Metadata and Open Access: Reliably Finding Content and Finding Reliable Content

Sommer Browning
University of Colorado Denver, sommer.browning@ucdenver.edu

Jean-Claude Guédon
Université de Montréal, jean.claude.guedon@umontreal.ca

Laurie Kaplan
ProQuest/Serials Solutions, lkaplan@proquest.com

Sommer Browning, Jean-Claude Guédon, and Laurie Kaplan, "Metadata and Open Access: Reliably Finding Content and Finding Reliable Content" (2013). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315314

Metadata and Open Access: Reliably Finding Content and Finding Reliable Content

Sommer Browning, Head of Electronic Access and Discovery Services, Auraria Library, University of Colorado Denver
Jean-Claude Guédon, Professor, Department of Comparative Literature, Université de Montréal
Laurie Kaplan, Director, ProQuest/Serials Solutions

Abstract

Metadata and open access publishing continue to be topics of debate and discussion in the popular media, blogs, and listservs. Different points of view exist among librarians, researchers, publishers, and others, and several examples will be presented regarding open access journals and articles and digital data from the perspective of metadata and accessibility. Open access content is the utmost accessible content, if students and researchers know how to find it and know how to judge whether what they find is worthy of inclusion in their research. The discussion will focus on how to make open access publications and articles more accessible. Questions the paper will strive to answer are:

- What metadata elements would help academic librarians and researchers find these resources within the larger databases, institutional repositories, and/or discovery services?
- How do librarians vet open access publications for research by students and faculty? How do they determine which titles to include in their catalogs and how to catalog them?
- What additional information would be helpful? What role could publishers of directories and providers of link, search, and discovery services play to that would lead to open access content?
- How can metadata better describe digital data and make it more accessible to researchers?

Introduction

Open access (OA) content has become increasingly important in the last decade and especially in the last few years. As more funding organizations stipulate that research must be made accessible as a requirement for funding, more data has become available. However, there are many ways that the data can be made accessible including institutional repositories, personal web sites, commercial sites, preprints, postprints, and articles in OA or commercial subscription journals.

Open access, which Peter Suber (n.d.) has defined as "digital, online, free of charge, and free of most copyright and licensing restrictions," became a reality between 2003 and 2004 with the first OA journal launched by the Public Library of Science and the beginning of the Directory of Open Access Journals at Lund University in Sweden. Today there are over 9,900 OA journals in the directory from 123 countries. Recently OA and the peer-review process have been popular topics of debate in the fields of publishing, librarianship, and research. One of the latest issues is concern about so-called "predatory publishers" who charge high fees for authors to publish their articles, claim to ensure peer review, but whose practices are questionable. Jeffrey Beall, a librarian at the University of Colorado in Denver, has a list that, in November 2012, included almost 450 publishers and 280 standalone journals that he claims are "potential, possible, or probable predatory scholarly open-access" entities. Are some of the DOAJ journals on Beall's list, leading unsuspecting researchers to potentially unreliable data? What is questionable—all of the research of all of the authors in these journals or just the publishing and peer-review process? Should the author's paper be ignored based on a poor choice of where to publish in the pursuit of making his or her data and analysis available online quickly? And how does a researcher know what materials to trust?

http://dx.doi.org/10.5703/1288284315314

There are several well-known commercial publishers who have joined the OA movement, creating fully open journals or hybrids that incorporate open articles in otherwise subscription journals. Are these publications more trustworthy? And many aggregators have begun to include OA journals in their collections in order to make available in one database all of the relevant subject-related materials for a researcher. Does the inclusion of an OA article by an aggregator confer legitimacy, or is it up to the researcher to determine the quality of the article?

The first section of the paper, by Sommer Browning, Head of Electronic Access and Discovery Services at the University of Colorado, Denver will discuss the access and discovery of OA resources, specifically how Auraria Library determines which OA materials to include in their catalog, how to catalog them, what providers can do differently to help provide access and discovery, and other points around this topic.

The second section, by Jean-Claude Guédon, Professor, Department of Comparative Literature at the University of Montreal, Quebec, Canada, will focus on the use of metadata by researchers. Producing metadata is deeply tied to the nature of the targeted documents. In the digital context, this issue is far more complicated than in print. It is suggested here that three concepts could help: sociology of documents, society of documents, and sociology proper. Furthermore, the concepts of studium and punctum introduced by Roland Barthes also help understand how documents are approached when they form part of complex devices such as a relational database. Finally, while human beings obviously play a role in the design of sociology of documents and a society of documents, it is suggested that, symmetrically, documents can help design and shape new kinds of human communities based on different forms of affinities.

