Information Services & Use 30 (2010) 1–16
DOI 10.3233/ISU-2010-0607
IOS Press

ICSTI 2010 Winter Workshop

Interactive publications and the record of science [*]

Brian McMahon [**]
International Union of Crystallography, Chester, UK

Abstract. The Proceedings of a one-day Workshop are described, in which publishers, publishing service providers, librarians, editors and authors met under the auspices of the International Council for Scientific and Technical Information (ICSTI) to survey new developments in interactive scholarly publishing, and to begin to identify the necessary infrastructure for including such interactive content within the long-term record of science.

Keywords: Electronic publishing, data visualization, interactive content, digital preservation, semantic processing

1. Background to Workshop

The idea for a Workshop to explore the role of interactive publications in the scientific record arose during the course of a technical project on interactive journal articles commissioned by the International Council for Scientific and Technical Information (ICSTI). This technical project focused on the technologies of interactive three-dimensional visualizations of data described in scientific research articles. Several new and emerging technologies exist that can produce such visualizations, but their adoption by academic journal publishers is still tentative. Among those identified during the course of the project were: arbitrary helper applications launched from a browser to visualize different types of data according to the specific requirements and conventions of different subject domains; PDF plugins allowing manipulation of three-dimensional representations in the context of a typeset page; and embedded Java applets served and controlled by the publisher, again with scope for different applications to suit different domains.

The reasons for publishers' reluctance to embrace such technologies are many. They include: difficulties in incorporating new methods of preparing and disseminating content within existing workflows; uncertainty over the real value of novel presentations to readers and subscribers; concern over the long-term archiving of software implementations. The suggestion was made that a Workshop to address these concerns would be of great interest to publishers, librarians, software developers and scientists alike – the sort of interdisciplinary audience that ICSTI is well placed to address.

The one-day Workshop took place at the Université Pierre et Marie Curie as an adjunct to the annual business meetings that ICSTI conducts in Paris. As well as the regular ICSTI members, the public meet-

[*] Paper presented at the ICSTI Interactive Publications Conference 2010.

[**] Address for correspondence: Brian McMahon, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK. Tel.: +44 1244 342878; Fax: +44 1244 314888; E-mail: bm@iucr.org.

ing attracted representatives of other publishers, scientific databases, research management agencies and libraries. The scope of the meeting extended beyond data visualizations to other interactive technologies that add value to scientific publications, but that challenge the production and archiving processes in a similar way.

The day was organised in four sessions, exploring respectively: interactive visualizations (the initial focus of the technical project); publisher-added value with enriched content and semantic linking; aspects of the archival problem; and the future evolution of the scholarly journal in the web era.

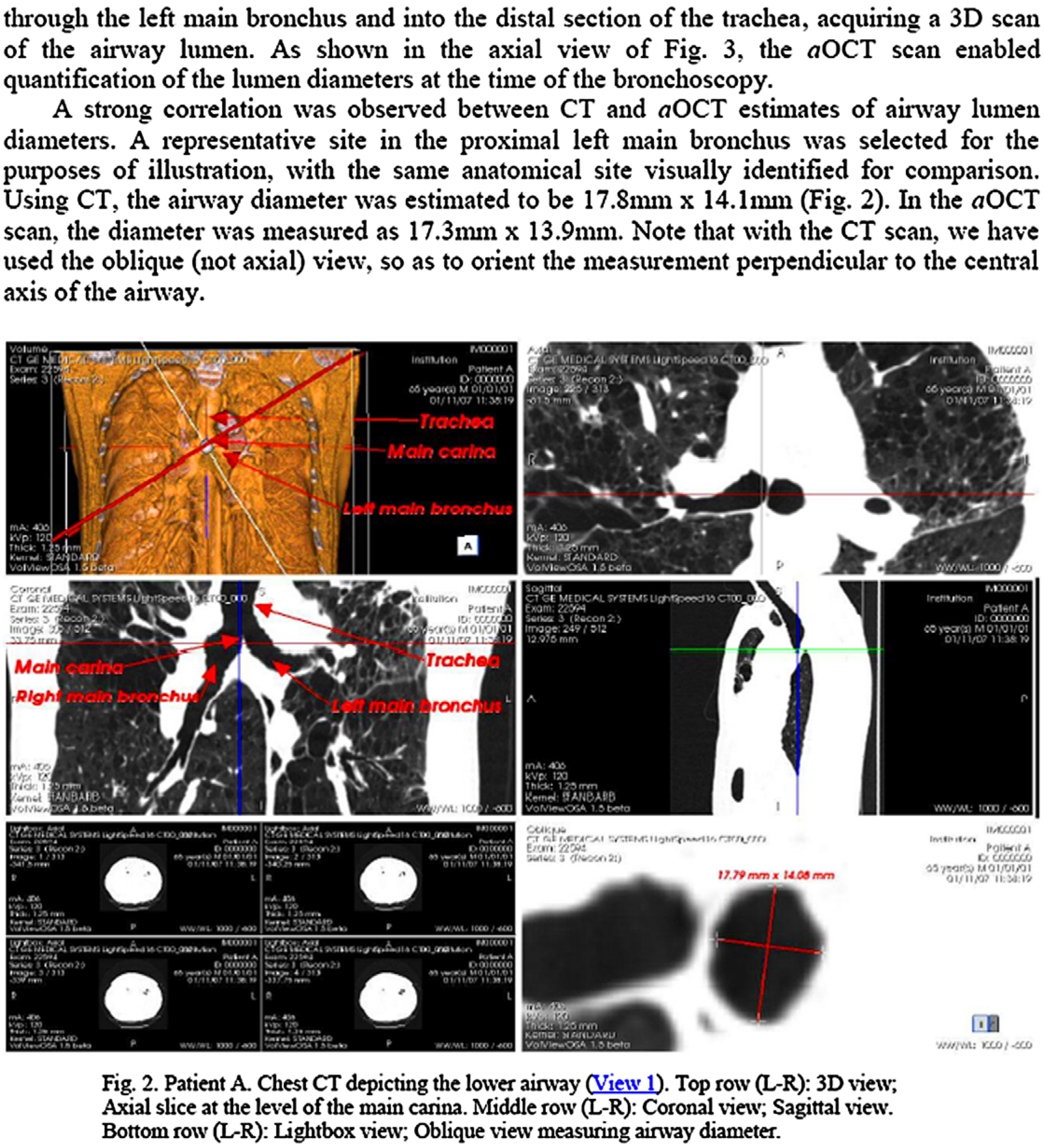
2. Session I: Interactive visualizations

The purpose of the meeting was described by Session Chair **Elliot Siegel** (National Library of Medicine (NLM)), who began by addressing directly the concern of many publishers that novel presentations were of limited value to readers. A joint Elsevier/NLM study had been undertaken to design and conduct a controlled experiment for evaluating the benefits of an interactive journal article against a conventional publication of the same article. An article from the journal *Urology* was selected, and an enhanced version prepared that contained both user-invoked features and presentational improvements. Members of a group of 51 medical students were randomly assigned the conventional ('control') or the enhanced ('experimental') article, and their knowledge gain on reading the article tested by pre- and post-experiment questionnaires. The students were also asked to rate their acceptance of the article format using Likert-scale psychometric questions. The results of the experiment revealed several unexpected findings. The first finding was that the dependent measure of knowledge acquisition showed no difference overall between the control and experimental groups. There was, however, significant gain on the content accessible directly through user-invoked interactive features. Statistically significant variations were correlated with student year and gender (second-year students performed best with both types of article; female students benefited more from the enhanced article), and the acceptance was greater for the experimental article.

Clearly, such an experiment is of limited generality, but illustrates the need to consider the interaction of user demographics, quite apart from a simple comparison of benefits deriving from interactive technologies as enablers of scholarly information dissemination and user education. The study does indicate that there is measurable benefit to the cohort under investigation. One aspect of the methodology is that transformation of an existing article has limited scope for supplying significant new information to the reader; the presentations in this section of the Workshop described initiatives with the specific objective of increasing the information content through interactive visualizations.

