W3C Provenance Incubator Group Teleconference 12 February 2010

Agenda:

- Discussion on provenance requirements for eGovernment

Present:

- James Cheney
- Kai Eckert
- Irini Fundulaki
- Yolanda Gil
- James McCusker
- Paolo Missier
- Luc Moreau
- Paulo Pinheiro da Silva
- Christine Runnegar
- Jun Zhao

Invited participants:

- John Sheridan, Office of Public Sector Information, UK National Archives; co-chair of W3C eGovernment Interest Group
- Hans J (Jochen) Scholl, University of Washington, President of the Digital Government Society of North America
- Yigal Arens, University of Southern California's Information Science Institute, Treasurer of the Digital Government Society of North America

Summary:

YG started by welcoming and introducing JS. JS is a member of the W3C eGov Interest Group and works in eServices and strategy at the UK National Archives.

John Sheridan:

- www.data.gov.uk is equivalent to www.data.gov
- The UK site uses linked data standards for publishing government data
- JS and his team are looking at how best to express provenance in the linked data context (e.g. how best to apply it to geographical data, statistical data, data about legislation etc.)
- Messages for the Provenance XG:
    - The #1 issue when a URI is requested – i.e. What is this resource? Who made it? Where did it come from?
    - We need reasonably simple patterns for expressing provenance
- Background to UK government data:
    - Data has been collected/created over many years
    - There is variable data quality within data sets
    - It is typical to have incomplete data
    - Example: The UK government Statute Book contains data going back 20 years, data that has gone through a conversion process, data that has legal status, data that has editorial status but no legal status – the published data is linked data, has been taken out of different systems and processed
- We want to be able to tell consumers of linked govt data:
    - This is the original source
    - This is what we did to produce this information

so they can understand where the data comes from

- We also want to convey to those persons the capacity that the creator of the data was acting in when the data was created. There is presently no comprehensive list of all parts of the UK government. We are constructing a list from multiple sources. We want to publish the government data with details regarding which government organisation it came from and in which statutory capacity that organisation was acting when the data was created.
- Provenance is a key issue
- We are able to convey provenance information but we do not have great patterns to quickly pick up and use. We have looked at existing provenance vocabulary and the Open Provenance Model but we are looking for something simple. We are having to establish patterns now. Our main priority is developing patterns for algorithms for how the data is processed, and as much as possible, for the underlying data. It is a complex and important problem.

Discussion:

Q: How does www.data.gov.uk compare with www.data.gov?

JS: The main similarity is that they are both catalogues of government data sets. However, the UK version aimed more at developers who want to use the data rather than ordinary citizens wishing to merely view the data. Transparency is important. We allow programmatic access and provide linked data sets in RDF (Resource Description Framework).

HS: Another difference may be that the US has >150,000 data sets and the UK is behind in quantity. The massive amount of US government data available creates a problem in itself. This means that what is published, what is available and the quality may not be always as one would expect.

YA: The US version is more focussed on providing information to citizens, not just developers. The US government does make an effort to describe the format, how often it is updated, the source etc.

JZ: I am working with Olaf (Hartig) on Provenance Vocabulary. Is there some documentation or a summary so we can understand where we can improve and fix your requirements?

JS: There is not much documentation as yet.

I would like to have a more detailed conversation about that. The core of our requirements is simple vocabulary to express things about provenance, and particularly the process that data has gone through to become linked data and to connect who created the information and in what capacity were they acting. We would like simple and practical patterns. We do not want our own vocabulary. We need patterns that we can apply to use cases (for example: we have statistical data in standard format, converted to RDF, excel spreadsheet csv files that have been converted into models using code, etc.)

We would like to explore these issues with provenance vocabulary people as we ramp up our linked data. They might want to look at some of our internal documentation and give us ideas as to how we can use your staff. Some funding may also be possible.

YG: It would be of interest to the Provenance XG to have a special group forum to work with you on this. It is very important. I would be happy to have a telecom devoted to this or maybe JZ and others interested on working on this with you could arrange something.

JS: We have very tight time frames, but we are keen to experiment. If they would like to give us their technology to try in the next couple of weeks, we could reflect on how it works. We just need a sense of who we should engage with and work with in a real and practical context to see what works. In the areas of linked data and versioning there is a huge amount you can do with current provenance approaches, but there is too much flexibility = too much choice = too hard. We need more practical approaches and to deal with real data.

In converting data to put it on the web, one of the people we have been working with is Jenny Tennison http://www.jenitennison.com/blog/node/133. A good start would be to reflect on this work and where you have done work on provenance vocabulary, how it might help us.

YG: I will send an email on Monday advising those people who would like to follow up with you.

JS: My email is john.sheridan@nationalarchives.gsi.gov.uk. There is a strong political push behind doing this work. Even if it is not perfect, we need something good enough that we can implement and use now.

YG: These issues have come up in our discussions. We may not have used the same words, e.g. in role and capacity of the entity. What about republishing? Has this come up as an issue?

JS: When we are creating URIs for the Statute Book, we would like to able to make statements like "this data was created by this entity in accordance with this power under such and such act". We have bits of this, but we would like the capacity to make more sophisticated statements alongside attribution and details of what we did to manage the data.

This is top of our list of issues – How we have a pattern to say this in linked data? We are very keen to try using the work that the community has done up to this point, learn with you and share our experiences with you.

Hans J (Jochen) Scholl:

- University of Washington, President of the Digital Government Society of North America
- My interest is focused on digital government research and I co-organise core conferences in that domain.
- Government is one of the larger providers and holders of data. Starting with the www.data.gov website and now the www.data.gov.uk website, we are entering a new stage of what we can do with public information and how we can use it. It provides transparency to government.
- There are 220 scholars dedicated to this particular area of digital government. I cannot speak to specifics of provenance at this point because I have been involved in another project, but it offers a tremendous opportunity to research (given large quantities of data).

