The technology of open access

Chris Awre, Integration Architect, University of Hull

Enabling open access

In a presentation to a symposium on 'Free Culture & the Digital Library' in October 2005 Simeon Warner looked back over the 14 years since the beginning of the arXiv, the first major pre-print archive (Warner, 2005). This archive initially aimed to facilitate the sharing of pre-print articles between scholars in the high-energy theoretical physics community, though it has now grown to encompass a varied spectrum of subjects including mathematics, computer science and quantitative biology, as well as many other branches of physics. Email had already been used to distribute pre-prints of articles between interested scholars and research groups, but arXiv provided a place where these pre-prints could be deposited, organised and subsequently disclosed to the wider community. Originally both deposit and distribution was also by email, though this was quickly followed by ftp and not long after by Web interfaces to support this interaction.

Email and ftp have been two developments among many that have assisted scholars in communicating with each other and sharing information. Technology has long been harnessed to support this process: the origin of the Internet, ARPANET, was built between four US universities in 1969 to support the work of the Advanced Research Projects Agency (Leiner et al., 2003). The development and wide adoption of the World Wide Web, another initiative originally developed to support scholarly communication within the high energy physics community (Berners-Lee, 1990), has arguably, though, provided more and greater opportunities to support scholarly communication than any other development since ARPANET was put in place. Journal publishers were quick to take advantage of this, providing web access to the electronic equivalent of printed journals. However, as arXiv had shown, the Internet and the Web can be used to facilitate scholarly communication in other ways as well: the boundaries of the printed journal publication are no longer limits in the networked world.

The advent of the open access movement has been chronicled elsewhere in this book. The Budapest Open Access Initiative proposed in 2002 (Chan et al., 2002) suggested two complementary strategies through which open access might be achieved, taking full advantage of the networked opportunities that had arisen. These were self-archiving into repositories (BOAI1), as demonstrated by arXiv, and the production of open access journals (BOAI2), titles that facilitated the structured dissemination of research articles in a non-subscription environment. This chapter focuses on the technology that underpins these approaches and the ongoing development of solutions to further the exchange of scholarly communications in a world of networked access.

The Open Archives Initiative

The success of arXiv stimulated similar activity in other subject fields: CogPrints, covering psychology, linguistics, neuroscience and computer science; RePEc, focused on

economics; and the NDLTD, addressing the disclosure of theses and dissertations. As the number grew it became apparent that it would be valuable for open access archives to cooperate to enable easier access across them by researchers and others wishing to access their contents. In October 1999 a meeting in Santa Fe, USA led to the Santa Fe Convention of the Open Archives Initiative (Van de Sompel and Lagoze, 1999), subsequently renamed the Open Archives Initiative (OAI) and its Protocol for Metadata Harvesting (OAI-PMH), now at version 2.0 (Lagoze and Van de Sompel, 2003). The Initiative is a series of organisational principles and technical specifications to facilitate a level of interoperability between e-print archives. The underlying mechanism to enable interoperability is metadata harvesting, where metadata from different e-print archives can be harvested into a central service or services that can then be searched independently. Bowman et al. (1995) had originally described this architecture as part of the Harvest project.

The OAI established at an early stage two separate roles or participants in the harvesting model: data providers, which make available the data from an e-print archive or collection for harvesting; and service providers, which carry out the harvesting and provide end-user services based on these harvested collections. Data provision is an integrated part of many repository systems (see later) used to store content and associated metadata, though separate data provider software tools are also available. An early, but now well-established, service provider system is the open source Java-based Arc, developed at Old Dominion University (Liu et al., 2001, Liu et al., 2005), and now used widely by other service providers, notably the ePrints UK initiative (Martin, 2003). The Arc service provider can in turn be harvested by other service providers, and thus act as an aggregator data provider service as well as provide end-user search access. Other open source tools to enable both data providers and service providers are listed on the OAI website: of note are harvester software tools in Perl and PHP and a tool, DP9, that allows web crawlers such as Google to access metadata exposed for harvesting by OAI data providers. The OAI website also lists a number of existing data and service providers.

