Towards Scholarly HTML

Author: Peter Sefton, Manager Software Research and Development Team, Australian Digital Futures Institute, University of Southern Queensland.

Abstract: The editors of this special edition originally asked if Institutional Repositories are shaping and changing scholarly communications. While the Open Access movement has had a profound impact, there is at least one area where they are not; while repositories are web based, they almost always contain only print-oriented materials which are not fully utilizing the fabric of the web and utilizing its ability to link documents to each other and to data.

This article will explore some of the factors that have contributed to this situation, then report on work undertaken at the University of Southern Queensland to make repositories more web-like, using a publishing system which was originally devised at the university for publishing course materials to multiple formats. It concludes with a description of a new model for changing scholarship by focusing on post graduate theses and new journals in an open access context.

The problem, paper based thinking

Two[1] types of software packages had a profound effect on publishing before the World Wide Web. Beginning in the late 1970's, the word processor provided for efficient production of text – words could be rearranged and edited without having to be re-typed from beginning to end or altered using physical means such as cut and paste of typed text, or obscuring typescript with white paint and then typing or writing over it. Then from the mid-1980's, desktop publishing began to democratize access to typesetting tools. Because of this desktop publishing revolution, word processors began to ship with a "What You See Is What You Get" (WYSIWYG) view, which encouraged text production and formatting to become one and the same operation coombs et al give a useful summary of the state of the art in 1987[2]. Before the web, authors had become used to editing within the constraints of the A4 or US Letter page. It has been argued that this WYSIWYG desktop publishing revolution has had a counter-productive effect on the progress of publishing, and in particular scholarly publishing by promulgating what Sorgaard and Sandahl call the 'paper metaphor'.[3]

The World Wide Web is now the key distribution for scholarship, offering new potential for documents which are seamlessly integrated with machine-readable data, and with human-readable visualization services for data as discussed by Murray-Rust and Rzepa in their paper introducing the term datument to represent this new kind of eScholarship[4]. Since the web arrived in the mid 1990's, writing tools such as Microsoft Word have failed to adapt to a web-based publishing environment[5], with typical web export from word processors producing HTML which is far from standards compliant and unsuitable for use on journal websites or in institutional repositories. I will show below how this works with the dominance of the paper metaphor to reinforce the role of the PDF as the currency for research reporting.

Of course word processors are merely the most common writing tool for the academy, not necessarily the best. Many proposals have been put forward for structured authoring using XML (and before that SGML) which has worked well in industries such as legal publishing and the military where specialist editorial teams can be trained and supported. Norman Walsh[6] captures the essence of the advantages of structured authoring in a contribution to a debate in the Journal of Digital Information. While the principle is sound in a theoretical sense, experience at USQ with an end to end XML publishing system for course materials has not been encouraging – the completed system had close to zero uptake from academic staff[7]. It is true that many commercial publishers use XML in their production systems, but it is unusual for authors to

1 extra

2 James H. Coombs, Allen H. Renear, and Steven J. DeRose, "Markup systems and the future of scholarly text processing," *Commun. ACM* 30, no. 11 (1987): 933-947, http://portal.acm.org/citation.cfm?id=32206.32209.

3 P. Sørgaard and T. I. Sandahl, "Problems with Styles in Word Processing: A Weak Foundation for Electronic Publishing with SGML," *Proceedings of the 30th HICSS*.

4 P. Murray-Rust and H. S. Rzepa, "The Next Big Thing: From Hypermedia to Datuments," *Journal of Digital Information* 5, no. 1 (2004): 248, http://jodi.tamu.edu/Articles/v05/i01/Murray-Rust/?printable=1.

5 Peter Sefton, "eResearch for Word users?," in , 2008.

6 N. Walsh, "XML: One Input--Many Outputs: a response to Hillesund," *Journal of Digital Information* 3, no. 1 (2002).

contribute XML; in most cases authors submit word processing files and these are converted to XML behind the scenes. I reviewed the state of the art and the (very limited) literature on how word processing might be integrated into back-end publishing systems in a paper for the Australian World Wide Web conference[8], and in 2008 for the Australasian eResearch conference[9]. Since then Microsoft Research has released previews of a Microsoft Word based tool which is claimed to produce XML conforming to the National Library of Medicine schema, but there is so far no evidence of the tool being used by typical authors.

