Learned Publishing, 21, 29–38
doi: 10.1087/095315108X254476

Data, disciplines, and scholarly publishing

Christine L. BORGMAN
University of California, Los Angeles

ABSTRACT. Data are becoming an essential product of scholarship, complementing the roles of journal articles, papers, and books. Research data can be reused to ask new questions, to replicate studies, and to verify research findings. Data become even more valuable when linked to publications and other related resources to form a value chain. Types and uses of data vary widely between disciplines, as do the online availability of publications and the incentives of scholars to publish their data. Publishers, scholars, and librarians each have roles to play in constructing a new scholarly information infrastructure for e-research. Technical, policy, and institutional components are maturing; the next steps are to integrate them into a coherent whole. Achieving a critical mass of datasets in public repositories, with links to and from publisher databases, is the most promising solution to maintaining and sustaining the scholarly record in digital form.

This article is an extended version of the author's presentation, 'How Data Differ by Discipline', at the ALPSP International Scholarly Communications Conference: The Transformation of Research Communication, London, 13 April 2007.

Photo: Ed Swinden

Christine L. Borgman

Introduction

While publications continue to be essential products of scholarship, data are becoming part of the scholarly record in their own right. Improving the ability to use and reuse research data is a central goal of e-science and cyber-infrastructure programs in the UK, US, and elsewhere.[1,2] However, notions of what it means to 'publish' data are far less mature than are notions of publishing journal articles and books.[3] Definitions of 'data' vary widely between disciplines and between individual research specialties. A new conversation is under way among scholars, librarians, and publishers about which data will be of most future use to whom, and how to capture, preserve, curate, and make those data accessible over the short and long term. In 2007, the ALPSP International Scholarly Communications Conference scheduled an entire session on data publishing, reflecting the growing concerns of publishers about their roles and responsibilities in providing access to research data.

Data-intensive scholarship

Cyberinfrastructure, *e-science*, and *e-research* are shorthand terms for new forms of data-intensive, information-intensive, collaborative, distributed forms of scholarship. Funding agencies in the US, Europe, Asia, and elsewhere are making massive investments in research on new tools, services, and collaboration frameworks that support computationally driven research.[1,2,4] While the sciences have led other disciplines in e-research, they are not alone. The e-social sciences now have their own conference, and the requirements for e-research in the humanities are the subject of several influential policy reports.[5–7]

Today's scholarship is distinguished by the extent to which its practices rely on the generation, dissemination, and analysis of data. These practices are themselves distinguished both by the massive scale of data production and by the global dispersion of data resources. The rates of data generation in most fields are expected to increase even faster with new forms of instrumentation such as embedded sensor networks in the sciences, mass digitization of texts in the humanities, and the digitized traces of human behavior available to the social sciences. Digital scholarship is spawning its own new set of research questions about how to manage the 'data deluge', about the changing nature of scholarly practices, and about economic and policy models to sustain access to research data.[3,8–10]

from a social perspective, however, data is not at all a simple or straightforward concept

Role of data in the value chain

Research data often have great value in and of themselves. They can be used to leverage research investments, whether by replicating or verifying research findings or by asking new questions with extant data. Scientific observations from multiple sites can be combined for longitudinal and comparative research. Social surveys can be mined to identify trends and to make comparisons. Historical records can be analyzed for scholarly inquiry, education, or genealogy. Data are also the 'glue' of collaborative research. Scholars work together to generate data, and those data are an essential product of the collaboration.

Data are even more valuable if they can be linked to the resulting publications and to other associated objects such as field notes, grant proposals, and software models. Finding technical and social means to make these links is among the great challenges of e-research. Technical developments such as the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) improve the ability to discover scholarly documents and data on the Internet.[11] The Object Reuse and Exchange (ORE) project is building a layer on top of the OAI that will enable links between parts of compound objects.[12,13] These efforts, in turn, build upon earlier interoperability technologies embraced by the publishing community such as Digital Object Identifiers, OpenURL, and CrossRef.[14–19]

From a technical perspective, *data* is a simple concept, as exemplified by this widely cited definition:

> *Data*: A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.[20]

From a social perspective, however, *data* is not at all a simple or straightforward concept. Notions vary widely about what are 'data', to whom, when, and for what purposes. Observations that are research findings for one scientist may be background context to another. Data that are adequate evidence for one purpose (e.g. determining whether water quality is safe for surfing) are inadequate for others (e.g. government standards for testing drinking water). Similarly, data that are synthesized for one purpose may be 'raw' for another.[3,8] Data do not become 'evidence' until someone makes a claim that some data are evidence of something.[21] These are among the many complexities of data practices that are being explored in research on scientists, computer scientists, and engineers.[22–24]

Disciplinary differences in access to digital content

The value chain of data, publications, and other objects will be constructed at different rates, and in different ways, across the disciplines (discussed at length in Borgman[3]). To incorporate content into a digital chain, it must first be in digital form. Non-digital content can be linked into the online chain if digital surrogates are available, such as bibliographic records or other forms of metadata. The proportion of publications, data, and other objects that are in digital form is far from uniform across the disciplines.

