#CitizenshipRepr>.

<#CNo> a skos:Concept;
    skos:notation "2";
    skos:prefLabel "No";
    skos:inScheme <#CitizenshipRepr>.

<#CUnknown> a skos:Concept;
    skos:notation "8";
    skos:prefLabel "Unknown";
    skos:inScheme <#CitizenshipRepr>.

<#CNIU> a skos:Concept;
    skos:notation "9";
    skos:prefLabel "NIU (not in universe)";
    skos:inScheme <#CitizenshipRepr>.
```

Any universe of a variable definition is a subset of the universe of the entire study. In our example, two questions are addressing the universe of persons, the third question is addressing a specific subset of the universe of persons.

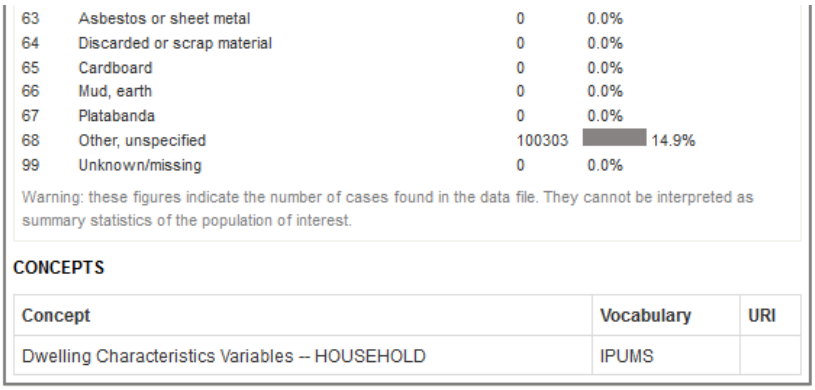
EXAMPLE 8

```
<#UniversePerson> a disco:Universe;
    skos:definition "All persons."@en ;
    skos:narrower <#Universe>.

<#UniverseNonArgentines> a disco:Universe;
    skos:definition "Foreign-born persons who reside in Argentina."@en ;
    skos:narrower <#Universe>;
```

```
        skos:narrower <#UniversePerson>.
```

At the bottom of the screen showing the variable detail, we can see that the variable for roofing material is associated with a high-level concept, “Dwelling characteristics variables.” (Figure 8.)



63	Asbestos or sheet metal	0	0.0%
64	Discarded or scrap material	0	0.0%
65	Cardboard	0	0.0%
66	Mud, earth	0	0.0%
67	Platabanda	0	0.0%
68	Other, unspecified	100303	14.9%
99	Unknown/missing	0	0.0%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

CONCEPTS

Concept	Vocabulary	URI
Dwelling Characteristics Variables -- HOUSEHOLD	IPUMS	

Fig. 8 Concept-Variable Link

Also in Disco, DDI concepts can be hierarchically structured

EXAMPLE 9

```
<#IpumsCS> a skos:ConceptScheme;
    skos:hasTopConcept <#IpumsC1>.

<#IpumsC1> a skos:Concept;
    skos:prefLabel "Demographic Variables - PERSON"@en, "Variables démographiques - PERSONNE"@fr;
    skos:inScheme <#IpumsCS>.

<#IpumsC2> a skos:Concept;
    skos:prefLabel "Nativity and Birthplace Variables -- PERSON"@en;
    skos:inScheme <#IpumsCS>.
```

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

EXAMPLE 10

```
<#Dstat1> a disco:DescriptiveStatistic;
    disco:frequency 13314444;
    # is that correct?
    disco:percentage 49.97;
    disco:statisticsVariable <#AR80A401>;
    disco:statisticsCategory <#SexM>;
    disco:statisticsDatafile <#Datafile>.

<#Dstat2> a disco:DescriptiveStatistic;
    disco:frequency 1336270;
    disco:statisticsVariable <#AR80A401>;
    disco:statisticsCategory <#SexF>;
    disco:statisticsDatafile <#Datafile>.
```

Next we find some general information about the data files produced by this study (Figure 9).

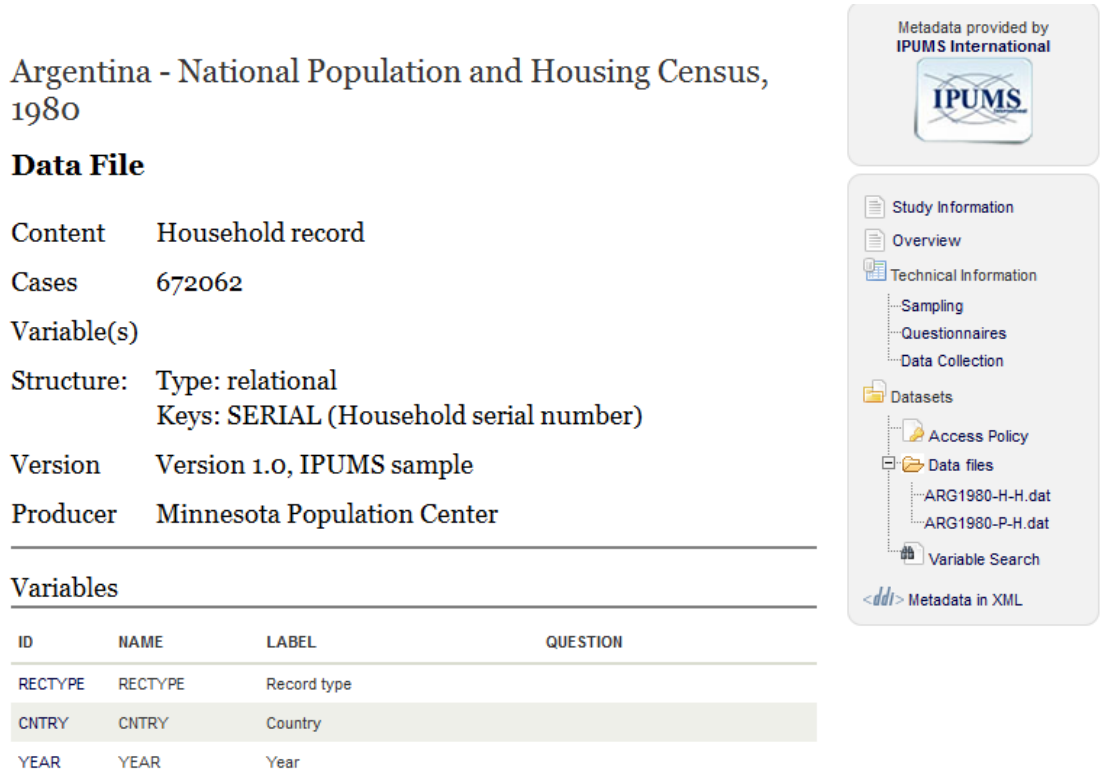


Fig. 9 General Data Set Information

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

EXAMPLE 11

```
<#Datafile> a disco:Datafile;
    dcterms:identifier "ARG1900-P-H.dat";
    dcterms:description "Person records"@en;
    disco:caseQuantity 2667714;
    dcterms:format "ascii";
    dcterms:provenance "Minnesota Population Center"@en;
    owl:versionInfo "Version 1.0, IPUMS sample"@en;
    dcterms:spatial [
        # This is the DC-strictly compatible way to do it
        a dcterms:Location;
        rdfs:label "Argentina, national coverage"@en
    ];
    dcterms:temporal "PeriodOfTime"@en;
    dcterms:subject "To be defined"@en.
```

4. 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. The key criteria for a study are: a single conceptual model (e.g. survey research concept), a single instrument (e.g. questionnaire) made up of one or more parts (ex. employer survey, worker survey), and a single logical data structure of the initial raw data (multiple data files can be created from this such as a public use microdata file or aggregate data files). 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" (to use the DDI term). In this case, a number of Study objects would be collected into a single StudyGroup.

Studies (Study) may be contained in at most 1 StudyGroup and groups of studies may include 0 to n studies. Studies (Study) may have 0 to n instruments (Instrument) relationships to instruments (Instrument). Particular instruments (Instrument), however, are connected with exactly 1 Study. Studies (Study) may have DataFile connections with 0 to n data files (DataFile) and data files (DataFile) must have 1 to n DataFile relationships to studies (Study). Studies (Study) are associated with 0 to n variables (Variable) using the object property Variable. On the other hand, variables (Variable) must be related to 1 to n studies (Study). Studies (Study) may have 0 to n logical data sets (LogicalDataSet) (product) and logical data sets (LogicalDataSet) must have 1 to n product relationships to studies (Study).

4.1 Coverage, References to DDI-XML Files, and Kind of Data

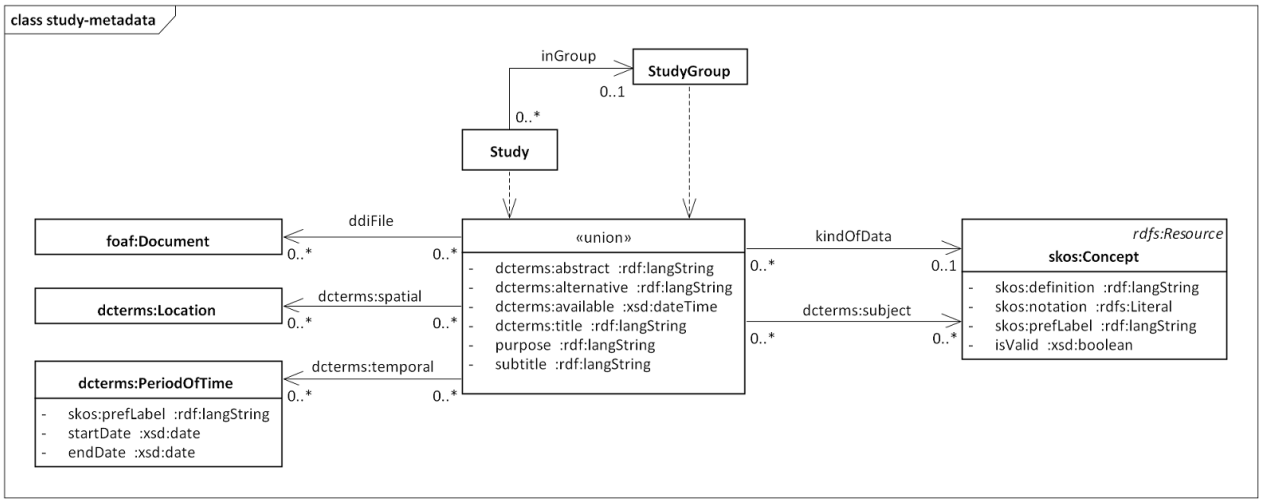


Fig. 10 Coverage, References to DDI-XML Files, and Kind of Data

Studies (`Study`) or groups of studies (`StudyGroup`) (the **union of `Study` and groups of studies (`StudyGroup`)**) may have different datatype properties. Studies (`Study`) or groups of studies (`StudyGroup`) may have an abstract (`dcterms:abstract`), a title (`dcterms:title`), a subtitle (`subtitle`), an alternative title (`dcterms:alternative`), a purpose (`purpose`), and information about the date and the time since when the `Study` is publicly available (`dcterms:available`). Studies (`Study`) or groups of studies (`StudyGroup`) may have multiple object properties. The object properties `kindOfData` and `dcterms:subject` guide to `skos:Concepts`. `kindOfData` describes, with a string or a term from a controlled vocabulary, the kind of data documented in the logical product(s) of a `Study`. Examples include survey data, census/enumeration data, administrative data, measurement data, assessment data, demographic data, voting data, etc. You can use `dcterms:subject` to describe the topical coverage of studies (`Study`) and groups of studies (`StudyGroup`). `ddiFile` to `foaf:Documents` which are the DDI-XML files containing further descriptions of the `Study` or the `StudyGroup`. Use `dcterms:temporal` for temporal coverages related to the union of studies (`Study`) and groups of studies (`StudyGroups`). For the spatial coverage use `dcterms:spatial`. The cardinalities of all the object properties are in both directions 0 to n. The only exception is that studies (`Study`) and groups of studies (`StudyGroup`) may have 0 or 1 `kindOfData` relationships to `skos:Concepts`.

4.2 Relationships to Agents

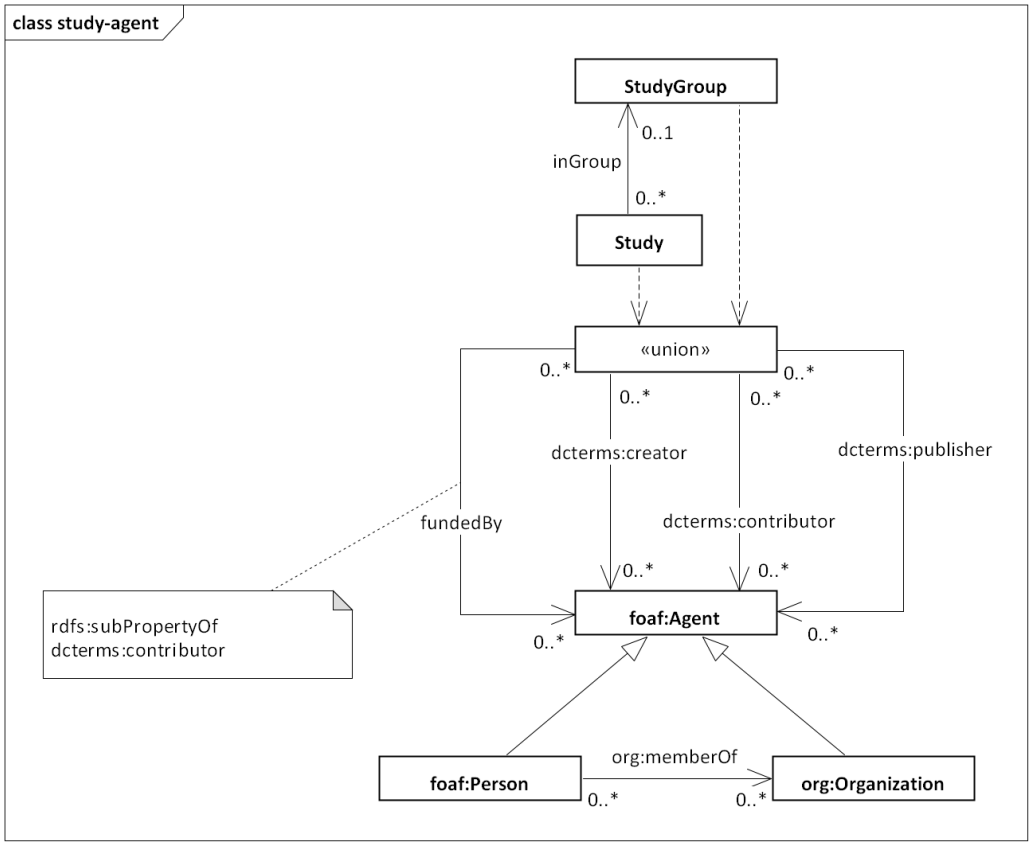


Fig. 11 Relationships to Agents

Creators (`dcterms:creator`), contributors (`dcterms:contributor`), and publishers (`dcterms:publisher`) of Studies (`Study`) and groups of studies (`StudyGroup`) are `foaf:Agents` which are either `foaf:Persons` or `org:Organizations` whose members are `foaf:Persons`. Studies (`Study`) or groups of studies (`StudyGroup`) may be funded by (`fundedBy`) `foaf:Agents`. The object property `fundedBy` is defined as sub-property of `dcterms:contributor`. The cardinalities of these object properties are in both directions always 0 to n.

4.3 Analysis Units and Universes

Universe is the total membership or population of a defined class of people, objects or events. There are two types of population, target population and survey population. A target population is the population outlined in the survey objects about which information is to be sought. A survey population (also known as the coverage of the survey) is the population from which information can be obtained in the survey. **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.

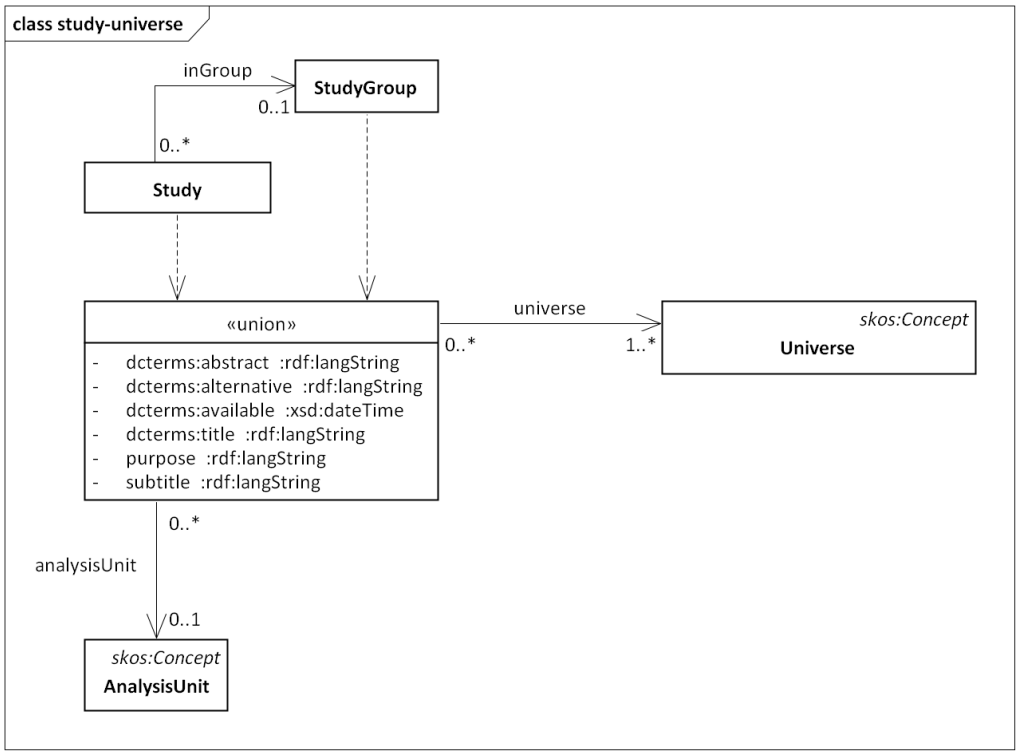


Fig. 12 Study, Universe and AnalysisUnit

Studies (Study) and groups of studies (StudyGroup) must have 1 to n universes (Universe) and 1 particular Universe may be in a Universe relationship with 0 to n unions of Studies (Study) and groups of studies (StudyGroup). Universes (Universe) are sub-classes of skos:Concepts. For universes (Universe) you can state definitions using skos:definition. The union of Study and StudyGroup may have 0 or 1 AnalysisUnit reached by the object property AnalysisUnit and a specific AnalysisUnit may be in a AnalysisUnit relationship to 0 to n studies (Study) or groups of studies (StudyGroup). AnalysisUnit is specified as a sub-class of skos:Concepts.

5. General Metadata

5.1 Identification

In DDI, a lot of entities hold particular identifiers. This can be identifiers for different versions of DDI, but also persistent identifiers for, e.g. persons or organizations, that are encoded in a particular identifier scheme, e.g. ORCID or FundRef. In general, such identifiers can be added to each entitiy in DDI-RDF, since every entity is defined as an rdfs:Resource. General metadata elements which can be used on every resource in a DDI-RDF description include:

- skos:prefLabel (rdf:langString): the preferred label of this element
- adms:identifier (rdfs:Resource, adms:Identifier): the identifier of this element

Each Disco resource must have an identifier (see figure below). The identifier is stated using the object property adms:identifier pointing from any rdfs:Ressource to 1 to n idetifiers (adms:Identifier). The class adms:Identifier can include the actual identifier itself and information on identifier scheme, its version, and its agency.

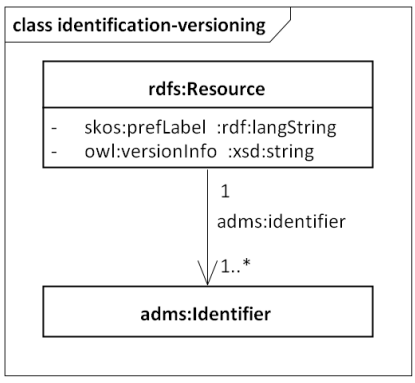


Fig. 13 Identification

EXAMPLE 12

