Modeling a Shared National Cross Digital Repository

Jean-Gabriel Bankier
bepress, jeangabrielbankier@yahoo.com

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

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

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Jean-Gabriel Bankier, "Modeling a Shared National Cross Digital Repository" (2013). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284315312>

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.

Modeling a Shared National Cross Digital Repository

Jean-Gabriel Bankier, President and CEO, bepress

Abstract

The United States Office of Science and Technology Policy (OSTP) delivered an open access directive in early 2013 mandating that federally funded research articles and associated data must be made accessible online to the public, free of charge. In response to this mandate, university and library organizations proposed the Shared Access Research Ecosystem (SHARE). With experience in the design and creation of large-scale federated repositories, bepress, best known for our hosted institutional repository platform Digital Commons, can offer unique insight into what makes a federated repository successful. This paper will outline the attributes of a successful shared national cross digital repository.

At bepress, we have thought a lot about large shared digital repositories and have come up with a model for what we believe a U.S. national repository should include. Before we dive into the story, however, it is worth sharing some background about bepress to provide context for this paper. First, bepress was founded in 1999 by University of California Berkeley faculty members who were frustrated with the state of scholarly communications and wanted to improve it for their peers. Being founded by faculty, with one of the founders currently serving as our board chair, means that we approach everything we create with a faculty-centric focus. For example, Digital Commons was built to be a publishing platform as well as an institutional repository (IR). This allows libraries to provide a variety of services to faculty, including workflows for publishing books, journals, and conference proceedings, among others.

Second, bepress is committed to trying new things. We strive to push the frontier of scholarly communications. We were, for example, the first to try the software-as-a-service model for institutional repository software. At the time, around the year 2002, the general thinking in the scholarly communications community was that repository software had to be built and managed locally. Now, with the advantages of the software-as-a-service model more widely understood, all of the major IR platforms like DSPACE, Fedora, and Eprints have software solution providers offering IR services in the cloud.

Third, we have been working with shared, aggregated repositories for many years. We built our first integrated, federated repository over ten years ago for the discipline of law. In November of 2012, we launched our latest iteration of a federated repository called the Digital Commons Network. The Digital Commons Network includes the work of more than 300 institutions and over 800,000 full-text objects, making it one of the largest databases of open access research in the world. While we had the ability to launch the Digital Commons Network earlier, it was crucially important to us to wait until the community of universities and colleges that use our platform had at least 500,000 articles in their repositories combined before deciding to aggregate our content.

When we launched the Network in November of 2012, there was little discussion of anything similar on a national scale. That changed in February of 2013 after the United States Office of Science and Technology Policy (OSTP) issued an open access directive mandating that federally funded research articles and associated data must be made accessible online to the public, free of charge. According to one estimate by Heather Joseph (2013), Executive Director of the Scholarly Publishing and Academic Resources Coalition (SPARC), this could include as many as 200,000 articles a year across the government agencies. With each agency independently evaluating options and partners for compliance, the OSTP announcement kicked off a series of discussions among a wide ranging group of stakeholders,

including groups of publishers and university organizations.

Out of these discussions at the university level emerged the Shared Access Research Ecosystem (SHARE). Proposed jointly by the Association of Research Libraries (ARL), the Association of American Universities (AAU), and the Association of Public and Land-grant Universities (APLU), SHARE represents a unique partnership between a national library organization and two national higher education organizations. The proposal put forth by the SHARE steering committee advocates the development of a system of cross-institutional repositories in response to the OSTP mandate. The community believes that universities could serve as the best partners and offer the best solution to government agencies' needs to comply with the open access directive. The thinking works as follows: universities produce the majority of research the mandate seeks to make public and have already invested heavily in hosting that content online in individual institutional repositories. The next step of organizing the content and providing a federated view of the content by government agency is a natural one. As John C. Vaughn, Executive Vice President of AAU, put it, "If we're going to be building these repositories anyway and want to interconnect them for our own purposes, we have got the framework of a system that could manage the content and provide the access that the OSTP directive is calling for" (Vaughn, as quoted in Schwartz, 2013).

Despite their great potential impact on the mission of the university, institutional repository programs rarely get the attention from senior administrators on campus that they deserve. Scholarly communications remain primarily a library passion. However, because two of the three associations involved in SHARE are university organizations rather than library organizations, the idea of a national portal caught the imagination of many presidents and provosts around the country, bringing the discussion of institutional repositories center stage. Eager to participate in a highly visible and important national initiative, senior administrators made

inquiries about their campus repository's current capabilities and future plans.