Yigal Arens

- University of Southern California's Information Science Institute, Treasurer of the Digital Government Society of North America
- I am involved in that society as well. I started research back in 1990 in information integration.
- The US govt has entered a new era in this respect. It has made a policy decision to provide data it collects to citizens through the data.gov website. It has made much data available and plans to make more. I recommend you look at the website. The government makes an effort to provide a lot of meta data about the datasets that are there. But, there is data from a multitude of agencies. Consequently, there is a need to standardise data collection and ontologies. They have not solved that problem yet.

- The work on data integration did not pay much attention to provenance issues because they are too hard, but it has become more of an issue now that more data sets available. The focus of earlier research on data integration was on seamless data access - 2 agencies might collect the same data (e.g. price of gasoline) that they want to make available as one answer without requiring one to know which agencies collected the data. There have been some efforts to take provenance into account and to pass some of that information through but there has been nothing comprehensive.

HS:

- The main thrust is just – get these things out!
- Some things online were not ready. We would have liked to have cleaned them up first.
- It is interesting how much evolution you can see on the www.data.gov website. In the beginning, some data sets were without much description and only a few tools. Much has happened since – missing pieces from early criticism (i.e. no tools) has been addressed. This process is ongoing. It is amazing to see how many data sets are available and that searches can be done by agency. A lot is going on but when you talk about data from various sources, there is no vocabulary and no metadata. There is still a lot to do, particularly when it comes to a tool for end-user (man on the street).
- I did not research on provenance but it is a major theme here.

YG: Did you look at tracking provenance in data integration problems? Which problems do you say were not addressed in the past?

YA:

- Traditionally government provided paper or PDF files. There were always lots of caveats related to data that originated with paper (footnote, boxes, notes) - not your standard database way of including data but it was important information. At times, we had to develop a whole tool basically to analyse that information and provide it forward to the person querying the data. When trying to sum data from different places, it is sometimes hard to interpret information that was passed forward. In the end, we rely on someone who gets an answer to the query to assess whether it will affect the validity of the data. This is one example of the difficulties with the data.
- The hope expressed by government officials regarding www.data.gov was that just as individuals have been able to create new products from data freely available on the Internet (e.g. using Google map data), people will do this with the data.gov data. For example, this is what www.fundraise.org did with data about donations to political campaigns. But there is no standardisation or provenance of that site or sites like it.

Floor opened for questions:

YG: What makes government data different from other data on the web? JS alluded to the possible use of data in a court of law. That might make the govt data more important.

HS: I guess it is the authority that government data commands, i.e. that you can use data published by the government in court. It is official data. Other data on the web is not like that. Citizens can hold government accountable for the data on their web. Now people can scrutinize the data and hold governments responsible. We can expect some litigation regarding the data. Data is data, but the authority that government carries is the difference.

JS notes on IRC:

- Authority is the key attribute of government data - it is used to make decisions or enforce laws, or to shape markets (stock markets etc)
- e.g. the inflation rate
- also expressing provenance is one of the things we can do that helps overcome the socio-cultural issues inside government inside government that deters data publishing
- finally, government is a large user of its own data - for example for evidence based policy making
- if we are using data to inform policy decisions, it is important that we know where that data came from and how it was collected
- the statisticians have done much work on this - and of course, governments tend to employ a lot of statisticians

YA: The huge quantity makes it unusual. Potentially all public domain data should be put on the web. But there are issues that come up in collecting the micro data from providers such as implied and explicit privacy and confidentiality agreements. We need to aggregate data to ensure that confidential information is not disclosed but in principle all government data is public.

LM: I am interested in JS's comment that they looked at OPM, described it as over engineered and need something more simple. From a designers point of view, OPM is not over engineered. We can provide justification for each part of the ontology. I do not know how to resolve that issue. That group might push for standardisation. Do we go for something more general use which might be described as over engineered or more specific and targeted? We thought it important to have different levels of description. Maybe that is one of the ways that can address JS's concerns, i.e. high level description of provenance of data and then delve into more detail if needed.

YG: LM has been leading OPM.

JS: LM is absolutely right. The difference between looking at a model which is very expressive and looking at how the model can be used in simpler less expressive ways. We do not want to lose expressiveness of the underlying ontology. If I can repeat the problem, it is: how do we simply deal with this? I agree with you. We need to qualify "over engineering". We want to find things we can use simply. If it is complex, the learning curve steepness is a barrier. Removing that barrier would facilitate wider adoption.

LM: I agree. Not enough efforts have been made regarding how to use and deploy provenance in particular contexts.

JM: I think you are arguing under engineered not over engineer. You are asking, e.g. what do I define as an artefact? OPM is a subset of the one you say is easier. It does not nail it down. It does not go far enough.

JS: That is a really great comment. From my perspective, the question is - How we can apply what has been created by community pragmatically without huge efforts? - because we need to move quickly and want subsets for particular problems, i.e. patterns that we can use for particular circumstances. We appreciate that can be difficult. That's partly why I am here. Unless there are very easy options, adoption will be low, particularly by government. A low ramp on is the key to widespread adoption by governments.

JS notes on IRC:

- re, over engineering, Luc is right about the trade-offs, what we need is something simple to express, which doesn't necessarily mean the underlying ontology has to too simplistic

JM: Perhaps OPM with a profile with linked data context stuff and potential profiles. OPM is just meant to be the top part. It is an interesting issue.

LM: I am very interested in following this issue. We came up with profiles in OPM for this purpose. I will follow up.

YG: Invited JS to attend the next Provenance XG telecom which will be on OPM