Linking Data: Linking Lives - the creation and display of Linked Open Data for Archives

Jane Stevenson

This paper describes two Linked Data projects: Locah and Linking Lives. It introduces some of the basic concepts around Linked Data, and refers to some of the challenges of transforming archival descriptions into Linked Data. It describes both the transformation of data from EAD and the creation of an end user interface based on Linked Data and drawing in external data sources.

The Archives Hub (http://archiveshub.ac.uk) is a JISC funded service that brings together descriptions of archives held across the UK. One of the great strengths of the Hub is the ability for researchers to make connections. They can search for people, organisations, places or subjects across over 27,000 collection descriptions and thousands of series and item level entries. They can also search serendipitously, the index links within the Hub facilitating a lateral search that can take a user across the wealth of content so that they can discover new knowledge for their research.

Linked Data is about making connections. Specifically, it is about connecting structured data on the Web. This immediately seemed to be a very good fit with our aims. In March 2010 the JISC (www.jisc.ac.uk) put out a call for proposals to "expose digital content for education and research", looking for projects that would enable structured data to be made available on the Web, in particular Linked Data. Our proposal was to create Linked Data for the Archives Hub, and we intended to implement best practice to create "5 star" Linked Data (see http://www.w3.org/DesignIssues/LinkedData.html). This work could potentially enable researchers to make new links between diverse content sources, encouraging new connections between people and events, to reveal more about our history and society. Securing this funding gave us an opportunity to explore what Linked Data has to offer for the discovery of archives.

Modelling Our Data

We started off by playing around with ideas for modelling our data. The Hub data is in Encoded Archival Description (EAD), XML markup for archival finding aids. But EAD is quite permissive and there are a substantial number of tags which can be used in various ways, so we made the decision early on that we were going to model Archives Hub EAD, which has a reasonably defined set of conventions, rather than try to think about EAD in more generic terms. However, we still had our work cut out; with over 200 contributors, often using different means to create descriptions, we have substantial variations in content to deal with.

Linked Data uses the Resource Description Framework (RDF) which is based upon the idea of simple "triple statements" that make assertions about things within the data. When considering how to represent EAD data within this framework, the first step was to take a step back from the nitty-gritty of the EAD XML document and think about the relationships we might construct to represent the information. We needed to get away from thinking in terms of documents, and think instead in terms of what those documents are saying about "things in the world". It is a change in mindset from a document-centric view of finding aids, and it can be quite hard to make that leap.

A consequence of thinking differently about the entities within an archival descriptions is that you start to appreciate just how much information is there – information that is accessible to a

human reader browsing a description, but not structured in a way that a machine can easily process. Names, places, subjects, events, book titles, etc., are often not explicitly identified for machines to be able to take advantage of them.

Entities and relationships

Taking a fragment of our data model as an example, and thinking about the relationship between three common entities - the archive, the creator and the repository - gives a sense of how "things" are modelled within RDF.

We refer to the archive within in our model as the *archival resource*. Clearly this is something we want to represent and say something about. It has a relationship with another key entity, the agent who is the creator of the archive, in EAD parlance, the *originator*. The *repository* or institution where the archive is held is a third entity (also an agent) that we want to represent and say something about. This gives us three entities which have a relationship to each other. We want to think about what those relationships are and make this explicit.

Fig 1: diagram of relationships between three entities

The square boxes indicate entities (what you might call real world things). Once these are established, you need to think about the relationships between the entities, indicated in the diagram above. The aim is to build up on this principle, to include all of the entities and relationships, so that you have a basic model for your Linked Data.

Defining Relationships

The relationship between entities, such as the archive and the holding repository, is stated with a "predicate" or property. It is good practice to use terms from vocabularies (such as Dublin Core and FOAF) where possible, rather than create your own terms, because the idea is to start to connect up different data sources, and if you use the same set of relationships you can start to connect datasets more effectively.

In the example above, we want to state that the repository "provides access to" the archive. There could be existing terms that we might use here, but for this particular relationship we created our own property of "accessProvidedBy" because we felt it was a relationship that was not captured in the vocabularies we looked at. The relationship can be stated in both directions, but this is not essential for Linked Data. In choosing properties to use you have to exercise judgment about appropriate terms and think about semantics. In some cases you may feel that the terms available in established vocabularies do not accurately represent the

relationship; in other cases it may be straightforward to use an existing property. For example, you might use the Dublin Core term "publisher" (dc:publisher) for the creator of the finding aid, or you might use the Dublin Core "date" element for the time period associated with the resource (dc:date).

