Title	SCOVO: Using Statistics on the Web of Data
Author(s)	Hausenblas, Michael
Publication Date	2009
Publication Information	Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, Danny Ayers "SCOVO: Using Statistics on the Web of Data", Semantic Web in Use Track of the 6th European Semantic Web Conference (ESWC2009), 2009.
Item record	http://hdl.handle.net/10379/443

SCOVO: Using Statistics on the Web of Data

Michael Hausenblas[1], Wolfgang Halb[2], Yves Raimond[3], Lee Feigenbaum[4], and Danny Ayers[5]

[1] DERI, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
michael.hausenblas@deri.org
[2] Institute of Information Systems & Information Management,
JOANNEUM RESEARCH, Steyrergasse 17, 8010 Graz, Austria
wolfgang.halb@joanneum.at
[3] BBC Audio & Music interactive
London, United Kingdom
yves.raimond@bbc.co.uk
[4] Cambridge Semantics,
PO Box 425003, Cambridge, MA 02142, United States
lee@cambridgesemantics.com
[5] Talis Ltd,
Knights Court, Birmingham Business Park, B37 7YB, United Kingdom
danny.ayers@talis.com

Abstract. Statistical data is present everywhere—from governmental bodies to economics, from life-science to industry. With the rise of the Web of Data, the need for sharing, accessing, and using this data has entered a new stage. In order to enable proprietary, closed-world formats, to enter the Web of Data, we propose a framework for modelling and publishing statistical data. To illustrate the usefulness of our approach we demonstrate its application in real-world statistical datasets.

1 Motivation

Statistical data is present everywhere—from governmental bodies to economics, from life-science to industry. With the rise of the Web of Data, the need for sharing, accessing, and using this data has entered a new stage. Available technologies are either not compatible with the Semantic Web or are too complex to be useful for a range of use cases. We have identified the need for a more general and flexible solution to the problem of modelling and publishing statistics on the Web.

Our motivation stems from ongoing work in three distinct efforts, namely riese ("RDFizing and Interlinking the EuroStat Data Set Effort") [11, 12], U.S. Census Bureau's annual Statistical Abstract of the United States, and making the UN data accessible on the Web of Data. In riese, we provide statistical data about European citizens. One of our main use cases of riese was in the context of an advertising analysis application [23] allowing for example a market researcher to better and faster understand a certain market or product. Further, in the U.S.

Census Bureau's annual Statistical Abstract of the United States case, one of the authors and representatives of the U.S. Environmental Protection Agency were exploiting the semantics implicit in spreadsheets published yearly since 1878 by the U.S. Census Bureau. This data corpus is a comprehensive collection of social, political, and economic statistics compiled from information from over 250 agencies. The goal here was to enable a fast and efficient publishing of the statistics on the Web. Currently, this means making MS Excel documents and PDF documents available for download from the Web.

The paper is structured as follows: Firstly, we review related efforts in section 2. Then, in section 3 we discuss issues with representing statistical data and derive requirements for a generic modelling. The core of the work is presented in section 4—where we propose a modelling framework for statistical data—and section 5 in which two reference implementations are discussed. Finally, we conclude our work and outline future steps in section 6.

2 Related Work

Representing statistical data has a long tradition, hence a plethora of proposals and solutions exists [3]—mainly driven by governmental and international institutions dealing with high volumes of data. In the 1990's the U.S. Bureau of the Census has developed a statistical metadata content standard, which allows to describe all aspects of survey design, processing, analysis, and data sets [15]. Later, the "United Nations Economic Commission for Europe" (UN-ECE) has developed guidelines [27] covering search, navigation, interpretation, and post-processing of statistical data in their realm. A standard proposed by the International Organization for Standardization (ISO) is the "Statistical data and metadata exchange" (SDMX) [14]. More recently, the OECD has released a report on the management of statistical metadata at the OECD [18]. Another related effort is the Data Documentation Initiative (DDI)[6], which aims at establishing an XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioural sciences. As we have already pointed out in [11], there are known attempts concerning the modelling and use of statistical data on the Web of Data [4, 10, 24, 25]. However, unlike earlier attempts such as [16, 19], we aim at a light-weight solution enabling a quick uptake and wide deployment.

The **Web of Data** is understood as the part of the Web where the linked data principles are applied. The basic idea of linked data was outlined by Sir Tim Berners-Lee [5]. The Linking Open Data (LOD) community project[7] is an open, collaborative effort applying the linked data principles. It aims at bootstrapping the Web of Data by publishing datasets in RDF on the Web and creating large numbers of links between these datasets [6].

We finally highlight the modelling issue with n-ary relations on the Web of Data. The data model of the Web of Data is RDF, hence modelling n-ary

[6] `http://www.ddialliance.org/`
[7] `http://linkeddata.org/`

relations is a non-trivial task. In 2006 the W3C Semantic Web Best Practices and Deployment Working Group has published a note dealing with this issue [17]; we will use this as a base for our framework.

