The Shape of the Scientific Article in The Developing Cyberinfrastructure

Clifford Lynch, Coalition for Networked Information (CNI)

CTWatch QUARTERLY
August 2007

Introduction

For the last few centuries, the primary vehicle for communicating and documenting results in most disciplines has been the scientific journal article, which has maintained a strikingly consistent and stable form and structure over a period of more than a hundred years now; for example, despite the much-discussed shift of scientific journals to digital form, virtually any article appearing in one of these journals would be comfortably familiar (as a literary genre) to a scientist from 1900. E-science represents a significant change, or extension, to the conduct and practice of science; this article speculates about how the character of the scientific article is likely to change to support these changes in scholarly work. In addition to changes to the nature of scientific literature that facilitate the documentation and communication of e-science, it's also important to recognize that active engagement of scientists with their literature has been, and continues to be, itself an integral and essential part of scholarly practice; in the cyberinfastructure environment, the nature of engagement with, and use of, the scientific literature is becoming more complex and diverse, and taking on novel dimensions. This changing use of the scientific literature will also cause shifts in its evolution, and in the practices of authorship, and I will speculate about those as well here.

A few general comments should be made at the outset. First, I recognize that it is dangerous to generalize across a multiplicity of scientific disciplines, each with their own specialized disciplinary norms and practices, and I realize that there are ample counterexamples or exceptions to the broad trends discussed here; but, at the same time, I do believe that it is possible to identify broad trends, and that there is value in analyzing them. Second, as with all discussions of cyberinfrastructure and e-science, many of the developments and issues are relevant to scholarly work spanning medicine, the biological and physical sciences, engineering, the social sciences, the humanities, and even the arts, as is suggested by the increasingly common use of the more inclusive term "e-research" rather than "e-science" in appropriate contexts. I have focused here on the sciences and engineering, but much of the discussion has broader relevance.

Finally, it's crucial to recognize that the changes to the nature of scholarly communication and the scientific article are not being driven simply or solely by technological determinism as expressed through the move to e-science. There are broad social and political forces at work as well, independent of, but often finding common cause or at least compatibility with, e-science developments; in many cases, the transfigured economics and new capabilities of global high-performance networking and other information technologies are, for the first time, making it possible for fundamental shifts in the practices and structures of scholarly communication to occur, and thus setting the stage for political demands that these new possibilities be realized. Because the same technical and economic drivers have fueled much of the commitment to e-science, these other exogenous factors that are also shaping the future of scholarly communication are often, at least in my view, overly identified with e-science itself. Notable and important examples include the movements towards open access to scientific literature; movements towards open access to underlying scientific data; demands (particularly in the face of some recent high-profile cases of scientific fraud and misconduct) for greater accountability and auditibility of science through structures and practices that facilitate the verification, reproducibility and re-analysis of scientific results; and efforts to improve the collective societal return on investment in scientific research through a recognition of the lasting value of much scientific data and the way that the investment it represents can be amplified by disclosure, curation

and facilitation of reuse. Note that in the final area the investments include but go beyond the financial; consider the human costs of clinical trials, for example.

Scientific Articles and their Relationships to Data

The vast majority of scientific articles present and/or analyze data. (Yes, in mathematics, and some parts of theoretical physics, they do something else, and if, when and how this particular sub-genre of writings will be transformed is a fascinating question that deserves an extensive discussion in its own right. But that is another question, for another time.) As this data becomes more complex, more extensive, more elaborate, more community-based, more mediated by software, the relationships between articles and the data upon which they are based is becoming more complex and more variable. And recognize that implicit in these relationships are a whole series of disciplinary norms and supporting organizational and technical cyberinfrastructure services.

To what extent should articles *incorporate* the data they present and analyze, and to what extent should they simply *reference* that data? The issues here are profoundly complex. First, there's the question of whether the data is original and being presented for the first time; certainly it is commonplace to draw upon, compare, combine and perhaps reinterpret data presented in earlier articles or otherwise made available, and here the right approaches would presumably be citation or similar forms of reference. Repeated publication of the same data is clearly undesirable.

For newly publicized data there are a range of approaches. Some journals offer to accept it as "supplementary materials" that accompany the article, but often with very equivocated commitments about preserving the data or the tools to work with it, as opposed to the article proper. Not all journals offer this as an option, and some place constraints on the amount of data they will accept, or on the terms of access to the data (e.g., subscribers only).