At the 2013 ARL Fall Forum in Arlington, Virginia, this October, excitement for a shared U.S. national repository idea was palpable among the attendees. During the session "Examining Our Collective Capacity," speakers Sayeed Choudhury, Associate Dean for Research Data Management and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries at Johns Hopkins University; James Hilton, Dean of Libraries and University Librarian at the University of Michigan; and Bradley Wheeler, Vice President for Information Technology and Chief Information Officer at Indiana University, made the case in support of the idea.

According to Wheeler, universities should be involved in a national portal for the sake of maintaining control. As he sees it, content needs to be stewarded by academic institutions to ensure it remains accessible, and only libraries have a mission to ensure persistent access. As we see it, retaining ownership also means that institutions would get credit for their contributions; something that may prove particularly important as university funding is increasingly scrutinized and institutions are being challenged to prove their monetary worth. Furthermore, when institutions outside the library and academic communities take control, that control comes with the authority to do what they will. They may remove hyperlinks in the articles, restrict access by requiring readers to log in, post advertisements, or charge for access or distribution. In his talk, Hilton focused on the need for speed and the importance of scale. He stressed that the library community must build such a federated repository before the content is owned by someone else and access is restricted, as a result. To illustrate Hilton's point, there are numerous sites currently clamoring for faculty's intellectual output. Academic social networking sites such as academia.edu are growing by providing faculty incentives to self-archive and engage with the site. In order to become the destination site for researchers, universities and libraries must form a critical mass of content. The take-away from the session was clear to everyone

National Repository Name	Number of articles
Recolecta (Spain)	1,570,000*
Pionier Network (Poland)	1,500,000
JAIR0 (Japan)	1,210,000
Digital Commons Network	800,000
RCAAP (Portugal)	500,000
NARCIS (The Netherlands)	340,000
HAL (France)	240,000
DIVA (Sweden)	140,000
NORA (Norway)	40,000
RIAN (Ireland)	30,000
*includes metadata only articles	

Table 1. National Repositories with Total Number of Articles in Each as of September 2013

in the room: the scholarly communications community is in a race to become the go-to site, and if the universities and libraries do not do it, and do it big, someone else will.

The U.S. is not a trail blazer when it comes to thinking about a national open access repository of scholarship. Many other countries already have national repositories, most of them in Europe, which we can look to for lessons as we build a U.S. model.

Norwegian Open Research Archives (NORA) in Norway and Repositório Científico de Acesso Aberto de Portugal (RCAAP) in Portugal, for example, provide links to member repositories from participating institutions and a search box for querying content. Although this is an easy and common approach in national repository design, it is insufficient. Even when there are more advanced search options built into a site, repositories like this miss key opportunities for creating value and better discoverability of content. Perhaps these national federated repositories seem to fall short because their primary goal is to make the corpus of national research simply *available* to the public.

Elements of a Successful National Repository

How do you begin thinking about a national repository that moves beyond simple access and offers something valuable to all types of users? We frame the discussion around three terms: interesting, discoverable, and engaging. Or, in

other words, a successful national repository must be interesting so scholars like it, discoverable so scholars can find it, and engaging so scholars return to it.

Make It Interesting

To make a repository interesting you must organize the materials in a compelling manner, add an informative interface, and, if possible, make it interactive.

HAL (Hyper Articles en Ligne), a repository in France, offers a useful example of how to think about making a national repository “interesting.” HAL’s design allows users to browse by subject, narrowing in on their specific area of interest. Organization by subject matter is an important attribute for exploration, given that authors tend to value discipline-based organization over institutional organization. The value of organization of the research into subject-based collections is supported by the findings that subject repositories are better able to meet researchers’ needs, whereas librarians and institutions, not surprisingly, feel that institutional repositories are preferable (Nicholas, Rowlands, Watkinson, Brown, & Jamali, 2012).

