Repositories post 2010: embracing heterogeneity in AWE, the Academic Working Environment

[This paper has been accepted for the main track of the Open Repositories Conference with very strong reviewer feedback. I'll be there.]

Peter Sefton
sefton@usq.edu.au
University of Southern Queensland

Duncan Dickinson
Duncan.Dickinson@usq.edu,au
University of Southern Queensland

Open Repositories July 2010, Madrid, Spain

Abstract: The organizers of the fifth international conference on Open Repositories list nine polar dichotomies that represent “The Grand Integration Challenge” for the repository community/movement. In this paper we take up the challenge. We do so in the context of a program of work being undertaken at our institution to build infrastructure for the academy in general, working towards a modular 'Academic Working Environment' (AWE) which encompasses both teaching and learning on one hand and research on the other. Repositories and the ecosystem of services and workflows that surround them play a key role in this emerging system.

1 Introduction

This presentation uses the conference organizers' description of current issues in repositories as a way to structure discussion about a program being undertaken at the Australian Digital Futures Institute; The Academic Work Environment (AWE). The presentation will look at each of the integration challenges listed on the call for papers: the web and the repository, knowledge and technology,wild and curated content, linked and isolated data, disciplinary and institutional systems, scholars and service providers, ad-hoc and long-term access, ubiquitous and personalized environments, the cloud and the desktop.

2 AWE: Meeting the grand challenges for repositories

The Academic Work Environment (AWE) is a label for a set of services and computational systems that support the academic enterprise. The acronym was coined to capture under a single identifier a range of research and development work going on in a group tasked with pragmatic, practical work on workflows and computer systems for eLearning and eResearch.

2.1 The web and the repository & The cloud and the desktop

It is telling that *the web vs the repository* is posed as a dichotomy or a challenge. While repositories are, by default, web-based systems the vast majority of document content housed in repositories is in a non-web format, PDF. Furthermore, vast amounts of potential repository content such as data-files are excluded from the repository, and from the web/cloud because of the lack of services that assist academic users in making their content available on the web. The Academic Work Environment addresses both of the issues identified above with systems to:

1. Allow academic documents to be made available in HTML as well as PDF. This work, based on The Integrated Content Environment (ICE) was presented at Open Repostiories 2009 (Sefton, Downing, & Day, 2009). Uptake on this has been very slow, but repositories need to start using web formats if they are to be part of the web and fulfill the promise of integrated documents and data as envisaged by Murray Rust and Rzepa (Murray-Rust & Rzepa, 2004).
2. Begin to close the gap between the desktop work environment and the repository via an application which bring a web-based repository view to all the files that a researcher/educator is using; allowing them to describe them, back them up and have them routed to appropriate repository for works-in-progress and completed research outputs as appropriate. (Dickinson & Sefton, 2009; Sefton, 2009)

Other work in this area includes Microsoft Research's work on embedding SWORD – named OfficeSWORD[1] – repository deposit into their word processing product as well as some attempts to allow structured authoring and semantic web-authoring in the browser (Fernicola, 2009; Fink et al., 2010).

2.2 Knowledge and technology

Whilst individual web-pages may contain information that people can use to construct knowledge, the technology of the web itself is largely ignorant to the concept of information and the notion of a body of information being spread across the network. Authors may hyperlink to provide readers with further context but the web browser is a dumb-terminal that understands nothing about the actual information embedded in the content. Search engines provide a basic entry point to the information network but their reliance is on key phrases and not the context and meaning understood by a community of practice (Neumann & Prusak, 2007). The challenge now is to open up the information existing in online forms – including documents and data – so that it is findable[2], a concept that spans more than just matching keywords:

> Any system aiming to integrate heterogeneous data on an ad hoc basis and present this to users will need to adopt sophisticated models of relevance, quality, and trust that are sensitive to the user's current task and its context. (Heath Heath, 2008 p. 91)

Heath's discussion here is focused on the Semantic Web. Based on technologies from groups such as the W3C[3], the Semantic Web provides a framework within the Web that gives web browsers and search engines the ability to interact with information. Navigation in this model is based on assembling meaning rather than merely providing presentation services (Heath, 2008). The challenge, then, is to get the semantics into the Semantic Web. Within the AWE we are looking at methods that allow researchers to easily provide semantic information in their data and documents. The Anotar[4] project presented a framework for adding semantics through the well understood concept of tagging and extending this by allowing taxonomies to be utilised. Created as an easy to adopt toolkit, Anotar is being adapted to systems such as The Fascinator, Moodle and WordPress. Initial pilot work has also been undertaken to provide facilities for adding semantic information to word processed documents so that semantic "mark up" can be created with the document rather as a deferred technical effort (Sefton, 2009).

