WP 4
1 March 2010

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

OPEN ISSUES ON THE SEMANTIC WEB

Invited Paper

Submitted by US Bureau of Labor Statistics[1]

I. INTRODUCTION

1. In 1999, Tim Berners-Lee, who is credited with the invention of the World Wide Web, laid out his vision of the Semantic Web[2]:

> 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.

The main idea is that rather than having Web pages that are readable by people, they need to be readable by a computer. Humans are capable of using the Web to carry out tasks such as finding the height of Mt. Everest, reserving a table at a favorite restaurant, and searching for the best prices for tires for a car. However, a computer cannot accomplish these tasks, because web pages are designed to be read by people, not machines. The Semantic Web is a vision of information that is understandable by computers, so that they can perform more of the tedious work involved in finding, combining, and acting upon information on the Web[3].

2. The Semantic Web demands much more in the way content is organized than what the Web currently requires. As such, it is designed to work with 2 new kinds of information artifacts:

- Web Services – Basically, the W3C defines a web service as a software system designed to support interoperable machine-to-machine interaction over a network. More generally, a service is a set of events with a defined interface.
- Ontologies – An ontology is a set of concepts, the relations among them, and a means for computing with them. The purpose is to be able to reason, i.e., make inferences, with the concepts in an ontology.

Web services and ontologies are designed to work in conjunction with each other. They are the foundation from which the Semantic Web springs.

---

[1] Prepared by Daniel Gillman – The opinions expressed in this paper are due to the author and do not necessarily represent official policies of the US Bureau of Labor Statistics
[2] See http://en.wikipedia.org/wiki/Semantic_Web
[3] This paragraph taken from the Wikipedia page for the Semantic Web: http://en.wikipedia.org/wiki/Semantic_Web

3. This short description is not very detailed, and there is much more we can say. However, we want to talk about how the Semantic Web might help the official statistics community. One of the most important purposes of a statistical agency is to produce data for its users. Some users only care about data sets that come from one survey, and they might care about subsequent reissues of it over time. Other users are interested in using and combining data from multiple sources, either from a single data producing organization or from many and possibly over a long time frame. The problems of how to find, process, analyze, combine, and harmonize data from multiple sources is one the Semantic Web is envisioned to facilitate.

4. This paper contains an exploration of some of these issues. First, we describe a scenario for combining data from multiple sources. Next, we describe some of the technologies that are associated with the Semantic Web. Third, we discuss the ways in which Semantic Web techniques might differ from metadata management principles the statistical offices have written about and implemented for years. Throughout, we try to determine which of the steps in the scenario the Semantic Web might reasonably be able to solve that can't be done automatically already (i.e., what is the value-added for the Semantic Web). We identify 3 kinds of problems in the scenario that range in difficulty and may pose problems for the Semantic Web.

5. The paper does not go into a lot of technical detail. It is intended to be a stimulus for continued debate.

II. SCENARIO

6. Here we present a scenario for finding, analyzing, and harmonizing data form multiple sources. This scenario is taken from the article "America's Safest Cities" by Zack O'Malley Greenburg in the 26 October 2009 issue of *Forbes Magazine* (http://www.forbes.com/2009/10/26/safest-cities-ten-lifestyle-real-estate-metros-msa.html). The scenario is described in the following excerpt from the article:

> To determine our list of America's safest cities, we looked at the country's 40 largest metropolitan statistical areas across four categories of danger. We considered 2008 workplace death rates from the Bureau of Labor Statistics; 2008 traffic death rates from the National Highway Traffic Safety Administration; and natural disaster risk, using rankings from green living site SustainLane.com. It devised its rankings by collecting historical data on hurricanes, major flooding, catastrophic hail, tornado super-outbreaks, and earthquakes from government agencies including the National Oceanic and Atmospheric Administration, the United States Geological Survey, the Department of Homeland Security, the Federal Emergency Management Agency, and private outfit Risk Management Solutions. We also looked at violent crime rates from the FBI's 2008 uniform crime report. The violent crime category is composed of four offenses: murder and non-negligent manslaughter, forcible rape, robbery, and aggravated assault. In cases where the FBI report included incomplete data on a given metro area, we used estimates from Sperling's BestPlaces.

7. As seen in the list of the rankings of the MSAs in the *Forbes* article, the overall ranking was determined by adding the ranks in each of the 4 measured categories. Cities are listed in increasing order of sum, with the lowest sum given the highest rank.

