Distr.
GENERAL

WP.27
7 April 2010

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (iv): Innovation and related issues including census systems

SDMX and the semantic web: implications for publishers of statistical data

Prepared by Simon Field, Office for National Statistics, United Kingdom

I. Introduction

1. This paper explores the emergence of the Semantic Web, and in particular of Linked Data, and their potential impact on the publication of official statistics. With the growing adoption of SDMX as a basis, not just for exchanging aggregate statistical data between agencies, but also for disseminating such data to the wider audience of consumers of statistical data, the paper examines the relationship between SDMX and Linked Data.

2. The paper includes a report on a workshop hosted by the Office for National Statistics on 4th and 5th February 2010, which initiated an international collaborative investigation of issues surrounding the publications of statistical data in Linked Data form. Opinions expressed in this paper are solely those of the author, but the paper draws on the expertise and collective wisdom of the 34 participants of that workshop, and the wider international community that has come together to pursue the topic in its wake.

II. Background

A. The Semantic Web

3. In 1999, Tim Berners-Lee wrote the following:

"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize." (Berners-Lee & Fischetti 1999)

4. This can be viewed, not so much as an extension of the human-readable web of HTML pages that we all navigate from browsers, but rather as a parallel web of machine-readable semantic information, in which software systems traverse the links looking for things that have related meanings.

B. Linked Data

5. In July 2006, Sir Tim Berners-Lee published a short “Design Issues” note entitled “Linked Data”, in which he proposed four simple rules for publishing data in the Semantic Web (Berners-Lee 2006). These have now become known as the "Linked Data Principles". They are:

a) Use URIs (Uniform Resource Identifiers) as names for things.
b) Use HTTP URIs so that people can look up those names.
c) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
d) Include links to other URIs so that they can discover more things.

6. This approach adopts two of the concepts that have led to the massive expansion and success of the World Wide Web over the past fifteen years, and applies them to a web of linked data:

a) The human readable web consists of pages, each of which has a unique name that is resolvable via a browser, irrespective of where the page is physically located. The first two principles propose adoption of the same concept and technologies to describe all “things” in the world - people, places, artefacts, facts, datasets. The term “web of linked data” is really short-hand for “web of linked data about things”.
b) Pages in the human readable web contain hyperlinks to other pages, enabling the user to traverse the web of related pages by following the links. The fourth principle applies the same concept to the web of linked data. The use of Resource Description Framework (RDF) Triples to represent these links (RDF Triples are also used to represent the data itself) means that links are typed, and are therefore themselves semantic (unlike HTML hyperlinks in the human readable web).

7. An RDF Triple is a very simple description that can be represented in XML, consisting of a *subject,* a *predicate,* and an *object.* Here is a statement that can be represented as an RDF Triple:

Simon Field knows Paul Richards

8. In this example, *Simon Field* is the subject, *knows* is the predicate, and *Paul Richards* is the object. When represented in RDF, the subject and predicate will be URI references, while the object may be either a URI reference or a string literal. The subject URI will therefore return some relevant information about *Simon Field* when de-referenced, the predicate URI will return some relevant information about the relationship type *knows,* and the object URI will return information about *Paul Richards*.

9. SPARQL (**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage) is an RDF Query Language. It became an official W3C Recommendation in January 2008 (Worldwide Web Consortium 2008). It is SQL-like, but is specifically for querying RDF graphs and dealing with RDF triples. A simple example, taken from Wikipedia, shows a SPARQL query to find all the country capitals in Africa (wikipedia.org 2010):

