Interactive Publications and the Record of Science

ICSTI Winter Workshop

Université Pierre et Marie Curie,
4 place Jussieu 75005 Paris, France
Monday, February 8, 2010

The Web is by nature an interactive environment; yet online journals are mostly static, befitting their traditional role as a never-changing scholarly record. However, this traditional role is increasingly challenged as browser technologies leap forward, dynamic visualization and presentation tools proliferate, and primary data are linked to research articles. In an important and timely workshop, publishers, publishing service providers, librarians, editors and authors meet for a one-day workshop under the auspices of ICSTI (International Council for Scientific and Technical Information) to survey the most exciting and challenging of the new developments, and to begin to identify the necessary infrastructure for including interactive content within the record of science.

Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:15-09:00</td>
<td>Check-in and continental breakfast</td>
</tr>
<tr>
<td>09:00-09:10</td>
<td>Welcome (Session Chair: Elliot Siegel, NLM)</td>
</tr>
<tr>
<td></td>
<td><em>An introduction to the purpose of the workshop and its genesis as a technical project of ICSTI on interactive visualizations. A brief account will be given of a 2008 NLM–Elsevier project comparing knowledge gain and satisfaction amongst medical students with standard and interactive versions of the same paper.</em></td>
</tr>
<tr>
<td></td>
<td><strong>Elliot R. Siegel</strong> is Associate Director for Health Information Programs Development at the US National Library of Medicine, with responsibility for the Library’s International Programs, Planning and Evaluation, and Outreach functions. At NLM since 1976, he has led research programs in biomedical informatics (decision support, computer-mediated communication, performance measurement, and consumer health), and developed new E-Health initiatives for underserved populations in the US and Africa. Elliot is co-Editor-in-Chief of the journal <em>Information Services and Use.</em></td>
</tr>
<tr>
<td>09:10-09:30</td>
<td>Interactive Science Publishing: a joint OSA-NLM project (Mike Ackerman,</td>
</tr>
</tbody>
</table>

I. Interactive visualizations

The purpose of this session is twofold: (1) to showcase examples where journals use interactive (usually three-dimensional) data visualization tools as an integral component of the scientific discourse and demonstrate their benefit; but (2) also to define the ‘costs’ involved in terms of providing infrastructure for creating, editing, reviewing and archiving such interactive content.

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:10-09:30</td>
<td>Interactive Science Publishing: a joint OSA-NLM project (Mike Ackerman,</td>
</tr>
</tbody>
</table>
Interactive Scientific Publishing (ISP) has been developed by the Optical Society of America with support from the NIH National Library of Medicine. It is an initiative that allows authors to publish large 2D and 3D datasets with original source data that can be viewed and analyzed interactively by readers. ISP provides the software for authors to organize and publish source data while offering readers the viewing and analysis tools.

Michael J. Ackerman is the National Library of Medicine’s Assistant Director for High Performance Computing and Communications, providing guidance for NLM's telemedicine, distance collaboratory, advanced networking and imaging interests. He has a PhD in biomedical engineering from University of North Carolina, Chapel Hill. He was a research physiologist in the Naval Medical Research Institute, and later head of the Institute's Biomedical Engineering and Computing Branch. Dr Ackerman came to NLM in 1987, where he has had a distinguished career in applying interactive technology to medical education (including leading the ‘Visible Human’ Project), and overseeing the Library's non-bibliographic data bases.

Breaking out of 2D: interactive PDFs (Michelle Borkin, Harvard University)

Interactive PDFs have obvious appeal to the reader and the publisher; they need only widely available (and comfortingly familiar) software to view, and they require no extra effort from the publisher to create or archive. They offer many advantages to both reader and author including providing 3D interactivity. How easy are they to create, and can they be used or designed for specific domains? What are the implications of using specific closed-source or proprietary software? What improvements does it bring to the scientist?