8. It is true that one of the referenced data sources, SustainLane.com, is not a US Government web site, however their rankings are based on US federal, state, and local government data as well as other public data sources. See the "About the Study" section on their web site for more details. As the paper contains an illustration, the inclusion of this source should not detract from the paper's main purpose. So, this scenario is to find data on rates of workplace fatalities, traffic fatalities, violent crimes, and natural disaster risk as proxies for "livability" among the MSAs in the US and rank them based on these rates. If, for instance, the traffic fatality rate in some city is lower than most, then it is a more desirable place to live.

9. The following data sets appear to have been used for this study:

- Bureau of Labor Statistics (BLS):
  - Workplace deaths – http://www.bls.gov/iif/oshwc/cfoi/cfoi_msa_2008.pdf
- National Highway Traffic Safety Administration (NHTSA):
  - Traffic deaths – http://www-nrd.nhtsa.dot.gov/Pubs/811170.PDF
  - Note: based on city, not MSA
- Federal Bureau of Investigation (FBI):
  - Violent crime – http://www.fbi.gov/ucr/cius2008/data/table_06.html

- Alternative crime rates at BestPlaces – http://www.bestplaces.net
- Natural Disaster Risks (SustainLane.com):
  - Natural disaster risk – http://www.sustainlane.com/us-city-rankings/categories/natural-disaster-risk
  - Note: Apparently, Boston was used as a proxy, because Providence is not listed
- Census Bureau
  - Population by MSA for 2008 -- http://www.census.gov/popest/metro/CBSA-est2008-annual.html
  - Definitions of MSA – http://www.census.gov/population/www/metroareas/metrodef.html

10. We need to assume the user doesn't know which data sets to use beforehand. They have to be discovered. Discovery is basically a question of when (2008), where (MSA), and what (subjects: violent crime, traffic fatalities, workplace fatalities, and natural disaster risk). Later, we'll discover we need to know the population of each MSA in 2008, also.

11. There are 4 search queries:
- "workplace fatalities" 2008 msa
- "traffic fatalities" 2008 msa
- "violent crime" 2008 msa
- "natural disaster risk" 2008 msa

In none of the cases does a simple Google® search turn up the right data set at the top of the list or even on the first page of links. The results of using the advanced search capability and limiting the search to the .gov domain increases searching accuracy, however the direct links (those listed above) to the data sets are not on the top Google® pages in any of the cases. What is displayed instead is a link to another page (at the agency Web site) on which the direct link to the correct data set is found. But, several questions arise from this. How does the user know which Web page to look at to find the link to the data? How does the user know that the US federal government is the source for most of that data? More to the point, how did we know to limit our search to the .gov domain? Additionally, searching the .gov domain will not locate the SustainLane.Com web site either. How is that Web site found?

12. These are difficult questions, and it is not clear that Semantic Web technology can answer them. These questions will be revisited later, so in order to refer back we label these questions the **discovery problem**.

13. Assuming we can find the data, then there is a relatively simple question, and that is does each data set have all the data elements that are needed to answer the original questions? This is the **meaning problem**. Let's look at each of the data sets we discovered to see if we have the right data elements, and we will discover a new problem along the way:

- At BLS, there is no workplace fatality rate by MSA, therefore it has to be calculated
  - The fatality totals for each MSA, which is the data element available, is not suitable for comparing risks
  - To calculate a rate, population data for each MSA must be used
  - Data for population by MSA is at the Census Bureau, but that has to be discovered
- At NHTSA, traffic fatality data is based on city, not MSA
  - This time there are rates, but now a qualitative decision must be made that city data are a reasonable proxy for MSA data
    - This is a very hard question that we revisit later, and we call this the **judgement problem**
- At FBI, violent crime data is listed by MSA and contains a rate based on fixed population size

    - This time there is nothing more we need to do
- At SustainLane.Com, the natural disaster risk data are by city, not MSA
    - There is not enough city data for every MSA, so not only do we have the judgement problem (can city data substitute for an MSA), we have the additional problem of deciding whether some other city can serve as a proxy for an MSA
    - No risk factor is provided; instead an ordinal number ranking of the cities by risk is given
        - Here we have to know that when we order the MSAs by some rate (e.g., workplace fatalities) we end up with an ordinal ranking

14. Note here that standards are required to make this work. Either standards for data set formats and data dictionaries or standards for the output of the services that access the formats and dictionaries are needed. Given that standards are in use, it is easy to see that the meaning problem is straightforward. Links are established between a data set and all the metadata needed to describe it. This will include a data dictionary, format, and other information.

