The Open Resource Scholarly Network: new collaborative partnerships between academics, libraries, archives and museums

Gavan McCarthy
Director
Australian Science and Technology
Heritage Centre
University of Melbourne
gavan@austehc.unimelb.edu.au

Joanne Evans
Deputy-Director
Australian Science and Technology
Heritage Centre
University of Melbourne
joanne@austehc.unimelb.edu.au

Abstract:
The Australian Science and Technology Heritage Centre (Austehc) has been collecting and disseminating information about the history of Australian science, technology and medicine, including data about archival resources, on the assumption that scholarly practice and the creation of new knowledge was based on free access to, and the citability of, existing knowledge. Despite the advent of enabling electronic network technologies, it appears that this assumption is not universally accepted. In this paper, we explore by way of real examples the benefits that come from the open sharing of information and knowledge, not just for researchers but for cataloguers, archivists, web publishers and other informatics professionals.

Introduction

Since 1985 the staff of the Australian Science and Technology Heritage Centre (Austehc) and its predecessor the Australian Science Archives Project, have been collecting and disseminating information about the history of Australian science, technology and medicine. This has involved the collection of data about people, places, organisations, archival resources and publications. Our web publication of this information has been grounded on the assumption that scholarly practice, including the creation of new knowledge, was based on free access to, and the citability of, existing knowledge.

The advent of electronic network technologies has enabled us to realise our goals in ways that were not even dreams in the earlier environment.

However, not all players in the scholarly information and publishing realms have responded in the same way. Despite these new technologies, which should be making resources much more readily accessible, many valuable resources are locked up (discoverable, perhaps, but not citable) behind closed database walls or are available only on a user-pays basis. In many cases, these resources were previously available freely through research libraries.

Austehc has spent the last few years developing database driven web-publishing tools to support an open resource scholarly electronic network. These tools are being offered to the community at no cost, under the open source philosophy, if they are used for public good and education purposes.

In this paper we explore the concept of the open resource scholarly network, review the functional requirements that would underpin such a network, and illustrate by way of real examples the benefits that come from the open sharing of data, information and knowledge, not just for researchers but for cataloguers, archivists, web publishers and other informatics professionals.

The Open Resource Scholarly Network

The Open Resource Scholarly Network is a phrase that was coined by Austehc in 2001 to embody a philosophy that we had been endeavouring to support since our inception in 1985 (McCarthy & Evans, 2001). When we first utilised the Internet (1992) and then the web (1994) we had assumed, like many others, that the ideals proposed by Tim Berners-Lee and the other founders of the web were widely, if not universally accepted. Although we were well aware of a tendency in academic circles to control access to knowledge in order to obtain or maintain positions of power and authority, we naively believed that we could use the web to circumvent this aberration and create free and equitable access to information.

As the 1990s progressed we became increasingly aware that the adage “knowledge is power” had been replaced by its economic rationalist equivalent “knowledge is money”. Many web based technologies designed to facilitate improved control over the source material at the manager’s or publisher’s end paid little attention to the needs of the user to both locate and then relocate those resources or tell other people about their existence. Furthermore, as the commercial potential of the web to deliver services to fee paying or purchasing customers was realised this tendency was extended to include the control and restriction of the user to meet the profit requirements of the service provider. Unfortunately, many software vendors and the emerging web developers were guided by the cyclic interaction of the commercial and technological imperatives to create products for web publications that were not at all suited to academic or scholarly purposes. This tendency included systems that were accessing information from databases and more recently XML sources.

As was evidenced at the recent conference of academics, scholars, archivists and librarians who met at the University of Sydney for *Computing Arts 2001*, the influence of commercially focused or supply-side technologies is prevalent. The use of resources in a scholarly fashion, in other words, for the purposes of knowledge generation and communication does not appear to have entered into the consciousness of some producers or publishers. In some cases the scholars themselves have been victims of the dynamic state of the technology and have focused primarily on their consuming interest in textual analysis, whereas in other cases librarians and archivists wishing to make resources more widely available to the public have received poor technological advice. The conference also produced some excellent examples of the opposite, that is, beautifully constructed and navigable resources that were discoverable, electronically citable, in some cases to paragraph level, and mounted on the web in ways that will enable persistence.

The role of citation as the foundation of scholarly practice has a long, interesting and contextually complex history. It evolved rapidly in parallel to the introduction of printing and gained a particular strength in the age of enlightenment and the emergence of modern scientific practice. It established a connectedness between ideas and thoughts, it created a web of links between authors and it built bridges between geographically or culturally disparate communities. It also built a web of connectedness that was highly contextualised. Authors cited other works for a purpose. They were selective. They built a localised web of links that brought authority to their arguments and mapped them into both the synchronic and diachronic evolution of their selected discourse.

