Third Conference on European Historical Bibliographies
3-4 December 2009, The Hague, Institute of Netherlands History

Peter Salt (Bibliography of British and Irish History)
spsalt@rhsbib.u-net.com

Connecting the Royal Historical Society Bibliography to other online resources

This text should be read in conjunction with the PowerPoint presentation used at the conference. Numbered headings refer to the slides in the PowerPoint presentation.

1. Opening slide

2. Introduction

Firstly, some background. This section of the conference is addressing the question of whether the national historical bibliography is an endangered species, but you could say that the Royal Historical Society Bibliography is not even an endangered species, but an extinct one, as it will cease to be published at the end of 2009. However, it will be evolving into something new: the Bibliography will be re-launched in January 2010 as the Bibliography of British and Irish History, or BBIH. BBIH will be published by Brepols as a subscription service. However, although Brepols will publish the database, the Royal Historical Society (RHS) and Institute of Historical Research (IHR) remain closely involved with the running of the Bibliography and will continue to contribute to the cost of compiling the data. The editorial staff will remain the same; in particular, Ian Archer, who represented the project at the first and second conferences, is still the Academic Editor – and, incidentally, he is sorry not to be able to attend the conference this year.

3. Contract signing

So BBIH will be a three-way partnership between the IHR, the RHS and Brepols. Slide 3 symbolizes this, by showing the signing of the contract for the new service. Professor Miles Taylor, the Director of the Institute of Historical Research, is on the left; Professor Colin Jones, the President of the Royal Historical Society is in the centre; Chris VandenBorre from Brepols is on the right.

In the background, you can see a prototype of the new BBIH interface. Where possible, the new interface will be used to illustrate this paper, but most of the techniques that I will describe have already been used in the RHS Bibliography. Indeed, I will refer to one or two things that we have done for the RHS Bibliography that won't be part of BBIH.

4. Summary

My theme is the links between the Bibliography and other online resources.
- I'm going to start with some simple links to online reference services.
- I will then look at links to library catalogues and to Google Books that use a simplified form of OpenURL.
- And then I will take that a step further by looking at the use of OpenURL link resolvers.
- Finally, I'll look at exposing the content of the Bibliography on the web.

Can any of these technologies help to address one of the threats to historical bibliographies that we are addressing today – the increase in online publication and the resulting belief that all knowledge can be found through a Google search?

I apologise if I'm covering ground that's familiar to some of you, or even to all of you. However, if the ideas and techniques are familiar, I hope that we can compare our experiences in discussions later.

5. Linking to reference works

Let's begin with the technically simple links to online reference services. Chris VandenBorre has already given a quick demonstration of the links to the Oxford Dictionary of National Biography. I'm therefore going to look at the links to the British National Register of Archives (NRA). These links work in the same way as the Dictionary of National Biography links. The NRA is a service maintained by the British National Archives which provides information about the whereabouts of people's manuscripts, whether these are in the National Archives, in other public collections, or in private collections that the National Archives have listed.

So, let's suppose that we've done a search for Emmeline Pankhurst, the campaigner for votes for women.
Slide 5 shows the detailed results page for a record that deals with her.
You can see the Dictionary of National Biography link, and also one to Who was Who, another Oxford University Press product. To the right, is the link to the NRA that I want to concentrate on.
We can see at the bottom of the screen the URL to which it will link. This uses Emmeline Pankhurst's reference number in the NRA database (22058), which you can see at the end of the link – we have this information stored in our authority list of personal names and we copy it into each record about her when we export the database for publication.

6. NRA page for Emmeline Pankhurst

So if we follow this link we land on the NRA page for Emmeline Pankhurst. It's a deep link – it doesn't simply take us to the site, it takes us straight to the relevant page. This is a useful service for our users, of course; while they are searching our Bibliography for secondary works about a person, they can quickly look to see information about that person's manuscripts.

However, there's a more selfish advantage for the Bibliography:

7. NRA link back to Bibliography

As you can see, the National Register of Archives offers a link to the Bibliography. This isn't simply a link to our home page but is once again a deep link that takes the user straight to the results for Emmeline Pankhurst.

8. Bibliography results for Emmeline Pankhurst

This brings some traffic to the Bibliography. And the presence of links on other sites also raises the Bibliography's standing with search engines such as Google.

So, although our first concern in providing the linkage is to help the user, we also bring traffic to the Bibliography.

9. Summary of links to reference services and online reviews

Slide 9 summarizes the reference services to which we offer deep links.

And it also shows the various online review sites to which we offer similar links. These don't all offer reciprocal links, although the Institute of Historical Research's online reviews site does.

The problem with these links is that they depend on us holding the relevant information about the other sites. We need to know the id numbers of personal names

in the NRA, for example, and to store this information in our database. Quite a lot of work is involved in keeping these links up to date, and ensuring that they are reciprocated.

