

# CHICAGO COLLABORATIVE

## EXECUTIVE SUMMARY

Fall Meeting: November 10, 2010

### ENSURING PERSISTENT ACCESS TO THE SCIENTIFIC RECORD

Marriott Wardman Park Hotel

Washington, D.C.

The fall meeting of the Chicago Collaborative was a special facilitated discussion of invited participants and Collaborative representatives on archiving and preservation. Clifford Lynch, Executive Director of the Coalition for Networked Information, facilitated the informal and wide-ranging discussion. A short business meeting followed the special archiving and preservation session which was entitled “Ensuring Persistent Access to the Scientific Record.” See Appendix A for a roster of special guests, a roster of Chicago Collaborative representatives, and the introductory material received by attendees in advance of the meeting.

Participants agreed to discuss archiving and preservation issues in three broad categories: 1. e-journals, 2. underlying research data, and 3. everything else (grey literature, multimedia educational materials, gene sequences, etc.).

Preservation is a joint effort between publishers and the library community. In the very long run, the library community has to take final responsibility, but the partnership with publishers has to be much closer. Archiving recorded knowledge is important – it’s a given – and the real questions have to do with navigating in an environment of limited resources, setting priorities for archiving with limited resources, and balancing the generation and curation of research data.

#### 1. E-JOURNALS

**Introduction (Lynch):** There are several mature e-journal archiving solutions. These include Portico, CLOCKSS, NLM PubMed Central (PMC), and Koninklijke Bibliotheek (KB) in the Netherlands. These are not as widely adopted as they need to be and ongoing issues include small society participation in e-journal archiving solutions, and issues identifying and gaining participation among the very small startup journals. The scholarly community needs a way to indicate when a journal is being electronically archived, like the infinity symbol that became a standard sign for acid-free paper.

Chicago Collaborative Fall Meeting

November 10, 2010

Page 1

**Discussion:** Portico has 120 publishers with 12,000 journals, but there are many more e-journals – perhaps 22-24K? It is very difficult to find one-off titles and very difficult to use their data. Where are the partnerships in something like that? Libraries can help by posting the names of journals or publishers to the Portico website. Some small publishers can be found through their technology partner (Allen Press, Atypion, MetaPress, HighWire —there are fewer than 10 of these technology partners). Work with SSP, ALPSP to find these journals. Once found, the cost is often more than the organization can pay. It is going to have to be free or actually paid for by the preserving organizations. Need to evolve standards that are easy for them to adopt. Portico approaches small publishers and even helps prepare their metadata. Generate methods of funding this—not funding from the publishers. Small journals are using OJS, Microsoft Word, home-grown solutions. Need to evolve standards—even something as simple as Dublin CORE. Include preservation solutions in courses on open access platforms that are given by librarians at an institutional level; try to get people to use some kind of standard. Go back to authoring tools. Universities are also hosting small journals without thought of archiving. If the library is making this part of their institutional repository, there is often more awareness.

David Gillikin of NLM drew a distinction between “archivable” versus “preservable”. An archiving solution doesn’t necessarily mean that the resources will also be preserved. Preservation is migration of the data so that it is always accessible using current technologies. Not all archive solutions have a plan for updating formats as needed.

NLM has worked with African countries to develop journals; they are not at an archiving stage yet. The amount of research being done at doctoral levels in India and China is another issue which will change, or at least broaden, Western focus. Standards need to be developed now. Does NLM have a preservation strategy? The recently launched Digital Collections site is for preservation, and the other NLM preservation strategy is PubMed Central (PMC). In addition there are multiple backup sites. NLM’s “Citing Medicine” is a tool that provides guidelines on citing resources, and the use of the DOI is an industry standard. Generally, NLM is a big proponent for national standards.

Risk and economic ramifications: Many small organizations are working in portable document format (PDF), not markup language. But archiving organizations can probably work with PDF. David Rosenthal and Vicky Reich are working on this issue. What is the risk if we fail? How do we assess risk? Are the commercial publishers a low risk or a big risk?