3 Requirements and Issues

3.1 Issues with Modelling Statistical Data

Independent of the original format of the data (such as a table in an Excel sheet, etc.), the issues discussed in the following need to be addressed properly to ensure a lossless representation of the statistical data.

Handling of Multiple Dimensions It is very often the case that a data item has several dimensions. Roughly, two types of dimensions can be identified, (i) generic dimensions, such as location and time, and (ii) domain specific dimensions. For example we may be interested in the reliability of flights (domain specific) from Cambridge, MA, US to London, UK (both are locations) between 2003 and 2007 (time period). It is crucial that a vocabulary aiming at representing statistics is capable of denoting such dimensions and allows to attach as many as needed to a single data item.

Reusability and Uptake Statistics are no ends in themselves; rather they are **about** something—be it money, flights, death rates or the consumption of YouTube videos. It is therefore essential being able to reuse existing information; both on the schema as on the instance level. Related to reusability is the issue of community uptake: Most statistical metadata formats are rather complex, yielding a small deployment.

Structural vs. Domain Semantics Two kinds of semantics come into mind when modelling statistics:

- structural semantics, stemming from how statistics are presented, such as grouping into time-periods, primary dimensions, etc.;
- domain semantics, stemming from the domain the statistic is about (money, airports, etc.).

Performance and Scalability Issues As discussed elsewhere [13] performance and scalability issues may arise from the way data is encoded and served.

3.2 Requirements

Based on the issues listed above we state the following requirements for our framework:

1. The framework *must* be directly usable on the Web of Data. This implies for example that a vocabulary used in the framework must be expressed in RDF. This requirement addresses the issue of structural and domain semantics, as well as ensuring reusability;

2. The framework *must* be extensible both on the schema level and the instance level, enabling reusability;
3. The framework *should* be light-weight, addressing uptake, and performance and scalability issues.

It has to be noted that the first requirement does not stem from the issues discussed earlier—it was rather introduced to benefit from the large deployment base of domain vocabularies[8] as well as available tools and systems[9].

4 Statistical Modelling Framework

Driven by the requirements we propose a modelling and publishing framework for statistics on the Web of Data consisting of:

- a core vocabulary for representing statistical data
- a "workflow" to create the statistical data

The framework is depicted at a glance in Fig. 1. The lower half is the generic part, defined in this work, the upper part is the application-specific part, depending on the technologies used to implement the framework.

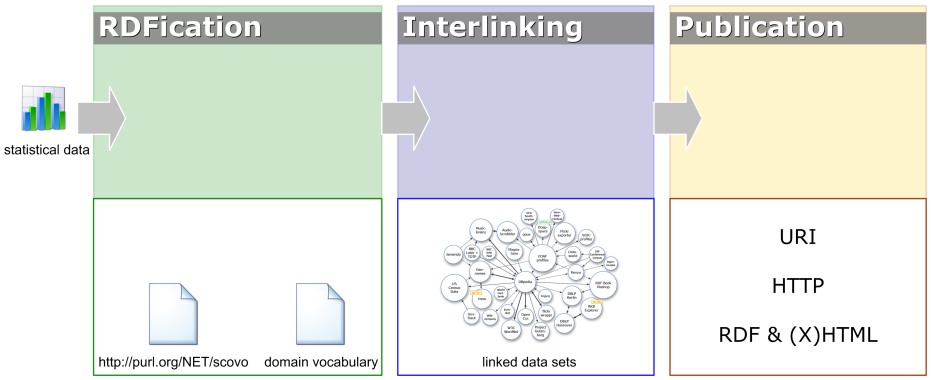

Fig. 1. The Statistical Modelling Framework

4.1 Statistical Core Vocabulary (SCOVO)

One of the main contributions of our work at hand is the Statistical Core Vocabulary (SCOVO)[10]. SCOVO defines three basic concepts:

[8] `http://schemacache.test.talis.com/`
[9] `http://esw.w3.org/topic/SemanticWebTools`
[10] `http://purl.org/NET/scovo`

– a dataset, representing the container of some data, such as a table holding some data in its cells;
– a data item, representing a single piece of data (e.g. a cell in a table);
– a dimension, representing some kind of unit of a single piece of data (for example a time period, location, etc.)

A statistical dataset in SCOVO is represented by the class `Dataset`; it is a SKOS concept [22] in order to allow hooking into a categorisation scheme. A statistical data item `Item` belongs to a dataset (cf. inverse properties `dataset` and `datasetOf`). An `Item` is subsuming the `Event` concept, as defined in the Event ontology[11]. The Event ontology essentially adopts the view from Allen and Fergusson [2]:

> [..] events are primarily linguistic or cognitive in nature. That is, the world does not really contain events. Rather, events are the way by which agents classify certain useful and relevant patterns of change.