For certain types of data, specific communities — for example crystallographers, astronomers, and molecular biologists – have established norms, enforced by the editorial policies of their journals, which call for deposit of specific types of data within an international disciplinary system of data repositories, and have the article make reference to this data by an accession identifier assigned upon deposit in the relevant repository. Clearly, this works best when there are well agreed-upon structures for specific kinds of data that occur frequently (genomic sequencing, observations of the sky, etc.); it also assumes an established, trustworthy and sustainable disciplinary repository system. Indeed, we have already seen the emergence of what might be characterized as a "stub article" that in effect announces the deposit of an important new dataset in a disciplinary repository and perhaps provides some background on its creation, but offers little analysis of the data, leaving that to subsequent publications. This allows the compilers of the dataset to have their work widely recognized, acknowledged, and cited within the traditional system familiar to tenure and promotion committees.

Another alternative is for the authors to store the underlying data in an institutional repository. While in some ways this is less desirable than using a disciplinary repository (due to potentials for economies of scale, easy centralized searching of material on a disciplinary basis, and for the development and maintenance of specialized discipline-specific software tools, for example) the institutional repository may be the *only* real option available for many researchers and for many types of data. Note that one function (and obligation) of each institutional repository is to provide the depositing researcher with a persistent accession identifier that can be used to reference the data.

Recognize that over time individual researchers may move from institution to institution and will ultimately die; technical systems evolve, organizations change mission and responsibilities, and funding models and funding agency interests and priorities shift — any of which can cause archived data to have to be migrated from one place to another or reorganized. The ability to resolve identifiers, to go from citation to data, is

highly challenging when considered across long time horizons. The research library community collectively has made a clear commitment to the long-term preservation of access to the traditional scientific literature; the assumption of similar ultimate responsibility for scientific and scholarly data today is still highly controversial.

Just because a dataset has been deposited into a repository does not automatically mean that other researchers (or indeed the public broadly) can have access to it. This is a question ultimately of disciplinary norms, of requirements imposed by funding agencies, of university policies, and of law and public policy. What the e-science environment does is to make these policies and norms much more explicit and transparent, and, I believe, to advance a bias that encourages more rather than less access and sharing. And there is still work to be done on mechanisms and legal frameworks — for example, the analogs of the Creative Commons type licenses for datasets are under development by Science Commons and others, but are at a much less mature stage than those used for journal articles themselves, with part of the problem being that the copyright framework that governs articles is much more consistent globally than laws establishing rights in datasets and databases.[1] Also to be recognized here are certain trends in the research community – most notably university interests in technology transfer as a revenue stream, and the increasing overreach of some Institutional Review Boards in restricting the collection, preservation and dissemination of materials dealing in any way with human subjects - which run very much counter to the bias towards greater openness.

Setting aside the broad issue of the future of peer review, a particularly interesting set of questions involves the relationships between traditional peer review and the data that underlies an article under review. It's often unclear the extent to which peer review of an article extends to peer review of the underlying data; even when the policies of a journal are explicit on this, I think it's likely readers don't have well-established expectations. Will there be a requirement that reviewers have access to underlying data as part of the review process, even if this data may not be immediately and broadly available when the article is published? A recent examination of editorial and refereeing policy by *Science* in the wake of a major incident of data falsification suggests that at least some journals may take a more aggressive and indeed even somewhat adversarial position with the authors of particularly surprising or high-visibility articles.[2] And post-publication, there's a very formalized means of correcting errors in published articles (or even withdrawing them) that's now integrated into the online journal delivery systems (though not necessarily other open-access versions of articles that may be scattered around the net). Data correction, updating and maintenance take place (if at all) through separate curatorial mechanisms that are not synchronized to those for managing the article literature.

Visual Presentation of Data in Scientific Articles

The chemist Peter Murray-Rust speaks passionately and powerfully of the ways in which traditional presentations of information in scientific articles, such as graphs and charts, actually obscure or destroy data, invoking scenes of readers employing rulers to try to estimate the actual values of coordinates of points in a graph. [3] Clearly, in a digital environment, it would be much better to be able to move directly and easily between the underlying table of numerical values and their graphical representation, for example. Note such a table really is, in my view, intellectually an integral part of the article rather than underlying data (it may in fact be a complex derivative or reduction of the underlying data, or it might duplicate it). While the example of a two-dimensional graph is fairly straightforward, one can imagine a wide range of more specialized visualizing tools operating on various forms of structured data.

