April 2006

I Hear the Train A Comin' -- LOCKKSS and Portico

Greg Tananbaum

The Berkley Electronic Press, greg@bepress.com

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Tananbaum, Greg (2006) "I Hear the Train A Comin' -- LOCKKSS and Portico," *Against the Grain*: Vol. 18: Iss. 2, Article 42.

DOI: <https://doi.org/10.7771/2380-176X.4671>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

one that I think is obvious but inaccurate, and one that I think is less obvious but more useful.

Costing Staff Time as a Function of Salary Budget

Suppose that your serials department employs three full-time classified staff at a total cost (including benefits) of \$100,000 per year. Assuming a forty-hour work week, that amounts to 120 hours per week, 52 weeks per year. If you subtract four weeks annually for personal leave and holidays, you're left with 48 weeks per year at 120 hours per week, which means that for \$100,000 you're getting 5,760 hours of work at roughly \$17.50 per hour. (Their hourly wage will actually be lower than this number, because we're taking vacation time and benefits into account in measuring the cost of an hour of their time.) For simplicity's sake, let's round up to \$18.

The temptation would be to think that this figure — the direct salary cost — accounts for the value of your staff's time, and therefore that you should set workflow priorities based on that \$18 value: the message to your staff would be "don't spend an hour doing anything that isn't worth \$18."

But there's a problem here, and that is the fact that this valuation doesn't take into account opportunity cost. A staffer could spend an hour doing a task worth \$30, and you might think you're well ahead of the game — but if there was another task awaiting his attention that was worth \$50, then you have a problem.

This is why I suggest that instead of thinking of the value of our staff's time as a unit of salary, we look at it instead as a unit of resource available for the management of the materials budget.

Costing Staff Time as a Unit of Collections-Budget Management

By this measure, we'd calculate the value of staff time in this way: given a serials budget of \$1,300,000 and annual available work hours of 5,760, it takes an hour of staff time to manage about \$225 of the serials collection.

Now obviously, this is far from a perfect measure — if a staff member takes a sick day, you don't lose \$1,800 from your serials budget, or lose access to \$1,800 worth of information. But I think it's a better way of thinking about the *importance of applying staff time to tasks that really matter*. Instead of simply measuring staff time in terms of direct labor costs, it takes into account the actual value of what is being managed with staff time. To put it another way: staff time is worth much more

than the salary we invest in it.

Now, am I suggesting that each of us follow our staff around with a calculator and constantly analyze the value of each of their tasks, keeping a running tally to determine whether they should be claiming or calling vendors or processing orders at a particular moment in time? Of course not. The point of the very imperfect mathematical exercise in which I indulged above wasn't to arrive at a strictly accurate measurement of the value of staff time, but rather to demonstrate the huge difference between the value we attribute to an hour of staff time if we think mainly in terms of salary, and the value we attribute to it when we think mainly in terms of the value of the work that needs to be done.

When it comes to deciding how we and our staff should spend our time at work, it's becoming increasingly important that we look in a very hardheaded way at the value of our time and the value of our tasks. Our time is becoming increasingly expensive. Are we still spending time on processes that are of decreasing value, while failing to take up new tasks or processes that are of increasing value to our patrons? No two libraries are likely to answer that question in exactly the same way. But we should all be asking it of ourselves, and we should be asking it on a regular basis.

I Hear the Train A Comin' — LOCKSS and Portico

Column Editor: **Greg Tananbaum** (President, The Berkeley Electronic Press, 2809 Telegraph Avenue, Suite 202, Berkeley, CA 94705; Phone: 510-665-1200 x.117; Fax: 510-665-1201) <greg@bepress.com> www.bepress.com

When The Berkeley Electronic Press was preparing our first electronic journals in 2000, I talked to dozens of librarians around the world to better understand the market we were entering. I was amazed that the consistency of their questions transcended library size, location, and budget. One question we always heard was whether the journals would be peer reviewed (Yes). Another was would they have the conventions of traditional journals, such as ISSNs, volumes, and issues (Yes, Yes, and Yes). The most difficult question to answer was how could the library be assured of the long-term availability of our content? We were a startup, after all, and our content was exclusively electronic. My answer, something to the effect of "We have money in the bank and a solid business plan," was insufficient to quell these concerns. Even sympathetic librarians wondered what would happen if an earthquake pushed California into the Pacific. What would become of the content the libraries had paid for, the content to which we were guaranteeing perpetual access in our licenses?