An event is then defined in this ontology as the way by which cognitive agents classify arbitrary time/space regions. Our `Item` concept is subsuming this `Event` concept—a statistical item is a particular classification of a time/space region. Dimensions of a statistical item are factors of the corresponding events, attached through the `dimension` property, pointing to an instance of the SCOVO `Dimension` class.

This model is easily extensible by defining new factors and agents pertaining to the actual statistical data. For example, we can relate to a statistical data item the institutional body responsible of it as well as the methodology used. A `Dimension` can have a minimum (and respectively a maximum) range value, captured through the `min` and `max` properties.

The Statistical Core Vocabulary (depicted in Fig. 2) is currently defined in RDF-Schema. It is possible to express SCOVO in OWL-DL, if advanced reasoning is of necessity. Although we have depicted the range of both `:min` and `:max` in Fig. 2 being of literal value, we emphasize that in the RDF-Schema the ranges have not been specified in order to allow an extension for domain-specific purposes. Hence, this can be seen as a kind of recommendation for the default case.

Example To demonstrate the usage of SCOVO, let us assume we want to model airline on-time arrivals and departures. The input in our example is the "Table 1047. On-Time Flight Arrivals and Departures at Major U.S. Airports: 2006"[12] (cf. Fig. 3) from the US Census data set. Every airport, for each time period has an on-time arrival percentage and an on-time departure percentage.

[11] `http://purl.org/NET/c4dm/event.owl`
[12] `http://www.census.gov/compendia/statab/tables/08s1047.xls`

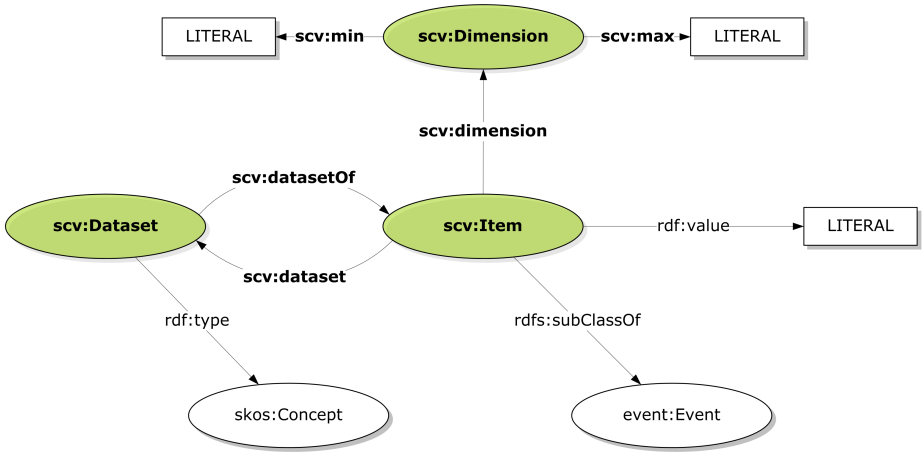

Fig. 2. The Statistical Core Vocabulary (SCOVO).

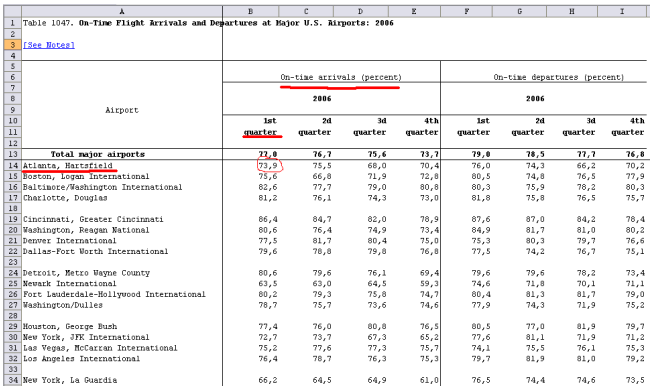

Table 1047. **On-Time Flight Arrivals and Departures at Major U.S. Airports: 2006**

[See Notes]

