Semantic Statistics: Bringing Together SDMX and SCOVO

Richard Cyganiak
Digital Enterprise Research Institute
NUI Galway, Lower Dangan
Galway, Ireland
richard@cyganiak.de

Simon Field
Office for National Statistics
Cardiff Road
Newport NP10 8XG
simon.field@ons.gsi.gov.uk

Arofan Gregory
metadata technology
5335 North Nina Drive
Tucson, Arizona AZ 85704
arofan.gregory@metadatatechnology.com

Wolfgang Halb
Institute of Information Systems
JOANNEUM RESEARCH
Graz, Austria
wolfgang.halb@joanneum.at

Jeni Tennison
The Stationery Office
Mandela Way
London SE1 5SS
jeni.tennison@tso.co.uk

ABSTRACT

Whether it's population, income, unemployment or interest rates, statistical data is a fundamental source of information for analysis and visualisations. Many publishers of statistics use SDMX to represent statistics and make them available through web services. The linked data principles of identifying items with HTTP URIs and representing data using RDF provide some benefits (though also some costs) for statistical publishing. This paper describes how the SDMX information model can be used with linked data and RDF and describes some ongoing work to explore the impact of doing so.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Information Systems Applications

General Terms

Design, Standardization

Keywords

statistics, linked data, RDF, SDMX, SCOVO, open data

1. INTRODUCTION

Statistical data is the life blood of the interesting mash-ups and visualisations we see on the web. It provides the raw numbers that designers love to turn into graphs and charts. More importantly, its analysis enables policy makers to make predications, plan and adjust. So much of the data that we have is statistical data. But how can we bring together the existing standards for transferring statistical data and the linked data approach [1]? What advantages does this bring?

The current standard for statistical organisations to produce their statistical data is through Statistical Data and Metadata Exchange (SDMX) [2]. This standard covers everything from how to represent statistical data in flat files and as XML, to the definitions of the dimensions and attributes of observations, to how to discover statistical dataset flows through a central registry. SDMX is used by organisations such as the U.S. Federal Reserve Board, the European Central Bank, Eurostat, the WHO, the IMF, and the World Bank. The Organisation for Economic Cooperation and Development (OECD) and the UN expect the publishers of national statistics, such as the Office for National Statistics (ONS) in the UK, to produce their statistical data using SDMX so that these can be aggregated on an international level.

At the same time, the UK Government has made a commitment to make public data available on the web using linked data standards to enable its widespread re-use. The government publishes large volumes of statistical data and the Office for National Statistics has invested heavily in technology for managing and publishing statistical data on the web. Individual departments and agencies are also publishing increasing volumes of statistical data, in specialist XML-based formats such as LGDx[1], in CSV, as well as in other formats that generally curtail their wider combination and re-use, such as Excel or PDF.

The UK Government has to reconcile the requirement to share statistical data with other statistics authorities using SDMX with the wider requirement to enable re-use of statistical data on the web using linked data standards. The challenge is how to marry these different requirements in a pragmatic way for publishers and consumers alike. Publishers are concerned about publishing data responsibly and avoiding the accidental mistinterpretation of statistical data while the majority of consumers care about the ease of accessing, querying and processing statistical data.

In practice, this means publishing statistical data using HTTP URIs for datasets, time series and individual observations. These enable both publishers and third parties to annotate and reference statistical data on the web, which helps to build trust with those engaging in conversations about the data. Using the RDF data model enables consumers to query statistical data in standard ways and to enhance statistical data by mixing it with other linked data.

LDOW 2010, April 27, 2010, Raleigh, North Carolina.

[1] http://neighbourhood.statistics.gov.uk/dissemination/Info.do?page=nde.htm

Pragmatically, however, we must adopt approaches that utilise and build on existing data standards and technology investments rather than replacing them. While this paper focuses on the domain of official statistics, a similar situation can be observed in many fields: organisations have heavily invested in existing domain-specific standards. Bridging from these standards to the linked data universe is a prerequisite for getting valuable domain-specific data into the linked data web.

