Learned Publishing, 22, 85–94
doi:10.1087/2009202

Semantic publishing: the coming revolution in scientific journal publishing

David SHOTTON

© David Shotton 2009

ABSTRACT. Recent developments in Web technology can be used for semantic enhancement of scholarly journal articles, by aiding publication of data and metadata and providing 'lively' interactive access to content. Such semantic enhancements are already being undertaken by leading STM publishers, and automated text processing will help these enhancements become affordable and routine. Publisher, editor, and author all have primary roles in that process; an incremental approach is needed. Publication of data and metadata to the Web make possible added-value 'ecosystem services'; semantic publishing will bring substantial benefits to scholarly communication.

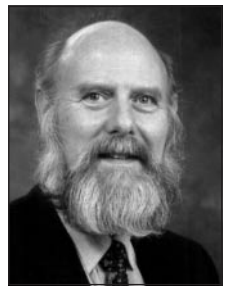
Introduction

Effective communication of research results is central to efficiency within the scientific research lifecycle of hypothesis formulation, experimentation, interpretation, and publication, since the findings of one academic research project inform the hypotheses developed for the next. As the primary dissemination channel and public record of new research results, journal publication is a vital ingredient of the scholarly workflow, and its key commodity, the original research article, is of primary importance out of all proportion to its intrinsic worth.

Scholarly journal publishing has undergone a digital revolution over the past decade, with massive uptake of online provision. However, the fundamental structure of the research article has remained relatively unaltered by this digital revolution, and online academic publishing has yet to realize the potential offered by the World Wide Web.

The Web has in recent years shaped revolutionary change within the scientific community, as a truly disruptive technology. New ways of digital communication between researchers (preprint archives, blogs, videoconferences, etc.) now challenge inherited assumptions about the roles of key players (researchers, funders, publishers, librarians, etc.) in the scholarly publication cycle. However, the scope of this paper is limited to consideration of the peer-reviewed journal article itself, and the semantic enhancements to it that are already finding their way into publishing practice.

Web documents are primarily designed to be viewed by people. However, from its outset, Tim Berners-Lee, the Web's inventor, had a vision for it as a 'Web of Data', structured in such ways that computers could assist humans in the ever-expanding tasks of



David Shotton

information discovery, digital data integration and knowledge management. Despite considerable initial overhyping, this 'Semantic Web'[1,2] (http://www.w3.org/2001/sw/) is now starting to deliver real benefits.[3] It does not require complex artificial intelligence to interpret human ideas, but 'relies solely on the machine's ability to solve well-defined problems by performing well-defined operations on well-defined data'.[2] As Berners-Lee prophetically stated in 2001, 'The semantic web will profoundly change the very nature of how scientific knowledge is produced and shared, in ways that we can now barely imagine.' These benefits are now being made possible by providing machine-readable metadata for journal publications and other data sources, using agreed semantic web standards that permit computers to assist in the tasks of information discovery and integration.

semantic enhancements increase the intrinsic value of journal articles, by increasing the ease by which information, understanding and knowledge can be extracted

However, while the use of semantic web technologies can certainly contribute importantly to semantic publishing, and were used in the work described below, semantic publishing itself is much broader, and includes, for example, more effective use of that most basic and mundane of all Web features, the hyperlink, and the intelligent use of interactivity by exploiting the capabilities of Javascript.

In the present context, I define 'semantic publishing' as anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers. Among other things, it involves enriching the article with appropriate metadata that are amenable to automated processing and analysis, allowing enhanced verifiability of published information and providing the capacity for automated discovery and summarization. These semantic enhancements increase the intrinsic value of journal articles, by increasing the ease by which information, understanding and knowledge can be extracted. They also enable the development of secondary services that can integrate information between such enhanced articles, providing additional business opportunities for the publishers involved. Equally importantly, readers benefit from more rapid, more convenient and more complete access to reliable information.

