Open Data Foundation

Open Data and Metadata Standards: Should We Be Satisfied with “Good Enough”?

Arofan Gregory, Executive Manager, The Open Data Foundation
16 June, 2011

Abstract

This paper provides a brief description of how the collaboration between open government initiatives, the Linked Data/RDF community, and experts in the field of statistics and research data management could enormously improve the usability of quantitative government data on the Web.

The Situation Today

Today, we are seeing increasing pressure for data collected by governments at taxpayer expense to be published in a form which is accessible to citizens. In the US, we have the “data.gov” initiative; in the UK, there is a similar initiative called “data.gov.uk”, and other initiatives of this type are coming into existence in many countries.

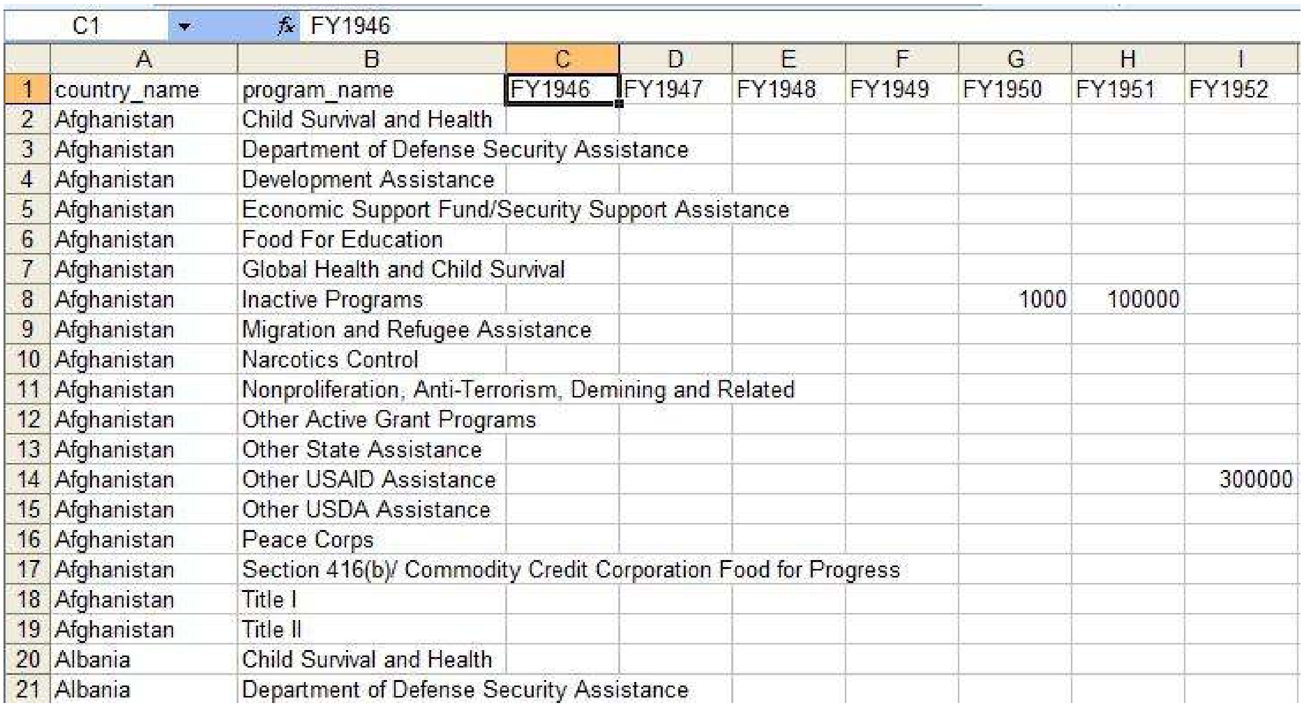
At the same time, there are many new tools for data visualization, such as those promulgated by Google. Ideally, generic data visualization tools would provide access to the government data published on the Internet. While this is happening, in a small way, the widespread access to government data which is the goal of the open data initiatives is not being fully realized.

What we see very often on sites such as data.gov are data files – typically in Excel, CSV, or similar spreadsheet formats – which are publically available, but are not very useful, lacking enough information for those who access them to use the data they contain. Often, potential users must do a lot of homework before they can even begin to understand the contents of the files. In some cases, there are RDF versions of the data files available, which is an improvement on the CSV formats, but even here the use of RDF alone is not enough to give data users all the information they need.

To give a simple example, if we go to the data.gov site in the US, we can find a data set labeled “US Overseas Loans and Grants (Greenbook)”. We can download a CSV file or an RDF file. Below is a snapshot of what the file looks like when opened in Excel.

If we look at the file, we see the column headers “country name”, “program name”, and then a series of column headers “FY 1946”, “FY 1947”, “FY 1948”, and so on. In the table, we have numbers such as “1000”. Now, one can easily guess what “country name” and “program name” represent, and it is not a huge stretch to guess that “FY” stands for “Fiscal Year”. There may be a lack of clarity about exactly what the definition of the countries or programs are – presumably these questions can be completely answered by reference to the “Green Book” referred to in the title of the data set.