In this paper, we'll first describe SDMX and the data model it uses. We'll then talk about how publishing statistical data as linked data provides some advantages that aren't realised within the more traditional SDMX publishing pattern. We'll go on to describe how the SDMX data model and process model map on to linked data concepts, and thus how organisations such as the ONS can publish statistical data as linked data without disrupting their existing tool chains.

2. SDMX

The Statistical Data and Metadata Exchange (SDMX) Initiative was organised in 2001 by seven international organisations (BIS, ECB, Eurostat, IMF, OECD, World Bank and the UN) who remain the governing sponsors. The goals for the initiative were to realise greater efficiencies in statistical practice, with a focus on employing current technology to enhance efficiency, improve quality, and address other challenges. These organisations all *collect* significant amounts of data, mostly from the national level, to support policy. They also *disseminate* data at the supra-national and international levels.

There have been several important results from this work: two versions of a set of technical specifications - ISO:TS 17369 (SDMX) - and the release of several recommendations for structuring and harmonising cross-domain statistics, the *SDMX Content-Oriented Guidelines*. All of the products are available at www.sdmx.org. The standards are now being widely adopted around the world for the collection, exchange, processing, and dissemination of aggregate statistics by official statistical organisations. The UN Statistical Commission recommended SDMX as the preferred standard for statistics in 2007.

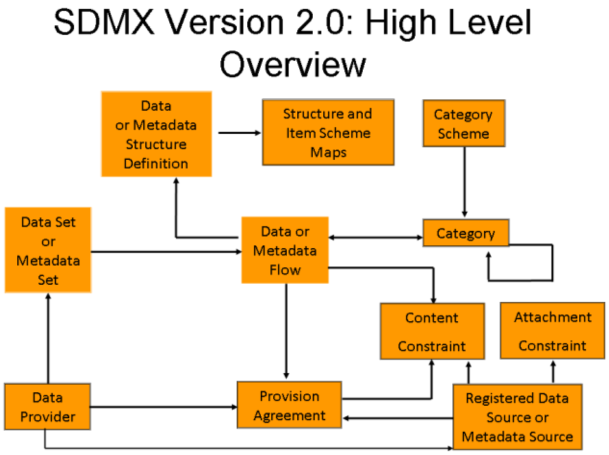
SDMX Version 2.0: High Level Overview
Data or Metadata Structure Definition
Structure and Item Scheme Maps
Category Scheme
Data Set or Metadata Set
Data or Metadata Flow
Category
Content Constraint
Attachment Constraint
Data Provider
Provision Agreement
Registered Data Source or Metadata Source

Figure 1: Schematic high-level view of the SDMX Information Model

SDMX is currently being employed by several producers of important data sets. To name a few examples, it is being used and adopted by the U.S. Federal Reserve Board, the Federal Reserve Bank of New York, the European Central Bank, Eurostat, the Bank for International Settlements, the OECD, the UN (for the Millenium Development Goals indicators), the World Health Organization, UNESCO (for education statistics), the IMF, the World Bank, the Food and Agriculture Organization, and many others, including numerous national-level statistical organisations and central banks. In some cases, it has become a standard mechanism for the dissemination of statistical data and metadata (OECD.stat is a good example). In others, it is used as an internal standard for processing, or as a means of supporting data collection and production. Adoption has been increasing rapidly throughout the official statistical world.

SDMX emphasises the *SDMX Information Model* shown at a high level in Figure 1 - a meta-model of the important aspects of collection, processing, exchange, and dissemination of aggregate data. All of the technology artefacts of SDMX are implementations of the SDMX Information Model. The model is based on an earlier standard, GESMES/TS, which used the UN/EDIFACT flat-file syntax. SDMX in its current version has expanded this model to include a view of the entire process of statistical production. This model is the result of implementation and analysis of statistical processes in many national, supra-national, and international organisations, and has been effectively used to support these functions in many implementations.