The potential of electronic networks, and in particular the concept of hypertext devised by Ted Nelson a number of decades ago, to advance scholarly practice and knowledge generation is significant. But as recent experience has shown, this is not a trivial political, social or cultural advance. The new communications technologies are indeed double-edged swords. With an increased ease in production and publication, and the consequent improvement in equity and access, comes an explosion of resources creating a massively large, contextually unstructured and inhuman information space. The development of the web as a place for scholarly purpose and indeed education at all levels is far from being realised. The potentials of open electronic scholarly networks are far reaching but they will only be realised if the key functional requirements are articulated and acted upon.

There is the possibility of building highly structured and contextualised spaces on the web, but to understand this we need to tackle three key concepts (they are dealt with in more detail later):

- scale-free complex networks - an approach that stems from the study of natural, evolving and dynamic complex systems that are composed of entities and defined relationships; (Barabasi 2001);
- the small world effect. - also known as ‘six degrees of separation’ is a useful by-product of scale-free complex networks for the purposes of building a cultural heritage information infrastructure (*How the Oracle of Bacon Works* 2001*)*; and

- artificial neuronal networks - a system-learning paradigm also based on entities and defined relationships with the potential to aid in analysis and contextualised discovery.

When combined and implemented these principles will underpin a scholarly web space that is self-sustaining, human friendly and is also deeply connected with traditional academic traditions. More importantly, it improves equitable interactive access to knowledge for broader cultural and social purposes. (McCarthy 2001)

The Functional Requirements

For those of us who want to contribute knowledge and information to the public space offered by the web, as opposed to those that are seeking to use the commercial potential of the web, what do we do?

The key principles or functional requirements of web object publication that will support the building of the Open Resource Scholarly Network are citability, coherence, communicability and endurance. Adherence to these principles will not only enable the building of connectedness at the micro level but will enable the building of larger scale structures and architectures based on the concepts of scale-free complex networks.

Citability

For a web object to be citable the user needs to know key information about the 'who, what, where and when' of the object origination. In the print world, this was the imprint data, and in the web world, it is what has become known as "resource discovery metadata".

Ideally, this information should be in both a human readable form and in a machine processable form. A standard for the machine processing form that has gained currency is the Dublin Core Metadata Initiative (2001). It seems strange that this point has to be made, and to this audience we may be preaching to the converted, but there are still so few web publishers implementing this key requirement.

The Dublin Core metadata set is used as the basis for the Australian government web metadata standards, the Australian Government Locator Service or the AGLS (National Archives of Australia 2001). An important aspect of citability is the "where". On the web, this is given by the URL or Uniform Resource Locator (World Wide Web Consortium 2001). Unfortunately, despite the enormous power of the URL and its structure, not all web publishing systems generate citable URLs. Although in theory an eight character URL that utilised only the 36 alphanumeric characters could uniquely address over 2,821 billion individual web objects, it is not uncommon to find URLs of the form:

<http://www.???.vic.gov.au/web/root/domino/cm_da/???ncor.nsf/frameset/???+Corporate?OpenDocument&[/4A25676D00279618/BCVIEW/AB5F9BE1DDCEAB074A25685D0016F9E0?OPENDOCUMENT]

What is clear, as a necessary requirement of the human interface, is that URLs be both human readable and machine processable. The issue of persistence of URLs, a core requirement for citability, has been identified and addressed by "The PURL Team" (2001) and the National Library of Australia (2001) amongst others, but awareness of its critical importance remains fragmentary within the world of contemporary web developers. As the National Library states: ‘A persistent identifier is a name for a resource which will remain the same regardless of where the resource is located. Therefore, links to the resource will continue to work even if it is moved.

Coherence

Coherence refers to the hyperlink navigation from an object that will take you to the related web objects that the user must be able to access to understand the context, jurisdiction or framework in which the object has been published. The beauty of hypertext is that you can jump in at any point. Hence all pages need navigational elements to get to the context, or some elements of context embedded in the page. You cannot assume that users know as much about your site as you do. They always have difficulty finding information, so they need support in the form of a strong sense of structure and place.

In other words, designers must have a good understanding of the structure of the information space and communicate this structure explicitly to the user. This understanding must extend beyond the technological framework to a deep understanding of the structure of the content and the wide range of users who will be attempting to gain meaning from the information.