Online access to scholarly publications

The sciences continue to be the most advantaged of the disciplines in the proportion of their scholarly publications that are online, and the humanities the least advantaged. By 2005, 93% of science, technology, and medicine journals, 84% of humanities journals, and 90% of all journals currently published were reported to be available online.[25–27] The proportion of a journal's online content varies widely. Newer journals may be online in their entirety. Older journals may have converted and posted all back issues or may have converted selectively. In yet other cases, older issues are online but the most recent ones are not. The citation indexes of Thomson Scientific, collectively known as the *Web of Knowledge*, are a good proxy for online availability of publications. Bibliographic coverage (i.e. digital surrogates) in the *Science Citation Index* dates back to 1900, in the *Social Sciences Citation Index* to 1956, and in the *Arts and Humanities Citation Index* only to 1975. Scopus, the competing Elsevier product, indexes science, technology, medicine, and social sciences literature only for the last 10 years, and does not claim to cover the humanities at all.[28]

Notably, the use of scholarly literature by discipline is the inverse ordering of that available in the *Web of Knowledge*. Scientists have the longest time period of content available online, yet concentrate their reading and citing in very recent publications. Conversely, online coverage is the shallowest in the humanities, where reading and citing span a much longer time range. Literature in the humanities notoriously goes out of print long before it goes out of date. The social sciences fall in between the sciences and humanities both in availability and in usage of literature.[29–32]

Two observations on these distributions should be of particular significance to publishers. One is that the *Web of Knowledge* is a 'long-tail' business model, to use the terminology popularized by Chris Anderson,[33,34] similar to that of Amazon and Netflix, whereas Scopus is a 'hits' business model akin to brick-and-mortar distribution. The long-tail model attracts consumers of older and less popular content, creating new markets. The other observation is that the mass digitization of books undertaken by Google Print, the Open Content Alliance, and a growing number of other international projects may shift the market considerably. Humanists, presently the least advantaged discipline, will benefit most from the online availability of older, out-of-copyright materials. They are hungry for the content farther down the tail. As those resources come online, e-research in the humanities may explode.[35–37]

Online access to scholarly data

While the form and content of journal articles and books varies considerably between the disciplines, far more consistency exists than is the case with data. Even within disciplines, types of data can range widely. Examples of scientific data include:

- Ecology: weather, ground water, sensor readings, historical records
- Medicine: X-rays, test results, pathology reports
- Chemistry: protein structures
- Astronomy: spectral surveys
- Biology: specimens
- Physics: events, objects

even within disciplines, types of data can range widely

Other forms of data include records such as laboratory or field notebooks and spreadsheets. What is common across scientific fields is that scientists tend to use data that were created for research purposes, whether they generate those data themselves or use data acquired from collaborators or other scientists. Other important sources are public data repositories such as the Protein Databank, Sloan Digital Sky Survey, and IRIS (seismology).[38–40]

Social science data also vary widely between fields, encompassing opinion polls, surveys, interviews, laboratory experiments, field experiments, demographic records, census records, voting records, and economic indicators. Many, if not most, social scientists collect their own data. Many also acquire research data from collaborators, other social scientists, or repositories such as the Survey Research centers or other data archives.[41–43] Social scientists differ from natural scientists by drawing upon data that were not produced by or for research purposes, such as government records, corporate records, or economic statistics.