```
Example code for the usage of adms:identifier and adms:Identifier
```

See section '[Asset Description Metadata Schema (ADMS)]' for more information.

5.2 Versioning Information

Use of the `owl:versionInfo` property is recommended to indicate the version number and/or additional versioning text of entities.

Any entity can have version information. As you can see in the next UML class diagram, the property owl:versionInfo has rdfs:Resource as domain. As a consequence, each DDI object can have attached versioning information. However, the most typical cases are:

- Version of the metadata (e.g., DDI file or RDF file), where the subject is the URL of the file
- Version of the study (e.g., as a study goes through the life cycle from conception through data collection, etc.), where the subject is a `Study`
- Version of the data files, where the subject is a `DataFile`.

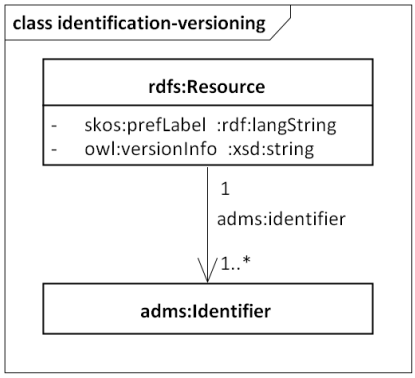


Fig. 14 Versioning Information

5.3 Relations to DDI-XML Files

Since the Discovery Vocabulary only covers a subset of an original DDI-XML file, it may be worthwile to have a relationship to the original DDI-XML file. Such a relationship can be represented using `dcterms:relation`. This way, every element can be related to any `foaf:Document`. The cardinalities are in both directions 0 to n.

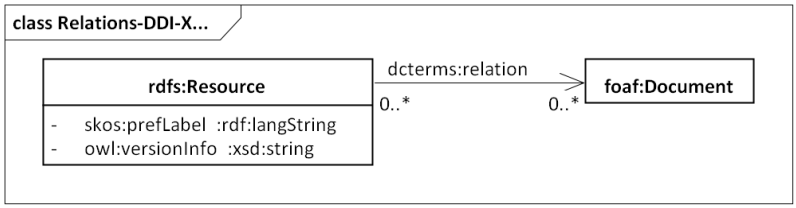


Fig. 15 Relations to DDI-XML Files

5.4 Access Rights Statements and Licenses

Every logical dataset may have access rights statements and licensing information attached to it. For those purposes, the Dublin Core properties `dcterms:accessRights` and `dcterms:license` are used.

Access rights are defined in a **`dcterms:RightsStatement`** object, which may reference an external document stating the access rights in more detail (`rdfs:seeAlso`). For `dcterms:RightsStatements` descriptions (`dcterms:description`) and labels (`skos:prefLabel`) can be assigned:

EXAMPLE 13

```
ex:Dataset1 a disco:LogicalDataset ;
  dcterms:accessRights ex:AccessRights1 .
ex:AccessRights1 dcterms:description "Everybody may see access this document." ;
  rdfs:seeAlso <http://www.example.org/access.html> .
```

License information is captured in a **`dcterms:LicenseDocument`**, which is a subtype of `dcterms:RightsStatements`:

EXAMPLE 14

```
ex:Dataset1 a disco:LogicalDataset ;
  dcterms:license ex:License1 .
ex:License1 dcterms:description "Published under Open Content License." ;
  skos:prefLabel "OCL 1.0" ;
  rdfs:seeAlso <http://opencontent.org/opl.shtml> .
```

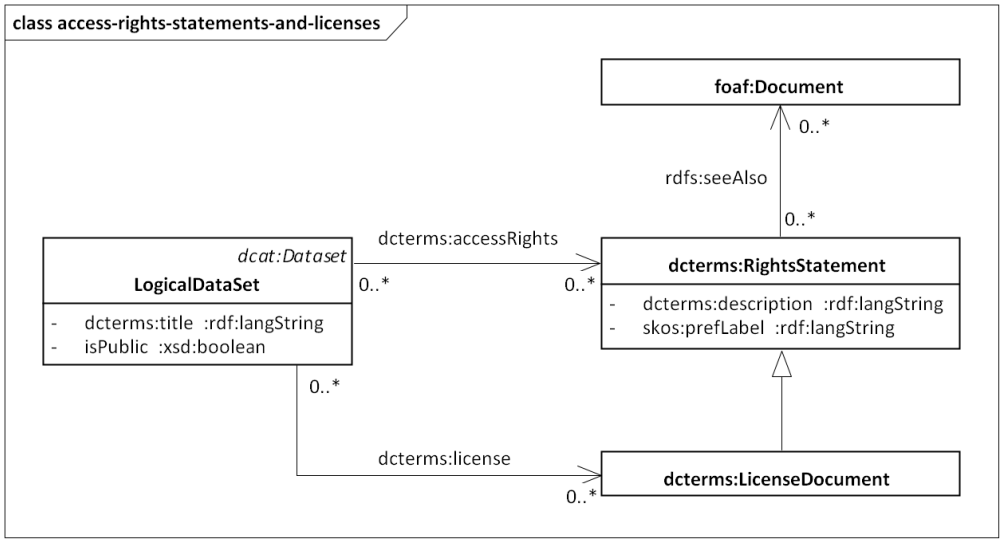


Fig. 16 Access Rights Statements and Licenses

Logical data sets (`LogicalDataSet`) may have `dcterms:accessRights` relationships to `dcterms:RightsStatements` and `dcterms:license` connections with `dcterms:LicenseDocument`. `dcterms:RightsStatements` is associated with `foaf:Documents` using the object property `rdfs:seeAlso`. The multiplicities for these object properties are in any case 0 to n.

5.5 Coverage of Studies, Logical Datasets, and Data Files

Coverage comprehends the key features of the scope of the data (e.g. geographic product occupation). Studies (`Study`), logical datasets, and data files may have a spatial, temporal, and topical coverage. Unlike in DDI-XML, there is no dedicated Coverage type in DDI-RDF. In contrast, spatial, temporal, and topical coverage are directly attached to the respective study, logical dataset, and datafile.

For spatial coverage, **`dcterms:spatial`** is used, pointing to any geographic location (`dcterms:Location`):

EXAMPLE 15

```
ex:Study1 dcterms:spatial <http://sws.geonames.org/2921044/> .
```

In this example, Geonames is used to refer to a spatial region, in this case, the country Germany. Geonames provides URIs for continents, countries, regions, and cities, among others, and is therefore a possible option to use for describing spatial coverage.

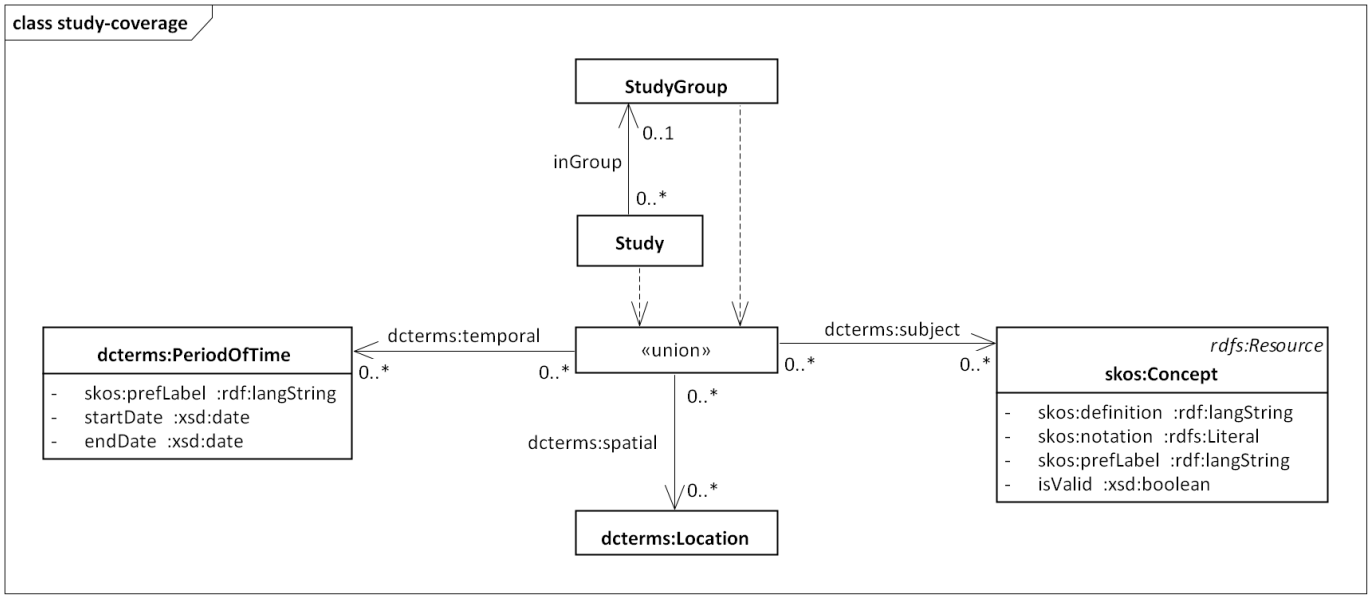


Fig. 17 Study Coverage

For temporal coverage, **`dcterms:temporal`** is used pointing to `dcterms:PeriodOfTime`. For time periods, labels can be attachted (`skos:prefLabel`). It is also possible to define start (`startDate`) and end dates (`endDate`). A possible way to describe temporal coverage is the use of the W3C time ontology:

EXAMPLE 16

```
ex:Study1 dcterms:temporal [
  a time:Interval ;
  time:hasBeginning [ time:inXSDDateTime
    "2012-01-01T00:00:00+01:00"^^xsd:dateTime ];
```

```
    time:hasEnd [ time:inXSDDateTime
      "2012-01-31T23:59:59+01:00"^^xsd:dateTime ] ] .
```

This example describes a study that has been conducted between January 1st and January 31st.

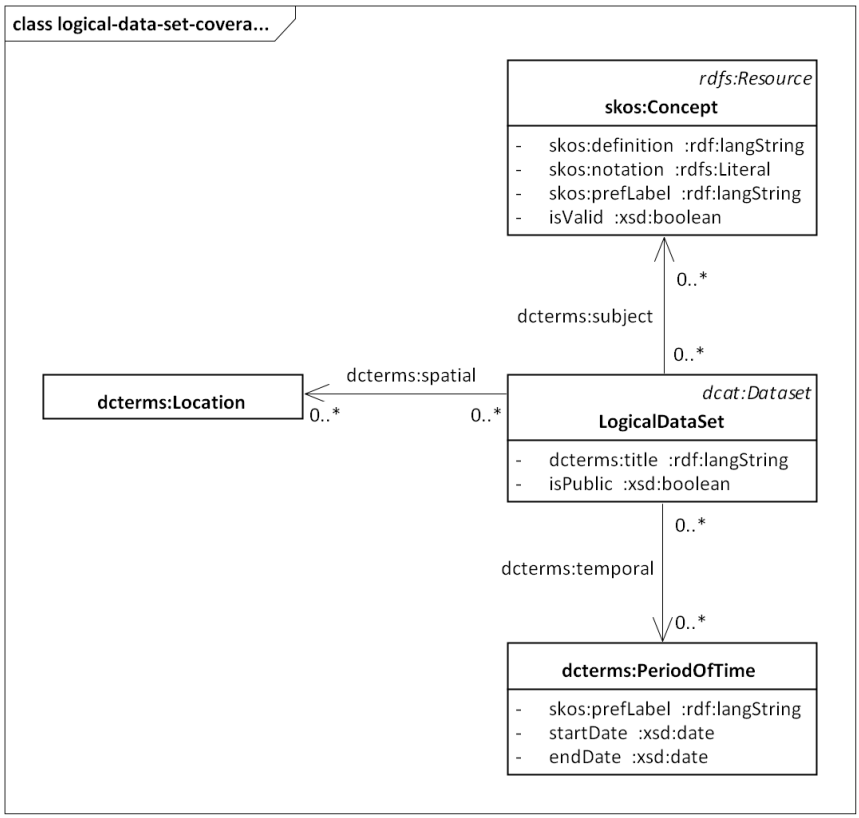


Fig. 18 LogicalDataSet Coverage

Topical coverage can be expressed using **`dcterms:subject`**. DDI-RDF foresees the use `skos:Concept` for the description of topical coverage:

EXAMPLE 17