Adopting the OAI model is relatively straightforward, but does still require that the data provider be implemented fully, a task better suited to organisations than individuals. Two approaches have emerged to allow individual researchers to provide their outputs on open access. The team behind the Arc harvester developed the Kepler framework (Maly et al., 2001), which makes use of ‘archivelets’ to enable the publication of outputs and make them available for harvesting rapidly from a local PC rather than an institutional server. The executive managing the OAI itself has also developed a specification for OAI Static Repositories, which allows a locally stored XML file to be made available for harvesting by a remote service.

The OAI-PMH can quickly enable the sharing of metadata in the most circumstances, though it is accepted that it cannot meet every need yet. Usage has identified areas where improvements might be made for the future. The protocol is limited to XML files currently, requiring data conversion where relevant, and cannot work with RDF. It also makes specific use of the HTTP protocol, where a more abstract model would allow

greater flexibility in the network transport protocol used. The format for metadata to be harvested is, by default, unqualified Dublin Core. The protocol makes this mandatory, though only as a lowest common denominator, and leaves open the possible use of more complex metadata schemas. There is increasing experimentation with more detailed metadata formats (e.g., Richardson and Powell, 2003, Bird and Simons, 2003), though many OAI-PMH transactions continue to use Dublin Core as their basis.

The OAI does not sit in isolation in enabling open access. When building services the protocol can be combined with a number of other digital library protocols and standards to enable a range of functionality. The IMesh project developed a module that allows OAI records to be delivered using RSS (Duke, 2003), whilst there are also synergies and complementarities between OAI-PMH and SRW/U (Sanderson et al. 2005). The Ockham project has also described how a number of "light-weight" protocols can be combined to add value to services for the end-user (Xiang and Lease Morgan, 2005). There is much potential in how the OAI-PMH can be used that remains to be revealed, including the possibility of harvesting content as well as metadata.

Implementing self-archiving

At the heart of the first proposed BOAI strategy is a place to store content that will be made available through open access. By virtue of depositing e-prints etc. in this place they can be disclosed for others to view. The arXiv is an example of such a place, and researchers voluntarily add their pre-prints to this in order to foster the sharing and discussion of ideas. As noted earlier, a range of other subject-related and other archives have emerged since arXiv. These have not always limited themselves to pre-prints of potential journal articles, but encompass a wide range of documents and other resources that those in the community are willing to disclose.

E-print archives have been built on top of many technical platforms. The three main components required are a place to store the materials, a mechanism for depositing them and a mechanism for allowing access to them by others. Additional functionality may be provided alongside this. The terms 'archive' and 'repository' have both been applied to this package of functionality and have also been applied to the software available to support such systems. arXiv has been built on a platform of in-house development and the incorporation of tools as required. Others have made use of dedicated repository software. For example, the E-LIS archive for library and information science is built on top of EPrints software (Medeiros, 2004), developed at the University of Southampton as part of the Open Citation Project, which also examined and developed tools to enable citation linking from e-print archives (Hitchcock et al., 2002).

The establishment of e-print archives for subject communities has been gradual since the origins of arXiv. Since 2002 there has also been a great deal of activity in establishing and promoting institutional e-print archives (often labelled institutional repositories). Early repository initiatives at the Universities of Nottingham and Edinburgh both used the EPrints software (Pinfield, et al., 2002). At about the same time, SPARC in the US commissioned a report to investigate the potential of institutional repositories (Crow,

2002) and Cliff Lynch from the Coalition for Networked Information described the benefits institutions would gain from establishing a repository: enabling alternative scholarly communication paths was prominent amongst these (Lynch, 2003). The interest in institutional repositories led to the Open Society Institute producing a report on available open source software packages (Crow, 2004). This report offers a good starting point in consideration of open source software packages: it is noteworthy that the majority of e-print repositories use one of these systems, predominantly EPrints or DSpace, a collaborative development between MIT and Hewlett-Packard. Current usage of these systems can be viewed through the Repository of Open Access Repositories (ROAR) or Directory of Open Access Repositories (DOAR). The commercial sector has, though, also developed repository software that can assist open access, for example ProQuest's Digital Commons, Innovative's Symposia, and BioMed Central's Open Repository service. The latter is notable for providing a hosted service for institutions or organisations that are unable to implement their own system.