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Africa .  }
```

Figure 1 - Example SPARQL Query

10. A schema language, RDF Schema (also known as RDFS) is used to provide structure to RDF statements. It allows for the definition of classes and sub-classes, and can be thought of as a language to specify reusable patterns for describing similar things in RDF. These specifications are known as Ontologies, or RDF Vocabularies.

11. The Worldwide Web Consortium's Linking Open Data project aims to make data freely available to anyone on the web of linked data (Worldwide Web Consortium 2010). In October 2007, it had created just over two billion RDF triples. Today, two and a half years later, it consists of 13.1 billion RDF triples, interlinked by around 142 million RDF links. See Figure 2 below for a graphical representation of the datasets that have been published and interlinked by the project so far.

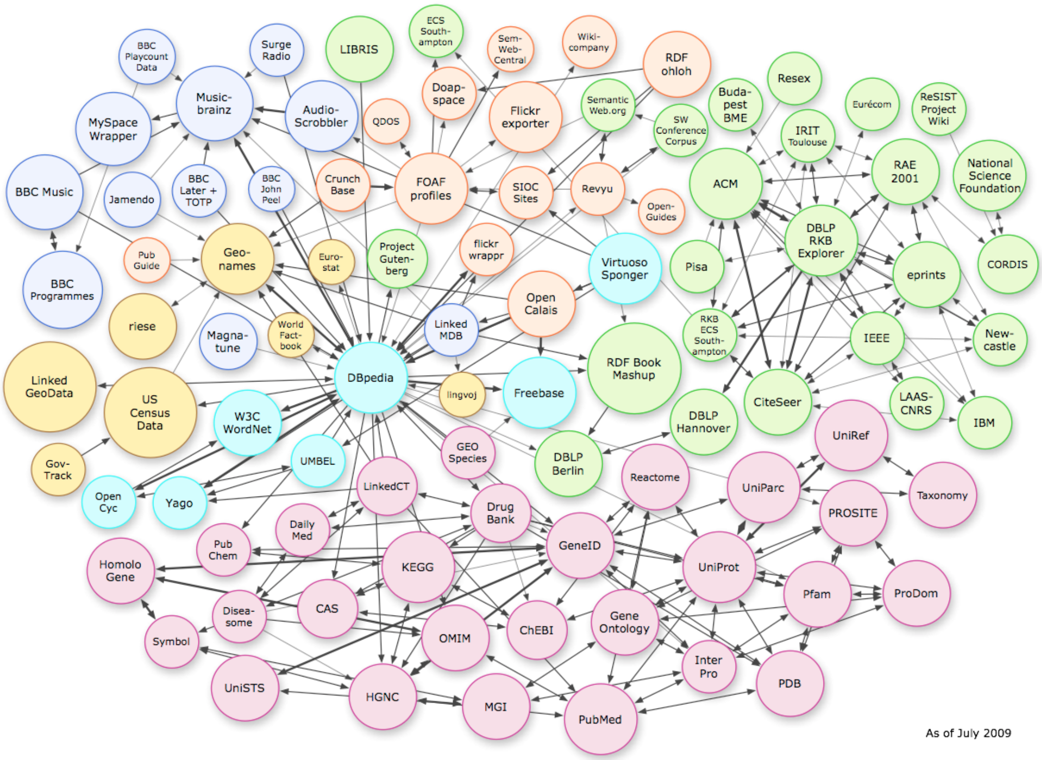

Figure 2 - Linking Open Data Datasets as of July 2009 (source: Linking Open Data project)

12. As with the early years of the human readable web, the linked data world today is mostly consumed by programmers and data developers. First generation end-user tools include browsers (e.g. Tabulator, Marbles) and search engines (e.g. Falcon, Sindice) and there are a small but growing number of domain specific applications (e.g. Revyu, DBpedia Mobile, DERI Pipes).

13. It is suggested, therefore, that the web of linked data can already be seen as a substantial global interconnected data resource growing at a considerable pace, supported by a maturing base of tools and end-user applications, though it remains in its early phase of development.

C. SDMX

14. This paper is being presented at an international meeting of publishers of statistical information. It is assumed that readers of this paper are broadly familiar with the Statistical Data and Metadata Exchange (SDMX) standard, and so it is beyond the scope of this paper to offer a broad introduction to the subject, which can be found at the standard's home page (SDMX Sponsors 2010a).

15. The standard consists of a number of assets, and is described as "an initiative to foster standards for the exchange of statistical information". Its exchange formats SDMX-EDI and SDMX-ML are underpinned

by a common data model that describes the universe of aggregate statistical dataset publication and exchange, and these exchange format standards are further supported by Content-oriented Guidelines that aim to ensure that different organisations publishing similar information will apply the standards in a common fashion.

16. The use of SDMX has been encouraged by its sponsors who have to receive statistical information from other organisations such as National Statistical Institutes (NSIs). Whilst adoption has been relatively slow, due to the need for NSIs to update the systems that generate publishable outputs, a growing number of NSIs have recognised that the greatest value of SDMX lies beyond its exchange formats, in its underlying data model which provides a common generic but comprehensive model of the world of aggregate statistical datasets.

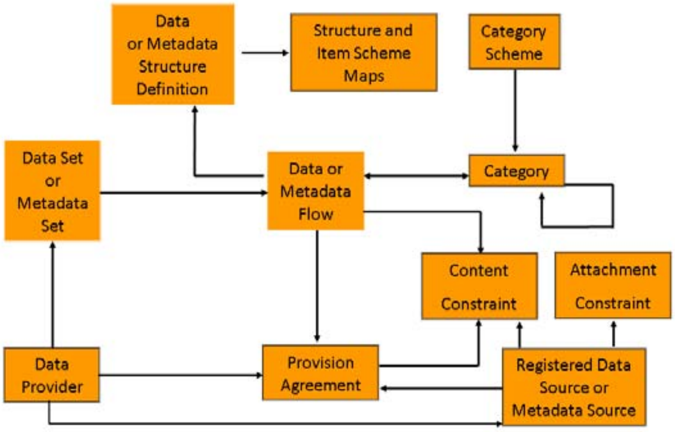

Figure 3 - schematic high-level view of the SDMX Information Model

17. SDMX is therefore becoming an international standard for describing, handling, and disseminating aggregate statistical datasets to all users of published statistics. The SDMX sponsors are a subset of these users, but recent years have seen the development of tools and reusable code for graphically displaying data held in SDMX-ML format to end users. This trend has been encouraged by the sponsors, some of whom are among the publishers of widely used open source assets for these purposes (SDMX Sponsors 2010b) (SDMX Sponsors 2010c).

III. Finding a common path

A. The dilemma