The relationship between the archival resource and the creator led to much discussion in our team and a blog post on the nature of an archival creator (http://archiveshub.ac.uk/blog/2010/07/who-is-the-creator/). For some resources the nature of the creator is clear; for the author of a book you could use dc:creator. But the creator in the archival sense is not necessarily the author; it is the person or organisation that was responsible for creating or accumulating the collection as a whole. Because of this unique role we decided to create our own term, which meant that we could define it ourselves.

Biographical and Administrative History

Another example of a decision around modelling involved the "biographical or administrative history", known as bioghist in EAD parlance. We decided to make this a *resource* in its own right, which means giving every bioghist a URI. We could have simply used a text value (known as a *literal*), but using a URI enables us to provide information about it as a distinct resource; it can become a *subject* in its own right as well as being the *object* of a statement. For example, we can say that the bioghist is *about the originator* (creator) of the archive. We could potentially analyse the text of the bioghist as a source of further information, and as it has its own URI we can be explicit about the source of the data. Giving the bioghist a URI also makes it into a resource that others can refer to – they can use our URI in their own statements.

Index Terms as Entities

The Hub supports the provision of the following index terms: subjects; personal names; family names; corporate names; place names; book titles; genres; functions. We also recommend the use of recognised thesauri or authority files from which the terms should be drawn. We represented these terms as entities within our model, establishing the relationship between the archival resource and the entity as "associated with" because the relationship is by nature non-specific.

Taking the Papers of Beatrice and Sydney Webb as an example of a typical archive (http://archiveshub.ac.uk/data/gb0097passfield), you can see that there is a huge amount of information within the description. Topics referred to include: the history of socialism; Charles Booth; the Fabian Society; the Labour Party; the suffrage movement; the USSR in 1932; the London County Council around the 1890s and 1900s; the London School of Economics; employment insurance. The list goes on. Ideally, what we want to do is make these topics easy to identify – to discover. The indexes (within the controlled access area) provide these terms as structured entries, along with personal names, corporate names and geographical names. Most Hub descriptions similarly include key subjects and names as structured terms, which is a great bonus when creating Linked Data.

We modelled the names within the controlaccess area as concepts. So, "Beatrice Webb" and "George Bernard Shaw" are concepts within an archive description. In particular, they are concepts according to a set of rules, such as the National Council on Archives Rules. The *concept* of Beatrice Webb according to the NCA Rules (the notion of the person within this rules system) is related to the *physical entity* of Beatrice Webb (the person), and we can include a triple statement to establish this. This conceptualisation adds a level of complexity, but it can be useful to think in terms of the distinction between the actual physical entity and a representation of that entity. The FOAF vocabulary (http://www.foaf-project.org/) has the

property of "foaf:focus" to represent the relationship between the conceptualisation and the thing conceptualised (person, place etc), to support exactly this convention.

URI Patterns

If you do not want to provide further information about an entity, but simply want to state the value, then you can use a *literal* value. In our case for example, access conditions, appraisal and accruals were just given literal values. But at the heart of the Linked Data approach is the principle that all the "things" we want to say anything about should be named using a URI, and URIs should use the http URI scheme, so that they can be easily looked up or "dereferenced" in order to obtain some information provided by the URI owner about the thing. In most cases, we created URIs under the "archiveshub.ac.uk" domain that Mimas owns. We chose these URIs and put in place the mechanisms to ensure that their dereferencing results in the provision of some useful information. We used UK Cabinet Office guidelines in the construction of our URIs (Designing URI Sets for the UK Public Sector), so that they would be constructed according to recognised good practice, and provide further information about the entity.

The pattern that we used is:

http://{domain}/id/{concept}/{reference}

So, for example, the URI for the concept of a person would be:

{root}/id/person/rules/{personid}

And an example of this is:

http://data.archiveshub.ac.uk/id/person/ncarules/skinnerbeverley1938-1999artist

As well as constructing our own URIs, we used some existing URIs, defined by other agencies who also provide dereferencing for their terms. For example, we used lexvo.org for our URIs for language, so "English" has the URI of http://lexvo.org/id/iso639-3/eng.

