The Web of Linked Data: Realizing the Potential for the Social Sciences

Arofan Gregory
Open Data Foundation

Mary Vardigan
Inter-university Consortium for Political and Social Research

Abstract
The Linked Data Web holds great promise for social science researchers, enabling efficient discovery of data and increasing the ease with which disparate datasets can be merged. Further, quantitative data sources can be easily connected with non-quantitative resources such as research papers. However, this vision will not be realized if best practice in documenting social science data is not combined with the use of the new technologies.

The key to realizing the transformative potential of the Linked Data Web for SBE sciences is collaboration among technologists and data managers and producers, based on the emerging standard metadata models used for quantitative SBE data. The most prominent of these is the Data Documentation Initiative (DDI). If these models can be realized as standard ontologies for the publication of research data as Linked Data, then the wealth of SBE data found in data archives and in government organizations become discoverable and available to create new knowledge. Without such standard ontologies, the creation of generic tools for working with quantitative Linked Data will not be possible, and we will fail to realize the potential of the technology for the SBE sciences.

Introduction

The Semantic Web/Linked Data technologies offer many possibilities for enhancing the discovery and use of social science data. It becomes possible to find and merge data coming from disparate sources, and to more effectively utilize the Internet for all types of data-related activities. Ultimately, the new technologies promise to bring greater transparency in government, more effective utilization of data in research, and better evidence-based policy. This is a compelling vision for the next decade.

However, the realization of this vision is complicated. Problems that have traditionally challenged data producers are not necessarily solved by the new technologies, and in some cases they are magnified. If the data community and technologists can work together effectively, many of these difficult issues can be solved, but neither group will be able to realize the promise of this vision if they do not work together, benefiting from each other's knowledge and experience.

Fortunately, the needed collaboration has begun, and if it can be successfully carried forward, then the new technologies will truly serve the users of quantitative data, and all the stakeholders will benefit. This paper outlines the road-map for such collaboration.

What is Linked Data?

Linked Data is about using the Web to connect related data that were not previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."

Linked data uses structured ontologies, or vocabularies, which are the explicit formal specifications expressing relevant terms in a domain, relationships among them, and their behavior. Software tools that discover and manipulate datasets or other information must understand the associated vocabulary in order to make full use of them.

The Vision of the Semantic Web

It is important to understand the benefits of Linked Data in the Semantic Web. When we think about the Web, the first issue is how to find what we are looking for. This is especially true for data, even with an increasing number of sites acting as portals for data. The *discovery* of data is a major challenge. And, once discovered, it is often difficult to know what the data are and how to use them – how are they organized and structured? How were they collected? What format are they in and how can they be processed?

Further, for data to be really useful in research, one may want to compare them with data from other sources, so we need to understand how comparable different data from various sources are.

An ideal vision of using Linked Data to produce a rich browsing environment for social science data in 2020 might look something like this:

A researcher discovers in an online journal an article about the relationship between obesity and income in a specific geographic area. She reads the article and links to a table with references to specific income and obesity variables from the underlying dataset, with links to full descriptions of each variable (name and label, associated question, responses, and frequencies) in context and links to download the data and browse additional documentation. This dataset is part of a multi-wave longitudinal study, and the user can view these variables in other years to compare them. She is able to recreate the table using data from other years. The researcher is also able to browse other variables in the dataset and to explore similar variables in other datasets ("More Like This") that might differ slightly in terms of how the questions were posed to respondents.

Let's say the researcher wants to know about the relationships among income, obesity, and the proximity of fast food restaurants in the region of interest. A quick search reveals data on fast food, and the researcher is able to understand quickly whether the analysis she wants to perform is scientifically sound given the nature of the disparate datasets because the data include "smart" variables that know what they can combine with. She successfully merges the data and is presented with a visualization of the results in the

form of a complex table and a map. This happens in a secure environment to minimize disclosure risk.

The researcher can also follow links to other articles that used these data and specifically the variables of interest. She can link to a profile of the researcher and discover additional publications on the topic of interest from the researcher's CV. She might even initiate a collaboration with that researcher, producing new knowledge linked to existing information on the Web and adding to the rich Web of linkages in her domain.

In short, through Linked Data for the social sciences, users should be able to easily discover the existence of data, and to determine what the data contain, how they are structured, and how they can be used. The data should be well-documented. They should be of known quality and provenance, and should exist in relationship to other versions of the same dataset, so that corrections and updates can be known. If comparable/ scientifically compatible with other datasets, then that should be something that is easily determined, and it should be possible to easily merge the data.

Linked Data Examples

A noteworthy effort in the realm of the Semantic Web and Linked Data is the Vivo project, originally developed at Cornell, http://www.vivoweb.org/. The VIVO national network of scientists facilitates the discovery of researchers and collaborators across the country. In fact, much of the infrastructure required for the collaboration described above is already in place as a result of this successful foray into the Semantic Web.