There are many aspects to implementation of an institutional repository, including both technical and non-technical aspects (Grieg and Nixon, 2005). Technical planning at an early stage is vital, however, to ensure the repository is capable of supporting its intended needs. Technical architecture and metadata are key to this planning.

Technical architecture

Planning for an institutional repository to allow open access to e-prints requires consideration of wider repository needs. Within the institution there may be different views required onto the repository: these specific needs could be addressed through alterations to the user interface or separate installations of the repository software, each with their own view onto the relevant content. The nature of the content being stored in the repositories will also have an impact. E-print repositories that focus solely on copies of peer-reviewed published papers (providing open access to these) can be set up separately to those for pre-prints or other materials, or they can all be included in one repository and flagged accordingly.

In an open access environment, it is also important to consider how any one repository will be accessed alongside others. Service providers were discussed earlier. The University of Glasgow investigated the use of a local OAI harvester to provide a single view across their repositories and have also been able to expose their repositories to Google (Nixon et al., 2005). This approach has also been adopted by the OAIster service provider with Yahoo!. With the flexibility of being able to move metadata (and potentially content itself) around using OAI-PMH, there is scope for individual repositories to be included in a wider federation through which content can be accessed and delivered in a flexible manner. Work on the aDORe architecture at the Los Alamos National Laboratory has highlighted many of the issues (and requirements) needed to enable this (Van de Sompel et al., 2005).

Metadata

Metadata has been at the core of cataloguing and information discovery systems for many years: catalogues hold metadata about a library's holdings and bibliographic databases hold metadata about a variety of different materials. This metadata has been used largely to describe physical content, though the metadata schemes employed, often MARC, have been adapted to describe digital content where required, often at a local level. Implementing an institutional repository offers an opportunity to re-visit how digital content should be described, and appropriate metadata scheme(s) put in place. A number of alternatives have been developed to meet the needs of the managing digital content (Jeevan and Nair, 2004), and the purpose and role of the repository will influence the choice.

For an e-prints repository Dublin Core metadata provides a means of describing articles to support interoperability between repositories and open access to them through appropriate services, and is, of course, mandated as the minimum requirement for use with the OAI-PMH. Metadata quality is an important part of facilitating management and access and this requires careful attention in the implementation of a self-archiving environment (Barton et al., 2003). The ePrints UK project have proposed some recommendations for how to describe e-prints using Dublin Core to encourage standardisation (Powell et al., 2003).

Implementing open access journals

The second strategy of the BOAI revolves around the production of journal titles that do not charge for subscription or access. A number of models have emerged from this strategy to provide free access to e-prints over the Web. The first port of call in discovering which open access journals exist is the Directory of Open Access Journals, based at Lund University in Sweden: as of February 2006 over 2000 titles are listed in this Directory. The technology underpinning these titles varies, as an open access journal can range in complexity from a simple web page to a fully interactive database-driven service. However, two mechanisms have emerged that can help facilitate the generation of open access journals.

E-prints in repositories provide a source of material for an open access journal. Indeed, generating a journal from repository content can be a value-added mechanism of providing more structured access to the repository's contents. The emphasis can come from both directions. The journal can be based on repository contents, for example the Lund Virtual Medical Journal which is based on the Lund University institutional repository LU:research, or the repository can hold e-prints submitted for inclusion in the journal from, for example the Journal of eLiteracy at the University of Glasgow. Overlay journals, as these titles are sometimes referred to, have also been set up over subject-based repositories: Advances in Theoretical and Mathematical Physics is based on submissions to arXiv and additional titles based on arXiv's holdings have also been established.

It has been argued that true overlay journals amalgamate from across more than one repository [55]. The American Institute of Physics and American Physical Society offer

a series of virtual journals that bring together content from other publications, though these are not open access [56]. However, in a discussion at the 3rd CERN Workshop on Innovations in Scholarly Communications it was considered that more content and greater consistency is required in institutional repositories to fully support this model of overlay journals [57].