Considerable effort and analysis may be required to analyze, interpret, and compare these types of data.[44]

Humanities data are the most varied of all, as almost any record of human activity can be considered data. Examples include newspapers, photographs, letters, diaries, books, articles; birth, death, and marriage records; church and court records; school and college yearbooks; and maps. The boundary between data and publications is much more vague in the humanities than in the other disciplines. An old book might be read as literature or might be treated as a source of data to be mined for names, places, and events. The mass digitization of books and records offers a wealth of data and text-mining opportunities for humanities researchers. Humanists are less likely to generate new data by observations, as in the sciences and social sciences. They are more likely to search libraries, archives, and public records, to acquire data from other scholars, or to search public data repositories such as the Beazley Archive, the Perseus Digital Library, or the Arts and Humanities Data Service.[45–48] Thus, humanists are much more dependent upon external data sources than are scholars in other disciplines, which presents a host of access and intellectual property issues for this community.[7]

humanists are much more dependent upon external data sources than are scholars in other disciplines

Disciplinary differences in access to data

Scholarship advances by sharing, exchange, and access. The idea of 'open science', that scholarship exists only after being reviewed by peers and openly distributed, dates back to Francis Bacon (1561–1626). Incentives to publish journal articles, books, and conference papers proceed from these basic premises. Scholars are judged by the quality of their publications for the purposes of hiring, tenure, and promotion.

Access to data that were produced with public monies is the highest priority for e-research, as these data are seen as a public investment to be leveraged for the good of the larger community. However, the means by which data are to be published or otherwise made available is not yet entirely clear, and considerations vary by discipline. The concerns for access to data are similar to those for publications; thus open access models offer a helpful comparison.

Scholarly incentives for open access to publications

The open access movement has stimulated an examination of the implicit incentives for choices of publication venue. Scholars balance multiple considerations in deciding where to submit their work. Prestige of the journal, conference, or publisher is a primary concern, despite its subjective character. The prestige of publishers, or 'symbolic capital' as discussed by Thompson,[49] is becoming an ever more fluid notion as the number of competing outlets for scholarly research increases. Scholars generally want to maximize the recognition of their research, which involves not only prestige but accessibility and visibility. Other factors include speed of publication and requirements by some universities and funding agencies to make their publications openly available within a certain timeframe.

Recognition of work is often measured by citation rates, whether by citations to individual publications or to journals in which the work is published. The ISI Impact Factors, which rank journals by their citation rates in the ISI *Web of Knowledge* databases, remain very influential. However, the value of those Impact Factors is being questioned in the digital environment.[50] Citation indicators have become more comprehensive through 'webometric' analyses of references made on the World Wide Web.[51,52] Some of the new online open access journals, such as those published by the Public Library of Science, are already cited highly. A new Mellon-funded project has compiled a very large corpus of user transactions drawn from online services of libraries, publishers and aggregators to compare usage-based and citation-based impact factors.[53,54]

Among the most controversial issues in open access is the degree to which free access to publications increases citation rates. Open access proponents cite mounting evidence that articles 'self-archived' on websites or repositories, or published in open access journals, are cited more frequently than those accessible only via subscriptionbased journals.[55–60] A current bibliography, with some annotation, is maintained by the Open Citation Project.[61] Charles Bailey[62] earlier published a more extensive bibliography of open access studies and reports. One publisher-led study surveyed OA and non-OA journals about their citation rates[63] and another examined recent literature on citation rates of self-archived articles.[64,65] Both concluded that evidence for the citation advantage of open access is weaker than its proponents claim. The Publishing Research Consortium, 'a group of publishing societies and individual publishers, which supports global research into scholarly communication in order to enable evidence-based discussion',[66] sponsored the latter study among many others.

Both proponents and opponents of open access dispute each other's research methods for measuring citation rates, revisiting age-old debates about the reliability and validity of bibliometrics.[67] Usage-based impact factors, which complement both citation-based and page-rank measures, offer new ways to assess the value of a publication to the scholarly community. What is hard to dispute, however, is that the 'principle of least effort' is a consistent conclusion of research on information-seeking behavior.[68] Given a choice between quality publications that are easy to obtain and those that are difficult to obtain, most seekers will opt for the former. In this complex environment, authors want their work to be as accessible, as valued, and as permanent a part of the scholarly record as possible.

Scholarly incentives for open access to data

Publication of articles, books, and papers is still the primary means by which scholars' research is evaluated. Publication of data, *per se*, is recognized as a scholarly contribution in only a few fields. Yet scholars are being encouraged, and sometimes required by funding agencies, to make their research data available. Their options for doing so depend on whether their research field has established data repositories, whether their university libraries or other campus entities are accepting responsibility for curating data, or whether they or their research partners have established community repositories. If none of these options exists, scholars can post datasets on their personal websites for downloading, or can release them upon request, on a case-by-case basis. In some cases, journals will post datasets online as appendices to articles. The situation becomes ever more complex as the line blurs between incorporating data in scholarly publications and making references from publications to datasets.[10] Thus the infrastructure for making data available is far less mature than is the case for publishing reports of research.[3]

the infrastructure for making data available is far less mature than is the case for publishing reports of research

