In March 2010 the JISC 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. We secured funding for a proposal to create linked data for the Archives Hub, and the Linked Open Copac and Archives Hub (LOCAH) project was the result of this. Running over one year, it aimed to output linked data, provide views on the data, and offer a SPARQL endpoint for querying the data—as well as documenting the process through the blog. We provided a stylesheet for the transformation of Archives Hub Encoded Archival Descriptions (EAD) into RDF XML, which is available from the linked data site that we created: LOCAH Linked Archives Hub.

It seemed to us that the next logical step in the linked data journey was to create some kind of proof of concept. While the premise behind linked data is that you open up your data for others to consume and thereby 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.

The Archives Hub is a JISC funded service that brings together descriptions of archives held across the UK. One of the most important strengths of the Hub is the ability for researchers to make connections. They can search for people, organizations, places, or subjects across 25,000 collection descriptions and hundreds of thousands of series and item level entries. They can search serendipitously, as the index links within the Hub facilitate a lateral search that can take a user across the wealth of content so that they can discover new knowledge for their research.
based entirely on linked data (that is, data provided in RDF and linked to other data sources) 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).

Linking Lives is therefore about focusing on individuals as a way into both archival collections and other relevant data sources. The Archives Hub data is rich in information about people, organizations, 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 visualizations. 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 discrete parts of the interface, which is important if we want to be able to clearly identify the source of the data. (See Figure 1.)

The name appears 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 center, 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 catalog the

---

Figure 1: Wireframe for the Linking Lives interface
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 personalization, by enabling end users to pick and choose boxes and move them around. This remains an option, but may not be doable 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:

1. You can find ways round inconsistencies through the transformation process itself.
2. You can address the problems at the 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, and the local reference for the archive—for example:

<table>
<thead>
<tr>
<th>GB 983 UWA</th>
<th>GB Country Code</th>
<th>983 Repository Code</th>
<th>UWA Local Reference</th>
</tr>
</thead>
</table>

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; 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 around 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 construct 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 project, but it is time consuming and has delayed the progress of our project somewhat.

This 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 standardized “extent” entries for the size of the archive, we could envisage a visualization that would show where the biggest concentrations of archives on any particular topic or person are. But these entries are highly variable because in the UK there is no recognized standard for this content, so you can have anything from “10 boxes” to “5 linear meters” 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 is 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 we have looked at include DBPedia, OpenLibrary, 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). 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 webpages (See for example: 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 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 us 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.

A Technical Perspective
The Linking Lives interface is a web application loaded onto a user’s web browser (the client) from our server.

As such, there were two obvious strategies we could employ for collecting the data together within the application:

1. **Let the server do the data collection.** In this scenario, we identify which data we want to link together and the server follows the links and contacts all the relevant external websites. Once it has completed its collection, the server sends the complete webpage back to the client.

2. **Let the client do the data collection.** In this scenario, the server sends back an empty interface webpage, and tells the client which online sources hold all the data and how to lay it out on the page. The client then makes its own requests to multiple external sources and updates parts of the interface webpage, using AJAX when the requests come back in.

There are pros and cons to each strategy:

**Pros**
- **Save time:** The client makes all the requests individually. If any one source is down, then just that part of the data will be delayed and the rest of the data can carry on loading into the interface.

**Cons**
- **Cross-domain:** Cross-domain scripting is a potential security risk as it allows code from other sources to run as part of the original website.

**Bottleneck effect:** The server becomes a middleman in the process of loading all the data. While it is necessary to query the server for the initial interface and the linked data it holds, all the requests to other sources do not really need to go through the server and they create a bottleneck.

**User demand:** Increased user demand for the Linking Lives interface could cause a performance hit on the server.
follows the links and contacts all the relevant external websites. Once it has completed its collection, the server sends the complete webpage back to the client.

2 Let the client web browser do the data collection. In this scenario, the server sends back an empty interface webpage, and tells the client which online sources hold all the data and how to lay it out on the page. The client then makes its own requests to multiple external sources and updates parts of the interface webpage, using AJAX when the requests come back in.

There are pros and cons to each strategy. In the first scenario, the server becomes a middleman in the process of loading all the data. While it is necessary to query the server for the initial interface and the linked data it holds, all the requests to other sources do not really need to go through the server and they create a bottleneck. One notable effect of such a bottleneck would be that if an external data source was down, or performing slowly, the loading of the entire interface would be delayed while the server waited for the response. It is also conceivable that increased user demand for the Linking Lives interface could cause a performance hit on the server. Another consequence of using the server in this fashion is that it must decide on all the queries it is going to make of remote sources and run them before the user sees anything.

