



Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

Title	Modelling of Statistical Linked Data
Author(s)	Iqbal, Aftab
Publication Date	2011
Publication Information	Jindrich Mynarz and Richard Cyganiak and Aftab Iqbal and Michael Hausenblas (2011) Modelling of Statistical Linked Data Znalosti Slovakia,
Item record	<a href="http://www.deri.ie/sites/default/files/publications/mynarz_j_et_al_2011.pdf">http://www.deri.ie/sites/default/files/publications/mynarz_j_et_al_2011.pdf</a> ; <a href="http://hdl.handle.net/10379/4494">http://hdl.handle.net/10379/4494</a>

Downloaded 2024-03-20T11:02:25Z

Some rights reserved. For more information, please see the item record link above.



# Modelling of Statistical Linked Data

Jindřich Mynarz<sup>1</sup>, Richard Cyganiak<sup>2\*</sup>, Aftab Iqbal<sup>2\*\*</sup>, and Michael Hausenblas<sup>2\*\*\*</sup>

<sup>1</sup> National Technical Library,  
Technická 6,  
160 80, Praha 6 - Dejvice, Czech Republic  
`jindrich.mynarz@techlib.cz`

<sup>2</sup> Digital Enterprise Research Institute,  
NUI Galway, Lower Dangan  
Galway, Ireland

`*richard@cyganiak.de`, `**aftab.iqbal@deri.org`,  
`***michael.hausenblas@deri.org`

**Abstract.** There exists different approaches on publishing statistical data on the Web as linked data. In this paper we will provide a survey of existing approaches for expressing statistical data as linked data. The review of their advantages will be provided along with the description of how they compare to each other underscoring their main differences. This will include the discussion of the important steps of the data modelling process covering up best practices for dealing with legacy data.

## 1 Introduction

The aim of this paper is to compare the existing approaches for representing statistical data in RDF [KCM04] for linked data [Aye07] distribution. The statistical data we are about to discuss are the aggregated data that are based on *microdata* created in the process of data collection (e.g., survey data). We will focus on the statistical data in RDF exclusively.

Having data in RDF comes with a several benefits that differentiate it from non-RDF data formats. In RDF, the data is decoupled from the layout which means the layout has no effect on the interpretation of the data. In the case of having a dataset in a table, the layout has significant influence on the way the information in such a dataset can be read and interpreted. It is the layout of the rows and columns that to a great degree defines the interpretation of the data.

The clear separation of data and presentation is an important design feature of RDF. Compared to tabular data formats the interpretation of RDF does not rely on the data being properly laid out. This separation makes it possible to re-contextualize the dataset by embedding or intergrating it with another dataset and to build a novel applications which serve the data to end-users in their own way of presentation.

RDF is a flexible, schema-less data format. This helps to avoid the shortcomings of rigidly specified data formats which may be overly inclusive or overly

exclusive, and therefore result in a sparse matrix, or restricted expressivity respectively [PPT<sup>+</sup>01]. The flexible nature of RDF allows to avoid these shortcomings by having the data schema adapted to the requirements of a particular dataset.

The inclusiveness of RDF enables to combine and integrate RDF datasets together. This means that having the statistical data in RDF opens the possibility of connecting it with other, not necessarily statistical data, and in turn makes the data linkable and re-usable by others. In this way the data are made to be more web-friendly.

The publication of statistical data on the Web as linked data opens the access to them to a whole array of consumers for whom the established mechanisms of finding a way to statistical datasets are too complicated and unfamiliar. The dissemination standards in the field of statistics that are already in use serve well for the data exchange among the offices for national statistics, however, linked data publication model serves well to exchange data with the wider web community. The dissemination of the information based on the data harvested is one of the main goals of the offices of national statistics. In the light of this aim, linked data can be seen as another way of disseminating statistical data which is remarkable for the possibilities it opens on the side of accessing data.

The other benefit that RDF has with respect to exposing of the data is that it makes precise and complex queries feasible. SPARQL<sup>3</sup>, the RDF query language, provides a very flexible way of viewing the data on a level of high granularity and retrieving precisely defined subsets of the dataset. The affordance of efficient querying can be employed in producing flexible data transformations which serve as a basis for further re-purposing of the data.