Deterrents to open access to data

Lacking adequate infrastructure in most fields, current scholarly practice tends to discourage researchers from sharing data beyond their collaborators. Reasons for not sharing or not contributing data to repositories can be grouped into four categories.[3] First is that scholars are rewarded for publication, not for data management. Second, and closely related, is the effort required to document data. Describing and tracking data for one's own use, and the use of labmates and other current collaborators, is far simpler than documenting them for use by unknown others. Even making data available for private exchange with known users requires richer explanations of the methods by which the data were collected, cleaned, analyzed, recorded, and interpreted. To make data available for public repositories, they may also need to be organized in compliance with community standards for metadata and ontologies.

Third is the concern for competition and priority of claims. Scholars collaborate, but they also compete. The first to publish a finding or a new interpretation will get the faculty position, the research grant, the patent, the graduate students, or perhaps the Nobel Prize. Scholars typically will not release data until the findings are published, or until they are finished mining the data, which may take longer.

Lastly, and most complex, are concerns about intellectual property. Researchers rarely own their data in the legal sense, but they usually have control over their data.

Data that result from grants funded by governments may fall into the public domain, for example. In other cases, scholars may be analyzing data obtained from public or private sources. In corporate situations, such as pharmaceutical research, the funding corporation usually would be the legal owner. Regardless of the legal realities, scholars may be reluctant to release their data out of concerns for misuse, misinterpretation, or 'free riders' – exploitation by those who did not invest the effort in data production such as obtaining grants and supervising the research.

These four deterrents to sharing data play out somewhat differently across the disciplines, following the differences in data sources outlined above. Data about living people, which is the usual arena in the social and health sciences, require anonymization. Priority for discovery, whether for a new planet, particle, or program, is usually handled by placing embargoes on data until publication of findings. Release of data from collaborative projects can be hampered by lack of agreement about who has the authority to release them or by conflicting requirements of collaborators' institutions or legal jurisdictions. Investigators who acquire data from third parties, whether public or private, may not have the authority to release those data due to contracts that specify their application and use. Potential reusers may need to go back to the original source. Humanists, who rely most heavily on data from third parties, may find that permissions are not available, transferable, or affordable. These challenges are greatest for scholars of contemporary culture who wish to use art, music, mass media, articles, books, and other content that is under copyright.

data resulting from publicly funded research are the highest priority for public access

Building a scholarly infrastructure for data

Scholarly publishers are not alone in assessing their future role in the scholarly information infrastructure, of which data are but one component. A few issues are becoming clear for all concerned. One is that data resulting from publicly funded research, in all disciplines, are the highest priority for public access. These data are 'public goods', in the economic sense – they can be used an infinite number of times without being consumed ('non-rival'), and it is difficult, if not impossible, to exclude anyone from using them ('non-excludable').[69,70] Many other types of data have scholarly value, but the economic and policy issues are even more complex and beyond the scope of discussion here.

Policy issues for open access to publicly funded data are being addressed at the national and international levels, through individual governments and multinational agencies such as CODATA.[71–75] Mechanisms are being established, for instance, to assist research partners in making agreements on the ownership, access, use, and reuse of data, and on conditions for release of data to others. Intellectual property agreements that promote the sharing of data, while retaining some degree of control, are being promoted by these agencies and by new entities such as the Science Commons.[76,77]

Sustainable institutional models for access to research data are also part of the public discussion. Libraries and archives have institutional mandates for long-term curation and preservation, whereas repositories within disciplines are often supported by fixed-term research contracts. The challenge lies in combining institutional models, funding sources, and expertise most effectively. Data curation is expensive for reasons similar to those which apply to publishing: peer review, editorial processes, technical support, and maintenance. Data also are much less 'self-documenting' than are publications. Without metadata and descriptions of research methods and of the context for data collection, they may simply be tables of numbers, lists of codes, pretty pictures, or boxes of rocks.[2,73,78–80]

Also required for a scholarly infrastructure for data are technical means to discover them and to establish links between related objects to form the value chain of scholarly content. Here, too, considerable progress is being made. Standards and tools mentioned above, including the Open Archives Initiative, OpenURL, CrossRef, and Digital Object Identifiers, underpin the scholarly infrastructure for digital publications. These methods can be extended to data. Digital Object Identifiers are already being used for data,[18] and the Object Reuse and Exchange project builds upon established mechanisms to create compound objects.

