The Library as Publisher?: Collecting and Curating for Digital Repositories

Abstract

The traditional role of libraries as aggregators, curators and disseminators of resources has been profoundly challenged by the notion of libraries as publishers of content. While publishing models include the ideas of aggregating, curating and disseminating content, these terms have radically different definitions in each context (libraries v. publishers). While we understand the motivation of publishers and their role in selecting and distributing content, much less is understood of the role that libraries should, or could play in the chain of "publishing" involved in their hosting of institutional repositories.

This paper will explore:

- the idea of publishing in the context of new models of library sponsored resource delivery;
- the challenges facing libraries in identifying, selecting and curating content
- emerging models of IRs especially in regard to discipline-based content

Presentation

SLIDE 1 TITLE

Under the dictionary definition of publish, publisher, etc…(Merriam-Webster Collegiate 11th edition) we, libraries that host institutional repositories or other kinds of digital libraries, qualify as entities who publish as in:

Publish: To make generally known, to disseminate to the public, to produce or release for distribution.
Publisher: one that publishes something, esp. a person or corporation whose business is publishing.
The 11th edition then ups the ante when they define:
Publishing; the business or profession of the commercial production and issuance of literature, information, musical scores or sometimes recordings, or art…

But we, in academic libraries, do not necessarily share the business model of electronic data publishers – we are not Ebsco or Elsevier or Proquest, although we do share some traits as far as scholarly and academic content –we do not seek profit from academic pursuits. We do, however, need to live up to the services we have advertised and developed with promises of ever-ready and perpetual access to scholarly materials.

Like e-data publishers, we do need to establish and maintain submission protocols, policies and the personnel and services to support this, but we don't need to review material, send it out for editorial decisions, peer review, revisions and re-formatting.

Like e-data publishers, we do need to justify the existence of these projects… we do not necessarily need to sell advertising in order to support our efforts.

SLIDE 2. Examples of web searches that have brought searchers to our IR
Like e-data publishers, we also build and maintain digital platforms that offer some form of structured searching; although with OAI-based metadata harvesting protocols the vast majority or our searchers (>80%) are arriving at our IR pages from search engines and referring websites, not from searches within the IR. This is both re-assuring, that metadata harvesting works well, and slightly disturbing when searches are sampled in our analytics

Without delving into the history of open-access, the pressures of ever-escalating journal and database packages or any of the motivations that have escalated libraries involvement in the dissemination of digital data –I want explore this morning the situation we are in today –where many institutions have adolescent or mature IRs or other digital dissemination programs and what effect this has and will have on libraries as these services continue to develop and evolve.

I have randomly defined two areas to frame our discussion his morning –these are 'Meta-librarianship' and plain old 'Librarianship.' This latter term refers to the operational structures and procedures that routinely occur within our trade, while meta- librarianship I am defining as those issues which reside outside of or "above" those operations, as in for example, relations between libraries and scholars

META LIBRARIANSHIP
Relationships between library, scholars and research have been profoundly and irrevocably altered with the evolution of libraries form collectors to disseminators. We have positioned ourselves as repositories and curators in perpetuity (digital preservationists please hold the tomatoes). We have "outreachers" outreaching for stuff to ingest to these repositories. We have copy rightists proselytizing new models of scholarly communication in order to attract business and content; and, in some ways, we have posed a direct challenge to commercial for-profit electronic publishing entities through the establishment and populating of our repositories, and this has already altered the playing ground and will continue to do so as libraries take on an even bigger role.

Now –there are necessary caveats and cautions:
I have yet to hear a department chair or dean tell junior faculty person, "Go on home and play with your kids, just shove some stuff in that repository thingy and that tenure thing will take care of itself." The pressure to publish and establish tenure is still not on the side of free and open distribution from the start of the publishing process –that remains squishily outside of our purview –for now…

The other caution I would raise is that in so many ways, we are duplicating our efforts, as we have often duplicated efforts in building traditional book collections. Scholars publish –perhaps in paper, or electronic or both, the vendor provides access, we provide access in our repositories, etc…
In the case of particular types of information, for example digital geospatial data, there are many entities–mainly government agencies for the most part, the US Dept of Agriculture, US Geological Survey and a multitude of state agencies who have been collecting and disseminating this data and doing a creditable job of it for some time now. We often duplicate those efforts as well.

SLIDE 3 Doesn't she know it is all FREE on the Web?
Our relationship with the world outside of libraries has also irrevocably changed vis-à-vis our relationship with our patrons. Who are our clients now? Anyone with a connected machine really. Via web-service delivery we must include anyone capable of reaching us via the web as a customer.
Their expectations are that everything should be free and always available; and that trend is not diminishing. The trends of bit torrent, kazaa, napster –grokster, peer to peer file sharing are the trends that are not going away, but gaining in popularity and influence. So we will continue to be challenged to host "free scholarship" which is vastly different than hosting licensed and "not free" content.
Which brings me to a brief discussion of how this affects the day-today business of librarianship.

LIBRARIANSHIP

We are now not only collecting both the fuel for scholarship, as we always have done, but we are hosting that which is produced from the ensuing combustion. The challenges we face day-to-day are not necessarily all new ones :
Preservation : We have grappled with the problem of information impermanence for centuries, but now it's more complicated I think we could agree. What are we digitally preserving and how? In some ways it seems we are busy collecting and hoping, as we did with print materials, that eventually we will get back to preserving and protecting these collections of objects. We are a long way off from establishing coherent methods and the pile of data is getting bigger.