Michelle Borkin is a researcher at Harvard’s School of Engineering and Applied Sciences. She works on creating new approaches to interdisciplinary scientific imaging, data exploration and image analysis with a focus on 3D visualization. As part of the ‘Astronomical Medicine’ project she collaborates with both astronomers and doctors to help them analyze and visualize their multidimensional data. She co-authored the 2009 Nature paper which was the first use of a 3D PDF for graphics in an article in a major scientific journal.

Accessing the data: going beyond what the author wanted to tell you (Brian McMahon, International Union of Crystallography)

Structure reports in IUCr journals depend on a standard data file format. Structure visualization uses both helper applications (e.g. Mercury, but the end-user has complete freedom of choice) and embedded applets (Jmol). Tools are provided for creating/editing dynamic content, and the handling of such content is completely integrated in the editorial/production workflow. Data validation is an integral part of the peer review process. There are concerns about long-term access to functionality that depends on how particular software programs behave, but archiving the basic
data provides more scope for retaining such functionality.

Brian McMahon has worked for the International Union of Crystallography since 1986, initially as a technical editor and subsequently in the development of computerised applications in publishing, including journal production workflow, web design and content provision, data checking and evaluation, and the provision of interactive services to journal authors and readers. He has been involved for many years in the IUCr’s programme to develop CIF, a standard information exchange and archival framework.

10:40 Coffee break

II. Adding value with enriched content and semantic links

This session surveys some of the growing ways in which value is added to the online publication by exposing the semantic content to computerised application: hyperglossaries, thesauri, taxonomies, chemical substructure display and search. Sometimes the value is added (expensively) by the publisher through semantic markup, sometimes it can be inferred by dynamic post-processing. How important is it to retain this type of added value in long-term archives?

11:00-11:20 Project Prospect and the place of primary data (Richard Kidd, Royal Society of Chemistry)

Project Prospect is an RSC initiative to add extensive semantic markup to chemistry publications. It improves literature searching, enhances the information content of an article by annotating chemical content, and opens the door to database searching of chemical structures. Reliable tagging of content is an intensive editorial process (for example, many chemicals referred to in an article may not be named, but simply referred to as something like ‘13b’); but adding accurate machine-readable information adds tremendous value to an article.

Richard Kidd is Manager of Editorial Production Systems at the RSC in Cambridge, UK. He has a background in Chemical Engineering and works within the Publishing division to implement new technologies into the publication process. Over the past few years this has involved switching to an advanced XML-based journals publication workflow, the digitization and delivery of the RSC Journals Archive, and most recently launching Project Prospect to structure researchers’ science within journal articles.

11:25-11:45 Semantic linking in the Concept Web (Jan Velterop, Concept Web Alliance)

Another approach to semantic markup is explored in applications of the Concept Web Alliance. Instead of relying on at-source markup, readers can stream content through
a large knowledge-based annotation server that overlays annotations, glossaries, hyperlinks etc. deduced from on-the-fly parsing and mapping to discrete concepts organised in a large triples store. The Alliance allows participants to explore many of the new semantic web technologies, maximizing the potential for knowledge discovery while at the same time removing both redundancy and ambiguity from available knowledge.

**Johannes (Jan) Velterop** is a science publisher. He was originally a marine geologist and became a science publisher in the mid 1970s, working successively for Elsevier, a Dutch regional newspaper *De Twentsche Courant*, Academic Press, and *Nature*. He moved on to help establish BioMed Central, the first commercial open access science publisher. He was one of the group who first defined 'open access' in 2001 in Budapest, a meeting resulting in the Budapest Open Access Initiative. In 2005 he joined Springer Science+Business Media as Director of Open Access. In 2008, he left Springer to join Knewco, a company that uses semantic technology to accelerate scientific discovery. Since January 2009 he has been involved in the Concept Web Alliance as one of the initiators.

11:50-12:10

**Visualizing and citing dynamic datasets** (Toby Green, OECD)

OECD is pioneering work in citing and visualizing economic and statistical data sets. These are available outside of journal publications, but can also be referenced and linked into from scholarly articles. Many of the data sets of interest here are growing continuously, so again the challenge is to properly cite a dynamic object, and to recreate a particular time slice or other subset of a complex and large data set.