18. As indicated above, many publishers of statistical datasets have begun to adopt SDMX at the core of their web-based data dissemination systems. Shared tooling and visualisation code has accelerated this trend, and SDMX-ML is set to become a common format for publishing multi-dimensional datasets. It is perhaps unfortunate that, just as this standard begins to gather momentum, the web of linked data on the semantic web emerges to present a potentially conflicting vision of how data can be shared and used on the Internet.

19. Paragraph 4. above suggests that the Semantic Web can be viewed as a parallel web for machines rather than humans. Does this mean that publishers of statistical datasets will now have to publish to both of these worlds? How much duplication of effort will this involve? How will both worlds remain consistent with each other? There have been early experiments in publishing statistical data in linked data form (Joanneum Research 2010) (Tauberer 2007) , and these have given rise to a first attempt to create an RDF Vocabulary for statistical datasets; the Statistical Core Vocabulary (SCOVO) (Hausenblas et al. 2009). SCOVO is intended to be a lightweight simple representation of statistical data, and does not aim to cover the considerable detail represented by the SDMX information model.

20. In June 2009, the Prime Minister of the United Kingdom appointed Sir Tim Berners-Lee as expert advisor on public information delivery. The Prime Minister described this appointment as follows:

"So that Government information is accessible and useful for the widest possible group of people, I have asked Sir Tim Berners-Lee, who led the creation of the world wide web, to help us drive the opening up of access to Government data in the web over the coming months." (Cabinet Office 2007)

21. A key element of Sir Tim's project for the UK Government has been the creation of *data.gov.uk*, which was launched in January 2010 (Cabinet Office 2009), following a successful pilot during the preceding six months. Whilst the focus of the web portal is on "making public data public" in any open format, adoption of the semantic web of linked data has been a stated long-term aim. Some government datasets have already been published in this format, and in December 2009 the UK Government, clearly reflecting the advice given by Sir Tim, made a formal commitment to making all of its published datasets in linked data form:

"Any 'raw' dataset will be represented in linked data form" (HM Government 2009)

22. The use of the word 'raw' is somewhat misleading and unfortunate, but would appear to be intended to mean "machine readable", and not the more generally accepted definition of "unprocessed" (and most certainly not disclosive unit-level data). The implication is therefore that, in future, aggregate statistical datasets, alongside other public data, will be represented in linked data form.

23. The Office for National Statistics, as the UK's NSI, has recently committed to making all of its published aggregate datasets available in SDMX-ML format, via an Application Programming Interface (API) and has embarked on a redevelopment of its dissemination systems to facilitate this. It now faces the dilemma of how best to use its limited resources to serve up datasets in formats and standards that may appear to compete or conflict with each other. Or is it possible to exploit the best of both the linked data and SDMX worlds simultaneously?

24. It is suggested that the United Kingdom is not alone in facing this dilemma. Public Data initiatives have been launched in many other countries, including the United States of America (United States Government 2009), Australia (Australian Government 2009) and New Zealand (New Zealand Government 2009). And if linked data goes some way towards achieving Sir Tim's vision for the semantic web, then it is clear that all publishers of statistical data will have to consider how their outputs can have a presence in this new world.

B. The Workshop