The ARROW project based at Monash University in Australia has as part of its remit the development of an e-press, to be built alongside and supported by the repository [58]. This 'overlay' activity doesn't just rely on repository contents, though, but proactively seeks to use the repository as part of the e-publication process. More specific development of systems to support the running of an open access journal has also taken place, for example as part of the Public Knowledge Project in Canada [59]. Their Open Journal System supports many of the processes involved in formal journal publication, but for an open access environment.

Looking ahead

The technology to enable the establishment of e-print archives and repositories is now relatively mature in its ability to support open access scholarly communication. Much development focuses on the policy framework within which these technologies sit. However, there are also many technical investigations in areas that will enhance open access scholarly communication still further. Two are briefly described here.

Following on from the Open Citation Project mentioned earlier citation analysis of open access articles is attracting growing attention. The project itself led to the development of the Citebase citation search index [60] based on harvested open access e-prints. A report by ISI in 2004 noted that open access articles were being cited highly alongside those published through toll journals [61] and they are releasing the Web Citation Index, which will cover institutional repositories as well as open access journals. These tools will enable ongoing analysis of the impact of open access publishing.

Harvesting e-prints has largely been limited to metadata about them: the OAI-PMH is about metadata harvesting after all. However, there is scope to harvest the full content of an e-print and related materials where relevant. Van de Sompel et al. have described a potential path to allow complex objects to be harvested using OAI-PMH, containing both content and metadata [62]: this has the potential for enabling a step up in the ability to communicate over the network and share research outputs.

In conclusion, technical advances and the underpinning network have opened up the development of new techniques to support scholarly communication. It is likely that such advances will continue and support future scholarly communication and research through open access and collaboration.

References

Barton, J., Currier, S. and Hey, J. (2003) 'Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice', *DC-2003 (2003 Dublin Core Conference) Supporting Communities of Discourse and Practice - Metadata Research and Applications*, Seattle, Washington, USA, 28th Sep-2nd Oct 2003. Available at *http://www.siderean.com/dc2003/201_paper60.pdf* [accessed 27/2/06].

Berners-Lee, T. (1990) *Information management: a proposal*. Available at *http://www.w3.org/History/1989/proposal.html* [accessed 17/2/06].

Bird, S. and Simons, G. (2003), 'Building an Open Language Archives Community on the OAI foundation', *Library Hi Tech* 21 (2): 210-218.

Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U and Schwartz, M.F. (1995) 'The Harvest Information Discovery and Access System', *Computer Networks and ISDN Systems*, 28 (1/2): 119-125.

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J-C., Hagemann, M., Harnad, S., Johnson, R., Kupryte, R., La Manna, M., Rév, I., Segbert, M., de Souza, S., Suber, P. and Velterop, J. (2002) *Budapest Open Access Initiative*. Available at *http://www.soros.org/openaccess/read.shtml* [accessed 21/2/06].

Crow, R. (2002) *The case for institutional repositories: a SPARC position paper*. Available at *http://www.arl.org/sparc/IR/ir.html* [accessed 21/2/06].

Crow, R. (2004) *A guide to institutional repository software*, 3rd edition. Available at *http://www.soros.org/openaccess//software/* [accessed 21/2/06].

Duke, M. (2003) 'Delivering OAI records as RSS: an IMesh toolkit module for facilitating sharing', *Ariadne*, Issue 37: October 2003. Available from *http://www.ariadne.ac.uk/issue37/duke/* [accessed 20/2/06].

Grieg, M. and Nixon, W. (2005) 'DAEDALUS: delivering the Glasgow ePrints service', *Ariadne*, Issue 45: October 2005. Available from *http://www.ariadne.ac.uk/issue45/greig-nixon/* [accessed 21/2/06].

Hitchcock, S., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C. and Harnad, S. (2002) 'Open citation linking: the way forward', *D-Lib Magazine*, 8 (10): October 2002. Available from *http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html* [accessed 21/2/06].

Jeevan, V.K.J. and Nair, S.S. (2004) 'A brief overview of metadata formats', *DESIDOC Bulletin of Information Technology*, 24 (4): 3-11.

Lagoze, C. and Van de Sompel, H. (2003) 'The making of the Open Archives Initiative Protocol for Metadata Harvesting', *Library Hi Tech*, 21 (2): 118-128.