Clarifying legal aspects of data sharing, establishing sustainable institutional models, and implementing technical standards for discovery will create more incentives to share data. In turn, the scholarly reward system will adapt to more computing- and data-intensive research environments in all disciplines. The longer-term goal for e-research is to achieve critical mass in the amount of research data publicly available for reuse. Whether repositories are established by research fields, by universities, or both is less an issue than whether their content is readily discoverable and usable. To be discoverable and usable, data need to be described consistently and curated to standards acceptable to the scholarly communities they serve.

The components of a scholarly information infrastructure are maturing at different rates. The challenge is to combine them into a coherent whole that will create large, public repositories of research data that can be mined and combined, used to ask new questions in new ways, and used to replicate and verify prior research. In the absence of comprehensive solutions, local solutions are emerging that threaten to fragment the scholarly record. Partial solutions include posting datasets on investigator websites or including them as supplemental materials in publisher databases. Neither of these approaches is scalable. Individual investigators are unlikely to document their data in ways that make those data easily discoverable or interpretable. Nor are individual investigators likely to take long-term responsibility for migrating those data to new software and hardware platforms or to new metadata and ontology structures in their fields. Without these investments, the data will decay, perhaps quickly. Including data as supplements to a journal article at least makes those data discoverable in association with that article. Such datasets are scarcely more likely to be in standard formats or to be upgraded as technology and standards evolve, however. In neither case are data stored in these ways necessarily discoverable by search engines. Data that are embedded in fixed formats such as PDF are especially problematic, as their structure is stripped away.[81]

Creating publicly accessible repositories, with links to and from publisher databases, is a promising solution for most types of data. More accessible methods of representing data within publications are also needed, as Lynch[10] has discussed. Piecemeal approaches to augmenting individual publications do not scale and do not establish the value chain, as the relationship between data and publications is rarely one-to-one. Proprietary databases of public research data are infeasible from either a policy or a market perspective. Conversely, publishers can partner with authors, scholarly societies (many of whom also are publishers), and funding agencies to support technologies, standards, practices, and the establishment of public data repositories. Links between such repositories and publisher databases add value to each. Readers can follow links to the data on which publications are based. People and computers alike can search for data and then follow links to the publications that describe those data. Establishing direct links between elements in the value chain of scholarship will thereby enhance the usefulness and discoverability of each element.

the longer-term goal for e-research is to achieve critical mass in the amount of research data publicly available for reuse

In sum, publishers, scholars, librarians, archivists, universities, and funding agencies have much to gain by working together to build the scholarly information infrastructure for e-research, of which research data are a critical component. Each has a role to play. By supporting public access to data and building links from publications to data, publishers will increase the visibility of their own products, while contributing to the larger public good.

References

1. Hey, T. and Trefethen, A. 2005. Cyberinfrastructure and e-Science. *Science*, 308: 818–21. http://dx.doi.org/10.1126/science.1110410
2. Cyberinfrastructure Vision for 21st Century Discovery. Washington, DC, National Science Foundation, 2007. http://www.nsf.gov/pubs/2007/nsf0728/ (accessed 17 July 2007).
3. Borgman, C.L. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA, MIT Press, 2007.
4. Atkins, D.E., Droegemeier, K.K., Feldman, S.I.,

Garcia-Molina, H., Klein, M.L., Messina, P., Messerschmitt, D.G., Ostriker, J.P., and Wright, M.H. Revolutionizing Science and Engineering Through *Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure*. Washington, DC, National Science Foundation, 2003. http://www.nsf.gov/cise/sci/reports/atkins.pdf (accessed 18 September 2006).