While organizing content by subject is a good approach, HAL’s user design comes up short. When browsing in HAL, users are presented with a series subject folder containing lists of hundreds of articles. This organization system is not only aesthetically lacking, it also does not effectively encourage user engagement. Like the designers of

Figure 1. The Digital Commons Network Discipline Wheel

Figure 2. A Commons Page in the Digital Commons Network

HAL, we believe in discipline as the underlying organizational principle for content, but a successful repository has to make subject browsing more interesting for researchers. Browsing exposes researchers to related content in their field, increasing opportunities for discovery of important material. The Digital Commons Network leverages this discoverability by using a three-tiered taxonomy—ten disciplines in the top tier and more than 1,000 subdisciplines in the other two—to organize works, helping researchers locate and engage with their specific area of research quickly and easily.

Unlike HAL, bepress introduced a graphical browse element to replace the folder structure. The

starburst graph—or, as we call it, the discipline wheel—employs a simple organizational principle. The ten major disciplines are represented by different colors on the top tier of the wheel. When you hover over one with the cursor, you see not only the name of the discipline and the number of sub-disciplines within it, but also the number of works in the discipline, the total number of downloads of those works, and the number of institutions represented in the collection. The wheel changes size and shape as you navigate, helping you explore more effectively and making the experience fun and interactive. In an article published by the *Library Journal*, “Uncommonly Open: The New Digital Commons Network,” author Matt Enis states, “...from a design perspective, this

Figure 3. A Visual Breakdown of Institutional Contributions within the Mechanical Engineering Discipline

colorful wheel plays an important role in communicating the vision and purpose of DCN and the institutional repositories served by Digital Commons” (Enis, 2013). The deeper you dive into the wheel, the more detail you are given.

You can jump off the wheel at any time by selecting a discipline or subdiscipline, at which point you’ll be brought to a new page designed specifically for the discipline, which we call a “commons page.”

Each commons page includes a short and manageable list of subdisciplines, when applicable, and the most popular articles, institutions, and authors within that discipline. Clicking on any of the article links on the page brings the user to a full-text download of the work; however, it is easy to navigate back and forth between different commons pages and the discipline wheel at any time.

One of the more unique features of the Digital Commons Network is its commitment to clearly representing the contributing institutions in any given discipline. Most national repositories either do not do this well, or they do not do it at all. With the Digital Commons Network, each article listed on the commons pages is displayed with the name and logo of the originating institution. This information is interesting to users, as it provides additional context about the research. A reader might be curious about the work of a specific institution or know a certain institution to be excellent in a particular field, piquing their interest in works they might not otherwise have discovered. While this is not the primary objective, institutional

branding is important to libraries and institutions. We believe academic institutions deserve credit and recognition for their contribution and scholarship, and the Digital Commons Network is specifically designed with that in mind.

Pushing that concept even further, researchers can also view the breakdown of content in a discipline by institution. This is displayed visually in a pie chart, with each institution represented by their relative contribution to the discipline. Each institution’s section on the chart is also linked to their respective IR, providing a direct path to read the scholarship with one click. Breaking down the content in this manner allows users to see how each individual institution’s contribution to a discipline compares to other institutions’ contributions within that specific field. We believe that this presentation is a more interesting interface for readers to interact with, leading to greater discoverability of content. Additionally, this display also increases exposure of institutions and their schools by showing where their work is having the largest impact in expertise, and once again granting them recognition for their contribution to a larger field. This visual approach also shows the impact an institution is having on the larger corpus of research by specific discipline in a way that simply listing participating institutions cannot.

Make It Discoverable

We put a lot of effort into making the Digital Commons Network interesting. But what good is an interesting repository if it is not also a discoverable one? Driving people to a shared national repository

Figure 4. Google Search Results for the Terms “Cultural Anthropology Articles”

by making it highly discoverable is an essential part of our approach. There are two main ways to increase visibility online: through strong Search Engine Optimization (SEO) tactics and by ubiquitous interconnection—linking and cross-search integration.

We know that much of the traffic to IRs comes from search results and that many of those results are generated by very specific, narrow searches in Google that match specific language in a research paper. But how much more traffic could we create if IRs also ranked highly in the results for all of those less specific, more generalized Google searches? General search terms like “mechanical engineering,” or “cultural anthropology” rarely direct traffic to a school’s IR for several reasons. First, because individually the work from one institution does not have a large enough presence and carry enough weight online to compete with everything else on the Internet related to a search term as broad as “mechanical engineering.” The more broad the search term, the more competitive the space. Second, because work from institutions does not generally include such broad terms in the metadata, the abstract or the heading of the PDF, they are not typically picked up as keywords when Google scans the Internet and, therefore, do not rank high in the search results. Or, to put it more bluntly, when it comes to generalized Google searches within a discipline, individual papers and series within institutions

repositories simply do not have enough “Google juice” to rank highly in search results. But what would happen if the Google juice from all of those individual institutions was combined into one?