The current state of online journal publishing

Advantages

For the publisher, online electronic manuscript submission obviates the need for traditional typesetting from printed manuscripts; if the journal is entirely online, there is no need for printing and physical distribution either. Peer reviewing is accelerated because postal delays are taken out of the equation, and electronic publication permits faster production, cheaper publication costs, and easier distribution. It also enables commercial alternatives to single journal subscriptions, including 'big deals', pay-per-view, and virtual journals, and provides the opportunity to create added-value search, browse, and summarization services. For authors, the benefits of online manuscript submission and revision are significant, while the easing of page limits for purely online journals, and the possibility of publishing supplementary information, means that additional data can be published that would formerly not have seen the light of day. Readers benefit from the increased ease of access to online journals, revolutionizing reading habits that no longer have to be synchronized with the opening hours of the print library. Efficient alerting and search services obviate the need to scan journal contents pages, and reference linking via DOIs (Digital Object Identifiers; http://www.doi.org/) assists exploration and knowledge acquisition.

Disadvantages

However, the benefits of online publication come at a cost. Publishers need competent ICT staff to keep abreast of rapidly developing technology, placing small publishing houses at a competitive disadvantage, while managing such change at any scale in what traditionally has been a fairly conservative industry can be uncomfortable. The effect on authors is surprising: despite easy online

access to a greater variety of journals, they have developed tunnel vision, citing articles that are less diverse and more recently published than was the case in the days of purely paper-based journals, since they are undertaking less journal browsing that was formerly necessary for knowledge discovery.[4] Readers face a different problem, since they have access to vastly more journals than ever before, which contributes to the ever-increasing data deluge with which they have to cope, while librarians, finding their traditional roles being progressively bypassed by online resources, are having to reinvent themselves as the knowledge navigators of the digital age.

Shortcomings

PDF

Online versions of journal articles are either presented as conventional HTML Web pages, generally with little in the way of semantic mark-up, or are delivered as PDF files. As the electronic embodiment of the printed page, the PDF document is both familiar and easy for humans to read, but it is antithetical to the spirit of the Web, being static rather than interactive, and difficult for machines to read, thus inhibiting the development of services that can link information between articles.

DOIs

Articles are increasingly being assigned DOIs, but this convenient standard is not being fully exploited to provide live links from online citing articles to the cited articles in their reference lists, despite the fact that anything that increases click-through to other articles reciprocally enhances journal usage. More generally, insufficient use is made of hypertext links to provide readers with access to other useful information sources.

Supplementary material

The ability to publish supplementary material with a journal article is of undeniable benefit. However, the lack of standards between journals, the less-than-rigorous quality control, and the inability of Web search engines to access and index much supplementary material, mean that its use is suboptimal, often being of value only in the context of the article itself.

Mark-up, data, and metadata

With a few shining exceptions, online journals currently provide no semantic mark-up of text that would facilitate increased understanding of the underlying meaning. Perhaps their most significant shortcomings are that there is little or no access to the data contained within them in actionable form, or to metadata describing the articles themselves. So much more could be done: while simple metadata may facilitate discovery of an article, suitable semantic mark-up of results within an article could enable automated meta-research, i.e. the undertaking of truly novel science.

the PDF document is both familiar and easy for humans to read, but it is antithetical to the spirit of the Web

The feasibility of routine semantic publishing

What publishers are already doing

Several publishers and other stakeholders have already taken the lead in showing what can be done:

Downloadable XML

Some publishers, such as the Public Library of Science (PLoS), already make their publications available as downloadable XML, in addition to PDF.

Downloadable datasets

Perhaps the clearest example of a publisher providing access to raw numerical data, by permitting downloads of spreadsheets, is given not by an STM publisher but by *SourceOECD* (http://www.sourceoecd.org/), the On-line Library of Statistical Databases, Books and Periodicals of the Organisation for Economic Co-operation and Development (OECD; http://www.oecd.org/). Not only do their online statistical tables have an Export Excel tab that creates and downloads an Excel spreadsheet from the data currently being observed, but, since April 2008, OECD.Stat also includes the ability to pro-

duce dynamic graphics under the control of the user, which can bring the data alive visually.