The Center for Research Libraries (CRL) has established archiving standards and certification (see: <http://www.crl.edu/archiving-preservation/>). Portico has been certified. CLOCKSS is going to go through process. Guidance around best practices and standards is needed.

There is a need for a database of archived journals such as exists in Europe (PEPPERS [need reference]). This may be an opportunity to work with CrossRef as a registry so that part of the metadata they record indicates that something has been archived somewhere.

Chicago Collaborative Fall Meeting

November 10, 2010

What is a definition of an archived publication? Does it include the links, erratum, retractions? In some fields (engineering, computer science), the abstract collections are the way the science moves forward. Scholarly journals are fairly rigorous about versioning. For archives, is the version that was first published that which is archived, or is the archive the whole trail of information about that article (e.g., erratum, retraction) that exists? The content is preserved by archiving systems, but the access system is not preserved (the links, etc). There is also a need for disaster preparedness systems, not simply an archiving system. Preservation now is at the article level, not the journal level. Doesn't include links to anything that the original article is linked to, not just the erratum links. Systems now archive the traditional article, but the traditional article will cease to exist in the next decades.

Post cancellation access is another slice of archiving. Delivery is a nightmare. How much involvement do archiving services want regarding the access piece? CLOCKSS made a decision is to have an open access system for triggered content. It is harvestable, so that if someone wants to create an access platform to that content, they can do so.

Historians are concerned that they will not be able to access the look and feel of the original journals or websites. But what does that mean for an online journal? Everyone sees a different page now with personalized ads, blogs, etc.

Who will provide the standards and tools for preserving, including authoring tools? The preservation entities are talking about this all the time, including the British Library and KB. These are not regular meetings, but meetings are generated around specific topics. Umbrella organizations include the Digital Preservation Consortium (morphed to Digital Library Federation: <http://www.clir.org/dlf.html>).

How can libraries help educate researchers and students? A single communications methodology is not going to work. To the extent that libraries talk to faculty about open access and journal start-ups, they can put archiving on that agenda. Also, digitization of theses can start this conversation. Graduate students are particularly hungry for information that is not published anywhere else. How can library organizations move into this? Portico has a lot of material about preservation on their website, and it is all freely reusable. Portico puts on regional forums for librarians which cover preservation issues. The issue needs to be raised up a level as a set of "Best Practices".

TRANSFER was set up by UKSG (see: <http://www.uksg.org/transfer/about>) to address the issue and establish a code of practice for journals transferring from one publisher or platform to another. Spreading the word through the peer review systems or open source publishing platforms may be reasonable communication strategies.

## **2. UNDERLYING RESEARCH DATA**

**Introduction (Lynch):** Another major piece of the scholarly record is the underlying research data. There is a growing awareness that the data generated by the research process is an important outcome and

Chicago Collaborative Fall Meeting

November 10, 2010

needs to be preserved and made accessible, but many questions remain about roles and responsibilities. Players include funding agencies and their requirements for data sharing (e.g., NIH, NSF, and Wellcome Trust) and the role of institutional repositories. Publishers are uneasy about commitments for access, preservation, and reuse of supplementary data. There is a distinction between different types of scientific data. Some data such as massive astronomical data sets may have no people or commercial interests associated with it, but other scientific data such as that funded by NIH will have all sorts of constraints related to it -- people, privacy, commercial interests, IRB, HIPPA – which makes this data the most challenging.

**Discussion:** The underlying data needs to be thought about as a strategic product. There is a need to be clearer about reusability of data. NIH has been requiring data management plans; NSF will require funders to submit data management plans (see: [http://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=116928](http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928)). NSF is obsessed with starting new science, hard to get them to fund preservation; the NIH mindset is quite different.