Airport	On-time arrivals (percent) 2006 1st quarter	2d quarter	3d quarter	4th quarter	On-time departures (percent) 2006 1st quarter	2d quarter	3d quarter	4th quarter
Total major airports	**72,8**	**76,7**	**75,6**	**73,7**	**79,0**	**78,5**	**77,7**	**76,8**
Atlanta, Hartsfield	73,9	75,5	68,0	70,4	76,0	74,3	66,2	70,2
Boston, Logan International	75,6	66,8	71,9	72,8	80,5	74,8	76,5	77,9
Baltimore/Washington International	82,6	77,7	79,0	80,8	80,3	75,9	78,2	80,3
Charlotte, Douglas	81,2	76,1	74,3	73,0	81,8	75,8	76,5	75,7
Cincinnati, Greater Cincinnati	86,4	84,7	82,0	78,9	87,6	87,0	84,2	78,4
Washington, Reagan National	80,6	76,4	74,9	73,4	84,9	81,7	81,0	80,2
Denver International	77,5	81,7	80,4	75,0	75,3	80,3	79,7	76,6
Dallas-Fort Worth International	79,6	78,8	78,8	76,8	77,5	74,2	76,7	75,1
Detroit, Metro Wayne County	80,6	79,6	76,1	69,4	79,6	79,6	78,2	73,4
Newark International	63,5	63,0	64,5	59,3	74,6	71,8	70,1	71,1
Fort Lauderdale-Hollywood International	80,2	79,3	75,8	74,7	80,4	81,3	81,7	79,0
Washington/Dulles	78,7	75,7	73,6	74,6	77,9	74,3	71,9	75,2
Houston, George Bush	77,4	76,0	80,8	76,5	80,5	77,0	81,9	79,7
New York, JFK International	72,7	73,7	67,3	65,2	77,6	81,1	71,9	71,2
Las Vegas, McCarran International	75,2	77,6	77,3	75,7	74,1	75,5	76,1	75,3
Los Angeles International	76,4	78,7	76,3	75,3	79,7	81,9	81,0	79,2
New York, La Guardia	66,2	64,5	64,9	61,0	76,5	74,4	74,6	73,5

Fig. 3. On-Time Flight Arrivals and Departures at Major U.S. Airports: 2006.

In listing 1 an excerpt[13] of the modelling of the airline on-time arrivals and departures is shown[14]. We note that the example has `http://example.org/on-time-flight#` as its base along with the prefix `ex:` (line 1). Lines 4 to 16 define domain-specific entities, such as a `ex:TimePeriod`, an `ex:Airport`, etc. From line 18 to 22 the one and only dataset (`ex:ontime-flights`) is defined, corresponding to the entire Excel table used as an input. Further, from line 24 on an exemplary data item is defined, stating that the on-time arrival of the

[13] The full example in RDF/XML is available at `http://sw.joanneum.at/scovo/otf-example-full.rdf`

[14] Note, that in all of our examples the well-known prefixes such as rdf:, rdfs:, etc. have been omitted due to readability reasons.

```
@prefix ex: <http://example.org/on-time-flight#> .
@prefix scv: <http://purl.org/NET/scovo#> .

ex:TimePeriod rdfs:subClassOf scv:Dimension; dc:title "time period" .

ex:Q12006 rdf:type ex:TimePeriod; dc:title "2006 Q1";
          scv:min "2006-01-01"^^xsd:date ;
          scv:max "2006-03-31"^^xsd:date .

ex:OnTime rdfs:subClassOf scv:Dimension; dc:title "on-time ..." .

ex:ota rdf:type ex:OnTime; dc:title "on-time arrivals" .

ex:Airport rdfs:subClassOf scv:Dimension; dc:title "airport" .

ex:AtlantaHartsfield rdf:type ex:Airport; dc:title "Atlanta, ..." .

ex:ontime-flights rdf:type scv:Dataset ;
                  dc:title "On-time Flight Arrivals ..." ;
                  scv:datasetOf ex:atl-arr-2006q1 .

ex:AtlantaHartsfield-ota-2006-q1 rdf:type scv:Item ;
                                 rdf:value 74 ;
                                 scv:dataset ex:ontime-flights ;
                                 scv:dimension ex:Q12006 ;
                                 scv:dimension ex:ota ;
                                 scv:dimension ex:AtlantaHartsfield .
```

Listing 1. Modelling flight on-time arrival statistics.

"Atlanta, Hartsfield" airport in the first quarter of 2006 was round 74%. This corresponds to the highlighted cell in Fig. 3.

```
1  SELECT ?airport_name ?percent_ontime ?period ?ontime_type WHERE {
2    ?item rdf:type scv:Item ;
3          scv:dimension ?airport ;
4          scv:dimension ?time_period ;
5          scv:dimension ?ontime ;
6          rdf:value ?percent_ontime .
7    ?airport rdf:type ex:Airport; dc:title ?airport_name .
8    ?time_period rdf:type ex:TimePeriod; dc:title ?period .
9    ?ontime rdf:type ex:OnTime; dc:title ?ontime_type .
10   FILTER (?percent_ontime > 85)
11 } ORDER BY DESC (?percent_ontime)
```

Listing 2. SPARQL query for high-performing airports.