Open Access from a Library Perspective, Sommer Browning

Auraria Library uses OA resources carefully and with a certain amount of trepidation for a few reasons. First of all, much of the discussion around OA journals is centered on scientists and researchers. Of the 45,000 students Auraria serves, the vast majority are undergraduates and commuters, and 80% hold full- or part-time jobs. Auraria Library serves three institutions on one downtown Denver campus: The University of Colorado, Denver; Metropolitan State University of Denver; and Community College of Denver. Auraria's typical patron is not a physicist or a neurosurgeon. The typical library user is more likely enrolled in an associate's degree program or pursuing a Master of Education. This unique demographic informs the library's relationship to OA content.

Secondly, neither the library nor the three institutions Auraria serves have OA policies. Without this kind of institutional directive, faculty are not explicitly encouraged to publish in OA journals. Librarians get few faculty inquiries about OA journals. Oftentimes the questions only concern Auraria librarian Jeffrey Beall's list of predatory publishers, publishers that use deceit and unfair publishing practices.

Finally, the discovery and cataloging relationship with OA resources is not always positive. Auraria uses Millenium, Serials Solutions, and Summon; merely navigating and troubleshooting data across these systems is a complex, time-consuming endeavor. The unreliability of OA metadata and the difficulty in contacting OA publishers makes troubleshooting these resources more onerous. Frankly speaking, time and resources force Auraria to fix what it pays for before it fixes what it does not. However, despite these factors, Auraria recognizes the tremendous value in OA content and provides discovery to thousands of OA resources. It does this mostly by tracking these databases in the Serials Solutions KnowledgeWorks.

Discovery Issues

OA journals have all the usual electronic access issues that other journals have—broken links, missing content, platform changes—but they are more difficult to troubleshoot for various reason. Finding and contacting responsible bodies can be difficult. Many OA journals do not have the

financial backing of large institutions and do not have customer service departments, so finding a working e-mail address can be a challenge. Also, since time and staff are in high demand, librarians are not always willing to "go the extra mile" they usually do when troubleshooting paid resources; checking back in a week to see if a missing issue was uploaded is not an efficient workflow for these resources.

Sometimes discovery problems are not in the library's control at all but are caused by other departments and vendors. Recently, Serials Solutions announced that it is removing HathiTrust from Summon for a 4–6 week maintenance period. In another recent circumstance, Scirus, an Elsevier-run, OA index of millions of scientific items, suddenly became inaccessible because the campus IT department blocked it from all campus computers. The site had scored a bad reputation with campus IT's security software. The Discovery Librarian spent nearly a full day's work corresponding with campus IT, library IT, and reference librarians to try and regain access. A week after access was restored, it was announced that the tool is going to be decommissioned in early 2014. Incidents like these do not help us promote OA materials.

Vendors and OA publishers could help libraries promote OA materials by improving metadata for these resources. The improvements can be divided into two categories: Metadata to find or exclude OA resources and metadata to assist in troubleshooting OA resources.

Metadata to Find or Exclude Open Access Resources

It may be impossible to convey every nuance about how "open" a resource is; however, something more generic, a nod to the way they are categorized in the Serials Solutions Knowledgeworks, for instance, would be a simple, productive step in making OA more visible and usable. This could take the form of an OA tag or icon next to an article's citation in the discovery layer. An icon could bring the existence of OA to the attention of our librarians, faculty, and students, perhaps piquing their interest.

The type of metadata needed to add an icon to OA materials, would also make faceting and limiting a search to OA materials possible. This would be an easy way to identify leading OA journals in a particular field. Being able to limit to OA materials could also have implications for collection development, renewals, and subscriptions. A facet like this would also make it possible to exclude OA materials; this could benefit collection assessment and accreditation reporting.