To illustrate the potential divide between the author's version and the publisher's; Elsevier, the publisher of this journal recently ran a competition, *Article 2.0*[10] to show the future of a scientific article. The competition winner shows that a journal article may be the web locus for discussion, annotation and semantic relationships, but this competition was built on XML source documents which are created and held by the publisher so there is no way that a typical institutional repository could easily provide the same services. This is a case where the publisher is shaping scholarly communications, or at least exploring how to do so, but a lack of tools means that repositories are unlikely to be able to do likewise. This creates a distinct divide between the publisher's more richly marked-up version and the version held by the author in word processing format or the typesetting system LaTeX[11], neither of which allow high quality HTML unless the author has used a particular set of templates and/or macros and has access to specifically conversion software. So there is no way for most author manuscripts - which are commonly deposited in Institutional Repositories - to be turned into usable web content, let alone with links to data and semantic-web content. The best most authors could hope for with their version would be to convert it to PDF and deposit in a repository, while the publisher can do much more with the article.

Against this background, our work in shaping scholarly communications in the Australian Digitial Futures Institute (ADFI) USQ has been focused on three areas:

1. Empowering authors with content creation using tools that are not constrained by the paper metaphor so that no matter what happens on the publisher's side of the transaction authors can use and re-use their work as flexibly as possible.
2. Work on integrating with institutional repositories, particularly in packaging using the Open Access Initiative protocol for Object Reuse and Exchange (OAI-ORE)[12].

---

7 Peter Sefton, “eResearch for Word users?,” in , 2008.

8 P Sefton, “The integrated content environment,” in *AUSWEB 2006* (presented at the AUSWEB 2006, Noosa: Southern Cross University, 2006), http://eprints.usq.edu.au/archive/00000697/01/Sefton_ICE-ausweb06-paper-revised-3.pdf.

9 Peter Sefton, “eResearch for Word users?,” in , 2008.

10 Elsevier, “Elsevier Article 2.0 Contest,” Publisher website, *Elsevier*, 2009, http://article20.elsevier.com/contest/home.html.

11 P.T. LaTeX3, “LaTeX2E for authors,” 2001, http://www.latex-project.org/guides/usrguide.pdf.

12 SLYCAT IONSREPORT, “OAI-ORE specifications,” *Scholarly Communications Report* 12, no. 1 (2008): 5-5.

3. Most importantly, providing the above tools for use by post graduate students writing theses; feeding new users into the academy who know how to create 'Article 2.0' type content for themsleves.

These themes are discussed below, showing how the technology now exists for authors to transcend the limitations of paper formatting. With these in place, there will be scope for truly open journal processes to emerge both with open and toll-access publishers and the discontinuity between the author's version and the publishers will disappear.

The Integrated Content Environment

The Integrated Content Environment (ICE) is an open source system created at USQ for producing courseware[13]. In early 2009 it became a core system, considered an essential part of the operations of the university. In the ICE system, efforts have been made to make sure that authors are not constrained by the "paper metaphor"; it provides the ability to preview a document continuously in web format during its production, thus encouraging the use of styles, and empowering users to create structured documents in their word processor. ICE has one very important property, while it is in some sense an XML based publishing system, the main XML format it targets is XHTML[14], while print versions are currently produced by using word processing software in an completely automated process. This provides many of the benefits of flexible delivery using XML, with low development costs for print production and a lot of flexibility for authors to control formatting details while still creating structured documents.

ICE Features

ICE has several key features:

- Automated document conversion with a rapid feedback cycle so authors can see their content in a web context as they write.
- In line threaded annotation in a web view. The screenshot in Illustration 1 shows a colleague commenting on draft version of this paper.
- Protocols for linking to data that support a document and make research more reproducible as well as provide live visualization tools. This has been demonstrated in the TheOREM project [note we will reference a forthcoming paper at Open Repositories 2009 here – details TBC]

13 P Sefton, "The Integrated Content Environment for Research and Scholarship," Project site, *ICE Website*, 2006, http://ice.usq.edu.au/introduction/ice_rs.htm.