```
ex:Study1 dcterms:subject [
  a skos:Concept ;
  skos:prefLabel "Alcohol consumption" ] .
```

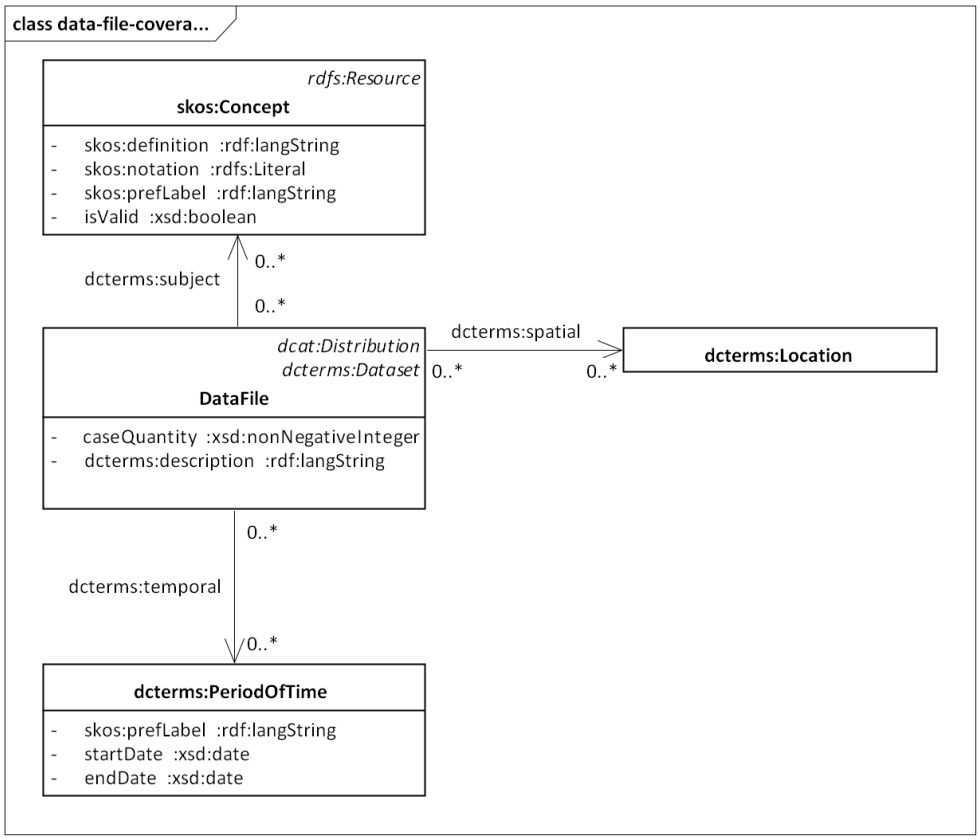


Fig. 19 DataFile Coverage

The multiplicities for each of the three object properties `dcterms:subject`, `dcterms:temporal`, and `dcterms:spatial` are in any case 0 to n.

5.6 Other General Dublin Core Metadata Properties

The following elements from Dublin Core may be used to describe general metadata of DDI-RDF elements (see the DC definitions for more detailed descriptions):

- `dcterms:abstract` (used with `Study`): an abstract of the study
- `dcterms:alternativet` (used with `Study`): an alternative name for the study
- `dcterms:available` (used with `Study`): the date (or date range) at which this study has or will become available
- `dcterms:title` (used with `Study`, `LogicalDataSet`): the element's title
- `dcterms:description` (used with `VariableDefinition`, `DataFile`, `Instrument`, `Variable`, `dcterms:RightsStatement`): a human readable description of the element
- `dcterms:provenance` (used with `DataFile`): defines the provenance information for the data file. The object is a `dcterms:ProvenanceStatement`.

6. Data Sets, Data Files, and Descriptive Statistics

Data sets have two representations in our model: 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. In our model the **`LogicalDataSet`** represents the content of the file (its organization into a set of variables (`Variable`)). The `LogicalDataSet` is an extension of the `dact:DataSet` class. Physical, distributed files are represented by the class **`DataFile`**, which is itself an extension of the `dcat:Distribution`. **`DescriptiveStatistics`** , i.e. `SummaryStatistics` as well as `CategoryStatistics`, are accociated with data files (`DataFile`) by the object property `statisticsDataFile`.

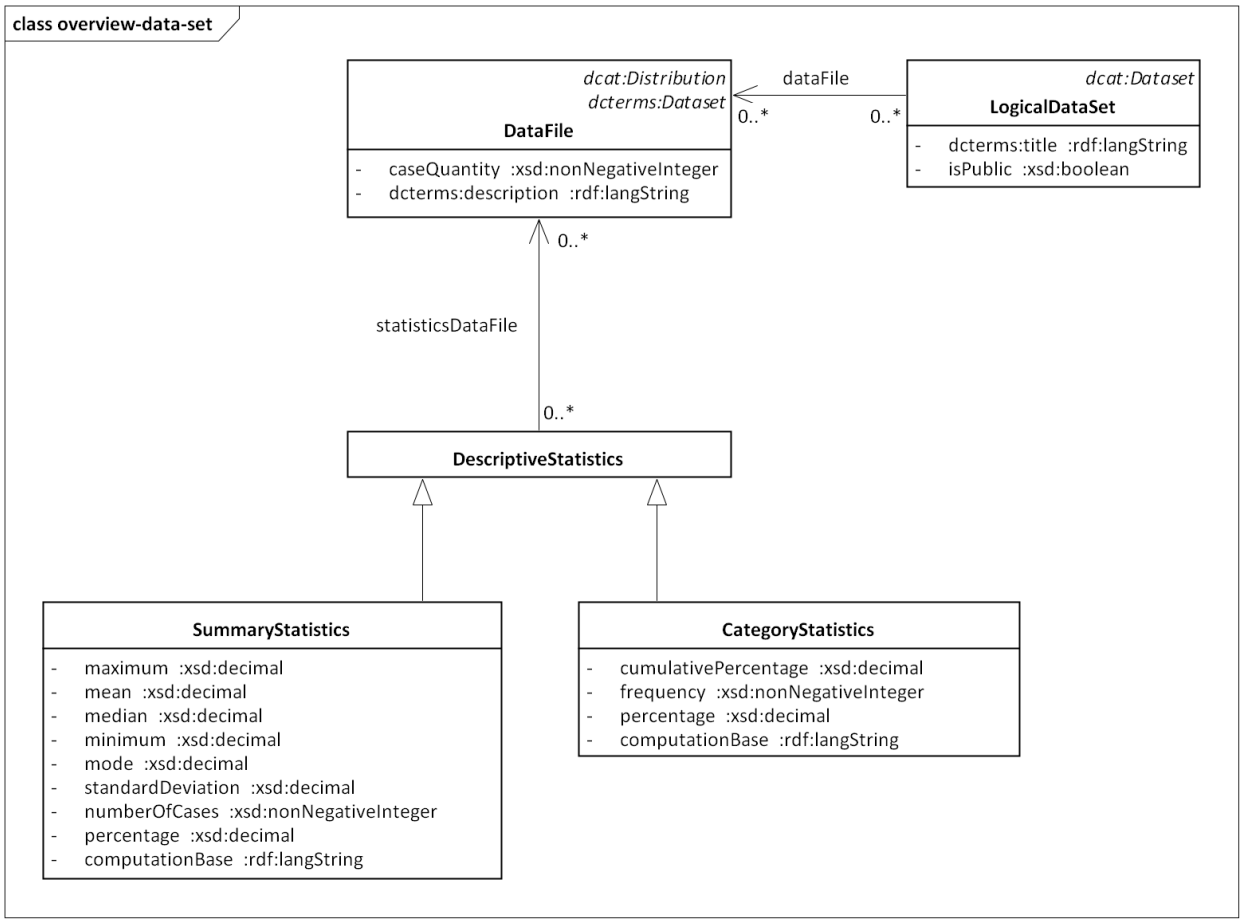


Fig. 20 Overview: Data Sets, Data Files, Descriptive Statistics

Logical data sets (`LogicalDataSet`) and data files (`DataFile`) are connected using the object property data files (`DataFile`). A specific logical data set (`LogicalDataSet`) may be linked to 0 to n data files (`DataFile`) and a particular `DataFile` may be connected with 0 to n logical data sets (`LogicalDataSet`) via `DataFile`. `DescriptiveStatistics` are accociated with data files (`DataFile`) by the object property `statisticsDataFile`. A concrete `DescriptiveStatistics` object may have `statisticsDataFile` relationships to multiple (0 - n) data files (`DataFile`). Data files (`DataFile`), however, may have 0 to n `statisticsDataFile` relations to `DescriptiveStatistics` instances.

6.1 LogicalDataSet

Each study has a set of logical metadata (**`LogicalDataSet`**) associated with the processing of data, at the time of collection or later during cleaning, and re-coding. `LogicalDataSet` represents the microdata dataset.

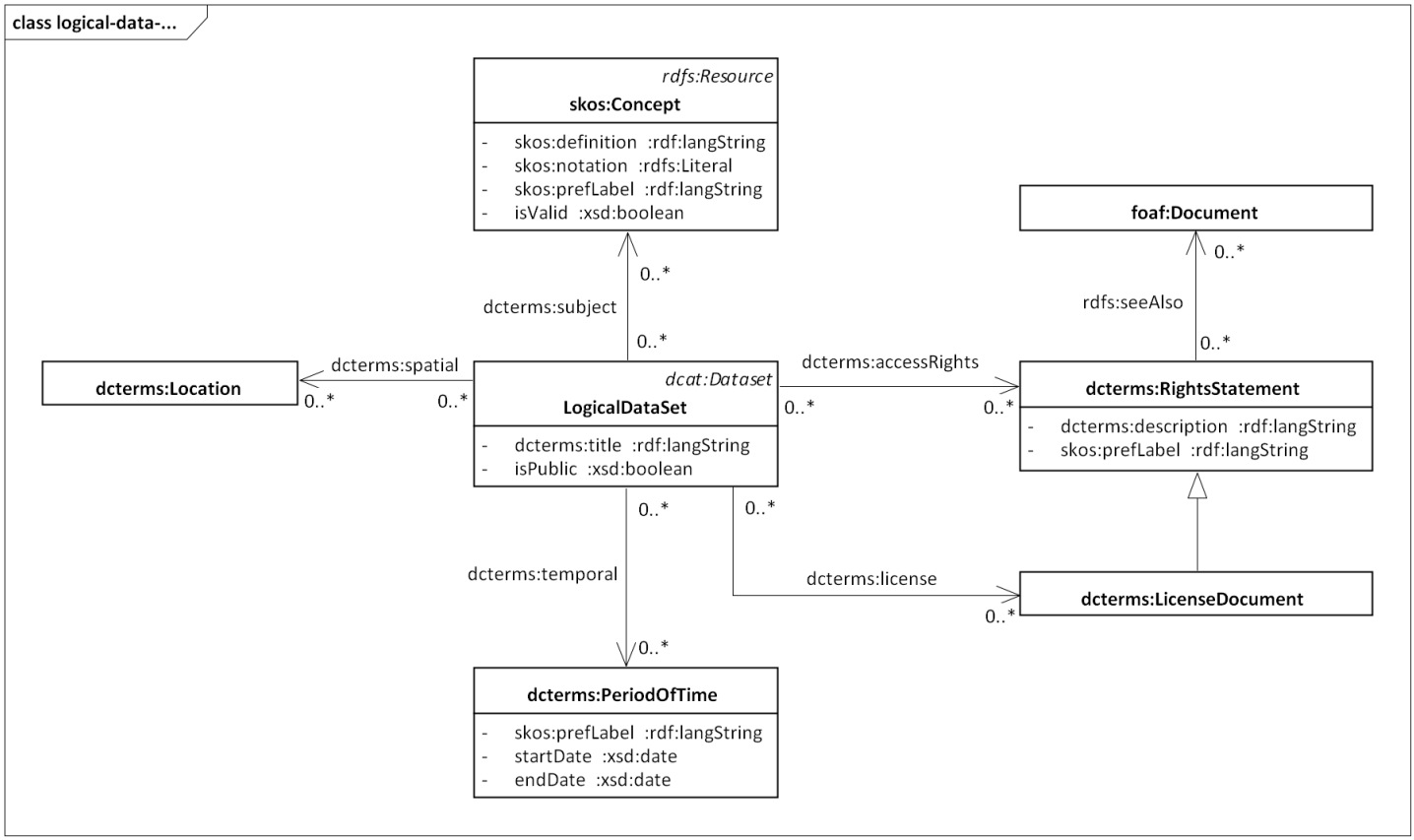


Fig. 21 LogicalDataSet

`LogicalDataSet` is defined as a sub-class of `dcat:Dataset`. You can state a title (`dcterms:title`) and a flag indicating if the microdata dataset is publicly available (isPublic). You can specify access rights (`dcterms:accessRights`) and LicenseStatements (`dcterms:license`) for microdata datasets. For a `LogicalDataSet` the three dimensions of coverage can be specified: Spatial (`dcterms:spatial`), temporal (`dcterms:temporal`), and topical (`dcterms:subject`). The cardinalities of the object properties `dcterms:spatial`, `dcterms:temporal`, `dcteerms:subject`, `dcterms:accessRights`, and `dcterms:license` are 0 to n. Microdata datasets may have `Instrument` associations to multiple (0 - n) instruments (`Instrument`) and instruments (`Instrument`) are connected with multiple (0 - n) logical data sets (`LogicalDataSet`). Each `LogicalDataSet` has exactly 1 `Universe` (`Universe`) and one specific `Universe` may be in multiple (0 - n) `Universe` relations to logical data sets (`LogicalDataSet`). Logical data sets (`LogicalDataSet`) may contain (`containsVariable`) 0 to n variables (`Variable`) and variables (`Variable`) must be contained in 1 to n logical data sets (`LogicalDataSet`). Logical data sets (`LogicalDataSet`) can be aggregated (`aggregation`) to 0 to n data sets (`qb:DataSet`) and data sets (`qb:DataSet`) can be aggregations of 0 to n logical data sets (`LogicalDataSet`). At last, logical data sets (`LogicalDataSet`) refer to 0 to n data files (`DataFile`) using the object property data files (`DataFile`) and data files (`DataFile`) may be linked to 0 to n logical data sets (`LogicalDataSet`). The class `qb:DataSet` is defined in the RDF Data Cube Vocabulary. 0 to n data sets (`qb:DataSet`) may point to multiple (0 - n) variables (`Variable`) (`inputVariable`).

EXAMPLE 18

```
<#Dataset> a LogicalDataSet;
    dcterms:accessRights <AccessRights>;
    disco:dataFile <#Datafile>;
    disco:instrument <#Questionnaire>;
    disco:containsVariable <#AR80A401>, <#AR80A402>, <#AR80A404>, <#AR80A407>, <#AR80A411>.
```

6.2 DataFile

The collected data result in the microdata represented by the **`DataFile`**. 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. data files (`DataFile`), which are also `dcterms:Datasets` as well as `dcat:Distributions`, represents all the physical distributed data files containing the microdata datasets.

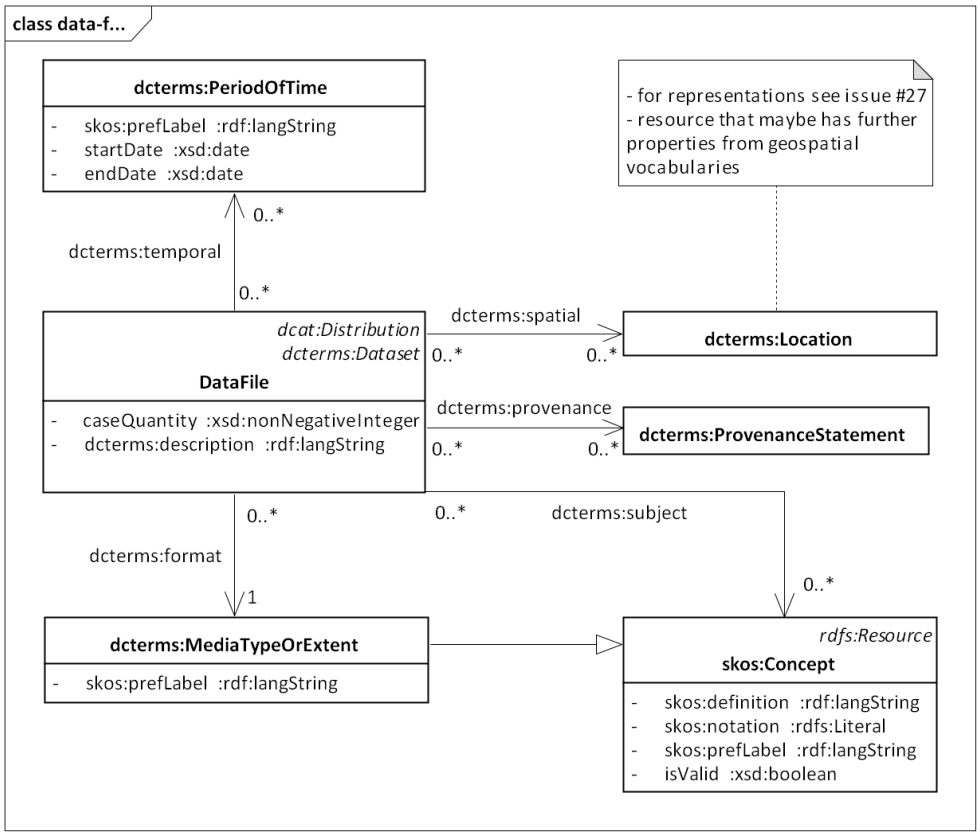


Fig. 22 DataFile

EXAMPLE 19

```
<#Datafile> a disco:Datafile;
  dcterms:identifier "ARG1900-P-H.dat";
  dcterms:description "Person records"@en;
  disco:caseQuantity 2667714;
  dcterms:format "ascii";
  dcterms:provenance "Minnesota Population Center"@en;
  owl:versionInfo "Version 1.0, IPUMS sample"@en;
  dcterms:spatial [
      # This is the DC-strictly compatible way to do it
      a dcterms:Location;
      rdfs:label "Argentina, national coverage"@en
  ];
  dcterms:temporal "PeriodOfTime"@en;
  dcterms:subject "To be defined"@en.
```

It is possible to describe data files (`DataFile`) (`dcterms:description`). Data files (`DataFile`), case quantities (`disco:caseQuantity`) and versions (`owl:versionInfo`) can also be stated. Using the object property `dcterms:format`, data files (`DataFile`) formats can be defined. Data files (`DataFile`) must have exactly 1 `dcterms:format` relationship to an instance of the class `dcterms:MediaTypeOrExtend` which is a sub-class of `skos:Concept`. Specific formats can be assigned to multiple (0 - n) data files (`DataFile`). Provenance information can be assigned to data files (`DataFile`). Data files (`DataFile`) may have multiple (0 - n) `dcterms:provenance` relationships to `dcterms:ProvenanceStatements`. `Dcterms:ProvenanceStatements`, however, may have 0 to n `dcterms:provenance` relations to data files (`DataFile`). The topical, spatial, and temporal coverage of data files (`DataFile`) is realized by the object properties `dcterms:subject`, `dcterms:spatial`, and `dcterms:temporal`, all with the cardinalities 0 to n on both sides.

6.3 DescriptiveStatistics

An overview over the microdata can be given either by the descriptive statistics or the aggregated data. **`DescriptiveStatistics`** may be minimal, maximal, mean values, and absolute and relative frequencies. `qb:DataSet` originates from the RDF Data Cube Vocabulary, an approach to map the SDMX information model to an ontology. A `qb:DataSet` represents aggregated data (also known as macrodata) such as multi-dimensional tables. Aggregated data are derived from microdata by statistics on groups, or aggregates such as counts, means, or frequencies. **`SummaryStatistics`** pointing to variables and **`CategoryStatistics`** pointing to categories and codes are both descriptive statistics.

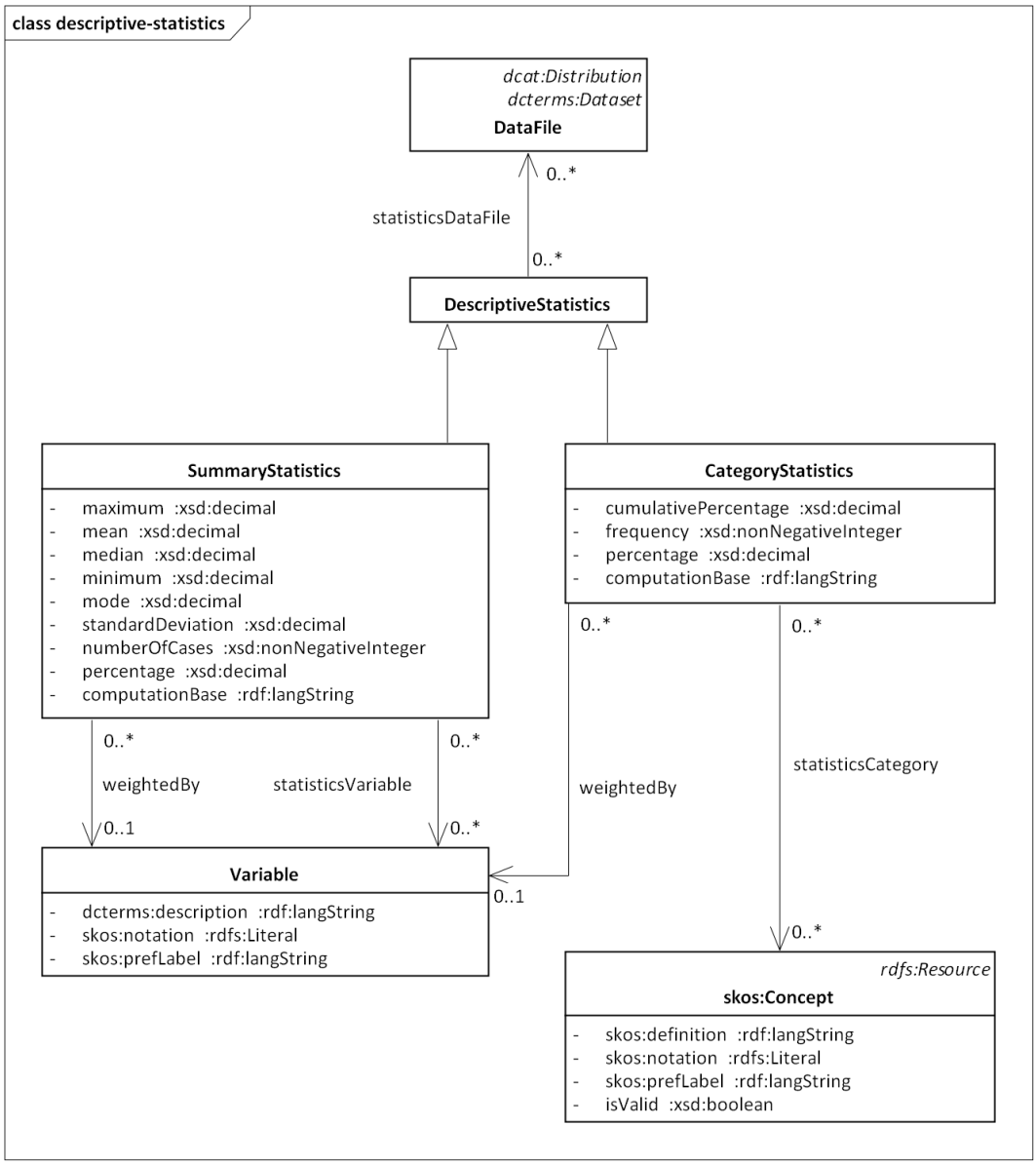


Fig. 23 DescriptiveStatistics

`DescriptiveStatistics` may have `statisticsDataFile` relations to 0 to n data files (`DataFile`) and data files (`DataFile`) may be in 0 to n `statisticsDataFile` relations to `DescriptiveStatistics` individuals. `SummaryStatistics` point to 0 to n variables (`Variable`) using the object property `statisticsVariable`. Variables (`Variable`), however, may be in 0 to n of such relationships to `SummaryStatistics` objects. `CategoryStatistics` may be connected with 0 to n `skos:Concepts` using the property `statisticsCategory` and `skos:Concepts` representing codes (values) and categories (value labels) may be in 0 to n of such relationships. `SummaryStatistics` and `CategoryStatistics` may have a `weightedBy` relation to a `Variable`.

EXAMPLE 20

```
<#Dstat1> a disco:DescriptiveStatistic;
  disco:frequency 13314444;
  # is that correct?
  disco:percentage 49.97;
  disco:statisticsVariable <#AR80A401>;
  disco:statisticsCategory <#SexM>;
  disco:statisticsDatafile <#Datafile>.