Unique and Persistent Identifiers for Archives

We used the Archives Hub persistent URI for the archival resource, which enables the end user to follow through to the description of the resource. This choice led us to think more carefully about Archives Hub persistent URIs, and in fact, we ended up embarking upon a substantial project to improve the consistency of the content and markup for these references. This is an example of where Linked Data work can have a major impact upon the source data, because it exposes weaknesses in the data. The end result is an improved dataset, but it has been a time-consuming and sometimes problematic process for an aggregator with over 200 contributors and hundreds of thousands of references to ensure consistency across the board. A major issue we found was the repetition of "unique" identifiers. This appeared to be largely due to human error, and was manifest particularly in very long descriptions, but we realised that it could also happen where the combination of repository code and local reference when brought together created identical references. For example, a repository code of 133 and a reference of 4MS1 would create the same URI as the repository code of 1334 and a local reference of MS1

http://data.archiveshub.ac.uk/id/archivalresource/gb1334ms1

A further issue was with very long identifiers, that do not make for very coherent URIs (we felt that human readability was a consideration, albeit not an essential requirement). We

found ways to address both these problems, and also used the opportunity to audit the Hub references for other issues that we could resolve.

Introducing Linking Lives

The Locah project aimed to output Linked Data, to provide views on the data and a SPARQL endpoint for querying the data, to document the process through the blog and to provide a stylesheet for the transformation of Archives Hub EAD into RDF XML.

It seemed to us that the next logical step in the Linked Data journey was to create some kind of proof of concept. Whilst the premise behind Linked Data is that you open up your data for others to consume, and provide the potential for innovative ways to combine different datasets, we felt that we needed a pro-active approach, developing our own front-end; something to demonstrate the potential benefits of Linked Data for end-users. We wanted to build on the initial investment in the Locah project and put Linked Data to the test in a "real life" scenario. Our proposition was that this could potentially connect archives more effectively to the wider information landscape, bringing them together with other sources to benefit researchers. It is important to state that for these reasons we wanted to have an interface based entirely on Linked Data (that is, data in RDF) rather than a hybrid approach, which could include non Linked Data sources.

Why Linking Lives?

We discussed a number of ideas around which we could create an interface. The obvious options were to base it around subjects, events or names. We decided on a biographical approach because it would clearly be of value to researchers, we felt it would be relatively easy to scope, and we had already done some matching of names within the Archives Hub to names in external datasets. Our Linked Data output includes statements using the <sameAs> property where we specify within our Linked Data that "x person in the Hub data is the same as y person in VIAF" (the Virtual International Authority File, http://viaf.org/).

Linking Lives is therefore about focussing on individuals as a way into both archival collections and other relevant data sources. The Archives Hub data is rich in information about people, organisations and events, and we wanted to highlight this, as well as putting the data within the context of a range of data sources in order to provide a biographical perspective, in contrast to the more traditional interface for archives that focuses on the collection description. Researchers do not usually have an archive collection in mind when they start their research, and they may not be familiar with primary sources. A biographical resource is a familiar starting point that can lead them to relevant collections and help them to make connections between people and events.

Interface Design

We decided to create a simple interface where one page would represent one person. We have had a number of ideas about ways to present the data, and we have tried out some visualisations. But we wanted something sustainable and extensible, where we could pull in a variety of external data types - text, images and links. Our interface uses the content boxes that are a familiar feature on many websites, and using these enables us to present different data sources as discreet parts of the interface, which is important if we want to be able to clearly identify the source of the data.

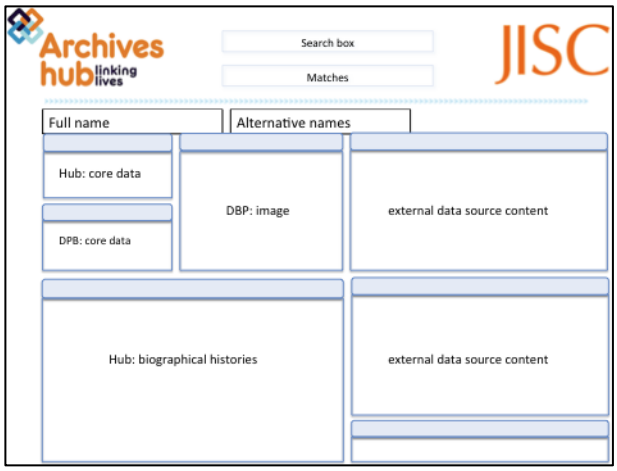

fig 2: wireframe for the Linking Lives interface

The name displays at the top of the main display, and below this a box contains key information that comes from the archive descriptions – life dates, occupation or status, family name and title. We decided to add place of birth and death as additional core information, provided by DBPedia (see below). We placed the image in the centre, as we felt this would make the interface more visually engaging. We intend to have a tab to list alternative names, which are provided by various sources, including VIAF.

We put a large box on the left-hand side to contain the all-important biographical notes for each individual that are typically created by archivists when they catalogue the material. Beyond these key boxes, we decided that we would explore different options and experiment with the data that we could bring into the interface. This meant we did not have to decide on the final content, and indeed, it means that we can continue to add content over time, beyond the end of the project.