The paper is structured as follows: in Section 2, we will briefly discuss the existing approaches for representing statistical data in RDF. Section 3 will address some key issues which are identified in existing approaches. Finally in section 4, we will conclude our paper.

## 2 State of the Art

To the current date there exists multiple approaches to the task of representing statistical data in RDF. Some of them are dataset-specific, others are based on a particular vocabulary that used to describe the data. In the following we will present an overview of the existing efforts in converting statistical data into RDF.

There were different approaches to produce RDF from datasets by Eurostat<sup>4</sup>, the aggregator of statistical data from the member countries of European Union. One of these approaches was developed at *Freie Universität Berlin*<sup>5</sup> as a wrapper for the Eurostat data in existing relational database using the D2R Server [BC06], which serves to expose such data as RDF on the Web. The second

---

<sup>3</sup> <http://www.w3.org/TR/rdf-sparql-query/>

<sup>4</sup> <http://ec.europa.eu/eurostat>

<sup>5</sup> <http://www.fu-berlin.de/en/>

approach in converting Eurostat data into RDF is *riese*<sup>6</sup> developed at *Joanneum Research*<sup>7</sup>. Riese stands for *RDFizing and Interlinking the EuroStat Data Set Effort* and it was an initiative with an aim of making the data coming from Eurostat available in RDF [HRH08]. The third approach on publishing Eurostat data into RDF is hosted at *OntologyCentral*<sup>8</sup>. The application that exposes this dataset acts as a wrapper that transforms the original data into RDF at real-time. Eurostat itself hosts an RDF export of the hierarchical list of *the Nomenclature of territorial units for statistics* from the year 2008, which contains a conceptualization of the geography of Europe, along with the dictionary of country codes, and a listing of Eurostat related legal acts.

One of the most extensive RDF conversions of statistical data was done with the U.S. Census dataset<sup>9</sup> of the year 2000. The result of this effort was one of the first datasets boasting with more than a billion RDF triples [TAU07].

Statistical data constitute a significant part of government linked data projects. It can be found in the U.S. **data.gov** datasets [VLH<sup>+</sup>10] or in the British initiative **data.gov.uk** [Ten09]. In these efforts, statistical data are used to describe various public service domains such as education, agriculture, or public finance. The path to **data.gov.uk** was paved by the previous successful project *EnAK-Ting*<sup>10</sup> that has also made available statistical datasets dealing with topics such as population, crime, or CO<sub>2</sub> emissions, for which it provides mash-ups with data visualizations<sup>11</sup>. There are also many datasets that have not originated from the field of official statistics, but have published data of a statistical kind. Among these are the *Linked Environment Data* dataset from *Federal Environment Agency* in Germany [BSCR10] and statistical data from an Italian university [Pir10]. The *Linked Environment Data* project combined SCOVO and SKOS vocabularies to express statistics about several projects of German Federal Environment Agency. In case of the Italian university statistics, an extended version of SCOVO was used to describe statistics about student activities at universities.

On the other hand there are approaches that are not only defined by one particular dataset that uses them, but they are formalized as stand-alone RDF vocabularies. These include most notably SCOVO<sup>12</sup> – The Statistical Core Vocabulary, and the Data Cube vocabulary [CRT10]. SDMX/RDF<sup>13</sup> also belongs to this type of vocabularies, but, in addition, it is based on an established standard from the field of statistics; SDMX – Statistical Data and Metadata Exchange.<sup>14</sup> These vocabularies fit into an evolutionary lineage that demonstrates the process of refining the tools for expressing statistical data in RDF. The vo-

<sup>6</sup> <http://riese.joanneum.at/about.html>

<sup>7</sup> <http://www.joanneum.at/>

<sup>8</sup> <http://ontologycentral.com/>

<sup>9</sup> <http://www.census.gov/>

<sup>10</sup> <http://www.enakting.org/>

<sup>11</sup> <http://www.enakting.org/gallery/index.html>

<sup>12</sup> <http://sw.joanneum.at/scovo/schema.html>

<sup>13</sup> <http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html>

<sup>14</sup> <http://sdmx.org/>

cabulary used in the *riese* project can be seen as a direct precursor of SCOVO, which in turn served as a basis for the SDMX/RDF and Data Cube vocabularies, which adopted the SCOVO’s basic concepts like modelling of the dimensions and observations as separate resources.