14 W3C, "XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition)," 2002, http://www.w3.org/TR/xhtml1/.

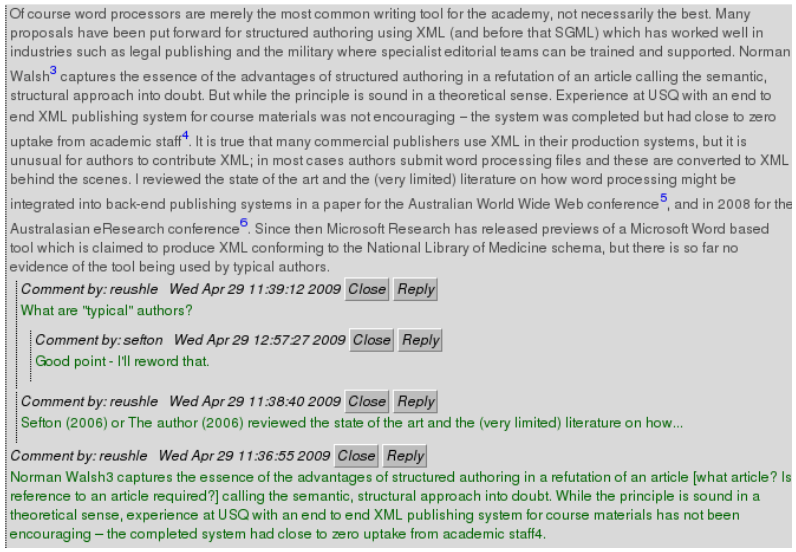


Illustration 1: A screenshot of the ICE inline annotation system in a web browser, showing threaded comments

ICE can be used in three modes:

1. As a complete distributed content management system as used for course creation at USQ.
2. As a central content management system where groups can collaborate on content.
3. As a set of software components that can be embedded in other applications,. This has been done at USQ to integrate ICE conversion services into the Moodle learning management system, and some experiments have been undertaken with the Open Journal Systems software.

Journals in ICE

The e-Journal of Instructional Science and Technology (e-JIST) is an open access journal published by USQ which served as a demonstration. The final issue serves as an example of how ICE can be used in journal production.

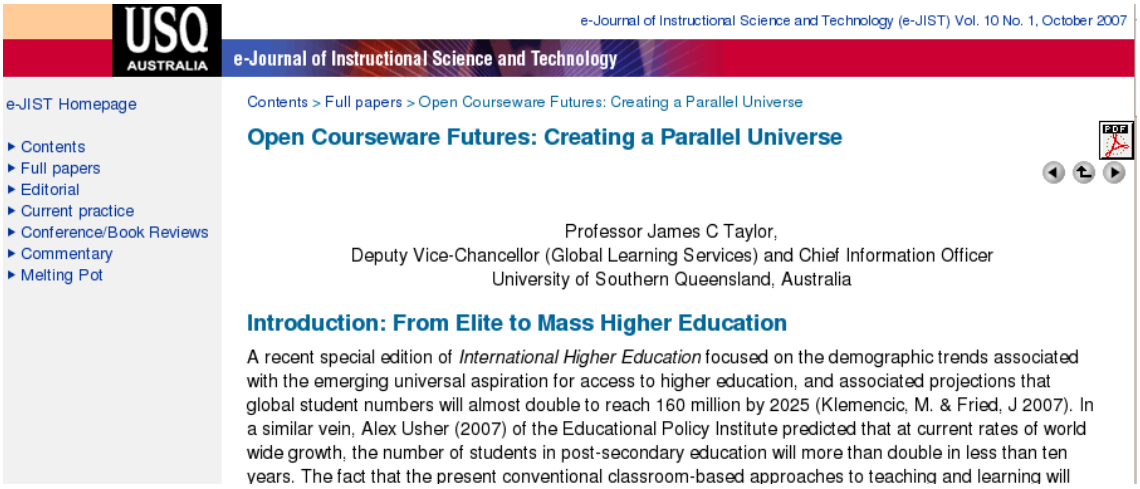


Illustration 2: A screenshot of the eJist journal in a web browser

This screenshot of a paper[15] in Illustration 2 shows that a low-end journal may be produced in HTML and PDF from single source files, but it does not deal with authoring or submission processes.

We now have a nascent collaboration with the Public Knowledge Project.