The presentation *Interactive Science Publishing: a joint OSA–NLM project* by **Michael J. Ackerman** (NLM) and **John Childs** (Optical Society of America (OSA)) was presented by Elliot Siegel. It described efforts to build a publishing infrastructure suitable for hosting and delivering authors' own databases and providing tools for readers to interact with such data.

The essential elements of the infrastructure included storage for the datasets on a platform that is open access, fully citable and archived within OSA's 'InfoBase' database; and the development in association with a commercial software company of a visualization tool for rendering the various datasets associated with an article. The visualization software ('Interactive Science Publishing' or ISP) is well suited to medical imaging, but encompasses many imaging functions that are found in diverse optical applications. Thus, the user may arbitrarily rotate and zoom projections of three-dimensional objects, change image brightness and contrast arbitrarily or according to stored presets, select different two-dimensional sections, or interact with the data in many other ways.



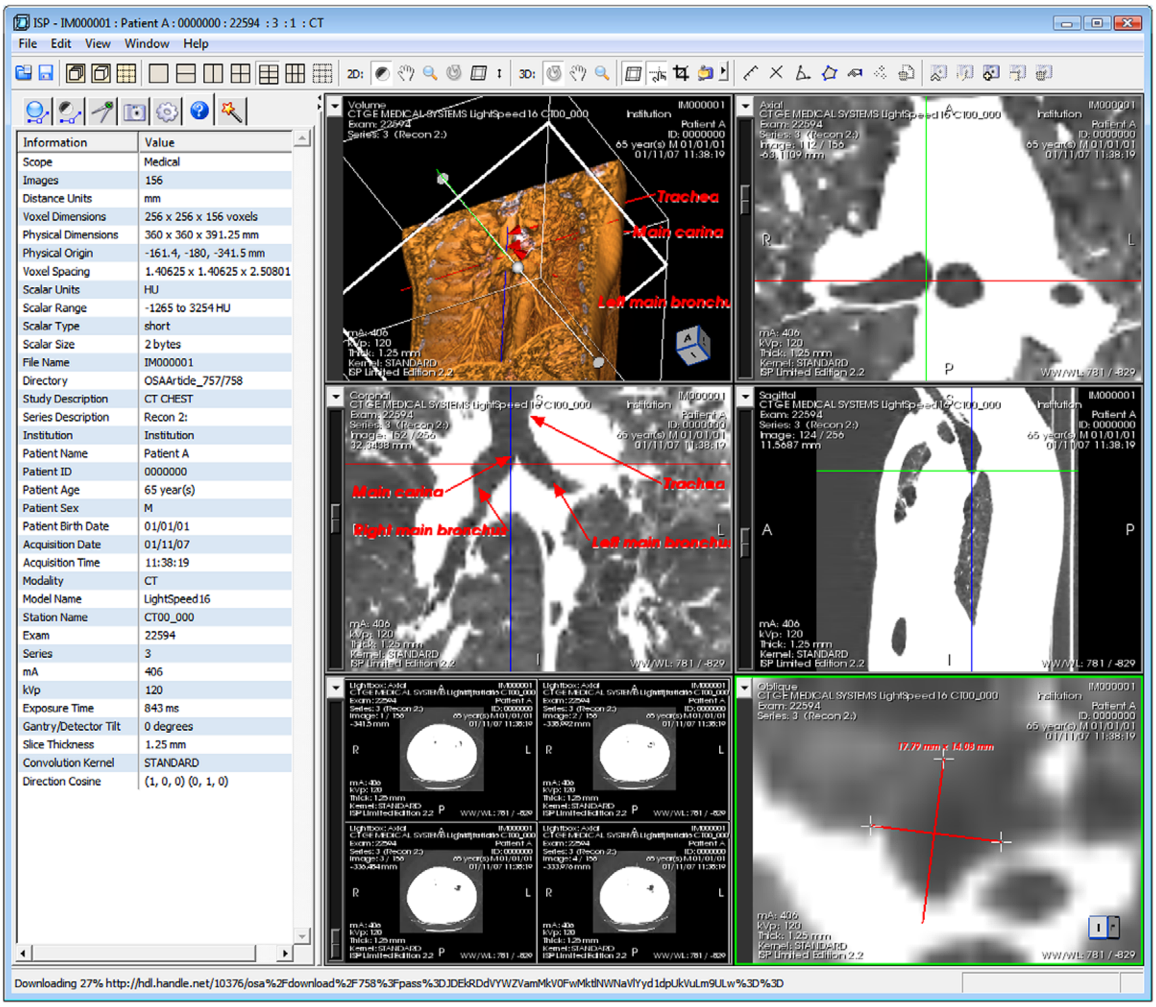
through the left main bronchus and into the distal section of the trachea, acquiring a 3D scan of the airway lumen. As shown in the axial view of Fig. 3, the *a*OCT scan enabled quantification of the lumen diameters at the time of the bronchoscopy.

A strong correlation was observed between CT and *a*OCT estimates of airway lumen diameters. A representative site in the proximal left main bronchus was selected for the purposes of illustration, with the same anatomical site visually identified for comparison. Using CT, the airway diameter was estimated to be 17.8mm x 14.1mm (Fig. 2). In the *a*OCT scan, the diameter was measured as 17.3mm x 13.9mm. Note that with the CT scan, we have used the oblique (not axial) view, so as to orient the measurement perpendicular to the central axis of the airway.

Fig. 2. Patient A. Chest CT depicting the lower airway (View 1). Top row (L-R): 3D view; Axial slice at the level of the main carina. Middle row (L-R): Coronal view; Sagittal view. Bottom row (L-R): Lightbox view; Oblique view measuring airway diameter.

Fig. 1. A figure caption in an article involving medical imaging hyperlinks to an associated dataset.

In practice, the dataset is most usually accessed through a hyperlink in the caption of a figure within the PDF version of the article (Fig. 1). The hyperlink refers to a persistent identifier through which the dataset may be located and downloaded (in the OSA implementation the persistent identifiers are supplied through the HANDLE SYSTEM® of the Corporation for National Research Initiatives [9]). The downloaded file advertises itself to the reader's operating system as being of the type associated with the ISP application, and the ISP software is launched (if it is not already open). The initial visualization of the data within ISP is in most cases identical to that shown in the corresponding static figure in the article. However, the reader is now free to manipulate the images in any way permitted by the application. Figure 2 shows examples of the types of change that can be made.



ISP - IM000001 : Patient A : 0000000 : 22594 : 3 : 1 : CT

Information	Value
Scope	Medical
Images	156
Distance Units	mm
Voxel Dimensions	256 x 256 x 156 voxels
Physical Dimensions	360 x 360 x 391.25 mm
Physical Origin	-161.4, -180, -341.5 mm
Voxel Spacing	1.40625 x 1.40625 x 2.50801
Scalar Units	HU
Scalar Range	-1265 to 3254 HU
Scalar Type	short
Scalar Size	2 bytes
File Name	IM000001
Directory	OSAArticle_757/758
Study Description	CT CHEST
Series Description	Recon 2:
Institution	Institution
Patient Name	Patient A
Patient ID	0000000
Patient Age	65 year(s)
Patient Sex	M
Patient Birth Date	01/01/01
Acquisition Date	01/11/07
Acquisition Time	11:38:19
Modality	CT
Model Name	LightSpeed16
Station Name	CT00_000
Exam	22594
Series	3
mA	406
kVp	120
Exposure Time	843 ms
Gantry/Detector Tilt	0 degrees
Slice Thickness	1.25 mm
Convolution Kernel	STANDARD
Direction Cosine	(1, 0, 0) (0, 1, 0)

Downloading 27% http://hdl.handle.net/10376/osa%2Fdownload%2F758%3Fpass%3DJDEkRDdVYWZVamMkV0FwMktlNWNaVlYyd1dpUkVuLm9ULw%3D%3D

Fig. 2. The hyperlink from the figure caption illustrated in Fig. 1 downloads an associated dataset and launches a visualization program as a separate helper application. The initial view within the helper software is identical to the static image in the article. The reader may however manipulate the content. In this example, the three-dimensional section of the lower airway (top left panel) has been oriented in an oblique view, and a different section (identified by the heavy white rectangle) is chosen for depiction in the lower right panel. The contrast of the sectional views has also been changed by the reader.

