All Aboard: Toward a Machine-Friendly Scholarly Communication System

HERBERT VAN DE SOMPEL
Los Alamos National Laboratory

CARL LAGOZE
Cornell University

"The current scholarly communication system is nothing but a scanned copy of the paper-based system."

This sentence, which we used for effect in numerous conference presentations and eventually fully articulated in a 2004 paper [1], is still by and large true. Although scholarly publishers have adopted new technologies that have made access to scholarly materials significantly easier (such as the Web and PDF documents), these changes have not realized the full potential of the new digital and networked reality. In particular, they do not address three shortcomings of the prevailing scholarly communication system:

- Systemic issues, particularly the unbreakable tie in the publication system between the act of making a scholarly claim and the peer-review process
- Economic strains that are manifested in the "serials crisis," which places tremendous burdens on libraries
- Technical aspects that present barriers to an interoperable information infrastructure

We share these concerns about the state of scholarly communication with many others worldwide. Almost a decade ago, we collaborated with members of that global community to begin the Open Archives Initiative (OAI), which had a significant impact on the direction and pace of the Open Access movement. The OAI Protocol for Metadata Harvesting (OAI-PMH) and the concurrent OpenURL effort reflected our initial focus on the process-related aspects of scholarly communication. Other members of the community focused on the scholarly content itself. For example, Peter Murray-Rust addressed the flattening of structured, machine-actionable information (such as tabular data and data points underlying graphs) into plain text suited only for human consumption [2].

A decade after our initial work in this area, we are delighted to observe the rapid changes that are occurring in various dimensions of scholarly communication. We will touch upon three areas of change that we feel are significant enough to indicate a fundamental shift.

AUGMENTING THE SCHOLARLY RECORD WITH A MACHINE-ACTIONABLE SUBSTRATE

One motivation for machine readability is the flood of literature that makes it impossible for researchers to keep up with relevant scholarship [3]. Agents that *read* and *filter* on scholars' behalf can offer a solution to this problem. The need for such a mechanism is heightened by the fact that researchers increasingly need to absorb and process literature across disciplines, connecting the dots and combining existing disparate findings to arrive at new insights. This is a major issue in life sciences fields that are characterized by many interconnected disciplines (such as genetics, molecular biology, biochemistry, pharmaceutical chemistry, and organic chemistry). For example, the lack of uniformly structured data across related biomedical domains is cited as a significant barrier to translational research—the transfer of discoveries in basic biological and medical research to application in patient care at the clinical level [4].

Recently, we have witnessed a significant push toward a machine-actionable representation of the knowledge embedded in the life sciences literature, which supports reasoning across disciplinary boundaries. Advanced text analysis techniques are being used to extract entities and entity relations from the existing literature, and shared ontologies have been introduced to achieve uniform knowledge representation. This approach has already led to new discoveries based on information embedded in literature that was previously readable only by humans. Other disciplines have engaged in similar activities, and some initiatives are allowing scholars to start publishing entity and entity-relation information at the time of an article's publication, to avoid the post-processing that is current practice [5].

The launch of the international Concept Web Alliance, whose aim is to provide a global interdisciplinary platform to *discuss, design, and potentially certify solutions for the interoperability and usability of massive, dispersed, and complex data,* indicates that the trend toward a machine-actionable substrate is being taken seriously by both academia and the scholarly information industry. The establishment of a machine-actionable representation of scholarly knowledge can help scholars and learners deal with information abundance. It can allow for new discoveries to be made by reasoning over a body of established knowledge, and it can increase the speed of discovery by helping scholars to avoid redundant research and by revealing promising avenues for new research.

INTEGRATION OF DATASETS INTO THE SCHOLARLY RECORD

Even though data have always been a crucial ingredient in scientific explorations, until recently they were not treated as first-class objects in scholarly communication, as were the research papers that reported on findings extracted from the data. This is rapidly and fundamentally changing. The scientific community is actively discussing and exploring implementation of all core functions of scholarly communication—*registration, certification, awareness, archiving, and rewarding* [1]—for datasets.

For example, the Data Pyramid proposed in [6] clearly indicates how attention to trust *(certification)* and digital preservation *(archiving)* for datasets becomes vital as their application reaches beyond personal use and into the realms of disciplinary communities and society at large. The international efforts aimed at enabling the sharing of research data [7] reflect recognition of the need for an infrastructure to facilitate discovery of shared datasets *(awareness).* And efforts aimed at defining a standard citation format for datasets [8] take for granted that they are primary scholarly artifacts. These efforts are motivated in part by the belief that researchers should gain credit (be *rewarded)* for the datasets they have compiled and shared. Less than a decade or so ago, these functions of scholarly communication largely applied only to the scholarly literature.