2.3 Wild and curated content

The AWE approach to wild and curated content is to gradually tame and domesticate the content, by allowing it to be husbanded by a series of 'curation events'. Using the Fascinator Desktop, The initial creator may label data items or sets in simple ad-hoc terms using tags such as "My Thesis" or "Anthropology 101 Course Notes". The incentive to do this is that when they do so, the items will be (a) backed up appropriately and (b) routed to collaborators automatically. This represents what we might call an emergent workflow; where object state changes result in items making progress through required stages.

We also provide a more intentional kind of curation via 'acts of publishing' where a data owner can push content across curation boundaries (the term coined by ARROW project members (Treloar, Groenewegen, & Harboe-Ree, 2007)) for various reasons;

- a draft can be pushed to a blog (remember that, using ICE content services, all word processing documents in the Academic Working Environment are web-ready and can be delivered as HTML)
- a finished paper or approved draft can be pushed to the institutional repository. Work is under way to enable this to happen in a way that (a) data is also deposited and (b) both document and data are web-ready and can be viewed in-browser as native web content, as well as being available for print .
- We are building systems for the Australian National Data Service to allow data to be registered with Research Data Australia [NOTE: this contract is not confirmed at time of writing but we expect to be able to demo solutions at conference time].

---

1 http://officesword.codeplex.com/

2 http://en.wikipedia.org/wiki/Findable

3 http://www.w3.org/2001/sw/

4 http://www.purl.org/anotar/

2.4 Linked and isolated data & Ad-hoc and long-term access

Bootstrapping the linked-data web remains a grand challenge, but we are attempting to address it in work on The Fascinator Desktop by providing URIs (the formal name for links) for data while it is still in isolation in a lab or on a laptop computer for example. Managing identifiers through the lifecycle of a digital object is not easy when you consider that a researcher may have their digital files spread across multiple desktop, portable and mobile devices and that, in the messy landscape of desktop filesystems, filenames are changed at a whim and multiple versions may exist throughout a system.

Within The Fascinator, every item will have a URI from the moment it is discovered on the user's desktop.

Creating a URI for desktop files is a similar approach to the SemDeskURI Scheme suggested by members of the Nepomuk[5] project (Sauermann, 2008). However, instead of relying on a new protocol (in the case of SemDeskUri this is "*desktop:///*"), URIs in The Fascinator will utilise "*http://*". For example, whilst "http://localhost/dickinso@usq.edu.au/research/data" will not allow remote users to access the resource, it does provide for contextual identification.

In terms of ad-hoc vs long-term access, under this scheme we would expect resources to move from an isolated desktop-web view of digital objects (remembering that our systems give people a web view from the very creation of the object) to ad-hoc team views where the provenance of items is preserved by keeping both desktop and team-URIs, through to more formally created views. As content is routed from the desktop to shared repositories the plan is to keep the URIs, so that the metadata for an item contains all of the known identifiers that we have for it.

2.5 Disciplinary and institutional systems / Scholars and service providers

Whilst much of this presentation has focused on the technical aspects of "The Grand Integration Challenge", the complexity of institutional and individual engagement must be considered. This is a challenge that threads through numerous areas of academia, including postgraduate researcher skills, central ICT provision, Government reporting requirements, Library systems, research project management etc. Indeed, it presents a "wicked problem"[6] for the academic research sector and is one that various stakeholders approach with a carrot and/or stick approach.

From the carrot side, we are working with motivated pilot users and faculties to build the AWE to meet their needs. Another carrot, of course, is grants - such as those offered by the Australian National Data Service – to teams developing solutions. The stick often presents itself as institutionalised mandates ("Data-sharing culture has changed," 2009; Australian Government, 2007) but for many researchers, the technical complexities around data management are an intrusion into their time (Henty, Weaver, Bradbury, & Porter, 2008). As mentioned earlier, a central goal of the AWE is to provide services that meet the various stakeholder demands but don't interfere with the researcher's core tasks.