To me, resolving this problem implies a somewhat "thicker" layer of software mediating between the machine-readable representation of articles in the cyberinfrastructure environment and the human reader. Today, articles are most typically delivered to readers in very unexpressive, semantically limited forms such as PDF or HTML, which are rendered by a PDF viewer or a web browser, respectively. As we build out the collaboratories and virtual workspaces to house the activities of our virtual organizations within the cyberinfrastructure, I hope that we will see a new generation of viewing and annotation tools deployed,

presumably working on semantically rich XML document representations that will allow us to move beyond the kind of practices that Peter Murray-Rust so appropriately highlights as impeding the progress of scholarship.

The issues here are not limited to graphs and charts. Let me just give two other examples to illustrate the range of potential opportunities.

The astronomer Robert Hanisch gives a talk[4] that includes this compelling example: An article includes an image of an astronomical object captured at a specific frequency range; a reader would like to place this object into the context of the available archive of astronomical observations and see what additional observations might be available at other wavelengths, for example. The process of manually re-creating what is, in effect, the digital provenance of the published image proves to be quite arduous, though once the source image in the archive context is established, it's easy enough to check for the availability of additional imagery for the same region. Clearly, it would be very desirable to generate the trail of digital provenance as the image is prepared for publication, and to make an appropriate representation of that provenance available in the article along with the image to facilitate exactly the kinds of exploration Hanisch describes.

Finally, we are coming to a more sophisticated understanding of photography. A photographic print, or its reproduction on a printed page (or, indeed, a digital simulacrum of a printed page), is in effect a *reduction* of the image stored in a photographic negative or, more recently, an image dataset captured by a digital camera. Adjustments in focus, dynamic range, contrast, and similar parameters can yield radically different images from the same dataset. Here, again, it would clearly be desirable to have software tools that allow one to toggle between the rendering of an image dataset and the underlying dataset, and to be able to manipulate the key parameters that control the rendering.

All of these examples share common themes and raise common questions. The source article becomes more richly structured and exposes its semantics more explicitly. At the same time, the article becomes more and more intractable for humans to read without the correct set of mediating software. This means that the requisite mediating software must be highly reliable, simple to use, and ubiquitously available. Specialized software for individual journals or individual publishers will not, in my view, reach the necessary critical mass.

It's easy to underestimate the problem of maintaining the level of quality and flexibility inherent in today's visually oriented presentations. Consider the humble graph — it's not just a table of coordinate pairs; it has a caption, labels and markings on the axes, a default scale, perhaps a layer of annotations on some of the regions of the graph. We don't want to lose any of these capabilities.

In terms of deployability and acceptance there are also questions about the potentially increased workload involved in preparing this new generation of scientific articles. While most publisher databases are already using XML based representations of articles internally and are carrying an increasing amount of tagging that can be put to good use, a careful analysis of the distribution of responsibilities between authors and publishers in areas such as the preparation of various kinds of "illustrations" is called for, as well as consideration of the tools that might be available to help the author.

Scientific Literature that is Computed Upon, Not Merely Read by Humans

In the previous section, we explored a few of the ways in which human readers may expand and extend their interactions with the scientific literature through the mediation of a new generation of software. But the use of the corpus of scientific literature is already changing in other ways as well: not only do human beings read (and interact with) articles from the scientific literature *one article at a time*, but we are also seeing the deployment of software that computes upon the entire corpus of scientific literature (perhaps restricted by discipline, and perhaps also federated with proprietary and/or unpublished literature or auxiliary databases).

Such computation includes not only the now familiar and commonplace indexing by various search engines, but also computational analysis, abstraction, correlation, anomaly identification and hypothesis generation that is often termed "data mining" or "text mining."

The implications of this shift are extensive and complex, but I want to sketch a few implications here. First, there will be greater demand for the availability of scientific literature corpora as part of the cyberinfrastructure, and for these corpora to be available in such a way — both technically and in terms of licensing, legal and economic arrangements — so as to facilitate ongoing computation on the literature by groups of collaborating researchers, including groups ("virtual organizations") assembled often fairly casually from across multiple institutions. The barriers here are formidable: most commonly, access arrangements for publisher-controlled literature are established on an institutional basis; these licenses often specifically prohibit large-scale duplication of the text corpora for this kind of computational use; and today most publishers do not provide technical provisions for arbitrary computation of the type envisioned here.

