University 2.0

John Unsworth

October 10, 2007

The Information Railroad:

One of the major challenges facing universities in the next decade is to reinvent themselves as information organizations. Universities are, at their core, organizations that cultivate knowledge, seeking both to create new knowledge and to preserve and convey existing knowledge, but they are remarkably inefficient and therefore ineffective in the way that they leverage their own information resources to advance that core activity. In what follows, I will explore some of the ways that the university could learn from what is now widely called "Web 2.0" -- a term that is meant to identify a shift in emphasis from the computer as platform to the network as platform, from hardware to data, from the wisdom of the expert to the wisdom of crowds, and from fixity to remixability.

In approaching this topic, however, I would like to begin with a summary of a very interesting article on what will seem like a very different topic. This article was published back in the spring of 1986, well before Web 1.0, by Ronald J. Zboray: it is entitled "The Transportation Revolution and Antebellum Book Distribution Reconsidered" and it was published in American Quarterly 38.1 (and today, you can find it on JSTOR, and its first page is also on Google).

As is customary in academic writing, Zboray begins by recapitulating the accepted understanding which his article is meant to modify. He writes,

> The transportation revolution has generally been credited with nationalizing--some may even say homogenizing--literary life in the antebellum United States. Prior to improvements in modes of transport, book production was highly decentralized, with numerous secondary cities supplying reading matter to their immediate hinterlands. The regional orientation of production inspired by this decentralization was, according to the traditional view, disrupted early in the

> nineteenth century by the first wave of the transportation revolution. A truly national reading public came into being and with it, presumably, a truly American literature.

His second paragraph then summarizes the modification he proposes to that view:

> While this view certainly does hold a great deal of truth it contains much oversimplification as well. The broad concept of the 'transportation revolution' obscures the special role the railroad played in changing patterns of literary dissemination in antebellum America. The improvements in road and water transport that characterized the early transportation revolution did little to facilitate the circulation of literature on a national scale. Books continued to move along many of the same avenues and in the same tenuous, seasonal manner as they had in the days of the colonial book peddlers. But it was the railroad that improved the regularity of communications upon which the emergent discount/commission relationship between central publisher and local bookseller depended. Also, the year-round regularity of rail communication permitted a national periodical literature in which publishers could advertise their books.

So, while the railroad was important in moving books more easily, it was (Zboray argues) at least as important as a communications technology that lowered what we would now call the transaction costs of publishing, and that allowed advance marketing to reach a broad audience. Moving the stuff, moving the money, moving the marketing: each of these was made easier by a single technology--the railroad--but if we assume that railroad service was the same in all parts of the country, then our railroad-based understanding of the changes in publishing would fail by assuming homogeneity instead of understanding heterogeneity and its effects. On that last point, Zboray says:

> the regional orientation of book dissemination did not diminish as much as the traditional view would have it. Instead, the coming of rail transformed the regional orientation of literature, so that it conformed to the different levels of rail development in the North, South, and West. With conditions of literary distribution differing in each of the three regions, the very idea of a truly national reading public in antebellum America may itself be an oversimplification.

Information Friction:

The Zboray article is a nuanced discussion of what I would call "information friction" in the 19th century--the factors impeding the movement of information in various forms from one place to another, and the lubricating effect of a new technology on the co-efficient of friction for different materials interacting in a system.[1]

Now I would like to focus on information friction in the present, and in a system that, like the one Zboray discussed, includes publishing, but also has other important components. And while for Zboray the larger frame in which publishing existed was the American Reading Public, for my purposes that larger frame is something more concrete, though perhaps no more homogenous, namely the 21st-century university. My argument will be that universities, out of which the internet (and its low-friction protocols) originally emerged, seem in their own information practices to gravitate to monolithic information systems which promise to be seamless but in practice prove to be less flexible and ultimately less innovative and interesting than the granular and remixable information services now often called Web 2.0. And while these monolithic systems may offer more control to administrators, they prevent innovation by faculty, staff, and students, and they ultimately make the university comparatively inefficient as an information organization. By "comparatively" here, I mean compared to other kinds of information organizations--for example, Amazon, Ebay, Facebook, FedEx, Flickr, Garmin, Google, the Internet Archive, Nasa, the National Oceanic and Atmospheric Administration, PayPal, Random House, Second Life, UPS, the US Postal Service, the Weather Channel, Yahoo, YouTube, etc.. At programmableweb.com you can track a mashup timeline: the number they knew about increased from 1800 to 2400 in a recent six month period, and the "mashup matrix" is 125 APIs long on each axis as of October, 2007. Google alone has 27 APIs for mashups; Yahoo has 24; how many does your university have?

