RESEARCH ARTICLE

Whose text, whose mining, and to whose benefit?

Christine L. Borgman

Director, Center for Knowledge Infrastructures, University of California, Los Angeles

Keywords: data, information retrieval, methods, publishing, scholarship, text

ABSTRACT

Scholarly content has become more difficult to find as information retrieval has devolved from bespoke systems that exploit disciplinary ontologies to keyword search on generic search engines. In parallel, more scholarly content is available through open access mechanisms. These trends have failed to converge in ways that would facilitate text data mining, both for information retrieval and as a research method for the quantitative social sciences. Scholarly content has become open to read without becoming open to mine, due both to constraints by publishers and to lack of attention in scholarly communication. The quantity of available text has grown faster than has the quality. Academic dossier systems are among the means to acquire more quality data for mining. Universities, publishers, and private enterprise may be able to mine these data for strategic purposes, however. On the positive front, changes in copyright may allow more data mining. Privacy, intellectual freedom, and access to knowledge are at stake. The next frontier of activism in open access scholarship is control over content for mining as a means to democratize knowledge.

1. DATA, TEXT, AND MINING

Scholarship has become datafied as text, images, sound, video, numerical observations, and other forms of intellectual materials meld together as born-digital content. While extant cultural artifacts such as older books, paper archives, and physical objects are unlikely to be replaced by digital records, the scholarly research about those materials will be published as digital objects, whether journal articles, books, "papers," videos, data sets, or other entities.

Paradoxically, the proliferation of digital content has made scholarly information harder to find. In the days of print publication, libraries cataloged books meticulously, providing multiple points of entry to authors, titles, subjects, and other bibliographic elements. Variant forms of author names were cross-referenced and clustered under a curated authority record. Online catalogs, starting in the latter 1970s, offered Boolean search capabilities that exploited these multiple indexes. Journal articles were described by indexing and abstracting services, often providing extensive subject-analytic metadata drawn from discipline-specific thesauri. The I&A services, as they were known, offered elaborate search functions that exploited these metadata and thesauri. User interfaces were cumbersome, but in the hands of experts, these bibliographic databases could be mined with great scholarly sophistication (Borgman, 2000, 2007, 2015; Borgman, Moghdam, & Corbett, 1984).

Today's search is dominated by keyword strings, flattening out the rich structure of earlier digital library systems. Users type a few words into a search engine, leaving the combinatorics to proprietary algorithms whose rules are known only to the companies that deploy them. Even search engine providers may be hard pressed to explain precisely how any given set of results

an open access journal

Citation: Borgman, C. L. (2020). Whose text, whose mining, and to whose benefit? *Quantitative Science Studies*, *1*(3), 993–1000. https://doi.org/10.1162/qss_a_00053

DOI: https://doi.org/10.1162/qss_a_00053

Corresponding Author: Christine L. Borgman
Christine.Borgman@ucla.edu

Handling Editors: Loet Leydesdorff, Ismael Rafols, and Staša Milojević

The MIT Press

are retrieved, given the use of machine-learning techniques that adapt continuously to changes in individual profiles, in auction algorithms that rank results by advertiser payment, and in proprietary knowledge graphs.

As a result of these and other changes in information retrieval, many scholars are finding that the best way to mine databases of text and other content with sufficient sophistication is to write their own algorithms and scripts. Searching databases, web archives, and other digital content is now known generically as *text data mining* (TDM), although the search may include more than text (McDonald & Kelly, 2014). When the content being searched is open, these methods may be known as *open content mining* (Murray-Rust, Neylon, et al., 2010).

TDM requires as much technical sophistication on the part of researchers in the quantitative social sciences as was required of librarians in earlier days of information retrieval. TDM is gaining in popularity in the social sciences to model behavior and policy, in the sciences to extract data from publications, and in the humanities to explore history, culture, linguistics, philology, and more. Data mining can regain many of the advantages of sophisticated ontology-based tools of an earlier era by giving the searcher fine-grained and transparent control over the search process, at scale.