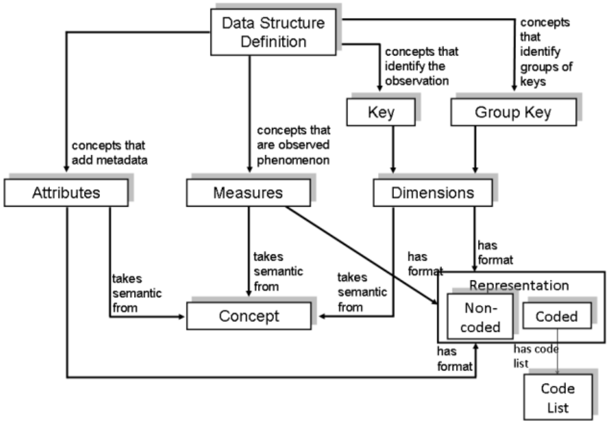

Figure 2: Schematic high-level view of data structures in the SDMX Information Model

The SDMX Technical Specifications describe two major syntaxes: SDMX-EDI, which uses the flat-file UN/EDIFACT syntax; and SDMX-ML, which is broader in scope and offers XML formats for many types of statistical data and metadata. In both cases, users configure the formats to work with the statistical concepts of importance to their data and metadata, providing a flexible, generic basis from which to work, which remains conformant with the standard model.

As well as specifying the formats for exchanging statistical data, SDMX defines a services-based architecture centred around the deployment of queryable web-services and coordination enabled by the uses of an optimised set of registry services.

Tools are becoming widely available for working with SDMX, as freeware, as open-source, and in statistical tools offered by commercial vendors.

3. STATISTICAL LINKED DATA

Linked data takes a different approach from SDMX. Rather than a centralised repository that can resolve URNs for the discovery of datasets, linked data simply uses HTTP URIs so that information can be found using the usual web architecture. Linked data emphasises the assignment of identifiers to all the instances in a data model, which in the SDMX case includes:

- code lists and codes
- concept schemes and concepts
- datasets and dataset flows
- time series and sections
- individual observations

Identifying each of these resources allows statements to be made about them. An individual anomalous observation might be annotated, for example. Datasets can be annotated to indicate the provenance of the statistics, including how they were collected and what processing they have been through. Being published as linked data also means that these items are accessible programmatically via the web. For example, this means that the detailed semantics of a particular dimension can be located easily, simply through resolving its identifier on the web.

What's more, statistical datasets can link into the wider web of data, leading to more powerful ways of filtering and querying statistical data. For example, statistical observations often reference the geographical area to which the statistic applies. Extra analysis can be made possible by resolving information about that area such as the political make-up of its council, how rural it is or the size of its police force. Similarly, information about the containment of an area inside others supports the aggregation of statistics for larger areas.

The Statistical Core Vocabulary (SCOVO) [3] demonstrates these principles and how they can apply to statistical data. It is a lightweight RDF vocabulary for expressing statistical data. Its relative simplicity allows easy adoption by data producers and consumers, and it can be combined with other RDF vocabularies for greater effect. The model is extensible both on the schema and the instance level for more specialized use cases. SCOVO's origins are in the *riese* ("RDFizing and Interlinking the EuroStat Data Set Effort") project [4, 5]. It has also been used to express statistics about RDF datasets [6, 7], in early efforts to convert UK government data to RDF[2], to publish Italian university statistics[3], and in other contexts.

SCOVO defines three basic concepts:

- 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 *scovo:Dataset*, which is a SKOS concept [8] in order to allow hooking into a categorisation scheme. A statistical data item *scovo:Item* belongs to a dataset. An *Item* is subsuming the *Event* concept, as defined in the Event ontology[4]. 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, one can relate to a statistical data item the institutional body responsible for 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.

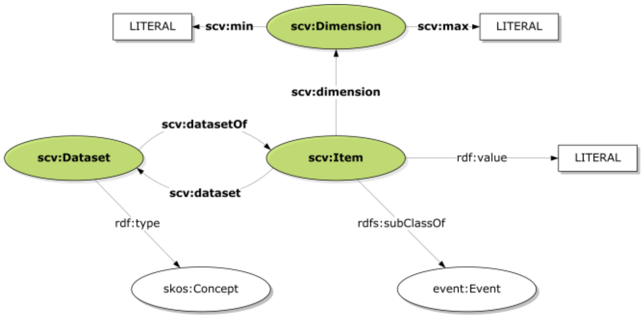

Figure 3: SCOVO data model

The Statistical Core Vocabulary (depicted in Fig. 3) is defined in RDF Schema.