I will admit that, as an administrator, I have sometimes thought that universities, like religious

1 "Information friction" is also discussed by David Glazer in an entry in his blog, at: http://dglazer.blogspot.com/2005/11/coefficient-of-information-friction.html. When writing this, Glazer was the unemployed former CTO at Open Text; he is now director of engineering for Google.

orthodoxies, may be designed to inhibit change in order to ensure the perpetuation of the values they represent, in which case what I'm discussing here would be a feature, not a bug. But be that as it may, it is demonstrably true that universities exhibit relatively high information friction, and it seems worth asking what would be different if that friction were reduced.

Most faculty have, I think, had the experience of going to a conference in some far-away place, only to discover that they share a research interest with someone else at their own university—but at the same time, are any two discipline's conference abstracts prepared in the same way, and could we actually cross-reference them? That's anecdotal evidence that we need better information exchange within and across universities. So is the fact that, as interdisciplinary educational programs become more important, it is surprisingly difficult to inventory courses and collocate syllabi in a given topic areas across colleges, schools, and departments. For that matter, could our students import their class schedules into their personal calendars? And how many of us work on campuses where there are multiple incompatible calendaring systems in use? And how often can we actually follow citations in online journals to articles in other online journals as reliably as we could, with print, by trudging through the stacks? What methods do we have of understanding the intellectual heritage of those around us (who was my colleague's disseration advisor?) and how can we know who else is connected to them in that heritage? Where is the RSS for CVs? What tools do we have for assessing the nature, extent, and value of informal collaborations and co-authorship? What tools do we have for actually managing the threads and throes of our email lives, now so totally out of control?

I have been a part of a community-based standard's development process, in the Text Encoding Initiative, and I am familiar with both the value and the shortcomings of that approach to solving the kinds of problems I have just enumerated. Certainly there is a very real value in having the discussions that inform such standards, and in embodying the outcomes of those discussions in guidelines, at least when the subject matter is of central disciplinary and intellectual interest to scholars (so, in the case of TEI, the ontology of literary and linguistic texts), but there are well-known problems with approaching information integration as a matter of uniformity in metadata or markup—beginning with the problem

that in the real world, even when you aim for uniformity, you do not achieve it, so tools and techniques that depend on it will break. Beyond that problem, there is the fact that there will not always be intrinsic intellectual interest in ontological debates around all of the data-types that are important to universities as information organizations (calendars? CVs? Syllabi?).

Sometimes, in these cases, there are other imperatives that drive, or try to drive, integration—usually administrative ones. In the two universities where I have worked in the last fifteen years, hundreds of millions of dollars, quite literally, have been spent to license, customize, and deploy enterprise resource systems that aim to integrate all of the administrative functions of the university under one schema, one ontology, one monolithic information infrastructure. Searching for "information friction" on Google, I actually found one such system, though not one aimed at universities:

> ComFrame removes the friction that inhibits smooth operations, seamless communications and optimal productivity. We create solutions that break down enterprise information processing barriers, helping you move forward.

Even in this brief description, "seamless" is a red flag: the promise of "seamlessness" is premised on uniformity, on total control over the generation and use of information resources: the notion is that you run an empire, and your problem is unruly principalities—if you could only subjugate them all to one information regime then the emperor's new clothes would be...seamless. However, the reality of our information ecosystems today is that they are not closed systems but open ones: no university, for example, generates and controls all of the information that is important to its faculty, its students, or its staff. That is a simple and undeniable fact, and from that follows the observation that a 'seamless' information management environment can only be some kind of terrarium, artificially closed off from the world around it. What we need instead, I would argue, are "seamy" information systems, designed for cobbling together new information services, showing (even foregrounding) their discontinuities, but doing so in ways that encourage people to think about new ways to draw this or that bit of information into this or that information context and provide this or that information service.

An excellent example of this kind of flexible and innovative information service in a university context is something from the University of Wisconsin-Madison's library, now also being adopted at my university, called BibApp. Strictly speaking, BibApp involves more pre-fetched and pre-processed data than mash-ups usually do, but it is certainly in the spirit of mash-ups, in that it cobbles together a number of information sources to do something interesting and new. Here's a brief sketch of what it does, and what you can do with the result.