Another factor that could improve the discoverability of OA materials would be metadata transparency. Because of cuts to cataloging departments, because of the very nature of electronic resources, because of the sheer number of cataloging records e-resource packages have, libraries have had to give up control of their metadata. These metadata are crucial in making discovery decisions, and oftentimes it is difficult to analyze the metadata libraries receive from vendors and publishers. Some questions vendors and publishers could answer are: Where is the metadata coming from? Cannibalized OCLC records, in house catalogers, vendors? What standards are publishers and vendors using for their metadata, for example National Information Standards Organization (NISO) or Library of Congress standards? It is also difficult to analyze and inspect the metadata before libraries load it into their systems. For instance, one must look at bibliographic metadata title by title in the Serials Solutions KnowledgeWorks because there is no reporting function that assesses the quality of the records. Similarly, discovery systems do not allow libraries to easily make customizations to these data, such as inputting local notes, or massaging data so it works best with a particular ILS. The inability to control metadata has implications beyond discovery; it can affect collection management, circulation, and acquisitions decisions.

Metadata to Assist in Troubleshooting Open Access Resources

Oftentimes troubleshooting journals requires data such as contact information, e-mail addresses, and names of editors and corporate bodies. Troubleshooting also frequently considers browser

compatibility, accessibility issues, and the presence of pop-ups. Providing data that addresses these concerns would help all electronic access problems and, most especially, those concerning OA. Auraria currently is experiencing searching errors for all products that use a certain platform. If Serials Solutions could provide data on which resources run on this platform, creating a list of these resources and appending a user note next to them would create much better access for and transparency to our users. Perhaps this kind of data changes often and would be difficult to maintain, but it is exactly the kind of information librarians need on a daily basis. Perhaps giving librarians more control over their metadata would encourage them to update these ever changing technological issues.

How Can Librarians Help?

Librarians, in particular, Discovery Librarians, already help vendors and publishers of OA content by reporting problems with broken links, missing content, and bad metadata. However, librarians can go beyond this by discussing their needs with vendors and publishers and becoming familiar with two forthcoming National Information Standards Organization (NISO) standards. Both the *Specification for Open Access Metadata and Indicators* and the *Open Discovery Initiative: Promoting Transparency in Discovery* aim to promote discovery and transparency in the metadata and indexing vendors and publishers provide. Librarians should feel empowered to discuss these standards with vendors and publishers and should continue to be vigilant about metadata quality.

Metadata for Digital Documents: A Researcher's Perspective, Jean-Claude Guédon

Introduction

The issue of producing metadata in a digital context is a good deal more complicated than what most presentations on the topic cover. The reason for this is that digital documents are not simply printed documents stored on digital media and transmitted digitally. In other words, digital documents are a great deal more than PDFs; in fact the PDF format is derived from the printing protocol. Postscript clearly falls in the category of "digital incunabula"—an expression that Gregory Crane coined quite a few years ago.

A recent article on tagging photographs begins to explain the complexity of the emerging new world. Picking up after an older essay by Roland Barthes, the authors use the photograph of a French African explorer, accompanied by two young Africans, to point out that, besides the obvious hero-worship objective of the photo, there are details one may focus on. For example, one of the two young boys has his arms crossed, and such a detail could—why not?—be tagged. But, in so doing, this particular photo may suddenly find itself associated with other photos only because all include a person with crossed arms. Barthes had analyzed that situation purely in terms of foci of interest by distinguishing between *stadium*—the ostensible major theme of the photo, generally identified in a title—from the *punctum*—particular point of view or perspective selected by some observer. Why is the digital context important in such a discussion? Simply because, if you put the colonial photo studied by Barthes inside a device such as Flickr, the punctum is strongly foregrounded since many other photos are similarly tagged. On the other hand, in a print world, noting the crossed arms will remain an isolated, somewhat idiosyncratic choice with few possible consequences.

In short, the cultural meaning of a digital document will tend to reach well beyond its ostensible meaning and functions, depending on the platform where it resides, the software that structures content into some kind of database, and the means to disseminate it (as well as the limits on such acts of dissemination, such as intellectual property rights that must be respected). To analyze this issue further, I shall appeal to three key terms: a sociology of documents, a society of documents, and a simple sociology. In so doing, some of the missing points will hopefully become clearer.

Sociology and Society of Digital Documents

The phrase "sociology of texts" is well known among bibliographers since D. F. McKenzie's book by the same title has generated a continuing stream of debates from the time of its original

publication. It essentially refers to the fact that ignoring the steps that have concretely accompanied the production of any document necessarily truncates its meaning. Adding the ways in which it is disseminated, preserved, and restored adds a great deal more to the cultural meaning of a document. For example, the quasi miraculous preservation of Lucretius' *De Rerum Natura*—itself the object of a recent book by Stephen Greenblatt—testifies to the fact that this book was anathema to Christianity and could survive only as a clandestine member of any manuscript collection. The result is that only one copy has survived, and that detail, in itself, adds a great deal of meaning to its place and role in the Renaissance: some people knew of it, through allusions or discussions preserved from Antiquity, but no one had ever read it completely. And this poetic rendition of Democritean atomism played a key role in countless debates around materialism and atheism.