Since the launch of this infrastructure in 2008, a number of special issues of OSA journals have been published, featuring a few dozen enhanced articles with links to over 300 datasets. Although the datasets are separately citable and can be downloaded directly, it has been found that most accesses are through the hyperlinks embedded within the articles. Feedback from authors, editors and usability experts has been positive. Reader feedback (as assessed through online questionnaires voluntarily completed) has also been positive from those who get past the initial learning curve, but there have been reports of installation and navigation problems, and inadequate help facilities. Nonetheless, the project continues to be developed.

The preparation of such enhanced figures is complex, and places a significant burden on the authors. However, authoring/editing functionality is built into the ISP software that is used as the default viewer (although long-term authoring functionality is currently absent from the free version), and so it may be expected that the community will find the procedures easier as it becomes more familiar with the tools. Current versions of the software are available only for Microsoft Windows and Mac computing platforms, and are released to the community as time-limited evaluation versions. An advantage of this approach is that it permits roll-out of enhanced software as it becomes available, although users can get annoyed when they are forced to install such upgrades. In principle, other software could be developed (for example, by open-source developers) to provide the same functionality, since the underlying data sets are freely available.

Audience discussion after the presentation focused on the need for peer review of datasets that were published, in a form such as this, as essential components of the article. There was concern that such a requirement would unduly increase the burden of work faced by reviewers.

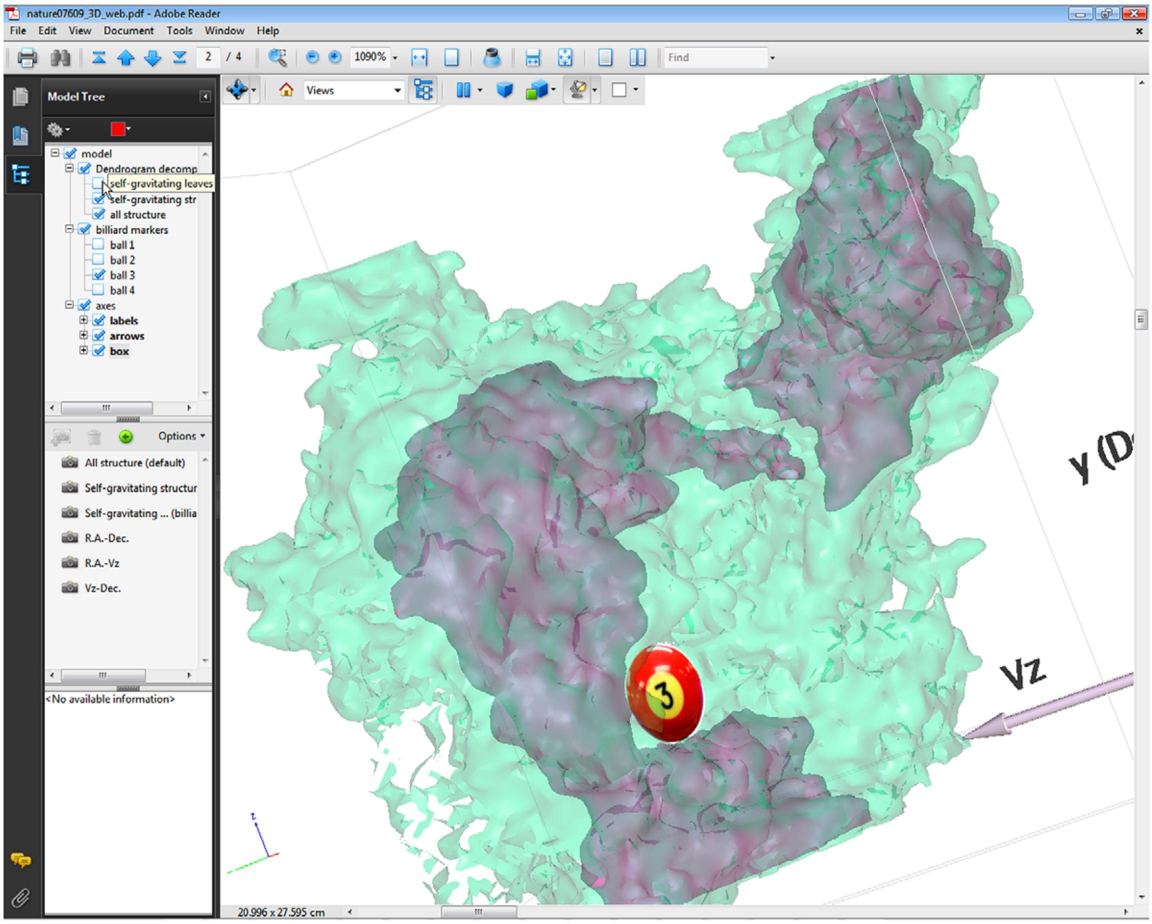
A different approach to the presentation of interactive figures was demonstrated by **Michelle Borkin** (School of Engineering and Applied Sciences, Harvard, MA, USA) in *Breaking out of 2D: interactive PDFs*. Here, the three-dimensional visualization sits within the PDF document, and does not require additional software for its manipulation (assuming the reader is using the most common PDF rendering software, Adobe Acrobat Reader).

The portable document format (PDF) developed by Adobe Systems Inc. has been able to embed, view and navigate three-dimensional (3D) content since the release of the Adobe Acrobat 7.0 family of tools in 2005 [1]. The technology derives from materials licensed from the company Right Hemisphere (which specialises in visual product and business information for manufacturing companies), and adopts many of the rendering conventions, software interfaces and file formats of computer-assisted design (CAD) applications. The usual method of integrating a 3D figure into a PDF document is to use the Adobe Acrobat 9 Pro Extended software. This can, however, import a range of CAD or virtual reality modelling language (VRML) formats, allowing greater flexibility in producing the upstream 3D model.

An example of a research paper using such visualization to powerful effect was a *Nature* astronomy Letter [3] depicting self-gravitating structures in a region of star formation. The interactive figure (Fig. 3) allows the reader not only to reorient and zoom the view, but also to show or suppress individual features. In this case the surface renderings derived from volumetric astronomy data are not straightforward to create, and use 3D Slicer software adapted within the Harvard University 'Astronomical Medicine' project [2], a collaborative effort developing visualization technologies suitable for both astronomy and medical imaging.

The visual information content of interactive 3D figures such as those in the *Nature* article is very high and certainly enhances the reader's perception of the 3D model. However, in this technology the 3D model itself is constructed from geometric primitive objects, and thus represents a pre-rendered abstraction from the original data that the author wishes to present. The PDF article contains the data to be visualized within itself; those data therefore do not have to be located and retrieved separately, as in the case of the OSA enhanced figures. On the other hand, they are not primary data; and they are not easily accessible for peer review using other tools.