12:15-12:35

**The Article of the Future** (Emilie Marcus, Cell Press)

Cell Press and Elsevier have launched a project, Article of the Future, as an ongoing collaboration with the scientific community to redefine how the scientific article is presented online. The goal is to fully exploit online capabilities, allowing readers individualized entry points and routes through the content, while using the latest advances in visualization techniques.
Emilie Marcus is Editor of *Cell* and Editor-in-Chief and VP of Content Development at Cell Press. She came to the position in 2003 after a graduate and postdoctoral research career at Yale University, where she received her PhD, and then at the Salk Institute, and 5 years of editorial experience at the journal *Neuron*. Under her direction in 2005 Cell Press adopted a 12 month open archive policy for all published journal content. She is also responsible for launching the *Leading Edge* section in *Cell* as a forum for discussion and debate on global policy issues. Recently, she has pioneered a new format for the presentation of scientific papers online called the ‘Article of the Future’ which takes advantage of the opportunities of the web to break free from the linear structure of the printed article.

12:40  
*Lunch break*

13:45-13:55  
**Introduction to afternoon session** (Session Chair: John Helliwell, University of Manchester)

*In crystallography, a data-intensive science, the coupling between text and primary data is seen as a very important part of a scientific publication and both aspects are being systematically enhanced. Is this approach followed in all science publications where optimising links to the experiment to the benefit of the reader should surely be followed? How can the relationship be specified and maintained? Who will store and provide organised long-term access to the data sets?*

John R. Helliwell trained in physics and molecular biophysics and is now Professor of Structural Chemistry at the University of Manchester. He is former Editor-in-Chief of the journals of the International Union of Crystallography (IUCr) and Past President of the European Crystallographic Association. His research involves crystallography methods developments applied to structural chemistry and biology. His role with the IUCr gave him an overview of the scientific policy and associated technical developments in structural science publishing.

### III. The archival problem and infrastructure for solutions

This session considers what needs to be in place to allow the added value and functionality of interactive publications to be retained and made available into the future. Is there a place (or a need) for emulation of legacy software environments? Do we yet have a consensus on how to package, identify and interlink the independent components of a complex article (e.g. linking article text, figures of various types – including animations, movies, audio annotations, data sets, procedural scripts)? Can we handle distributed articles – text on a publisher’s web site, associated data in a subject repository? Can we identify and retrieve slices through large archived data sets? Do we have any idea how to approach changing data sets? What is actually worth keeping for posterity anyway?
13:55-14:15  **What needs to be archived and what needs to be done?** (Richard Boulderstone, British Library)

*An important role of the archivist is to decide what actually needs to be archived – one can never retain everything. What elements of interactive publishing are most important as candidates for long-term preservation? Can libraries advise on the most effective container formats for composite documents? Can libraries support long-term access to interactive content through software emulation, or can they help to commission abstraction layers that define interactivity in formats independent of specific operating systems or software?*

**Richard Boulderstone** is the Director of e-Strategy and Information Systems at The British Library. He is a former CTO and Product Development Director at a number of international information providers, and has led the creation of many information based products both in the USA and UK. He is currently leading The British Library’s efforts to create a large-scale digital object management system that will become the primary repository for the Library’s, and hence the UK’s, legal deposit collection of electronic resources.

14:20-14:40  **Maintaining a persistent scholarly citation record when content is protean and identity is cheap** (Geoffrey Bilder, CrossRef)

*CrossRef can bring experience with a number of publishers to the subject of providing persistent identifiers to data sets and linking them to related publications. There are different ways of describing compound documents (articles and their supplementary material, data sets, multimedia content), with distinct DOIs linked through appropriate compound-document schemas. Among current CrossRef initiatives in providing long-term and unambiguous linking tools, the quest for persistent author identifiers suggests the possibility of granular citations of data or nano-publications.*

**Geoffrey Bilder** is Director of Strategic Initiatives at CrossRef, and has over 16 years experience as a technical leader in scholarly technology. He co-founded Brown University’s Scholarly Technology Group in 1993, and was subsequently head of IT R&D at Monitor Group, a global management consulting firm based in Cambridge, Massachusetts. From 2002 to 2005, he was Chief Technology Officer of Ingenta, and just prior to joining CrossRef was a Publishing Technology Consultant at Scholarly Information Strategies, where he consulted extensively with publishers and librarians on emerging social software technologies and how they may affect scholarly and professional researchers.