BibApp starts by using the Rails framework to build a generic bibliographic database for storing information about publications. Next, you pull in faculty information from the campus LDAP (Lightweight Directory Access Protocol) server, you sort those individuals into generic groups (representing things like departments) using information that also comes from LDAP. Next, using the names of faculty and what a human being knows about the standard online publications or disciplinary databases for the departments in which those faculty work, you get their published papers and the subject-headings under which those papers are listed in the disciplinary database. Using the published papers, you parse citation records, look for co-authors, go to Sherpa and get archival data for authors.

Having assembled and stored this information, here are some of the things you can do. With subject-headings, you can generate tag clouds showing you what are prominent topics for individuals, research groups, or departments; since you have publication dates as well, you can actually show an evolving timeline of topics of interest for an individual, research group, or department, and you can show what journals commonly publish research from a given individual or group. With publisher-supplied DOIs (Digital Object Identifiers) you can generate URLs, get the file, or link to the citation. You can search for a subject and find faculty members who work in that area; zeroing in on one faculty member, you can follow link to his or her home page (listed in LDAP), see his or her types of publicaitons (80% journal articles, 20% conference presentations, what subjects are of interest lately). You can expand out from one faculty member to the department, the college, or the campus, to see who that person is publishing with at the university, and you can then see what groups those people are in and what those groups are doing. If you wanted to, you could use that information to generate network

graphs of publishing patterns. Presumably, you could also have RSS (Really Simple Syndication) newsfeeds for people, for groups, if those were listed in LDAP, or the BibApp host could serve them up (or, for that matter, provide RSS feeds for subjects of interest).

In their YouTube-style presentation on code4lib.org (http://code4lib.org/2007/larson) the developers of BibApp, Eric Larson and Nate Vack, admit that this is still experimental code, but part of the point here is that it will always be that, and it should always be that. BibApp is not a "seamless, enterprise-wide solution," it is a seamy stitched-together mashup. On the other hand, it didn't take three to five years to put in place, nor did it cost a hundred million dollars: BibApp represents work done in less than a year with a $10,000 grant and a few people "willing to work weekends."

One of the challenges for a project like BibApp, as its developers freely admit, is that apparently simple things like names of persons are not simple in practice, especially when you want an automated process to use them as identifiers or disambiguators. You may have a lot of Michael Smiths on your campus, and when these Michaels publish, their names may look exactly alike or—at least as often—the problem may be that the same Michael Smith sometimes publishes as Michael Smith, sometimes as Michael J. Smith, sometimes as M.J. Smith, etc. Likewise, citations are not always marked off from other text in a way that would make them easy to pull out, nor are they formatted consistently across books, journals or disciplines. Real data is messy, in other words. These are problems that people doing bibliometrics have been dealing with for a long time, and there are contextual strategies for mitigating some of the mess, but you could also allow authors or other users to suggest corrections, and you could use visualization techniques to spot outliers (the one article by Michael Smith about veterinary medicine among a host of others on mechanical engineering) and examine them. In the monolithic systems approach, or the top-down classification and encoding approach, the usefulness of the system depends on the accuracy of the data or the data representation, but in a mash-up, people are willing to trade some of that accuracy for increased functionality and the flexibility to extend that functionality when interesting new opportunities present themselves. Perhaps most important is that the application itself be fault-tolerant and not require highly structured data.

I have another example, conceptually related in some ways to BibApp, and possibly even a service that could be tacked on to it. This example is a project of my own, currently in mothballs, called BRAIN, which stands for Better Repositories Are Information Networks (see http://brain.lis.uiuc.edu). BRAIN is a peer-finder for institutional repositories, and it was designed to provide an incentive for faculty to deposit their materials in those repositories. Here's how that incentive scheme works. When a scholar deposits a document in an institutional repository that participates in BRAIN, he or she gets back a list of documents from any open-access repository or journal that best match the material deposited, based on several different relevance measures (coincidence of citations, overlapping vocabulary, plus a variety of full-text clustering techniques). Because data-mining can be time-consuming, we serve up the relevance information to participants by email, rather than as a real-time web service, but we can format that email to provide links directly to full-text content of relevant open-access materials. BRAIN itself would not republish any of the material it aggregates, beyond the metadata. We have built a prototype system for BRAIN, mostly from open-source software, with some key pieces contributed by other academic software developers (Michael Jensen at the National Academies Press, and Dave Eichmann at the University of Iowa), and plenty of glue-code written by students at UIUC, again in under a year, with a small grant. You can try it online, though you'll find that relevance is problematic depending on how well the subject area of your submission matches what's in the repository, and particularly so in the measures that depend on citation extraction, which needs more work. Even this small experiment exposed some interesting problems, though, in the real world of institutional repositories: two in particular come to mind.