Working with images, audio, and video

The *New England Journal of Medicine* (*NEJM*) has a highly interactive website (http://content.nejm.org/) showcasing new articles, but also providing a variety of audio and video publications, an image challenge (http://image-challenge.nejm.org/) that challenges readers to make correct medical diagnoses from displayed images, and 'drag-and-drop' images that make it easier for readers to create PowerPoint presentations from published journal images.

Structured Digital Abstracts are machine-readable summaries of the key data and conclusions of articles

The tabbed article

On the *NEJM*'s beta site (http://beta.nejm.org/), the publisher is 'pursuing new ideas in publishing and showcasing innovative ways to present information for use in medical education, research, and clinical practice'. Here a new tabbed interface is provided to the journal article, in which different parts of the article can be selected using tabs, thereby attempting to break away from the traditional linear narrative. Disappointingly, these tabs are presently limited to 'Abstract', 'Article' and 'References', with no breakdown of the article into Introduction, Methods, Results, etc., and these tabs scroll out of sight when reading down the page. The equivalent sectional functionality provided by the non-scrolling link set at the top of the enhanced *PLoS Neglected Tropical Diseases* article (http://dx.doi.org/10.1371/journal.pntd.0000228.x001) described by Shotton *et al.*,[5] and discussed below, provides much better functionality while occupying less screen real-estate.

Reference management

The Nature Publishing Group provides a free online reference management tool for clinicians and scientists, *Connotea* (http://www.connotea.org/about), that enables one quickly and easily to save, organize, tag, share, and download bibliographic references. This has the particular advantage of auto-completion of the bibliographic record if given a DOI, or if the item is from a list of major publishers.

Structured Digital Abstracts

Many journals now break up their normal Abstracts into sections, typically 'Background', 'Methodology/Principal findings', and 'Conclusions/Significance'. However, Structured Digital Abstracts (SDAs) are something quite different; they are machine-readable summaries of the key data and conclusions of articles. Following papers by Seringhaus and Gerstein,[6] Gerstein *et al.*,[7] and Seringhaus *et al.*[8] concerning their desirability, the editors of *FEBS Letters*, in collaboration with their authors and the curators of the MINT Database (http://mint.bio.uniroma2.it/) that records details of protein–protein interactions extracted from published papers by expert curators, recently implemented SDAs for *FEBS Letters* papers describing protein–protein interactions. These SDAs are XML-encoded summaries appended to the articles' conventional abstracts, detailing the names of the interacting proteins, with unique protein identifiers and links to MINT and the Universal Protein Resource Uniprot (http://www.uniprot.org/), and the types of protein–protein interaction involved, defined from the HUPO Proteomics Standards Initiative's Molecular Interaction (MI) Controlled Vocabulary (http://www.psidev.info/index.php?q=node/31). Between April and December 2008, 63 papers have so far been published in the *FEBS Letters* with SDAs. The *FEBS Letters* SDA is based on the MIMIx standard (Minimum Information required for reporting a Molecular Interaction experiment; http://www.mibbi.org/index.php/Projects/MIMIx), developed by experts in the protein interaction field, one of several minimal information ontologies that have recently been developed as part of the MIBBI Project (Minimal Information for Biological and Biomedical Investigations; http://www. mibbi.org/).[9] If other journals are to develop SDAs specific to their domains of interest, these will need to be based on similar minimum information standards, to ensure interoperability.

