

The Continued Importance of XML Production in Scholarly Publishing

Written by Simon Inger & Chris Beckett, Scholarly Information Strategies Ltd. under the commission of SPI Publisher Services, 11147 Air Park Road, Suite 4, Ashland, VA. 23005 USA.

© Scholarly Information Strategies Ltd. 2003.

ABSTRACT: XML-based production offers both those publishers reviewing their production and distribution strategies, and those already using XML the best long term electronic publishing investment strategies. This paper provides a reminder of the value of XML, by summarising the activities and initiatives in the scholarly information chain that rely on structured data tagged for both meaning and presentation.

Introduction

For many years, leading publishers have invested both time and money in the extensive tagging of their research material, as a precursor to online publication. These investments were made, in the belief rather than knowledge, that ultimately they would result in a significant competitive advantage.

At the same time the low cost route to electronic publishing typically involving production of PDF from typesetter files with tagged bibliographic headers (and sometimes tagged references) has remained attractive for many publishers looking for a cost-effective entry into electronic publishing. However, it now appears that the flexibility inherent in a more richly-tagged, XML-based workflow is finally finding business applications in scholarly publishing that justify the additional investment.

Tagging is entering a new era of importance and this, in association with revised production processes designed for end-to-end print and electronic simultaneous publication, will help to maximise the benefits for publishers.

From course packs to PDAs, from archives to content mapping, well-tagged scientific literature is about to enjoy a new future.

Archival in Perpetuity

It has been said that providing a perpetual archive of electronic content is easy. All you have to do is archive the content in any form you like, archive the software necessary to read back the data-archive, archive the machine upon which the software will operate and finally archive the engineer who can fix it all when it goes wrong. Although somewhat flippant, this certainly illustrates the point often ignored about the future of the scholarly archives.

There are two main, alternative approaches to archiving. The first involves archiving the data, the application software, the delivery interface and the necessary machinery. The second seeks to migrate the data to whatever is the current format and delivery environment, thereby severing any dependency upon specific hardware and software configurations, and upon the engineering expertise to maintain it. Creating the data in such a way as to enable this migration is not an insignificant task and depends crucially upon extensive tagging, ideally for meaning as well as for display.

In choosing an archival format, one must consider that the proprietary nature of the PDF format places it at a disadvantage to an XML archive. For PDF to persist in the long term requires that all of its features and format details are placed in the public domain so that future generations of engineers can develop the technology to

render PDF files on computers of the future. On the other hand, XML is a public domain standard, which is therefore easier to migrate, and moreover, since it does not dictate the document appearance, has the advantage of allowing the content to be re-purposed in the future, perhaps in ways yet to be envisaged. This is not yet possible with PDF. For this reason, this makes an XML archive the more reliable and responsible route for a scholarly publisher seeking to meet long term archival needs.

The Digital Preservation Testbed founded by the National Archives and the Ministry of the Interior and Kingdom Relations in the Netherlands reports in its White Paper *XML and Digital Preservation*¹ that perhaps choosing both XML and PDF as preservation standards would help to mitigate the risk of the demise of one of those standards. However it further observes that it would be easier to regenerate PDF from an XML archive than the other way around, illustrating this with the observation that “It is easier to bring a mammoth back to life using its preserved DNA (read XML) than on the basis of a photo (read PDF).”

In the Harvard E-journal Archiving Project study into the feasibility of the creation of a single archival DTD², Inera, the researching body, found that it was possible and desirable to develop a single DTD for the creation of a uniform E-journal archive of full text. The concept recommended by the report is for a DTD to be developed into which publishers would be able to convert their SGML or XML, rather than all publishers having to produce content in the first instance conforming to the archival DTD.

Library Archives

The lack of a durable archive is one of the key reasons cited by librarians for their unwillingness to move to an e-only model. Overcoming this justified concern would allow many publishers ultimately to dispense totally with print products.

Besides the long term archival goal of some very major libraries, there is an additional archival need for many more libraries to provide long term local access to subscribed material. This arises in part because of the current business models adopted by some publishers.