A different approach is taken by the journals of the International Union of Crystallography (IUCr), as illustrated by **Brian McMahon** in *Accessing the data: going beyond what the author wanted to tell you*. Authors of articles describing crystal structures are required to deposit the structural data, either with the journals, using the domain standard Crystallographic Information File (CIF) format, or with the Protein Data Bank (PDB), an established data curation centre, in the case of biological macromolecular structures. These structural data sets are routinely validated and checked for scientific consistency, either as part of the journal standard peer review procedure, or as part of the PDB curation task. The structural


nature07609_3D_web.pdf - Adobe Reader
File Edit View Document Tools Window Help
2 / 4
1090%
Find
Model Tree
Views
Options
y (D
Vz
3
20.996 x 27.595 cm

- ☑ model
  - ☑ Dendrogram decomp
    - □ self-gravitating leaves
    - ☑ self-gravitating str
    - ☑ all structure
  - ☑ billiard markers
    - □ ball 1
    - □ ball 2
    - ☑ ball 3
    - □ ball 4
  - ☑ axes
    - ☑ labels
    - ☑ arrows
    - ☑ box

All structure (default)
Self-gravitating structur
Self-gravitating ... (billia
R.A.-Dec.
R.A.-Vz
Vz-Dec.

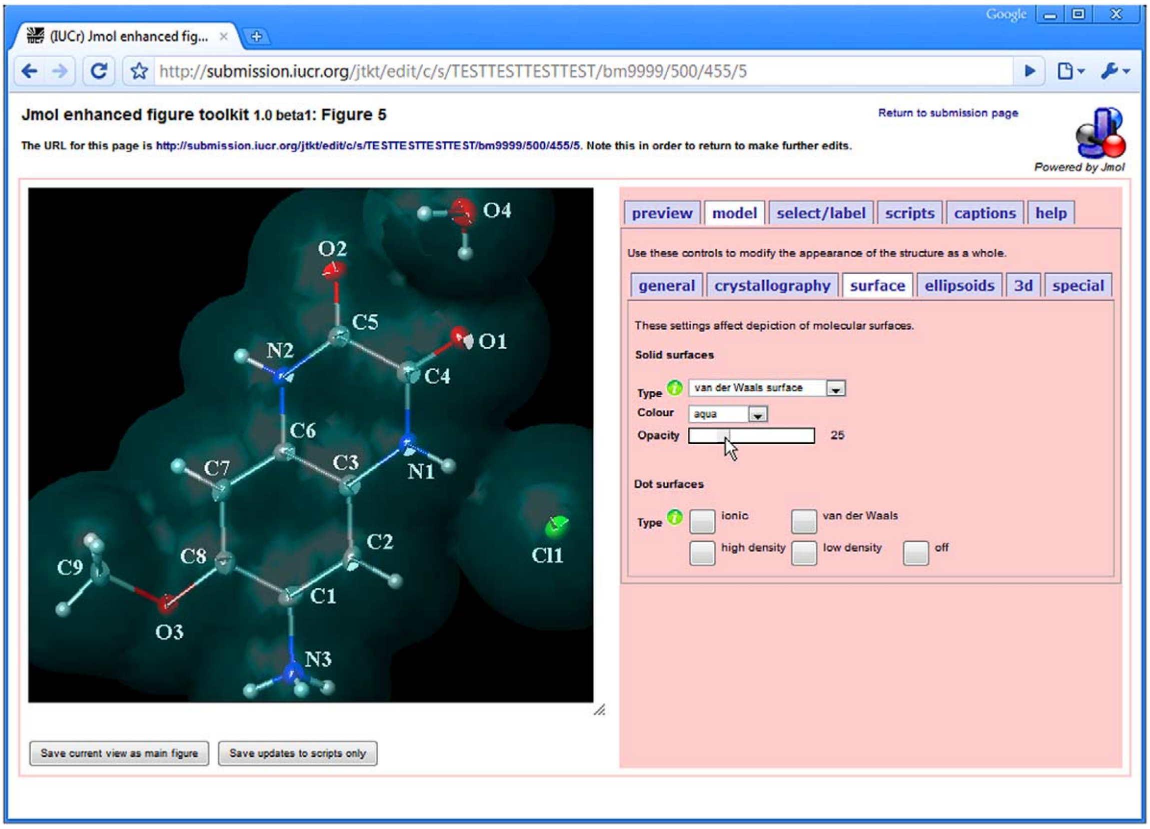
<No available information>

Fig. 3. An interactive figure in an astronomy article. The reader is showing or hiding individual components of the 3D model using the hierarchical tree view at the left of the screen.

data sets are always freely available for download as supporting information for the research article. The journals provide links from their table of contents to an embedded Java molecular visualization of each structure in a standard orientation using the program *Jmol* [6]. This has been standard practice for several years.

Recently, the IUCr journals have developed a toolkit designed to make it easy for authors to create bespoke visualizations of the structure using the same *Jmol* application [7]. Jmol is a powerful tool, but has a steep learning curve. The idea is to provide an untutored author with graphical controls on a web page that allow the straightforward generation of a specific view of the structure in *Jmol*. Once the desired view is achieved, a single keystroke saves the complete graphics state (in the form of a normalized sequence of *Jmol* scripting commands) to the journals server computer. Server-side utilities provide storage for the dataset and its associated scripts, and integrate completely with the journal production, submission and review systems.

Hence, an intending author can use the toolkit (Fig. 4) to create an interactive visualization that is presented to the referee during the standard peer review process. The author can, if necessary, make further edits to the figure within the submission system environment, using the same toolkit. If the


(IUCr) Jmol enhanced fig...
http://submission.iucr.org/jtkt/edit/c/s/TESTTESTTESTTEST/bm9999/500/455/5
Return to submission page
Powered by Jmol
O4
O2
C5
O1
N2
C4
C6
C7
C3
N1
C9
C8
C2
Cl1
C1
O3
N3
preview
model
select/label
scripts
captions
help
general
crystallography
surface
ellipsoids
3d
special
Save current view as main figure
Save updates to scripts only

Jmol enhanced figure toolkit 1.0 beta1: Figure 5

The URL for this page is http://submission.iucr.org/jtkt/edit/c/s/TESTTESTTESTTEST/bm9999/500/455/5. Note this in order to return to make further edits.

Use these controls to modify the appearance of the structure as a whole.

These settings affect depiction of molecular surfaces.

Solid surfaces

Type: van der Waals surface

Colour: aqua

Opacity: 25

Dot surfaces

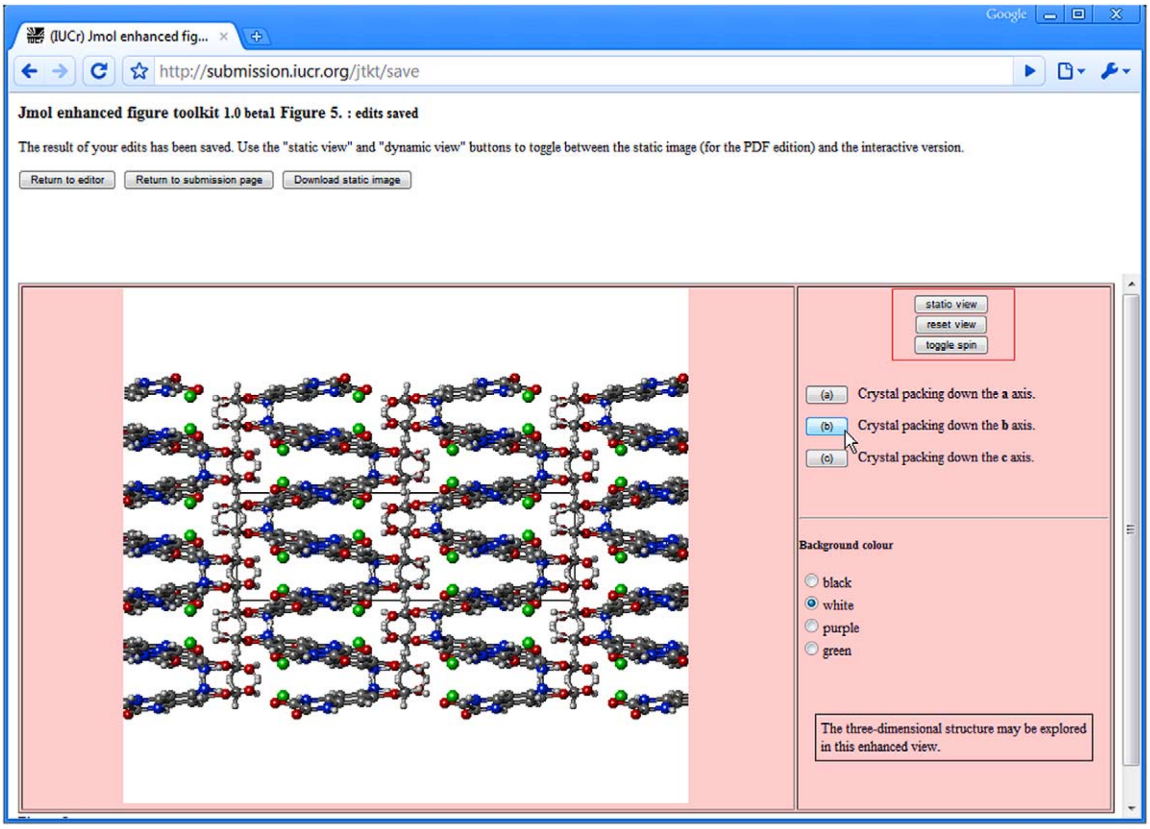
Type: □ ionic □ van der Waals □ high density □ low density □ off