Semantic mark-up of text

The Royal Society of Chemistry's journals, for example any recent paper from *Molecular Biosystems* (try http://dx.doi.org/doi:10.1039/b613673g), offer an Enhanced HTML version in which a user toolbox is provided as a semitransparent overlay at the top right-hand corner of the article. Among the tools presented there is the ability to highlight the text with terms from the IUPAC Compendium of Chemical Terminology – the 'Gold Book', from a dictionary of chemical compounds, and from a number of relevant ontologies: the Gene Ontology (GO), the Sequence Ontology and the Cell Type Ontology. When the reader clicks on any one of these highlighted terms, he or she is taken to a separate page providing relevant information and links. For a GO term, this includes the definition of the term, its GO ID number, a list of synonyms, and a list of other RSC articles referencing this term, while for a chemical name it includes a structural diagram of the chemical, a list of synonyms, its IUPAC International Chemical Identifier (InChI code; http://old.iupac.org/inchi/), a downloadable XML file describing the chemical using the Chemical Mark-up Language (http://cml.sourceforge.net/), and links to search for this chemical in *PubChem* and *SureChem* patents. This development, known as the *RSC Project Prospect*, won the 2007 ALPSP/Charlesworth Award for Publishing Innovation, and is thought to be the first major application of semantic web technologies in science publishing.

Acknowledging the need to do more

Perhaps the clearest acknowledgement that STM publishers need to do more in terms of enriching their current journal offerings, to utilize more fully the great potential of the Web, came recently from Elsevier, who launched two competitions, Article 2.0 and the Elsevier Grand Challenge. The Article 2.0 competition (http://article20.elsevier.com), for total prize money of $4,000 (first prize $2,500), provided access to approximately 7,500 full-text XML scientific articles (including images) and challenged contestants, in very general terms, to use these to 'demonstrate how scientific research articles should be presented on the Web to meet their needs', including integrating the article into existing applications or combining it with other web service APIs. The Elsevier Grand Challenge: Knowledge Enhancement in the Life Sciences (http://www.elsevier grandchallenge.com/) was similar but grander, as the name implies, and was described as a contest 'to improve the way scientific information is communicated and used'. Here the total prize money was $50,000 (first prize $35,000) and competitors, given open access to Elsevier's life science journal corpus and databases, were required to submit a demonstration and a paper describing and prototyping a tool 'to improve the interpretation and identification of meaning' in online journals and text databases relating to the life sciences. In both cases, Elsevier sought to obtain ideas for semantic enhancement from the larger community that they could then use to enhance the value of their publications, requiring from participants the right of first refusal for the exclusive commercial development of any finalist's submissions. That Elsevier chose to invest in these contests is witness to the fact that we are in the opening phases of a scientific publication revolution.

Elsevier sought to obtain ideas for semantic enhancement from the larger community

Reflections on our own work of semantic enhancement

In Shotton *et al.*[5] we describe the semantic enhancements that we have made to a recent article by Reis *et al.*[10] from the journal *PLoS Neglected Tropical Diseases* (*PLoS NTD*), to demonstrate the range of possibilities that publishers should be considering. The semantically enhanced article can be found at http://dx.doi.org/10.1371/journal.pntd.0000228.x001, with online technical descriptions provided at http://dx.doi.org/10.1371/journal.pntd. 0000228.x009. These semantic enhancements include live DOIs and hyperlinks, semantic mark-up of textual terms with links to further information, interactive figures, a reorderable reference list, and two novel types of semantic enrichment: Citations in Context, using a Supporting Claims ToolTip, and Tag Trees. We created a document summary containing document statistics; a study summary, a tag

cloud and tag trees of the marked-up named entities; and a numerical analysis of the citations within the article. The study summary itself, which is the human-readable equivalent of a structured digital abstract, specifies the disease studied, its causative agent, the purpose of the study, and the number of human subjects involved; the indicator of prior infection and the assay used to detect it; the dates of the study and the name and location of the study site; and the study's principal findings. In addition, we published downloadable spreadsheets containing data from within the article, enriched these with provenance information, and demonstrated various types of data fusion (mashups) with information from other research articles and with maps. We also published machine-readable metadata both about the article and about the references it cites, for which we developed a Citation Typing Ontology, CiTO (http://purl.org/net/cito).