25. On 4th and 5th February 2010, the Office for National Statistics hosted a workshop at the National School of Government, Sunningdale, entitled "Publishing Statistical Datasets in SDMX and the semantic web". The 34 invited participants were drawn from four overlapping communities:

a) SDMX expertise (including two of the standard's sponsoring organisations)
b) Linked Data expertise (including authors of the SCOVO vocabulary)
c) Publishers of statistical datasets (including NSIs from Australia, Norway, Portugal and the UK)
d) Consumers of statistical datasets

26. The stated aim of the workshop was "to better understand how dataset publishers can make the best of both the SDMX and semantic web worlds in serving their customers" (Office for National Statistics 2010c). What quickly emerged was a consensus that publishers of statistical datasets had much to gain by having their information represented in the linked data world, and that the linked data world had much to gain from the in-depth experience embodied in the SDMX standard in general, and the SDMX data model in particular. In other words, there is much to be gained from greater collaboration between the SDMX and Linked Data communities, and both publishers and consumers of statistical data would benefit from the results.

C. The recommendations

27. A summary of the workshop's conclusions has been put online, together with a list of all the participants (Office for National Statistics 2010d). It found that:

a) there is a need for a Linked Data representation of the SDMX model;
b) the model is well placed for adoption by those seeking to make data available using Linked Data standards, with some simple steps. To do this all data entities should be addressable by a dereferenceable http URI;
c) the community would benefit from an authorised Linked Data representation of the SDMX model;
d) the approach developed in SCOVO is compatible with the SDMX model.

28. It is perhaps worth examining the second finding in more detail before considering the activities that flowed from the workshop. At first sight, the conclusion that “all data entities should be addressable by a derefenceable http URI” would seem to be an onerous requirement, since it demands that every cell of every dataset, and every meaningful combination of cells, should have a unique reference. For a large multi-dimensional dataset, this could amount to many hundred million URIs. This is indeed what is required, and whilst it does represent adoption of the first two Linked Data Principles, it is not a requirement that is specific to Linked Data. The same requirement would have to be satisfied by any publisher wishing to offer a RESTful API (Fielding 2000) to access its dataset contents, irrespective of the format of the data to be returned.

29. The participants proposed to take the following next steps to continue and broaden the collaboration that began during the workshop:

a) We will discuss the outcomes of this workshop with four communities: SDMX, Linked Data, statistical data publishers and consumer of statistical data, and promote engagement in further work.
b) To move the alignment of the SDMX model and Linked Data forward, we will collaborate to develop a draft reference model of SDMX for datasets in RDF for consideration.
c) We will collaborate on developing a recommended style for URI design for use in APIs to find, obtain and query statistical data.
d) We will endeavour to create example implementations that demonstrate some of these findings.

IV. Starting the journey

A. A collaborative effort

30. The next steps outlined in paragraph 29 above propose the development of a draft reference model of SDMX for datasets in RDF, and the development of a recommended style for URI design. Work on these commenced during the workshop, but the participants were keen that continued work should involve a much wider group of collaborators (as shown in 29 a. above).

