Building the Web of Data

Position paper for the W3C Workshop on the Web of Things
Phil Archer, W3C Data Activity Lead <phila@w3.org>.

Introduction

The rise and success of the open data movement has lead to many countries, regions and cities publishing their data through portals: data.gov, data.gov.uk, dados.gov.br etc. Scientific research and cultural heritage are among other areas where data is made increasingly available for reuse by others (e.g. dnadigest.org, europeana.eu) . The use of the Web as a data sharing platform by commerce is a little way behind this: manufacturers do not (yet) routinely publish their product data although exciting work at GS1, the standards body behind product bar codes and more, has the potential to change that. The Web of Things has the potential to radically change the way people think about data on the Web and to increase both its volume and variety.

The W3C Data Activity

In December 2013, W3C enacted a plan that had been developed over the previous two years. The Semantic Web and eGovernment Activities (W3C's name for collections of related working groups) were merged to form the Data Activity. Its overall vision is that people and organizations should be able to share data as far as possible using their existing tools and working practices but in a way that enables others to derive and add value, and to utilize it in ways that suit them. Achieving that requires a focus not just on the interoperability of data but of communities. In other words, the Data Activity is chartered to promote the implementation of multiple data formats including, but not limited to, Semantic Web technologies.

An immediate expression of this was the formation of the CSV on the Web Working Group. The majority of data on the Web is published in tabular format, principally CSV, and the new WG will make this more readily processable. The extensive use cases and the Tabular Data Model show the approach being taken: metadata about CSV files will allow automatic discovery of annotations at every level from the table as a whole, down to rows, columns and individual cells. Annotations can include datatypes, provenance and more. Linkage from the CSV data to the metadata will be via a URL.

Of potential greater interest in the context of the Web of Things is that the link can go the other way, i.e. **from** the metadata **to** to data. That will again be via a URL which of course can include parameters that a service would use to define the tabular data to be returned. The Hydra Community Group lead by Markus Lanthaler is interesting in this context. It is defining a vocabulary for describing RESTful APIs. Community Groups are designed to allow communities to form and to do initial work in a specific area, such as developing a vocabulary or testing out new ideas, but they don't create formal W3C standards.

Is Hydra something that should be fully standardized by W3C? Are there other vocabularies that the WoT needs to make data integration easier?

Another expression of the Data Activity's mission is the Data on the Web Best Practices Working Group. This group is about developing the ecosystem of data on the Web including, but not limited to, open data. How should data be published? What guarantees of quality, persistence and provenance are needed to foster trust in data? What return on investment do publishers need to encourage them to publish, or continue to publish, their data?

The DWBP is developing vocabularies to encourage this exchange. One of these is an extension to DCAT to cover quality and persistence, whilst the other is focused on data usage. The carrot for developers here is that by describing their application machine readably their work becomes more readily discovered. By citing the dataset(s) they use, they encourage the continued provision of that data. Straightforward usage of data in an application is one class of usage but another is citation of data in scientific research. The work of the Force 11 group on Data Citation is of direct relevance therefore.

One of the important inputs to the DWBP WG is an EC-funded Thematic Network, Share-PSI 2.0. Coordinated by ERCIM/W3C it comprises 44 partners from 25 countries working on implementing the revised EC's revised Directive on Public Sector Information. A workshop series is eliciting experience in this area that will help to inform the working group and ensure that its outputs are relevant in the real world.

Among the areas of best practice under discussion is the design of persistent URIs. Many Web sites are designed to be ephemeral with publishers routinely removing material from the Web whether as a deliberate act or as a side effect of a site re-design. For data URIs, if nowhere else, this is malpractice. If the Web of Data has one central tenet it is that HTTP URIs should be used as identifiers. The features will be well known to participants in the workshop but they are worth highlighting. HTTP URIs are:

1. Globally unique: the same URI identifies the same thing wherever you see it.
2. Dereferencable: you can look them up and get back either the identified thing itself or metadata about that thing following a simple redirect.
3. Technology neutral: data returned from a URI can be in any format including JSON, XML, Turtle, ATOM, YAML, HTML etc. via content negotiation.
4. URIs are functional: they can be APIs to any data system.

No other system offers this combination out of the box. If there is a weakness in the system it is that the culture of ephemeral Web pages runs deep in many people's experience. If URIs are to be persistent they need to be *designed and managed for persistence* and the W3C Data on the Web Best Practices WG will offer advice on this, based on existing work such as the Study on Persistent URIs conducted under the European Commission's ISA Programme.

URIs are defined as being dumb strings with zero semantics but patterns are useful for developers. The little-used POWDER standard offers a semantically-valid way to make statements about groups of resources based on URI patterns — which may or may not be useful when making statements about thousands of sensors. In the context of the Web of Things, are there particular factors to bear in mind for URI design?

An aspect of the Linked Data/RDF world that causes some surprise and disquiet amongst those using the technologies for the first time is that it adheres to the Open World Assumption. Your RDF triples may be syntactically valid but there's no standardized method to validate the data against a particular model. Emphasis on *standardized method*: in fact there are any number of methods of doing this, notably SPIN, Resource Shapes and W3C's

Eric Prud'hommeaux has been working on what amounts to an extension of Resource Shapes called Shape Expressions. This work ties in with long standing efforts at DCMI to develop Application Profiles and the expectation is that a new Working Group will be formed around this topic imminently.

Finally, and perhaps of most relevance to the Web of Things, a workshop held in March this year looked at Linking Geospatial Data. Run under the EC's Smart Open Data project, it was organized by W3C with the Open Geospatial Consortium, the UK Government, Ordnance Survey and Google. The outcome of the workshop was that there is a strong desire for a new working group to be formed in which OGC and W3C will work closely together. Among the deliverables being considered for the proposed new WG are full standards for:

1. GeoJSON/GeoJSON-LD (in cooperation with the original authors);
2. the OWL Time Ontology (stuck at Working Group Draft since 2006 but widely used);
3. the Semantic Sensor Network Vocabulary (developed by a W3C Incubator Group in 2011).

What role should the Web of Things play in this (likely) work? NB. the likely WG under discussion here is separate from the Geolocation Working Group which is about to start work on a level 2 geolocation API to cover geofencing and a device orientation API.

Conclusion

The Web of Things promises a Web of Services built on top of the Web of Data. A lot of work is going on, or is already planned to be done, within the Data Activity. However, the use cases arising from the WoT are likely to be quite different from those coming from the open data movement, scientific research data and cultural heritage communities that form the majority of those involved already. Just as it is important for data not to be locked away in silos, it is important for the Web of Things community and the Web of Data communities not to be in separate rooms.

$Id: buildingthewebofdata.html,v 1.2 2014-05-09 15:02:48 phila Exp $