In the second scenario, the client makes all the requests individually. If any one source is down, then just that part of the data will be delayed and the rest of the data can carry on loading into the interface. Additional queries can be made on the fly; information from one source can be used to generate a query on another source, or even to amend and update a query that has already run. All of this can be going on while the user has something to look at on the screen.

The problems with the second scenario come in the form of increased complexity in the interface logic and the cross-domain problem. Cross-domain scripting is a potential security risk as it allows code from other sources to run as part of the original website. Sites may not have the capability to accept requests like this, or they may block webpages that try to load further content from other webpages. This problem could potentially make this solution untenable, which may be a significant problem for an open linked data approach.

There are a number of workarounds to the cross-domain problem. The W3C have a recommended solution involving remote websites supplying an extra piece of header information that confirms that data from their page can be loaded into other pages as long as it is properly requested. As long as this feature—known as Cross-Origin Resource Sharing (CORS)—is enabled on remote servers, the second scenario is possible.

We took the decision to implement this second option, despite the extra work involved, as it provided for a more flexible and effective solution and makes the design more open ended.

Problems of Identity
One of the biggest challenges around our linked data work has been identifying individuals—a particular focus for us because Linking Lives is based upon people. The URIs used to identify persons in the Linked Archives Hub dataset have their origins in the names of persons occurring in the Archives Hub EAD XML documents.
Within those documents, person names occur in two contexts:

1. **Personal names as index terms**
   
   The first context is that of personal names added to the description by the cataloger as index terms, on the basis that they may be useful for the purposes of retrieval/search/browse.
   
   An index term for one individual may occur several times within the Archives Hub data. For example, *Webb, Martha Beatrice, 1858-1943, social reformer*, occurs in three different EAD XML documents. This name is taken from the National Register of Archives held in the UK (the NRA), so this is cited as the source of the index term. For this term the URI is:

   http://data.archiveshub.ac.uk/id/person/nra/webbmarthabeatrice1858-1943socialreformer

   Use of the rules may lead to different descriptors being used for the person, so we have the URIs:

   http://data.archiveshub.ac.uk/id/person/ncarules/webbmarthabeatrice1858-1943
   neepottersocialreformerandhistorian

   http://data.archiveshub.ac.uk/id/person/ncarules/webbmarthabeatrice1858-1943socialreformer

   These use the UK National Council on Archives (NCA) Rules. As different forms of the name can legitimately be used to refer to the same person, our current transformation process means that we end up with multiple URIs for one individual.

   In addition to this, use of the name within the URI does not avoid any issues of ambiguity. It is very unlikely with a name like *Martha Beatrice Webb*, but it is very possible with many names within archive descriptions, as they do not always include life dates and so you may have something like *Mary Jones, b 1901* and *M Jones, 1901-1980* in two different archive descriptions, both adhering to the same rules for name construction and referring to the same person. You may also have *John Smith, b 1945, engineer* in two different descriptions, which would create the same URI, but it may not be the same person.

   A further problem is that names may change when death dates are added. This means the subsequent re-transformation of the data will generate a different URI from that generated by the previous process using the initial form of the name.

2. **Personal names as creators**

   Personal names are also found within an EAD entry for the name of creator (or originator) of the archive—the agent(s) responsible for the creation or bringing together of the resources described. In the Hub EAD data, the names are not marked up to distinguish the name of a person from that of an organization. Furthermore, this entry is a free text entry, and usually the commonly used form of the name is given, so it is not easy to map this to the index entry.

   An example of a URI generated for this data is:

   http://data.archiveshub.ac.uk/id/agent/gb97/webbmarthabeatrice1858-1943wifeof1stbaron
   passfieldsocialreformerandhistorian
In addition to this, use of the name within the URI does not avoid any issues of ambiguity. So you may have something like Mary Jones, b 1901 and M Jones, 1901-1980 in two different archive descriptions, both adhering to the same rules for name construction and referring to the same person. You may also have John Smith, b 1945, engineer in two different descriptions, which would create the same URI, but it may not be the same person.

We include the repository reference (gb97), so that the name is effectively the person as represented within that repository.

We used a number of processes to identify candidate matches within the Hub dataset between “agents” (generated from the creator/origination context) and “persons” (generated from the index terms context). A degree of manual checking was then used to assess the accuracy of these candidate matches before creating “sameAs” relationships to indicate that the URIs refer to the same person.