Fig. 4. During submission of a research article, an author establishes a specific rendering of a molecular structure in *Jmol* using the IUCr enhanced figure toolkit. *Jmol* uses the stored structural data to calculate the amplitude of atomic displacements from their mean values (the loci are represented as ellipsoidal objects) and to superimpose a translucent space-filling representation of the atomic van der Waals radii.

article is accepted for publication, the enhanced figures will be published from the journals' full-text HTML publishing platform without any manual intervention by journal editorial staff. The system will also auto-generate a static image of the initial graphics state of the visualization. This is used in the PDF version of the article as an archival representation of the enhanced figure; it also functions as a failover rendering if readers of the full text do not have Java and JavaScript actively running in their browsers.

In fact, the toolkit also allows authors with little or no knowledge of HTML, JavaScript and the *Jmol* scripting language easily to create complex figures with multiple parts and alternative orientations or styles of representation (Fig. 5). The author can therefore create multiple specific 'preferred views' of the data.

However, when the enhanced figure is published, the reader is not constrained to just these views. A right mouse click in the image field creates a pop-up menu of *Jmol* functions, thus allowing the reader extensive control over the rendering style, the ability to measure distances and angles in the model, to overlay numerous colour coding schemes (e.g., atoms coloured by element type, or by amplitude of their displacements from their mean positions), or to show distinct crystallographic symmetry operations. Indeed, a console function allows a reader with expert knowledge of *Jmol* to modify the representation arbitrarily. The reader may also view or capture the underlying primary data directly, if so desired.


(IUCr) Jmol enhanced fig...
http://submission.iucr.org/jtkt/save
Return to editor
Return to submission page
Download static image
static view
reset view
toggle spin
(a)
(b)
(c)

Jmol enhanced figure toolkit 1.0 beta1 **Figure 5. : edits saved**

The result of your edits has been saved. Use the "static view" and "dynamic view" buttons to toggle between the static image (for the PDF edition) and the interactive version.

(a) Crystal packing down the **a** axis.

(b) Crystal packing down the **b** axis.

(c) Crystal packing down the **c** axis.

Background colour

○ black
● white
○ purple
○ green

The three-dimensional structure may be explored in this enhanced view.

Fig. 5. The author can use the enhanced figure toolkit to construct complex web pages showing different views of the structure. Whenever the author saves the current edit, the finished version of the figure is displayed exactly as it will appear when published online. Note the ‘toggle spin’ widget provided as a standard utility by the toolkit software.

The different visualization technologies demonstrated in this session are probably all attractive to different communities, or in different applications. They also offer different opportunities and challenges for validation and archiving. The crystallographic community is particularly fortunate in having a domain-wide standard that permits routine deposition, validation and exchange of datasets. Since deposition and visualization are both routine, it has also established an infrastructure for identifying (by digital object identifiers (DOI)), archiving (in journals or data banks) and linking articles with associated data that might usefully serve as a model for other communities.

3. Session II: Adding value with enriched content and semantic links

The second session explored other ways in which the reader might interact with an article, focusing less on data visualization and more on the additional value available through presentational reorganisation and semantic markup.

Data visualization does remain an important element of the value-added services that the Organisation for Economic Cooperation and Development (OECD) provides for users accessing its books, papers and reports online. **Toby Green** (OECD Publishing), in his presentation *Publishing datasets: visualizing and citing*, demonstrated how statistical tables and datasets relevant to OECD publications were available for download and inspection by subscribers. Tools were made available allowing graphing, histogram building, trend depiction and correlation of statistical data, and a number of examples were provided of how such graphics-driven analysis could rapidly and easily highlight trends or relations that were not obvious in a conventional tabulation.

In many respects, the OECD infrastructure parallels the OSA InfoBase described in the first session. In particular, the datasets of interest can be individually cited (through DOIs). In organising its datasets, the OECD uses hierarchies of collections, with levels similar to 'serial', 'book or issue' and 'chapter' level, and DOIs are assigned at each level as they would be to conventional book or issue publication. In this hierarchy, subsets of a parent data set are each considered as discrete data publications, and assigned distinct identifiers.

However, unlike many of the datasets considered in the first session, statistical tables can grow with time, and in some cases can be modified in other ways than by addition of new data. In managing such alterations, the OECD will maintain the same identifier for a dataset that changes just by growth. The associated metadata is used to explain the time-dependent change in content. Where a dataset changes in other ways, the new version is assigned a *new* identifier. The associated metadata explains the change, but also provides a link back to the pre-revision dataset. The principles behind these management policies are described in an OECD white paper [4], and they significantly enrich the emerging requirements for adequate identification and retrieval of specific data values used in, or associated with, a specific research result. They may not, however, be fully extensible to the enormous time-series data collections found in some of the earth, space and physical sciences, where the number of possible rearrangements of data to define distinct subsets with unique identifiers can grow combinatorially to unimaginable values.

Supplemental data sets are becoming an increasingly important component of the scientific literature, whether available from a central resource or deposited directly with the publisher. In *Evolving the scientific article* **Emilie Marcus** (Cell Press) described the novel online presentation of an article that has been adopted by the journal *Cell*. The original goals behind this project had been to rethink the online presentation of an article, developing a hierarchical presentation of text and figures allowing readers to drill down through layers of content based on their level of expertise and interest; to provide multiple mechanisms for conveying the core content of the article; to redefine the unit of publication to address the problem of burgeoning supplemental data; to involve the scientific community in developing and refining the ideas; and to integrate video, sound and animation into the scientific article.

The new format uses many of the features of compact and effective web presentation that have evolved in fields outside of scholarly publishing. While a more or less conventional PDF version of the article is still produced, suitable for printing, the online presentation organises the sections and subsections of the article in a tabbed interface, with nested tabs as appropriate to reflect the hierarchical structure of the article. The online medium allows richer use of graphics and multimedia than conventional print: where feasible, a graphical abstract is presented to enhance the 'take-home message' of the publication. Video abstracts and author interviews are also provided in some cases. Different tabs allow a synoptic view of all the figures and tables relevant to the article, whether these appear in the conventional published text or as supplemental documents. There is rich linking between the components of the article: when a figure or table is viewed, snippets of text within the article can be viewed alongside, to show the context in which it has been referenced. There is also real-time reference analysis. When the reference list is viewed, dynamic citation lists are generated that show the current number of citations for each reference stored in major bibliographic databases.