the real questions arising are whether the added value achieved was worth the effort invested

These enhancements required about 10 person-weeks of effort, most of which was taken in understanding, deciding, and prototyping exactly what to do and how best to do it, since this was a new area of endeavour for us. For the publisher, the real questions arising from this work are whether the added value achieved was worth the effort invested, and which if any of these enhancements could be brought into mainstream STM journal publishing in an affordable manner. While we, for the purpose of this demonstration, undertook this work manually and post-publication, one key to answering these questions is to consider which enhancements could be provided by publishers and editors, which by authors, and which could be automated.

Publishers' roles in semantic publishing: freeing data

The Brussels Declaration on STM Publishing (http://www.stm-assoc.org/brussels-declaration/), published in November 2007 by the international scientific, technical and medical (STM) publishing community, states:

> STM publishers are committed to change and innovation that will make science more effective, and believe that raw research data and datasets submitted with a paper to a journal should, wherever possible, be made freely accessible to other scholars.

In subscribing to this, STM publishers have already aligned themselves with the aims of semantic publishing, and are seeking ways to implement these commitments in an affordable manner. Lest some subscription-access publishers be anxious about giving away information associated with their published articles, it may be instructive to look at three examples where giving away data and metadata has brought financial benefit.

- The first is the most obvious. Many STM publishers, particularly in the life sciences, have for years been making the bibliographic 'header' information and the abstract of each published article freely available to services such as PubMed, bringing incalculable benefits in terms of online search services allowing readers more easily to find the articles in which they are interested. It is impossible now to imagine an age when that was not so.
- The second is presented by Amazon (http://www.amazon.com), which has an open API and makes freely available all its book metadata, thereby gaining substantial revenue on subsequent click-throughs from third-party services to purchase books.
- The third is described in the influential report of Weiss,[11] who contrasts the fundamental differences in the policy and funding models for public sector information – exemplified by meteorological data – in the USA as compared to Europe, a region of comparable size and governmental investment in weather forecasting. The USA embraces a policy of unrestricted free access to such public sector information, while meteorological agencies in Europe treat their information holdings as commodities used to generate short-term revenue by their sale. The result of the US 'open access' model has been the rapid growth of secondary information services unfettered by copyright or other restrictions, that has not occurred in Europe due to the restrictive government 'subscrip-

tion access' information practices. As a result, European information service providers are increasingly frustrated at the competitive advantages enjoyed by their American counterparts, while the US Treasury accrues taxes from the secondary publishing and service activities that far exceed any revenues generated in Europe through 'cost recovery' sales of meteorological information.

So what data should the publishers be making freely available? Clearly they should provide the datasets that underlie the figures and tables in their articles, and machine-readable provenance information about the article itself. But machine-readable reference lists should also be made available, so that citation networks can be created, analysed, and used to promote reader traffic to both citing and cited articles, to the benefit of the publishers concerned. Furthermore, publishers already have extensive sectional mark-up for their articles within the XML created during the publication process, many using a recognized *de facto* international standard, the National Library of Medicine's XML document format. It would be hugely advantageous if this information was also made available online, rather than being discarded upon creation of PDF versions of the articles.

Fortunately, with the introduction of the new ACAP open standard (Automated Content Access Protocol; http://www.the-acap.org/) that enables publishers to express terms under which automated access to website content can be regulated, and the increasing employment of Creative Commons licenses regulating rights for reuse, publishers have the means to specify clearly which data are to be made freely available. The open question of who should host the data published to the Web in this manner – whether publishers should each host the datasets relevant to their own publications, or whether there should be independent data repositories, equivalent to SourceForge (http://sourceforge.net) for open access software – will be decided in practice, but this is a secondary concern. The important thing is to get the relevant data onto the Web, no matter where they are hosted!