But what are the numbers in the table? Are these US Dollars (one assumes so)? But have the data been adjusted for inflation? Are they individual dollars? Thousands of US dollars? In order to get definitive answers to these questions, the user will need to do some homework.



C1 fx FY1946

	A	B	C	D	E	F	G	H	I
1	country_name	program_name	FY1946	FY1947	FY1948	FY1949	FY1950	FY1951	FY1952
2	Afghanistan	Child Survival and Health							
3	Afghanistan	Department of Defense Security Assistance							
4	Afghanistan	Development Assistance							
5	Afghanistan	Economic Support Fund/Security Support Assistance							
6	Afghanistan	Food For Education							
7	Afghanistan	Global Health and Child Survival							
8	Afghanistan	Inactive Programs					1000	100000	
9	Afghanistan	Migration and Refugee Assistance							
10	Afghanistan	Narcotics Control							
11	Afghanistan	Nonproliferation, Anti-Terrorism, Demining and Related							
12	Afghanistan	Other Active Grant Programs							
13	Afghanistan	Other State Assistance							
14	Afghanistan	Other USAID Assistance							300000
15	Afghanistan	Other USDA Assistance							
16	Afghanistan	Peace Corps							
17	Afghanistan	Section 416(b)/ Commodity Credit Corporation Food for Progress							
18	Afghanistan	Title I							
19	Afghanistan	Title II							
20	Albania	Child Survival and Health							
21	Albania	Department of Defense Security Assistance							

So, if I go to Google and type in “Green Book Overseas Loans” I find a website where I can determine that yes, these are figures in individual US dollars. I can even see the source of the data. I am still unable to answer the question “Have these figures been adjusted for inflation?” with 100% certainty, even though in this case I can guess that they probably have not. Many other potential questions remain unanswered, as well.

Is this data file useful to me? Of course – it is far better to have *some* access to this data than to have *no* access to it. But is it as useful as it could be? No, it is not.

Ideally, all of the data found on this site would have a standard set of metadata associated with it, so that any type of user – be it a human being or a web application of some type – could fully understand the data and use it for visualizations and other purposes, immediately, and with no additional investigation or homework needed. If the data were fully documented in this way, I could even begin to compare and merge data from different sources.

Best Practice in the Field of Data Management

Within the field of data management, there exist some good standard metadata models which could be used to help open data initiatives realize this vision. There are two major models: The Statistical Data and Metadata Exchange (SDMX) model, which is used by official statistical organizations to mark up their aggregate data in XML; and the Data Documentation Initiative (DDI), which is used as a metadata model for describing survey data and other types of micro-data. It is important to understand that both SDMX and DDI focus primarily on quantitative data – this is a different use of the term “data” than that prevalent in the Linked Data community, where “data” can be other types of information as well.

Both of these standards are being increasingly adopted across the globe, and they provide a capacity for the creation of generic tools which can work with any data, so long as the metadata is available in the standard forms.

Today, both DDI and SDMX are XML standards, but there has been an interesting development within the data.gov.uk project. Here, the UK’s Office of National Statistics brought together SDMX experts, experts in the field of Linked Data and RDF, and members of the UK open data community. After working together for some months, an RDF vocabulary was developed, based on the SDMX model. The results of this work are now under review, titled the “Data Cube” vocabulary. (This can be found at http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html).

Further, there is a workshop being organized by some of the members of the DDI community, working with experts from the Linked Data/RDF community, to produce a companion vocabulary for describing micro-data. The experts who have been creating the Data Cube vocabulary will be involved in this DDI-centric work as well.

What Does This Mean?

What we are seeing is this: people from the open data community are starting to join forces with technology experts and experts in the data management field. We have three standards communities involved here: the Linked Data/RDF recommendations from W3C, the SDMX standard, and the DDI standard. This combination is a powerful one.

RDF has claimed the attention of those in the open data movements, but RDF alone is not enough: it needs vocabularies based on proven best practice around the use of statistical and research data. This point has been clearly demonstrated by the work being done on the SDMX-based Data Cube vocabulary.

The collaboration between these different communities is significant because it could provide a means of getting far more value from our open data. If we want to realize the power of existing best practice within the field of statistical and research data, and apply it to open data, we now have that opportunity.

Users of open data could deploy generic tools for accessing, understanding, and comparing data coming from many sources. While many different technologies promise this functionality, in truth it is only the combination of technology and expertise coming from these different communities which can make it a reality.

For this to happen, we need first to develop the needed RDF vocabularies, and then to adopt them as we publish data to the Web. This work has started, but its success is not yet guaranteed. At this point we have only the opportunity – without support and adoption, this vision of really useful open data will not be realized.

We must ask ourselves this: if we are going to work to provide access to government data, shouldn't that access be the best possible? There is no good reason why not.

To learn more, you can contact the author at:
agregory@opendatafoundation.org .

Also, please visit the websites of the SDMX Initiative (www.sdmx.org) and the DDI Alliance (www.ddialliance.org), as well as that of the Open Data Foundation (www.opendatafoundation.org).