One of our ideas is to add an element of personalisation, by enabling end-users to pick and choose boxes and move them around. This remains an option, but may not be do-able within the timescale of the project.

The Challenges of the Source Data

Working with aggregated data from so many sources, created over a long period of time and often migrated between different systems is a challenge. The data is inevitably inconsistent and there are errors that interfere with the data processing. There are, broadly speaking, two alternative approaches to working with problematic data – you can find ways round inconsistencies through the transformation process itself, or you can address the problems at source. We have written about some of the issues with the data that we have faced on our blog; the biggest issue has been with the identifiers for the archives themselves.

The full identifier for the archive comprises the ISO code for the country, the UK Archon code for the repository (http://www.nationalarchives.gov.uk/archon/) and the local reference for the archive, e.g.:

GB 983 UWA

On the Hub the primary role of this reference is to be a visual indicator displayed to end-users, so a level of inconsistency in the make-up of the reference within the XML document might not be a problem as long as we display it correctly, and the only part the end-user really needs to see is the local reference. But there is a lack of consistency in the structure of these

identifiers and how the country code, repository code and local reference are marked up in the XML. Sometimes the country code and repository code are not included, and we have to work round this, but it is far harder to work with such a level of inconsistency in Linked Data because we want to create unique and persistent URIs out of the content.

We made the decision to go back to the Archives Hub data and get a level of consistency, addressing any problems with duplicates and very long local references, which do not create very practical URIs. This work will be of benefit beyond the Linked Data work, but it is time-consuming and has delayed the progress of our project somewhat.

It is only one of a number of areas where the potential for working with Linked Data is hampered by inconsistencies. For example, if we had standardised "extent" entries, for the size of the archive, we could envisage a visualisation that would show where the biggest concentrations of archives on any particular topic or person are. But these entries are very variable because in the UK there is no recognised standard for this content, so you can have anything from "10 boxes" to "5 linear metres" to "photographs and drawings in 3 outsize boxes".

Working with External Datasets

When working with data that comes from external sources you have no control over the data. You may have problems if it is inconsistent or if it changes. This is one of the major issues with Linked Data. By building an end-user interface that will become part of the Archives Hub service, we should be able to get a very practical perspective on what this might mean over time.

The persistence of URIs has often been cited as an issue with Linked Data, and although it is certainly not a problem unique to the Linked Data approach, it does become particularly problematic when the aim is to present a coherent and consistent information source that relies upon external URIs. So far we have not had any problems, as the URIs have been maintained, but we believe that this is an issue that needs to be monitored and assessed over time.

We have had variable success with linking to different datasets and pulling in data. To do this you need relevant content and you need the right "hooks" to pull it into the interface. We found that a number of data sources do not provide all of their data as Linked Data. Simply looking at the web interface can be misleading – you have to dig into the RDF and see what is there. For example, VIAF provides a list of selected titles for authors, but this information is not included within the Linked Data. In addition, some data sources do not provide a SPARQL interface, which is what it typically used to query data. So far we have struggled to find Linked Data that includes connections between people. For example, a simple statement that "x person knows y person". Our hope was to include these types of relationships, as we wanted to build up a resource that would show connections between people.

We created our own Wiki in order to list different datasets and provide summary notes about them. Datasets have looked at include DBPedia, OpenLibrary, the Virtual International Authority File (VIAF), Freebase, BBC Programmes and Linked Open British National Biography (BNB). It is unlikely that we will be able to add data from all of the datasets we assess within this project, even if they all have relevant and useful data, because of time constraints. But we can continue to use the Wiki to monitor potential data sources, and add them at a later date. We may also make the Wiki public in order to share our experiences and findings.

We agreed from the outset that we wanted to bring in data from Wikipedia (DBPedia being the Linked Data version of Wikipedia, http://dbpedia.org/About). But, as with many other

external datasets, we have hit one significant problem – not all records on Wikipedia have the same information. So, for example, we have provided for space for an image of the individual, but we will not always have that image available. We are considering options for ways to address this issue, and we may take the same approach as the BBC, which includes Wikipedia content in its web pages (e.g. http://www.bbc.co.uk/nature/life/Felidae). The BBC makes clear where the content is from and invites readers to edit the Wikipedia article.

Understanding the Interface

With our interface, we want to show that archives can benefit from being presented not in isolation, but as a part of a fuller picture, alongside different data sources, to create a rich biographical resource. People do not always find dedicated archives sites easy to use. The hierarchical nature of archives, and the nature of collections (which can be anything from one item to a vast collection of items in various different media) can make them difficult to represent online. Combining them with other sources and presenting them in a different way may facilitate interpretation, but it is essential to evaluate this hypothesis, to find out how researchers react to what they are presented with and whether they believe it is useful for their work.