The SPARQL query from listing 2 can be used to explore high-performing airports. With "high-performing" we define in this context airports with on-time arrivals or departures higher than 85%. Lines 2 to 6 provide the generic pattern for an `Item`. Note how in line 7 and 8 the dimensions are constrained. Line 9 of listing 2 basically expresses "give me all kinds of on-timeness", and finally line 10 implements the "high-performance" filter criteria.

airport_name	percent_ontime	period	ontime_type
Cincinnati, Greater Cincinnati	88	2006 Q1	on-time departures
Salt Lake City International	88	2006 Q2	on-time departures
Cincinnati, Greater Cincinnati	87	2006 Q2	on-time departures
Salt Lake City International	87	2006 Q3	on-time departures
Salt Lake City International	86	2006 Q2	on-time arrivals
Portland International	86	2006 Q3	on-time departures
Cincinnati, Greater Cincinnati	86	2006 Q1	on-time arrivals
Portland International	86	2006 Q2	on-time departures

Fig. 4. Results for high-performing airports.

The query result—depicted in Fig. 4—shows the list of high-performing airports along with the time period, starting with the best airport in terms of "on-timeness". We note that the complete example, including the exemplary queries in an executable form, is available at `http://purl.org/NET/scovo`.

4.2 Workflow—Good Practice Rules

In this section we discuss the overall workflow as shown in Fig. 1. Based on our findings from publishing real-world statistical datasets, the following should be seen as strong advises, helping to avoid failings and to enable a quick adoption.

RDFication In the very first step, the data needs to be converted into an RDF-based form. This is equally true for the schema level as for the instance level. The schema level (e.g. XSD, etc.) is a typical starting point which is followed by the conversion of the actual data in a second step. While creating and populating the ontology with instances several issues arise.

URI Design It has to be ensured that every entity has a URI assigned, which is usually referred to as "URI minting"; cf. [7] for a more detailed discussion on URI design. For example `http://dbpedia.org/resource/Airport` has been minted by DBpedia to represent the concept of an airport.

More specifically we recommend using HTTP URIs in order to be compliant with the linked data principles: When dereferencing the aforementioned URI for "Airport" it yields

```
curl -I http://dbpedia.org/resource/Airport
HTTP/1.1 303 See Other
Server: Virtuoso/05.00.3028 (Solaris) x86_64-sun-solaris2.10-64  VDB
Content-Type: text/html; charset=UTF-8
Date: Thu, 08 May 2008 10:32:29 GMT
Location: http://dbpedia.org/page/Airport
```

basically telling us that the "concept URI" redirects to an information resource at `http://dbpedia.org/page/Airport`, see also [21].

Domain Ontologies As already mentioned, statistics are always about a certain domain. In order to use domain vocabularies together with SCOVO, several "hooks" can be used:

- Subclassing the SCOVO-`Dimension` class. In most cases it is sufficient to use this technique to incorporate domain-specific concepts, for example

  ```
  ex:Airport rdfs:subClassOf scv:Dimension
  ex:AtlantaHartsfield rdf:type ex:Airport
  ```

  from the example in listing 1;
- Use the built-in support for `event:Event` and `skos:Concept`. The latter is of particular help if an existing taxonomy or thesaurus is used as a base. The earlier can be used to capture more information pertaining to the creation of a particular statistical item;
- Defining sub-properties of using SCOVO-`min` and `max`. Whenever the need arises to more explicitly declare what kind of range is intended, this technique can be used (e.g. an `xsd:date`).

Interlinking Classes and instances of the domain vocabulary *should* be interlinked to existing LOD entities. The rationale behind is that any dataset can be enriched through this at low costs. For example, to connect the airports to the LOD datasets, one could use the query from listing 3 to find according targets in DBpedia (note that this query can be executed at `http://dbpedia.org/snorql/`).

```
SELECT ?airports_state ?airport
WHERE {
 ?airports_state skos:broader
 <http://dbpedia.org/resource/Category:Airports_in_the_United_States> .
 ?airport skos:subject ?airports_state ;
          <http://dbpedia.org/property/name> ?name .
 FILTER regex(?name, "Atlanta", "i")
}
```

Listing 3. Interlinking airports to DBpedia.

The result of the query from listing 3 may subsequently be used to enrich our example, that is adding for example the triple

```
ex:AtlantaHartsfield owl:sameAs
<http://dbpedia.org/resource/Hartsfield-Jackson_Atlanta_International_Airport>
```

in order to express that the two URIs are actually identifying the same thing. With this interlinking we have significantly broadened the possibilities for querying our dataset; for example we could issue a location-based query with the geo-data from DBpedia or could further follow down the path to other LOD datasets containing even more information related to the Hartsfield airport.