5. *E-resources for Research in the Humanities and Social Sciences. A British Academy Policy Review*. London, The British Academy, 2005. http://www.britac.ac.uk/reports/eresources/ (accessed 30 September 2006).
6. *ESRC National Center for e-Social Science*. 2006. http://www.ncess.ac.uk/ (accessed 22 April 2006).
7. Unsworth, J., Courant, P., Fraser, S., Goodchild, M., Hedstrom, M., Henry, C., Kaufman, P.B., McGann, J., Rosenzweig, R., and Zuckerman, B. *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for Humanities and Social Sciences*. New York, American Council of Learned Societies, 2006. http://www.acls.org/cyberinfrastructure/cyber.htm (accessed 17 July 2007),
8. Bowker, G.C. *Memory Practices in the Sciences*. Cambridge, MA, MIT Press, 2005.
9. Hey, A.J.G. and Trefethen, A. The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, and A.J.G. Hey (eds), *Grid Computing: Making the Global Infrastructure a Reality*, Wiley: Chichester, 2003. http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf (accessed 20 January 2005).
10. Lynch, C.A. 2007. The shape of the scientific article in the developing cyberinfrastructure. *CT Watch Quarterly*, 3(3). http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/ (accessed 17 July 2007).
11. Van de Sompel, H., Nelson, M.L., Lagoze, C., and Warner, S. 2004. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10(12). http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html (accessed 5 October 2006).
12. *Object Reuse and Exchange*. 2006. http://www.openarchives.org/ore/ (accessed 15 November 2006).
13. Van de Sompel, H. and Lagoze, C. 2007. Interoperability for the discovery, use, and re-use of units of scholarly communication. *CT Watch Quarterly*, 3(3). http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/ (accessed 17 July 2007).
14. *OpenURL and CrossRef*. 2006. http://www.crossref.org/02publishers/16openurl.html (accessed 3 October 2006).
15. Chudnov, D., Cameron, R., Frumkin, J., Singer, R., and Yee, R. 2005. Opening up openURLs with autodiscovery. *Ariadne*, 43. http://www.ariadne.ac.uk/issue43/chudnov/ (accessed 29 September 2006).
16. Hellman, E. 2003. OpenURL: making the link to libraries. *Learned Publishing*, 16(3): 177–81. http://dx.doi.org/10.1087/095315103322110950
17. *The Digital Object Identifier System*. 2006. http://www.doi.org (accessed 5 October 2006).
18. Paskin, N. 2005. Digital object identifiers for scientific data. *Data Science Journal*, 4(1): 1–9.
19. *CrossRef*. 2006. http://www.crossref.org/index.html (accessed 26 July 2006).
20. *Reference Model for an Open Archival Information System*. Recommendation for Space Data System Standards, 2002, pp. 1–9. http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed 4 October 2006).
21. Buckland, M.K. 1991. Information as thing. *Journal of the American Society for Information Science*, 42(5): 351–60. http://dx.doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3
22. Borgman, C.L., Wallis, J.C., and Enyedy, N. Building digital libraries for scientific data: an exploratory study of data practices in habitat ecology. In J. Gonzalo, C. Thanos, M.F. Verdejo, and R.C. Carrasco (eds), *10th European Conference on Digital Libraries, Alicante, Spain*. Berlin, Springer Verlag, 2006, pp. 170–83.
23. Borgman, C.L., Wallis, J.C., and Enyedy, N. 2007. Little Science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*. http://www.springerlink.com/content/f7580437800m367m/ (accessed 29 September 2007).
24. Wallis, J.C., Borgman, C.L., Mayernik, M., Pepe, A., Ramanathan, N., and Hansen, M. Know thy sensor: trust, data quality, and data integrity in scientific digital libraries. In L. Kovacs, N. Fuhr, and C. Meghini (eds), *11th European Conference on Digital Libraries, Budapest, Hungary*. LINCS 4675. Berlin, Springer Verlag, 2007, pp. 380–91.
25. Cox, J. and Cox, L. *Scholarly Publishing Practice: The ALPSP Report on Academic Journal Publishers' Policies and Practices in Online Publishing*. London, Association of Learned and Professional Society Publishers, 2003.
26. Garson, L.R. 2004. Communicating original research in chemistry and related sciences. *Accounts of Chemical Research*, 37(3): 141–48. http://dx.doi.org/10.1021/ar0300017
27. Cox, J. and Cox, L. *Scholarly Publishing Practice: Academic Publishers' Policies and Practices in Online Publishing*. London, Association of Learned and Professional Society Publishers, 2005. http://www.alpsp.org/ngen_public/article.asp?id=200&did=47&aid=269&st=&oaid=-1 (accessed 27 September 2007).
28. *Scopus in Detail*. 2006. http://www.info.scopus.com/detail/what/ (accessed 31 March 2006).
29. Meadows, A.J. *Communicating Research*. San Diego, Academic Press, 1998.
30. Tenopir, C. and King, D.W. 2002. Reading behaviour and electronic journals. *Learned Publishing*, 15: 259–65. http://dx.doi.org/10.1087/095315102760319215
31. Tenopir, C. and King, D.W. *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*. Washington, DC, Special Libraries Association, 2000.
32. Cronin, B. *The Hand of Science: Academic Writing and its Rewards*. Lanham, MD, Scarecrow Press, 2005.
33. Anderson, C. *The Long Tail: Why the Future of Business is Selling Less of More*. New York, Hyperion, 2006.
34. Anderson, C. 2004. The long tail. *Wired Magazine*, 12(10). http://wired.com/wired/archive/12.10/tail_pr.html (accessed 17 September 2006).
35. *Open Content Alliance*. 2005. http://www.opencontentalliance.org/ (accessed 25 November 2005).
36. Crane, G. 2006. What do you do with a million books? *D-Lib Magazine*, 12(3). http://www.dlib.org/dlib/march06/crane/03crane.html (accessed 17 August 2006).

