Scientific research communication: the promise and current realities of enhanced publications

MacKenzie Smith, MIT Libraries

There has been an upsurge of interest from the scientific community in the creation of "enhanced publications" – Web-accessible networks of resources related to scientific research publications. Peer-reviewed journal articles (and their close cousins, white papers and technical reports) are still the gold standard for communicating the results of scientific research and experimentation, but in this age of digital Web publishing neither authors nor their readers still find them entirely sufficient. Advances in publishing technology are raising expectations: readers want interactive user interfaces for visualizing results, searching and browsing tools, collaboration tools, and linkages between text, multimedia, data, and tools to work with all of the above.

The recent report of the 2020 Science Group *Towards 2020 Science* stated that "our findings have significant implications for scientific publishing, where we believe that even near-term developments in the computing infrastructure for science which links data, knowledge, and scientists will lead to a transformation of the scientific communications paradigm.[1]" But a recent editorial in Nature asserted that "Web tools now allow data sharing and informal debate to take place alongside published papers. But to take full advantage, scientists must embrace a culture of sharing and rethink their vision of databases. [2]" The vision of the Science Commons will smooth the way to achieving just this, and the time is right to start.

Alongside these publishing trends, both scientists and their funders (notably the NSF and NIH) are beginning to take more seriously the need to archive research data in ways that support rediscovery and reuse by people other than the original creators. Data is expensive to produce and often has a useful afterlife (examples include microarray data from bioinformatics and astronomical sky survey data, among many others). But useful data archiving is difficult, and long-term preservation of research data is of unknown difficulty and expense.

Scientific research data includes many highly diverse data types and formats. Experimental research data requires subject expertise to describe, interpret, and use (of course, this was also true of journal articles). Some trends, including the inexorable transformation of data files into one or other XML-based file formats, and the steady march of computer systems towards Service-Oriented Architecture and Web Services in particular, is lowering the barriers to interoperability of technology for these different types of data.

With new options for online data publishing offered Open Access repositories run by research institutions we begin to see ways in which research articles (published or

---

[1] Towards 2020 Science, 2006, Microsoft Corporation.
http://research.microsoft.com/towards2020science/downloads/T2020S_Report.pdf

[2] Nature, 438(7068), December 1, 2005 *Let data speak to data*
http://www.nature.com/nature/journal/v438/n7068/full/438531a.html

unpublished) can evolve into networks of information about the research they describe by implementing standard means of interoperation between content repositories (e.g. library-run DSpace archives and publishers content management systems) and between types of content in them (e.g. research papers and related datasets). There are various means to accomplish this, such as using simple RDF graphs to represent the networks of digital resources, but several issues that must first be resolved, which are detailed below.

Libraries and publishers are both interested in ways of supporting this type of "enhanced publication" and some in particular are beginning to work on the elements of the publication network, the standards to identify and link them, and the tools to search and display them. Creating such publication networks would be a great leap forward for researcher in many disciplines, and begin to leverage the infrastructure afforded by the Web and the Semantic Web. Researchers are beginning to see the value of such "enhanced publications" to allow for simpler means to corroborate results, dispute conclusions, or as possible source materials for future works so that valuable data doesn't need to be re-created.

The network of resources related to scholarly articles might include related articles (from the same journal or issue, or from other journals), related e-prints, white papers or technical reports, information about people and their institutions, research data, and tools for working with all of this (for example, social bookmarking or other collaboration tools, searching tools, data manipulation and visualization tools). We have all these now, individually and separately, but no way to predictably link them all up. So what's missing from our environment that would support this?

- An *ontology* for complex digital objects that represents a shared understanding of the parts of the enhanced publication, their relationships, and their other properties or attributes. We could start with a model for journals, issues, articles, and supplementary material. We could then add a clustering model such as FRBR[3] to allow for multiple versions of the parts. Representing distinct versions, either of the articles or the underlying data, presents significant challenges since we don't yet have a handle on how to compare different copies that are available on the Web. In fact, there is a NISO working group currently considering what versions of articles are important to separately label and track (e.g. the e-print, the pre-publication version, and the final published version) and whether they should be separately identified or linked via standard identifiers such as DOIs using agencies like CrossRef.

- A system for reliably *identifying* all the parts. These identifiers need not be related to each other in any a priori way, but should be citable in standard Web environments (e.g. URIs). Every piece of information that we want to talk about should have one of these identifiers attached to it. Where data is concerned, analysis is needed to determine whether it's sufficient to identify the dataset in its entirety, or whether we need ways to identify components of the datasets (e.g. individual results in the data).

[3] http://www.ifla.org/VII/s13/frbr/frbr.pdf

- A subcomponent of the part identification problem is a way of assigning every *person and organization* a unique identifier, i.e. the author, researcher, university, research institution, etc. This is absolutely necessary in order to unambiguously cite them as part of the network and support clustering of, for example, all the articles by a particular author, or from a particular department. One way to achieve this might be to use the national authority files (e.g. the Library of Congress' Name Authority File) but provide ways to extend and enhance the data included in them to allow for new people, and new properties about them, to be added. Perhaps the OCLC Virtual International Authority File (VIAF[4]) could be leveraged for this.

- Descriptive metadata for all the parts will be necessary, so they can be found on the Web to create our networks, and so that readers can find relevant enhanced publications. Over time different practices have evolved in the publishing, library, and data archiving communities for how to describe content (e.g. publishers often use PRISM metadata, libraries tend to like Dublin Core metadata, and data archives often have custom, unique schemas to describe their contents). None of this metadata interoperates well today, but by using common data models such as RDF (perhaps with RDF/A to convert them from their native XML schemas) then interoperability in our networks can be achieved.