More important to the changing nature of the individual article as opposed to the literature as a whole, the computational techniques that are applied to the current literature base make extensive use of heuristics (as well as various auxiliary databases, dictionaries, ontologies and other resources). Basically, they use algorithms to guess (with increasingly good accuracy) whether "Washington" in a given bit of text refers to a person, a state, or a city (and if so which one), whether something is the name of a gene, a chemical compound, a species, or other entity of interest. As we create new literature going forward, it makes sense to specifically tag some of these terms of interest to allow more accurate computation on the articles. Clearly, also, there are interesting possibilities of retrospectively tagging older literature, or even running the current best heuristics to provisionally tag the older literature, and then selectively (and perhaps collaboratively) applying human analysis to review provisional tags that are most suspect (Greg Crane and his colleagues at the Perseus Project have run some fascinating pilots of this type in doing markup of classical texts). The questions here are what entities merit tagging, how standards are set for tagging such entities, and what combination of author and publisher should take responsibility for actually doing the markup? There are delicate issues about how to manage the evolution of the tagging over time and also how to manage it across disciplines in such a way to facilitate interdisciplinary work. There's a difference between viewing the presence of tags as conclusive positive information and being able to count on the absence of a tag as conclusive negative information, for example.

Paths of Change and Adoption

What's sure to happen? What is contingent or uncertain? What's already happening?

The linkages between articles and underlying data are a very real and active (though far from fully settled) area of development today. All the models — supplementary materials as part of articles, disciplinary repositories and institutional repositories — are in use today.

The *problems* raised by visual displays of data in scientific articles are more widely recognized, and frustrate more readers each day. Some problems — graphs for example — are easily solved in prototypes. Others, such as the appropriate way to express data provenance, remain active and challenging research areas. But the real challenges here are about deployment, scale, adoption, and standards. We have decades of examples of systems that support the creation and use of digital documents that are far more powerful and expressive than those in common today, but for a wide variety of reasons these systems never reached critical mass in deployment and adoption, which has proven hugely difficult to predict or ensure at Internet scale.

Text mining is a reality today, at least on a limited basis, and producing some results of real value. Here, I suspect, the barriers to progress will be more around business models for those journals that aren't open access (some open access journals actually package up a compressed archive of all their articles and invite

interested parties to simply copy the files and compute away; clearly this is not going to be as straightforward for a commercial publisher).

One thing is clear: in the cyberinfrastructure, the future of the article will be shaped not only by the requirements to describe changing scientific practices, but also by the changing nature of the authoring and usage environments for the scholarly literature that describes this changing science.

Further Reading

There are now a wealth of publications dealing with data and e-science; particularly valuable is the article "The Data Deluge" by Hey & Trefethen;[5] the NSF/NSB report on Long-Lived Datasets;[6] the report of the NSF/ARL workshop "To Stand the Test of Time;"[7] the current NSF Office of Cyberinfrastructure vision document,[8] and, a UK perspective.[9] On institutional repositories see the Lynch article in the *ARL Bimonthly Report*;[10] for more information on the implications of literature mining, see for example the article in the book *Open Access: Key Strategic, Technical, and Economics Aspects*.[11]

[1]Guide to Open Data Licensing - http://okfn.org/wiki/OpenDataLicensing
[2]See Donald Kennedy's editorial at http://www.sciencemag.org/cgi/content/summary/314/5804/1353 and the report itself at http://www.sciencemag.org/cgi/content/full/314/5804/1353/DC1
[3]See Peter Murray-Rust's Blog, http://wwmm.ch.cam.ac.uk/blogs/murrayrust/, and the various publications and presentations referenced there.
[4]See, for example, http://www.arl.org/sparc/meetings/ala06/ for PowerPoint and audio.
[5]Hey, T., Trefethen, A. "The Data Deluge: An e-Science Perspective," *Grid Computing: Making the Global Infrastructure a Reality*. (Chichester: Wiley, 2003), pp. 809-24. http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf
[6]National Science Board. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," National Science Foundation, 2005. http://www.nsf.gov/pubs/2005/nsb0540/start.jsp
[7]Association of Research Libraries. "To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering," *Association of Research Libraries*, 2006. http://www.arl.org/pp/access/nsfworkshop.shtml
[8]National Science Foundation Office of Cyberinfrastructure — http://www.nsf.gov/oci
[9]Lyon, L. "Dealing with Data: Roles, Rights, Responsibilities and Relationships (Consultancy report)," UKOLN and the Joint Information Systems Committee (JISC), 2006. http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx
[10]Lynch, C.A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report*, Vol 226 (February 2003), pp. 1-7. http://www.arl.org/newsltr/226/ir.html
[11]Lynch, C.A. "Open Computation: Beyond Human-Reader-Centric Views of Scholarly Literatures," *Open Access: Key Strategic, Technical and Economic Aspects*. Neil Jacobs (Ed.), (Oxford: Chandos Publishing, 2006), pp. 185-193. http://www.cni.org/staff/cliffpubs/OpenComputation.pdf

URL to article: http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/