Open access publishing is a parallel trend, where scholarly publications are available to readers without charge. A growing proportion of new scholarly articles (and books to a lesser extent) is publicly available immediately or within a few months of initial release. In principle, open access publishing should make much more content available for TDM, which in turn, would facilitate open content mining. In fact, open access publishing does not appear to be advancing the scale of TDM. The failure of these two trends to converge is the subject of this article.

2. OPEN DATA, CLOSED DATA, AND MINABLE DATA

Researchers have sought technical access to proprietary databases of published materials since the earliest days of online databases in the latter 1970s, yet publishers continue to write contracts with university libraries based on assumptions of human readership. By the time of Google Books and the associated author lawsuits, around 2005, we learned that publishers wished to restrict "non-consumptive use" of scholarly content (Duguid, 2007; Leetaru, 2008; Nunberg, 2010). Throughout this period, the move toward open access to journal articles accelerated, with arXiv launching in 1991 (Ginsparg, 2011) and PubMed Central in 2000 (PMC Overview, 2018). Numerous other discipline-specific preprint servers, institutional repositories, and commercial services designed to distribute or redistribute open access versions of scholarly publications have been launched since. Concurrently, open access to publications became mandatory or highly recommended by many funding agencies and universities, in the United States and internationally (Borgman, 2015; Boulton, Babini, et al., 2015; Enserink, 2016; Piwowar, Priem, et al., 2018; Rabesandratana, 2019; Willinsky, 2018).

As a consequence of open science policies and practices, a growing amount of digital content is available as open access for downloading, whether in open access journals, data archives, institutional repositories, library catalogs, preprint servers (such as arXiv, SocArXiv, and bioRxiv), government databases, social media, web portals, public agencies, or elsewhere. Open access to content does not necessarily mean that these data are minable, however. In many, if not most cases, these user interfaces presume a human user who is capable of reading a web page, searching for content, and selecting individual items for download. The number of records that may be downloaded for local mining may also be limited. Robots may or may not be allowed to search open access databases. Scholars and libraries are pressing for greater mining privileges of

journals, books, and other intellectual resources (Lammey, 2014; Senseney, Dickson, et al., 2018; Van de Sompel, 2013; Van de Sompel, Rosenthal, & Nelson, 2016; Williams, Fox, et al., 2014).

2.1. Open to Read vs. Open to Mine

Open science, in policy and in concept, is intended to improve transparency, accountability, and access to knowledge by providing open access to publications, data, and software; stewarding collections of scholarly resources for the long term; and making research data more findable, accessible, interoperable, and reusable (FAIR) (Borgman, 2015; Boulton et al., 2015; European Union Publications Office, 2018; Wilkinson, Dumontier, et al., 2016). While open science policies and practices have made great headway in increasing access to publications for reading and to research data for downloading, making scholarly content available for data mining is rarely a stated priority. Thus, the scholarly communication paradox: Open access to text for reading may not yield open access to text for mining.

The scholarly communication paradox can be traced to the early days of the internet and digital publishing. Activists' goals for open access to scholarly materials were to democratize access to knowledge and to limit the role of big publishers to control access to scholarly content via expensive contracts. Whereas open access proponents viewed digital publishing as a liberating technology, commercial publishers saw economic efficiencies and new markets (Borgman, 2007; Harnad, 1991, 1999, 2005; Suber, 2012; Willinsky, 2006).

Conflicts between democratization and publisher control intensified as open access to publications became the norm. To make articles available free of charge to readers, commercial publishers developed new business models that require authors to pay several thousand dollars (or euros) to make a single article open access. Subscription charges to university libraries continue, despite these author fees, which has led to new rounds of negotiation between publishers and universities. Several large countries and university systems recently terminated contracts with large publishers when talks broke down (Ellis, 2018; Kwon, 2017; UC and Elsevier, 2019; Yeager, 2018).