In addition, there have been several implementations of Linked Data technology with aggregate statistical data. For example, tables from the US Census have been published in the form of RDF triples (the base technology of the Linked Data Web), at http://www.rdfabout.com/demo/census/. Data from Eurostat have been re-published in Linked Data form in a project that resulted in the creation of SCOVO (the Core Statistical Vocabulary): http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html. And the SDMX standard for time series data now has, in addition to its XML representation, an RDF expression, leading the way for more publication of semantic statistics on the Web.

The Semantic Web as originally envisioned by Tim Berners-Lee has been slow to arrive but new Linked Data implementations are evidence of good progress and have created a greater sense of momentum.

The DDI Standard

What is needed for Linked Data to work for social sciences microdata is agreement on the form of the vocabulary describing such data – in essence, standardized metadata. In the world of social, behavioral, and economic research, the standard that has emerged is the Data Documentation Initiative (DDI). This agreed-upon format for exchanging information (data and metadata) is based not on the Linked Data technologies (i.e., RDF) but on the popular XML technologies.

An international committee has shaped the DDI specification and continues to refine it. DDI represents the understanding and expertise found among data producers, archivists, and researchers of their common business – the creation and dissemination of useful data. As such, the DDI model is a good one for any representation of the data, whether that is in the XML in use today, or in a Linked Data representation, or in any other form. It is likely that the data themselves would never be expressed in a Linked Data form, but rather as SAS, SPSS, Stata, or another statistical package format.

Discussions have taken place regarding not only the underlying Linked Data vocabularies, but also the tools that will be needed to make quantitative data published into the Web of Linked Data useful. The key here is to have a generalized interface to data – termed a "generic API" – so that tools can work on any dataset they discover, rather than just on the single set of data they were designed to work with. This issue is very closely related to the creation of agreed vocabularies – if there are no tools, having agreed vocabularies is unimportant; on the other hand, the generalized interface is essentially nothing more than a standard implementation of those vocabularies.

Linked Data Issues

Issues of confidentiality and data provenance are not specific to Linked Data expressions of data, but Linked Data magnifies these concerns. Microdata collected from individuals inherently carry the risk of disclosure, despite the best efforts to anonymize them, and the risk of identifying respondents increases along with the number of links provided. Given this, it makes sense to focus initial efforts at Linked Data on social science metadata until security questions are resolved.

Lack of information about provenance is often a problem in the Web. Good quality metadata should be clear in terms of provenance, but the Linked Data community is also developing techniques for addressing issues of data provenance. Notable among these is the Provenance Vocabulary Core Ontology Specification (http://trdf.sourceforge.net/provenance/ns.html).

Merging data appropriately is another complex matter, especially when the data span scientific domains. Modeling the data and metadata in the right way is very important in order to allow for generic tools to be created. The benefits of having Linked Data decrease unless we have very broadly applicable standards in this area.

The Way Forward

It is clear that the best Linked Data expression for quantitative data can only be crafted with input from all the stakeholders. There is a clear path forward from here to maximize the benefits of the Linked Data technology with respect to quantitative data of all sorts, even if all of the answers are not yet visible. All of the stakeholders in this picture must be engaged to provide a solution that will meet the many and various requirements they present. The dialogue must include not only data producers but representatives of the data.gov initiatives, data archives and libraries, professional journals, statistical offices, the user community, and researchers themselves. Once a reasonable output from this work exists – or, perhaps, as the outputs manifest themselves – there needs to be wide-

ranging public review, which is well-promoted. Without this, it will be difficult or impossible to attract the attention of and engage with the implementers of Linked Data technologies.

It will ultimately be necessary to find a home for these products as freely-available technical standards that can be broadly adopted. A strategy must be found for publishing, governing, and maintaining this work in a way that is responsible and responsive to the needs of all stakeholders. Without a solid international imprimatur, acceptable to all parties, the standards will become less useful to statistical organizations and governmental bodies through concerns about reputational risk.

An agreed-upon strategy for the basis of a distributed architecture leveraging the Linked Data Web must be mapped out and implemented. The technology tools for providing good visibility to data exist, but the foundation work needed to support them has yet to be done. This is a key infrastructure task, but one that is not possible without having a shared set of models: the Linked Data vocabularies that have been developed and approved by the stakeholders.

Several questions remain. How do we fund and organize this work? Getting the attention of the stakeholders has not been the problem – these are issues that are currently very much in focus. What organization is suitable for governing and maintaining this work? Existing standards bodies may or may not have all the needed qualifications and capabilities. Will this require a joint effort between several standards bodies in the form of a memo of understanding or other agreement?

We do not have the answers to these questions, but they are questions that we will need to explore. What is apparent is that, unless we undertake that exploration in a collaborative way, we will not realize the promise that the new technologies hold for the SBE sciences.

References

Cygniak, Richard, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. "Semantic Statistics: Bringing Together SDMX and SCOVO." Paper presented at the Linked Data on the Web 2010 Conference, April 27, 2010, Raleigh, North Carolina, USA. http://events.linkeddata.org/ldow2010/papers/ldow2010_paper03.pdf

Google Group on "Publishing Statistical Data." http://groups.google.com/group/publishing-statistical-data

Data Documentation Initiative (DDI) Alliance. http://www.ddialliance.org

Creative Commons License