Publication When publishing the dataset, one needs to make choices on the formats to be used for the data. While certain circumstances may require the usage of specialized and/or proprietary formats such as PDF or the SPSS file format, there are four basic technologies that we (unsurprisingly) see central to our setup: URIs, HTTP, RDF and (X)HTML; every publishing system on the Web of Data *should* use these as primary technologies. We have discussed URI minting above. Regarding HTTP—beside its basic transport function—we encourage people to use light-weight REST interfaces. One particular issue, however, is how to deploy the metadata. Several options exist, we list some widely used in the following:

- use an RDF standalone format such as RDF/XML, N3, etc. along with 303 redirects or links such as described in [21];
- use XHTML+RDFa[15] for both humans and machines (see also [11]);
- SPARQL-endpoints and RDF dumps [20, 13].

Note that in practice very often the approaches listed above are used in combination. For example offering an RDF dump (in N-Triples) for semantic search engines such as Sindice [26] along a SPARQL-endpoint for cross-site query is a typical pattern.

To allow semantic search engines to efficiently and effectively process the dataset it is advisable to use proper announcement mechanisms such as the semantic crawler sitemap extension protocol [8].

4.3 Comparison with other approaches

The following table presents a comparison between three different approaches for modelling statistics in RDF. The comparison is based on [16] and highlights some differences between the modelling from the D2R Server for Eurostat[17], the 2000 U.S. Census Data [25], and SCOVO itself.

The most distinguishing feature of SCOVO is the ability to express complex statistics over time while still keeping the structural complexity very low. Both other approaches are not capable of representing historical data and only provide statistics for one point-in-time. From the table below we conclude further that SCOVO seems to be the best combination of flexibility and usability, allowing to recreate the data-table structures with a reasonable degree of fidelity in another environment (that is, on the Web). Additionally, in our understanding SCOVO is more aligned with the linked data principles, compared to D2R Eurostat and 2000 U.S. Census.

[15] `http://www.w3.org/TR/rdfa-syntax/`
[16] `http://www.thefigtrees.net/lee/blog/2008/03/modeling_statistics_in_rdf_a_s.html`
[17] `http://www4.wiwiss.fu-berlin.de/eurostat/`

	D2R Eurostat	2000 U.S. Census	Scovo
Expressivity	Simple, limited	Complex	Complex
Modelling of time	Point-in-time*	Point-in-time	Over time
Historic data	No*	No	Yes
Easy table (re-)generation	No	Yes	Yes
Access to actual statistical data	Using individual predicates	Using individual predicates	Value attached to items
Location of classifying features	Concatenated in the name of the predicate	Concatenated in the name of the predicate and a chain of predicates	Attached to items
Structural complexity	Flat model	Complex, related via individual predicates; sometimes inconsistent	Relatively flat model; all information attached to item; datasets as container
Use of blank nodes	No	Yes	No
Different predicates	Many	Many	Few
Knowledge needed for query	Predicate names of desired dimensions	Predicate names of desired dimensions and nesting chain	Name or URI of desired dimensions
What is modelled?	Real-world, ignoring statistical artefacts such as time, table, sub-tables, etc.	Statistical domain in question	Statistics in general
Use of deref.-able URIS	Very limited	Limited	Each statistical item has an explicit URI

* The use of different named graphs for different points in time is planned in D2R.

5 Usages in the Wild

5.1 Eurostat Data—riese

In riese[18], the "RDFizing and Interlinking the EuroStat Data Set Effort" [11, 12] we have RDFized, interlinked, and published the Eurostat dataset on the Web. First of all the dimensions and dataset hierarchies defined in Eurostat get translated to RDF and interlinked. The actual data is translated to RDF on-the-fly from the raw Eurostat tables. Both human users and machines can access the data from the same location thanks to embedding RDF on the human-readable pages with XHTML+RDFa. Additional access methods (as described below) are available as well.

RDFication The translation to RDF is performed using SWI-Prolog. The SWI-Prolog Semantic Web Library provides an infrastructure for reading, querying and storing Semantic Web documents; additionally the Prolog-2-RDF (p2r) modules[19] and

[18] http://riese.joanneum.at
[19] http://moustaki.org/p2r/

individually defined mappings are used for translating the input data to RDF. The input data consists of a table of contents in HTML defining the overall structure of the datasets, dictionary files for resolving the approximately 80,000 different data codes used, and the actual data itself in tab-separated values format. The Eurostat data consists of more than 4,000 datasets containing roughly 350 million data values/items.

In riese HTTP URIs are used, which are compliant with the linked data principles. For instance the currency dimension "Euro" has been minted with the "concept URI" of `http://riese.joanneum.at/dimension/currency/eur` which can be dereferenced for accessing the information resource that describes the concept. Accordingly datasets and individual items have an URI assigned for unambiguous identification of all resources.

Several domain ontologies are being re-used in riese. Geographical dimensions such as countries, etc. for instance make use of the Geonames ontology as they are subclasses of `geonames:Feature` which allows more expressive descriptions. It is hence easily possible to express further classifications such as the class of the geographical dimension (e.g. country, administrative division, etc.). Furthermore, the riese schema re-uses the Event ontology in the same way as described above.