III. SEMANTIC WEB TECHNOLOGIES

15. Here we briefly take a look at some of the technologies that are associated with the Semantic Web and show how they fit together. There are three main kinds:

- Web services
- Ontologies
- Knowledge representation languages

In paragraph 2, we said the Semantic Web requires two new kinds of information artifacts: web services and ontologies. The knowledge representation languages are a kind of bridge between these two. We discuss this a little more below.

16. Web services take all the actions in the Semantic Web. There are several kinds of web services, but that discussion is beyond the scope of this paper. If some operation needs to be performed, a Web service is called to do it. Not everything needs to be programmed that way; this is conceptually what is going on. An example of a web service is the action of displaying the data dictionary for a data set.

17. Ontologies are the means by which knowledge of some subject domain is organized. As defined above, they contain sets of concepts and the relations between them. Concepts are identified and related to other concepts for some purpose. For instance, in an ontology for METIS, there will be a concept for the CMF[4]. That concept has a partitive relation (a whole–part relation) with each of the 4 parts of the CMF.

18. Ontologies are formal in the sense that they support logical operations, inference making, and possibly theorem proving. In order to take advantage of this feature, a language for representing the knowledge contained in the concepts and relations is required. These knowledge representation languages are based on logical principles. Some use simple logical systems, such as Resource Description Framework (RDF), which just allows for relations and connections between objects. RDF is based on a simple triple: subject – predicate – object. All knowledge is represented using interlocking triples, i.e., the object of one triple may the subject of many others.

19. The Web Ontology Language (OWL)[5] is more powerful than RDF and actually is in several different versions, each based on a different logical system, and with different powers of inferencing. These are OWL-Full, OWL-DL, and OWL-lite. OWL-Full is similar to RDF-Schema, and has much of the power of first order logic[6], the system in which mathematics is based. See, for instance, *Model Theory* by Chang and Keisler for a

[4] See http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework
[5] See http://en.wikipedia.org/wiki/Web_Ontology_Language
[6] See http://en.wikipedia.org/wiki/First-order_predicate_logic

thorough mathematical description of first order logic. OWL-DL and OWL-lite are description logics[7], which are less powerful but easier to implement. Whatever logical system is used determines what kinds of statements can be made about the subject domain represented. However, it is important to point out that one of the commonly used programming languages (e.g., FORTRAN, C, C++, Java, LISP) have the same formality or can be used in that way. The reason people use knowledge representation languages is they feel the syntax of these languages make it easier to express the formal statements that are needed.

20. Groups representing a subject matter area, such as official statistics in some country, might get together to build an ontology of their subject domain. To make this useful for the Semantic Web, the ontology needs to be represented in some knowledge representation language. Then, a web service is designed to read the formal statements about the ontology and reach conclusions based on them. Together, the web services and ontologies are expected to turn the Semantic Web into a very powerful distributed system of knowledge.

21. The use of a knowledge representation system turns an ontology from a model of concepts into something that is computational. The usual set of knowledge representation languages do not need to be used. However, they are suited to the task.

22. Generally, the work required to build anything but a trivial ontology is huge. After a decision is made about the level of formality required, the concept system must be built. Definitions, delimiting characteristics, relations, etc all have to be worked out. This is time consuming, expensive, and very labor intensive. Then, the difficulty of mapping the concept system to statements in a knowledge representation language may be equally difficult. One problem is insuring the consistency of all the statements created. Much of this work requires expert help, both from the point of view of terminology and knowledge representation.

23. One common error is to claim a Semantic Web application just because RDF (or OWL) was used to model information. Knowledge representation languages are just modeling paradigms. They don't cause a system to have some quality that isn't already designed into the system to begin with. In addition, common knowledge representation languages are not the only way to represent a formal system for web services to interact with. As stated in paragraph 19, the common programming languages can be used for the same purpose. It just might be harder to write that code.

IV. SEMANTIC WEB AND METADATA MANAGEMENT

24. Clearly, metadata will play a central role in the Semantic Web. In the scenario described in section II, there are many instances where metadata are needed. Data dictionaries, data set formats, concept definitions, and data set locators are just a few of the metadata categories that were identified. However, in this section we investigate in what ways the Semantic Web represents something new.