The sociology of texts, consequently, provides an axis of analysis, a perspective of studies, that provides a privileged position to the issue of how a document comes to be, comes to be preserved, and comes to be accessed and used. Simply adding these concerns to the issues of metadata points to the richness of the field and the possibilities for countless numbers of new studies.

The society of texts proceeds from an entirely different perspective. The term, of course, is derived from Marvin Minsky's celebrated work on artificial intelligence. Starting from elementary information processing units, Minsky strives to demonstrate that, through suitable combinations and associations, something like an artificial intelligence can be synthesized, so to speak. However, the metaphor as transposed to documents in general, and digital documents in particular, refers to the fact that the very material form of documents may make it more or less difficult to relate to other documents. For example, manuscript scrolls, which really act as little more than frozen memory, tend to make texts behave as isolated, unique entities. Allusions and references to other texts may be found in the text, but retrieving those will be difficult; all the more difficult that, by definition, copies are few in the manuscript world, and they are widely dispersed. In the codex world, new working habits developed, including marginal notes and commentaries. The result was a slightly stronger link between various documents, although copies remained few and just as widely dispersed as before. In such contexts, it is easy to see why libraries, such as that in Alexandria, are so crucial.

With print, the scene changed drastically. Much greater numbers of copies made it far more probable that related texts could be found in the same place or could be identified and obtained. In a sense, Ramelli's wheel showing a Renaissance scholar playing with a kind of office-sized Ferris wheel with books on its shelves symbolizes the completion of a dream and desire: that of making full use of the codex form and of print to engineer the best conceivable society of texts within this technological context.

With digitization, of course, linking anything with anything became the mantra of hypertext, and the transposition of this philosophy on the Internet gave us the World Wide Web: a gigantic society of documents has been evolving with lightning speed since Tim Berners-Lee unleashed this technical concept into the world.

Metadata for a Sociology and a Society of Texts

Obviously, the dual axis of analysis just outlined leads to different kinds of metadata and also to different kinds of producers of metadata. While institutional entities such as libraries or publishers may well provide the more traditional forms of metadata that come with categories, ontologies, etc., thus taking charge, more or less, of Barthes' *studium* perspective, users of all kinds may well prefer to address the *punctum* perspective and point to all the details that they find interesting in any document, however quirky, unusual, or unique. Within documentary spaces as vast and multidimensional as the Web, this interest for the *punctum* will be greatly needed, and, at the same time, it will never fully exhaust the documentary scene of the Web. As was just noted, the distinction between *studium* and *punctum* suggests a division of labour between information specialists and users (for example, researchers, like myself). The information specialist, in this

perspective, would offer the more established, traditional, and well-known pathways through which to navigate the sociology and the society of documents. Readers and researchers (and perhaps—why not?—spy agencies, as well) will explore the *punctum* and will try to make sense of encountered forms of tagging, etc., that may float like a cloud around textual objects. At the same time, it may be that, at some point in their development, *studium* and *punctum* will be able to exchange roles, as Boullier and Crepel argue in the article cited earlier, especially if all these documents are located within a relational database. Let us remember that a relational database simply allows the almost unlimited multiplication of points of view.

What About Sociology Proper?

All that has been said so far appears to leave human beings in a somewhat uneasy and ambiguous situation. So far, like shadowy figures, they have appeared only as temporary and secondary actors. Documents, as presented here, seem to enjoy a fair degree of autonomy; they look as if they could self-create and could relate to each other without much human intervention. This is silly, of course; yet, a kernel of truth remains attached to this vision. It is silly in the first instance because the production, storage, preservation, and dissemination of documents, until recently at least, has remained firmly in the hands of humans. This is particularly clear when the objective is to destruct documents. In the manuscript world, this objective is not too difficult to reach, but the exceptions that exist nonetheless set limits to such a desire. Lucretius *De Rerum Natura* is one such case, as are the Gnostic gospels discovered in Egypt in 1945. The latter were obviously hidden to avoid being burnt, but the job was a little too good since it took about 19 centuries to discover them accidentally. The destruction of printed books is possible but always more difficult, and this difficulty points to a growing autonomy of the technical agency. Finally, with the digital world, the destruction of any document is almost impossible, as no one knows how many copies were made, for example, in the process of sending that document from one computer to another over the Internet and where they sit. In the digital world, as has always been the case with written documents, copying is the way to resist annihilation, but it reaches new orders of magnitude. This point is all the more crucial that every digital document, when taken singly, is particularly fragile and vulnerable. Indeed, a tendency to evaporate allied to rapid technical obsolescence make individual digital documents hard to preserve.