Eventually, we scrambled to create archiving arrangements with OCLC and the California Digital Library. We were thus able to say that in the event of a catastrophic event

— financial or climatological — those who had paid for our journals would be able to pull copies to their local servers. These ad hoc arrangements, however revealed a weakness in the Internet-era publishing world. It was apparent that an independent resource, trusted by publishers and libraries alike, was necessary to help secure the long-term archiving of scholarly journals in their electronic form. Call it "assisted archiving."

Fast forward to 2006, and we find not one, but two such resources have emerged. Both **LOCKSS (Lots of Copies Keeps Stuff Safe)** and **Portico** provide a mechanism to preserve licensed digital materials. **LOCKSS** was first out of the gate, launching its production code in 2004. Operating at **Stanford** and funded in part with **Mellon** and **National Science Foundation** support, **LOCKSS** is "a community-based effort addressing libraries' need for affordable, robust, reliable means of preserving many different genres of digital content." More than 110 research libraries worldwide, along with 70 publishers of all shapes and sizes participate in the program.

From the libraries' standpoint, participation entails the installation of a "**LOCKSS Box**" on a dedicated PC. The **LOCKSS Box** can be configured in a few minutes with modest tech-

nical resources. Once installation of the open source software is complete, the PC is transformed into a digital preservation appliance that performs four primary functions. (1) It collects newly published content from targeted e-journals using a Web crawler similar to those used by search engines. Note that "targets" can only include titles to which the library otherwise has access (i.e., journals to which they subscribe). (2) It continually compares the content it has collected with the same content collected by other appliances, and repairs any differences. (3) It acts as a Web proxy or cache, providing readers in the library's community with access to the publisher's content or the preserved content as appropriate (with steps taken to ensure content is readable by new Internet browser versions). (4) Finally, it provides a Web-based administrative interface that allows the library staff to target new journals for preservation, monitor the state of the journals being preserved, and control access to the preserved journals.

With the exception of the cost of purchasing the dedicated PC and the human cost of running an update of the software twice a year, the libraries have no outlay associated with operating the **LOCKSS Box**. They may

continued on page 85

I Hear the Train A Comin' from page 84

choose to join the **LOCKSS Alliance**, a strategy group that discusses and helps shape **LOCKSS** policies and direction. **Alliance** membership currently ranges from \$1,080 to \$10,800 figures annually, tied to **Carnegie** classifications. **LOCKSS** expects these fees to drop as more members join the **Alliance**.

Note that before the **LOCKSS Box** can preserve a journal, the publisher must give permission for the **LOCKSS** system to collect and preserve the journal. Permission is granted through a Web page called a **LOCKSS** publisher manifest. A publisher manifest permits the **LOCKSS** software to collect, and preserve the content. The manifest allows the publisher to control, by content unit (e.g., volume), the timing of the content preservation. In other words, the ability for the library to preserve an electronic journal is tied to the timeliness with which the publisher updates the manifest. Once in the manifest, the material may be preserved by any **LOCKSS** library participant that subscribes to the publication.

The material preserved on the **LOCKSS Box** is meant to serve first and foremost as a dark archive. Content will be served to an institution's readers when a publisher's Website is not available under terms of original license agreement. To return to the Megaquake scenario in my introduction, imagine that a catastrophic event wipes out **The Berkeley Electronic Press**. Because our materials are cached in the **LOCKSS Box**, a subscribing library could simply redirect requests for our content to the local **LOCKSS** version until our own servers returned to operational status.

The **LOCKSS** model is predicated on a diffuse set of redundancies (presumably not simply because the acronym for My One Copy Keeps Stuff Safe isn't as cool as the one they chose). Whereas **LOCKSS** bestows upon its participating libraries the responsibility to maintain local backups, **Portico** has opted for a centralized solution. Formally launched in 2005 with support from **JSTOR**, the **Library of Congress**, the **Mellon Foundation**, and non-profit **Ithaka**, **Portico's** stated goal is to ensure enduring access to scholarly e-journals via a single, secure archive.

Content for the **Portico** archive comes directly from members of the scholarly publishing community who have agreed to contribute to the archiving service. In its first months of operation, **Portico** has secured the participation of close to a dozen publishers. Participating publishers deliver their content to **Portico**. The data are then normalized by **Portico** in a format based on the open standard **Journal Archiving and Interchange DTD** developed by the **National Library of Medicine**. This normalization facilitates both preservation and future migration. **Portico** has pledged to migrate publisher data to new archival formats that emerge over time, and to supply the migrated data back to the publishers upon request.