<#Dstat2> a disco:DescriptiveStatistic;
  disco:frequency 1336270;
  disco:statisticsVariable <#AR80A401>;
  disco:statisticsCategory <#SexF>;
  disco:statisticsDatafile <#Datafile>.
```

Available category statistics types are frequency, percentage, and cumulativePercentage. Available summary statistics types are frequency, percentage, maximum, mean, median, minimum, mode, and standardDeviation.

There are two properties which describe details of a category or summary statistic value, computationBase and weightedBy.

computationBase expresses if the cases - which are the basis of the computation of a statistics value - are valid, invalid or the total of both. The usage of computationBase for frequency differs from the usage for the percentage statistics and the summary statistics. A distinction regarding computationBase doesn't apply to frequency as category statistic. The following table describes the details of usage of computationBase in dependency of the respective statistics type.

Table 1: Description of Statistics of Valid/Invalid Cases

Statistics Type	computationBase			
	valid	invalid	total	not used
Category Statistics Type				
frequency	n/a	n/a	n/a	X
percentage	X	x	X	n/a
cumulativePercentage	X	x	X	n/a
Summary Statistics Type				
percentage	X	x	n/a	n/a
Any other summary statistics type	X	x	X	n/a

Legend: X – used frequently, x – not used frequently, n/a – not applicable

weightedBy defines the weight variable of a category or summary statistic computation respectively value. It can also be used to indicate if a weight variable is used but the related variable is not known. weightedBy may be assigned to a category statistic value or to a summary statistic value.

Table 2. Description of Statistics of Non-weighted/Weighted Variables

Statistics Value of ...	Value of weightedBy
unweighted variable	not used
weighted variable Weight variable is not known.	Reference to blank node
weighted variable Weight variable is known.	Reference to weight variable

ISSP 2011 (INTERNATIONAL SOCIAL SURVEY PROGRAMME)

The following example shows different categories of an ISSP data set and the values of the related summary and category statistics. Each category is defined as a skos:Concept and the used name is <issp:category_X>, which is the corresponding category value in the frequency table above (see Figure 23, second column).

The category <issp:category_1> is the category with the code 1 (skos:notation '1'), the category label ‘Yes, have partner; live in same household’ (skos:preflabel 'Yes, have partner; live in same household') and which is valid (disco:isValid true). <issp:XYZ_1> defines the frequency (disco:frequency '15893') of the category <issp:category_1> (disco:statisticsCategory <issp:category_1>).

PARTLIV Living in steady partnership					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes, have partner; live in same household	15893	60,6	63,7	63,7
	2 Yes, have partner; don't live in same household	1089	4,2	4,4	68,0
	3 No partner	7983	30,5	32,0	100,0
	Total	24965	95,2	100,0	
Missing	0 Not available (GB)	936	3,6		
	7 Refused	66	,3		
	9 No answer	249	,9		
	Total	1251	4,8		
Total		26216	100,0		

Fig. 24 Example Category Statistics: Frequency Table of Variable PARTLIV (ISSP 2011)

WRKHRS Hours worked weekly					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	,0	,0	,0
	2	11	,0	,1	,1
	3	15	,1	,1	,2
	...				
	36	197	,8	1,4	24,9
	37	332	1,3	2,3	27,2
	38	457	1,7	3,2	30,4
	39	203	,8	1,4	31,8
	40	3430	13,1	24,1	55,9
	41	63	,2	,4	56,4
	42	544	2,1	3,8	60,2
	43	172	,7	1,2	61,4
	...				
	90	19	,1	,1	99,1
	91	12	,0	,1	99,2
	92	2	,0	,0	99,2
	93	1	,0	,0	99,2
	94	1	,0	,0	99,2
	95	3	,0	,0	99,3
	96 96 hours and more	106	,4	,7	100,0
	Total	14237	54,3	100,0	
	Missing	0 NAP (Code 2 or 3 in WORK)	11033	42,1	
		98 Don't know; TW: Time varies	385	1,5	
		99 No answer	561	2,1	
		Total	11979	45,7	
	Total		26216	100,0	

Fig. 25 Example Category Statistics: Frequency Table of Variable WRKHRS (ISSP 2011)

Descriptive Statistics						
	N	Range	Minimum	Maximum	Mean	Std. Deviation
WRKHRS Hours worked weekly	14237	95	1	96	41,74	14,265
Valid N (listwise)	14237					

Fig. 26 Example Summary Statistics: Descriptive Statistics of Variable WRKHRS (ISSP 2011)

```
@prefix issp: <http://www.issp.org/>

<issp:Category_1>
  a skos:Concept;
  skos:notation '1';
  skos:preflabel 'Yes, have partner; live in same household';
  disco:isValid true.

<issp:Category_3>
  a skos:Concept;
  skos:preflabel 'valid total';
  disco:isValid true.

<issp:Category_2>
  a skos:Concept;
  skos:notation '0';
  skos:preflabel 'Not available (GB))';
  disco:isValid false.

<issp:Category_4>
  a skos:Concept;
  skos:preflabel 'missing total';
  disco:isValid false.

<issp:XYZ_1>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_1>;
  disco:frequency 15893.

<issp:XYZ_2>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_2>;
  disco:frequency 936.

<issp:XYZ_3>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_1>;
  disco:percentage 60.6.
  disco:computationBase 'total'.

<issp:XYZ_4>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_2>;
  disco:percentage 3.6;
  disco:computationBase 'total'.

<issp:XYZ_5>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_1>;
  disco:percentage 63.7;
  disco:computationBase 'validOnly'.
```

```
<issp:XYZ_6>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_1>;
  disco:cumulativePercentage 63.7;
  disco:computationBase 'validOnly'.

# optional: harmonized CategoryStatistics resource if computationBase and category is the same
<issp:XYZ_13>
  a disco:CategoryStatistics;
  disco:statisticsCategory <issp:Category_1>;
  disco:percentage 63.7;
  disco:cumulativePercentage 63.7;
  disco:computationBase 'validOnly'.

<issp:XYZ_7>
  a disco:SummaryStatistics;
  disco:statisticsVariable <issp:PARTLIV>;
  disco:numberOfcases 24965;
  disco:computationBase 'validOnly'.

<issp:XYZ_8>
  a disco:SummaryStatistics;
  disco:statisticsVariable <issp:PARTLIV>;
  disco:percentage 95.2;
  disco:computationBase 'total'.

<issp:XYZ_9>
  a disco:SummaryStatistics;
  disco:statisticsVariable <issp:PARTLIV>;
  disco:numberOfcases 1251;
  disco:computationBase 'missingOnly'.

<issp:XYZ_10>
  a disco:SummaryStatistics;
  disco:statisticsVariable <issp:PARTLIV>;
  disco:percentage 4.8;
  disco:computationBase 'missingOnly'.

<issp:XYZ_11>
  a disco:SummaryStatistics;
  disco:statisticsVariable <issp:PARTLIV>;
  disco:numberOfCases 26216;
  disco:computationBase 'total'.

<issp:XYZ_12>
  a disco:SummaryStatistics;
  disco:statisticsVariable <issp:WRKHRS>;
  disco:numberOfCases 14237;
  disco:minimum 1;
  disco:maximum 96;
  disco:mean 41.74;
  disco:standardDeviation 14.265;
  disco:computationBase 'validOnly'.
```

7. Variables, Variable Definitions, Representations, and Concepts

When it comes to understanding the contents of the data set, this is done using the Variable class. Variables (**Variable**) provide a definition of the column in a rectangular data file, and can associate it with a **Concept**, and a Question. Variables (Variable) 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. Variable definitions (**VariableDefinition**) encompasse study-independent, re-usable parts of variables like occupation classification.

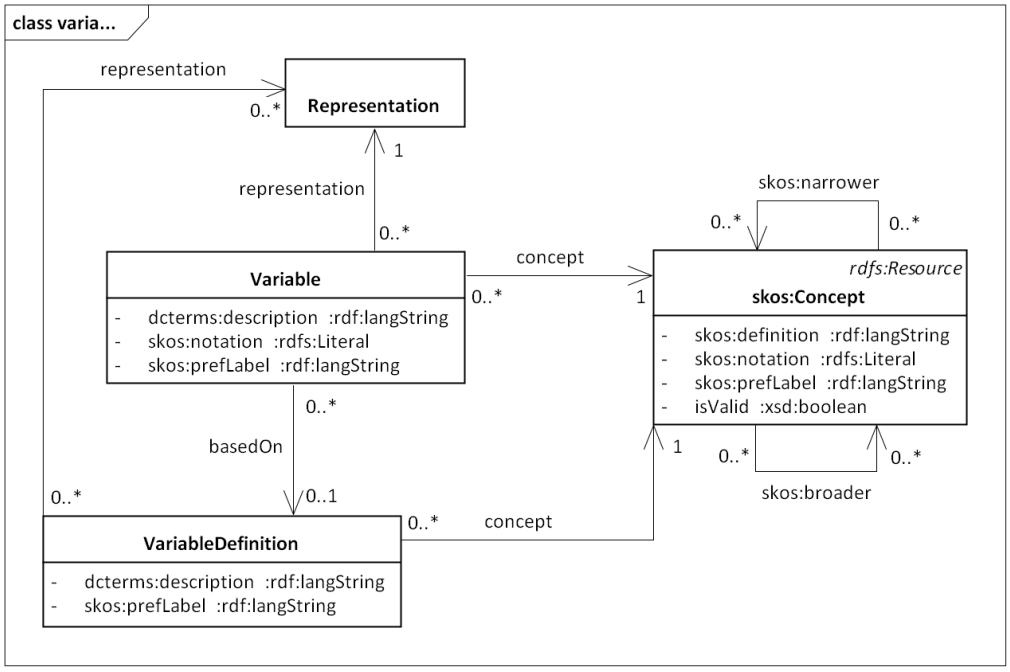


Fig. 27 Variables, Variable Definitions, Representations, and Concepts

Variables (Variable) may be based on (basedOn) 0 or 1 variable definitions (VariableDefinition) and variable definitions (VariableDefinition) can be in 0 to n basedOn relationships to variables (Variable). Both variables (Variable) and variable definitions (VariableDefinition) have Representation object properties with the class Representation as

range. Variables (Variable) must have exactly 1 Representation and variable definitions (VariableDefinition) may have 0 to n Representation connections to Representation. On the other hand, representations have 0 to n links to variable definitions (VariableDefinition) and to variables (Variable). Variables (Variable) as well as variable definitions (VariableDefinition) have both 1 connection to the concept which should be measured. Concepts have 0 to n relationships to variables (Variable) and variable definitions (VariableDefinition) using the object property concept.

7.1 Variable and Variable Definition

Variables provide a definition of the column in a rectangular data file. 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.
VariableDefinitions encompasse study-independent, re-usable parts of variables like occupation classification.

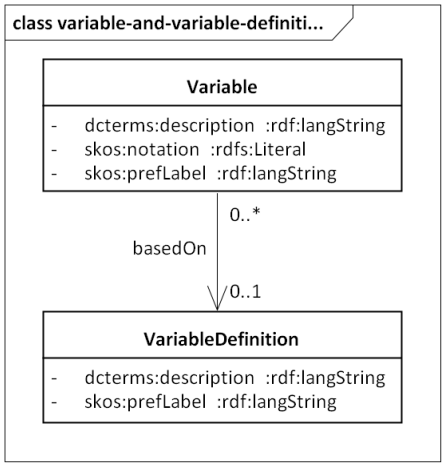


Fig. 28 Variables and VariableDefinitions

Variables (Variable) can be described (dcterms:description), skos:notation is used to associate names to variables and labels can be assigned to variables via the datatype property skos:prefLabel. Variable definitions (VariableDefinition) can also be described using dcterms:description. Labels can be assigned to variable definitions (VariableDefinition) via the datatype property skos:prefLabel. Variables (Variable) may be based on (BasedOn) 0 to 1 VariableDefinition. BasedOn also connects variable definitions (VariableDefinition) with 0 to n variables (Variable). Variables (Variable) and variable definitions (VariableDefinition) are connected with exactly 1 skos:Concept via Concept. skos:Concept have this connection to 0 to n variables (Variable) and variable definitions (VariableDefinition). Variables (Variable) are represented by 1 Representation and variable definitions (VariableDefinition) are represented by multiple (0 - n) representations (Representation). Representations (Representation) may be linked to 0 to n variables (Variable) and their definitions. Variables (Variable) may have (Question) 0 or more questions (Question) and questions (Question) may be associated with 0 to n variables (Variable). Universe is used to link 1 Universe to 0 to n variables (Variable) and 0 to n universes (Universe) to 0 to n variable definitions (VariableDefinition).

The following example illustrates the three variables Sex, Age and Citizenship.

EXAMPLE 21

```
<#AR80A401> a disco:Variable;
  dcterms:identifier "AR80A401";
  skos:prefLabel "Sex"@en, "Sexe"@fr;
  dcterms:description "This variable indicates the person's gender."@en;
  disco:basedOn <#SexVD>;
  disco:question <#QuestionGender>.

<#AR80A402> a disco:Variable;
  dcterms:identifier "AR80A402";
  dcterms:description "This variable indicates the person's age in years."@en;
  skos:prefLabel "Age"@en, "Âge"@fr.
  disco:basedOn <#AgeVD>;
  disco:question <#QuestionAge>.

<#AR80A407> a disco:Variable;
  dcterms:identifier "AR80A407";
  dcterms:description "This variable indicates whether or not the person
                       is a naturalized citizen of Argentina."@en;
  skos:prefLabel "Citizenship"@en, "Citoyenneté"@fr;
  disco:basedOn <#CitizenshipVD>;
  disco:question <#QuestionCitizenship>.
```

The three variables refer to universe, representations and concepts in their VariableDefinition.

EXAMPLE 22

```
<#SexVD> a disco:VariableDefinition;
  disco:universe <#UniversePerson>;
  disco:representation <#SexRepr>;
  disco:concept <#IpumsC1>;
  skos:prefLabel "Sex"@en, "Sexe"@fr;
  dcterms:description "Sex data element"@en.

            <#AgeVD> a disco:VariableDefinition;
  disco:universe <#UniversePerson>;
  disco:representation <#AgeRepr>;
  disco:concept <#IpumsC1>;
```

```
    skos:prefLabel "Age"@en, "Sexe"@fr;
    dcterms:description "Age data element"@en.