Supplemental data: everyone is uneasy about it. Publishers are not sure what archiving solutions exist. Sage Bioinformatics Commons and GenBank are examples. Institutional repositories are also dealing with research data at various points in its life cycle. Biomedical data often has commercial implications and has people implications—identifiable patients, privacy issues, IRB.

Most data proposals of funding agencies are not prescriptive, but ask researchers what they would like. Or funders will indicate that they need to keep data for a specified time period, for example 3 years past grant funding or past publication, mostly for reproducibility of work. Longer term preservation is more about reuse. Need to help faculty think about data lifecycles—how long to keep it and why keep it that long. Publishers also have an important role to play. Some are getting more aggressive about this for ethical reasons.

What are copyright implications? Data can't be copyrighted, so it's mainly about possession. Attribution is not a legal issue; it is simply a custom. Faculty handbooks will often address this. There are overlays on institutional policies having to do with grants. There are sometimes contractual overlays (e.g., disease advocacy organizations are very aggressive about making data public as soon as possible). HIPAA has become very active about constraining use. How to balance privacy and depersonalizing of data with the desire to capture data for potential reuse? IRB would insist that researchers have to spell out EVERY potential use of data, and it would be difficult to get agreement from patients. There are also state requirements.

The Society for Neuroscience is disallowing supplemental data now, partly because of the ambiguity about whether supplemental data is peer reviewed. NISO and NFAIS are working on best practices. A business group and technical group are working on this. Is supplemental data peer reviewed? Publishers' policies need to be recorded and tracked over time. There can be a fuzzy line between the article and the supplemental data—depending on the type of research, one or the other is primary.  
Chicago Collaborative Fall Meeting

November 10, 2010

There is a distinction between research and quality improvement. Quality improvement and comparative effectiveness studies generate huge data sets. Real time reporting is an issue and massive data sets will keep evolving over time.

Those running institutional repositories are now worrying about whether the data is documented and whether data validation and curation are occurring. There is a Federal group for digital data sharing and interoperability called the Interagency Working Group on Digital Data (see: [http://cendi.dtic.mil/presentations/01-06-09\\_CENDI\\_Presentation\\_IWGDD\\_Status.pdf](http://cendi.dtic.mil/presentations/01-06-09_CENDI_Presentation_IWGDD_Status.pdf)). AAHSL is working with AAMC in this area. Many institutions are waiting for standards to be developed. Example of this is electronic dissertations and theses. Graduate studies administrators having trouble advising students. The CTSA funded institutions are taking the lead on this.

**Data citation and referencing issues:** There is a need to cite out from an article to a data set and a need to create a citation to data in supplemental files and data residing in institutional repositories.

There is a need to give researchers credit for use of data. Data cite work is gaining some momentum. It uses the DOI. It is trying to be a CrossRef for citing data. There are unresolved DOI issues; are journal supplemental data given separate DOIs and if so, are they linked to the article DOI?

University of California system has EZ ID (see: <http://n2t.net/ezid>) which is a tool that faculty members can use to create a DOI-like reference for a digital object. The Merit system allows the data to be exposed.

Purdue and MIT are doing a lot of work on how to do data sharing. NSF as part of its cyber-infrastructure program has Interop Grants to help out communities that want to do data sharing (see: [http://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=503141](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141)). There is a model of data sharing on Alzheimer's at NIH and academic institutions and industry.

Is there a role for ICMJE Editors in police the standard, like clinical trials? Should scientific societies play a role in alerting members? NISO and the eScience initiative should probably be involved.

### 3. EVERYTHING ELSE

**Introduction (Lynch):** In addition to the published literature and the underlying research data, there are other scholarly outputs that need to be preserved. These include extensive teaching materials, complex multimedia materials, grey literature, gene sequences, ACCME requirements, and commentary or polling related to articles.