Users need to know where they are and where they can go. They also need a good search feature since even the best navigational support will never be enough. The lack of signals and visual guides provided in a two dimensional analogue view or screenshot of a digital object creates challenges for users raised on the tradition of a highly evolved three dimensional world of print and manuscript.

A good illustration of the issues of citability and coherence can be found by comparing the *Peoplescape* entry for Ada A’Beckett which can be located through [http://www.peoplescape.com.au/home.cfm?a=stories&o=explore] with the equivalent entry in *Bright Sparcs* at [http://www.asap.unimelb.edu.au/bsparcs/biogs/P000996b.htm].

Communicability

Communicability refers not only to the ability of a web object to present its content in a human readable form but also in a standardised structured form that enables the machine processing of elements of its content. It is taken for granted that content should be expressed in a human readable form that assists the user to interpret the content. Indeed, hypertext markup language (HTML), the language of the web, has been geared specifically to this end.

This context of visualisation, the larger framework in which we perceive information objects, is an area that has been of interest to book and print designers for hundreds of years. Processes, principles and standards are now highly evolved and there are general communities of understanding which enable meaningful interpretation within defined language and cultural settings. However, the limited reach of print has tended to keep print objects within contextually bounded spaces, whereas web objects are accessible in a multitude of cultural, social, political and linguistic spaces thus creating a whole new set of challenges.

HTML, the lingua franca of the web, is focused on meeting the needs of human visual communicability but not to the machine processing of the content. XML, or eXstensible Markup Language, has the capability to provide this service. The promise of XML has been slow to be realised. The technological challenges have proven to be non-trivial and it has only been in highly constrained and relatively simple systems that XML has been an effective transmitter of semantic content. There have been some flagship examples of the effective use of XML in the scholarly environment in closed systems but these have required intense intellectual rigour both in understanding the technology and the content, long-term commitment to the development of the resources and access to large sums of money. It is difficult for many academics to put all these requirements together at the same time and for most, access to the funding for expensive technology is never likely to be available.

Endurance

How do we deal with the problem of the long-term endurance of web objects? This has been exercising the minds of many over the last few years and we will not address this issue in detail here. However, our experience has shown that the simple answer, that is an uncomplicated, explicit, flat framework for the publication of web objects of long term value based on unique identifiers, has been sustainable over a seven year period and non-problematic to maintain. We do not anticipate any immediate problems in continuing this practice into the foreseeable future. The National Library of Australia has recently implemented a strategy and system for managing persistent identifiers which seeks to deal explicitly with this issue in the large-scale digital library environment.

Larger Scale Architectures

However, our particular research interest is in how we can build large-scale open resource information architectures on the web that will support the location of system-independent web objects in contextually meaningful environments. From our perspective as archivists we are particularly interested in how information gathered by archivists in the normal activities they undertake can be used to build local, national and international information networks that are not only interoperable and systematically interconnected but empower a localised and focused web presence for individual archival organisations.

Scale-Free Complex Networks

For those with experience with and knowledge of fractal networks and structures and their visual rendering through Mandelbrot diagrams, the ideas behind scale-free complex networks will be familiar. The web has become a fruitful publication forum for Mandelbrot enthusiasts, for example Stepney (2001), who have provided an interesting array of resources that explore these ideas. These networks occur at many levels throughout nature and are one of the key aspects that sit behind self-organising systems, systems that result from processes based on the iterative interaction of objects and the relationships they form. They appear in chemical systems, biochemical process and biological process, as well as in human constructs such as large-scale electricity grids and human social, political and cultural networks. Indeed recent analysis of the web has shown it to be an evolving dynamic complex network with scale-free qualities. The conscious development of the web in this direction will greatly improve its usability and sustainability.

The Small World Effect

The small world effect is also known as of ‘six degrees of separation’. The history of the six degrees of separation can be tracked back to Stanley Milgram, a sociologist at Harvard University who surprised the world with a bold claim: any person in the world can be traced to any other by a chain of five or six acquaintances. The idea was further advanced in public knowledge through John Guare’s Broadway play and movie *Six Degrees of Separation*. This notion of the small world is based on the larger structure of human society that is drawn from the relationships formed between people. Analysis (Barabasi 2001) has shown that is indeed a scale free complex network.

It is a meaningful network for humans because it is built on relationships that people construct to make their lives work. There are limits on the number of relationships and the size of local/personal networks that people can maintain. This establishes a natural human limitation on the size of the clusters that form. What enables the small world effect to really work are the links that join those clusters to each other.

Archivists identify and define the individuals that are involved in the creation of the records they manage. By the process of appraisal and the economic limits of the archival process, this is a quite narrow select cluster of people but with links to many more people that are referenced in the records preserved. Archivists also collect information about the clusters people form in the form of corporate bodies and families. Recently, persons, corporate bodies and families have been described as entities.