<#CitizenshipVD> a disco:VariableDefinition;
    disco:universe <#UniverseNonArgentines>;
    disco:representation <#CitizenshipRepr>;
    disco:concept <#IpumsC2>;
    skos:prefLabel "Citizenship"@en;
    dcterms:description "Citizenship data element"@en.
```

7.2 skos:Concept and skos:ConceptScheme

SKOS defines the term **`skos:Concept`**, which is a unit of knowledge created by a unique combination of characteristics. In context of statistical (meta)data, concepts are abstract summaries, general notions, knowledge of a whole set of behaviours, attitudes or characteristics which are seen as having something in common. Concepts may be associated with variables and questions. A **`skos:ConceptScheme`**, also defined within the SKOS namespace, is a set of metadata describing statistical concepts. `Skos:Concept` is reused to a large extent to represent DDI concepts, codes, and categories.

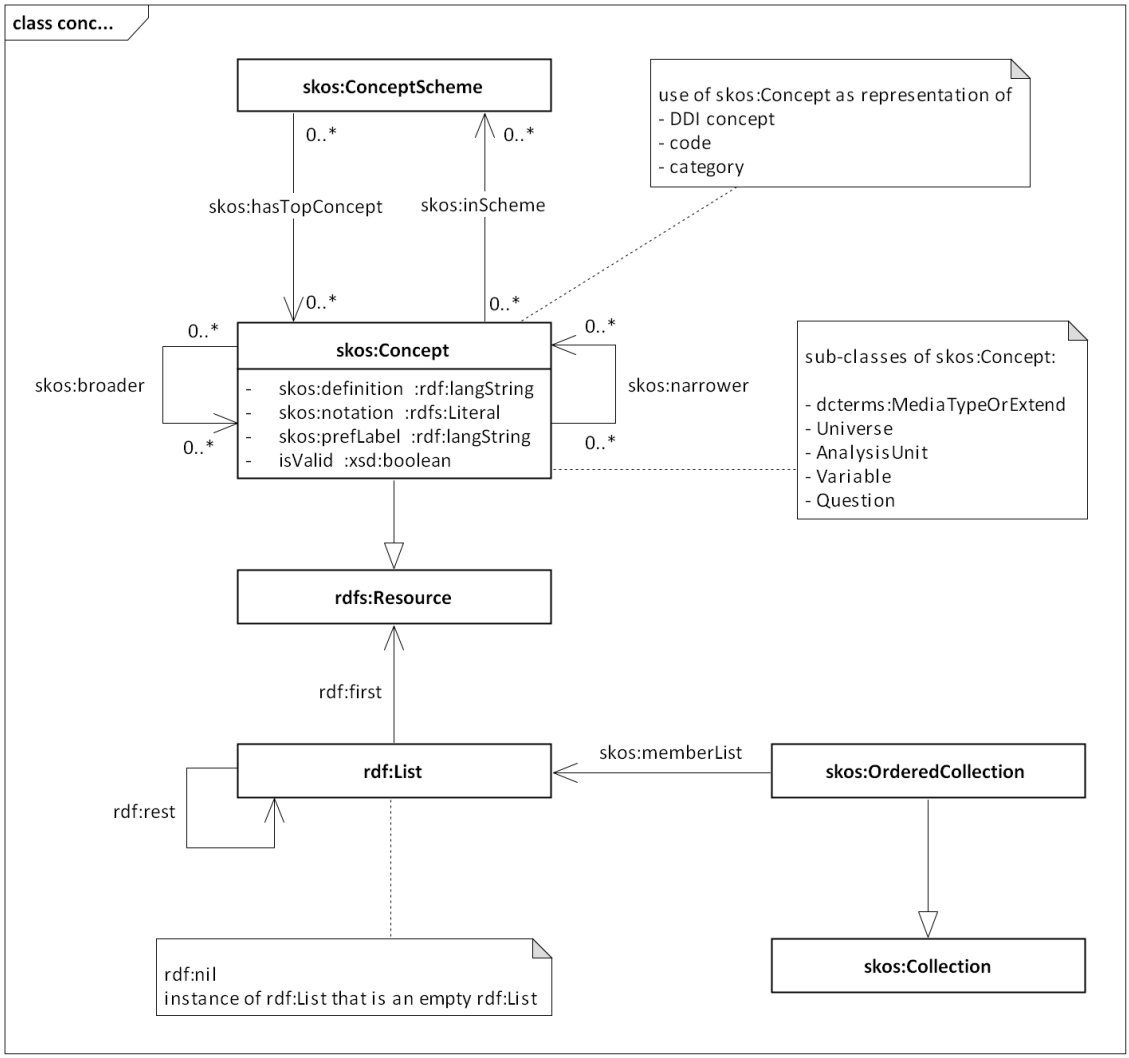


Fig. 29 skos:Concept and skos:ConceptScheme

DDI concepts can be described using `skos:definition`. Furthermore, you can describe code values (`skos:notation`) and category labels (`skos:prefLabel`). Hierarchies of DDI concepts can be built using the object properties `skos:broader` and `skos:narrower`. The domains and the ranges of `skos:broader` and `skos:narrower` are `skos:Concept`. The cardinalities are in both directions 0 to n. `Skos:Concept` may be organized in 0 to n `skos:ConceptSchemes` by means of `skos:inScheme`. `skos:ConceptSchemes` may have multiple (0 - n) `skos:Concept` as parts. The top concept in a specific ConceptScheme is indicated by `skos:hasTopConcept` pointing to 0 to n top `skos:Concept`. A specific `skos:Concept` may be the top concept to multiple (0 - n) `skos:ConceptSchemes`.

EXAMPLE 23

```
<#SexRepr> a skos:ConceptScheme, disco:Representation;
    skos:hasTopConcept <#SexM>, <#SexF>.

<#SexM> a skos:Concept;
    skos:notation "1";
    skos:prefLabel "Male"@en, "Homme"@fr;
    skos:inScheme <#SexRepr>.

<#SexF> a skos:Concept;
    skos:notation "2";
    skos:prefLabel "Female"@en, "Femme"@fr;
    skos:inScheme <#SexRepr>.
```

ISSP 2011 (INTERNATIONAL SOCIAL SURVEY PROGRAMME)

PARTLIV Living in steady partnership					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes, have partner; live in same household	15893	60,6	63,7	63,7
	2 Yes, have partner; don't live in same household	1089	4,2	4,4	68,0
	3 No partner	7983	30,5	32,0	100,0
	Total	24965	95,2	100,0	
Missing	0 Not available (GB)	936	3,6		
	7 Refused	66	,3		
	9 No answer	249	,9		
	Total	1251	4,8		
Total		26216	100,0		

Fig. 30 Example Category Statistics: Frequency Table of Variable PARTLIV (ISSP 2011)

```
@prefix issp: <http://www.issp.org/>

<issp:Category_1>
  a skos:Concept;
  skos:notation '1';
  skos:preflabel 'Yes, have partner; live in same household';
  disco:isValid true.

<issp:Category_2>
  a skos:Concept;
  skos:notation '2';
  skos:preflabel 'Yes, have partner; don't live in same household';
  disco:isValid true.

<issp:Category_3>
  a skos:Concept;
  skos:notation '3';
  skos:preflabel 'No partner';
  disco:isValid true.

<issp:Category_4>
  a skos:Concept;
  disco:isValid true.

<issp:Category_5>
  a skos:Concept;
  skos:notation '0';
  skos:preflabel 'Not available (GB))';
  disco:isValid false.

<issp:Category_6>
  a skos:Concept;
  skos:notation '7';
  skos:preflabel 'Refused';
  disco:isValid false.

<issp:Category_7>
  a skos:Concept;
  skos:notation '9';
  skos:preflabel 'No answer';
  disco:isValid false.

<issp:Category_8>
  a skos:Concept;
  disco:isValid false.
```

7.2.1 Uses of skos:Concept

In this sub-section, we describe all possible uses of the class `skos:Concept`.

- **Code values:** Code values are represented using the datatype property `skos:notation` with `skos:Concept` as domain.
- **Category labels:** Use `skos:prefLabel` and the domain class `skos:Concept` to describe category values
- **DDI concepts:** DDI concepts are described by the property `skos:definition` pointing from `skos:Concept` classes.
- **Hierarchies of DDI concepts:** Hierarchies of DDI concepts can be built using the object properties `skos:broader` and `skos:narrower`. The domains and the ranges of `skos:broader` and `skos:narrower` are `skos:Concept`.
- **Organization in `skos:ConceptSchemes`:** `Skos:Concepts` may be organized in `skos:ConceptSchemes` by means of `skos:inScheme`. The top concept in a specific ConceptScheme is indicated by `skos:hasTopConcept` pointing to top `skos:Concept`.
- **Topical coverage:** Topical coverage can be expressed using `dcterms:subject`. DDI-RDF 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.
- **Category linked to `CategoryStatistics`:** `CategoryStatistics` like frequencies and percentages are associated to the respectve Category using the object property statisticsCategory. `skos:Concept` represents categories.
- **Concepts of questions: Questions (`Question`)** are associated with concepts via the object property concept.
- **`Universe`:** Each universe is also a `skos:Concept`. Therefore the properties defined for `skos:Concept` can be reused for universes.
- **Collection Mode: Questionnaires (`Questionnaire`)** may have multiple collection modes which are represented by `skos:Concept`.
- **Concepts of variable definitions:** `Variable` definitions are associated with concepts via the object property concept.
- **Concepts of variables: Variables (`Variable`)** are linked to concepts via the object property concept.
- **Kind of data:** KindOfData describes, with a string or a term from a controlled vocabulary, the kind of data documented in the logical

product(s) of a Study. Examples include survey data, census/enumeration data, administrative data, measurement data, assessment data, demographic data, voting data, etc. The range of kindOfData is skos:Concept
- **Format of data files:** Using the object property dcterms:format, data files (DataFile) formats can be defined. Data files (DataFiles) must have exactly 1 dcterms:format relationship to an instance of the class dcterms:MediaTypeOrExtend which is a sub-class of skos:Concept.
- **AnalysisUnit:** Each analysis unit is also a skos:Concept. Therefore the properties defined for skos:Concept can be reused for analysis units.

7.2.2 Uses of skos:OrderedCollection and skos:ConceptScheme

In DDI, variables, questions, and categories are typically organized themselves in a particular order. For obtaining this order, skos:OrderedCollections are used. For example, a collection of variables is represented as being of the type skos:OrderedCollection containing multiple variables (each represented as skos:Concept) in a skos:memberList.

ISSP 2011 (INTERNATIONAL SOCIAL SURVEY PROGRAMME)

PARTLIV Living in steady partnership

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes, have partner; live in same household	15893	60,6	63,7	63,7
	2 Yes, have partner; don't live in same household	1089	4,2	4,4	68,0
	3 No partner	7983	30,5	32,0	100,0
	Total	24965	95,2	100,0	
Missing	0 Not available (GB)	936	3,6		
	7 Refused	66	,3		
	9 No answer	249	,9		
	Total	1251	4,8		
Total		26216	100,0		

Fig. 31 Example Category Statistics: Frequency Table of Variable PARTLIV (ISSP 2011)

The following example shows a ordered collection of categories in a abbreviated as well as a complete syntax.

```
@prefix issp: <http://www.issp.org/>

<issp:XYZ_1>
  a disco:Variable;
  skos:notation 'PARTLIV';
  skos:preflabel 'Living in steady partnership';
  disco:representation <issp:OrderedCollection1>.

# abbreviated syntax:
<issp:XYZ_2>
  rdf:type skos:OrderedCollection;
  skos:memberList (
    <issp:Category_1>
    <issp:Category_2>
    <issp:Category_3>
    <issp:Category_4>
    <issp:Category_5>
    <issp:Category_6>
    <issp:Category_7>
    <issp:Category_8> ).

# complete syntax:
<issp:XYZ_2>
  rdf:type skos:OrderedCollection;
  skos:memberList [
    rdf:first <issp:Category_1>; rdf:rest [
    rdf:first <issp:Category_2>; rdf:rest [
    rdf:first <issp:Category_3>; rdf:rest [
    rdf:first <issp:Category_4>; rdf:rest [
    rdf:first <issp:Category_5>; rdf:rest [
    rdf:first <issp:Category_6>; rdf:rest [
    rdf:first <issp:Category_7>; rdf:rest [
    rdf:first <issp:Category_8>;
    rdf:rest rdf:nil.] ] ] ] ] ] ] ].
```

If no order inside a collection of variables and questions is necessary, they are represented as unordered skos:ConceptSchemes. The classes Variable and Question are defined as sub-classes of skos:Concept.

7.3 Representation

The **Representation** of a variable is the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set. Representation is one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, and sex: male coded as 1). Questions (ResponseDomain), variables (Variable) (Representation), and variable definitions (VariableDefinition) (Representation) may have representations. Representation is defined as sub-class of the union of **rdfs:Datatype** (e.g. numeric or textual values), **skos:ConceptScheme**, and **skos:OrderedCollection**, as for example questions may have as response domain a mixture of a numeric response domain containing numeric values (rdfs:Datatype) and an unordered code response domain (skos:ConceptScheme) as well as an ordered code response domain (skos:OrderedCollection).

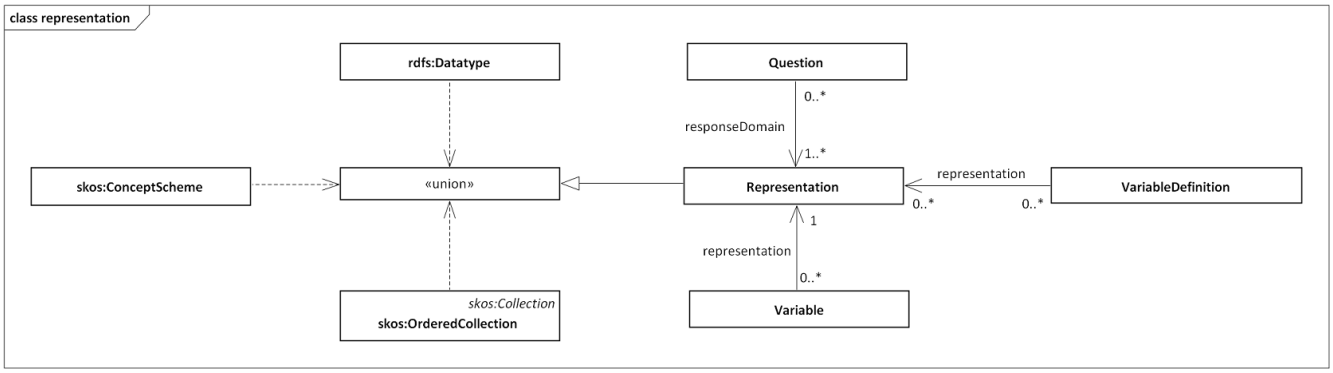


Fig. 32 Representation

Questions (Question) (responseDomain), variables (Variable) (representation), and variable definitions (VariableDefinition) (representation) may have representations. Questions (Question) must have 1 to n representations (representation), variables (Variable) must have exactly 1 Representation, and variable definitions (VariableDefinition) may have 0 to n representations (Representation). Each Representation can be in 0 to n Representation relationships with questions (Question), variables (Variable), and variable definitions (VariableDefinition).

The following example shows the representations of the three previously introduced variables Sex, Age and Citizenship. All of them refer to the particular concepts.

EXAMPLE 24

```
<#SexRepr> a skos:ConceptScheme, disco:Representation;
  skos:hasTopConcept <#SexM>, <#SexF>.

      <#AgeRepr> a skos:ConceptScheme, disco:Representation;
skos:hasTopConcept <#Age0>, <#Age1>, <#Age99>.

      <#CitizenshipRepr> a skos:ConceptScheme, disco:Representation;
skos:hasTopConcept <#CYes>, <#CNo>, <#CUnknown>, <#CNIU>.
```

8. Data Collection

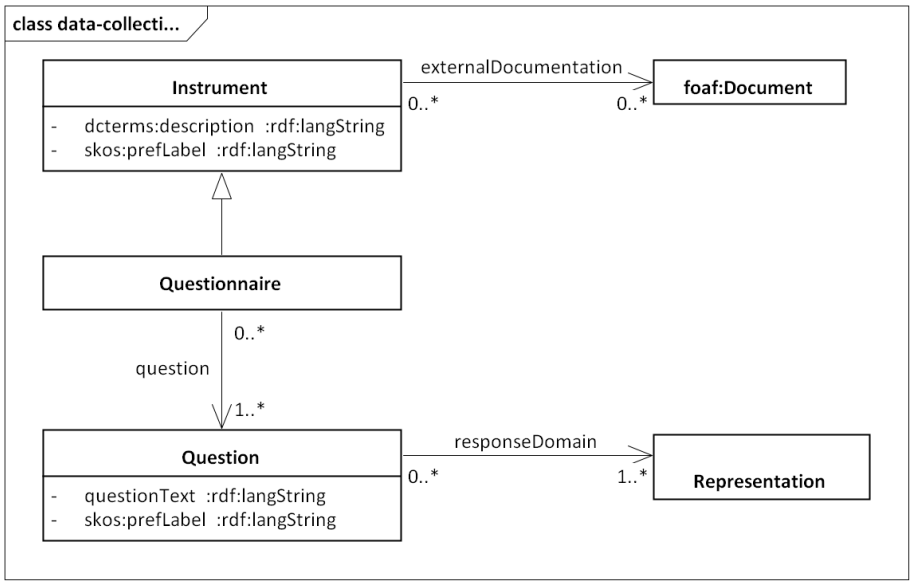


Fig. 33 DataCollection

8.1 Instrument

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.

You can describe (dcterms:description) instruments (Instrument) and associate labels (skos:prefLabel) to instruments (Instrument). Instruments (Instrument) may have (externalDocumentation) multiple (0 - n) external documentations which are of the type foaf:Documents. Foaf:Documents may be external documentations of 0 to n instruments (Instrument). collectionMode are special instruments having at least 1 (1 - n) collection mode (Question), which is a skos:Concept. A specific collection mode can be associated with 0 to n questionnaires (Questionnaire). Questionnaires (Questionnaire) must contain 1 to n questions (Question) using the object property Question. Particular questions (Question) may be contained in 0 to n questionnaires (Questionnaire).

The following example illustrates a questionnaire with three example questions. The questions are defined the next section.

EXAMPLE 25

```
<#Questionnaire> a disco:Questionnaire;
  disco:question <#QuestionGender>;
  disco:question <#QuestionAge>;
  disco:question <#QuestionCitizenship>.
```

8.2 Question

A **`Question`** is designed to get information upon a subject, or sequence of subjects, from a respondent.

Questions (`Question`) have a question text (`questionText`), a label (`skos:prefLabel`), exactly 1 universe (`Universe`), multiple (1 - n) concepts (`concept`), and at least 1 response domain (`responseDomain`). Representations (`Representation`) may have 0 to n `responseDomain` relations to questions (`Question`). Particular universes (`Universe`) may be connected with 0 to n questions (`Question`). `Skos:Concepts` are associated with 0 to n questions (`Question`).

EXAMPLE 26

```
<#QuestionGender> a disco:Question;
  disco:questionText "2. Is the person a man or a woman? [] Man, [] Woman"@en.

<#questionAge> a disco:Question;
  disco:QuestionText "3. What is his or her age? _ _ Mark the age in
                      completed years at the date of the census for those younger than
                      one year old mark 00. For those younger than 10 years old, mark 01,
                      02, 03, etc. For those older than 99 years old, mark 99."@en.