**Discussion:** There is medically related information on the web that is not published in scholarly medical journals. Material includes patient information resources from societies (e.g., American Heart Association), patient blog sites, websites updated sporadically, and popular press pieces about diets, prescription drugs, online pharmacies, smoking control, etc. Archiving issues include versioning when

Chicago Collaborative Fall Meeting

November 10, 2010

the content is simply a website that is updated sporadically and the preservation challenges of grey literature including blogs, social networking sites. Are there any special archiving requirements for continuing medical education content or and other teaching materials?

Citizen journalism is done through blogs, Flickr, Facebook, etc. Google is not in the preservation business and the Internet Archive takes snapshots. Many social networks have made it difficult to harvest them. Twitter, however, has decided it wants to be archived and has given some content to Library of Congress and Google.

Bloggers want to be read and preserved. The Library of Congress preserves blogs selectively, and asks permission first. Internet Archive archives blogs if they don't have a "don't-archive" flag. There are only 4 or 5 blogging platforms, so they could put a flag or button there that allows bloggers to choose to be auto-archived.

The Internet Archive takes snapshots of some pages at random times, but is also doing depth archives in certain fields at certain institutions. The University of California has a web archiving service that is topical.

List of questions familiar to traditional publishers should be given to those running institutional repositories. They might not agree on the same answers, but the list of questions is important for them to have.

**SPECIAL MEETING ADJOURNED:**

Special guests and Chicago Collaborative representatives were thanked for their participation in the special Chicago Collaborative meeting on archiving and preservation. Following a group lunch with all attendees, Chicago Collaborative representatives held a brief business meeting.

## CHICAGO COLLABORATIVE

### EXECUTIVE SUMMARY

Fall Business Meeting: November 10, 2010

Marriott Wardman Park Hotel

Washington, D.C.

#### Follow Up on Special Meeting

Representatives indicated that they had a far greater appreciation for the archiving and preservation landscape. There was a sense that many issues remained following the special meeting, and continuing attention to this area will be needed. There was a desire to continue a dialogue with NLM regarding its PMC preservation policy. Other outstanding issues included the need for education of small journals and societies about archiving and preservation. The issue of standards and best practices remain open issues with NISO the obvious organization to develop standards.

Representatives agreed that the notes from the meeting should be provided to all participants.

#### Education Activities Update

Brief updates were presented on the various educational activities of the Chicago Collaborative.

- Biomed Pub 101: A paper resulting from the June 2010 NASIG presentation will be available. At the recent Charleston Conference, not many people of the approximately 35 people in the audience were librarians. The section on ethics was eliminated due to time constraints. Presenters were asked to post content to the Charleston conference website. The intention is to post the December 7, 2010 planned webinar to the CC website and request Charleston to link to it. The December 7 webinar will be available in 3 RML regions and was the result of discussions at the New Mexico meeting on scholarly communication.
- MLA CE Course on Biomedical Publishing: The MLA proposal deadline for CE courses to be presented in 2010 is Dec 2010. Representatives agreed that it would be worthwhile to submit a proposal for a 4-hour MLA CE course. Jean will submit. Representatives suggested another approach for BMP 101 to appeal to librarians who want/need to be publishers. Give the CE Course a new title: "So you want to be a publisher? Find out what's really involved." This might be the best hook – better than BMP 101. Further justification for the MLA CE course is the interest being shown in registrants for the December 7, 2010 webinar. Currently, approximately 50 have registered.

Chicago Collaborative Fall Meeting

November 10, 2010

- Libraries 101. Volunteers are needed to fully develop this CC education offering. In addition to library faculty, publisher input is needed since this offering will be developed with the publishing community in mind. Norman and Paul agreed to assist with the development and Scott agreed to continue to work on this course. The audio from the Society for Scholarly Publishing session in June in San Francisco is available on the SSP website and this content may help with the further development of Libraries 101

### **Spring Meeting Date**

The Chicago Collaborative spring meeting will be in Chicago, most likely in April 2011, but avoiding April 9-13, 2011, and Passover.