31. Immediately following the workshop, an open web-based collaborative forum was created, hosted by Google Groups (Office for National Statistics 2010b). This environment supports discussion threads, wiki-like pages, and shared documents. In the month following the workshop, membership of the forum has grown to over 60 participants from across the world, representing all four communities. Work on the common data model and a common URI style are progressing, so the following sections represent a snapshot of work in progress, reflecting the contributions of many participants. For the current state of this work, readers are invited to visit the forum at http://groups.google.com/group/publishing-statistical-data.

B. A common data model

31. Paragraph 16 above suggests that the greatest asset in SDMX is its data model, and this view is supported by the conclusions of the workshop. Work has commenced on identifying those parts of the model that will bring greatest value to the Linked Data world, and on extending earlier work, especially SCOVO, to create a suitable RDF vocabulary so that the model can be applied to RDF representations of statistical datasets and their metadata. Figure 4 below shows a high level overview of the initial scope, and Figure 5 shows a more detailed schematic of the vocabulary for datasets and data structure definitions. The structure of these should be instantly recognisable to readers who are already familiar with the SDMX data model.

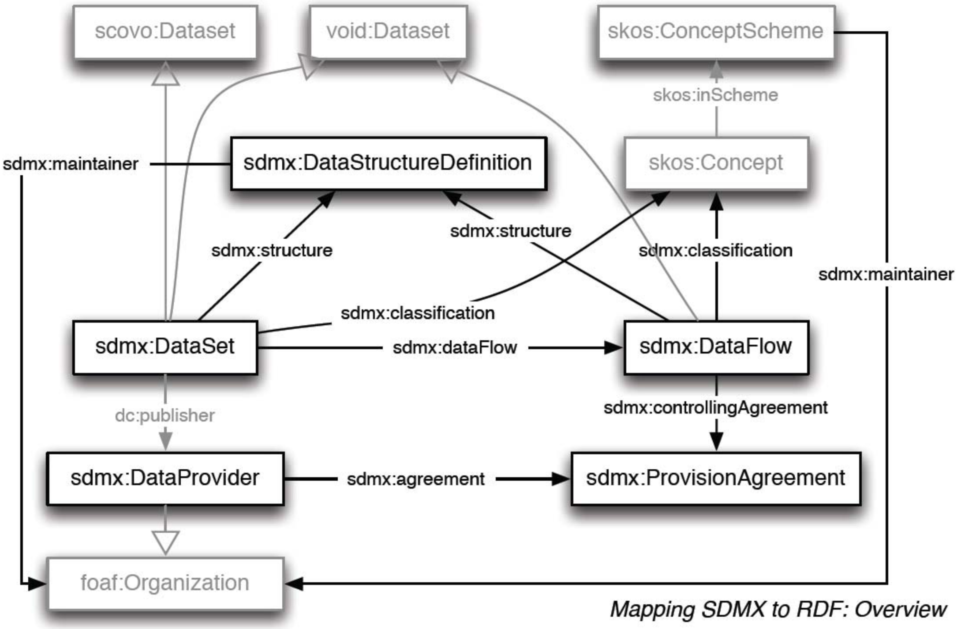

Figure 4 - Overview of mapping SDMX to RDF

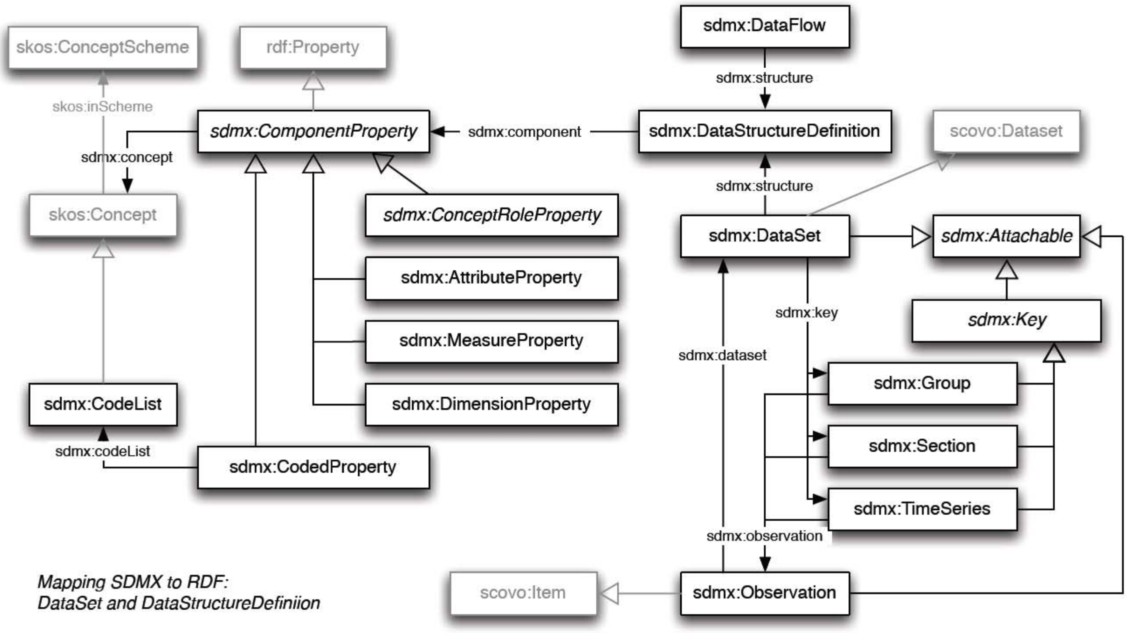

Figure 5 - Mapping Datasets and DSDs to RDF

32. It will be noted that whilst parts of the model are able to inherit from existing classes, leveraging SKOS and SCOVO, there are a number of new classes proposed that are specific to the SDMX model. It is suggested that, once this work reaches an appropriate level of maturity, including some sample implementations, the model will achieve wider adoption if it can achieve the status of an authorised representation. Whilst it is clearly premature to make a formal proposal today, we believe that future adoption of an RDF Vocabulary representation of the SDMX data model as part of the SDMX standard would extend the reach of SDMX into the emerging Semantic Web, while also benefiting the semantic web with an authoritative approach to dealing with statistical datasets.

33. The greatest beneficiary of this will be the data publishers themselves. Many are organising their aggregate data repositories around the SDMX model, and are serving SDMX-ML representations of their published datasets to their customers. If the semantic web of linked data adopts the same data model for its representation of statistical datasets, the effort for a publisher to offer its published aggregate dataset in a variety of formats, including RDF and SDMX-ML and SDMX-EDI, will be minimised.

C. A common URI style

34. Paragraph 28 above highlights the fact that URI design is an issue for any publisher offering a RESTful API, and not an issue specific to Linked Data. The workshop concluded that a standardised style of URI design would be of great value to any organisation wishing to offer an API to enable its customer to find, query and obtain its published statistical datasets. This recommended style could apply equally for provision of data in RDF and SDMX-ML formats (and indeed other popular web-oriented formats such as JSON). A first draft of this recommended style has already been published (Office for National Statistics 2010a).

V. What is our destination?

A. Open issues

35. Whilst the workshop participants agreed that they “found no barrier to bringing aggregate statistical data to a web of linked data, using the SDMX model” during the two days, there remain a number of open issues that need to be addressed.