EXPOSURE OF PROCESS AND ITS INTEGRATION INTO THE SCHOLARLY RECORD

Certain aspects of the scholarly communication process have been exposed for a long time. Citations made in publications indicate the use of prior knowledge to generate new insights. In this manner, the scholarly citation graph reveals aspects of scholarly dynamics and is thus actively used as a research focus to detect connections between disciplines and for trend analysis and prediction. However, interpretation of the scholarly citation graph is often error prone due to imperfect manual or automatic citation extraction approaches and challenging author disambiguation issues. The coverage of citation graph data is also partial (top-ranked journals only or specific disciplines only), and unfortunately the most representative graph (Thomson Reuters) is proprietary.

The citation graph problem is indicative of a broader problem: there is no unambiguous, recorded, and visible trace of the evolution of a scholarly asset through the system, nor is there information about the nature of the evolution. The problem is that relationships, which are known at the moment a scholarly asset goes through a step in a value chain, are lost the moment immediately after, in many cases forever. The actual dynamics of scholarship—the interaction/connection between assets, authors, readers, quality assessments about assets, scholarly research areas, and so on—are extremely hard to recover after the fact. Therefore, it is necessary to establish a layer underlying scholarly communication—a grid for scholarly communication that records and exposes such dynamics, relationships, and interactions.

A solution to this problem is emerging through a number of innovative initiatives that make it possible to publish information about the scholarly process in machine-readable form to the Web, preferably at the moment that events of the above-described type happen and hence, when all required information is available.

Specific to the citation graph case, the Web-oriented citation approach explored by the CLADDIER project demonstrates a mechanism for encoding an accurate, crawlable citation graph on the Web. Several initiatives are aimed at introducing author identifiers [9] that could help establish a less ambiguous citation graph. A graph augmented with citation semantics, such as that proposed by the Citation Typing Ontology effort, would also reveal why an artifact is being cited—an important bit of information that has remained elusive until now [10].

Moving beyond citation data, other efforts to expose the scholarly process include projects that aim to share scholarly usage data (the process of paying attention to scholarly information), such as COUNTER, MESUR, and the bX scholarly recommender service. Collectively, these projects illustrate the broad applicability of this type of process-related information for the purpose of collection development, computation of novel metrics to assess the impact of scholarly artifacts [11], analysis of current research trends [12], and recommender systems. As a result of this work, several projects in Europe are pursuing technical solutions for sharing detailed usage data on the Web.

Another example of process capture is the successful myExperiment effort, which provides a social portal for sharing computational workflow descriptions. Similar efforts in the chemistry community allow the publication and sharing of laboratory notebook information on the Web [13].

We find these efforts particularly inspiring because they allow us to imagine a next logical step, which would be the sharing of provenance data. Provenance data reveal the history of inputs and processing steps involved in the execution of workflows and are a critical aspect of scientific information, both to establish trust in the veracity of the data and to support the reproducibility demanded of all experimental science. Recent work in the computer science community [14] has yielded systems capable of maintaining detailed provenance information within a single environment. We feel that provenance information that describes and interlinks workflows, datasets, and processes is a new kind of process-type metadata that has a key role in network-based and data-intensive science—similar in importance to descriptive metadata, citation data, and usage data in article-based scholarship. Hence, it seems logical that eventually provenance information will be exposed so it can be leveraged by a variety of tools for discovery, analysis, and impact assessment of some core products of new scholarship: workflows, datasets, and processes.

LOOKING FORWARD

As described above, the scholarly record will emerge as the result of the intertwining of traditional and new scholarly artifacts, the development of a machine-actionable scholarly knowledge substrate, and the exposure of meta-information about the scholarly process. These facilities will achieve their full potential only if they are grounded in an appropriate and interoperable cyberinfrastructure that is based on the Web and its associated standards. The Web will not only contribute to the sustainability of the scholarly process, but it will also integrate scholarly debate seamlessly with the broader human debate that takes place on the Web.

We have recently seen an increased Web orientation in the development of approaches to scholarly interoperability. This includes the exploration or active use of uniform resource identifiers (URIs), more specifically HTTP URIs, for the identification of scholarly artifacts, concepts, researchers, and institutions, as well as the use of the XML, RDF, RDFS, OWL, RSS, and Atom formats to support the representation and communication of scholarly information and knowledge. These foundational technologies are increasingly being augmented with community-specific and community-driven yet compliant specializations. Overall, a picture is beginning to emerge in which all constituents of the new scholarly record (both human and machine-readable) are published on the Web, in a manner that complies with general Web standards and community-specific specializations of those standards. Once published on the Web, they can be accessed, gathered, and mined by both human and machine agents.

Our own work on the OAI Object Reuse & Exchange (OAI-ORE) specifications [15], which define an approach to identifying and describing eScience assets that are aggregations of multiple resources, is an illustration of this emerging Web-centric cyberinfrastructure approach. It builds on core Web technologies and also adheres to the guidelines of the Linked Data effort, which is rapidly emerging as the most widespread manifestation of years of Semantic Web work.

When describing this trend toward the use of common Web approaches for scholarly purposes, we are reminded of Jim Gray, who insisted throughout the preliminary discussions leading to the OAI-ORE work that any solution should leverage common feed technologies—RSS or Atom. Jim was right in indicating that many special-purpose components of the cyberinfrastructure need to be developed to meet the requirements of scholarly communication, and in recognizing that others are readily available as a result of general Web standardization activities.