The cancellation of publisher contracts has received far more public attention than has the quieter consolidation of infrastructure for scholarly communication. A small group of large publishers are consolidating the industry by purchasing smaller publishers and by acquiring technology and content companies across the spectrum of academic services (Posada & Chen, 2018). Of particular note is the purchase of open access preprint servers such as SSRN and Bepress by commercial publishers, rebranding community resources as corporate content. Academic authors who contributed papers to these repositories as community-based, not-for-profit enterprises are not happy (Cookson, 2016; Ellis, 2019; Elsevier, 2017; McKenzie, 2017; Pike, 2016). In sum, open access is not turning out to be the information commons that was envisioned by its pioneers (Benkler, 2004; Hess & Ostrom, 2007; Kranich, 2004; Lessig, 2001; O'Sullivan, 2008; Reichman, Dedeurwaerdere, & Uhlir, 2009; Reichman, Uhlir, & Dedeurwaerdere, 2016).

Intellectual property issues abound. Researchers who wish to mine texts, and libraries who have paid large sums for digital access to published content, often claim that text mining should fall under fair use protections of copyright. (Legal protections vary by country; "fair use" is a term specific to U.S. law.) Publishers, in turn, often claim that their contracts cover only "consumptive use" by human readers and that universities should pay additional fees for mining access. Complicating matters further, large text corpora may contain both public domain and copyrighted materials that are indistinguishable for mining purposes (Baldwin, 2014; Elkin-Koren, 2004; Elkin-Koren & Fischman-Afori, 2017; Levine, 2014; Senseney et al., 2018; Wilkin, 2017).

2.2. Mining Quantity vs. Quality

Researchers' ability to mine text is fraught with complications, above and beyond the intellectual property and contractual challenges. User interfaces to bibliographic databases provide minimal mining capabilities and may limit the number of records that can be downloaded. Researchers report missing records and a general lack of transparency in search results when they attempt to download files for TDM (Dickson, Senseney, et al., 2018; Senseney et al., 2018).

Data quality is another complication for TDM. Original articles typically provide accurate bibliographic descriptions, and may also include "please cite as" instructions. However, references to published articles, which are essential for bibliometrics or for integrating content across databases, are inherently dirty data due to the vagaries of how authors create reference lists. A bibliography in a journal article is far from the "necessary and sufficient" set of citations that might be assumed by bibliometric evaluations. Rather, it is often an idiosyncratic list of familiar sources, compiled based on what is handy when the publication is submitted. Too few authors are bibliographic purists who verify middle initials, dates, DOIs, and page, volume, and issue numbers (Borgman, 2015, 2016). Complicating matters further is the lack of agreement on bibliographic styles. At last count, Zotero offered about 9,500 journal styles for referencing, representing about 2,000 unique bibliographic styles (Zotero Style Repository, 2019).

One way to get cleaner data is to extract them from authors' curricula vitae, as authors have a vested interest in providing accurate lists of their own oeuvre. However, CVs tend to be closely held documents in many fields. While some individuals post their CVs on web pages, few are comprehensive or current. To the extent that authors consistently submit their publications to institutional repositories, which is also rare, these could become reliable sources for bibliographic data.

2.3. Privacy and Intellectual Freedom

As universities automate academic personnel processes, faculty dossiers become high-quality sources of bibliographic data. These digital dossiers are typically isolated from the public record for privacy protection. Individuals can give informed consent for specific uses of specific data, such as a dossier for hiring or promotion. In principle, bibliographic records could be separated from confidential review letters, allowing bibliographies to become public records that could be mined. In practice, this opportunity rarely arises, even as an opt-in or opt-out mechanism.