Editors' roles in semantic publishing: using domain expertise

The Royal Society of Chemistry's prize-winning Project Prospect relies heavily upon the domain expertise of its editors. Its Editorial Production Systems Manager, Richard Kidd, said 'To a great extent success is due to our technical editors, who have developed new skills to judge the meaning and context of terms – as they're experienced highly qualified chemical scientists, we feel that this is a great application of their skills and knowledge.' It is perhaps a tribute to this 'coming home' of editors to their academic roots that the 2007 impact factors for RSC Journals showed significant increases, reflecting in part the fact that the added value of these semantic enhancements is increasing the attractiveness of the RSC journals to both contributors and readers. How widely could this model of intensive editor involvement in semantic enrichment be spread to other journals?

Authors' roles in semantic publishing: imparting tacit knowledge

Authors know better than anyone else their domain of discourse, and the position of their article within it. Only the authors really know why they cite particular papers in their articles in preference to others, and the nature of both the citations and the cited articles. If that tacit knowledge could be captured using an online reference annotation tool such as Connotea, a modification to EndNote, or a new plug-in for Word, using terms from the Citation Type Ontology, the work of developing reference lists would essentially be done.

Microsoft has already developed a plug-in for Word 2007 (http://tinyurl.com/5szjly) to permit reading and writing of XML-based documents that comply with the National Library of Medicine's XML document format shared by many publishers and by PubMed Central. Additionally, Microsoft has just released a second plug-in permitting ontology-based semantic mark-up of named entities (http://tinyurl.com/abc4c7). Routine use of this would enable authors to add structural mark-up to articles with minimal effort.

only the authors really know why they cite particular papers in their articles

Furthermore, authors keep their raw data in spreadsheets, and could, if requested to do so by the journal, easily submit these with their manuscripts for Web publication. Thus, given the correct tools and incentives, authors could create much of the semantic metadata required during the course of article writing, with marginal additional effort.

Automated entity recognition and links to ontologies: the role of text mining

To enable mark-up of content to scale cost-effectively across the publishing world, it would be necessary to automate it. We experimented with the use of two automated systems to identify semantically meaningful terms in one article, Reis *et al.*[10] Our trial of the uBio Taxon Finder Web service for finding taxonomic names (http://www.ubio.org/index.php?pagename=xml_services) was highly successful. It found every instance of taxonomic names in the *PLoS NTD* text ('*Leptospira*', '*Leptospira borgspetersenii*', '*Leptospira interrogans*', '*Leptospira kirschneri*', '*Rattus norvegicus*'), but also returned one false positive, 'Strina', the name of one of the authors of Reference [52] in that article, showing that some human supervision is required. The uBio findIT algorithm did even better, also finding abbreviated forms (e.g. '*L. interrogans* serovar Autumnalis'), but with more false positives. Using the free Reuters OpenCalais automated text mining service (http://www.opencalais.com/) was less successful – while this was excellent at recognizing persons, place names, and institutions, as might be expected, it performed poorly for domain-specific biological terms, either failing to recognize them or classifying them incorrectly: for example, 'antibodies' were classified as a form of technology.

in the face of the digital data deluge, readers need all the help they can get

Other more sophisticated text mining and natural language processing tools are currently being developed to recognize textual instances and link them automatically to domain-specific ontologies. For example, the open source application AKTive Media (http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html) will enable other occasions of instances, identified by the user, of terms from one or more specified ontologies or controlled vocabularies to be automatically marked up across a whole corpus of documents, doing so in a provisional manner that then permits the author or editor to confirm or cancel each assignment. This may prove more difficult in some fields where nomenclature is ambiguous (e.g. genetics) than in others.

Our own experience in marking up the *PLoS NTD* article clearly showed the requirement for human intervention. For example, we wished to record 'slums' and 'slum environments' as types of habitat in which leptospirosis was likely to occur. However, blindly marking up every occurrence of phrases in which the word 'slum' appeared was not appropriate, since a 'slum dweller' is clearly a person, not a habitat. To guide our mark-up, we developed a set of simple heuristics (available at http://dx.doi.org/10.1371/journal.pntd.0000228.x010) that may be of assistance to others undertaking similar work.