Leiner, B.M., Cerf, V.G., Clark, D.D., Kahn, R.E., Kleinrock, L., Lynch, D.C., Postel, J., Roberts, L.G. and Wolff, S. (2003) *A Brief History of the Internet*. Available at *http://www.isoc.org/internet/history/brief.shtml* [accessed 17/2/06].

Liu, X., Maly, K., Zubair, M. and Nelson, M.L. (2001) '*Arc* - An OAI service provider for Digital Library Federation', *D-Lib Magazine*, 7 (4): April 2001. Available at *http://www.dlib.org/dlib/april01/liu/04liu.html* [accessed 20/2/06].

Liu, X., Maly, K., Nelson, M.L. and Zubair, M. (2005) 'Lessons learned with Arc, an OAI-PMH service provider', *Library Trends*, 53 (4): 590-603.

Lynch, C. (2003) 'Institutional repositories: essential infrastructure for scholarship in the digital age', *ARL*, no. 226: 1-7. Available at *http://www.arl.org/newsltr/226/ir.html* [accessed 21/2/06].

Maly, K., Zubair, M. and Liu, X. (2001) 'Kepler: an OAI data/service provider for the individual', *D-Lib Magazine*, 7 (4): April 2001. Available at *http://www.dlib.org/dlib/april01/maly/04maly.html* [accessed 20/2/06].

Martin, R. (2003) 'ePrints UK: developing a national e-prints archive', *Ariadne*, Issue 35: April 2003. Available at *http://www.ariadne.ac.uk/issue35/martin/* [accessed 20/2/06].

Medeiros, N. (2004) 'A repository of our own: the E-LIS e-prints archive', *OCLC Systems and Services*, 20 (2): 58-60.

Nixon, W., Drysdale, L. and Gallacher, S. (2005) *Search services at the University of Glasgow: PKP Harvester and Google*. Available at *https://dspace.gla.ac.uk/handle/1905/425* [accessed 21/2/06].

Pinfield, S., Gardner, M. and MacColl, J. (2002) 'Setting up an institutional e-print archive', *Ariadne*, Issue 31: April 2002. Available at *http://www.ariadne.ac.uk/issue31/eprint-archives/intro.html* [accessed 21/2/06].

Powell, A., Day, M. and Cliff, P. (2003) *Using simple Dublin Core to describe eprints,* ePrints UK report. Available at *http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/* [accessed 27/1/06].

Richardson, S. and Powell, P. (2003), 'Exposing information resources for e-learning – harvesting and searching IMS metadata using the OAI Protocol for Metadata Harvesting and Z39.50', *Ariadne* Issue 34: January 2003. Available at *http://www.ariadne.ac.uk/issue34/powell/* [accessed 20/2/06].

Sanderson, R., Young, J. and LeVan, R. (2005) 'SRW/U with OAI: expected and unexpected synergies', *D-Lib Magazine*, 11 (2): February 2005. Available at *http://www.dlib.org/dlib/february05/sanderson/02sanderson.html* [accessed 20/2/06].

Van de Sompel, H. and Lagoze, C. (2000) 'The Santa Fe Convention of the Open Archives Initiative', *D-Lib Magazine*, 6 (2): February 2000. Available at *http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html* [accessed 17/2/06].

Van de Sompel, H., Bekaert, J., Liu, X., Balakireva, L. and Schwander, T. (2005) 'aDORe: a modular, standards-based digital object repository', *The Computer Journal*, 48 (5): 514-535. Preprint available at *http://arxiv.org/ftp/cs/papers/0502/0502028.pdf* [accessed 21/2/06].

Warner, S. (2005) *The arXiv: 14 years of open access scientific communication*, Symposium on Free Culture & the Digital Library, Emory University, Atlanta, 14th October 2005. Available at *http://www.cs.cornell.edu/people/simeon/talks/Emory_2005-10-14/arXiv_history_talk.pdf* [accessed 17/2/06].

Xiang, X. and Lease Morgan, E. (2005) 'Exploiting "light-weight" protocols and open source tools to implement digital library collections and services', *D-Lib Magazine*, 11 (10): October 2005. Available at *http://www.dlib.org/dlib/october05/morgan/10morgan.html* [accessed 20/2/06].