However, these digital dossiers on academic staff are becoming rich sources to be mined by universities, publishers, and data analytics companies. When dossiers were paper files, academic personnel processes were entirely internal to universities. When they became digital files, a new market arose for data management and mining of these materials. Some of these academic analytic companies are independent or privately held; others are among the entities acquired by major publishers in recent years (Ellis, 2019; Posada & Chen, 2018). Rather than build their own infrastructure, universities are outsourcing many of their academic personnel services to these companies. Job applicants submit dossiers to websites, as do those who write their references. Candidates for tenure and promotion also upload their files to university portals on these systems. Dossier-hosting services have certain mining rights under their contracts with universities. Similarly, universities may mine these data for strategic purposes beyond the personnel action for which they were harvested. As faculty become aware of these systems and practices, concerns arise about who has access to their dossiers and how the data can be mined for making decisions about their careers, their departments, and their fields (Borgman, 2018a; Ellis, 2019).

The emerging academic analytics industry appears to be following the successful business models of Alphabet/Google, Facebook, and Amazon in aggregating vast amounts of data about people's lives. To the consumer, they promote the advantages of improving user experience with intelligent adaptation. To their business clients and investors, they promote the advantages of predictive analytics that can be deployed to strategic advantage. In the academic community, predictive analytics are being used to assess the performance of students and faculty, departments, universities, journals, research programs, and much more. The concentration of data by a few large players gives them a "god's eye view" of their domains, with minimal oversight or regulation (*Economist*, 2017).

A related concern is the ability of publishers to surveil uses of scholarly materials. Ownership of intellectual property carries a large set of rights and responsibilities, some of which are associated with privacy protection and intrusion. Corporate owners of scholarly publishing, mass media, and social media content deploy digital rights management (DRM) technologies to track uses and users in minute detail. These technologies have eroded traditional protections of privacy and intellectual freedom in libraries and other domains (Cohen, 1996; Lynch, 2017).

The ability of publishers and other database companies to surveil the uses of their content also has implications for intellectual freedom. To submit TDM queries to some of these systems, researchers may explicitly, or sometimes implicitly, be providing database owners with their research questions and methods. These constraints are of considerable concern to many researchers, who would prefer to search anonymously or to download text for local manipulation (Dickson et al., 2018). Among the motivations of HathiTrust Digital Library to build a research center is to facilitate TDM within the constraints of copyright law, with a rich array of tools (HathiTrust Digital Library, 2019). Another positive development is a shift in international copyright law to allow more TDM for scholarly and other purposes, on the grounds that these constraints would limit the growth of new data-intensive commerce (Samuelson, 2019).

3. DISCUSSION AND CONCLUSIONS

As scholarly information retrieval has degraded, from customized discipline-specific tools to generic search engines, TDM becomes researchers' best option for sophisticated information retrieval and content analysis. Open access publishing, despite making vastly more scholarly content available to read online, has not resulted in substantial improvements in open content mining. The lack of convergence of TDM and open access is due partly to a lack of foresight by activists who focused on human readers alone. TDM and robotic searching also democratize access to knowledge. The larger cause for the lack of convergence is the vested interests of publishers and other private stakeholders in maintaining control over intellectual property. These forms of control have proven lucrative, as more uses can be made of bibliographic data and scholarly materials through mining and combining with other intellectual assets (Posada & Chen, 2018).

Scholarly research with TDM methods, pioneered in the humanities in the 1960s, has benefited from advances in computation and data science. Researchers have deployed these methods, alone or in combination with other analytical tools, across academe. TDM is among the methods on which quantitative social sciences depends. The irony is that scholars produce the content that is valuable to mine, and build many of the tools on which these methods depend, and yet encounter ever more barriers in their efforts to exploit those texts in new ways. Universities collectively, and academic authors individually, have led the fight for more open access to knowledge for readers. University contracts with publishers are changing. Authors have more control over where

they submit their work, and more opportunity to post their work in open access repositories. Copyright law is allowing more data mining. Now is the time for activism on uses of our scholarly content. By enabling TDM on our works, individually and collectively, readers and researchers can make fuller use of scholarly knowledge. Scholars are overdue in asking, "Whose text, whose mining, and to whose benefit?"