Improving the effectiveness of automated term recognition and ontology linking will involve further research at a frontier of information management, between semantic technologies to provide controlled vocabularies that can be reasoned over, and text mining technologies to identify concepts for semantic annotation. However, many experts are confident that performance in this area will increase dramatically over the next few years, and will be routine in ten years time, albeit with the requirement for human editorial intervention if one is to avoid silly mistakes.

The development of 'ecosystem services'

In the face of the digital data deluge, readers need all the help they can get in interpreting the information contained within individual scholarly articles, in following leads from one article to others it cites, and in understanding the whole 'ecosystem' of publications in a particular domain of knowledge. Once basic metadata concerning articles are published to the Web, semantic technologies permit the development of added-value 'ecosystem services', above the level of the individual article, that facilitate this understanding. Value is created when information from two or more articles is integrated to permit

meta-analysis, or fused with data from other sources (e.g. maps). Our simple demonstrations of the potential of such data fusion services for infectious disease epidemiology, in the context of semantic enhancements to Reis *et al.*,[10] described in Shotton *et al.*,[5] and the more sophisticated work of the Royal Society of Chemistry cited above, illustrate how investment of effort at the time of publication can pay back in terms of significantly increased efficiency when data from many different reports need to be integrated to enable meta-analysis and mathematical modelling, since currently research efficiency is severely limited by the cost in terms of time and effort of acquiring such data manually from the published literature.

Conclusions

The research challenge and the need for collaboration

The whole area of semantic publishing seems ripe for pre-competitive multidisciplinary collaborative research and development, of the sort that individual research groups and publishing companies may have neither the capacity nor the skills to undertake individually. In particular, there is scope to develop and agree upon simple services and standards that can be widely adopted, along the lines of the Citation Typing Ontology, CiTO (http://purl.org.net/cito).[5] The best precedent for this is the inter-publisher collaboration, facilitated by CrossRef and the International DOI Foundation, that led to the widespread uptake of DOIs.

Principles for semantic publishing

While publishers may be anxious about the commitment involved in full-blown semantic enrichment of journal articles, it is important to emphasize the benefits of an agile incremental approach, and the value of quick wins, such as the inclusion of DOIs that actually link out to the referenced article, and of publishing downloadable datasets in the form of Excel spreadsheets rather than images. Starting experimentally with a single journal is almost certainly wiser than waiting until technology makes it easy to enhance all one's journals, since technologies will continue to evolve and experience gained will prove invaluable. Interaction with academic editors and research contributors will be essential to ensure the enhancements under consideration meet real user requirements and relieve existing pain points. Just getting data out there onto the Web is a primary virtue, and the essential prerequisite for the development of the ecosystem services discussed above. Available standards should be adopted where possible, but if no pre-existing standards exist, it will pay just to do whatever seems most sensible and wait for standards to emerge. Technological advances will make tasks progressively easier, particularly if open standards and open source tools are used, since the collaborative nature of their development leads to rapid progress once adopted. Finally, the next decade will see raw text decreasing in value, with a corresponding increase in the value of semantic services that will help readers to find actionable data, interpret information, and extract knowledge. Cross-disciplinary communication and linking are drivers of innovation, and publishers who continue to offer traditional 'text' journals, whether these be in print or on line, will lose out.