37. *Mass Digitization: Implications for Information Policy*. Report from "Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects", Symposium, University of Michigan, 2006. http://www.nclis.gov/digitization/MassDigitizationSymposium-Report.pdf (accessed 29 July 2006).
38. *Sloan Digital Sky Survey*. 2006. http://www.sdss.org/ (accessed 15 August 2006).
39. *Incorporated Research Institutions for Seismology*. 2006. http://www.iris.edu (accessed 30 September 2006).
40. *Protein Data Bank*. 2006. http://www.rcsb.org/pdb/ (accessed 4 October 2006).
41. *Survey Research Center, UC-Berkeley*. 2005. http://srcweb.berkeley.edu/ (accessed 24 May 2005).
42. *Survey Research Center, Institute for Social Research*. 2006. http://www.isr.umich.edu/src/ (accessed 4 October 2006).
43. *UK Data Archive*. 2006. http://www.data-archive.ac.uk/about/about.asp (accessed 5 October 2006).
44. Berman, F. and Brady, H. *Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences*. National Science Foundation, 2005. http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf (accessed 30 April 2006).
45. *Beazley Archive*. 2006. http://www.beazley.ox.ac.uk/BeazleyAdmin/Script2/TheArchive.htm (accessed 31 March 2006).
46. *Perseus Digital Library*. 2006. http://www.perseus.tufts.edu/ (accessed 22 April 2006).
47. Mahoney, A. 2002. Finding texts in Perseus. *New England Classical Journal*, 29(1): 32–34.
48. *Arts and Humanities Data Service*. 2006. http://www.ahds.ac.uk/ (accessed 28 September 2006).
49. Thompson, J.B. *Books in the Digital Age*. Cambridge, Polity, 2005.
50. Monastersky, R. 2005. The number that's devouring science. *Chronicle of Higher Education*, 52(8): A12–A17.
51. Thelwall, M. 2006. Interpreting social science link analysis research: a theoretical framework. *Journal of the American Society for Information Science & Technology*, 57(1): 60–68.
http://dx.doi.org/10.1002/asi.20253
52. Thelwall, M., Vaughan, L., and Bjorneborn, L. Webometrics. In B. Cronin (ed.), *Annual Review of Information Science and Technology*. Information Today, Medford, NJ, 2005, pp. 81–135.
53. Bollen, J. and Van de Sompel, H. 2008. Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1)
http://dx/doi.org/10.1002/asi.20746
54. Rodriguez, M.A., Bollen, J. and Van de Sompel, H. A practical ontology for the large-scale modeling of scholarly artefacts and their usage. *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. Vancouver, BC, Association for Computing Machinery, 2007, 278–87.
55. Harnad, S. and Brody, T. 2004. Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 10(6). http://www.dlib.org/dlib/june04/harnad/06harnad.html (accessed 30 September 2006).
56. Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., and Murray, S.S. 2005. The effect of use and access on citations. *Information Processing & Management*, 41(6): 1395–1402.
http://dx.doi.org/10.1016/j.ipm.2005.03.010
57. Lawrence, S. (2001). Free online availability substantially increases a paper's impact. Nature Web Debates. http://www.nature.com/nature/debates/e-access/Articles/lawrence.html (accessed 9 May 2005).
58. MacCallum, C.J. and Parthasarathy, H. 2006. Open access increases citation rate. *PLoS Biology*, 4(5): e176.
doi:10.1371/journal.pbio.0040176
59. Davis, P.M. and Fromerth, M.J. 2007. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2): 203–15.
http://dx.doi.org/10.1007/s11192-007-1661-8
60. Antelman, K. 2004. Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5): 372–82. http://eprints.rclis.org/archive/00002309/
61. *The effect of open access and downloads ('hits') on citation impact: a bibliography of studies*. The Open Citation Project – Reference Linking and Citation Analysis for Open Archives, 2007. http://opcit.eprints.org/oacitation-biblio.html (accessed 27 September 2007).
62. Bailey, C. *Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals*. Washington, DC, Association of Research Libraries, 2005. http://info.lib.uh.edu/cwb/oab.pdf (accessed 28 September 2006).
63. *The Facts about Open Access: A Study of the Financial and Non-Financial Effects of Alternative Business Models on Scholarly Journals.* Kaufman-Wills Group LLC. Worthing, Association of Learned and Professional Society Publishers, 2005. http://sippi.aaas.org/Pubs/FAOAcompleteREV.pdf (accessed 27 September 2007).
64. Craig, I.D., Plume, A.M., McVeigh, M.E., Pringle, J. and Amin, M. 2007. Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics*, 1: 239–48.
http://dx.doi.org/10.1016/j.joi.2007.04.001
65. *Publishing Research Consortium*. http://www.publishingresearch.net/Citations.htm (accessed 12 November 2007).
66. *Publishing Research Consortium.* http://www.publishingresearch.net/index.html (accessed 12 November 2007).
67. Borgman, C.L. (ed.) *Scholarly Communication and Bibliometrics*. Newbury Park, CA, Sage, 1990,.
68. Case, D.O. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*, 2nd edn. San Diego, Academic Press, 2006.
69. Nelson, R.R. 1959. The simple economics of basic scientific research. *Journal of Political Economy*, 67: 323–48.
http://dx.doi.org/10.1086/258177
70. Dasgupta, P. and David, P.A. 1994. Toward a new economics of science. *Research Policy*, 23(5): 487–521.
http://dx.doi.org/10.1016/0048-7333(94)01002-1
71. David, P.A. (2003). The economic logic of 'open science' and the balance between private property rights and the public domain in scientific data and information: a primer, in *The Role of the Public Domain in Scientific Data and Information*, Washington, DC, National Academy Press, pp. 19–34. http://siepr.stanford.edu/papers/pdf/02-30.html (accessed 30 September 2006).
72. *CODATA-CENDI Forum on the National Science Board Report on Long-Lived Digital Data Collections*. 2005. http://www7.nationalacademies.org/usnc-codata/