ACKNOWLEDGMENTS

This paper is an expanded version of a discussion paper written for Data Mining with Limited Access Text: National Forum in 2018 (Borgman, 2018b; Dickson et al., 2018). Thank you to the organizers of the forum for the invitation, and to Michael Scroggins and Morgan Wofford of UCLA for comments and discussion on earlier drafts.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

No funding was received for this research.

REFERENCES

Baldwin, P. (2014). *The copyright wars: Three centuries of trans-Atlantic battle*. Retrieved from http://press.princeton.edu/titles/10303.html

Benkler, Y. (2004). Commons-based strategies and the problems of patents. *Science*, *305*(5687), 1110–1111. https://doi.org/10.1126/science.1100526

Borgman, C. L. (2000). *From Gutenberg to the global information infrastructure: Access to information in the networked world*. Cambridge, MA: MIT Press.

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: MIT Press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 93–115). Retrieved from https://www.degruyter.com/view/product/379257

Borgman, C. L. (2018a). Open data, grey data, and stewardship: Universities at the privacy frontier. *Berkeley Technology Law Journal*, *33*(2), 365–412. https://doi.org/10.15779/Z38B56D489

Borgman, C. L. (2018b). Text data mining from the author's perspective: Whose text, whose mining, and to whose benefit? *ArXiv:1803.04552 [Cs]*. Presented at the National Forum: Data Mining Research Using In-copyright and Limited-access Text Datasets, Chicago, IL. Retrieved from http://arxiv.org/abs/1803.04552

Borgman, C. L., Moghdam, D., & Corbett, P. K. (1984). *Effective online searching: A basic text*. New York: Marcel Dekker.

Boulton, G., Babini, D., Hodson, S., Li, J., Marwala, T., Musoke, M. G. N., … Wyatt, S. (2015). *Open data in a big data world: An international accord* [Outcome of Science International 2015 meeting]. Retrieved from ICSU, IAP, ISSC, TWAS website: https://twas.org/sites/default/files/open-data-in-big-data-world_ short_en.pdf

Cohen, J. E. (1996). A right to read anonymously: A closer look at "copyright management" in cyberspace. *Connecticut Law Review*, *28*, 981–1039.

Cookson, R. (2016). Elsevier buys research sharing website. *Financial Times*, May 17. Retrieved from https://www.ft.com/content/807f7714-1c48-11e6-b286-cddde55ca122

Dickson, E., Senseney, M., Namachchivaya, B., & Ludäscher, B. (2018). *IMLS National Forum on data mining research using in-copyright and limited-access text datasets: Discussion paper, forum statements, and SWOT analyses*. Retrieved from https://www.ideals.illinois.edu/handle/2142/100055

Duguid, P. (2007). Inheritance and loss? A brief survey of Google Books. *First Monday*, *12*. Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847

Economist. (2017). The world's most valuable resource is no longer oil, but data, May 6. Retrieved from http://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource

Elkin-Koren, N. (2004). The Internet and copyright policy discourse. In H. Nissenbaum & M. E. Price (Eds.), *Academy & the Internet* (pp. 252–274). New York: Peter Lang.

Elkin-Koren, N., & Fischman-Afori, O. (2017). Rulifying fair use. *Arizona Law Review*, *59*(1). Retrieved from http://arizonalawreview.org/rulifying-fair-use/

Ellis, L. (2018). In talks with Elsevier, UCLA reaches for a novel bargaining chip: Its faculty. *The Chronicle of Higher Education*, December 12. Retrieved from https://www.chronicle.com/article/In-Talks-With-Elsevier-UCLA/245311

Ellis, L. (2019). Elsevier's presence on campuses spans more than journals. That has some scholars worried. *The Chronicle of Higher Education*, April 3. Retrieved from https://www.chronicle.com/article/Elsevier-s-Presence-on/246048