The first problem BRAIN had with the real world was that Institutional Repositories are a largely undifferentiated mix of primary materials from library holdings and scholarship or research findings from current faculty. There's no attempt made to separate or distinguish the two, which means that if you go out to hoover up everything you can find in open-access institutional repositories, which we did, you get a lot of stuff but only a relatively small amount of it is scholarly or scientific literature of recent vintage.

The second problem BRAIN had with the real world was that a lot of people apparently put image-only pdf files in their repositories. To do vocabulary comparisons, citation extraction, and full-text clustering, BRAIN wants to use PDFBox to extract the text from PDF files, but that doesn't work if the PDF is a picture of a text rather than a text.

These are the sort of problems that will probably decrease over time, but the first one might only decrease if there were information services for which it was important to distinguish recent scholarship from older library holdings, and the second might only decrease if the percentage of recent scholarship in institutional repositories increased. Both depend to a significant extent on the end user —as a source of demand in the first case, and as a supplier of content in the second. The kind of information services that are going to get that end user involved, and that are going make demand felt and encourage contributions, are likley to look more like mashups than like traditional library catalogue systems, or enterprise resource programs of any kind.

Increasing the motivation to deposit materials in institutional repositories may be especially important in the humanities: recent surveys indicate that "the number of humanities documents in institutional repositories is currently far lower than that in STM disciplines" (see http://eprints.rclis.org/archive/00005180/). That may be because scholars in those disciplines are still considerably less wired than their colleagues in science and engineering, on almost any campus I know. To return for a moment to Zboray, and to the differential effect of the railroad on publishing in more and less “railed” communities in antebellum America, the fact that there is still significantly higher information friction in humanities and, to some extent, social science departments than in science and engineering departments on the very same campuses is, I would argue, producing a digital divide within those campuses. The humanities are the equivalent of the underindustrialized South with its devalued currency and its genteel poverty, steadily losing ground to other parts of the campus, where the trade is in gigabits and petabytes. At the University of Illinois, Green St. is the Mason-Dixon line: north of it are the many mansions of the engineering campus, south of it are the hamlets of the humanities and social sciences.

Happily, though, the cost of deploying University 2.0 does not need to be great, and even the poor can participate. What we need more than big science or big servers are good ideas about interesting things that faculty, staff, and students could do with the information produced in, by, and about universities. WhoShouldIHaveLunchWith.app, perhaps, or SixDegreesOfMyAdvisor.app, or ShowEmergingFields.app, or WhatConferenceShouldIAttend.app, or WhatJournalsPublishOnMyTopic.app, and so on, and on. If the university makes its information accessible in the right way, people will build these things—sometimes people in the library, but also people elsewhere on campus, or simply elsewhere. There are issues of privacy, copyright, and all the usual sources of information friction, but even partial information could be useful, even metadata could serve many of these purposes. It would help, of course, if universities promoted some basic kinds of interoperability, not overfitted, but very simple. A standard calendar event expressed as an RSS feed; a recommended rdf tag or two for CV or syllabi. These would not have to be monolithic systems or all-eoncompassing standards, but they could be functional requirements for vendors who want to sell things like calendar systems to campuses, and they could be recommended to faculty and departments on the basis of the effort-to-utility ratio that can be demonstrated even with imperfect data.

I want to close by repeating something I said a couple of years ago, in a lecture on "Vernacular Computing" that I gave in London as the "Vodaphone Fellow" at Kings College, London. When I wrote this, I wasn't thinking of Web 2.0, which hadn't become a buzzword at that point, but now that it's here, this passage seems more true than when I wrote it. I said,

> Fifteen years ago, the challenge before us was to imagine how new technology might provide a new platform for the practice of scholarship in the humanities, but today our challenge is the reverse. It is no longer about opening the university and inviting the public in: it's about getting out where they already live, and meeting the public in the information commons, on the same terms that everyone else does. In fact, it's almost too late for us. We will find that hard to believe, ensconced (as we all are) in solid-seeming residential universities, with long histories and the expectation of a long future—but older institutions on more solid foundations have been

> swept away or radically transformed in cultural upheavals of the past. In spite of the inertia of these institutions, which we all know so well, the forces of change outside the institution have much greater inertia, and all of the practical furniture of our daily academic lives could easily be gone, or changed beyond recognition, in a generation.

So, while I recognize and give full weight to the inertia of universities, inertia doesn't necessarily mean that you remain in place, especially when the friction between you and the world around you suddenly decreases. We're at such a moment today, and it should be possible, with just a small push, to move the university forward into University 2.0.