14:45-15:05  **Bridging the gap between data centres and publishers** (Jan Brase, German National Library of Science and Technology)

*TIB has growing experience in assigning persistent identifiers to data sets, and linking them to primary publications, and has given some thought to standards for citing data sets with the aim of securing appropriate scholarly credit. With the foundation of DataCite – the International Data Citation Initiative, founded in December 2009 by*
Jan Brase has a degree in Mathematics and a PhD in Computer Science. Since 2004 he has been coordinating the non-commercial DOI registration agency at the German National Library of Science and Technology (TIB). So far TIB has registered DOI names for over 600,000 datasets and several thousand additional non-textual information resources. Since January 2010 he is also Managing Agent of DataCite. Based on the approach of TIB, the aim of DataCite is to establish a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects.

15:10  Tea break

IV. W(h)ither journals?

The final session will explore the direction in which these new developments are taking the whole practice of scholarly communication. Conventional journals have been the mainstay of formal scientific discourse for over 300 years. The technological explosion of the early 21st Century offers a bewildering variety of new ways to communicate. How will they complement or supplant the traditional publication?

15:30-16:05  The nature of scholarly publishing in the new century (Timo Hannay, Nature Publishing)

Nature is experimenting energetically with new technologies and techniques of social networking. It is bringing to the scholarly publication elements of interactive review, critique, comment and feedback, and is helping to develop new tools for tracking new literature and sharing scholarly information. Which of these experiments represent passing fads, and which reflect real changes in the communication of scholarly information that will be adopted as reputable practice by the community?

Timo Hannay is Director of Web Publishing at Nature Publishing Group (NPG). His primary responsibility is the creation of innovative products and services that help scientists to make the most of the web as a communication medium.

16:10-16:45  Dumbing down or opening new horizons? (Phil Bourne, University of California San Diego)
Multimedia innovations such as SciVee (‘YouTube for Scientists’) improve the immediacy and impact of scientific reporting, but do they carry the risk of ‘dumbing down’ the information content? Are they a useful adjunct to the written word, or will they replace it in some circumstances? What new challenges of citability and archivability do they raise? What new tools are needed by authors to compose in the new interactive media? Will they be willing to learn? Will the new ways of communicating science change the way scientists work, and indeed how they think?

Philip E. Bourne is a Professor in the Department of Pharmacology and Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California San Diego, Associate Director of the Protein Data Bank and an Adjunct Professor at the Burnham Institute. He is a Past President of the International Society for Computational Biology and an elected fellow of the American Medical Informatics Association. He is the Founding Editor-in-Chief of the open-access journal PLoS Computational Biology and a long standing member of the National Science Foundation, National Institutes of Health and Genome Canada panels responsible for reviewing proposals relating to computational biology.

16:50-17:30  Panel Discussion

Interactive journal articles project

This Workshop forms part of an ICSTI technical project surveying current practice and policy among scientific, technical and medical publishers in incorporating interactive visualizations in their online journal content. For more information about the project, access to an online questionnaire for publishers, and links to examples of interactive publications, visit http://scripts.iucr.org/icsti-project

About ICSTI

ICSTI, The International Council for Scientific and Technical Information, offers a unique forum for interaction between organizations that create, disseminate and use scientific and technical information. ICSTI's mission cuts across scientific and technical disciplines, as well as international borders, to give member organizations the benefit of a truly global community.

http://www.icsti.org

ICSTI acknowledges with thanks the assistance of Professor Yves Epelboin at Université Pierre et Marie Curie in accommodating this Workshop.