While much of the improvement in usability of the article is the result of more effective presentation, the integration of the supplemental data and associated multimedia annotations with the full text, as well as the rich and effectively presented links between references, figures, tables, captions and other components of the article offer a significant advance in describing and managing multi-component documents. Cell Press is also developing additional value-added content. ‘Enhanced SnapShots’ are reference guides for important topics in cell and molecular biology, typically presented as flow charts or connected networks, where additional information not available in print can be layered onto HTML or interactive PDF documents as animations, embedded captions or dynamic visuals. ‘Reflect’ is a pilot scheme using embedded semantic tagging to trigger pop-up windows giving more information about individual proteins, genes and molecules.

Enhanced richness in content was further discussed by **Richard Kidd** (Royal Society of Chemistry (RSC)). In *Semantics and standards in chemistry*, he described some of the projects of the RSC that aimed to provide the reader of a chemistry article with more effective access to the structures and reactions that the article described. Project Prospect was launched in 2007 to generate editorial markup of chemical content building on existing standards, such as the InChI unique chemical identifier, the ChEBI public database of some compounds and groups of compounds, public gene, sequence and cell ontologies, and the IUPAC *Gold Book* [8], a dictionary of chemical terms. With the assistance of various software tools, the journal editors could mark up chemical terms, compound names and generate links to databases of chemical structures and to machine-readable descriptions of those structures. The editorial tools included text-mining applications; but the subtleties of text descriptions (distinct chemical names can be identified in text simply by generic descriptions or by numerals referencing unstructured material elsewhere in the article) make such annotation a time-consuming and hence expensive task.

Recent efforts have focused on semantic annotation pre-publication (by the authors) or post-publication (using ‘the power of the crowd’ [5]). The RSC has been collaborating with Microsoft Research, Science Commons and the University of California, San Diego to develop an ontology add-in, and with Microsoft Research and the University of Cambridge to develop an authoring tool for chemical structure, *Chem4Word*, both add-ins to Word 2007. These will shortly be made available to authors in an effort to capture more semantic content from the start of the publishing process.

While this may provide great benefits for future publications, there remains a huge amount of legacy content that lacks such markup; there is also much published material that contains errors that would benefit from appropriate online annotation. The ChemMantis project aims to provide readers with tools to provide such markup, and to deposit published structures into the ChemSpider database. Such initiatives are very new and still under early development, but the ambition is to allow a progressive semantic enhancement of the entire backfile of RSC journals (extending to 1841) through community curation. To be effective, such a programme must have good validation tools, and will depend upon, as well as contribute to, distributed chemical knowledge bases.

In *Semantic linking and the Concept Web*, **Jan Velterop** (Concept Web Alliance) suggested a different approach to semantic markup in the life sciences. The Concept Web Alliance (CWA) is an open collaborative community of publishers, librarians, informaticians, scientists and other stakeholders interested in the challenges flowing from the production of unprecedented volumes of academic and professional data. Such challenges include storage, interoperability and analysis of massive and disparate data sets.

Among the tools under development to tackle these challenges are formal knowledge bases built from triples of ‘concepts’, by which is meant distinct and uniquely identifiable encapsulations of real or abstract objects and relationships. For example, a particular chemical compound will have a single identifier, independent of its name, synonyms, terminology in other languages, chemical formula or other representations. Establishing a comprehensive store of such concepts will be a very large undertaking, because of the amount of disambiguation and elimination of redundancy that will be necessary. However, in principle, from a concept store can be build triples of concepts, the members of which respectively take on the role of subject, predicate and object. Such triples form ‘assertions’, and they can also be stored within the same electronic information systems.

To what uses can such a knowledge base be put? With sufficiently powerful concordances, a scientific article can be parsed on the fly and marked up with annotations referencing the concepts and assertions associated with each term in the article. Hence the full text delivered by a publisher can be overlaid with rich semantic hyperlinks. These might typically be realised as pop-ups allowing the reader to locate references to a term in online dictionaries, databases of structures or of product suppliers. This parallels some of the functionality of the RSC’s Project Prospect, for instance – but with the difference that the markup is applied *a posteriori*. (On the other hand, contributing publishers could supply the markup *a priori* – during the authoring or editorial processes. This should reduce ambiguity; but it does have the possible disadvantage of reflecting the publisher’s view of the most appropriate markup at a single point in time, when the article was produced.)

Of course, assertions could be constructed arbitrarily from triples of concepts; how is the validity of any assertion to be judged? The possibility arises of nanopublications: assertions with associated metadata attributes that can describe such things as who claims authorship of the statement, who approves of it, who curates it, whether it has been peer reviewed, and so forth. All such assertions would themselves be managed as concept triples. The idea seems far removed from the accepted ways of transmitting knowledge in the scholarly literature. But by abstracting the semantic content of published articles, the approach offers the possibility of discovering new knowledge through machine analysis of the concepts extracted from the otherwise overwhelming volume of scientific information.

4. Session III: The archival problem and infrastructure for solutions

Two themes that emerged from the morning sessions were that interactive publications almost invariably involved increased complexity in the description of a document (with concomitant need to link reliably between document components), and that the volume of scholarly information (understood as literature, associated data and semantic databases) would therefore grow even faster than librarians and archivists were anticipating.

In prefatory notes for the afternoon sessions, the intended Session Chair **John Helliwell** (University of Manchester), who was unable to attend, explained that even a discipline such as crystallography, that already required authors to deposit data, could soon see a massive growth in storage requirements. While crystallography journals and data banks currently stored processed experimental data (‘structure factors’) as well as the result datasets (CIFs), there was increasing pressure to archive the raw diffraction images generated by the experiment. The reasons included didactic and validation needs; but there was also an awareness that the raw datasets included additional science that could not be extracted by current techniques, but that lay latent for future exploitation; and so were a valuable resource in their own right. Such datasets were orders of magnitudes larger than those currently archived by the publishers and data banks, and might perhaps be preserved at national experimental facilities. Again, robust techniques would be needed to identify and link relevant data to publications in a distributed ‘publication’ environment.

One organisation with extensive experience of managing identifiers for digital content is CrossRef, whose representative **Geoffrey Bilder** acknowledged the scale of the problem in his presentation *Maintaining a persistent scholarly citation record when content is protean and identity is cheap.* Many challenges are being raised by distributed publications, where supporting data is stored remotely from the published article (and where the article may indeed rely on data collected or curated by other researchers), and by semantically enhanced articles such as those envisaged by the Concept Web Alliance.

One of the most significant of these challenges (and possibly most difficult to solve) is the due identification and crediting of work contributed by individuals. CrossRef has an interest in this area, and has been studying the requirements for a persistent identifier of an individual person, so that authorship (or other roles and responsibilities) can be traced across the use of different forms of a name, different historical affiliations and so on. For scholarly communication, the requirements of such an identifier include that it should support the creation of a clear and unambiguous record of scholarly communication in any medium; it should transcend discipline, geographic/national and institutional boundaries; and it should identify 'contributors', not just 'authors'. The requirements are technical but also include social components: an acceptable contributor identifier should support reliable attribution in both formally and informally published literature (e.g., blogs, social networking contributions); it should be 'open' whilst complying with the privacy requirements of the individual as well as of various legal jurisdictions; it should be persistent; and it should be controlled by the contributor.

The reference to informally published literature reflects the growing importance in at least some disciplines of novel means of exchanging information between scientists. It is myopic to constrain 'scholarly communication' simply to traditional peer-reviewed published journals. The contributor identifier must connect the author of a journal article to the same person's contributions in data collections, multimedia presentations, blogs, web pages and whatever other online presence there might be. It can also function to some extent in helping to validate the quality of those contributions by association with a trusted publication record, even in media that are not traditionally peer reviewed.

And even in the 'traditional' publication environment, online publications are not as static as their paper forebears. While the article published in a scientific journal has an honoured role as the 'version of record' of the reported research, the mutability of online media has led to the development and recognition of multiple 'versions of record' – for example, an enhanced or a corrected one. CrossMark is a new initiative involving CrossRef that aims to annotate modifications and enhancements to a published article and provide a standard audit trail.