There are two main groups of vocabularies that were designed as general solutions for modelling statistical datasets in RDF: SCOVO and SCOVOLink,<sup>15</sup> and the most recent SDMX/RDF and Data Cube vocabularies.

SCOVO lays out statistical data in a data cube – a multi-dimensional logical space in which the observations are located. It was formulated as a simple, yet powerful, light-weight vocabulary, and this feature made it relatively easy to adopt. SCOVOLink is an extension of SCOVO that addresses the *domain semantics*, the way how a dataset refers to the things that it is about.

Data Cube vocabulary can be seen as an evolutionary step of SCOVO, from which it borrows the basic idea of a data cube with multiple dimensions, but it describes the cube with greater precision [CFG<sup>+</sup>10]. The design of Data Cube was informed by statistical expertise which makes it powerful enough to describe more complex data. SDMX/RDF is built on top of Data Cube and it represents a translation of the core parts behind the statistical standard SDMX to the RDF data format. It provides a layer over the Data Cube vocabulary that describes the *domain semantics*, dataset’s metadata, and additional information that is helpful in the exchange of statistical data.

The statistical datasets that are available in RDF have used a mixture of vocabularies, conversion or publishing mechanisms to convert the datasets into RDF. Table 1 gives an overview of the existing statistical datasets that are exposed in RDF and summarizes each of the examined datasets with its quantitative and qualitative parameters.

The approaches we have mentioned above have taken different decisions and standards on representing statistical data in RDF. In the following section we will address some key issues which we identified in existing approaches during the process of generating RDF out of the legacy statistical data.

### 3 Data Modelling

The most important thing during conversion of any data into RDF is to choose vocabularies and a way to model the data. There are generally two ways to address this. The preferred way is to pick and re-use one of the existing vocabularies that have been proven to work for such a purpose (e.g., SCOVO). However, if no vocabulary is available then one can create a vocabulary to address the needs of a particular data source. Such a vocabulary can be created either with an *ad hoc* design that aims to serve primarily for a concrete dataset or with long-term re-use in mind.

Based on the observation of existing approaches there have been recognized three choices with respect to the target that is being modelled [Fei08]. The