Beyond the role that humans can play in designing a sociology and a society of documents, one must finally conclude by moving the other way: How do documents affect a sociology and a society of humans? There begins perhaps the most exciting task for the metadata that readers and researchers really need: How can documents help form communities? How can documents help form identities? The answer lies in the possibility for machines to construct affinities between documents in such a way that the humans related to them will find themselves invited to join some new community form or will find themselves trapped into some form of profile. The latter suggestion bears some sinister consequences that we would do well to think about if we want to hold on to democratic values.

A quick example with PhD dissertations will help understand how documents can support the building of communities. Imagine librarians systematically producing the concordance of the dissertations in their local repository. They could then remove the 1,000–1,200 words that are most commonly used in a given language. Suppose further that, with what is left, attention would be paid to rare words used relatively frequently in each of these dissertations. By adjusting parameters such as numbers of words retained, frequency of use of these words, etc., one could imagine building a proximity metric between documents. Such a metric would clearly reveal affinities between these documents. The affinities could be refined according to *studium* or *punctum* so that proximity metrics could be mapped onto groups of recent PhDs. Transforming these abstract lists of names into research communities would need some social strategies that are well known and tested, such as conferences, summer schools, etc. The result

would be that young researchers would begin to understand better who their closest intellectual "neighbors" are. Something like the structuring of research areas could also begin to appear. But to achieve goals such as these, new metadata will be needed. It is the cost for allowing databases and complex algorithms to suggest new forms of associations and collaborations among human beings.

The sociology and society of documents needs to be taken into account in the production of any metadata. Symmetrically, the presence of large bodies of documents points to new ways to stimulate and facilitate the formation of new communities and new identities. Here, we are speaking about research, but, clearly, similar forms of reasoning can be applied to vastly different life situations, simply because we are increasingly living in a mixed world of humans and documents, all mediated by computers and algorithms.

Conclusions

While the conference presentation focused more on OA, this paper focuses on metadata and its complexities for digital data and for OA data. Metadata can bring disparate but related documents together and create new research possibilities. However, it can also frustrate a researcher when it is not as precise as it could be and leads to broken links or inaccurate data. Adding metadata to identify more elements about digital data, including a tag for OA publications, will be helpful in the future for discovery and acquisition and can help researchers, students, and librarians find any and all data.

References

Beall, J. (2013). *Scholarly open access: Critical analysis of scholarly open-access publishing.* Retrieved from http://scholarlyoa.com/publishers/

Boullier, D., & Crepel, M. (2013). Biographie d'une photo numérique et pouvoir des tags, *Revue d'anthropologie des connaissances*, *4*, 785–813.

Greenblatt, S. (2012). *The swerve: How the world became modern*. New York, London: W. W. Norton.

McKenzie, D. F. (1999). *Bibliography and the sociology of texts*. Cambridge: Cambridge University Press.

Minksy, M. L. (1986). The society of mind. New York: Simon and Schuster.

Neylon, C., Tananbaum, G., Pentz, E., Koscher, C., Bilder, G., Meyer, C., Smit, E., Shillum, C., Hulbert, T., Dylla, F., Weinrich, D., Tagler, J., Showers, B., Jacobs, N., Chvatal, D., Cox, L., Bide, M., & Devenport, T. (2012, December 18). *A proposed NISO work item: Specification for open access metadata and indicators*. Retrieved from http://www.niso.org/apps/group_public/download.php/9845/Open Access Metadata—Work Item for ballot.pdf

Open Discovery Initiative Working Group. (2013). *Open discovery initiative: Promoting transparency in discovery: A recommended practice of the National Information Standards Organization*. Retrieved from http://www.niso.org/apps/group_public/download.php/11606/rp-19-201x_ODI_draft_for_comments_final.pdf

Suber, P. (n.d.) *Peter Suber.* Retrieved from http://legacy.earlham.edu/~peters/hometoc.htm