Staffing & Workflows: It is critical that we scrutinize current operations and staffing and re-configure work units and workflows to meet these new obligations. We might repurpose existing staff, much like The Ohio State University which shifted serial check-in staff to half-time copyright analysts in order to ingest faculty materials as they began to populate their IR. We might hire staff with new skill-sets, and I noticed while I prepared for this talk that the Graduate School of Library and Information Science at the University of Illinois offers a specialization in Data Curation as part of their MLS program, which is already quite strong regarding digital libraries, metadata, etc…

But, we need to rapidly ramp up staffing and skills in areas that we have not necessarily been focused on in the past. Re-structure working units, so that staff with the expertise in server administration, preservation and scholarly communication can come together in nimble and pragmatic units to formulate and sustain access and discovery for digital data.

Finally, we need to convince administrators –at the campus level- that we are performing as trusted repositories that require the appropriate levels of support in order to sustain that trusted status.

Because what we have done is open the door, or should I say portal, for libraries to evolve as IT and web services are evolving and we have walked through it eagerly and in anticipation of this new role of digital archiving and disseminating, and I'm not sure we noticed the *abyss* that lies ahead.

SLIDE 4 Dante: Abandon Hope All ye Who Enter Here ("Lasciate ogne speranza, voi ch'intrate")
That abyss – is the somewhat inscrutable challenge encapsulated by the term "Data Curation." Dante's Divine Comedy lists nine circles in the Inferno ….I think, with the new role of digital data curation and dissemination we have a solid candidate for that 10th circle. Those are bags (in the slide) full of datasets.

We are no long talking, or should no longer be talking about storing, describing and serving up .pdfs of post-prints, grey literature, or even complex sound and image files. What we need to rapidly turn our focus to is the ingest, description and storage of large and often interactive datasets.
The curation of data –and by that I mean the persistent and perpetual storage and access to data is by far the most profound challenge yet that we have faced in this publishing business.
To quote from the I-school at Urbana Champaign "Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for re-use over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation."[i]

The scale of the challenge is unthinkably enormous, especially in light of the costs associated with developing cyber-infrastructure processes, soft and hardware. For libraries there are additional challenges –mostly based in the costs associated with personnel and building systems to handle the load. So what is this load? What is this stuff and how do we create the conditions for stewardship?

SLIDE 5 Data, Data, Watson!
As I mentioned earlier, there are types of data and datasets that are complex and require enormous amounts of computing capacity –not only for hosting, but for services to the user. Geopsatial data, human genome data, hydrologic data including attributes for geology, stream flow rates, water quality and geographic position to name a few present special challenges in order to allow for re-use, interactivity and the preservation of the native or initial work. Our scholars are increasingly dependant on rapid and seamless access to this type of data.

Not only are we challenged by the type of data but also by the platforms we have employed, which, for the most part will be limited to the description of the data –with a link to the object or stuff that must reside elsewhere. It is the metadata that will make or break successful access and use.

SLIDE 6 Everything…All The Time
This is an unfortunate quote from an Eagles song… but somehow appropriate. This is the new exabyte transmission device we have developed here at the UT Libraries. It was a soft launch…

A 2008 International Data Corporation white paper estimated that there were approximately 281 exabytes of digital data in existence in 2007 An Exabyte is 2.25 times 10 to the 21st power of bits.[ii]
That is a lot of stuff and the challenge of storing, accessing, managing, preserving some part of this so that our scholars can search, analyze, model, mine, manipulate and re-use it will require a massive collective and collaborative effort. It will be especially important for trusted, service-based, not-for-

profit entities (versus private corporations like that one which desires to be the world's library) to be actively engaged in building the cyber-infrastructure necessary for data curation, as this data will be crucial to serve the collective public good –much of this data is also publicly funded so there's that part too.

Private-Public partnerships, **AUTHENTIC** partnerships will be necessary in order to successfully employ a business model that benefits the collective good. Google and Microsoft are already building immense data storage centers, and while the price for storing data has rapidly diminished (the same IDC report estimates a terabyte of data in 1997 cost roughly $440,000.00 and in 2008 a terabyte drive was priced around $400.00) the ancillary costs for expertise, power, curation and metadata have increased.

We all know the rate at which digital data is created is far greater than our capacity to store it –so consequently, collaborative scholarly community appraisal for that data which is necessary for our stewardship will be the alpha-omega of our collection development policies. Necessarily then, new collaborations will form amongst scholars and those who support their work. Someone who works on fruit fly genetics has data, shares data and collaborates not with other biologists in his department, but with other scientists who share his research interests or specialty. New collaborative models then must evolve, collaborations between the data-producing community and the data archiving community.

To Conclude –I really have no conclusion, other than we are faced with a tremendous opportunity to re-imagine the work of the library and the librarian and it comes at a terrible time for budgeting new initiatives. I don't think we can use models that we have now, it has to be collaborative across and between institutions and disciplines. Thank you

[i] Master of Science: Specialization in Data Curation, University of Illinois GSLIS web page. Cited on January 28, 2010 http://www.lis.illinois.edu/programs/ms/data_curation.html

[ii] Gantz, J. The Diverse and Exploding Digital Universe.White paper. International Data Corporation, Framingham, MA, Mar. 2008; www.emc.com/collateral/analyst-reports/diverse-exploding-digitaluniverse.pdf.