At this time there is no predominant model for e-journal access. Some publishers allow libraries that have cancelled a subscription to the electronic content, to continue to access previous years' content to which they have subscribed. Others, however, require a current subscription in order to access the backfile, changing the subscription purchase model into effectively, a data lease model. Whilst a library remains a subscriber to any given journal, it gains access to all the content to which it ever subscribed in electronic form, but if it cancels its current subscription then access to the back catalogue is also lost. This gives libraries two specific problems to deal with.

Firstly the absence of a local archive of content means that all access is lost to backfiles if a library can no longer afford to subscribe to a title.

Secondly, that same loss and the absence of an archive mean that some universities may have to account for their journals differently. In the world of print journals, the subscription gave rise to the delivery of physical product and hence a tangible asset for the university. In transitioning to electronic content on a data lease model, that asset value is lost and each year's subscription spend no longer services an ever-increasing asset.

The provision of a local archive copy of the subscribed data can alleviate both of these issues, providing continued access and solving an accounting problem too.

Many publishers resist such requests for local content archiving for many diverse reasons. These reasons range from those of content security, to a desire to make sure that library patrons access current content from the publisher's web site, rather than a local version. When a user accesses a local version then the publisher no longer has knowledge of usage levels, and hence cannot compile accurate usage statistics for its content. Furthermore, there may be an impact on advertising revenue if usage of the publisher web site is so diluted.

Such fears are probably, in most cases, overstated since local access is not ideally what most libraries want. One of the key needs that would be met by the local archive is to provide access to back issues lost as a result of the publisher adopting a data lease model. If however, the publisher has already lost the revenue associated with the subscription because of the cancellation then what has the publisher to lose by allowing the local archive?

Allowing libraries to hold their own electronic archive not only resolves the problems mentioned above but also facilitates the transition to an electronic only model. The key point here is that once more, the most effective method of providing the local library archive copy will be in a form that the library may also be able to migrate to future formats. Therefore the same format arguments become true for the local library archive as for the national archives of scholarly content.

There is an alternative strategy to this approach being looked into in the UK at the moment by JISC (the Joint Information Systems Committee), a government-funded body in Higher Education³. However the approach under consideration in the 2003 workshop was one of archiving the permissions to access a publisher's site, rather than the data itself and to the authors of this paper at least, this seems overly complicated to succeed.

Repurposing Content in the Library

The current model of publisher-centric web site collections of content, forces libraries to provide access points to the literature that are counter-intuitive to the reader. The reader approaches the literature predominantly from a subject perspective, which is why libraries use subject classification schemes for organising print content, rather than shelving by publisher. The problems created by publisher-based collections of content is one of the reasons why a growing number of leading-edge libraries are beginning to demand, from publishers, the ability to reorganise and repackage content in ways, and for uses, that make more sense for the institution and their readers and researchers. In his paper at the UKSG conference⁴ David Seaman, Director of the Digital Library Federation (DLF⁵), called upon publishers to make content “malleable, reusable, standardised and reshapeable” giving the libraries the ability to “mix and match content” and to “build course-packs reliably from digital versions” as well as “support multi-formats like e-book, PDA, wireless access and text-to-speech”. These questions are just as much questions of licensing as they are of technology and XML, but assuming that in return for appropriate licensing fees libraries are allowed to use and re-use data in a less encumbered way, then the question of adequate flexibility of the content will then become highly relevant.

Also at UKSG, Mick Archer, Global Information Science and Library Project Manager of AstraZeneca⁶ reported their desire to have unlimited use of research content, be that for local intranet use, promotional use, for reformatting and push delivery to staff and data mining, for an all-inclusive licence fee. One of the concepts being explored at AstraZeneca is going beyond the organisation’s library portal, and instead delivering content of interest to researchers, directly to their own departmental portal. It is permission for this kind of activity that AstraZeneca now seeks from the copyright owners that will allow them to maximise the usefulness of the journals to which they subscribe.

Clearly to support these kinds of initiatives fully, publishers are going to need XML formatted content. PDF will not be a sufficient solution for these kinds of repurposing applications.

Learning Packages

The US Department of Defense came to the realisation some time ago that it was spending unnecessarily large amounts of money on the same or similar learning materials reformed into multiple formats for their staff training needs.

According to Seybold Reports⁷ “For the past four years the DoD has been working with hundreds of academic institutions, standards bodies, corporations and private organizations. Its goal: to develop an open-architecture

standard for a global distributed network of interoperable and reusable learning content.”