10. COPAC link from BBIH

There is an easier way, which is to build links that use the bibliographic information that is contained in our data in any case. This is, for example, a good way of linking to library catalogues, since these can be searched using the kind of bibliographical information that our records are bound to contain.

If you look at slide 10, showing a detailed record in BBIH, you can see that, among the links at the top of the record, is one to COPAC. COPAC is the combined catalogue of a consortium of major libraries in the UK and Ireland, including the national libraries for England, Scotland, Wales and Ireland, and most of the big university libraries. I know that other projects represented at the conference make similar links to, for example, the Karlsruhe Virtueller Katalog, but I'll explain briefly how it works.

You can see at the bottom of the screen the URL of this link.

11. COPAC link in detail

Slide 11 enables us to see the link more closely, in simplified form. I am showing only the essentials – and you can see that the essentials are very simple.

There's a base URL that addresses COPAC:

http://copac.ac.uk/wzgw

This is followed by a question mark and then the query string made up of value pairs that tell COPAC what to look for. Here we have just a single value pair, telling COPAC to look in the ISBN field ("isn="), and telling COPAC the ISBN to look for ("1855143488").

In essence, this is a very simple OpenURL, although it doesn't conform to the OpenURL standards.

12. Results on COPAC

Slide 12 shows the results of our link on COPAC. They also illustrate a good reason for using ISBN for making these links, as, of the two records on COPAC, one has the title slightly wrong.

Nonetheless, we do use titles for linking, as we don't always have ISBNs in our data, and of course many older books do not have an ISBN. Linking by title was particularly useful when we were looking for a way to link to the texts of pre-1900 publications that Google has been placing online in partnership with several major libraries. We were particularly keen to do this because of the continuing importance to modern scholars of nineteenth-century editions of manuscripts which, in many cases, have not been replaced.

13. Google Books links

Slide 13 is an example from the Royal Historical Society's own record series, published in 1848.

You can perhaps see from the link at the bottom that we don't use the full title. We put the title through a routine that strips out insignificant words, such as "the", "of" or "and". We also limit the query to the first six words. We found that, if we used the whole title, we were sometimes missing useful results because Google itself wasn't

using the complete title. On the other hand, of course, make the title too short and it matches records that aren't relevant. There's no hard and fast rule to this, obviously; it's just a matter of trying things out to see what works best. Indeed, we worked out the format for the links by experimenting with searches from the Google Books interface, and seeing what the URLs of the results pages looked like.

14.   Google Books results page

Google usually produces multiple hits so the first thing that users see on following the link is a Google results page, showing the various versions of the book available. You can quickly see, however, which ones offer "Full view", which means that the full text is available.

15.   Online text for Machyn diary

So one more click brings us to online full text.

As I said, we devised this for pre-1900 publications, but we decided to provide the Google Books links for all the books in the bibliography, using ISBN for the more modern works. Although the latter aren't available as full text from Google Books, because they are still in copyright, there's information such as publishers' summaries, tables of contents, links to reviews, and, of course, links to bookshops where users can buy the work!

Nonetheless, if we want to get at the *full* text of modern publications, and especially of journal articles, the material freely available on Google Books isn't going to be enough and we need to link to subscription sites.

16.   DOI link

One of the most reliable ways of doing this is to use the DOI, or Digital Object Identifier. These are produced by a not-for-profit organization called CrossRef which issues publishers with unique numbers and redirects queries that use those numbers to the publishers' own sites. So, instead of giving a link to the publisher's site as such, we give a link to CrossRef, and add the DOI number for the particular item, in this case an article from the *Historical Journal,* published by Cambridge University Press.

You can see the URL at the bottom: the initial http://dx.doi.org gets us to CrossRef; the next element, 10.1017 identifies Cambridge University Press. The final bit is entirely at the publisher's discretion and can take many forms, and can often be a meaningless number, which makes it difficult to check – this particular example falls into that category, although you can see that the ISSN of the journal forms part of it.

17.   Cambridge Journals Online page from CrossRef link

Slide 17 shows the results from clicking on the link. CrossRef resolves the DOI number into the address of the Cambridge University Press journals site and takes us straight to the abstract for the correct article, with links to get to the full text in a choice of formats – assuming, of course, that we have a subscription.

The advantage of this for us is that we don't have to worry about publishers redesigning their sites and changing all their URLs, or, indeed, publishers being taken over and acquiring completely new web addresses as a result. The publishers tell CrossRef about the changes, so that the DOI links remain valid, and we don't have to change anything.

And, as publishers digitize more and more of their older journal articles, the amount of material to which we can link in this way increases.