4. EXPRESSING SDMX IN RDF

While SCOVO addresses the basic use case of publishing statistical data in linked data form, its minimalist design is limiting, and it does not support important scenarios that occur in statistical publishing and have led to the development of the SDMX information model, such as:

- definition and publication of the structure of a dataset independent from concrete data
- data flows which group together datasets that share the same structure, for example from different national data providers
- definition of "slices" through a dataset, such as an individual time series or cross-section, for individual annotation
- distinctions between dimensions, attributes and measures

There are also features of SDMX which are rightly not addressed by SCOVO but can be expressed through other vocabularies or design patterns, such as:

- describing code lists, category schemes, and mappings between them using SKOS
- describing metadata and access details about datasets using Dublin Core and voiD [7]
- describing organisations using FOAF

In this section, a mapping from SDMX to RDF is described. It is based on SCOVO, with extensions that partly borrow from existing vocabularies and partly reside in a new SDMX vocabulary. The task of mapping SDMX to RDF is greatly aided by the fact that the SDMX standard is separated into an abstract information model (SDMX-IM) and concrete XML and UN/EDIFACT based syntaxes. We will list main structures of the SDMX information model (see Fig. 1 and Fig. 2), and sketch their

[2] http://www.jenitennison.com/blog/node/138

[3] http://sw.unime.it/loius/info.html

[4] http://purl.org/NET/c4dm/event.owl

translation to RDF. The section will conclude with first implementation experiences.

Mapping overview. Fig. 4 provides a high-level overview of the RDF model. At the core of SDMX is the *data structure definition (DSD)*, which describes the structure, or metamodel, of one or more statistical datasets. Individual datasets must conform to a DSD, and are represented by instances of the *sdmx:DataSet* class. The *sdmx:structure* property connects a dataset and its DSD. The *sdmx:DataSet* class is defined as a subclass of SCOVO's *scovo:Dataset* class, and also as a subclass of *void:Dataset*, so VoiD properties can be used to describe access methods (SPARQL endpoint, RDF dump, etc.) to the data. VoiD covers much of the same ground as SDMX's web service based *registry* module, which we therefore do not map to RDF.

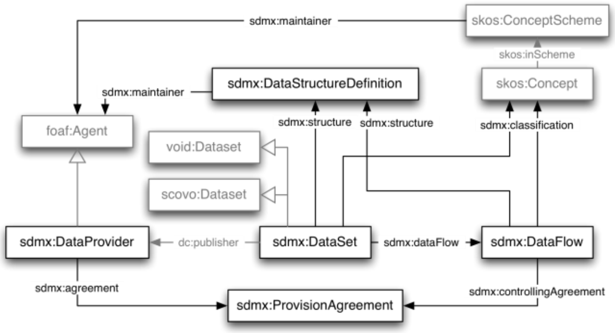

Figure 4: Mapping SDMX to RDF: Overview

Data providers. The organisation that publishes a dataset is given via the *dc:publisher* relationship. Organisations are represented as instances of *foaf:Agent*. Organisations are also used with the sdmx:maintainer property, which indicates the maintenance agency of various SDMX artefacts, such as DSDs, code lists, and category schemes.

Data flows and provision agreements. Two important scenarios in official statistics are the periodical publishing of datasets according to a schedule, and the aggregation of datasets from different data providers (e.g., European Union national statistics offices) into a larger collection for central dissemination (e.g., Eurostat). These scenarios are addressed via *sdmx:DataFlow*. A data flow represents a "feed" of datasets that all conform to the same DSD. Data flows are associated with *provision agreements*, which can be understood as commitments from an organisation to publish datasets into a data flow.

Data structure definition details. A DSD, also known as a *key family* in SDMX, describes the metamodel of one or more datasets (see Fig. 5). It defines *attributes*, *measures*, and *dimensions*, collectively called *components*. Measures name the observable phenomenon, such as *income per household*. Dimensions identify what is measured, such as of a particular *country* at a particular *time*. Attributes define metadata about the observations, such as the *method of data collection* or the *unit of measurement*. Components are *coded* if possible values come from a pre-defined code list (such as *country*), or *uncoded* otherwise. Code lists are mapped to a subclass of *skos:ConceptScheme*.