A benefit to organizing the greater body of work into subject-based commons like the Digital Commons Network is that it aggregates the research and supplies enough Google juice to rank highly in search results. This means that when people who typically search with more general terms—such as students, graduate students, government workers, and the general public—use Google, they are far more likely to find the repository high in the search results. And, because all of the articles in the Digital Commons Network link to the participating institutional repository, this means greater readership and greater impact. At its current size, the Digital Commons Network typically has enough Google juice to rank within the top five search results for some general searches such as “cultural anthropology articles.” A national repository, built in the model of the Digital Commons Network, we believe would likely have enough Google juice to return as the top result for nearly any general search within a discipline.

Another effective way to increase discoverability is through cross-linking and the creation of ubiquitous connections between participating repositories and the Digital Commons Network. By highlighting the Digital Commons Network on

Figure 5. An Institutionally Branded PDF Cover Page with Links to the IR and DCN

individual article pages with links, Digital Commons member repositories enable readers to discover interesting paths to browsing related content. Highlighting the relationship between the individual repository and the other repositories in the network with links also leads authors to become more engaged, by showing them that their work, their library, and their institution are all part of something larger. Contributing to their discipline is important to authors, and this kind of cross-linking breaks down the institutional silos and helps authors make a meaningful connection between their repository and their larger scholarly community. Currently, no other national repositories utilize this approach. Instead, users of other national repositories are taken directly to member repositories with no way to return to the national repository except for the browser's back button, setting up IRs as isolated nodes and

negating the feeling that the research is part of something bigger.

Another way to capitalize on the power of cross-linking is by adding links to the related discipline of a paper directly in the automatically generated PDF cover page. Seventy percent or more of all search traffic to repositories goes directly to PDFs. Highlighting links back to the institutional repository, and the federated repository takes advantage of that traffic and creates a connected path among repositories for the user to follow. PDFs are no longer dead-ends, but another stepping stone in the researcher's path to discovery. Finally, it is worth noting that cross-linking also significantly impacts SEO. There are currently over 1.6 million links from repositories to the Digital Commons Network, boosting that all-important Google juice even higher.

In addition to ubiquitous cross-linking throughout the repositories, we wanted the search boxes in Digital Commons to maintain a constant presence. As a result, there are close to a million search boxes across all of our member repositories, and for each one, a researcher has the option to search within that particular series, that particular repository, or across all member repositories. Having one million search boxes that can lead a researcher to an institution's repository in the whole network is a huge advantage.

Of course, this approach to increasing discovery introduces many new paths for readers to navigate away from a repository, which raises questions about traffic flow. Are individual repositories losing more traffic than they are gaining from this cross-linking and searching? No is the answer. On average, our studies have shown that twice as much traffic is directed *to* Digital Commons member repositories *from* the Network, rather than the other way around. Additionally, we have found that researchers who come to the network from one of the participating repositories browse twice the number of pages and stay twice as long as those who come from a search engine. It would seem that not only does cross-linking increase discoverability, it also appears to play a role in making the Digital Commons Network more interesting.

	Google	IRs in the Network
Pages per visit	1.74	3.10
Visit duration	1:16	2:50

Table 2. Number of Pages Browsed and Time Spent Browsing by Users Coming from a Search Engine Result Versus Those Coming from a Link in a Digital Commons Repository

Make It Engaging

The final piece in the federated repository trifecta is engagement. A researcher may have discovered your site and found it interesting, but the real question is whether or not they will return. That is where engagement comes in. Eighty percent of the visitors to the Digital Commons Network are new, and 20% are returning. For a young site like the Digital Commons Network, that is a pretty good ratio. Over time, however, we would like to see the percentage of returning visitors climb. The key to

creating a successful national repository is designing something that is engaging to authors so they will contribute more research and engaging to readers so they will keep coming back. We shaped our engagement strategy for the Digital Commons Network around three main themes: faculty promotion and recognition, competition, and community.

