ACA-AAQ Symposium, 26 May 2004
Reflections on the future of the archival community
Tim Hutchinson

Introduction

First of all, my thanks to Jerry O'Brien and the ACA for their invitation to speak during this session. While I'm billed as the speaker for the ACA, I obviously bring my own interests and background to this as well.

To provide some focus for a broad topic like the future of the archival community, I'd like to set my remarks in the general framework of two overlapping themes: making archives accessible to researchers and a wider public; and our increasingly online environment. To some extent this has started already, but I believe that our exposure to new user communities, and the public's general awareness of the archival community, will depend on our online presence. More importantly, fundamental changes in how web content is searched and delivered will affect a variety of archival services. At the same time, though, it is the archival community's more traditional principles and practices that will allow us to continue to have an important role to play.

First, a few thoughts about reference service.

In Dan Brown's novel *Angels and Demons* (the prequel to the popular *Da Vinci Code*), our hero finds himself in the Vatican Archives in the dead of night – which, of course, is when most archives are visited in books and movies. Finding the archives' database password-protected, our hero does not despair. He looks around the stacks, and "immediately discerns" the organization of the archives, and finds what he's looking for.

We should be so lucky.

Perhaps this scene would have been more realistic had the hero done a Google search, not found anything, and concluded that the document he was looking for did not exist.

Reference services – the Google generation

We are all familiar with the phenomenon of web search engines being perceived as all that's needed to find anything on the web. Put another way, if you can't find something via Google, is it worth knowing about? Of course, archivists and librarians are all too aware that doing research this way is incredibly restrictive. A question for the archival community, then, is whether we try to serve the Google generation on its own terms, and try to make all our web resources exposed to web search engines. Or do we try to draw them in in other ways? There are developing technologies – so far largely untested – that would enable archival resources now part of the "hidden web" to be exposed to traditional search engines. This would not solve the problem entirely, but I'm referring to the Open Archives Initiative. For those not familiar with this metadata harvesting protocol, I should point out that the "Archives" in the title refers to archives more in the computer sense of the word, and in fact very few archives are currently involved in this initiative. Of course, content other than databases are already available on archives' web sites, and exposed to search engines – such as virtual exhibits – and this provides another way in.

In some ways, the predominance of search engines like Google will preempt any debates over the convergence of cultural heritage databases – for examples, there have been debates about whether one should be able to search library and archival databases concurrently. More and more, I think we're recognizing that users, especially users who are not seasoned archival researchers, are more interested in finding the information they need. Technologies like the Open Archives Initiative give us the opportunity to make archival information available in a variety of ways. If anything, I would expect the situation we're in now to become even more pronounced. That is, information about archival resources must be easy to find on the web. There must be a variety of ways to find this information – even if ultimately the search leads to the original archival database.

Having said all that, the obvious desire for one-stop shopping does not necessarily have to be a negative thing. Surely an important goal is to make information about archives as accessible as possible. More generally than the Open Archives Initiative, the most promising development that would make content on the web more accessible and structured is the move towards XML. I'll return to this point later.

Changing online environment and electronic records mandate

An ongoing challenge for the archival community, it almost goes without saying, is the preservation of electronic records. I think it's fair to say that we're losing the battle so far. But now the rules are changing, because of a rapidly changing online environment. This changing online environment affects not only our approach to providing services, but also our preservation mandate.

The World Wide Web was invented in 1990, but it wasn't until 1993, with the development of Mosaic that it was accessible to a wider public. Version 2.0 (January 1994) introduced inline images and forms. Netscape then introduced tables, forms and JavaScript in 1995 – and this was really the last major upgrade to *browser* features, according Tim Berners-Lee, founder of the World Wide Web Consortium and developer of the first web browser. The revolution since then has been in how the web is used, and in delivery of resources on the server side, behind the scenes. We've moved from static web pages, to web pages delivered dynamically, and coming soon, some say, is the "semantic web" (more on that in a few minutes). Moreover, contrary to an earlier period when most institutional websites were simply electronic versions of print publications, much more content is being delivered solely on the web. In 1997, one of my class projects was to analyze the web site of our information studies graduate program, with a view to capturing its content for archival preservation. My group's conclusion at the time was that the school's web site was a record of how the school presented itself to the world, and therefore that quarterly snapshots of the site would suffice. I don't think that's a conclusion that could be drawn today, only seven years later.

The good news is that content on web sites, to a large extent, does not have the same challenges as some other electronic records. That is, a lot of the content is text based; and theoretically, at least, most technologies on the web are open source. On the other hand, it may be more difficult to capture the dynamic content simply by browsing the web site.

The short life span of websites (even pages on institutional sites) has not changed. Content also tends to change on a single document, so policy changes are not necessarily captured. This is a record creation and record keeping issue, too. Archivists' role in helping records creators to create reliable and authentic records on the web is no different than before – in fact it may be even more important. Especially if the copy on the web is declared to be the "official" copy. There's a similar issue with institutional *intra*nets.

Some archives have started to take steps to preserve institutional web sites, and as online content becomes even more central to institutions' activities we will, I'm sure, see more initiatives like this. I'd like to think that the structure and the spirit of the web will help us here – that is, there may be open source, decentralized solutions.

Positioning ourselves for the next leap forward

I've mostly been talking so far about our current situation, extrapolating forward.