One of the problems with this approach is that an application consuming the data still has to be prepared to work with these multiple URI aliases and, particularly with SPARQL, this can be quite cumbersome. To find all the data we hold about the person denoted with URI \( X \), an application has to search for patterns involving not just that known URI \( X \), but also any URI \( Y \), where URI \( Y \) is “sameAs” URI \( X \).

One approach to the repeatability problem would be to see the transformation stage as only the first part of a larger process, to keep track of the URIs generated over time, and build in a stage of processing to reconcile the URI generated this week from Scott, James, 1950-2012, Sir, biologist from the URI generated from Scott, James, 1950-, scientist in the previous version of the document six months ago. This perhaps then becomes simply a special case of dealing with multiple URIs for a single entity.

To avoid multiple URIs for one individual, it may well be that rather than publishing a set of “sameAs” triples, we should take a step further and consider consolidating our data to use a single URI for the person. But which version do we distill our multiple URIs down to? Or do we create a new URI for the individual? Should we instead think about creating a mapping to some sort of code and use that to construct a distinct URI? Or maybe more radically, but potentially more practical, would be to use existing external URIs in our data, such as the URIs from the VIAF name authority. However, it is unlikely that any resource would be able to provide URIs for all of the names in the Archives Hub dataset. In addition, there would be issues with control and persistence, as well as “dereferencing” the URI to provide information about the entity. But there would certainly be advantages to the principle of using the same URIs for the same entity across different datasets.

The issues surrounding identification of persons are many and complex. Our Linking Lives project has helped us to understand the practical implications of using our linked data, but we are not yet in a position to say that we have found a sustainable and reliable way to identify individuals. This is not ideal when you are trying to make something work in a practical, cost-effective way.

CONTINUED »
Conclusions

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 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 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. The inconsistencies within data created by so many institutions over such a long period are compounded by the complex nature of hierarchical EAD finding aids. This work requires a level of expertise in archival description as well as specialist skills in 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 are 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.

My feeling is that it needs to be easier to locate and probe data sources to ascertain the classes of things being described and the properties used to describe them—and it needs to be easier to link to these external sources. The CKAN Data Hub is one attempt to bring data together, but it is not comprehensive and not entirely easy to navigate. However, it must be recognized that working with open data in this way is not going to be easy, and a degree of investigation may be necessary to establish exactly what is being provided and how uniform it is. Simply connecting and cross-searching just two datasets using more traditional means can often prove to be challenging; with Linked Data the idea is to be able to access and connect numerous open datasets in RDF.
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 also seems to be significant and increasing interest from the LOD-LAM community (Linked Open Data for Libraries, Archives, and Museums). However, there are indications that linked data is still evolving too slowly to attract the level of investment necessary to make it a viable business enterprise and attract significant investment. (See the blog post by Tim Hodson of Talis.) Does the altruistic goal of opening up data to advance knowledge and benefit research provide a strong enough impetus to drive the linked data ideal?

Jane Stevenson (jane.stevenson@manchester.ac.uk) is Archivist and Archives Hub Manager at Mimas, based at The University of Manchester in the UK.

This article was written with help from Adrian Stevenson and Lee Baylis (Mimas) and Pete Johnston (University of Cambridge).

Simply connecting and cross-searching just two datasets using more traditional means can often prove to be challenging; with Linked Data the idea is to be able to access and connect numerous open datasets in RDF.

Archives Hub
archiveshub.ac.uk/

ARCHON Directory
www.nationalarchives.gov.uk/archon/

BBC Programmes
www.bbc.co.uk/programmes

British National Biography (BNB) Linked Open Data
thedatabub.org/dataset/bluk-bnb

CKAN Data Hub
thedatabub.org

Cross-Origin Resource Sharing (CORS)
www.w3.org/TR/cors/

Datalinks Wiki (Archives Hub Linked Open Data)
datalinks.wikia.com/wiki/Archives_Hub_Linked_Data

DBPedia
dbpedia.org/about

Freebase
www.freebase.com/

JISC
www.jisc.ac.uk

Library of Congress Bibliographic Framework Project

Linked Open Data for Libraries, Archives, and Museums
lodlam.net

LOCAH Linked Archives Hub
data.archiveshub.ac.uk/

OpenLibrary
openlibrary.org

SPARQL Query Language for RDF
www.w3.org/TR/rdf-sparql-query/

Too early, too slowly, timhodson.com [blog]. July 5, 2012
timhodson.com/2012/07/too-early-too-slowly/

Virtual International Authority File (VIAF)
viaf.org/