36. Sir Tim Berners-Lee, in a recent paper co-authored with Tom Heath and Christian Bizer, has identified “Trust, Quality and Relevance” as open research challenges (Bizer et al. 2009). Even if provenance knowledge can be relatively easily incorporated into a linked data model, the authors conclude that "from an interface perspective, the question of how to represent the provenance and trustworthiness of data drawn from many sources into an integrated view is a significant research challenge".

37. Dealing with multiple dataset versions and revisions will also introduce a complexity that can be dealt with at the model level, but might prove more difficult for semantic web browsers and search engines to deal with. Such tools, and their users, may find it difficult to be sure that their RDF graph contains the most recent version of some published data. This is further complicated by the fact that, in the Linked Data world, each RDF triple is considered to be a piece of information that can stand on its own. This may be true where simple factual information can be contained within a single triple, but it is not true for a multi-dimensional dataset that has been decomposed into a large number of RDF triples. Validation and verification of completeness are not currently well supported in RDF, but these are needed if the integrity of a dataset is to be assured. Without it, version information may be lost, or an observation (such as a population count) may be received without one of its dimension values, totally and misleadingly altering the meaning of that observation.

38. It is encouraging that the research community is aware of these issues, and the new forum is providing an environment for their discussion in the specific context of publishing statistical data. It is perhaps too early to conclude that RDF is a suitable vehicle for representing all of the detail of a dataset, and it may be that a hybrid approach will emerge, where dataset and dimensional metadata are published in

linked data form to facilitate discovery, search and data linking, but dataset content is published in SDMX-ML form, where the integrity of the content can be more easily assured, verified and validated.

B. Ease of publication

39. Whether the semantic web of linked data reaches down to individual observations, or remains at the level of dataset discovery and linking, adoption of a common model for representing datasets in both the semantic web and the "traditional" human-readable web, will be of considerable benefit to dataset publishers. A single IT implementation will be able to serve both worlds consistently, to a common level of quality, and with significantly lower costs than if different models are required for each of these worlds.

C. A universal API?

40. Adoption of a standardised approach to URI design has been proposed. This would establish a common pattern to URIs used by publishers to provide access to their published statistical datasets. With the adoption of common metadata standards, the terms used to populate this pattern would also become standardised, and a universal API for accessing statistical datasets would emerge. This might mean, for example, that an API call to a service offered by the UK's Office for National Statistics could be changed into an equivalent call, returning equivalent data, to the corresponding service offered by the Australian Bureau of Statistics merely by changing the first part of the URI from *statistics.gov.uk* to *abs.gov.au.*

41. Achievement of this exciting goal is not a technical challenge, nor does it depend in any way on the semantic web. The burden is on the ability of the global statistical community to agree some common naming and metadata standards to complement the technical standards and recommendations that are emerging from the collaboration that was initiated at the Sunningdale Workshop. We believe that further effort to achieve this level of agreement is worthwhile, and a universal API to access statistical data across the globe is a prize that is both highly desirable and, in time, achievable.