We have a group of students and researchers from The University of Manchester taking part in an evaluation of the Linking Lives interface. We wanted to ascertain their thoughts about the more traditional archival interface, and get a sense of their understanding of archives, so initially we asked them to visit the Archives Hub and give use their thoughts in response to a number of questions. Our intention now is to run a focus group with these participants where we introduce them to the new interface. We intend to incorporate their feedback into a modified design.

Aside from bringing together different data sources, one of the features of the new interface is that it brings together a number of biographical histories for any one person, if that person has created more than one archive. We are particularly interested to find out how researchers react to this: whether they find it useful and whether the inevitable repetition of information is seen as a distraction.

Conclusions

We have continued to find linked data work challenging, partly due to the fact that it is a new and developing area, with few templates or tools to utilize, partly due to the challenges of working with various external data sources and partly because of issues within our own data. We necessarily needed to take a lightweight approach to project management and to adopt an iterative technical development methodology because it was difficult to set clear objectives.

With limited time and resources for what turned out to be a more complex project than we had initially envisaged, we necessarily had to prioritize. One decision we made was to focus on the interface, rather than the search and navigation elements of the service. If the interface proves to be useful to end users we will continue to develop the search capability and look to integrate it more fully with the main Archives Hub service.

I would say that the biggest single factor in terms of additional work has been cleaning up our own data, and it is likely that other data providers will have similar challenges when outputting effective linked data. We did not have time to look in detail at as many external datasets as we would have liked, but more than this, the linked data space is constantly changing, so new data is created all the time, and improvements made to existing data. This makes it quite a moveable feast, and you have to make decisions about whether to go back to updated datasets and re-examine them, or stick with what you have. This may be a challenge in terms of maintaining the interface. We may find that the need to monitor the linked data

space takes up significant time. We will continue to maintain our linked data interface, and seek to add some more external data sources, and then we will monitor the result, see how much it is used, and how much effort we have to invest in ensuring it is current and all links are operable.

Part of the motivation behind Linking Lives is to assess whether linked data really does provide an alternative way forward. We believe that we are creating a useful and valuable resource, and we are successfully connecting to external datasets using Linked Data principles. Linking Lives enables us to give archives a different context, putting them into a broader knowledge domain, and we will be able to evaluate the response to this approach from researchers. Our hope is that it provides a useful case study for others who are undertaking similar projects.

We have not found it particularly easy to link to external datasets, and my feeling is that it needs to be easier to locate and probe sources to ascertain the classes of things being described and the properties used to describe them. With some datasets that are more "crowdsourced" in nature, There needs to be a clearer sense of the linked data out there and what it provides. The CKAN Data Hub (http://thedatahub.org/) is one attempt to bring data together but it is not comprehensive and not entirely easy to navigate. However, it must be recognised that working with open data in this way is not going to be easy. Simply connecting and cross-searching two datasets using more traditional means can often prove to be challenging; with Linked Data the idea is to be able to access all open datasets in RDF. We have found creating "sameAs" links takes time and effort, although increasingly there are tools to help. Again, it is important that tools are easily available. But the difficulty in assessing the linked data space is that things are still changing quickly and data constantly becomes available as linked data, or there is a re-issue with improved RDF. My feeling is that there needs to be more evidence of the benefits for non-technical end users, as the impression that is often given is that users need to understand RDF and SPARQL in order to utilize the data.

With big players like the Library of Congress committing more fully to linked data with the Bibliographic Framework project, a certain level of optimism in the promise of linked data is clearly still in evidence, and the community is continuing to expand and evolve. There does also seem to be significant and increasing interest and optimism from the LOD-LAM community (Linked Open Data for Libraries, Archives and Museums, http://lod-lam.net). Maybe linked data has evolved too slowly to attract the level of investment necessary to make it a viable business enterprise and attract significant investment as Tim Hodson suggested in his blog.. Could it be that "Big Data" processing offers a more attractive, and more practical option, or is the altruistic goal of opening up data to advance knowledge and benefit research still a strong enough impetus to drive the linked data ideal?

=======

Jane Stevenson (jane.stevenson@manchester.ac.uk) is Archivist and Archives Hub Manger at Mimas, based at The University of Manchester in the UK.
This paper was prepared with the help of Pete Johnston (Cambridge University Library) and Adrian Stevenson and Lee Baylis (Mimas).