We represent all components as instances of *rdf:Property*. We define subclasses of *rdf:Property* to indicate the particular kind of component, as well as whether it is coded, and the particular role it plays in the DSD (e.g, *TimeDimension*, *PrimaryMeasure*). Compared to SCOVO, the property-based modeling of dimensions allows for a more compact RDF representation of observations.

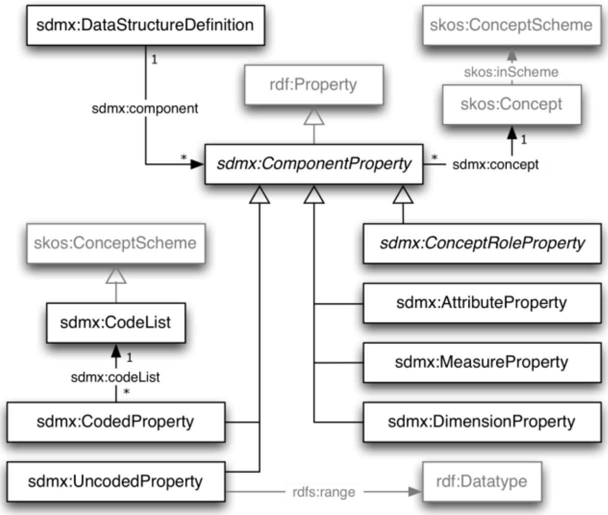

Figure 5: SDMX Data Structure Definition in RDF

Dimensions, attributes and measures in SDMX take their semantics from *concepts*. Concepts are items in *concept schemes*. By using standard concepts and code lists, data becomes comparable across datasets, DSDs, and providers. Concepts could be modeled as properties, and could be associated with components using *rdfs:subPropertyOf*. Instead, we model them as *skos:Concepts*, and introduce a new property for associating them with the component. This takes advantage of the easier management, wider reusability, and fine-grained mapping features of SKOS vocabularies compared to RDFS-defined properties.

Data set details. SDMX offers two approaches to organising the data inside a dataset. Either the dataset is a collection of time series (a set of observations that share the same dimension values except for the time dimension), or it is a collection of cross-sections (a set of observations that share the same dimension values except for one or more non-time "wildcard dimensions"). In our RDF mapping, we unify both models into a simpler yet more verbose model that can be more easily interrogated with SPARQL queries (see Fig. 6). The observation values are modeled as instances of *sdmx:Observation*, a subclass of *scovo:Item*. Each observation instance is directly connected to the *sdmx:DataSet* via the *sdmx:dataset* property. An observation must have a value for each dimension property defined in the DSD. The actual observation value is recorded using *rdf:value*.

The time series and cross-sections found in SDMX data are still translated to RDF, in order to make any metadata attached to them available in the RDF view. The same applies to *groups*, which are another organisational tool that can be used to apply metadata to sections of a dataset, for example to monthly, quarterly and annual timelines of the same measure.

Content-Oriented Guidelines. A key component of the SDMX standards package are the Content-Oriented Guidelines, a set of cross-domain concepts, code lists, and categories that support interoperability and comparability between datasets by providing a shared language between SDMX implementers. We have performed an initial mapping of the cross-domain concepts and the category scheme to RDF, although the result has not yet found a permanent home where we can guarantee stable URIs.