- Conventions are needed for structuring the original datasets that would be attached to publications, to make them easier to interpret and process with standard tools. These do exist in a few areas (e.g. SBML – Systems Biology Markup Language, or CML – Chemistry Markup Language) but are by no means common across all scientific disciplines. This is going to be a significant challenge… if data is un-normalized then its interpretation by researchers will be very difficult, if not impossible, but normalizing data will require many things: expertise in the data (protocols, instrumentation, methodology, etc.), expertise in the data model into which the data is to be converted, and, of course, agreement about what the best data model is for heterogeneous data. A graph-based data model such as RDF could be very helpful as a starting point, but there is much work still needed before we can claim true interoperability across every dataset under consideration.

- Protocols to move data around on the Web will also be necessary. Fortunately, flexible protocols already exist that can support this type of data. For example, the Atom publishing protocol[5] uses standard HTTP and XML (including RDF/XML) to support the publication of collections of Web resources in a very straightforward, widely supported manner.

[4] http://www.oclc.org/research/projects/viaf/
[5] http://www.atomenabled.org/

- Protocols for searching and data mining will be needed, such as the Open Text Mining Interface (OTMI) proposed by the Nature Publishing Group[6], and the SRU[7] or OpenSearch[8] protocols for search and retrieval of data on the Web.

- Other tools will need to integrate into these resource networks as well. For example, tools for taxonomic filtering by third parties, tools for collaboration between groups of researchers, tools for personal annotation of networks or individual resources in a network (e.g. social tagging tools like Connotea). Citation analysis tools that can take a publication and find related content on the Web using standards such as OpenURL.

- Finally, a legal framework to support of all this does not yet exist and will probably be quite complex. Mechanisms for sharing are emerging (e.g. the Creative Commons licenses, and various author/publisher agreements that support this sort of secondary publication), but are far from commonly used. Means to capture the legal status of each part of a research publication network is needed, as well as techniques to inform users of their rights to use or reuse each part of the network that comprises an enhanced publication.

As should be evident from the above discussion, many of the pieces of infrastructure that we need to build enhanced scientific research publications already exist. What has not yet happened is the knitting… taking publications that exist on publishers websites or in institutional repositories, and datasets that exist in subject-based or institutional archives, information about people and organizations that exists in Web-based databases, and tools that are provided by individual publishers, researchers, or their institutions, and beginning to create these resource networks.

Given that we're in striking distance of having the data and the infrastructure to achieve "enhanced publications", what should be the next steps?

First, we can identify a few candidate fields for enhanced publication… ideally those which are inherently data-intensive, have some degree of ability to structure their data for potential reuse, and have a culture of sharing (and the legal means to do so) already. Experiments can then be set up to take the parts of an enhanced publication and start to build the network of them, proposing solutions for each of the issues identified above. And new Web interfaces can be devised for these networks to allow researchers to explore them. The ensuing feedback will support refinement of the data model and the infrastructure to make it easier for scientists to find, read, and re-use these publication networks in their own research.

Assuming that enhanced publications will initially be created by traditional publishers and libraries as extensions of the current publishing and distribution supply chain, we can imagine a future stage of these publications that is author- or reader-supplied. For

[6] http://blogs.nature.com/wp/nascent/2006/04/open_text_mining_interface_1.html
[7] http://www.loc.gov/standards/sru/
[8] http://opensearch.a9.com/

example, linking the article to related articles or other material on the Web, merging the data with related data (i.e. “Web 2.0-style mashups” that merge different data sources to explore new connections there were not foreseen by the creators), adding more information that was created post-publication, and so on. Some of this could be supported by the emerging collaboration tools described above (e.g. social bookmarking tools like NPG’s Connotea), but more sophisticated authoring tools are clearly called for. The problem then becomes creating the right social constructs to make this useful. Authors, publishers, and libraries need to know that the “authoritative” publication is still available and clearly identified. Readers need to know what is authoritative and what isn’t, who authored what parts of the enhanced publication, and to what degree they can rely on any of the information they find there. Current publications tend to be static so that authorship, and the imprimatur of a publisher, can be clearly identified and readers are sure that what they find is an official version. This practice has already begun to erode on the Web, although we go to great lengths to make readers feel a confidence in what they find that is hard to insure in reality.

There is an enormous tension between the desire to harness the expertise and energy of readers, but without losing the quality and authoritativeness of authored, peer-reviewed work. We really need new modes of attribution for content that is included in Web publications, and tools to easily reveal them… imagine a browser that let you see “layers” of content, starting with everything, then peeling away the layer of links, metadata or other structures that were added by anonymous readers, then the layer of attributed additions (but not from the author or publisher), then the layer of additions from “trusted” experts in the discipline other than the author or publisher, then finally the original enhanced publication. The means of providing this type of attribution exist on the Semantic Web (e.g. as RDF properties) but the tools are lacking that let non-authors add their information to the publication, or allow readers to see who created what part of the publication and view only what they want too see.

In conclusion, we know that the current situation (of discrete, static, peer-reviewed print-like publications) is less than desirable and far less than what is achievable with new technology and social practices, so experimental enhanced publications are imminent. Much of the necessary infrastructure exists now to support this experimentation, but the grander vision of where the new type of publications may take us requires solving a number of problems. Each of these problems is challenging, but tractable if we get started soon.