Forum_on_NSB_Report.pdf (accessed 29 September 2006).

73. *Long-Lived Digital Data Collections*. 2005. http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf (accessed 1 October 2006)

74. Esanu, J.M. and Uhlir, P.F. (eds). *Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium*. US National Committee for CODATA, Board on International Scientific Organizations, Policy and Global Affairs Division, National Research Council. National Academies, Washington, DC, The National Academies Press, 2004. http://books.nap.edu/catalog/11030.html (accessed 30 September 2006).

75. Uhlir, P.F. The emerging role of open repositories as a fundamental component of the public research infrastructure. In G. Sica (ed.), *Open Access: Open Problems*. Monza, Polimetrica, 2006.

76. *Science Commons*. 2006. http://sciencecommons.org/about/index.html (accessed 6 October 2006).

77. Burk, D.L. Bioinformatics lessons from the open source movement. In H. Tavani (ed.), *Ethics, Computing, and Genomics: Moral Controversies in Computational Genomics*. Sudbury, MA, Jones and Bartlett, 2005, pp. 257–254.

78. Lyon, L. Dealing with data: roles, rights, responsibilities, and relationships. Bath, UKOLN, 2007. http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx (accessed 23 July 2007).

79. Hodge, G. and Frangakis, E. *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice*. Washington, DC, The International Council for Scientific and Technical Information (ICSTI) and CENDI: US Federal Information Managers Group, 2005, p. 84. http://cendi.dtic.mil/publications/04–3dig_preserv.html (accessed 30 September 2006).

80. Lord, P. and Macdonald, A. *E-Science Curation Report – Data Curation for E-science in the UK: An Audit to Establish Requirements for Future Curation and Provision*. JISC Committee for the Support of Research, 2003. http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf (accessed 1 October 2006).

81. Murray-Rust, P., Rzepa, H.S., Tyrrell, S.M., and Zhang, Y. 2004. Representation and use of chemistry in the global electronic age. *Organic and Biomolecular Chemistry*, 2(22): 3192–203.
http://dx.doi.org/10.1039/b410732b

Christine L. Borgman
Professor & Presidential Chair in Information Studies
University of California, Los Angeles
Email: Borgman@gseis.ucla.edu
Website: http://is.gseis.ucla.edu/cborgman/