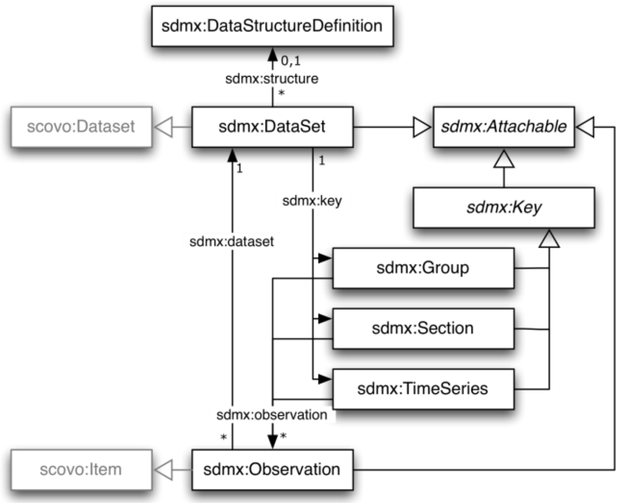

Figure 6: SDMX DataSet in RDF

Implementation experience. This mapping from SDMX to RDF can be carried out in a number of ways. For example, we have used XSLT to demonstrate mapping SDMX-ML (the XML dialect for SDMX) into RDF/XML. The mapping is easy to articulate, and the resulting linked data can be queried in a number of ways, to provide slices through the data that were not anticipated by the original publishers. On the down side, this approach to modeling statistical data does result in a large numbers of triples, with each observation resulting in a number of statements equal to at least the number of dimensions in the DSD, and consequently large file sizes.

5. CONCLUSIONS

The statistical publishing community and the linked data community share some common aims. Both seek to make data easy to locate and open it up for reuse, particularly for analysis and visualisation. From a linked data perspective, the world of statistical publishing is an extremely rich seam of data. For statistical publishers, linked data provides a way of addressing and retrieving information at a range of levels and in a distributed fashion.

By showing how to map from the standard SDMX information model into RDF, we hope to illustrate the ease with which statistical publishers could transition to providing data to both communities. There are already moves within the UK to create demonstrators that illustrate both the feasibility and the practical costs and benefits of publishing statistical data using linked data principles using this mapping, both from statistical data that is currently represented in SDMX and that represented in other forms such as LGDx and CSV.

Future work in this area includes formally capturing both the additional RDF vocabularies and concept schemes that are required to support SDMX and the ways in which this can work with other vocabularies to provide a mapping from SDMX. We will also be exploring the use of APIs aimed at web developers who will not typically be interested in learning the complexities of either SDMX or, indeed, RDF, but who simply want to access statistical data in simple and familiar ways. We believe that combining the experience and rigour behind SDMX with the web-based paradigm of linked data in this way will ultimately enhance the value both of statistical data and of the web of data.

6. ACKNOWLEDGEMENTS

This paper is based on the collaboration that was initiated in a workshop *Publishing statistical datasets in SDMX and the semantic web* hosted by ONS in Sunningdale, United Kingdom in February 2010. The completion of a draft reference model was one of several recommendations made by the participants, and this ongoing work continues in an open collaborative environment[5]. Taken together with the proposed collaboration to create a recommended style for URI design for use in APIs to find, obtain and query statistical data[6], we believe this work represents a key step towards bringing the worlds of linked data and official statistics together through the wider adoption of open standards. The authors would like to thank all the participants at that workshop for their input into this work.

The authors would also like to thank John Sheridan for his comments and suggestions on an earlier draft of this paper.

7. REFERENCES

[1] C. Bizer, T. Heath, T. Berners-Lee: Linked Data – The Story so Far. In International Journal on Semantic Web and Information Systems (IJSWIS), 2009

[2] International Organisation for Standardisation: ISO/TS 17369:2005 Statistical Data and Metadata Exchange (SDMX)

[3] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, D. Ayers. SCOVO: Using Statistics on the Web of Data. Proceedings of ESWC 2009 - 6th European Semantic Web Conference, p704-718, Heraklion, Greece, 2009

[4] 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.

[5] M. Hausenblas, W. Halb, and Y. Raimond. Scripting User Contributed Interlinking. In 4th Workshop on Scripting for the Semantic Web (SFSW08), Tenerife, Spain, 2008.

[6] A. Langegger, W. Wöß: RDFStats - An Extensible RDF Statistics Generator and Library. In DEXA Workshops 2009: 79-83

[7] K. Alexander, R. Cyganiak, M. Hausenblas: J. Zhao: Describing Linked Datasets. In: Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009

[8] Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, Semantic Web Deployment Working Group, 2009.

[5] http://groups.google.com/group/publishing-statistical-data

[6] http://groups.google.com/group/publishing-statistical-data/web/workshop-summary