Through research of its needs the SCORM standard was formed. SCORM is an acronym for the Sharable Courseware Object Reference Model and a useful summary of it is available from Randall House Associates, Inc.⁸

SCORM describes the framework for developing Web-based instructional materials in a way that will allow a global e-learning community to use the materials with a variety of systems. SCORM compliant material meets the so-called “GRIN test”: granular, reusable, interoperable and networkable. This requires that the content be independent of proprietary systems and software and allows it to be usable in many different learning scenarios.

SCORM compliance is a pre-requisite of selling courseware to US Government and therefore the inclusion of research articles within course-packs also requires SCORM compliance.

However its take-up is limited in the scholarly arena right now, since only a relatively small proportion of academic literature is destined for such course-packs.

PDA Delivery

Highwire⁹ is already supporting a number of initiatives involving the delivery of selected biomedical content to users of Palm PDAs. Bernhard Hecker, Journals Manager at Highwire, reported at the recent 2003 CSE conference¹⁰ that their Highwire Remote product is a custom application for handheld devices and Highwire repurposes selected content for this delivery medium. The application re-renders the content as a normal PDA document, not in the article’s original format.

PDA delivery continues to grow strongly in certain sectors, especially for practising physicians and Kent Anderson, Publishing Director at the New England Journal of Medicine (NEJM¹¹) reported at the same conference that the NEJM now has over 17,000 users reading full-text content on PDAs. While this still represents a small fraction of the overall user base of NEJM products, it is significant in that it continues to grow steadily with no promotion whatsoever.

It seems unlikely that PDAs will ever be used to read PDF files onscreen, and so access to this form of content delivery is thus limited to those who work within an XML environment.

Portals

The delivery of scholarly content into community-based portals is not new. Medscape¹² and CTSnet¹³, amongst others both include a variety of primary research content from primary publishers.

As portals continue to proliferate it will become increasingly important that content is available for repurposing within them.

MedBiquitous, which is the organisation that operates the CTSnet portal (the Cardiothoracic Surgery Network) is a membership body dedicated to the creation of collaborative technologies to support medical education.¹⁴ Its involvement extends from the definition of XML schemas for the interoperability of medical content through to the development and use of web services technologies to share applications across multiple web sites.

Most of the existing and successful portals tend to be built around biomedical content, partly due to the levels of sponsorship and advertising that are available in the medical arena and consequently publishers of biomedical content have most to gain from being able to repurpose their content easily and to interoperate with the portals.

The Semantic Web

There is much research being undertaken today about how to link web documents not just by the hyperlinks contained within them, but also by the meaning of the documents themselves. Search engines already go some way down this route with the “find more pages like this” features which are quite commonplace. But additional tagging in documents for meaning, topics and context will start to allow for more advanced content mapping engines to link documents together in wholly new ways.

Tim Berners-Lee first introduced this concept in his 1999 book *Weaving the Web: The Original Design and the Ultimate Destiny of the World Wide Web*¹⁵ and updated in *Scientific American* in 2001¹⁶

There are two important technologies in place to support the Semantic Web: eXtensible Markup Language (XML) and the Resource Description Framework (RDF). XML is currently best used to describe layout through an interpretation of the tagging used (for example an article title is rendered in a certain way on screen or in print), but since those tags are not standardised from publisher to publisher, they cannot be relied upon to say anything about what the data within them mean. However RDF tags, which can be encoded within XML, are a standardised way of structuring the meaning of a document such that computers can then interpret and link like objects across the web.

It appears to be universally accepted that the semantic web is both desirable and inevitable, as it will allow users to navigate the web in more intelligent ways than currently possible. Onward links from a web page will be dynamically and automatically changed as other similar pages on the web are created. Ossenbruggen discusses the current state of play in his paper

Hypermedia and the Semantic Web and sets out an agenda for further research.¹⁷

Semantic tagging also has relevance inside a single document such as a scholarly article. In a talk at the PSP Annual Conference 2003, Thane Kerner from Silverchair¹⁸ explained how careful use of semantic tags can aid navigation within a single document.