<sup>15</sup> <http://vocab.deri.ie/scovolink>

Dataset	Triples	Links	Domain	SPARQL endpoint	Vocabulary	Dump	Script	Year	Licence
World Factbook <sup>a</sup>	38640	n/a	geography, demography, finance	Yes	dataset specific	n/a	D2R Server	n/a	other
Eurostat (FU Berlin) <sup>b</sup>	8850	n/a	statistics	n/a	dataset specific	Yes	D2R Server	n/a	other
Riese <sup>c</sup>	5000000	100000	statistics	Yes	riese (dataset specific)	Yes	SWI Prolog + PHP	2008	other
U.S. Census data <sup>d</sup>	1002848918	n/a	demographics, geography	Yes	dataset specific	Yes	Perl	2007	Creative Commons Non-Commercial (Any)
Lotus <sup>e</sup>	586000000	311	government, university	Yes	SCOVO	Yes	Python	2010	n/a
OntologyCentral <sup>f</sup>	40000000	75	statistics	n/a	dataset specific	n/a	n/a	n/a	n/a
Eurostat <sup>g</sup>	19500	120	geography, government, law	n/a	dataset specific	Yes	n/a	2008	Creative Commons Attribution
EnAKTing NHS Dataset <sup>h</sup>	23828953	280	health	n/a	SCOVO	n/a	n/a	n/a	n/a
EnAKTing CO <sub>2</sub> Emission Dataset <sup>i</sup>	2316831	377	environment	n/a	SCOVO	n/a	n/a	n/a	n/a
EnAKTing Energy Dataset <sup>j</sup>	2316831	368	energy industry, transportation	n/a	SCOVO	n/a	n/a	n/a	n/a
EnAKTing Mortality Dataset <sup>k</sup>	12933	476	demographics	n/a	SCOVO	n/a	n/a	n/a	n/a
EnAKTing Population Dataset <sup>l</sup>	2316831	535	demographics	n/a	SCOVO	n/a	n/a	n/a	n/a
EnAKTing Crime Dataset <sup>m</sup>	n/a	n/a	crime	n/a	SCOVO	n/a	n/a	n/a	n/a
Analytics data.gov.uk Dataset <sup>n</sup>	865904	n/a	web analytics	Yes	SCOVO	n/a	n/a	n/a	Open Government Licence
Crime data.gov.uk Dataset <sup>o</sup>	41896	n/a	crime	Yes	SCOVO	n/a	n/a	n/a	Open Government Licence
Education data.gov.uk Dataset <sup>p</sup>	6619847	402076	education	Yes	SCOVO	n/a	n/a	n/a	Open Government Licence
Environment data.gov.uk Dataset <sup>q</sup>	185338	n/a	agriculture	Yes	SCOVO	n/a	n/a	n/a	Open Government Licence
Finance data.gov.uk Dataset <sup>r</sup>	14926	n/a	finance	Yes	SCOVO	n/a	n/a	n/a	Open Government Licence
Transport data.gov.uk Dataset <sup>s</sup>	329527661	909145	government, transportation	Yes	SCOVO	Yes	n/a	n/a	Open Government Licence
Statistics data.gov.uk Dataset <sup>t</sup>	343733	32284	geography, education, government	Yes	SCOVO	Yes	n/a	n/a	Open Government Licence

Table 1. Statistical datasets in RDF

- <sup>a</sup> <http://www4.wiwiw.fu-berlin.de/factbook/>  
<sup>b</sup> <http://www4.wiwiw.fu-berlin.de/eurostat/>  
<sup>c</sup> <http://riese.joanneum.at/data/>  
<sup>d</sup> <http://www.rdfabout.com/demo/census/>  
<sup>e</sup> <http://sw.unime.it/loius/>  
<sup>f</sup> <http://ontologycentral.com/2009/01/eurostat/>  
<sup>g</sup> <http://ec.europa.eu/eurostat/ramon/rdfdata/>  
<sup>h</sup> <http://nhs.psi.enakting.org/>  
<sup>i</sup> <http://co2emission.psi.enakting.org/>  
<sup>j</sup> <http://energy.psi.enakting.org/>  
<sup>k</sup> <http://mortality.psi.enakting.org/>  
<sup>l</sup> <http://population.psi.enakting.org/>  
<sup>m</sup> <http://crime.psi.enakting.org/>  
<sup>n</sup> <http://analytics.data.gov.uk/>  
<sup>o</sup> <http://crime.data.gov.uk/>  
<sup>p</sup> <http://education.data.gov.uk/>  
<sup>q</sup> <http://environment.data.gov.uk/>  
<sup>r</sup> <http://finance.data.gov.uk/>  
<sup>s</sup> <http://transport.data.gov.uk/>  
<sup>t</sup> <http://statistics.data.gov.uk/>

most straight-forward is the assumption that the statistical data models the real world. This design choice was adopted in *D2R Eurostat* dataset as its contents refer to real world entities. The second choice is to model only *a part of the real world* (i.e. domain) which is described in a dataset. The third choice is to choose *statistics* itself as the target being modelled. Approaches that adopt this choice will generally first build a model of statistics expressed with the *statistical artefacts* such as time, dimension, or table, and use it consequently to express statistics about the domain in question. Among the typical examples that apply this principle are the SCOVO and SDMX/RDF vocabularies.

Having decided on what is the object of modelling, there are mentioned two main distinct types of semantics that are modelled [HHR<sup>+</sup>09]. The first one is the *structural semantics* which takes into account how a structure within a dataset can be expressed using the means such as groups, slices, or aggregates. The second is the *domain semantics* that constitutes a part of the data model that enables to make claims about the topic the dataset is about.