On the other hand, there are snags with this arrangement. Collecting the DOIs is not too burdensome for new records, since we usually consult the publisher's abstract in order to get subject information, but it *is* time-consuming to go back over older records to see if DOI numbers are available.

Also, DOI numbers link to the publishers' sites. But they don't tell users about other online versions that may be available, for example from JSTOR, the American online journal archive, or from one of the journal aggregators, such as Ingenta or EBSCO. This doesn't matter if the user has a subscription to the publisher's version. But what if he or she doesn't have a subscription to the publisher's site, but does have access to the same journal through JSTOR, Ingenta or one of the other services? We need to get the user not simply to *a* copy online, but to the online copy to which he or she has a subscription. This is known as the "appropriate copy problem".

The answer lies in exploiting the OpenURL link resolvers that most major libraries now possess. These take bibliographical information and compare it with the online resources to which the library subscribes – they then return results which are limited to what's available at that particular library. They can include the results of local catalogue searches, as well as looking for online full text. The next slides show a couple of quick examples.

18. Find a copy

Slide 18 shows what a user sees if he or she is working in the central departments of London University. The orange "find a copy" button is one that such a user would be used to seeing on other resources.

19. ULRLS link resolver

And clicking on the orange button connects us to London University's link resolver, which gives the user information about online resources that reflects the subscriptions available in London – in this case they can see if the article is available from EBSCO or from JSTOR. And clicking on the EBSCO link, for example …

20. Article from EBSCO

… takes us to the abstract of the article provided by EBSCO, with a link to the full text.

21. Find it @ Oxford

If we were in Oxford University, on the other hand, we would see a different button, which again would be familiar to us from other resources in Oxford.

22. Oxford link resolver

And clicking on the "Find it @ Oxford" button of course connects us to *Oxford* University's link resolver, and we get information that reflects what's available in Oxford – EBSCO again appears but, instead of the JSTOR option that appeared on the London page, we have a link to the publisher's site. And clicking on this link …

23. Article from Wiley

… again takes us to the abstract of the article on the publisher's site, with a link to full text.

Behind this of course lies something very similar to the technology used to link to the National Register of Archives and to Google Books. That is, we use a base URL followed by a query string made up of value pairs that describe the article for which we are looking.

There are two complications.

24.   OpenURL standard

Firstly, the second part of the URL, the query string, has to be understood not by a single service, such as Google Books, but by any link resolver to which it may be sent. Fortunately, there are OpenURL standards to which the resolvers comply, and which define how we should do this. Although OpenURL itself was invented in Ghent, the standards are now maintained by OCLC – the Online Computer Library Center in Dublin, Ohio. The standard looks pretty complex, as the example in slide 24 shows, but the advantage is that there is a way of expressing most of the bibliographical information that you are likely to have, whether it is for a book, a journal article or for part of a book (see http://alcme.oclc.org/openurl/servlet/OAIHandler?verb=ListRecords&metadataPrefix =oai_dc&set=Core:Metadata+Formats). And link resolvers will generally work with as much or as little of this information as you provide.

25.   Link to COPAC and OpenURL link compared

So the simple link to COPAC that we looked at earlier can become something much more elaborate. Where the COPAC link conveyed just the ISBN, an OpenURL link can convey much more, such as the source of the query, the type of bibliographical item (called "genre"), the author surname and so on.

The principle is the same, though – a base URL followed by a question mark and then the query string which consists of one or more value pairs.

So the OpenURL standard enables us to compile the query string, the second part of the OpenURL. But how do we determine the appropriate base URL for the user? We need to know where he or she is in order to use the correct base URL for the resolver belonging to his or her institution.

26.   Identifying the user's link resolver

There are several possible answers to this. The one that BBIH adopts is to get subscribers to send information about their link resolvers to Brepols as part of the subscription process, along with the range of IP addresses that the subscriber uses. Then, when Brepols detect a subscriber coming from an IP address in that range, they can also provide links that go to the appropriate resolver.

That's a reasonable thing for a publisher to do, because the information can be shared across the range of services that they offer, and because they anyway need to know subscribers' IP addresses - this is usually the basis for giving them access to the Bibliography itself.

It's different for a single project, often providing a free service – as the RHS Bibliography used to do and as many of bibliographies represented at the conference still do.

A solution that's available in the UK is an OpenURL resolver registry maintained on behalf of the higher education sector at the University of Edinburgh (http://openurl.ac.uk/doc/index.html). This is designed to make information about link resolvers available to projects such as the Bibliography. The RHS Bibliography worked with it – when a user first entered the Bibliography, some VBScript on the opening page was used to collect information about his or her IP address and to send it to the registry; the registry then returned the resolver details. We could then incorporate these into any OpenURL links that we displayed for that user.