> *The Public Knowledge Project is a research and development initiative directed toward improving the scholarly and public quality of academic research through the development of innovative online publishing and knowledge-sharing environments.*[16]

With PKP we are looking at how ICE might be integrated with their suite of open access journal, conference and monograph work flow tools, particularly the Open Journal Systems (OJS)[17] to provide the same benefits outlined above for ICE, with authors able to see their content in both web and print formats as soon as they upload it, and with reviewers having the option to critique submissions online in a web browser.

Related to OJS, work by a consortium of Australian institutions[18] resulted in a packaging mechanism for journals, leading to the ability for journals created with the OJS to sent to a repository automatically; proof of concept work has been completed that shows that ICE templates coupled to a system like OJS could automate production of web and print versions of journal articles[19].

---

15 J. C. Taylor, "Open Courseware Futures: Creating a Parallel Universe," *e-Journal of Instructional Science and Technology (e-JIST)* 10, no. 1 (2007).

16 PKP, "Public Knowledge Project |," Project webstite, *Public Knowledge Project*, 2009, http://pkp.sfu.ca/.

17 "Open Journal Systems | Public Knowledge Project," Project website, *Public Knowledge Project*, 2009, http://pkp.sfu.ca/?q=ojs.

18 J. Pearce et al., "The Australian METS Profile–A Journey about Metadata," *D-Lib Magazine* 14, no. 3/4 (2008): 1082-9873, http://www.dlib.org/dlib/march08/pearce/03pearce.html.

19 USQ, "ICE Developers Blog » Blog Archive » Linking authoring tools to repositories," Project weblog, *Integrated Content Environment Blog*, 2007, http://ice.usq.edu.au/blog/2007/12/18/linking-authoring-tools-to-repositories.html.

Theses

While journal articles are most commonly distributed as PDF files, so today are electronic theses and dissertations. The ADFI has been involved in two projects which are attempting provide tools for thesis writing which result in web-integrated documents with PDF version for printing and with data integrated as best as they can be into the document and the production workflow. I presented an initial proof of concept for ICE as a thesis editor at the 2007 Electronic Thesis and Dissertation conference[20], containing a comparison with other thesis production processes based on XML formats, which led to a collaboration with Murray-Rust's team on the ICE-TheOREM[21] project which builds on work in ADFI on converting theses to ICE format. Internally, we are working with a small group of PhD candidates to supply an ICE server for thesis management, where candidates will be able to store their thesis as a set of word processing documents, along with data files. Supervisors will be able to comment on the work using an annotation system without having to touch the original documents.

Working with early career researchers is one way to bring about change in scholarly communications. Theses are under institutional and departmental control so it is possible to influence candidates to work with innovative systems, in the hope that practices which improve the dissemination and re-reproducibility of research will be taken up by these new researchers as they enter the academy.

In addition to a the cultural aspects of the work with PhD candidates there is an important technical dimension to the work on ICE-TheOREM, which is to formalize the way that systems such as content management systems and repositories can exchange content. The *ORE* in TheOREM is for Object Reuse and Exchange – a protocol that for the first me provides a web-native way to describe compound web resources such as theses consisting of multiple chapters, with associated images, style sheets and data files. The project has demonstrated how a thesis can be created and managed with a word processor and automatically packaged using an ORE resource map for automated ingest into a repository.

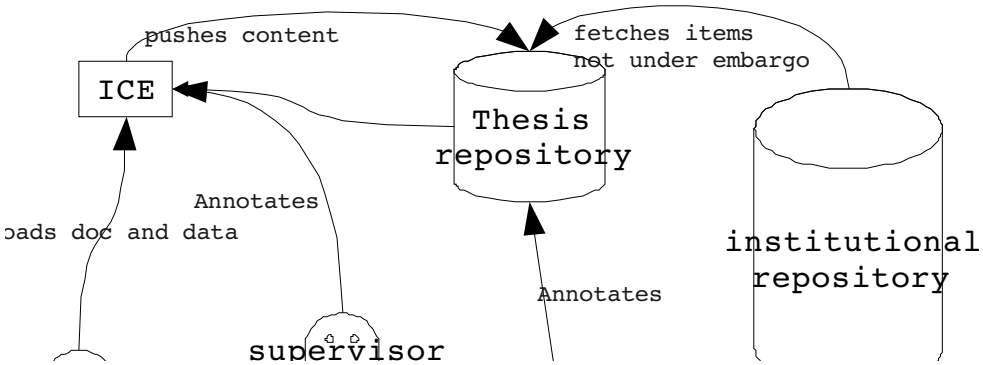
ICE-TheOREM has produced a new model for repository-aided thesis production and dissemination where an author can manage their thesis, have it backed up automatically, with a complete audit trail of stand-off annotations left by their supervisor(s). When the thesis is ready for examination, it can be placed in a research-office repository from where the examination process is managed. There is potential here for examiners to use the same kinds of annotation systems and evaluative forms as are used by supervisors, but PDF for reading in print form is also provided. Feedback from this system would need to be sent back to ICE, for action by the candidate, which is a feature we have yet to build. From the thesis repository, content is delivered to the institutional repository automatically via a pull protocol Atom Archive, as it comes off embargo.