The issue of *structural semantics* is addressed mostly in the latest practice for statistical datasets in RDF. SCOVO, for example, allows to group observation in a dataset via the `Dataset` class, but it lacks means to express structure inside of a dataset.

In contrast to this, SDMX/RDF and Data Cube vocabularies have strong expressive power when it comes to describing the internal structure of a dataset. Data Cube employs the notion of *slices* that allow to delimit a subset of a dataset for which some dimensions' can have fixed values. If the dimension without fixed value is *time*, the slice is referred to as *a time series*, otherwise it is called *a section*. Slices can be then organized by their relation to a dataset or by the relation they have to another slice (e.g., sub-slices).

The solutions that have been proposed for expressing *domain semantics* vary a great deal. This is heavily influenced by the choice the existing approaches made with respect to the target of their modelling, which constrains the possible means of expression for domain semantics. The core of this issue is the ability to model the domain that the statistical dataset is about and how to connect the observations in it with this model. In the common case, the domain of a statistical dataset consists of a set of real world objects; in the linked data terms *non-information resources*.

When creating the model of a domain the dataset is about, one can benefit from re-using the resources that are already exposed on the Web in the linked data fashion. This means it is possible to re-use such resource simply by using its URI and therefore build the description of a domain in question as a combination that merges in resources from external datasets. With the growth in size of the *linked data cloud*,<sup>16</sup> a few hubs that offers widely re-usable concepts have appeared, for example *DBPedia*,<sup>17</sup> an RDF version of the Wikipedia; or *Geonames*,<sup>18</sup> which supply concepts for geographic areas.

<sup>16</sup> <http://richard.cyganiak.de/2007/10/lod/>

<sup>17</sup> <http://dbpedia.org/>

<sup>18</sup> <http://www.geonames.org/>

The approaches that are discussed in this paper handles real world objects as separate resources. They identify them largely with URIs, even though in some cases *blank nodes* are used instead. Most of them are very brief about the type of the real world object and their expressive power for domain semantics is weak. Data Cube and SDMX/RDF vocabularies offer stronger means to describe the domain of a dataset. The approach employed in these vocabularies makes use of the parts of Simple Knowledge Organization System (SKOS)<sup>19</sup> to formalize the conceptualization of the domain in question. These vocabularies have taken over the `skos:Concept` class which is used to act as a substitute for the real world object the dataset describes. The concepts are grouped in *concept schemes* that serve as *codelists* from which the dataset’s dimensions draw on their values. Also, the dimensions have their own concepts that formulate what the dimension measures; for example, a dimension “year of age” is linked to the *concept* of “age”.

Besides the domain modelling, the gist of any statistical dataset is in the observed and aggregated values it contains. One of the distinctive factors that sets the existing approaches apart is the way they have chosen for modelling of such statistical observations. The main difference in this respect is the question whether to treat an observation as a separate resource, or if the observed values should be attached directly to the things they describe.

In some datasets, for example in the *D2R Eurostat* or *U.S. Census 2000* dataset, the observations are not resources on their own. Instead the observed values are attached directly to the proxies of real world objects from the domain described in the dataset.

The other option is to recognize observations as standalone resources. This is the way how it was done in the *OntologyCentral’s Eurostat* dataset where the observations are typed as instances of a class of observations that is defined for every dataset, for example a class for “GDP per capita in PPS”. Then there are the approaches that define observation as an instance of a special class which comes from the vocabulary rather than being re-defined for each of the datasets. Such classes are `riese:Item` in *riese*, `scv:Item` in SCOVO, or `qb:Observation` in Data Cube vocabulary.

Having described the general aspects of data modelling above we will now discuss some issues that one can encounters in the course of modelling statistical data in RDF. The following issues have been chosen because of their impact on the characteristics of a dataset.

