Observing Observatories:

Web Observatories should use Linked Data

Hugh Glaser

University of Southampton and Seme4 Ltd.
{hg@ecs.soton.ac.uk, hugh.glaser@seme4.com}

Abstract
Web Observatories are a major international scientific collaboration concerned with data sources of a heterogeneous nature, and often quite large. Of course, they are not the first such collaboration; the Web itself was born as a response to a similar scientific endeavor. It is therefore appropriate to look at other collaborative activities, and try to learn and use the lessons they have learnt.
We argue that Web Observatories should build in interoperability using current best practices right from the start. We also argue that Linked Data is a best practice, and can provide the basis for a research environment that will deliver the vision of a large group of cooperating Observatories, sharing data and research results to the benefit of all. In addition, we argue that the activity should not start with a major standardization process, but should grow around appropriate standards as required.

1 Introduction

In collaborating with scientists from subject domains, technology providers commonly find that the scientists see the subject matter and challenges of the subject as the major obstacles to progress, and relegate the technological support activities such as data representation, storage, communication, etc. to a role that will "just follow" once the main problem has been properly addressed. In fact, technology providers are often called in after data has been gathered, long after the start of the research activity, to address problems of interoperability, data representation, etc.. This way of proceeding, where the enabling technology is patched into existing systems, is a false economy, and reduces the value of the research: It is much more expensive to retrospectively revisit data acquisition to make the data that has been acquired fit for reuse and interoperability, and indeed it frequently turns out that data and metadata that is needed for reuse would have been useful for the substantive research, if it had been available from the start of the activity.

Of course, there are many challenges from a wide range of disciplines to build and run a Web Observatory, and these are rightly seen as the major challenges. But building for interoperability between Web Observatories (and other research activities) using best practice should be a basic principle of the projects, and should be an objective from the start.

In building a Web Observatory and running it, it can be frustrating that the places and sources being observed could make things much easier if they would build for interoperability and ensure that their data and metadata was published appropriately. It would be ironic indeed if it turned out that the only way of accessing what a Web Observatory did was by using web observatory techniques to observe it, rather than looking at the data and metadata it was publishing. One might ask "And why beholdest thou the mote that is in thy brother's eye, but considerest not the beam that is in thine own eye?" (Matthew 7:3, King James Version).

Space is too limited in this position paper to provide an explanation of Linked Data. The Wikipedia entry (`http://en.wikipedia.org/wiki/Linked_Data`) is a reasonable place to start for those who are not familiar with the subject. We only note that Linked Data is a way of publishing and consuming data that provides HTTP URIs to identify things and enables machine-interpretable formats to be returned when such URIs are retrieved from the Web. The standard model for the machine-interpretable content is the Resource Description Framework (`http://www.w3.org/RDF`) (RDF). Although a seemingly small change to the way things are done, having strong identifiers for everything, and the ability to retrieve a "meaning" is a fundamental change to the way in which things are published. Notably, it allows statements such as source and provenance to be made about data, as well as the experimental procedures used, along with agreement on the "meaning" of terms in vocabularies.

2 Related Systems

We choose some related systems as illustrations – there are of course others; in particular Life Sciences (Functional Genomics and related fields have long used Linked Data).

2.1 Astronomy: International Virtual Observatory Alliance

Since the Web Observatory activities are named by analogy with the field of Astronomy, it is appropriate to look first to that field for guidance.

Indeed, Astronomers have collaborated for generations to share their observations, and have more than a decade of experience of trying to create an environment in which researchers can seamlessly access material from a wide range of sources, the International Virtual Observatory Alliance (IVOA) (`http://www.ivoa.net`), which is an aggregating body for 19 mainly national VOs, founded in 2002.

Looking at the Technical Architecture for the IVOA (`http://www.ivoa.net/Documents/Notes/IVOAArchitecture/20101123`), we of course find that such a mature system has significant complexities. But significant components from the VO Core that facilitate the whole activity are based on *Resource Identifiers*, *Vocabularies*, *Semantics* and *Resource Metadata*. Although the IVOA was formally constituted after the early Semantic Web work, it actually predates Linked Data. Nevertheless, its structure around Resource Identifiers echoes the Linked Data principles, and already the Vocabularies standard for VOs specifies that RDF should be used, along with the Simple Knowledge Organization System

(http://www.w3.org/TR/skos-reference) (SKOS); it is clear that the IVOA is looking to Linked Data and Semantic Web Technologies to deliver their improvements to their already impressive interoperability capabilities.

2.2 Cultural Heritage: CIDOC CRM, Europeana

In the field of Cultural Heritage, such as Libraries and Museums, there are often significant datasets that cover related resources that benefit greatly from interoperability (although the datasets are not really big data at the same scale as Astronomy or the Web). We find that over the last few years, as with the IVOA, the standards have been reengineered to be able to use Semantic Web technologies as their basis. Thus, the well-established CIDOC CRM (http://www.cidoc-crm.org), which also predates Linked Data has recently been enhanced with standards for Linked Data identifiers.

Significant collaborative research activities in this area, such as Europeana (http://europeana.eu) and ResearchSpace (http://www.researchspace.org) have moved a long way towards using Linked Data to realize the true value of the data they are collecting, processing and republishing.