Elsevier. (2017). Elsevier acquires bepress, a leading service provider used by academic institutions to showcase their research (August 2). Retrieved October 28, 2019, from: https://www.elsevier.com/about/press-releases/corporate/elsevier-acquires-bepress,-a-leading-service-provider-used-by-academic-institutions-to-showcase-their-research

Enserink, M. (2016). In dramatic statement, European leaders call for 'immediate' open access to all scientific papers by 2020. *Science*. https://doi.org/10.1126/science.aag0577

European Union Publications Office. (2018). Turning FAIR data into reality: Final report and action plan from the European Commission expert group on FAIR data, November 26. [Website]. Retrieved December 6, 2018, from https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF

Ginsparg, P. (2011). ArXiv at 20. *Nature*, *476*(7359), 145–147. https://doi.org/10.1038/476145a

Harnad, S. (1991). Post-Gutenberg galaxy: The fourth revolution in the means of production of knowledge. *Public-Access Computer Systems Review*, *2*, 39–53. Retrieved from http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/harnad91.postgutenberg.html

Harnad, S. (1999). Free at last: The future of peer-reviewed journals. *D-Lib Magazine*, *5*(12). Retrieved from http://www.dlib.org/dlib/december99/12harnad.html

Harnad, S. (2005). The implementation of the Berlin Declaration on Open Access: Report on the Berlin 3 Meeting held 28 February–1 March 2005, Southampton, UK. *D-Lib Magazine*, *11*(3). Retrieved from http://www.dlib.org/dlib/march05/harnad/03harnad.html

HathiTrust Digital Library. (2019). Hathi Trust Research Center. Retrieved November 4, 2019, from https://www.hathitrust.org/htrc

Hess, C., & Ostrom, E. (2007). *Understanding knowledge as a commons: From theory to practice*. Cambridge, MA: MIT Press.

Kranich, N. (2004). *The information commons: A public policy report*. Retrieved from The Free Expression Policy Project, Brennan Center for Justice, NYU School of Law website: http://www.fepproject.org/policyreports/InformationCommons.pdf

Kwon, D. (2017). Major German Universities cancel Elsevier contracts. *The Scientist* (July 17). Retrieved from https://www.the-scientist.com/news-analysis/major-german-universities-cancel-elsevier-contracts-31208

Lammey, R. (2014). CrossRef's text and data mining services. *Learned Publishing*, *27*(4), 245–250. https://doi.org/10.1087/20140402

Leetaru, K. (2008). Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday*, *13*. Retrieved from https://journals.uic.edu/ojs/index.php/fm/article/view/2101/2037

Lessig, L. (2001). *The future of ideas: The fate of the commons in a connected world*. New York: Random House.

Levine, M. (2014). Copyright, open data, and the availability-usability gap: Challenges, opportunities, and approaches for libraries. In J. M. Ray (Ed.), *Research data management: Practical strategies for information professionals*. West Lafayette: Purdue University Press.

Lynch, C. (2017). The rise of reading analytics and the emerging calculus of reader privacy in the digital world. *First Monday*, *22*(4). https://doi.org/10.5210/fm.v22i4.7414

McDonald, D., & Kelly, U. (2014). Value and benefits of text mining. Retrieved July 31, 2018, from Jisc website: http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

McKenzie, L. (2017). Elsevier makes move into institutional repositories with acquisition of Bepress. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2017/08/03/elsevier-makes-move-institutional-repositories-acquisition-bepress

Murray-Rust, P., Neylon, C., Pollock, R., & Wilbanks, J. (2010). *Panton principles*. Retrieved August 30, 2013, from http://pantonprinciples.org/

Nunberg, G. (2010). Counting on Google Books. *The Chronicle Review*, December 16. https://www.chronicle.com/article/Counting-on-Google-Books/125735