In closed archival information systems, the creation of defined relationships between entities results in scale-free complex networks that can be utilised to provide a contextual framework for the interpretation of the records. By making these entities citable on the web an international scale-free complex network will evolve that could link all cultural heritage resources in a way that is meaningful to human users.

XML Enabled Interoperability and Encoded Archival Context

Recent work by the self-assembled international Encoded Archival Context Working Group has produced an XML document type definition (DTD) for the encoding of archival entities that defines the semantic elements forming the core identity of an entity. In July 2001, an alpha version of the DTD was released to the group for testing. Austehc has been a key player in this process as it has been building systems based on these concepts since 1987, reflecting the Australian archival tradition of separately identifying and documenting archival context entities.

One of the foundation aims of the DTD is that it will enable the sharing of key authority information about entities, for example: name, date and place of birth, date and place of death, functions and occupations, life and career events and a whole host of other variables that are located in time and space. These ‘factoids’ are often time-consuming and expensive to derive from records and secondary sources and at the moment much duplicate effort is being expended in their pursuit.

However, the EAC XML DTD could open the door to many other benefits including a dynamically evolving scale-free complex network which links all key related archival and heritage resources by as few as two or three links. This ultra-small world effect would be built on the identity mapping of the same entity represented in different jurisdictions and environments. It would enable the subject specific, or geographic supersets of entities (with links back to their origins) to be harvested and the creation of a variety of information infrastructure resources including national registers. It would also enable the interlinking of national registers (or indeed registers at any level) on an international level.

The Practical Requirements

This vision is not without its challenges and, within the technological limitations of the present age, is not without problems. Austehc has been building software explicitly for the lowest level user in the electronic world to assist those that wish to participate in what we see as the open resource scholarly network or which may also be viewed as the online version of cultural 'National Parks'. For us, this user, which was where we started out, usually has access to a personal computer and an Internet service provider that can serve simple html documents. They do not have deep technical knowledge or the time to learn it; they do not have that expertise on hand, but they do have a deep commitment to the resources they control and knowledge of their content.

The Web Academic Resource Publisher

The Web Academic Resource Publisher (WARP) is a database tool under development to enable the scholarly web publication of reference texts. Promoting more than just online reproduction of texts, the WARP facilitates the creation of a knowledge space, which becomes a research tool from which new connections, insights and ideas can be discovered and explored. A key aspect of the WARP is the functionality developed to use the index to enable cross-referencing within the resource and to link the resource to an appropriate external entity-based (EAC) register.

Austehc has used the WARP to publish:-

Technology in Australia 1788-1988. The online edition of this bi-centenary study by the Australian Academy of Technological Sciences and Engineering of the men, women and organisations involved in the development of technology in Australia. [http://www.austehc.unimelb.edu.au/tia/titlepage.html]

Science and the Making of Victoria. A Centenary of Federation project exploring the history and views of the Royal Society of Victoria since its inception in 1854 to the present day, and its role supporting science and technology in Victoria.
[http://www.austehc.unimelb.edu.au/smv/smv.html]

Victorian Patents and Patentees 1854 to 1904. A pilot project in collaboration with the State Library of Victoria to explore the publication of Victorian patent applications from 1854 to 1904. [http://patentsvictoria.net/]

Federation and Meteorology. A Centenary of Federation publication on the emergence of Australian meteorology as a science and the formation of the Bureau of Meteorology in 1908, paralleling the story of Australian Federation. Compiled by the Australian Science and Technology Heritage Centre and the Bureau of Meteorology [http://www.austehc.unimelb.edu.au/fam/title.html].

The WARP is currently under development with the aim of making it available under license at no cost for non-commercial, heritage and public good purposes.

The Online Heritage Resource Manager

The Online Heritage Resource Manager or OHRM is a context or entity based resource discovery and access system that links creators, archival and heritage resources and published materials within the one system. The evolution of the OHRM through its first practical implementation in 1987 and publication on the web in 1994 as *Bright Sparcs* has influenced the development of the EAC DTD and more recently has benefited enormously from the intellectual input of the EAC Working Group.

The logical structure of the OHRM, is based around:

- capturing key information about persons or corporate bodies in the Entity table;
- capturing key information about archival resources in the separate "ArcResource" table;
- capturing bibliographic citations of published resources, whether print, digital, online or offline in the "PubResource" table; and
- capturing the relationships of entities to other entities, archival and published resources in appropriate relationship tables, i.e. RelatedEntity, Earrship and Eprrship.