In addition, products like Themescape¹⁹ can be used to search through semantic tags across a number of documents to create a topic map, showing the relationships between a multitude of documents on a similar subject.

Such XML-dependent technologies are in their infancy, but begin to show the future of navigation through scholarly and other information.

The Arrival of STIX Fonts

Since 1997, a number of leading science publishers – American Chemical Society, American Institute of Physics, American Mathematical Society, American Physical Society, Elsevier and the Institute of Electrical and Electronic Engineers – have been working together to design, fund and manage a project to define a set of fonts that will cover the entire needs of the scientific publishing community. These Unicode STIX fonts (Scientific and Technical Information Exchange fonts²⁰) will become freely available to anyone and will greatly enhance the viewing of technical content within HTML, removing the need for the now common practice of creating images of formulae instead of rendering them through HTML.

According to Jennifer Ann Hutt, in her article in *Science Editor*²¹ the font set will be ready by the fall of 2003. These fonts should further enhance the quality of HTML delivery and in so doing further justify an investment in XML and enhance all of the other applications for XML, from archiving to repurposing for other devices.

Accessibility Issues

Many publishers around the world seem to remain unaware of the legislation in force in the USA and many European Union countries governing the accessibility of their web sites to users with varying levels of disabilities. The main elements of these legislative acts cover those with visual impairment.

One of the key features of accessibility compliance is the ability for the reader to be able to alter the point size used on screen to make reading easier. In addition, good descriptive alternates to graphics are essential for those with impaired vision. Within HTML this is achieved with a descriptive <ALT> tag associated with each graphic. Such tags can also be read by text-to-speech systems used by the visually impaired.

One of the publishers that continues to make their research content accessible is the Institute of Physics Publishing²².

While accessibility features are easier to implement appropriately within an XML environment, Adobe Acrobat also has downloadable accessibility plug-ins.²³

Production Methodology

Many publishers are now benefiting from an improved XML workflow that starts at the very front end of their production process, rather than a post-production conversion of Word or PDF files into XML.

By moving to an XML workflow straight from author submission, publishers can analyse and parse article submissions for errors at an earlier stage. This can include the quality control of cross-references, index terms, other internal linking or indeed spotting missing elements. In addition, by creating metadata for articles earlier in the production process, it becomes easier to support many of the pre-print initiatives that publishers are beginning to experiment with. Moreover there is growing evidence that this creates a cost saving opportunity over traditional workflows.

Bruce Rosenblum, CEO of Inera Inc., concluded his presentation at the Seybold Seminars in New York, 2002²⁴ by saying that XML workflows lead to lower costs, higher quality and faster production.

Although many publishers have invested in true end-to-end XML production, some are still guilty of making handcrafted changes to the final HTML that is output before loading onto their web pages. Sadly by making those changes they lose a considerable amount of the value of their XML archive, since the final online version can no longer be reproduced from the source XML. For the XML production investment to make economic sense and to allow a publisher to take advantage of all of the initiatives underway and summarised in this paper, one must be in a position to create the final online article automatically from the XML without any human intervention and with 100% quality assurance.

The careful development of an appropriate DTD that can be successfully used right through the production process appears to be key.

Conclusions and Recommendations

The benefits of having a quality XML archive of scholarly content are many. It allows publishers to more fully comply with major archival initiatives and generate additional revenues by repurposing content for alternative delivery platforms, by delivering content to portals and by making their documents more highly visible to readers on the web.

At the same time publishers need to guard against the hand crafting of corrections to HTML on their web

sites which invalidates the XML versions of their content as a true record for archival purposes, or indeed as a reliable source for others to repurpose their content into portals, course packs or other media.

Above all else there are two key factors in publishers' preparedness for all of the opportunities that this paper describes. Firstly, that a publisher's typesetter or online service provider creates a format of XML suitable for both long-term archival use to agreed standards and for repurposing into other formats. (If XML is created at any point in the production process, then it should be the XML that is archived, not just the HTML or PDF versions of it.) Secondly, as a safeguard, the publisher should hold a copy of all of its XML data archive or obtain a guarantee from whoever is holding that archive that it can be supplied, at little or no cost, back to the publisher for the fulfilment of any of the opportunities described in this paper.