### 3.1 Data Granularity

Data granularity defines the detail with which the data is sub-divided into parts. The issue of data granularity is important especially when it comes to using a dataset in machine extraction processes. In such cases it matters if the information that one wants to extract is in a separate node in the dataset’s structure, or if it is a part of another node. The information can be contained in literal

<sup>19</sup> <http://www.w3.org/2004/02/skos/>

values if the data is poorly structured. One such example is the value “78693011 m<sup>2</sup>” in the *U.S. Census 2000* dataset in which the unit of measurement is hidden inside the value instead of being expressed separately. Low data granularity implies less refining of the structure of dataset. This can be seen in the *riese* dataset where dates are treated as plain literals (e.g., “05/10/2007”) instead of making them adhere to XML Schema type and annotating them with the datatype property `xsd:type`.

High data granularity requires investing a lot of effort into the data modelling, which might turn out to be not the most efficient choice. The question how to choose an optimal data granularity is therefore a trade-off between the time spent on crafting the data structure and the benefits of affordances that highly granular data has.

### 3.2 Units of Measurement

As we have seen in the previous example, it is beneficial for the user of a dataset to have units of measurement expressed in a structured manner. Having explicitly stated units of measure that are used in a statistical dataset is a crucial requirement which is based on the necessity to enable comparing values across datasets. If the units of measurement are declared implicitly, the interpretation of the dataset becomes harder because it is not straight-forward to determine what is measured in a value.

In *D2R Eurostat* the units of measurement are part of the property names. For example, `eurostat:total.area.km2` property indicates that it uses *square kilometer* as a unit of measurement which is derived from the label of the property. The *U.S. Census 2000* dataset uses both untyped values and values with information about the units of measurement, e.g., “78693011 m<sup>2</sup>” where “m<sup>2</sup>” stands for square metre. This issue can also be illustrated on the *OntologyCentral’s Eurostat* dataset where one can find values such as “117 b” or “798045 p” for which the only possibility to determine what “b” or “p” stands for is to carefully examine the context surrounding such values. These implicit units of measurements are used inconsistently which results in a state in which some values do not have any unit specified, whereas others have it declared by the `:unit` property.

On the other side, there is the more explicit approach that has been chosen by the Data Cube and SDMX/RDF vocabularies. The units of measurements are modelled as separate resources and attached via the `sdmx-attribute:unitMeasure` property which points to a concept that serves as the unit of measurement. In this way the values in a dataset can be compared easily with each other.

### 3.3 Time

When it comes to modelling of *the time*, a clear distinction between the different approaches of expressing statistical data in RDF can be observed. The first group, including *D2R Eurostat* and *U.S. Census 2000* dataset, excludes the

dimension of time. Instead, in the case of *D2R Eurostat*, only the latest values are provided, and the *U.S. Census 2000* dataset makes the implicit assumption that all its contents date to the year 2000. The second group, including *riese* and SDMX/RDF, treats time as an individual dimension which allows it to cover up observations for different times in a time series. In *riese* the `dimension:Time` is used, in SDMX/RDF there is a specifically tailored `sdmx:TimeRole` concept denoting that the dimension's concept expresses time while the dimension itself may be expressed using `sdmx-dimension:refPeriod` property and the like. These means makes the vocabulary able to describe data that are bound to a particular period of time and produce a *time series*.

### 3.4 Identifiers

An issue which is particularly relevant from the linked data perspective is the choice of identifiers which are used to represent a resource in the dataset. Providing every resource with a URI is a fundamental linked data design principle.<sup>20</sup> Linked data publishing model recommends HTTP URIs. HTTP URIs utilize HTTP schema which allows them resolvable by any HTTP agent. URIs are globally unique identifiers which means that they can be used to identify a resource from any dataset.

Having a URI for every resource contained within a dataset makes specific references possible which can be then used to track the provenance of the source information it refers to. Also, a resource can be linked by its URI from external datasets. Vocabularies such as SCOVO, SDMX/RDF, or Data Cube are built with this requirement in mind and recommends the usage of URIs for every resource described in a dataset.

However, representing the resources in the dataset with URIs is not a matter of fact but rather a design choice. Most of the statistical datasets that we observed uses HTTP URIs. However, it is not a matter of course, for example in the *U.S. Census 2000* dataset non-HTTP URIs are employed (e.g., `<tag:govshare.info,2005:data/us>`).