2.3 Big Physics: CERN

It might have been expected that the cradle of the Web would be embracing the new technologies for its latest system, the Large Hadron Collider (LHC), but it seems this is not the case (http://home.web.cern.ch/about/computing) Although Linked Data is used in the original application area of document management, it is not used as a basis for the exchange of data. We can speculate that there are a number of reasons for this. Firstly, the lead time on such a large system is such that it was started (in 1998) long before Linked Data, and the investment is so great, with so many partners and suppliers that any adjustment would be prohibitively expensive. Secondly, the scale of the data and processing challenges are unique, and so the scientists chose to invest all their efforts in efficiency. Thirdly, the effort is clearly well funded, and could afford to build bespoke systems at the limits of the technology of the time.

But the negative side of this extreme technology is huge costs, lack of interoperability and enormous barriers to entry – it is a huge investment for a new partner to begin to collaborate on the data being used. An international Web Observatory infrastructure that chose to parallel the Grid Computing solution of the LHC would be unlikely to achieve many of the organizational and social objectives of the participants.

2.4 eScience: Wf4Ever

As the “little brothers” of LHC-like experiments, wider eScience has often embraced Linked Data. A particularly nice example is Wf4ever (http://www.wf4ever-project.org), which “addresses some of the challenges associated to the preservation of scientific experiments in data-intensive science”, including trying to define best practices. Almost everything can be mediated through Linked Data, and of particular

interest is the ability to document all aspects of experiments, including workflows and provenance (using Linked Data), so that others can understand, verify or even reinterpret the conclusions drawn.

3 Discussion

So what does this mean for Web Observatories?

Clearly the systems described here have all suffered from the changing technological base as the world of metadata representation evolved in the last decade. But like many other systems they have tracked and then adopted Semantic Web Technologies and then Linked Data. For the older systems this was a reaction to the silos they were building, by having internal, local vocabularies, local identifiers, effective metadata, and lacking the ability to easily exchange machine interpretable data and metadata. For the newer systems, it was a recognition that the ability to easily exchange the data and metadata was a crucial part of the non-functional requirements.

Doing the two first and perhaps easy bits of Linked Data, of using http URIs to identify resources is such a low-cost but high value activity that there should be no question of its deployment. Returning machine-interpretable meanings for the URIs in the form of RDF (we would recommend Turtle (`http://www.w3.org/TR/turtle`, `http://en.wikipedia.org/wiki/Turtle_(syntax)`) as the easiest form of RDF) can be more challenging, but if it is built-in from the start, then it is essentially just another format for delivery of the content.

So where is the linkage?

This is the topic that can be the most challenging. It is tempting to decide that the joint activities need joint vocabularies and common URIs before they can begin to publish their data and metadata. There is of course great benefit to having such commonality, as consumers can more easily interpret and aggregate results from different Web Observatories. But it can be a big mistake to look for too much commonality at the start. At its worst, years can go by while communities gather to standardize on vocabularies and ontologies:- years during which the science proceeds, gathering data and processing it and failing to make it available in any useful form to the community. Even if less global standardization is considered acceptable, adherence to rigid standards can have a negative affect on a developing research field such as Web Observation. Were it the case that Web Observation was an established field, with a data model and developed taxonomic classes, then it might possibly be sensible to attempt major standardization, but it is not.

On the other hand, there are standards that can be used, as well as common vocabularies that can be developed. These should be encouraged, as they make it easier for the consumers to access and consume the data and metadata. Typical standards such as SKOS and the Data Catalog Vocabulary (`http://www.w3.org/TR/2013/WD-vocab-dcat-20130312`) (DCAT) are likely to play a major part. But the Web Observatory projects should not be delayed while the different sites work to achieve commonality. Indeed, it is the nature of Linked Data and these technologies that there is no need for such a delay – the technologies themselves facilitate the later work of linkage

and commonality and vocabulary alignment without having to change the original publication of the data and metadata.

We therefore recommend that Web Observatories seek to publish their data and metadata as soon as they acquire and build it; of course where privacy, commercial, or other constraints exist that mean this is not socially possible, then this cannot be done, but the systems should be in place such that it can be published if or when the constraints no longer apply. It is easy to say that the data is not clean, or not of interest, or not quite ready yet, or even that no-one would be interested. To this the response should be the same that has been successful in the field of Open Data, as stated by Rufus Pollock (http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now) and echoed by Tim Berners-Lee (http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html): "No, we want the data raw, and we want the data now". That is, even if it cannot be made Linked Data, then it should still be published – this often ensures that the data owner realizes that publishing as Linked Data is little more onerous than the raw data, and that the perceived technical problems preventing publication were in fact social problems about publishing at all.

A final issue is whether the interoperability that we discuss is hard. If it is easy (and so can be left until later), then it clearly is easier to provide it early, so that the advantages can be gained early; if it is hard, then it is really important that planning for a challenging aspect of the research should be included at every stage. We believe that in fact it is relatively straightforward, and the costs of embracing the challenge early are relatively small.

4 Concluding Remarks

In the rush and excitement of new results and insights being gained from Web Observatories, it is important to think about reuse and the legacy.

In the modern scientific world, reuse and legacy depend on principled and documented experimental method, and agreement on communication standards. We urge Web Observatory researchers to ensure that their activities adhere to best practices of technology support for scientific method. We recommend that Linked Data technologies can deliver much of what is required.

This work is partially supported under *SOCIAM: The Theory and Practice of Social Machines*. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford and Edinburgh. Thank you to Kevin Page and Ian Brown for comments.