To address the issue of faculty promotion and recognition, we have created a service for scholars called SelectedWorks that is designed to increase the profiles of individual researchers and their institutions by allowing them to showcase their scholarly work and update interested parties about new content via an integrated mailing list. Behind the scenes, our Author Dashboard tool displays traffic overviews and detailed information about download activity, reader demographics, and referral sources. Monthly readership reports are automatically generated and e-mailed to authors, which provides them with a quick overview of relevant statistics and serves as a reminder to continue interacting with their SelectedWorks page. Engaging authors in this way allows them to clearly see the impact of their work and is highly effective at retaining their interest in the site, thus encouraging further submissions. The response from authors to this approach has been overwhelmingly positive. Cyndi Nienhaus, Assistant Professor of Religious Education at Marian University, had this to say about her readership report for an article she wrote and published in a journal within her institution's Digital Commons repository: "This is the most wonderful feature a journal can offer! To see how many times my article has been downloaded and where it was downloaded is fascinating."

When it comes to creating engagement through community, there is a successful road map already in place. By taking the best of what other social networking sites offer and adapting it to the Digital Commons Network, we are helping researchers to better connect with the site, the content, and each other. Our "Follow" feature allows readers to request periodic updates about new content from a specific author, series, publication, discipline, institution, and more. These Follow buttons appear throughout the Digital Commons Network and

member repositories and make it easy for readers to participate in the community, discover new content, and stay engaged. Authors are alerted whenever someone follows their work which helps foster a sense of community. The larger this community grows, the greater the impact these kinds of features will have.

Finally, the best way to illustrate how competition creates engagement is through a story. With a little cunning, Harrison W. Inefuku, Digital Repository Coordinator at Iowa State University, drove a significant increase in faculty participation to his IR. After looking at the pie chart displaying the institutional breakdown for the Agricultural Engineering commons and seeing that Iowa State's rival, University of Nebraska-Lincoln (UNL), was showing more strongly, Harrison decided to take action and use the spirit of competition to his advantage. He showed the chairman of the Department of Agricultural and Biosystems Engineering and others within the department how UNL was taking a bigger piece of the pie. From this, faculty were persuaded to respond positively by submitting more of their articles to the repository in an attempt to surpass UNL's representation in the discipline network; a goal they ultimately achieved. This success led Harrison to utilize the same tactic with other departments, and he is now happily overwhelmed by the response. The Digital Commons Network supports over 300 institutions. If a national repository were to encourage this same kind of competition among peers and see the

same kind of reaction, it would do wonders for accelerating the pace of self-archiving.

Conclusion

A national repository, as we view it from our experience with the Digital Commons Network, can magnify the impact of the university and college's investment by consolidating a scattered repository landscape into a more useful and efficient resource for showcasing the nation's scholarly content. By integrating these otherwise disparate caches of research and optimizing the interactions with all stakeholders, we have seen how a cross-institutional repository can greatly increase the discoverability of content, exposure for institutions and their libraries, and engagement with authors and readers. Universities are being pressed to demonstrate their worth to funders, the government, and the general public. Showing off the research a university produces is a critical part of proving the value of institutions and their libraries. We can all do this together without losing control of the content. We have demonstrated that it is both possible and rewarding through the ideas explored in this paper, but we need a critical mass to join us in order to achieve large-scale success. A U.S. national repository that is interesting, discoverable, and engaging will help support the survival of institutions across the world. At a time when universities and libraries are under fire, to not leverage an investment like this would truly be a missed opportunity.

References

- Enis, M. (2013, June 19). *Uncommonly open: The new Digital Commons Network*. Retrieved from <http://www.thedigitalshift.com/2013/06/discovery/uncommonly-open/>
- Joseph, H. (Moderator). (2013, October 10). Examining our collective capacity [Panel Discussion]. *ARL Fall Forum 2013*.
- Nicholas, D., Rowlands, I., Watkinson, A., Brown, D., & Jamali, H. R. (2012). Digital repositories ten years on: What do scientific researchers think of them and how do they use them? *Learned Publishing*, 25(3), 195–206. <http://dx.doi.org/10.1087/20120306>
- Schwartz, M. (2013, June 12). ARL launches library-led solution to federal open access requirements. *Academic Newswire*. Retrieved from http://lj.libraryjournal.com/2013/06/oa/arl-launches-library-led-solution-to-federal-open-access-requirements/#_