VI. Conclusions

42. This paper has provided a high level overview of the emerging semantic web of linked data, and highlighted some of the potential impact it may have on the world of statistical data publication. International collaboration to explore the mutual benefit of bringing the linked data and SDMX worlds closer together has been described, and all interested parties are encouraged to join the ongoing collaborative effort to develop and test solutions that will benefit all publishers and consumers of statistical data. Finally, we have indicated the potential for this work to contribute to the creation of a universal API for accessing the world's statistical data.

VII. Acknowledgements

43. This paper is a summary of the achievements of many people. In particular, the author wishes to acknowledge the contribution of all 34 participants in the Sunningdale Workshop, and the subsequent contribution of the growing membership of the "Publishing Statistical Data" Google Group.

VIII. References

Australian Government, 2009. data.australia.gov.au – beta. Available at: http://data.australia.gov.au/

Berners-Lee, T., 2006. Linked Data - Design Issues. Available at: http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. & Fischetti, M., 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor* 1st ed., Harper San Francisco.

Bizer, C., Heath, T. & Berners-Lee, T., 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, (Special Issue on Linked Data).

Cabinet Office, 2007. Pioneer of the World Wide Web to advise the government on using data. Available at: http://www.cabinetoffice.gov.uk/newsroom/news_releases/2009/090610_web.aspx

Cabinet Office, 2009. Unlocking innovation | data.gov.uk. Available at: http://data.gov.uk/

Fielding, R., 2000. Architectural Styles and the Design of Network-based Software Architectures. Available at: http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Hausenblas, M. et al., 2009. SCOVO: Using Statistics on the Web of Data. In European Semantic Web Conference 2009. Heraklion, Crete.

HM Government, 2009. Putting the Frontline First. Available at: http://www.hmg.gov.uk/frontlinefirst.aspx

Joanneum Research, 2010. RDFizing and Interlinking the EuroStat Data Set Effort - riese. Available at: http://riese.joanneum.at/

New Zealand Government, 2009. data.govt.nz - New Zealand government data online » Data.govt.nz. Available at: http://www.data.govt.nz/

Office for National Statistics, 2010a. Draft URL Structure - Publishing Statistical Data | Google Groups. Available at: http://groups.google.com/group/publishing-statistical-data/web/draft-url-structure

Office for National Statistics, 2010b. Publishing Statistical Data | Google Groups. Available at: http://groups.google.com/group/publishing-statistical-data

Office for National Statistics, 2010c. Publishing statistical datasets in SDMX and the semantic web. Available at: http://ons.eventbrite.com/

Office for National Statistics, 2010d. Workshop Summary - Publishing Statistical Data. Available at: http://groups.google.com/group/publishing-statistical-data/web/workshop-summary

SDMX Sponsors, 2010a. SDMX – Statistical Data and Metadata Exchange. Available at: http://sdmx.org/

SDMX Sponsors, 2010b. SDMX – Tools – ECB. Available at: http://sdmx.org/?page_id=133

SDMX Sponsors, 2010c. SDMX – Tools – Eurostat. Available at: http://sdmx.org/?page_id=52

Tauberer, J., 2007. The 2000 U.S. Census: 1 Billion RDF Triples. Available at: http://www.rdfabout.com/demo/census/

United States Government, 2009. Data.gov. Available at: http://www.data.gov/

wikipedia.org, 2010. SPARQL - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/SPARQL

Worldwide Web Consortium, 2010. Linking Open Data community project. Available at: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Worldwide Web Consortium, 2008. SPARQL Query Language for RDF. Available at: http://www.w3.org/TR/rdf-sparql-query/