From this database, both static and dynamic HTML output and EAC XML output can be generated to build open citable, scalable historic registers that will form the scale free complex networks that will support the open resource scholarly network. More details about the OHRM can be found at [http://www.austehc.unimelb.edu.au/ohrm/].

The Move to Open Source Licensing

Although these tools have been developed using proprietary database software we have licensed the software so that it can be used for public good purposes at no cost. Moreover, as we have not attempted to encrypt the code other developers may learn from our work and experiences. There are two important strategic considerations that underpin this approach: the development has been funded using public money; and it is only open systems that can survive the limitations of human and corporate life. Indeed, the Historic Manuscripts Commission in the United Kingdom which has had a major influence on the development of Austehc since 1985, has recently used some of our code to enhance their systems which are two orders of magnitude larger than ours.

Austehc now has the technical capability to move beyond proprietary systems but our user community does not. We are systematically working towards the use of full open source software components but are committed to supporting users who remain restricted to proprietary technologies. We acknowledge that this is not a trivial issue especially as the future of these systems is extremely unpredictable. The principles underpinning this approach to software development would form the basis of whole new paper.

Conclusion

The Australian Science and Technology Heritage Centre has uncovered a strong demand for the types of products outlined above. The Online Heritage Resource Manager (OHRM) is a contextual framework builder and a key tool for creating scalable information infrastructure that utilises and enhances the "natural" tendency for humans to create scale-free complex networks that have strong fractal forms based on standardised semantic content. It has revealed itself to be adaptable to many different contexts and environments. The potential in the OHRM is substantial and this has been significantly enhanced by the Encoded Archival Context XML DTD. The use of the OHRM in real situations with real data and with human operators with a range of foibles is robustly probing our ideas and concepts, and testing the structures and functions. This is challenging and immensely exciting.

The Web Academic Resource Publisher (WARP) has developed quickly and moved from prototype to generalised tool with remarkable ease. We have recently completed work on the bridging mechanisms between the outputs from the OHRM and the WARP. Although XML has promised much it has been slow to deliver. We are looking to develop standardised inputs and outputs in XML, in particular the use of the Text Encoding Initiative TEI XML DTD (TEI Consortium 2001) that can be easily customised to capture individual identity and design without sacrificing functional and upgrade requirements.

Politically, we see the need for community financial support, i.e. support from government, industry and individuals, to enable the building and maintenance of the open resource scholarly network. It is a bit like a virtual national park - it is a place for all people to visit, a place in which knowledge and experience can be shared, a place where all are welcome, a place where our heritage is conserved and transmitted.

References

Barabasi, Albert-Laszio 2001, ‘The physics of the Web’ *PhysicsWorld* Volume 14 Issue 7 Article 9, July 2001 <http://physicsweb.org/article/world/14/7/9> [Cited: 22 July 2001].

Dublin Core Metadata Initiative 2001, Making it easier to find information <http://dublincore.org/> [Cited: 22 October 2001].

How the Oracle of Bacon Works, Department of Computer Science, University of Virginia, <http://www.cs.virginia.edu/oracle/how.html> [Cited: 22 July 2001].

McCarthy, Gavan 2001, ‘Of networks, entities and relationships: utilising the small world effect in an archival setting’ in the session ‘Building bridges between heritage resources’, *2001 A Global Archival Odyssey,* Society of American Archivists Annual Meeting August 2001, Washington DC USA [not as yet published].

McCarthy, Gavan and Joanne Evans 2001, ‘The Open Resource Scholarly Network: a new era for historians, archivists and technologists.’ *Computing Arts 2001* Digital Resources for Research in the Humanities, 26th-28th September 2001 conference University of Sydney. To be published on the World Wide Web.

National Archives of Australia 2001, Australian Government Locator Service (AGLS) <http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html> [Cited: 2 August 2001

National Library of Australia 2001, Persistent Identifiers, April 2001
<http://www.nla.gov.au/initiatives/persistence.html> [Cited: 12 November 2001].

Stepney, Susan 2001, ‘Complexity and self-organisation’ Susan Stepney’s personal page, Logica Public Server <http://public.logica.com/~stepneys/complex/index.htm> [Cited: 26 October 2001

TEI Consortium 2001, ‘Text Encoding Initiative. Welcome to the TEI Website’ <http://www.tei-c.org/> [Cited: 26 October 2001]

The PURL Team 2001, PURL <http://purl.oclc.org/> 11 November 2001 [Cited: 12 November 2001].

World Wide Web Consortium 2001 <http://www.w3.org/Addressing/URL/url-spec.html> [Cited: 2 August 2001].