Once the content has been ingested and normalized, it sits in secure, redundant dark storage. It is periodically checked for decay, and select librarians at participating libraries are granted password-controlled access for verification purposes, but otherwise not accessed. **Portico** provides all libraries supporting the archive with campus-wide access to archived content only if specific trigger events occur. Examples include a publisher ceasing operations or discontinuing a title, a publisher no longer offering back issues, or a catastrophic and sustained failure of a publisher's delivery platform. In these instances, participating libraries will be granted IP access to the archived literature. Note that, like **LOCKSS**, the completeness of the **Portico** archive is dependent on the timely and accurate supply of data from the publishers.

The logistical aspects of a library's participation in **Portico** are minimal. Keeping an updated list of IP ranges on file at **Portico** is the primary activity. Annual participation fees are tied to a library's materials expenditures. At the low end, this translates to \$1,500 annually, and graduated along a 16 step scale to \$24,000 (although discounts apply to early adopters in 2006 and 2007). Publishers also pay an annual participation fee ranging from \$250 to \$75,000 depending on journal revenue.

The main difference between **LOCKSS** and **Portico** is the degree of centralization each brings to preservation. Other distinctions occur as well, of course. The focus of **Portico's** preservation effort is the intellectual content and not the look and feel of the journal. **LOCKSS** enables libraries to cache Web-delivered pages and images to capture both appearance and content. Whereas a trigger event within **LOCKSS** will unleash access only to the subset of **LOCKSS** library participants who subscribed to the affected journal, within the **Portico** model all library participants will have access to the affected journal, regardless of subscription history. (Note that **Portico** does offer a "perpetual access" feature via which a publisher may designate **Portico** as the mechanism they will use to provide ongoing access to library subscribers who have terminated a license.)

Portico and **LOCKSS** take different approaches to address a significant problem — the secure, long-term, and independent archiving of scholarly journals in their electronic form. Many publishers and libraries are content in the immediate term to experiment with both. This may be a sensible solution. If one believes in the value of assisted archiving, having a backup to the backup is a perfectly logical extension.

Back Talk from page 86

13. Contact all users of learning platforms like Blackboard, WebCT, etc., what they are doing with the digital files are not now actively using to teach their classes and suggest that they add them to the IR.

14. Ask the department(s) that process professional leave applications to send an IR brochure to all those who get travel support urging them to add their papers to the IR. Point out the value of triple dipping: give paper, put in IR, revise for publication.

15. Get all university administrators to encourage participation: To require that the intellectual output produced using University funds be placed in the repository. You need, however, to be careful about taking this latter step since many faculty members will automatically view with great suspicion anything a university's central administration mandates.

16. Look for departments that are publishing their own journals/periodicals and work with them to use the IR for that purpose.

17. Ask your university press to let you digitize their out-of-print books and add them to your IR.

18. Work with your university's archive to use the IR as a medium for exposing University related digital resources to the world.

Hopefully some of the above 18 are new to you or you have not yet done them. If you have other suggestions, please send them to me (ferguson@hkucc.hku.hk) and I will get **Katina** to list them for everyone's benefit. By the way, you are all invited to come to Hong Kong to help us celebrate the addition of our **millionth eBook**. For details, send an email to <mlyip@hkucc.hku.hk>.

In addition to surfing the Websites of repositories, I culled many of these ideas from the following sources:

Ashworth, S., Mackie, M., & Nixon, W.J. (2004). The DAEDALUS project, developing institutional repositories at **Glasgow University**: the story so far. *Library Review*, 53, 259-264.

Chan, L. and Costa, S. (2005). Participation in the global knowledge commons: Challenges and opportunities for research dissemination in developing countries. *New Library World*, 106, 141-163.

Genoni, P. (2004). Content in institutional repositories: A collection management issue. *Library Management*, 25, 300-306.

Gibbons, S. The Institutional Repository Issue. (July/August, 2004). *Library Technology Reports*, 40.

Horwood, L., Sullivan, S., Young, E., & Garner, J. (2004). OAI compliant institutional repositories and the role of library staff. *Library Management*, 25, 170-176.

Rogers, S.A. (2003). Developing an institutional Knowledge Bank at Ohio State University: From concept to action plan. *Portal: Libraries and the Academy*, 3, 125-136.