<#questionCitizenship> a disco:Question;
  disco:QuestionText "6. [Immigration status] Only for persons who have
                      usual residence in Argentina and were born in another country.
                      [Questions 6A and 6B asked only of persons born outside Argentina
                      and who currently reside in Argentina.] B. Are you a naturalized
                      citizen of Argentina? [] Yes [] No [] Unanswered"@en.
```

9. Use of Other 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: describ-ing metadata for citation purposes using the DCMI Metadata Terms (DCMI) [DCMI], de-scribing catalogues of datasets using the Data Catalog Vocabulary (DCAT) [DCAT], describ-ing aggregate data like multi-dimensional tables using the RDF Data Cube Vocabulary [RDF Data Cube Vocabulary], describing formal statistical classifications using the SKOS Extension for Statistics (XKOS) [XKOS], and delineating code lists, category schemes, mappings between them, and concepts like topics using the Simple Knowledge Organization System (SKOS) [SKOS]. Furthermore, the external vocabularies Friend of a Friend (FOAF) [FOAF], the Organization Ontology (ORG) [ORG], the Asset Description Metadata Schema (ADMS) [ADMS], and the PROV Ontology (PROV-O) [PROV-O] are used.

9.1 DCMI Metadata Terms (DCMI)

DCMI is reused in order to describe general metadata of Disco constructs such as a study abstract (dcterms:abstract), a study or dataset title (dcterms:title), a human readable description of a Disco construct (dcterms:description), provenance information for a data file (dcterms:provenance), or the date (or date range) at which a study will become available (dcterms:available).

9.2 Friend of a Friend (FOAF) and Organization Ontology (ORG)

With-in the context of Disco, FOAF as well as ORG are reused. 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 de-fined as sub-property of dcterms:contributor.

9.3 Asset Description Metadata Schema (ADMS)

Especially persons and organizations may hold one or more persistent identifiers of particular schemes and agencies (e.g. ORCID, FundRef) 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.

9.4 PROV Ontology (PROV-O)

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.

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 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)?

9.5 Simple Knowledge Organization System (SKOS)

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 statis-tical 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.

9.6 SKOS Extension for Statistics (XKOS)

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 [eXtended Knowledge Organization System]. While the use of XKOS is not required by this vocabulary, the two are designed to work in complementary fashion. SKOS properties may be substituted by additional XKOS properties.

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 descrip-tion 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 classifi-cation (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 queries 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).

9.7 Data Catalog Vocabulary (DCAT)

DCAT is a W3C standard 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

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

> EXAMPLE 27
>
> ```
> Example for the usage of dcat vocabulary
> ```

9.8 RDF Data Cube Vocabulary

The RDF Data Cube Vocabulary is a W3C standard for representing data cubes, that is, multidimensional aggregate data. A `qb: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 [Semantic Statistics]. This is not the case for Disco, which only describes the structure of a da-taset, 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 file, or in a proprietary statistical package file format) that is not represented in RDF.

10. From Literals to Globally Unique Identifiers

> ISSUE 1
>
> This section should talk about Issue #27.

11. Mapping from DDI-XML to DDI-RDF

In this section a detailed mapping from DDI Codebook and Lifecycle is provided. It allows an easy adoption of the DDI Discovery Vocabulary for existing DDI metadata. XSLTs for converting any XML output of DDI-C and DDI-L are available at the DDI-RDF-tools project page.

11.1 Overview of the Mapping from DDI-C and DDI-L to DDI-RDF

11.1.1 Studies and StudyGroups

#	property	domain class	range class	DDI-C	DDI-L
1	universe	union of Study and StudyGroup	Universe	X	X
2	dcterms:subject	union of Study and StudyGroup	skos:Concept		X
3	dcterms:temporal	union of Study and StudyGroup	dcterms:PeriodOfTime		
4	dcterms:spatial	union of Study and StudyGroup	dcterms:Location		
5	kindOfData	union of Study and StudyGroup	skos:Concept		X
6	analysisUnit	union of Study and StudyGroup	AnalysisUnit		
7	dcterms:abstract	union of Study and StudyGroup	rdf:langString	X	X
8	dcterms:alternative	union of Study and StudyGroup	rdf:langString	X	X
9	dcterms:available	union of Study and StudyGroup	xsd:dateTime		X
10	dcterms:title	union of Study and StudyGroup	rdf:langString	X	X
11	purpose	union of Study and StudyGroup	rdf:langString		X
12	subtitle	union of Study and StudyGroup	rdf:langString	X	X
13	ddiFile	union of Study and StudyGroup	foaf:Document		
14	fundedBy	union of Study and StudyGroup	foaf:Agent		
15	dcterms:creator	union of Study and StudyGroup	foaf:Agent		X
16	dcterms:contributor	union of Study and StudyGroup	foaf:Agent		
17	dcterms:publisher	union of Study and StudyGroup	foaf:Agent	-	X
18	instrument	Study	Instrument		X
19	inGroup	Study	StudyGroup		X
20	dataFile	Study	DataFile		X
21	variable	Study	Variable	X	X
22	product	Study	LogicalDataSet		X
23	owl:versionInfo	Study			
24	skos:definition	Universe	rdf:langString		X

11.1.2 General Metadata

#	property	domain class	range class	DDI-C	DDI-L

1	adms:identifier	disco:Study	adms:Identifier		X
2	adms:identifier	disco:StudyGroup	adms:Identifier		
3	adms:identifier	disco:AnalysisUnit	adms:Identifier		
4	adms:identifier	disco:Universe	adms:Identifier		
5	adms:identifier	disco:LogicalDataSet	adms:Identifier		
6	adms:identifier	disco:DataFile	adms:Identifier		X
7	adms:identifier	disco:DescriptiveStatistics	adms:Identifier		
8	adms:identifier	disco:SummaryStatistics	adms:Identifier		
9	adms:identifier	disco:CategoryStatistics	adms:Identifier		
10	adms:identifier	disco:Variable	adms:Identifier		X
11	adms:identifier	disco:VariableDefinition	adms:Identifier		
12	adms:identifier	disco:Question	adms:Identifier		
13	adms:identifier	disco:Instrument	adms:Identifier		
14	adms:identifier	disco:Questionnaire	adms:Identifier		
15	skos_prefLabel	rdfs:Resource	rdf:langString		
16	dcterms:relation	rdfs:Resource	foaf:Document		
17	dcterms:description	dcterms:RightsStatement	rdf:langString		
18	skos:prefLabel	dcterms:RightsStatement	rdf:langString		
19	rdfs:seeAlso	dcterms:RightsStatement	foaf:Document		
20	skos:prefLabel	dcterms:PeriodOfTime	rdf:langString		
21	startDate	dcterms:PeriodOfTime	xsd:date		
22	endDate	dcterms:PeriodOfTime	xsd:Date		
23	skos:prefLabel	dcterms:MediaTypeOrExtent	rdf:langString		
24	org:memberOf	foaf:Person	org:Organization		

11.1.3 Data Sets, Data Files, and Descriptive Statistics

#	property	domain class	range class	DDI-C	DDI-L
1	instrument	LogicalDataSet	Instrument		
2	dataFile	LogicalDataSet	DataFile		
3	aggregation	LogicalDataSet	qb:DataSet		
4	containsVariable	LogicalDataSet	Variable		
5	universe	LogicalDataSet	Universe	X	
6	dcterms:title	LogicalDataSet	rdf:langString		X
7	isPublic	LogicalDataSet	xsd:boolean		
8	dcterms:accessRights	LogicalDataSet	dcterms:RightsStatement		X
9	dcterms:license	LogicalDataSet	dcterms:LicenseDocument		
10	inputVariable	qb:DataSet	Variable		
11	caseQuantity	DataFile	xsd:nonNegativeInteger		X
12	dcterms:description	DataFile	rdf:langstring		
13	owl:versioninfo	DataFile	string		X
14	dcterms:temporal	DataFile	dcterms:PeriodOfTime		
15	dcterms:spatial	DataFile	dcterms:Location		X
16	dcterms:provenance	DataFile	dcterms:ProvenanceStatement		
17	dcterms:subject	DataFile	skos:Concept		
18	dcterms:format	DataFile	dcterms:MediaTypeOrExtend		
19	statisticsDataFile	DescriptiveStatistics	DataFile		
20	statisticsVariable	SummaryStatistics	Variable		
21	invalidcases	SummaryStatistics	xsd:nonNegativeInteger		
22	maximum	SummaryStatistics	xsd:decimal		
23	mean	SummaryStatistics	xsd:decimal		
24	median	SummaryStatistics	xsd:decimal		
25	minimum	SummaryStatistics	xsd:decimal		
26	mode	SummaryStatistics	xsd:decimal		
27	standardDeviation	SummaryStatistics	xsd:decimal		
28	validCases	SummaryStatistics	xsd:nonNegativeInteger		
29	weightedInvalidCases	SummaryStatistics	xsd:nonNegativeInteger		
30	weightedMean	SummaryStatistics	xsd:decimal		
31	weightedMedian	SummaryStatistics	xsd:decimal		
32	weightedMode	SummaryStatistics	xsd:decimal		
33	weightedValidCases	SummaryStatistics	xsd:nonNegativeInteger		
34	statisticsCategory	CategoryStatistics	skos:Concept		
35	cumulativePercentage	CategoryStatistics	xsd:decimal		
36	frequency	CategoryStatistics	xsd:nonNegativeInteger		
37	percentage	CategoryStatistics	xsd:decimal		
38	weightedCumulativePercentage	CategoryStatistics	xsd:decimal		
39	weightedFrequency	CategoryStatistics	xsd:nonNegativeInteger		
40	weightedPercentage	CategoryStatistics	xsd:decimal		

11.1.4 Variables, Variable Definitions, Representations, and Concepts

#	property	domain class	range class	DDI-C	DDI-L
1	skos:inScheme	skos:Concept	skos:ConceptScheme		
2	skos:hasTopConcept	skos:ConceptScheme	skos:Concept		
3	skos:broader	skos:Concept	skos:Concept		X
4	skos:narrower	skos:Concept	skos:Concept		
5	skos:definition	skos:Concept	rdf:langString		
6	skos:notation	skos:Concept	rdfs:Literal		X
7	skos:prefLabel	skos:Concept	rdf:LangString		
8	question	Variable	Question		X
9	universe	Variable	Universe	X	X
10	analysisUnit	Variable	AnalysisUnit		
11	concept	Variable	skos:Concept		X
12	representation	Variable	Representation		
13	basedOn	Variable	VariableDefinition		
14	dcterms:description	Variable	rdf:langString		X
15	skos:notation	Variable	rdfs:Literal		X
16	skos:prefLabel	Variable	rdf:langString		X
17	concept	VariableDefinition	skos:Concept		
18	universe	VariableDefinition	Universe		
19	representation	VariableDefinition	Representation		
20	dcterms:description	VariableDefinition	rdf:langString		
21	skos:prefLabel	VariableDefinition	rdf:langString		

11.1.5 Data Collection

#	property	domain class	range class	DDI-C	DDI-L
1	universe	Question	Universe	X	X
2	concept	Question	skos:Concept		X
3	responseDomain	Question	Representation		
4	questionText	Question	rdf:langString		X
5	skos:prefLabel	Question	rdf:langString		X
6	question	Questionnaire	Question		
7	collectionMode	Questionnaire	skos:Concept		
8	externalDocumentation	Instrument	foaf:Document		
9	dcterms:description	Instrument	rdf:langString		X
10	skos:prefLabel	Instrument	rdf:langString		X

11.2 Mapping from DDI-C to DDI-RDF

11.2.1 Studies and StudyGroups

#	property	domain class	range class	mapping
1	universe	union of Study and StudyGroup	Universe	/codeBook/stdyDscr/stdyInfo/sumDscr/universe
2	dcterms:subject	union of Study and StudyGroup	skos:Concept	
3	dcterms:temporal	union of Study and StudyGroup	dcterms:PeriodOfTime	
4	dcterms:spatial	union of Study and StudyGroup	dcterms:Location	
5	kindOfData	union of Study and StudyGroup	skos:Concept	
6	analysisUnit	union of Study and StudyGroup	AnalysisUnit	
7	dcterms:abstract	union of Study and StudyGroup	rdf:langString	/codeBook/stdyDscr/stdyInfo/abstract
8	dcterms:alternative	union of Study and StudyGroup	rdf:langString	/codeBook/stdyDscr/citation/altTitl
9	dcterms:available	union of Study and StudyGroup	xsd:dateTime	
10	dcterms:title	union of Study and StudyGroup	rdf:langString	/codeBook/stdyDscr/citation/titl
11	purpose	union of Study and StudyGroup	rdf:langString	
12	subtitle	union of Study and StudyGroup	rdf:langString	/codeBook/stdyDscr/citation/subTitl
13	ddiFile	union of Study and StudyGroup	foaf:Document	
14	fundedBy	union of Study and StudyGroup	foaf:Agent	
15	dcterms:creator	union of Study and StudyGroup	foaf:Agent	
16	dcterms:contributor	union of Study and StudyGroup	foaf:Agent	
17	dcterms:publisher	union of Study and StudyGroup	foaf:Agent	
18	instrument	Study	Instrument	
19	inGroup	Study	StudyGroup	
20	dataFile	Study	DataFile	
21	variable	Study	Variable	/codeBook/dataDscr/var/@id
22	product	Study	LogicalDataSet	
23	owl:versionInfo	Study		
24	skos:definition	Universe	rdf:langString	

notes

- **(1):** -

11.2.2 General Metadata

#	property	domain class	range class	mapping
1	adms:identifier	disco:Study	adms:Identifier	
2	adms:identifier	disco:StudyGroup	adms:Identifier	
3	adms:identifier	disco:AnalysisUnit	adms:Identifier	
4	adms:identifier	disco:Universe	adms:Identifier	
5	adms:identifier	disco:LogicalDataSet	adms:Identifier	
6	adms:identifier	disco:DataFile	adms:Identifier	
7	adms:identifier	disco:DescriptiveStatistics	adms:Identifier	
8	adms:identifier	disco:SummaryStatistics	adms:Identifier	
9	adms:identifier	disco:CategoryStatistics	adms:Identifier	
10	adms:identifier	disco:Variable	adms:Identifier	
11	adms:identifier	disco:VariableDefinition	adms:Identifier	
12	adms:identifier	disco:Question	adms:Identifier	
13	adms:identifier	disco:Instrument	adms:Identifier	
14	adms:identifier	disco:Questionnaire	adms:Identifier	
15	skos_prefLabel	rdfs:Resource	rdf:langString	
16	dcterms:relation	rdfs:Resource	foaf:Document	
17	dcterms:description	dcterms:RightsStatement	rdf:langString	
18	skos:prefLabel	dcterms:RightsStatement	rdf:langString	
19	rdfs:seeAlso	dcterms:RightsStatement	foaf:Document	
20	skos:prefLabel	dcterms:PeriodOfTime	rdf:langString	
21	startDate	dcterms:PeriodOfTime	xsd:date	
22	endDate	dcterms:PeriodOfTime	xsd:Date	
23	skos:prefLabel	dcterms:MediaTypeOrExtent	rdf:langString	
24	org:memberOf	foaf:Person	org:Organization	

notes

- **(1):** -

11.2.3 Data Sets, Data Files, and Descriptive Statistics

#	property	domain class	range class	mapping
1	instrument	LogicalDataSet	Instrument	
2	dataFile	LogicalDataSet	DataFile	
3	aggregation	LogicalDataSet	qb:DataSet	
4	containsVariable	LogicalDataSet	Variable	
5	universe	LogicalDataSet	Universe	/codeBook/stdyDscr/stdyInfo/sumDscr/universe
6	dcterms:title	LogicalDataSet	rdf:langString	
7	isPublic	LogicalDataSet	xsd:boolean	
8	dcterms:accessRights	LogicalDataSet	dcterms:RightsStatement	
9	dcterms:license	LogicalDataSet	dcterms:LicenseDocument	
10	inputVariable	qb:DataSet	Variable	
11	caseQuantity	DataFile	xsd:nonNegativeInteger	
12	dcterms:description	DataFile	rdf:langstring	
13	owl:versioninfo	DataFile	string	
14	dcterms:temporal	DataFile	dcterms:PeriodOfTime	
15	dcterms:spatial	DataFile	dcterms:Location	
16	dcterms:provenance	DataFile	dcterms:ProvenanceStatement	
17	dcterms:subject	DataFile	skos:Concept	
18	dcterms:format	DataFile	dcterms:MediaTypeOrExtend	
19	statisticsDataFile	DescriptiveStatistics	DataFile	
20	statisticsVariable	SummaryStatistics	Variable	
21	invalidcases	SummaryStatistics	xsd:nonNegativeInteger	
22	maximum	SummaryStatistics	xsd:decimal	
23	mean	SummaryStatistics	xsd:decimal	
24	median	SummaryStatistics	xsd:decimal	
25	minimum	SummaryStatistics	xsd:decimal	
26	mode	SummaryStatistics	xsd:decimal	
27	standardDeviation	SummaryStatistics	xsd:decimal	
28	validCases	SummaryStatistics	xsd:nonNegativeInteger	
29	weightedInvalidCases	SummaryStatistics	xsd:nonNegativeInteger	

30	weightedMean	SummaryStatistics	xsd:decimal	
31	weightedMedian	SummaryStatistics	xsd:decimal	
32	weightedMode	SummaryStatistics	xsd:decimal	
33	weightedValidCases	SummaryStatistics	xsd:nonNegativeInteger	
34	statisticsCategory	CategoryStatistics	skos:Concept	
35	cumulativePercentage	CategoryStatistics	xsd:decimal	
36	frequency	CategoryStatistics	xsd:nonNegativeInteger	
37	percentage	CategoryStatistics	xsd:decimal	
38	weightedCumulativePercentage	CategoryStatistics	xsd:decimal	
39	weightedFrequency	CategoryStatistics	xsd:nonNegativeInteger	
40	weightedPercentage	CategoryStatistics	xsd:decimal	

notes

- **(1):** -

11.2.4 Variables, Variable Definitions, Representations, and Concepts

#	property	domain class	range class	mapping
1	skos:inScheme	skos:Concept	skos:ConceptScheme	
2	skos:hasTopConcept	skos:ConceptScheme	skos:Concept	
3	skos:broader	skos:Concept	skos:Concept	
4	skos:narrower	skos:Concept	skos:Concept	
5	skos:definition	skos:Concept	rdf:langString	
6	skos:notation	skos:Concept	rdfs:Literal	
7	skos:prefLabel	skos:Concept	rdf:LangString	
8	question	Variable	Question	
9	universe	Variable	Universe	/codeBook/stdyDscr/stdyInfo/sumDscr/universe
10	analysisUnit	Variable	AnalysisUnit	
11	concept	Variable	skos:Concept	
12	representation	Variable	Representation	
13	basedOn	Variable	VariableDefinition	
14	dcterms:description	Variable	rdf:langString	
15	skos:notation	Variable	rdfs:Literal	
16	skos:prefLabel	Variable	rdf:langString	
17	concept	VariableDefinition	skos:Concept	
18	universe	VariableDefinition	Universe	
19	representation	VariableDefinition	Representation	
20	dcterms:description	VariableDefinition	rdf:langString	
21	skos:prefLabel	VariableDefinition	rdf:langString	

notes

- **(1):** -

11.2.5 Data Collection

#	property	domain class	range class	mapping
1	universe	Question	Universe	/codeBook/stdyDscr/stdyInfo/sumDscr/universe
2	concept	Question	skos:Concept	
3	responseDomain	Question	Representation	
4	questionText	Question	rdf:langString	
5	skos:prefLabel	Question	rdf:langString	
6	question	Questionnaire	Question	
7	collectionMode	Questionnaire	skos:Concept	
8	externalDocumentation	Instrument	foaf:Document	
9	dcterms:description	Instrument	rdf:langString	
10	skos:prefLabel	Instrument	rdf:langString	

notes

- **(1):** -

11.3 Mapping from DDI-L to DDI-RDF

11.3.1 Studies and StudyGroups

#	property	domain class	range class	mapping
1	universe	union of Study and StudyGroup	Universe	/ddi:DDIInstance/s:StudyUnit/r:UniverseReference/r:ID

2	dcterms:subject	union of Study and StudyGroup	skos:Concept	/ddi:DDIInstance/s:StudyUnit/r:TopicalCoverage/r:Subject
3	dcterms:temporal	union of Study and StudyGroup	dcterms:PeriodOfTime	
4	dcterms:spatial	union of Study and StudyGroup	dcterms:Location	
5	kindOfData	union of Study and StudyGroup	skos:Concept	/ddi:DDIInstance/s:StudyUnit/r:KindOfData
6	analysisUnit	union of Study and StudyGroup	AnalysisUnit	/ddi:DDIInstance/s:StudyUnit/r:AnalysisUnit
7	dcterms:abstract	union of Study and StudyGroup	rdf:langString	/ddi:DDIInstance/s:StudyUnit/s:Abstract/r:Content
8	dcterms:alternative	union of Study and StudyGroup	rdf:langString	/ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle
9	dcterms:available	union of Study and StudyGroup	xsd:dateTime	/ddi:DDIInstance/s:StudyUnit/r:Embargo/r:Date/r:SimpleDate
10	dcterms:title	union of Study and StudyGroup	rdf:langString	/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title
11	purpose	union of Study and StudyGroup	rdf:langString	/ddi:DDIInstance/s:StudyUnit/s:Purpose/r:Content
12	subtitle	union of Study and StudyGroup	rdf:langString	/ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle
13	ddiFile	union of Study and StudyGroup	foaf:Document	
14	fundedBy	union of Study and StudyGroup	foaf:Agent	/ddi:DDIInstance/s:StudyUnit/r:FundingInformation
15	dcterms:creator	union of Study and StudyGroup	foaf:Agent	/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Creator
16	dcterms:contributor	union of Study and StudyGroup	foaf:Agent	/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Contributor
17	dcterms:publisher	union of Study and StudyGroup	foaf:Agent	/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Publisher
18	instrument	Study	Instrument	/ddi:DDIInstace/s:StudyUnit/d:DataCollection/@id
19	inGroup	Study	StudyGroup	//s:StudyUnit/ancestor::g:Group[1]/@id
20	dataFile	Study	DataFile	//s:StudyUnit/pi:PhysicalInstance/@id
21	variable	Study	Variable	/ddi:DDIInstance/s:StudyUnit//l:Variable/@id
22	product	Study	LogicalDataSet	//s:StudyUnit/l:LogicalProduct/@id
23	owl:versionInfo	Study		
24	skos:definition	Universe	rdf:langString	c:Universe/c:HumanReadable

notes

- **(2):** inf code list is defined use it as the identifier
- **(9):** the date the study is available to the public
- **(13):** the URI to the DDI file(s) defined via param to the xslt
- **(21):** suggested for identification

11.3.2 General Metadata

#	property	domain class	range class	mapping
1	adms:identifier	disco:Study	adms:Identifier	/ddi:DDIInstance/s:StudyUnit/@id
2	adms:identifier	disco:StudyGroup	adms:Identifier	
3	adms:identifier	disco:AnalysisUnit	adms:Identifier	
4	adms:identifier	disco:Universe	adms:Identifier	
5	adms:identifier	disco:LogicalDataSet	adms:Identifier	
6	adms:identifier	disco:DataFile	adms:Identifier	//pi:PhysicalInstance/pi:DataFileIdentification
7	adms:identifier	disco:DescriptiveStatistics	adms:Identifier	
8	adms:identifier	disco:SummaryStatistics	adms:Identifier	
9	adms:identifier	disco:CategoryStatistics	adms:Identifier	