O'Sullivan, M. (2008). Creative Commons and contemporary copyright: A fitting shoe or "a load of old cobblers"? *First Monday*, *13*. Retrieved from http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2087/1919

Pike, G. H. (2016). *Elsevier buys SSRN.com: What it means for scholarly publication* (SSRN Scholarly Paper No. ID 2963709). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2963709

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

PMC Overview. (2018). Retrieved January 22, 2018, from https://www.ncbi.nlm.nih.gov/pmc/about/intro/

Posada, A., & Chen, G. (2018). Inequality in knowledge production: The integration of academic infrastructure by big publishers. In L. Chan & P. Mounier (Eds.), *ELPUB 2018*. https://doi.org/10.4000/proceedings.elpub.2018.30

Rabesandratana, T. (2019). The world debates open-access mandates. *Science*, *363*(6422), 11–12. https://doi.org/10.1126/science.363.6422.11

Reichman, J. H., Dedeurwaerdere, T., & Uhlir, P. F. (2009). *Designing the microbial research commons: Strategies for accessing, managing, and using essential public knowledge assets*. Washington, D.C.: National Academies Press.

Reichman, J. H., Uhlir, P. F., & Dedeurwaerdere, T. (2016). *Governing digitally integrated genetic resources, data, and literature: Global intellectual property strategies for a redesigned microbial research commons*. New York, USA: Cambridge University Press.

Samuelson, P. (2019). Europe's Controversial Digital Copyright Directive Finalized. *Communications of the ACM*, *62*(11), 24–27. https://doi.org/10.1145/3363179

Senseney, M., Dickson, E., Namachchivaya, B., & Ludäscher, B. (2018). Data mining research with in-copyright and use-limited text datasets: Preliminary findings from a systematic literature review and stakeholder interviews. *International Journal of Digital Curation*, *13*, 183–194. https://doi.org/10.2218/ijdc.v13i1.620

Suber, P. (2012). *Open access*. Cambridge, MA: MIT Press.

UC and Elsevier. (2019). Retrieved May 6, 2019, from Office of Scholarly Communication website: https://osc.universityofcalifornia.edu/open-access-at-uc/publisher-negotiations/uc-and-elsevier/

Van de Sompel, H. (2013). *From the version of record to a version of the record*. Opening Plenary Session presented at the Coalition for Networked Information (CNI) Spring 2013 Membership Meeting, San Antonio, Texas.

Van de Sompel, H., Rosenthal, D. S. H., & Nelson, M. L. (2016). Web Infrastructure to Support e-Journal Preservation (and More). *ArXiv:1605.06154 [Cs]*. Retrieved from http://arxiv.org/abs/1605.06154

Wilkin, J. P. (2017). How large is the "public domain"? A comparative analysis of Ringer's 1961 copyright renewal study and HathiTrust CRMS data. *College & Research Libraries*, *78*(2), 201–218. https://doi.org/10.5860/crl.78.2.201

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. Retrieved from http://dx.doi.org/10.1038/sdata.2016.18

Williams, L. A., Fox, L. M., Roeder, C., & Hunter, L. (2014). Negotiating a text mining license for faculty researchers. *Information Technology and Libraries*, *33*(3), 5–21. https://doi.org/10.6017/ital.v33i3.5485

Willinsky, J. (2006). *The access principle: The case for open access to research and scholarship*. Cambridge, MA: MIT Press.

Willinsky, J. (2018). The academic library in the face of cooperative and commercial paths to open access. *Library Trends*, *67*(2), 196–213. https://doi.org/10.1353/lib.2018.0033

Yeager, A. (2018). Sweden cancels agreement with Elsevier over open access. *The Scientist*, May 16. Retrieved from https://www.the-scientist.com/the-nutshell/sweden-cancels-agreement-with-elsevier-over-open-access-64405

Zotero Style Repository. (2019). Retrieved January 24, 2018, from https://www.zotero.org/styles