Six rules for semantic publishers

1. Start simply and improve functionality incrementally.
2. Expect greater things of your authors.
3. Exploit your existing in-house skills fully.
4. Use established standards wherever possible.
5. Publish raw datasets to the Web.
6. Release article metadata, particularly reference lists, in machine-readable form.

the next decade will see raw text decreasing in value, with a corresponding increase in the value of semantic services

The future of semantic publishing

It is inevitable that the trends already seen towards semantic publishing will increase, as new methodology makes this easier. Early adopters of semantic publishing will benefit by enhancing the desirability of their journals, leading to increased paper submissions, greater readership, and higher impact factors. Use of the Web as the platform, together with co-operative sharing of metadata and citation information, will create new business opportunities in the form of added-value services. Researchers will bene-

the publishing industry has no intention of being left behind

fit from better, faster, cheaper access to data related to publications, enhancing the capacity for *in silico* meta-research.

One famous view of progress in the online world is that attributed to John Gilmore: 'The Internet treats censorship as damage, and routes around it.' In this view, journals that fail to provide the quality and depth of information that readers come to expect will become increasingly marginalized, as the readers go elsewhere on the Web to find it. However, the selected examples given above, of what publishers are already doing in terms of semantic publishing, show that the peer-reviewed article is far from dead, and set gold standards for other publishers to follow. They show that the publishing industry has no intention of being left behind, and demonstrate that what is possible in the field of semantic publishing is limited only by one's imagination.

Acknowledgements

I thank my colleagues Graham Klyne, Alistair Miles, Katie Portwin and Jun Zhao for stimulating discussions that have influenced my thinking on these matters.
Funding: This work received no specific external funding.

References

1. Berners-Lee, T. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor.* New York, Harper Collins, 1999. Available at http://www.harpercollins.com/books/9780694521258/Weaving_the_Web/index.aspx.
2. Berners-Lee, T., Hendler, J. and Lasilla, O. 2001. The Semantic Web. *Scientific American* 284: 35-43. Available at URL http://www.sciam.com/article.cfm?id=the-semantic-web.
3. Shadbolt, N., Hall, W. and Berners-Lee, T. 2006. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21: 96–101. http://dx.doi.org/10.1109/MIS.2006.62.
4. Evans, J. 2008. Electronic publishing and the narrowing of science and scholarship. *Science*, 321: 395–9. http://dx.doi.org/10.1126/science.1150473.
5. Shotton, D., Portwin, K., Klyne, G. and Miles, A. 2009. Adventures in semantic publishing: exemplar semantic enhancement of a research article. (Submitted for publication.) Preprint available at http://purl.org/net/semanticpublication/Shotton_et_al_PLoS_enhancement_report.pdf
6. Seringhaus, M. and Gerstein, M. 2007. Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, 8: 17. http://dx.doi.org/10.1186/1471-2105-8-17
7. Gerstein, M., Seringhaus, M. and Fields, S. 2007. Structured digital abstract makes text mining easy. *Nature*, 447: 142. http://dx.doi.org/10.1038/447142a
8. Seringhaus, M. and Gerstein, M. 2008. Manually structured digital abstracts: a scaffold for automatic text mining. *FEBS Letters*, 582: 1170. http://dx.doi.org/10.1016/j.febslet.2008.02.073
9. Taylor, C., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R. and Ashburner, M. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26: 889–96. http://dx.doi.org/10.1038/nbt.1411
10. Reis, R., Ribeiro, G., Felzemburgh, R., Santana, F., Mohr, S., Melendez, A., Queiroz, A., Santos, A., Ravines, R., Tassinari, W., Carvalho, M., Reis, M. and Ko, A. 2008. Impact of environment and social gradient on *Leptospira* infection in urban slums. *PLoS Neglected Tropical Diseases*, 2: e228. http://dx.doi.org/0.1371/journal.pntd.0000228.
11. Weiss, P. 2002. Borders in cyberspace: conflicting public sector information policies and their economic impacts. Washington DC, US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, 2002. Available at http://www.weather.gov/sp/Borders_report.pdf.

David Shotton
Image Bioinformatics Research Group
Department of Zoology
University of Oxford
South Parks Road
Oxford OX1 3PS, UK
Email: david.shotton@zoo.ox.ac.uk