About the Authors

Simon Inger and Chris Beckett are both directors of Scholarly Information Strategies Ltd., a specialist consultancy in scholarly publishing based in the UK. Both Simon and Chris were formerly directors of CatchWord, the world's largest scholarly journal hosting services company prior to its sale to Ingenta plc in February 2001. Scholarly Information Strategies Ltd provides publishers and intermediaries with consultancy services focused on electronic publishing strategy, product assessment and audit; competitor analysis; outsourcing options; and sales and marketing strategies. See www.scholinfo.com

References

- ¹ Digital Preservation Testbed White Paper, *XML and Digital Preservation*, Den Haag, September 2002 at http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf
- ² *E-Journal Archival DTD Feasibility Study* by Inera Inc. for the Harvard University E-Journal Archiving Project at <http://www.diglib.org/preserve/hadtdfs.pdf>
- ³ Report on the JISC e-journals Licensing & Archiving Workshop held on 17th February 2003 http://www.jisc.ac.uk/index.cfm?name=pres_ejournals_report
- ⁴ David Seaman, Director of the Digital Library Foundation, UKSG Annual Conference, Edinburgh 2003, **From Isolation to Integration** http://www.diglib.org/presentations/uksg2003_files/frame.htm#slide0001.htm
- ⁵ Digital Library Federation <http://www.diglib.org/>
- ⁶ Mick Archer, AstraZeneca R&D, Charnwood, UK at the UKSG Annual Conference, Edinburgh 2003, **E-only in the Corporate Sector**

-
- ⁷ Mike Letts, **ADL and SCORM: Creating a Standard Model For Publishing Courseware**, *Seybold Reports*, Vol 2 No. 1 April 8th 2002
- ⁸ Randall House Associates, Inc, **SCORM**
<http://www.rhassociates.com/scorm.htm>
- ⁹ Highwire PDA delivery section is to be found at—
<http://www.highwire.org/customize/#myalerts>
- ¹⁰ Bernhard Hecker, Journals Manager, Highwire, at the Council of Science Editors Annual Meeting 2003 in Pittsburgh, PA, **The Wireless Frontier: PDA Delivery of Scientific Content**
- ¹¹ Kent Anderson, Publishing Director, NEJM, at the Council of Science Editors Annual Meeting 2003 in Pittsburgh, PA, **Handholding the Handhelds**
- ¹² Medscape is at <http://www.medscape.com/>
- ¹³ CTSnet, the Cardiothoracic Surgery Network is at <http://www.ctsnet.org/>
- ¹⁴ MedBiquitous, *Collaborative Technologies for Medical Education*, at http://www.medbiq.org/about_us/medbiq_whitepapers/whitepaper.pdf
- ¹⁵ Berners-Lee, T. and Fischetti, M. *Weaving the Web: The Original Design and the Ultimate Destiny of the World Wide Web*. London: Orion, 1999.
- ¹⁶ Berners-Lee, T., Hendler, J. and Lassila, O. **The Semantic Web** in *Scientific American* May 17th 2001.
- ¹⁷ Jacco van Ossenbruggen, Lynda Hardman, and Lloyd Rutledge of National Research Institute for Mathematics and Computer Science, Holland, **Hypermedia and the Semantic Web: A Research Agenda**
<http://jodi.ecs.soton.ac.uk/Articles/v03/i01/VanOssenbruggen/>
- ¹⁸ Silverchair. <http://www.silverchair.com>
- ¹⁹ Themescape presentation at the MicroPatent site at <http://www.micropat.com/0/pdf/themescape.pdf>
- ²⁰ STIX fonts at <http://www.stixfonts.org/>
- ²¹ Jennifer Ann Hutt, **New Comprehensive Font Set to Serve Sciences** in *Science Editor* March-April 2003, Vol 26 No 2 p49.
- ²² Institute of Physics Publishing <http://www.iopp.org/>
- ²³ Acrobat accessibility plug-ins information and downloads at <http://www.adobe.com/products/acrobat/solutionsacc.html>
- ²⁴ Bruce D. Rosenblum, CEO, Inera Incorporated, Seybold Seminars, **XML in Journal Publishing Today** at, http://seminars.seyboldreports.com/2002_new_york/files/presentations/082/rosenblum_bruce.ppt