But it's interesting to look back – very briefly – at where we've been. In terms of access to information about archival holdings, I entered the archival profession when RAD was effectively complete, and when online access to archival descriptions was becoming more common but when Archives Canada (as it's now called) was still being planned. Since there's no argument that the existence of RAD was a key factor in the development of Archives Canada – that is, we needed a content standard – I was curious to see if the prospect of a national database was originally articulated as a reason for the development of RAD, or whether it was just a useful by-product. Well, from the 1980 report of the Consultative Group on Canadian Archives, we have: "The lack of uniformity of descriptive and cataloguing methods seriously hinders the creation of an information system at the national level." This report proposed several projects, among them "a feasibility study on a national machine readable data bank on archival holdings". To the best of my knowledge this proposal was not pursued at the time – neither (and perhaps we should be thankful for this) was the "microfilm collection of the finding aids of every archives". Of course, with the earlier publication of the Union List of Manuscripts, the Canadian archival community was no stranger to standardization and community efforts to make our holdings available. My purpose here isn't to explore the history of Archives Canada, but rather to marvel at how far we've come, and at the vision of the archivists 25 years ago who put the wheels in motion towards a national archival database, when the available technology was relatively under-developed.

Having said that, with such rapid development in Internet applications over the last ten years, in some ways it's difficult to speculate about where we'll be in another ten years. Some upcoming developments seem, somehow, more predictable. The basis of this is increased implementation of XML, and therefore the potential for a more structured, searchable web. Some, however, see this as going much further, in the form of the so-called "semantic web."

Potential new developments – the semantic web

In a Scientific American article published in 2001, Tim Berners-Lee and his colleagues defined the semantic web as "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." They go on to say, "The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users." And from the World Wide Web consortium: "The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications."

It's this last point that I find a bit more feasible – re-use of data. But it seems that much of the work on the semantic web deals with software agents, and the idea that one such re-use could be by intelligent agents; or, to repeat the quote I just used, the agents will "carry out sophisticated tasks for users".

But let's ignore what I find feasible for a moment, and explore this idea a bit…

The scenario that's presented about semantic web agents is one where the agents undertake tasks otherwise done by a user. This would take advantage of properly structured web content so that, for example, the web agent could figure out when an archives is open to the public, or which e-mail address to use for reference enquiries.

So maybe a user will instruct the agent to look for archival material relating to a particular research topic. The agent goes out to a variety of archival web sites and databases; carries out searches; follows links to the "related material fields"; takes note of restrictions and additional finding aids not online; and submits a report to the user. Or does it get even more sophisticated than that? Does the agent in fact follow up with the archives, requesting copies of relevant material?

So if an intelligent agent came calling, could archives respond? The good news is that much of our descriptive data is already available in structured form. And much work has been done relating to ontologies – for example, mapping data elements from one standard to another, especially across disciplines; developing subject headings and other terminology indexes. Development of the semantic web is based on Extensible Markup Language (XML), and Resource Description Framework (RDF - a model, and syntax, for knowledge representation). In that light, the bigger challenge (and a huge stumbling block in efforts to create a truly "semantic web") is that archival descriptions hardly lend themselves to simple statements. For example, what is an intelligent agent to make of an administrative history? Will it have to be taught what a "fonds" is? On the other hand, more extensively structured statements about relationships between archival material, and archival material and its creators, would be much more easily parseable, and this is supported by the ISAAR authority data standard and the Encoded Archival Context model. A simpler example is the communication of information like an archives' public hours, or conditions on access.

I remain quite sceptical about the prospect of "intelligent agents" anytime soon. After all, this has been the stuff of science fiction for some time. On the other hand, the Internet and the Web seem revolutionary, too. When you have people like Tim Berners-Lee taking it seriously, and organizations like the World Wide Web Consortium actively working on it, this is something to watch. An important consideration is to be aware of what kind of information can be used by a so-called intelligent agent, and to keep our approach to providing information flexible enough to adapt to changing standards. So that when the intelligent agents do come calling, we're in a position to respond. Putting this in a less futuristic framework, the current focus of work on the semantic web relates largely to knowledge representation and how to make web content meaningful. This can only help improve our access to web resources – putting aside any ideas about intelligent agents.

A simpler example may help. If someone searches for "archives", it is currently almost impossible to distinguish between an archival institution, a set of records, non-current pages of a web site, previous postings of a listserv, or a journal title. More generally, a lot of searches for personal names yield results for genealogy pages. What if you wanted to search only the genealogy pages? Or search only the genealogy pages for marriages?

We've already begun to see some more structured web-based search possibilities. Even within Google, without requiring XML, we have categories such as news, usenet and image searches. (Though I wouldn't hold your breath waiting for the archives category!) There is an XML format for news and weblog feeds – to allow other web sites to re-use this content (which one of the goals of the semantic web).

Let me come back, then, to the comments I started with. The behaviour of the so-called Google generation is not likely to change. That is, more and more, people will expect to easily find information on the web. But the developing trend seems to be that the web will change to accommodate this and make this an effective strategy – even if this just means that information on the web will be presented in more structured, more effectively searchable ways. I'm confident that the archival community will continue to be able to take advantage of these changes.

Conclusion

So where does that leave archivists, and the archival community?

I do think that there will continue to be fundamental changes in how we communicate and make archives accessible to the broader public, largely focussed on web-based technologies. I've talked about the potential of the semantic web, but I would look at that as a representation of a much more concrete development – more structured web content.

In many institutional settings, this will also force the issue of electronic records. In the past, it has been easy to ignore this issue, but with more records being delivered on the web – and more and more tools that make it easy to create web content – it's obviously something that we will be ignoring at our peril. I'm referring both to our preservation and records management mandates.

At the same time, I don't want to leave the impression that the future of the archival community is based entirely on the Internet, and technology in general. In fact I would argue that it's our more traditional practices and principles – especially in providing contextual information and ensuring reliable and authentic records – that make our role even more important in an increasingly online environment. Indeed, a lot of web content would be most suitably handled in an archival way: "Context is everything", and all that. Will web standards and practices evolve to reflect this? As well, in an increasingly online world, the unique nature of archival material will put archives in an enviable position – not only as content providers for the web, but also as irreplaceable research centres in their own right.