10	adms:identifier	disco:Variable	adms:Identifier	//l:Variable/l:VariableName
11	adms:identifier	disco:VariableDefinition	adms:Identifier	
12	adms:identifier	disco:Question	adms:Identifier	
13	adms:identifier	disco:Instrument	adms:Identifier	
14	adms:identifier	disco:Questionnaire	adms:Identifier	
15	skos_prefLabel	rdfs:Resource	rdf:langString	
16	dcterms:relation	rdfs:Resource	foaf:Document	
17	dcterms:description	dcterms:RightsStatement	rdf:langString	
18	skos:prefLabel	dcterms:RightsStatement	rdf:langString	
19	rdfs:seeAlso	dcterms:RightsStatement	foaf:Document	
20	skos:prefLabel	dcterms:PeriodOfTime	rdf:langString	
21	startDate	dcterms:PeriodOfTime	xsd:date	
22	endDate	dcterms:PeriodOfTime	xsd:Date	
23	skos:prefLabel	dcterms:MediaTypeOrExtent	rdf:langString	
24	org:memberOf	foaf:Person	org:Organization	

notes

- **(1):** s:StudyUnit/r:Archive/a:ArchiveSpecific/a:Collection/a:CallNumber is also a candidate for identification

11.3.3 Data Sets, Data Files, and Descriptive Statistics

#	property	domain class	range class	mapping
1	instrument	LogicalDataSet	Instrument	
2	dataFile	LogicalDataSet	DataFile	
3	aggregation	LogicalDataSet	qb:DataSet	

4	containsVariable	LogicalDataSet	Variable	
5	universe	LogicalDataSet	Universe	
6	dcterms:title	LogicalDataSet	rdf:langString	//l:LogicalProduct/r:Label
7	isPublic	LogicalDataSet	xsd:boolean	
8	dcterms:accessRights	LogicalDataSet	dcterms:RightsStatement	ancestor::s:StudyUnit/a:Archive/a:DefaultAccess/a:AccessConditions
9	dcterms:license	LogicalDataSet	dcterms:LicenseDocument	
10	inputVariable	qb:DataSet	Variable	
11	caseQuantity	DataFile	xsd:nonNegativeInteger	//pi:PhysicalInstance/pi:GrossFileStructure/pi:CaseQuantity
12	dcterms:description	DataFile	rdf:langstring	
13	owl:versioninfo	DataFile	string	//pi:PhysicalInstance/@version
14	dcterms:temporal	DataFile	dcterms:PeriodOfTime	
15	dcterms:spatial	DataFile	dcterms:Location	pi:PhysicalInstance/r:Coverage/r:SpatialCoverage/@id | pi:PhysicalInstance/r:Coverage/r:SpatialCoverageReference/r:ID
16	dcterms:provenance	DataFile	dcterms:ProvenanceStatement	
17	dcterms:subject	DataFile	skos:Concept	
18	dcterms:format	DataFile	dcterms:MediaTypeOrExtend	
19	statisticsDataFile	DescriptiveStatistics	DataFile	
20	statisticsVariable	SummaryStatistics	Variable	
21	invalidcases	SummaryStatistics	xsd:nonNegativeInteger	
22	maximum	SummaryStatistics	xsd:decimal	
23	mean	SummaryStatistics	xsd:decimal	
24	median	SummaryStatistics	xsd:decimal	
25	minimum	SummaryStatistics	xsd:decimal	
26	mode	SummaryStatistics	xsd:decimal	
27	standardDeviation	SummaryStatistics	xsd:decimal	
28	validCases	SummaryStatistics	xsd:nonNegativeInteger	
29	weightedInvalidCases	SummaryStatistics	xsd:nonNegativeInteger	
30	weightedMean	SummaryStatistics	xsd:decimal	
31	weightedMedian	SummaryStatistics	xsd:decimal	
32	weightedMode	SummaryStatistics	xsd:decimal	
33	weightedValidCases	SummaryStatistics	xsd:nonNegativeInteger	
34	statisticsCategory	CategoryStatistics	skos:Concept	
35	cumulativePercentage	CategoryStatistics	xsd:decimal	
36	frequency	CategoryStatistics	xsd:nonNegativeInteger	
37	percentage	CategoryStatistics	xsd:decimal	
38	weightedCumulativePercentage	CategoryStatistics	xsd:decimal	
39	weightedFrequency	CategoryStatistics	xsd:nonNegativeInteger	
40	weightedPercentage	CategoryStatistics	xsd:decimal	

notes

- **(7):** not populated from DDI (could be set as an param to the xslt)
- **(17):** located in pi:PhysicalInstance/r:Coverage/r:TopicalCoverage (both subject and keyword)

11.3.4 Variables, Variable Definitions, Representations, and Concepts

#	property	domain class	range class	mapping
1	skos:inScheme	skos:Concept	skos:ConceptScheme	
2	skos:hasTopConcept	skos:ConceptScheme	skos:Concept	
3	skos:broader	skos:Concept	skos:Concept	c:Universe/c:SubUniverse/@id
4	skos:narrower	skos:Concept	skos:Concept	
5	skos:definition	skos:Concept	rdf:langString	c:Universe/c:UniverseName
6	skos:notation	skos:Concept	rdfs:Literal	c:Universe/c:MachineReadable [skos:notation is only used to represent codes]
7	skos:prefLabel	skos:Concept	rdf:LangString	c:Universe/r:Label [skos:notation is only used to represent categories]
8	question	Variable	Question	//l:Variable/r:QuestionReference/r:ID
9	universe	Variable	Universe	//l:Variable/r:UniverseReference/r:ID
10	analysisUnit	Variable	AnalysisUnit	
11	concept	Variable	skos:Concept	//l:Variable/r:ConceptReference/r:ID
12	representation	Variable	Representation	
13	basedOn	Variable	VariableDefinition	
14	dcterms:description	Variable	rdf:langString	//l:Variable/r:Description
15	skos:notation	Variable	rdfs:Literal	//l:Variable/l:VariableName
16	skos:prefLabel	Variable	rdf:langString	//l:Variable/r:Label
17	concept	VariableDefinition	skos:Concept	
18	universe	VariableDefinition	Universe	
19	representation	VariableDefinition	Representation	
20	dcterms:description	VariableDefinition	rdf:langString	
21	skos:prefLabel	VariableDefinition	rdf:langString	

notes

- **(12):** not sure where to map to in DDI 3.1
- **(13):** coming in DDI 3.2

11.3.5 Data Collection

#	property	domain class	range class	mapping
1	universe	Question	Universe	//l:Variable/r:UniverseReference/r:ID
2	concept	Question	skos:Concept	//l:Variable/r:ConceptReference/r:ID
3	responseDomain	Question	Representation	
4	questionText	Question	rdf:langString	//d:QuestionItem | d:MultipleQuestionItem/d:QuestionText/d:LiteralText/d:Text
5	skos:prefLabel	Question	rdf:langString	//d:QuestionItem/d:QuestionItemName | d:MultipleQuestionItem/d:MultipleQuestionItemName
6	question	Questionnaire	Question	
7	collectionMode	Questionnaire	skos:Concept	
8	externalDocumentation	Instrument	foaf:Document	
9	dcterms:description	Instrument	rdf:langString	d:Intrument/r:Description
10	skos:prefLabel	Instrument	rdf:langString	d:Instrument/r:Label

notes

- **(4):** question-text exists for multiple elements
- **(5):** the question name as label

12. Mappings

12.1 GSIM

12.2 Schema.org

A. Vocabulary Reference

1. Studies and StudyGroups

Class: **`disco:Study`**

A Study represents the process by which a data set was generated or collected.

Object Property: **`disco:variable`** **(*Domain:*`disco:Study` *-> Range:* `disco:Variable`)**

Indicates the Variable of a Study.

Object Property: **`disco:ddifile`** **(*Domain:*`disco:Study`, `disco:StudyGroup` *-> Range:* `foaf:Document`)**

points from a Study or a StudyGroup to the original DDI file which is a foaf:Document.

Object Property: **`disco:inGroup`** **(*Domain:*`disco:Study` *-> Range:* `disco:StudyGroup`)**

points from a Study to the StudyGroup which contains the Study.

Object Property: **`disco:universe`** **(*Domain:*`disco:Study`, `disco:StudyGroup`, `disco:VariableDefinition`, `disco:Variable`, `disco:Question`, `disco:LogicalDataSet` *-> Range:* `disco:Universe`)**

Indicates the Universe(s) of Studies, StudyGrous, VariableDefinitions, Variables, Questions, and LogicalDataSets.

Object Property: **`disco:fundedBy`** **(*Domain:*`disco:Study`, `disco:StudyGroup` *-> Range:* `foaf:Agent`*; sub property of:* `dcterms:contributor`)**

points from a Study or a StudyGroup to the funding foaf:Agent which is either a foaf:Person or a org:Organization.

Object Property: **`disco:dataFile`** **(*Domain:*`disco:Study`, `disco:LogicalDataSet` *-> Range:* `disco:DataFile`)**

points to the DataFile of a Study or a LogicalDataSet.

Object Property: **`disco:kindOfData`** **(*Domain:*`disco:Study`, `disco:StudyGroup` *-> Range:* `skos:Concept`)**

The general kind of data (e.g. geospatial, register, survey) collected in this study, given either as a skos:Concept, or as a blank node with attached free-text rdfs:label.

Object Property: **`disco:product`** **(*Domain:*`disco:Study` *-> Range:* `http://purl.org/linked-data/cube#LogicalDataSet`)**

Indicates the LogicalDataSets of a Studies.

Object Property: **`disco:instrument`** **(*Domain:*`disco:Study`, `disco:LogicalDataSet` *-> Range:* `disco:Instrument`)**

Indicates the Instrument of a Study or a LogicalDataSet.

Datatype Property: `disco:subtitle` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `rdf:langString` **)**

The sub-title of a Study of a StudyGroup.

Datatype Property: `disco:purpose` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `rdf:langString` **)**

The purpose of a Study of a StudyGroup.

Class: `disco:StudyGroup`
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 (to use the DDI term). In this case, a number of Study objects would be collected into a single StudyGroup.

Object Property: `disco:ddifile` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `foaf:Document` **)**

points from a Study or a StudyGroup to the original DDI file which is a foaf:Document.

Object Property: `disco:universe` **(*Domain:*** `disco:Study`, `disco:StudyGroup`, `disco:VariableDefinition`, `disco:Variable`, `disco:Question`, `disco:LogicalDataSet` ***-> Range:*** `disco:Universe` **)**

Indicates the Universe(s) of Studies, StudyGrous, VariableDefinitions, Variables, Questions, and LogicalDataSets.

Object Property: `disco:fundedBy` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `foaf:Agent`***; sub property of:*** `dcterms:contributor` **)**

points from a Study or a StudyGroup to the funding foaf:Agent which is either a foaf:Person or a org:Organization.

Object Property: `disco:kindOfData` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `skos:Concept` **)**

The general kind of data (e.g. geospatial, register, survey) collected in this study, given either as a skos:Concept, or as a blank node with attached free-text rdfs:label.

Datatype Property: `disco:subtitle` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `rdf:langString` **)**

The sub-title of a Study of a StudyGroup.

Datatype Property: `disco:purpose` **(*Domain:*** `disco:Study`, `disco:StudyGroup` ***-> Range:*** `rdf:langString` **)**

The purpose of a Study of a StudyGroup.

Class: `disco:AnalysisUnit` **Sub Class of:** `skos:Concept`
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.

Class: `disco:Universe` **Sub Class of:** `skos:Concept`
A Universe is the total membership or population of a defined class of people, objects or events.

Class: `disco:DataDiscoveryDocument`
Data discovery document in DDI. Dct:publisher is used for the agency.

2. Data Sets, Data Files, and Descriptive Statistics

Class: `disco:LogicalDataSet` **Sub Class of:** `http://www.w3.org/ns/dcat#Dataset`
Each study has a set of logical metadata associated with the processing of data, at the time of collection or later during cleaning, and re-coding. LogicalDataSet represents the microdata dataset.

Object Property: `disco:containsVariable` **(*Domain:*** `disco:LogicalDataSet` ***-> Range:*** `disco:Variable` **)**

points to Variable contained in the LogicalDataSet

Object Property: `disco:universe` **(*Domain:*** `disco:Study`, `disco:StudyGroup`, `disco:VariableDefinition`, `disco:Variable`, `disco:Question`, `disco:LogicalDataSet` ***-> Range:*** `disco:Universe` **)**

Indicates the Universe(s) of Studies, StudyGrous, VariableDefinitions, Variables, Questions, and LogicalDataSets.

Object Property: `disco:dataFile` **(*Domain:*** `disco:Study`, `disco:LogicalDataSet` ***-> Range:*** `disco:DataFile` **)**

points to the DataFile of a Study or a LogicalDataSet.

Object Property: `disco:aggregation` **(*Domain:*** `disco:LogicalDataSet` ***-> Range:*** `http://purl.org/linked-data/cube#DataSet` **)**

points to the aggregated data set of a microdata data set.

Object Property: `disco:instrument` **(*Domain:*** `disco:Study`, `disco:LogicalDataSet` ***-> Range:*** `disco:Instrument` **)**

Indicates the Instrument of a Study or a LogicalDataSet.

Datatype Property: **disco:isPublic** (***Domain:***disco:LogicalDataSet ***-> Range:*** xsd:boolean)

The value true indicates that the dataset can be accessed (usually downloaded) by anyone.

Class: **disco:DataFile** **Sub Class of:** **http://www.w3.org/ns/dcat#Distribution**
The class DataFile, which is also a dcterms:Dataset, represents all the data files containing the microdata datasets.

Datatype Property: **disco:caseQuantity** (***Domain:***disco:DataFile ***-> Range:*** xsd:nonNegativeInteger)

case quantity of a DataFile.

Class: **disco:DescriptiveStatistics**
SummaryStatistics pointing to variables and CategoryStatistics pointing to categories and codes are both DescriptiveStatistics.

Object Property: **disco:statisticsDataFile** (***Domain:***disco:DescriptiveStatistics ***-> Range:*** disco:DataFile)

Indicates the DataFile of a specific DesciptiveStatistics individual.

Class: **disco:SummaryStatistics** **Sub Class of:** **disco:DescriptiveStatistics**
For SummaryStatistics, maximum values, minimum values, and standard deviations can be defined.

Object Property: **disco:statisticsVariable** (***Domain:***disco:SummaryStatistics ***-> Range:*** disco:Variable)

Indicates the Variable of a specific SummaryStatistics individual.

Datatype Property: **disco:standardDeviation** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

standard deviation

Datatype Property: **disco:weightedValidCases** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:nonNegativeInteger)

weighted valid cases

Datatype Property: **disco:minimum** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

minimum

Datatype Property: **disco:weightedMode** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

weighted mode

Datatype Property: **disco:mode** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

mode

Datatype Property: **disco:validCases** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:nonNegativeInteger)

valid cases

Datatype Property: **disco:median** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

median

Datatype Property: **disco:mean** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

mean

Datatype Property: **disco:weightedInvalidCases** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:nonNegativeInteger)

weighted invalid cases

Datatype Property: **disco:weightedMean** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

weighted mean

Datatype Property: **disco:invalidCases** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:nonNegativeInteger)

invalid cases

Datatype Property: **disco:weightedMedian** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

weighted median

Datatype Property: **disco:maximum** (***Domain:***disco:SummaryStatistics ***-> Range:*** xsd:decimal)

maximum

Class: **disco:CategoryStatistics** **Sub Class of:** **disco:DescriptiveStatistics**
For CategoryStatistics, frequencies, percentages, and weighted percentages can be defined.

Object Property: **disco:statisticsCategory** (***Domain:***disco:CategoryStatistics ***-> Range:*** **skos:Concept**)

Indicates the skos:Concept (representing codes and categories) of a specific CategoryStatistics individual.

Datatype Property: **disco:weightedFrequency** (***Domain:***disco:CategoryStatistics ***-> Range:*** **xsd:nonNegativeInteger**)

weighted frequency

Datatype Property: **disco:weightedCumulativePercentage** (***Domain:***disco:CategoryStatistics ***-> Range:*** **xsd:decimal**)

weighted cumulative percentage

Datatype Property: **disco:weightedPercentage** (***Domain:***disco:CategoryStatistics ***-> Range:*** **xsd:decimal**)

weighted percentage

Datatype Property: **disco:cumulativePercentage** (***Domain:***disco:CategoryStatistics ***-> Range:*** **xsd:decimal**)

cumulative percentage

Datatype Property: **disco:frequency** (***Domain:***disco:CategoryStatistics ***-> Range:*** **xsd:nonNegativeInteger**)

frequency

Datatype Property: **disco:percentage** (***Domain:***disco:CategoryStatistics ***-> Range:*** **xsd:decimal**)

percentage

3. Variables, Variable Definitions, Representations, and Concepts

Class: **disco:Variable**
Variables provide a definition of the column in a rectangular data file. 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.

Object Property: **disco:analysisUnit** (***Domain:***disco:StudyGroup, disco:Variable ***-> Range:*** **disco:AnalysisUnit**)

analysis unit of a Study, a StudyGroup, or a Variable.

Object Property: **disco:representation** (***Domain:***disco:VariableDefinition, disco:Variable ***-> Range:***)

VariableDefinitions and Variables can have a Representation whose individuals are either of the class rdfs:Datatype (to represent values) or skos:ConceptScheme (to represent code lists).

Object Property: **disco:universe** (***Domain:***disco:Study, disco:StudyGroup, disco:VariableDefinition, disco:Variable, disco:Question, disco:LogicalDataSet ***-> Range:*** **disco:Universe**)

Indicates the Universe(s) of Studies, StudyGrous, VariableDefinitions, Variables, Questions, and LogicalDataSets.

Object Property: **disco:question** (***Domain:***disco:Variable, disco:Questionnaire ***-> Range:*** **disco:Question**)

Indicates the Questions associated to Variables or contained in Questionnaires.

Object Property: **disco:basedOn** (***Domain:***disco:Variable ***-> Range:*** **disco:VariableDefinition**)

points to the VariableDefinition the Variable is based on.

Object Property: **disco:concept** (***Domain:***disco:VariableDefinition, disco:Question, disco:Variable ***-> Range:*** **skos:Concept**)

points to the DDI concept of a VariableDefinition, a Variable, or a Question

Class: **disco:VariableDefinition**
VariableDefinitions encompasse study-independent, re-usable parts of variables like occupation classification.

Object Property: **disco:representation** (***Domain:***disco:VariableDefinition, disco:Variable ***-> Range:***)

VariableDefinitions and Variables can have a Representation whose individuals are either of the class rdfs:Datatype (to represent values) or skos:ConceptScheme (to represent code lists).

Object Property: `disco:universe` **(*Domain:*`disco:Study`, `disco:StudyGroup`, `disco:VariableDefinition`, `disco:Variable`, `disco:Question`, `disco:LogicalDataSet` *-> Range:* `disco:Universe`)**

Indicates the Universe(s) of Studies, StudyGrous, VariableDefinitions, Variables, Questions, and LogicalDataSets.

Object Property: `disco:concept` **(*Domain:*`disco:VariableDefinition`, `disco:Question`, `disco:Variable` *-> Range:* `skos:Concept`)**

points to the DDI concept of a VariableDefinition, a Variable, or a Question

4. Data Collection

Class: `disco:Question`

A Question is designed to get information upon a subject, or sequence of subjects, from a respondent.

Object Property: `disco:universe` **(*Domain:*`disco:Study`, `disco:StudyGroup`, `disco:VariableDefinition`, `disco:Variable`, `disco:Question`, `disco:LogicalDataSet` *-> Range:* `disco:Universe`)**

Indicates the Universe(s) of Studies, StudyGrous, VariableDefinitions, Variables, Questions, and LogicalDataSets.

Object Property: `disco:concept` **(*Domain:*`disco:VariableDefinition`, `disco:Question`, `disco:Variable` *-> Range:* `skos:Concept`)**

points to the DDI concept of a VariableDefinition, a Variable, or a Question

Datatype Property: `disco:questionText` **(*Domain:*`disco:Question` *-> Range:* `rdf:langString`)**

question text

Class: `disco:Instrument`

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.

Object Property: `disco:externalDocumentation` **(*Domain:*`disco:Instrument` *-> Range:* `foaf:Document`)**

points from an Instrument to a foaf:Document which is the external documentation of the Instrument.

Class: `disco:Questionnaire` **Sub Class of:** `disco:Instrument`

A questionnaire contains a flow of questions.

Object Property: `disco:collectionMode` **(*Domain:*`disco:Questionnaire` *-> Range:* `skos:Concept`)**

mode of collection of a Questionnaire

Object Property: `disco:question` **(*Domain:*`disco:Variable`, `disco:Questionnaire` *-> Range:* `disco:Question`)**

Indicates the Questions associated to Variables or contained in Questionnaires.

B. Combined UML Diagram

The following figure shows the object properties between the most important classes of the DDI-RDF Discovery Vocabulary. Additionally, the cardinalities of these object properties and class hierarchies are visualized.

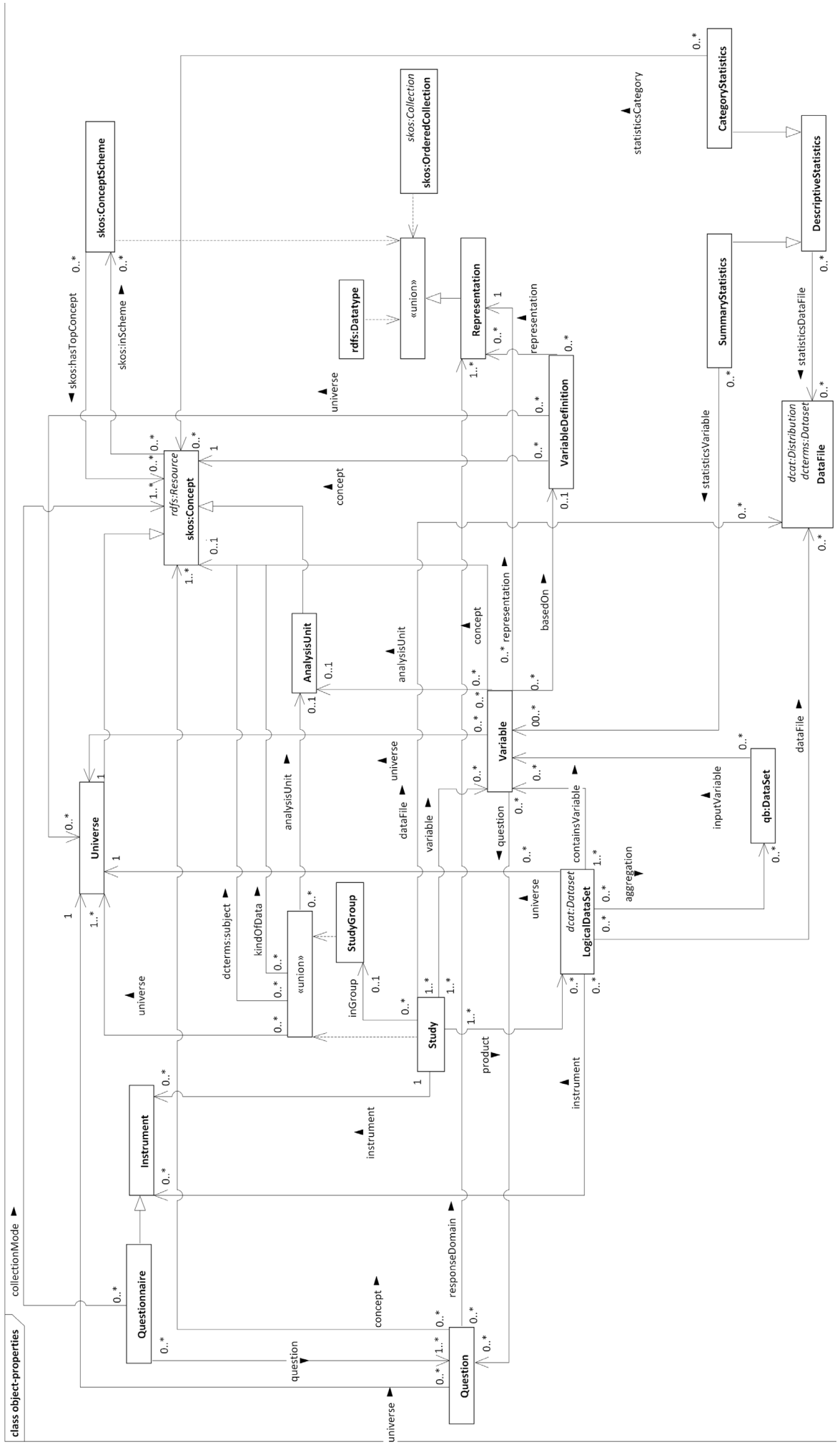


Fig. 34 Combined UML Diagram (object properties only)

A scalable version of this diagram can be found here.

C. Example Queries

Vompras, Gregory, Bosch, Capadisli, and Wackerow [Scenarios] have written a paper describing typical use cases associated with the DDI-RDF Discovery Vocabulary. The specification the DDI-RDF Discovery Vocabulary does not contain the full list of all the possible use cases. The complete list can be found in the mentioned paper. We now show a couple of representative use cases associated with the DDI-RDF Discovery Vocabulary.

Find studies from years 2000 and after about climate change.

EXAMPLE 28