25. As identified in section II, the **discovery problem** looms large. When a user is trying to answer some question involving data, finding relevant data is paramount. Without the data, nothing more can be done. How can the process of finding relevant data be more automated? Is the Semantic Web the right approach to solve this?

26. A relatively new development in the Semantic Web community is the idea of Linked Data. For the statistical community, this idea seems to be very close to what is needed to help with the discovery problem. Again, Tim Berners-Lee is credited with laying out the ideas, and he outlined four principles of Linked Data in his Design Issues: Linked Data note[8], paraphrased as follows:

1. Use URIs to identify things.
2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.

[7] See http://en.wikipedia.org/wiki/Description_Logic
[8] See http://www.w3.org/DesignIssues/LinkedData.html

3. Provide useful information (i.e., a structured description — metadata) about the thing when its URI is dereferenced.

4. Include links to other, related URIs in the exposed data to improve discovery of other related information on the Web.[9]

27. There are two main reactions to these 4 principles. The first is that they do not differ in any meaningful way from the metadata management and data dissemination principles the official statistical offices have promulgated for years (See METIS: 20 Years of Progress in this meeting). Providing URLs (a URI with an HTTP dereferencing mechanism) is the main way offices point to data sets on the Web. The second and more important reaction is they beg the question. The idea is to use Linked Data to find data, but once you know its URI, the problem is basically solved. The real question is how to find the right URI.

28. Ontologies are supposed to lead the user to the right URI. This is a much deeper claim. The question then becomes, is it possible to build an ontology that is relatively complete (it covers all the relevant topics, therefore no relevant data set will be skipped), is consistent (it doesn't lead to contradictions, i.e., an irrelevant data set is found), and adequately leads users to the right data (it is useful, i.e., it finds relevant data in a reasonable amount of time). It is possible to build a complete and consistent ontology that does nothing very useful.

29. On the other hand, registries as laid out in SDMX[10] and ISO/IEC 11179[11] are designed to solve the same problem. Other specifications describe registries as well. A registry mimics a library card catalog file, only it is an on-line resource. Each record contains enough information to describe the resource it references, but it provides this description briefly. The elements in each description are metadata elements, and one of the most common sets of such elements is the Dublin Core[12]. Many successful discovery applications have been built upon this framework. It is not yet clear why and how Linked Data may be a better solution.

30. The hardest problem identified in section II is the judgement problem. In examples of this problem, judgment decisions are made. For instance, the decision to substitute city estimates as a proxy for an entire MSA containing it is a judgment decision. The same is true for substituting data for a nearby city as a proxy for an MSA. More interesting for the debate about the Semantic Web is to look at these judgments as they concern what tool or data set to select. This is often presented as an easy problem, and it is not. Even though the SustainLane.Com data does not contain natural disaster risks for every city, let alone any MSA, we made the judgment to use them anyway. At this stage in the development of the Semantic Web, AI technology, and the human – machine debate, it is extremely difficult to see how problems of this sort will be solved automatically. The last sentence in paragraph 1 – "The Semantic Web is a vision of information that is understandable by computers, so that they can perform more of the tedious work involved in finding, combining, and acting upon information on the Web." – is true. And, as we are seeing, some tasks are not "tedious" and require human intervention to solve.

31. Lastly, the meaning problem may be the easiest to solve with respect to leading users to relevant descriptions. Meaning is what much of the METIS community is interested in organizing, managing, and presenting to users about surveys and data. This has been the subject of the METIS meetings since their inception. The difficult parts of providing meaning, also discussed in METIS over the years, are 1) deciding which attributes or categories are most informative, and 2) creating the metadata for meaning in the first place. Once we have these answers, we as statistical offices know what data to link the descriptions to. This means, from the user perspective, the problem is straightforward to solve. What is troubling is that problems that were touted as easy are in fact very difficult, while so-called hard problems turn out to be easy.

---

[9] Taken from the Wikipedia page on Linked Data – http://en.wikipedia.org/wiki/Linked_Data
[10] See http://www.sdmx.org
[11] See http://en.wikipedia.org/wiki/11179
[12] See http://dublincore.org

V. CONCLUSION

32. In this paper we discussed the Semantic Web, its technologies, and its applicability for statistical offices. This was done without forcing the discussion to a deep technical level. Even so, through the analysis of a scenario involving the discovery, use, and understanding of several statistical data sets, we tried to identify the limits of what the Semantic Web can do for statistical offices.