---

20 Peter Sefton, “An integrated approach to preparing, publishing, presenting and preserving theses,” in *ETD 2007* (presented at the Electronic Theses and Dissertations, Uppsala, 2007), http://eprints.usq.edu.au/archive/00002653/.

21 Neil Jacobs, “Departmental Thesis Management System development using the Integrated Content Environment (TheOREM-ICE),” 2008, http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/theorem-ice.aspx.

While the ICE-TheOREM work is at this stage a proof of concept it does illustrate an repository based publishing process.



Conclusion: a proposed new model

While ICE provides a 'missing link' in the form of tools for creating web content from word processors, and our work on ORE and repository integration provides packaging and delivery mechanisms to deliver data-supported papers into repositories and journals this technology alone is of little use when expectations are shaped by current practice:

- Institutional Repositories are expected to contain only PDF.
- Publishers ask for word or LaTeX documents which they then process into PDF and, maybe, HTML without any supporting data or complex visualization technologies.

There is no incentive for an author to create anything other than what is demanded of them by publishers, which is typically a source file that cannot be used to create anything but PDF.

Above I have shown some of the progress we are making towards a new scholarly communications process where research publications and data are seamlessly integrated as part of the web, beyond the paper metaphor. Below I outline a model of how a university such as USQ might position itself as a change-agent to drive scholarly publishing towards a more seamlessly web-based model. This approach would build on one of the key strengths of the the established distance education model of long-form course content materials delivered as web-native documents via a core supported IT system (ICE), working on two fronts:

1. To take the proof of concept work from ICE-TheOREM and to do for theses what we did for courseware at USQ; become the first institution in Australia with a mandate for all theses to be made available not just on the web in PDF but *of the web*, in HTML. The most obvious way for candidate to comply would be to use ICE but other toolchains could be used.
2. To make the courseware process more like a research workflow by introducing post publication peer-review for course content, thus turning an established workflow into a publishing model without attempting to change an existing sytem over which we have little influence.

The first of these is a straightforward policy decision which could be made by the university. There is precedent in that electronic submission of theses and deposit in the repository is already mandatory, as is the use of ICE for all new courses, so this would be a simple step, accomplished after a year or two of piloting. This is a natural step in the approach of changing scholarly publishing from below.

The Second involves rather more interaction with the outside world, and needs a little explanation. One of the issues faced by staff is that while they are responsible for writing course material it is not recognized as a contribution to the research literature and is not counted in government reporting as such. Because copyright in course materials is held by the university, and they are managed centrally it is not even clear at times what the authorship of a paper is. A proposed model would work like this:

1. An author creates courseware which is published as open courseware by the university.
2. The author writes a short paper abstracting a module or book from the courseware with an explanation of what the item represents, it might be a literature review, or contain instructional design which is the product of research into previous cohorts of students.
3. The author submits the paper to an existing journal or to a new kind of open access journal, which would be article-centric and arrange for peer review on a rolling basis as articles are submitted – with the output of the journal deposited directly into a repository of papers on pedagogical practice.
4. Reviewers would be able to recommend not only changes to a paper to make it publishable, but to the courseware item itself.

This model for influencing scholarly communications side-steps established incumbents on both sides of the publishing fence by working with new academics on one hand, and on the other a whole class of research which is under-valued. If such a model does succeed then we might be able to provide a positive answer to the question posed by the editors of this special edition of Serials Review about how repositories have influenced publishing models.