Whether there are similar registries in other European countries, I don't know – it would be interesting to hear if anyone is aware of any. An international one does

exist, although it is heavily biased towards the United States and Canada (http://www.oclc.org/productworks/urlresolver.htm). It is maintained by OCLC, and it is possible to send OpenURL queries to it. It will check the user's IP address and direct the query to an appropriate link resolver if it is aware of one. If it isn't, the query goes to OCLC's WorldCat which returns a library catalogue entry but then offers the user the chance to enter information about their location, and get information about holdings in any OCLC member libraries in their area. We provided these links for all users of the RHS Bibliography – in the new BBIH, they will be available if subscribers request them. One good thing about the OCLC registry is that links are easy to set up – one link does it all and there's no need to query the registry as with the UK example.

A final approach involves Context Objects in Spans (http://www.openly.com/openurlref/#coins). The OpenURL link is embedded in special code that can be identified by an OpenURL referrer extension to the Firefox web browser. It's then up to the user to configure the OpenURL referrer with information about their library's resolver. However, the OpenURL referrer hasn't been updated to make it compatible with the latest versions of Firefox. So I'm not sure if this method has a future – it would be interesting to know if anybody knows any more about it.

27.   Google results page

Before I sum up I will just touch quickly on one other aspect of web connectedness – making the content of the bibliography accessible to search engines, and more specifically to Google. Is this a way of capturing the attention of web users who don't know about specialized Bibliographies, and think that all knowledge is accessible through a Google search? Of course, Google should index your home page, but the aim is to get Google to index individual records. We achieved this by creating a Google Sitemap – an XML file that lists the URLs of all of our detailed records. We then set up a Google account through which we could tell them where to find the sitemap on our server.

Slide 27 shows an example of the results of a Google search for 'medieval Alcester' where an RHS Bibliography record comes first on the Google results page. 'Alcester', by the way, is a market town in the Midlands of England – my reason for choosing this rather obscure example is that, for a better-known topic, Google would return so many results that the RHS Bibliography results would be swamped amongst results from other sites!

Also, Google hasn't indexed everything, by any means – at the moment 64,000 of the 470,000 records submitted have been indexed.

So, the sitemaps haven't been quite as effective as we had hoped, especially as we also tried to boost our pages' standing with Google by giving them meaningful titles – you can see from the example here that both the page title and its descriptive metatag repeat information from the article title.

28.   Statistics for referrals

Nonetheless, at one level, it does work. Slide 28 shows that more traffic reaches us from google.co.uk than from any other site; indeed, more visits result from referrals from google.co.uk than come from the "no referrer" category. The effect would be even more clear if we put the British, American, Canadian and other Googles all together into one. And we have come across one user (a Historic Environment Record Officer for a local authority) who told us that she had found the Bibliography and realized that it would be a useful resource because she did a Google search for archaeology in her area and saw a Bibliography page in her Google results.

On the other hand, because our users don't log on, we can't tell how many of the people who find the Bibliography through Google searches do take a serious interest. Many may conclude that it's not what they were looking for and quickly go off to other sites – it has to be said that the readiness of social historians to write about gender and sexuality means that some of our records sound as if they might have pornographic content, and I suspect that some of the users who hit upon them are disappointed to find bibliographic records, especially as there are no illustrations.

Also, this type of exposure to Google may be suited only to a free service. So there aren't currently any plans to continue it with the BBIH subscription service.

We have also experimented with the Open Archives Initiative (OAI). However, not long after we announced our compatibility somebody threatened to use it to download the complete bibliography, so we switched our OAI server off again. It would be interesting to hear about others' experience with it.

29. Summary

Rather than dwell on this, however, I want to return in conclusion to the linking using OpenURL and similar technologies. Do users like these features? We asked users to complete a survey in 2008 and we found less enthusiasm for the OpenURL links than we had hoped – over a quarter of respondents had used links to COPAC, and around a fifth had used the OCLC resolver registry links, but only one eighth had used OpenURL links to library link resolvers. On the other hand, the proportion using OpenURL links to library link resolvers has grown considerably since our previous survey in 2005 – the proportion using OpenURL had doubled by 2008, compared with what it had been in 2005, so perhaps our users are discovering what the technology can do for them.

Certainly, the librarians to whom we have spoken do like the OpenURL linkage. We have been told that it gives BBIH an edge in the climate that this conference is designed to address – that is, in a world where bibliographies are in danger of being ignored because it is so easy to Google for online text. On the one hand, BBIH does what a traditional bibliography aims to do and that a Google search does not – BBIH covers the field systematically, including local journals and the products of small publishers that have little or no presence on the web; BBIH provides systematic indexing using a controlled language that should make for more consistent results. And to that we can add the possibility of searching by period covered, which isn't possible in a Google search. Yet, on the other hand, where text is available online, we are providing our users with the links and tools that will enable them to find it.