Interlinking The automated interlinking of country descriptions between riese and Geonames for instance is done using the ISO-3166 alpha2 country codes which are available in both datasets assuring that exactly the same resource is addressed. In the practical implementation this means that a search using the country code is performed in both datasets. According to the nomenclature used by Eurostat it is also possible to identify only country descriptions in the source dataset. The result from the target dataset is restricted to return only countries as well. Finally all matches are being interlinked using `owl:sameAs`. In this case it is possible to create exact matching high-quality interlinks. The generation of interlinks to other LOD datasets such as DBpedia, CIA Factbook, and Wikicompany follows a similar approach.

Publication All data on riese is accessible for both humans and machines (Semantic Web agents) equally. An Apache 2 Server with a set of PHP scripts is used to render the pages in XHTML+RDFa. A dump of the entire data is also available in RDF/XML which should be used by Semantic Web agents like indexers that want to crawl the entire content. Providing a simple file download enables the indexers to easily acquire all data and reduces the server load. Furthermore a SPARQL endpoint using SWI-Prolog and p2r is provided. It operates on the raw input data and maps the data on-the-fly. Experiments have shown that for average queries this approach offers acceptable response times at reasonable server cost.

5.2 In Other Vocabularies

SCOVO is used in voiD, the "Vocabulary of Interlinked Datasets" [1] to express information about the number of triples, resources and so forth. Using SCOVO in voiD allows a simple and extendable description of statistical information, however, a shortcoming has been identified: as `scovo:Items` are grouped into `scovo:Datasets`, there is an implicit assumption that all items in such a dataset share the same dimensions. This yields to complex SPARQL expressions, as it will often require a verbose check to make sure that an item has only certain dimensions and no others. An examplary usage of SCOVO in voiD is given below in listing 4. Further, we have gathered that

```
:DBpedia a void:Dataset ;
         void:statItem [
           scovo:dimension void:numberOfTriples ;
           rdf:value 212576239 ;
         ] ;
}
```

Listing 4. Usage of SCOVO in voiD.

SCOVO is used in the RDFStats framework[20], see Fig. 5 (kudos to Andreas Langegger for the screen shot), that generates statistics for datasets behind SPARQL-endpoints and RDF documents.

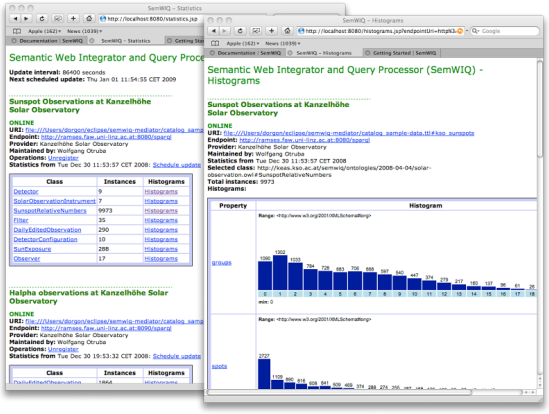

Fig. 5. SCOVO driving the statistics in RDFStats.

6 Conclusion and Future Work

We have proposed a vocabulary, SCOVO, and discussed good practice guidelines for publishing statistical data on the Web in this paper. The framework aims at supporting people to publish their statistics on the Web of Data in an effective and efficient manner. The framework includes good practice rules stemming from our experience in publishing linked datasets. Finally, we have demonstrated the implementation of our framework and discussed practical issues with it.

However, there are limitations we are aware of. We advocate simplicity, hence there exist edge cases where it is hard to find an appropriate semantic modelling. Take for example our `scovo:Dataset` concept. In the current representation it is not explicitly defined how the overall range of the dataset is expressed. This may yield performance problems when one determines to figure out the overall range of a dataset. It is for sure possible to concatenate single dimensions used on the `scovo:Item`-level—for example concluding from the range of the four quarters `ex:Q12006` to `ex:Q42006` that the dataset

[20] `http://semwiq.faw.uni-linz.ac.at/node/9`

actually is referring to the year 2006. Additionally, from the application of SCOVO in voiD we have learned that there is a demand for aggregates. Hence, we plan to add support for data aggregation in a future version of the SCOVO schema.

The United Nations provide a range of statistics about various domains[21]. We plan to publish this available data using the Talis Platform [9]. Due to the high volume of the data, scalability issues are at the centre of the entire design. For example, one part of the UN data set—the Commodity Trade Statistics Database (COMTRADE)—alone provides commodity trade data for all available countries and areas since 1962, containing almost 1.1 billion records. Further, our ongoing work focuses on broadening the deployment base available[22], making converters (from and to SCOVO) available, and extending the framework itself. One very important issue is what we call "statistical-presentation fidelity". When the data is present in a table with a certain layout, it turns out to be advantageous to not only repurpose and link the data, but also reuse the data table in the author's intended form.