The design of RDF also allows to use *blank nodes* as resources' identifiers. But in fact, blank nodes can be referred exclusively from the dataset in which they are minted because there is no way to resolve them on the global scale. This means blank nodes are not well suited for the linked data distribution of data and they are best used only for the resources which the dataset's producer does not expect to be re-used. This approach is employed in the *U.S. Census 2000* dataset.

**URI Patterns** One part of the data modelling is the design of URI patterns for the types of resources that have been established within a dataset. A URI pattern prescribes a template that declares how is the URI structured. The importance of URI patterns stems from seeing humans as their users. In a sense,

<sup>20</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

URIs following known URI patterns can serve as a simple query language to retrieve resources from a dataset. When one discovers a structure in a URI (e.g., a hierarchy) this insight can be used to modify the URI to get to another resource. For this particular reason are URI patterns therefore designed to produce *human readable* URIs. The recommended approach is to agree on a set of URI patterns within a specific sector [psa09]. Standardizing on a set of URI patterns among the offices of national statistics has another benefits for the user of a dataset. If there is a generally accepted way to structure URIs for certain types of resources, similar URI can be issued against multiple datasets to obtain similar results, i.e., the same type of statistical observations.

One of the most common practices in the design of URI patterns is to cluster URIs by the type of resource they identify. When URIs for a certain resource type follow a URI pattern the resources belonging to this type can be referenced consistently. For example, geographic areas can be grouped in one namespace, so their URI start with the same base URI that may be followed by the type of the area (e.g., “city”) and the name of the city after that (e.g., “dublin”).

A common URI pattern is to mimic the hierarchy of the resources that represent real world things in the URI. This may work well for the geographic areas where their URIs may include a path made of the broader areas. This practice can be found in the *the U.S. Census 2000* dataset or *OntologyCentral’s Eurostat* to name a few examples.

One of the design patterns that are may be used for the URIs of observations that make up the dataset is to put the values for each dimension in the URI. This reflects the location of the observation in the logical space of the dataset and produces a human readable URIs given that the dimensions’ values and well encoded. For example, a variation of this design was adopted by the *riese* dataset.

### 3.5 Re-use

The use of globally unique identifiers in RDF encourages design of data structures with re-use in mind. With such identifiers it is possible to repurpose existing resources in the design stage of a data model. For example, SDMX/RDF vocabulary provides a set of re-usable properties or concepts that is based on the *SDMX Content-oriented guidelines* that can be used as parts of a data structure definition [SDM09].

In some cases, a resource is similar or somehow related to another resource which can be expressed as a typed link that qualifies the relation between both resources. There are a number of different ways to describe the similarity between resources: the strongest one being the `owl:sameAs` property which expresses that the resources are equal. Less rigorous are the properties in the SKOS vocabulary that enable to specify how closely the resources match (`skos:exactMatch`, `skos:closeMatch`) with one another, or the ones that imply there is a hierarchical (`skos:broadMatch`, `skos:narrowMatch`) or an associative relationship (`skos:relatedMatch`).

This feature demonstrates the flexibility of RDF and entails that one can adopt only those parts of a vocabulary that are needed for a particular dataset. On the other hand it means that the schemas or vocabularies created as a mesh by linking to existing components can be used as a source to yet another purposes as well. By designing the schema with potential re-use in mind, one can increase the probability of establishing widely used standard for a particular domain.

## 4 Conclusion

In this paper we covered the current practices for representing statistical data in RDF. We have shown how one can tackle the common issues in the conversion of legacy statistical data into linked data while highlighting the main differences between the approaches that have been developed so far.

Having statistical data publicly available endows everyone with the ability to explore the data on their own. Users do not have to rely on the official reports produced by the departments of national statistics as they can yield useful insights directly from the data. Linked data is a technology that makes a step in this direction by making the statistical data available on the Web in a *web-friendly* way. It is also an opportunity for the field of statistics to join the greater Web community by integrating statistical with non-statistical data.