2.6 Ubiquitous and personalized environments

Two central goals of the AWE are to hide the technicalities around data management and the semantic web and to provide services that meet various stakeholder demands. These goals are both aimed at allowing the researcher to focus on their research. We envisage a mesh of repository services, building on the existing standards for linking repository content; the Fascinator Desktop work (and before it The Integrated Content Environment) introduces the idea of a personal desktop web; with services such as tagging and note-taking acting not only in their traditional role of assisting in the research process, but as triggers in emergent workflows.

A good example to illustrate how the personal and ubiquitous can meet and "intermesh" is our work on a general framework for annotations, Anotar. In work extending from that described in Dickinson and Sefton (2009), we are adding Anotar annotation services to The Fascinator. In the case of our work with USQ's Public Memory Research Centre, this facility will allow the researcher to tag photos in their desktop repository for release to an online repository for viewing. Research participants can log onto this repository and provide the researcher with essential information regarding the people and locations in various photos and movies through the use of taxonomy-based tagging. Furthermore, the open-ended nature of the annotations will provide an online forum for the participants to share their memories and

5 http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main1/

6 http://en.wikipedia.org/wiki/Wicked_problem

even debate points of view – providing a rich data-set for the researcher as well as a first-person public memory archive.

Whilst tagging and annotation services are provided (to some extent) by various online Web 2.0 systems, the AWE solution meets research-specific data management requirements by keeping research data in a way that adheres to the University's ethical clearance requirements. For the researcher and their community of participants, it provides a personalized environment that lets them focus on their own goals rather than the technical infrastructure.

3 Conclusion

The Academic Working Environment is not a single monolithic application with a “one size fits all” approach. Its name defines its structure: an environment of interoperable services that works with the researcher and doesn't hamper their efforts.

4 References

Australian Government. (2007). *Australian code for the responsible conduct of research*. Canberra, Australia: Australian Government. Retrieved from http://www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/r39.pdf

Data-sharing culture has changed. (2009, December 11). *Research Information*. Retrieved November 19, 2009, from http://www.researchinformation.info/news/news_story.php?news_id=553

Dickinson, D., & Sefton, P. (2009). Creating an eResearch desktop for the Humanities. Presented at the eResearch Australasia 2009, Manly, Australia. Retrieved from http://eprints.usq.edu.au/6090/

Fernicola, P. F. (2009). Incorporating Semantics and Metadata as Part of the Article Authoring Process. Retrieved March 1, 2010, from http://elpub.scix.net/cgi-bin/works/Show?152_elpub2009

Fink, J. L., Fernicola, P., Chandran, R., Parastitidas, S., Wade, A., Naim, O., Quinn, G., et al. (2010). Word add-in for ontology recognition: semantic enrichment of scientific literature. *BMC Bioinformatics*, *11*(1), 103. doi:10.1186/1471-2105-11-103

Heath, T. (2008). How Will We Interact with the Web of Data? *Internet Computing, IEEE*, *12*(5), 88-91.

Henty, M., Weaver, B., Bradbury, S. J., & Porter, S. (2008). *Investigating Data Management Practices in Australian Universities*. Retrieved from http://eprints.qut.edu.au/14549/

Murray-Rust, P., & Rzepa, H. S. (2004). The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information*, *5*(1), 248.

Neumann, E., & Prusak, L. (2007). Knowledge networks in the age of the Semantic Web. *Briefings in Bioinformatics*, *8*(3), 141-149. doi:10.1093/bib/bbm013

Sauermann, L. (2008, October 27). RFC-draft: SemDesk URI Scheme. Retrieved February 25, 2010, from http://dev.nepomuk.semanticdesktop.org/repos/trunk/doc/2008_09_semdeskurischeme/index.html

Sefton, P. (2009, April 8). Journal 2.0: Embedding semantics in documents. *ptsefton*. Retrieved February 26, 2010, from http://ptsefton.com/2009/04/08/journal-20-embedding-semantics-in-documents.htm

Sefton, P., Downing, J., & Day, N. (2009). ICE-theorem - end to end semantically aware eResearch infrastructure for theses. *University of Southern Queensland*. Retrieved August 24, 2009, from http://eprints.usq.edu.au/5248/1/ice-theorem-paper-OR09.htm

Sefton, P. M. (2009). The Fascinator - Desktop eResearch and Flexible Portals. Presented at the 4th International Conference on Open Repositories, Georgia Institute of Technology. Retrieved from http://smartech.gatech.edu/handle/1853/28483

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The Data Curation Continuum. *D-Lib Magazine*, *13*(9/10). doi:10.1045/september2007-treloar