As we look into the short-term future, we are reminded of one of Jim Gray's well-known quotes: "May all your problems be technical." With this ironic comment, Jim was indicating that behind even the most difficult technical problems lies an even more fundamental problem: assuring the integration of the cyberinfrastructure into human workflows and practices. Without such integration, even the best cyberinfrastructure will fail to gain widespread use. Fortunately, there are indications that we have learned this lesson from experience through the years with other large-scale infrastructure projects such as the Digital Libraries Initiatives. The Sustainable Digital Data Preservation and Access Network Partners (DataNet) program funded by the Office of Cyberinfrastructure at the U.S. National Science Foundation (NSF) has recently awarded funding for two 10-year projects that focus on cyberinfrastructure as a sociotechnical problem—one that requires both knowledge of technology and understanding of how the technology integrates into the communities of use. We believe that this wider focus will be one of the most important factors in changing the nature of scholarship and the ways that it is communicated over the coming decade.

We are confident that the combination of the continued evolution of the

Web, new technologies that leverage its core principles, and an understanding of the way people use technology will serve as the foundation of a fundamentally rethought scholarly communication system that will be friendly to both humans and machines. With the emergence of that system, we will happily refrain from using our once-beloved scanned copy metaphor.

REFERENCES

[1] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner, “Rethinking Scholarly Communication: Building the System that Scholars Deserve,” *D-Lib Mag.*, vol. 10, no. 9, 2004, www.dlib.org/dlib/september04/vandesompel/09vandesompel.html.

[2] P. Murray-Rust and H. S. Rzepa, “The Next Big Thing: From Hypermedia to Datuments,” *J. Digit. Inf.*, vol. 5, no. 1, 2004.

[3] C. L. Palmer, M. H. Cragin, and T. P. Hogan, “Weak information work in scientific discovery,” *Inf. Process. Manage.*, vol. 43, no. 3., pp. 808–820, 2007, doi: 10.1016/j.ipm.2006.06.003.

[4] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, K. Forsberg, Y. Gao, V. Kashyap, J. Kinoshita, J. Luciano, M. S. Marshall, C. Ogbuji, J. Rees, S. Stephens, G. T. Wong, E. Wu, D. Zaccagnini, T. Hongsermeier, E. Neumann, I. Herman, and K. H. Cheung, “Advancing translational research with the Semantic Web,” *BMC Bioinf.*, vol. 8, suppl. 3, p. S2, 2007, doi: 10.1186/1471-2105-8-S3-S2.

[5] D. Shotton, K. Portwin, G. Klyne, and A. Miles, “Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article,” *PLoS Comput.* Biol., vol. 5, no. 4, p. e1000361, 2009, doi: 10.1371/journal.pcbi.1000361.

[6] F. Berman, “Got data?: a guide to data preservation in the information age,” *Commun. ACM*, vol. 51, no. 12, pp. 50–56, 2008, doi: 10.1145/1409360.1409376.

[7] R. Ruusalepp, “Infrastructure Planning and Data Curation: A Comparative Study of International Approaches to Enabling the Sharing of Research Data,” JISC, Nov. 30, 2008, www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf.

[8] M. Altman and G. King, “A Proposed Standard for the Scholarly Citation of Quantitative Data,” *D-Lib Magazine*, vol. 13, no. 3/4, 2007.

[9] M. Enserink, “Science Publishing: Are You Ready to Become a Number?” *Science*, vol. 323, no. 5922, 2009, doi: 10.1126/science.323.5922.1662.

[10] N. Kaplan, “The norm of citation behavior,” *Am. Documentation*, vol. 16. pp. 179–184, 1965.

[11] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute, “A Principal Component Analysis of 39 Scientific Impact Measures,” *PLoS ONE*, vol. 4, no. 6, p. e6022, 2009, doi: 10.1371/journal.pone.0006022.

[12] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, and L. Balakireva, “Clickstream Data Yields High-Resolution Maps of Science,” *PLoS ONE*, vol. 4, no. 3, p. e4803, 2009, doi: 10.1371/journal.pone.0004803.

[13] S. J. Coles, J. G. Frey, M. B. Hursthouse, M. E. Light, A. J. Milsted, L. A. Carr, D. De Roure, C. J. Gutteridge, H. R. Mills, K. E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke, and M. Day, “An e-science environment for service crystallography from submission to dissemination,” *J. Chem. Inf. Model.*, vol. 46, no. 3, 2006, doi: 10.1021/ci050362w.

[14] R. Bose and J. Frew, “Lineage retrieval for scientific data processing: a survey,” *ACM Comput. Surv.* (CSUR), vol. 37, no. 1, pp. 1–28, 2005, doi: 10.1145/1057977.1057978.

[15] H. Van de Sompel, C. Lagoze, C. E. Nelson, S. Warner, R. Sanderson, and P. Johnston, “Adding eScience Publications to the Data Web,” *Proc. Linked Data on the Web 2009*, Madrid.