CrossRef was founded in 2000, and so has a decade of experience in assigning digital object identifiers (DOIs) to publications and associated supporting data. DataCite is an organisation founded at the end of 2009 to play a similar role in the registration of DOIs and development of standards, workflows and best practice for scientific primary data sets. In *Bridging the gap between data centres and publishers*, **Jan Brase** (German National Library of Science and Technology, TIB) explained the background to the formation of this organisation by a consortium of large (often national) libraries and information centres. It builds on TIB's existing experience of registering DOIs for large datasets, typically in the earth and marine sciences.

The new consortium includes among its concerns the rescuing of datasets from their status as 'second-class citizens'. Findings of a UK Research Data Service study [10] suggested that: data is difficult to manage after project funding ceases; informal networks provide the primary means of sharing; only 21% use a national or international facility; datasets are not included in impact analysis; and that other researchers may have difficulty in finding it or getting permission to use it. DataCite's mission is to improve the scholarly infrastructure around datasets and other non-textual information. Among its interests in this area is the development of standards for citing datasets.

One of the members of the DataCite consortium is The British Library (BL), whose Director eStrategy, **Richard Boulderstone**, addressed the question *What needs to be archived and what needs to be* *done?* For a large national library, the recognised technical challenges in archiving digital content are exacerbated by statutory duties restricting the library's freedom to select material for preservation. In its transition to a digital archive (in addition to its established expertise in storing hard-copy materials), the BL faces challenges in ingest, storage and access. Digital content is heterogeneous, increasingly complex in its content, encodings and relationships between component parts, and of arbitrarily large (and rapidly increasing) size. Interactive items pose challenges in capturing functionality and in subsequent faithful rendering, and there are questions of whether and how much the library should validate digital material that it is processing. Storage problems include certifying the long-term authenticity of items, guarding against loss or corruption, and resolving external references. The access challenges include: secure sharing with other legal deposit libraries; long-term access beyond the life of the original hardware and software platform; controlled access, honouring intellectual property constraints and privacy issues; and ensuring that the content can be easily located again.

Currently the BL is developing a multiple-site digital library system store that will provide sufficient digital storage over a high-speed secure network, with redundant copies and continuous validation and correction procedures. The current holdings (over half a million digital items occupying ~50 TB) are believed to be secure with respect to bit-level preservation, but considerable work is required to support content-level preservation, ensuring that users can render and use preserved content. The Planets Project is intended to deliver preservation modules by summer 2010, covering identification of at-risk content, support for file format migrations, and a technology watch service.

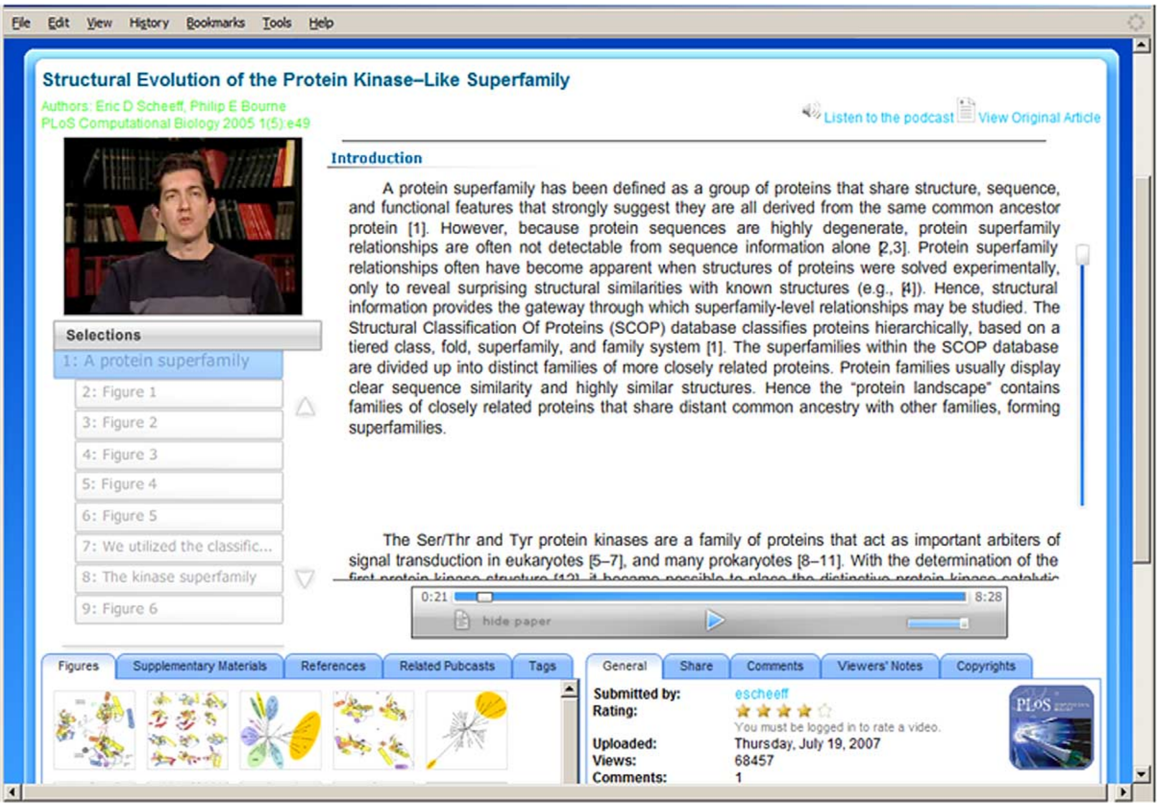
The Library still has significant issues to address concerning the ingest of content. They include a policy for updating dynamic content after initial deposit: currently a snapshot, version-based approach is used, but will this scale? There are debates over whether to archive published outputs, underlying data, or both; concerns over the need to undertake at least some validation of the increasingly diverse content in order to facilitate long-term access; and worries over how container formats may hide significant complexity (as in the example of 3D PDF files).

5. Session IV: W(h)ither journals?

In the course of the day, the apparently simple question 'how does one archive interactive publications?' had led into wide-ranging discussions of the nature and value of material illustrating, enhancing and even modifying the supposed record of science, and so the final session provided an appropriate opportunity for a more general discourse on the changing nature of the scientific journal, and its future status as the vehicle of record.

In *Experiences with rich media in the dissemination and comprehension of science*, **Phil Bourne** (University of California, San Diego (UCSD)) described the development of the SciVee service, a platform for video content including or commenting upon journal articles, conferences, research news and teaching sessions. The initial launch point was a laboratory experiment in which graduate students were invited to create video presentations of their research. Most enjoyed the experience, performed fluently in front of the camera, and anticipated becoming a new generation of 'sciencecasters'. The subsequent web-based platform www.scivee.tv has in two years acquired over 3500 video uploads, has 9000 registered users in 350 communities, and experiences over 75,000 unique visitors and 150,000 page views per month.

The platform hosts a variety of types of presentation. In the context of interactive publications, perhaps the most notable is the 'PubCast': a scientific publication (typically a peer-reviewed journal article) is



Structural Evolution of the Protein Kinase–Like Superfamily

Authors: Eric D Scheeff, Philip E Bourne
PLoS Computational Biology 2005 1(5):e49

Introduction

A protein superfamily has been defined as a group of proteins that share structure, sequence, and functional features that strongly suggest they are all derived from the same common ancestor protein [1]. However, because protein sequences are highly degenerate, protein superfamily relationships are often not detectable from sequence information alone [2,3]. Protein superfamily relationships often have become apparent when structures of proteins were solved experimentally, only to reveal surprising structural similarities with known structures (e.g., [4]). Hence, structural information provides the gateway through which superfamily-level relationships may be studied. The Structural Classification Of Proteins (SCOP) database classifies proteins hierarchically, based on a tiered class, fold, superfamily, and family system [1]. The superfamilies within the SCOP database are divided up into distinct families of more closely related proteins. Protein families usually display clear sequence similarity and highly similar structures. Hence the "protein landscape" contains families of closely related proteins that share distant common ancestry with other families, forming superfamilies.