Acknowledgements

The research reported in this paper was partially supported by the "Understanding Advertising" (UAd) project[23], funded by the Austrian FIT-IT Programme and the ICT-2007.1.2 ROMULUS project[24], partially funded under the 7th Framework Programme of the European Commission.

References

1. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, Spain, 2009.
2. J. F. Allen and G. Fergusson. Actions and events in interval temporal logic. Technical Report TR521, University of Rochester, Computer Science Department, 1994. Available at `http://citeseer.ist.psu.edu/allen94actions.html`. Last accessed February 2008.
3. J. Antoch. Environment for statistical computing. *Computer Science Review*, 2(2):113–122, 2008.
4. P. Assini. NESSTAR: A Semantic Web Application for Statistical Data and Metadata. In *International Workshop Real World RDF and Semantic Web Applications, 11th International World Wide Web Conference (WWW2002)*, 2002.
5. T. Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2007.
6. C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In *4th European Semantic Web Conference (ESWC2007)*, pages 802–815, 2007.
7. D. Booth. URI Declaration Versus Use. `http://dbooth.org/2007/uri-decl/`, 2008.

[21] `http://unstats.un.org/unsd/databases.htm`

[22] by for example publishing as RDF the OECD data at `http://www.sourceoecd.org/database/OECDStat`

[23] `http://www.sembase.at/index.php/UAd`

[24] `http://www.ict-romulus.eu/`

8. R. Cyganiak, R. Delbru, and G. Tummarello. Semantic Web Crawling: A Sitemap Extension. `http://sw.deri.org/2007/07/sitemapextension`, 2007.
9. I. Davis. The Talis Platform. `http://www.talis.com/applications/downloads/white_papers/TalisPlatform.pdf`, 2005.
10. A. Grossenbacher. Semantic Web: Basics, RDF, DC and the description of a statistical site. `http://tinyurl.com/2d5gta`, 2007.
11. W. Halb, Y. Raimond, and M. Hausenblas. Building Linked Data For Both Humans and Machines. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*, Beijing, China, 2008.
12. M. Hausenblas, W. Halb, and Y. Raimond. Scripting User Contributed Interlinking. In *4th Workshop on Scripting for the Semantic Web (SFSW08)*, Tenerife, Spain, 2008.
13. M. Hausenblas, W. Slany, and D. Ayers. A Performance and Scalability Metric for Virtual RDF Graphs. In *3rd Workshop on Scripting for the Semantic Web (SFSW07)*, Innsbruck, Austria, 2007.
14. ISO. Statistical data and metadata exchange (SDMX). Standard No. ISO/TS 17369:2005, 2005.
15. G.J. Lestina, W.P. LaPlant, D.W. Gillman, and M.V. Appel. Technical Development of the Proposed Statistical Metadata Standard. Report, Bureau of the Census, 1996.
16. S. McClean, W. Grossmann, and K. Froeschl. Towards Metadata-Guided Distributed Statistical Data Processing. In *Proc. of New Techniques and Technologies for Statistics (NTTS)*, 1998.
17. N. Noy and A. Rector. Defining N-ary Relations on the Semantic Web. W3C Working Group Note, W3C Semantic Web Best Practices and Deployment Working Group, 2006.
18. OECD. Management of Statistical Metadata at the OECD. Report, Organisation for Economic Co-operation and Development (OECD), 2006.
19. H. Papageorgiou, F. Pentaris, E. Theodorou, M. Vardaki, and M. Petrakos. Modeling statistical metadata. *Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on*, pages 25–35, 2001.
20. E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. W3c working draft 4 october 2006, W3C RDF Data Access Working Group, 2006.
21. L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web. W3C Interest Group Note, W3C Semantic Web Education and Outreach Interest Group., 2008.
22. Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference. W3C Working Draft, Semantic Web Deployment Working Group, 2008.
23. S. Softic and M. Hausenblas. Towards Opinion Mining Through Tracing Discussions on the Web. In *Social Data on the Web (SDoW 2008) Workshop at the 7th International Semantic Web Conference*, Karlsruhe, Germany, 2008.
24. H. Stuckenschmidt and F. van Harmelen. *Information Sharing on the Semantic Web*. Springer, 2005.
25. J. Tauberer. The 2000 U.S. Census: 1 Billion RDF Triples. `http://www.rdfabout.com/demo/census/`, 2007.
26. G. Tummarello, R. Delbru, and E. Oren. Sindice. com: Weaving the Open Linked Data. *Proceedings of the 6th International Semantic Web Conference 2007 (ISWC2007)*, 4825:552–565, 2007.
27. UN. Guidelines for Statistical Metadata on the Internet. Report, United Nations Economic Commission for Europe (UNECE), 2000.