```
SELECT ?studyTitle ?studyAbstract ?logicalDataSetTitle
WHERE {
  ?study a disco:Study ;
    dcterms:title ?studyTitle ;
    dcterms:abstract ?studyAbstract ;
    dcterms:subject [ skos:prefLabel "Climate Change" ] ;
    dcterms:temporal [ disco:startDate ?date ] ;
    disco:product ?logicalDataSet .

  ?logicalDataSet a disco:LogicalDataSet ;
    dcterms:title ?logicalDataSetTitle .

  FILTER (?date >= 2000)
}
```

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.

EXAMPLE 29

```
SELECT ?logicalDataSetTitle
WHERE {
  ?logicalDataSet a disco:LogicalDataSet ;
    dcterms:title ?logicalDataSetTitle ;
    disco:isPublic ?isPublic ;
    dcterms:accessRights ?rightsStatement .

  ?rightsStatement skos:prefLabel ?rightsStatementLabel .

  FILTER (
    ?isPublic = "true" &&
    ?rightsStatementLabel = "Data Liberation Initiative Community"
  )

  OPTIONAL {
    ?rightsStatement rdfs:seeAlso ?rightsStatementURL .
  }
  OPTIONAL {
    ?logicalDataSet dcterms:license ?licenseDocument .
  }
}
```

Find all studies with questions about commuting to work.

EXAMPLE 30

```
SELECT ?studyTitle ?studyAbstract
WHERE {
  ?study a disco:Study ;
    disco:instrument ?instrument ;
    dcterms:title ?studyTitle ;
    dcterms:abstract ?studyAbstract .

  ?instrument disco:questionnaire ?questionnaire .
  ?questionnaire disco:question ?question .
  ?question disco:questionText ?questionText .

  FILTER (regex(?questionText, "commut.*work"))
}
```

Find study groups where the study uses the species variable and has a variable defined as Bufo alvarius

EXAMPLE 31

```
SELECT ?studyGroupTitle ?studyGroupAbstract
WHERE {
  ?study a disco:Study ;
    disco:inGroup ?studyGroup ;
    disco:variable ?variable .

  ?studyGroup dcterms:title ?studyGroupTitle .
  ?studyGroup dcterms:abstract ?studyGroupAbstract .

  ?variable disco:concept ?variableConcept .
  FILTER (regex(?variableConcept, "species", "i"))

  ?variable disco:baseOn ?variableDefinition .
  ?variableDefinition disco:concept ?variableDefinitionConcept .
  FILTER (regex(?variableDefinitionConcept, "Bufo alvarius", "i"))
}
```

D. Acknowledgements

This work 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” at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in September 2011 organized by Richard Cyganiak, Arofan Gregory, Wendy Thomas, and Joachim Wackerow. 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) 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” at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in October 2012
- 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 (Colectica, USA)
- Humphrey Southall (Department of Geography, UK Portsmouth University)
- Wendy Thomas (MPC - Minnesota Population Center, USA and DDI Alliance Technical Implementation Committee)
- Johanna Vompras (University Bielefeld Library, Germany)
- Joachim Wackerow (GESIS - Leibniz Institute for the Social Sciences, Germany and DDI Alliance Technical Implementation Committee)
- Benjamin Zapilko (GESIS - Leibniz Institute for the Social Sciences, Germany)
- Matthäus Zloch (GESIS - Leibniz Institute for the Social Sciences, Germany)

We would like to thank the following organizations which have supported this work:

- DDI Alliance
- GESIS - Leibniz Institute for the Social Sciences
- Schloss Dagstuhl - Leibniz Center for Informatics

E. References

E.1 Normative references

[ADMS]
Asset Description Metadata Schema (ADMS), *http://www.w3.org/TR/vocab-adms/*

[DCAT]
Data Catalog Vocabulary (DCAT), *http://www.w3.org/TR/vocab-dcat/*

[DCMI]
DCMI Metadata Terms (DCMI), *http://dublincore.org/documents/dcmi-terms/*

[DDI-RDF-Discovery-Vocabulary]
Bosch, T., Cyganiak, R., Gregory, A., Wackerow, J. *DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data.* 2013. Proceedings of the WWW2013 Workshop on Linked Data on the Web. URL: http://ceur-ws.org/Vol-996/papers/ldow2013-paper-12.pdf

[eXtended Knowledge Organization System]
Dan Gillman, Franck Cotton, and Yves Jaques *eXtended Knowledge Organization System (XKOS).* 2013. METIS, Work Session on Statistical Metadata. URL: http://www.unece.org/stats/documents/2013.05.metis.html

[FOAF]
Friend of a Friend (FOAF), *http://www.foaf-project.org/*

[FundRef]
FundRef, *http://www.crossref.org/fundref/*

[Linked-Statistical-Data]
Bosch, T., Cyganiak, R., Wackerow, J., and Zapilko, B. *Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences.* 2012. International Conference on Dublin Core and Metadata Applications. URL: http://dcpapers.dublincore.org/pubs/article/view/3654

[ORCID]
ORCID, *http://orcid.org/*

[ORG]
Organization Ontology (ORG), *http://www.w3.org/TR/vocab-org/*

[PROV-O]
PROV Ontology (PROV-O), *http://www.w3.org/TR/prov-o/*

[RDF Data Cube Vocabulary]
RDF Data Cube Vocabulary, *http://www.w3.org/TR/vocab-data-cube/*

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

[Semantic Statistics]
Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J. *Semantic Statistics: Bringing Together SDMX and SCOVO.* 2010. Proceedings of the WWW2010 Workshop on Linked Data on the Web. URL: http://ceur-ws.org/Vol-628/ldow2010_paper03.pdf

[SKOS]
Simple Knowledge Organization System (SKOS), *http://www.w3.org/2004/02/skos/*

[XKOS]
SKOS Extension for Statistics (XKOS), *http://htmlpreview.github.io/?https://github.com/linked-statistics/xkos/blob/master/xkos.html*

E.2 Informative references