The Ser/Thr and Tyr protein kinases are a family of proteins that act as important arbiters of signal transduction in eukaryotes [5–7], and many prokaryotes [8–11]. With the determination of the

Submitted by: escheeff
Rating: ★★★★☆
You must be logged in to rate a video.
Uploaded: Thursday, July 19, 2007
Views: 68457
Comments: 1

Fig. 6. A SciVee pubcast in which an author provides a video commentary on the content of a peer-reviewed research article. The structured presentation includes the text, references and figures from the original publication.

combined with a parallel video, in which portions of the commentary can be synchronized with the corresponding figure, table or portion of text (Fig. 6). Informal usability experiments have been performed where half a class has been given a published article to read for the same time that it took the other half of the class to watch the pubcast of the same article. The members of the class were then asked to complete a multiple-choice questionnaire testing their comprehension of the material. It was found that the students who used the pubcast performed slightly better, and that the pubcast format was preferred. These results are not statistically rigorous, but they do suggest that the medium is of value where limited time can be dedicated to studying an article. It is estimated that scanning an article abstract can be done in about a minute; a careful reading through a full article takes 2 or 3 h; but a pubcast may take 5 or 10 min, with the potential benefit of direct access to the author's emphasis and interpretation.

To maximise the benefit of the pubcast, tools have been provided for authors to synchronise their video or audio content against the highlighting of specific content within the article. The user interface is intuitive, timeline-based and with clear visual feedback; however, it is only used in about 50% of cases – perhaps because its perceived value is low, perhaps because the web tools are still too cumbersome; but perhaps also because authors feel their time is better spent on preparing their next article.

Of course the technology is still new, and even enthusiastic adopters are on a learning curve. But there is as yet no clear answer to the question of whether such new ways of communicating science will change the way scientists work, or indeed how they think; however, it is likely that whatever changes do become established will owe more to the volume of publication than to the appearance of new media types.

Among the changes driven by increasing volume will surely be new forms of rapid communication. New media types will certainly facilitate capturing, disseminating and archiving more aspects of the intellectual memory of a working scientist or a laboratory group, and the publisher of the future should give thought to the best ways of handling these new opportunities.

Matt Day, presenting *The nature of scholarly communication in a new century* in association with **Timo Hannay**, considered how the Nature Publishing Group continued to be guided by *Nature*'s founding editorial in its mission *to place before the general public the grand results of Scientific Work and Scientific Discovery; and to urge the claims of Science to a more general recognition in Education and in Daily Life; and ... [to aid scientists] ... by affording them an opportunity of discussing the various Scientific questions which arise from time to time*.

Beyond the growing stable of journal titles and their web sites, the online medium provides many opportunities for informal communication and networking. Nature Network has been established as a professional networking website for scientists, providing discussion groups, forums and blogs that can be organised on the level of individuals, laboratories, departments, institutions or subjects.

Nature Precedings provides a platform for posting and discussing pre-publication research and preliminary findings. Connotea and Scintilla have been developed to aggregate, share and discover references and science news items. The new Scitable web site, described as ‘a collaborative learning space for science’, represents a recent entry into the teaching and learning environment by the recently formed Nature Education division, and provides facilities for groups of students and faculty to collaborate live and offline to teach and learn complex concepts.

In 2006 *Nature* experimented with a web-based trial to explore the level of interest in community peer review of manuscripts submitted for publication. The take-up in the trial was modest. Although some manuscripts received significant comments, in no case was the journal's editorial decisions modified by community input.

Nature is also experimenting with different online presentation models for conventional research articles, sharing some features with the Cell Press ‘Article of the Future’ approach, including semantic markup and rich hyperlinking to glossaries, structure databases and other external resources. Nature Publishing Group is active in building subject-specific collaborative gateways and databases, and it is providing machine-readable interfaces for search and discovery of content within the journals and associated data stores. Interest is also developing in the growth of hand-held devices, such as mobile phones, which can not only provide web access to journal articles, but are increasingly capable of downloading and functioning in the implementation of experimental protocols. Audio and video content is also appearing on the *Nature* website and in dedicated YouTube channels.

6. Concluding remarks

The ICSTI 2010 Winter Meeting demonstrated a wide range of publishing activities utilising the interactivity of the electronic medium. In a one-day meeting, it was clear that the topic was far from being exhausted. Indeed, the opportunities for interactivity continue to grow as information technology develops ceaselessly. At first glance, the challenge of preserving this interactivity in the record of science appears daunting.

But there are encouraging developments. Much interactivity arises through linking of different components of a publication. In many of the examples shown throughout the day, the link was to associated datasets. In other examples, multimedia enhancements, value-added thesauri or glossaries, relevant citations or related objects or concepts lie at the end of the hyperlinks. An essential component of preserving these relationships is to have a comprehensive, structured and standard description of a compound document and its constituent components. Here, the archiving infrastructure is making advances, with persistent identifier management systems, standard network interchange protocols, emerging metadata standards and – most importantly – a growing understanding of what is required. Digital library systems are maturing, with sufficient capacity, robustness and management to secure bit-level preservation of large volumes of content. However, more work is needed to ensure that access to the full richness of the scientific record can be guaranteed in future.

The introductory note to the Workshop programme began by stating 'The Web is by nature an interactive environment; yet online journals are mostly static, befitting their traditional role as a never-changing scholarly record'. Many of the presentations during the day challenged, implicitly or otherwise, that characterisation of online journals as 'static'. Indeed, scientific communication is increasingly dynamic, and the scientific record increasingly subject to constant updating and occasional revision. It is hoped that this Workshop will have raised awareness of many of the important factors that will need to be addressed in preserving that record in all its integrity, richness and dynamism.

References

[1] Adobe Systems Incorporated, *PDF Reference, sixth edition: Adobe Portable Document Format Version 1.7*, 2006, available at: http://www.adobe.com/devnet/acrobat/pdfs/pdf_reference_1-7.pdf.

[2] D. Alan, M. Borkin, A. Goodman, M. Halle, J. Kauffmann and R. Kikinis, IIC astronomical medicine project home, 2010, available at: http://am.iic.harvard.edu/.

[3] A.A. Goodman, E.W. Rosolowsky, M.A. Borkin, J.B. Foster, M. Halle, J. Kauffmann and J.E. Pineda, A role for self-gravity at multiple length scales in the process of star formation, *Nature* **457** (2009), 63–66.

[4] T. Green, We need publishing standards for datasets and data tables, *OECD Publishing White Paper*, OECD Publishing, 2009, doi: 10.1787/603233448430.

[5] J. Howe, The rise of crowdsourcing, *Wired*, 2006, available at: http://www.wired.com/wired/archive/14.06/crowds.html.

[6] *Jmol*: an open-source Java viewer for chemical structures in 3D, available at: http://www.jmol.org/.

[7] B. McMahon and R.M. Hanson, A toolkit for publishing enhanced figures, *Journal of Applied Crystallography* **41** (2008), 811–814.

[8] A.D. McNaught and A. Wilkinson, *Compendium of Chemical Terminology – The Gold Book*, 2nd edn, Blackwell Science, Oxford, 1997, available at: http://goldbook.iupac.org.

[9] S. Sun, L. Lannom and B. Boesch, RFC3650: Handle system overview, Request for comments, The Internet Society, 2003, available at: http://www.ietf.org/rfc/rfc3650.txt.

[10] UK Research Data Service, The data imperative: Managing the UK's research data for future use, 2009, available at: http://www.ukrds.ac.uk/resources/download/id/14.