## 5 Acknowledgements

The work of Jindřich Mynarz is partially supported by CSF grant no. P202/10/0761, Web Semantization.

## References

- [Aye07] D. Ayers. Evolving the Link. *IEEE Internet Computing*, 11(3):94–96, 2007.
- [BC06] Christian Bizer and Richard Cyganiak. D2R server - publishing relational databases on the semantic web. In *Poster at the 5<sup>th</sup> International Semantic Web Conference*, 2006.
- [BSCR10] Thomas Bandholtz, Till Schulte-Coerne, and Maria Rüther. Linked Environment Data: SCOVO-fying the environment specimen bank. 2010. <http://www.w3.org/egov/wiki/images/8/85/Isem2010-bandholtz.pdf>.
- [CFG<sup>+</sup>10] Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic statistics : bringing together SDMX and SCOVO. In *Linked Data on the Web 2010 : CEUR workshop proceedings*, 2010. [http://events.linkedata.org/ldow2010/papers/ldow2010\\_paper03.pdf](http://events.linkedata.org/ldow2010/papers/ldow2010_paper03.pdf).
- [CRT10] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The RDF Data Cube vocabulary. 2010. Last update 2010-07-14. <http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html>.
- [Fei08] Lee Feigenbaum. Modelling statistics in RDF : a survey and discussion. 2008. [http://www.thefigtrees.net/lee/blog/2008/03/modeling-statistics\\_in\\_rdf\\_a\\_s.html](http://www.thefigtrees.net/lee/blog/2008/03/modeling-statistics_in_rdf_a_s.html).

- [HHR<sup>+</sup>09] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. Scovo: Using statistics on the web of data. In *ESWC 2009 Heraklion: Proceedings of the 6<sup>th</sup> European Semantic Web Conference on The Semantic Web*, pages 708–722, Berlin, Heidelberg, 2009. Springer-Verlag.
- [HRH08] Wolfgang Halb, Yves Raimond, and Michael Hausenblas. Building linked data for both humans and machines. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008), (Beijing, China)*, 2008.
- [KCM04] G. Klyne, J. J. Carroll, and B. McBride. Resource Description Framework (RDF): Concepts and Abstract Syntax). W3C Recommendation 10 February 2004, RDF Core Working Group, 2004.
- [Pir10] Giovanni Pirrotta. Linking Italian university statistics. In *I-SEMANTICS '10: Proceedings of the 6<sup>th</sup> International Conference on Semantic Systems*, pages 1–10, New York, NY, USA, 2010. ACM.
- [PPT<sup>+</sup>01] Haralambos Papageorgiou, Fragkiskos Pentaris, Eirini Theodorou, Maria Vardaki, and Michalis Petrakos. Modelling statistical metadata. In *Proceedings of the 13<sup>th</sup> International Conference on Scientific and Statistical Database Management*, pages 25–35. IEEE, 2001.
- [psa09] Designing URI sets for the UK public sector : a report from the Public Sector Information Domain of the CTO Council's Cross-Government Enterprise Architecture. 2009. [http://www.cabinetoffice.gov.uk/media/301253/public\\_sector\\_uri.pdf](http://www.cabinetoffice.gov.uk/media/301253/public_sector_uri.pdf).
- [SDM09] SDMX. SDMX Content-oriented guidelines. 2009. [http://sdmx.org/wp-content/uploads/2009/01/00\\_sdmx\\_content-oriented\\_guidelines\\_2009.pdf](http://sdmx.org/wp-content/uploads/2009/01/00_sdmx_content-oriented_guidelines_2009.pdf).
- [TAU07] Joshua TAUBERER. The 2000 U.S. Census : 1 billion RDF triples. 2007. <http://www.rdfabout.com/demo/census/>.
- [Ten09] Jeni Tennison. Expressing statistics with RDF. 2009. <http://www.jenitennison.com/blog/node/132>.
- [VLH<sup>+</sup>10] Denny Vrandečić, Christoph Lange, Michael Hausenblas, Jie Bao, and Li Ding. Semantics of governmental statistics data. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.