Coalition for Networked Information

Responding to the Move of Content to the Cloud: GitHub, Figshare, YouTube, and Similar Services

Report of a CNI Executive Roundtable Held December 10 & 11, 2017

Published February 2018

Background and Synthesis

At the Fall CNI meeting in Washington, DC, we held two rounds of an executive roundtable on **Responding to the Move of Content to the Cloud: GitHub, Figshare, YouTube, and Similar Services.** We discussed a wide range of topics related to how campuses are developing institution-wide strategies for managing, curating, storing, preserving, and disseminating data and/or surrogates or born-digital representations of objects collected or created by members of their institutions in the context of these cloud-based services. We discussed how institutions are dealing with faculty and others in their community making their own, independent decisions to use various cloud-based services. We also explored some of the policy and risk management issues that are emerging from these developments. It is very clear that these are early days, and developments on the ground have outrun efforts to track, much less genuinely manage or respond to these developments at a policy level.

While the explosion of cloud-based computational and storage environments and the shift to cloud-based platforms for many application software and service providers has been very extensively discussed both in the higher education context and more broadly, there is a related series of developments that have been much less carefully explored, particularly from a strategic institutional perspective. Attention has been focused on high performance computing, formerly local commodity applications (e.g., email, Office 365, Google apps), research data management, and even developments in next generation digital learning environments (future learning management systems). We want to carefully distinguish between bulk bit storage (Amazon S3 or Glacier, for example) and much more content-specific services like YouTube or GitHub; in these discussions we are primarily focused on the latter.

We have seen the steady adoption of various services that now house a tremendous amount of content developed by the research and higher education communities: source code (GitHub, SourceForge); video (YouTube, Vimeo); images (Flickr, Instagram); research data (figshare, Center for Open Science, various disciplinary repositories); and undoubtedly other classes of material.

Some of these enterprises are commercial; others are run by not-for-profit groups. Some are struggling with long-term financial viability. In most cases the higher education institutions have no role in the governance or policies of these external enterprises, and

often they don't even have a negotiated contractual relationship (as opposed to a consumer targeted click-through license) because material have been placed on them by faculty, students, or even individual departments without the institution's formal involvement.

CNI executive director Cliff Lynch opened each roundtable by noting that it seems like we have seen a gradual, incremental uptake of all kinds of off campus services that store and act as distribution points for a great variety of types of material. Interestingly these activities are often happening on a grassroots basis by individual faculty, students, or groups who frequently decide to use a cloud service because it is easy and/or free, and they click through the license without much thought. These services are often used to store material that the institution has some responsibility for, such as research results. Common practice leads institutional staff members, not just faculty or graduate students, to use these services in the absence of any institutional policy, contract or review. This is not an entirely new story, of course, but it is extending to new areas. As institutions recognize the increasing use of cloud services by members of the community (for example, uptake of Dropbox for collaboration) the chief information officer (CIO) will read the license and realize that the use of the service violates institutional policy, or, in some cases, law or regulations related to contracts and grants. In the case of Dropbox, some institutions developed arrangements with Box that resulted in a more suitable contract for institutions, often through Internet2's NET+ program. However, CIO's found they were still facing an uphill battle to shift people from Dropbox to Box, and there has been mixed success in trying to drive its widespread adoption.

Lynch then gave a quick list of some of the content-specific platforms (not including commodity storage services) and reported on his tracking down of who actually owned them (figure 1).

Content/Service Type	**Service**	**Owner**
Images	flickr	Verizon, by way of their acquisition of Yahoo
Images	Pinterest	Private investors – venture capitalist
Images	Instagram	Facebook
Video	Vimeo	IAC Corp.
Code	SourceForge	BizX
Code	GitHub	Private, venture capital start-up
Code	figshare	Digital Science – Nature/Springer – Holtzbrinck
Data	Mendeley	Elsevier
Data	Zenodo	CERN
Full research lifecycle	Open Science Framework	Center for Open Science
Websites	Wordpress	Wordpress

Figure 1: Various online service platforms and the entities that own them.

This list gives a feel for the diversity of products and applications and the fact that many are commercially based. Some, like figshare, will let you host a local copy for a fee, but not all.

Lynch concluded by noting that those platforms owned by corporate interests are not necessarily focused on the best interests of the research and education community. Indeed, in many cases, the research and education community are not the primary users of the service, but only represent a tiny fraction with perhaps specialized interests and concerns that may be quite different from those of the general consumer. In most cases the higher education institutions have no role in the governance or policy formulation of these external enterprises, and often they don't even have a negotiated contractual relationship (as opposed to a distributed click-through license).

The materials placed into these services in growing quantities are critical to higher education institutions in various ways, and yet our institutions may not even know what resides on any given external service. Furthermore, in a significant number of cases, the university or college may actually have *obligations* (for example to funders) to preserve materials; they may not simply be operationally critical in the near-term. In addition, services come and go, or preferred services may change within various communities over time. But we often have no guarantee that the materials we have placed on these services will not vanish on very short notice, and we have no easy ways to migrate material, and no backup strategies for the services. A fundamental error is the confusion of access with longer-term stewardship of material, where absent very detailed contracts and risk analyses, these services cannot be readily trusted.

Only a small number of institutions attending the roundtables were making efforts to collect data on the use of cloud services by members of their community, usually through monitoring network routers. Only one institution mentioned interviewing researchers about their practices. This lack of strategies for even systematically monitoring and tracking what the campus community was doing was especially worrisome.

When a particular service is heavily used by members of an institution, a vendor may initiate contact with central IT to seek to negotiate a campus solution. Many noted that what they referred to as "policing" conversations with faculty – warning them of consequences of using unauthorized cloud services – were not helpful. One described a situation where the institution had set a policy for its researcher to not use Dropbox and in response, researchers began to open personal accounts in order to circumvent the policy.

Many institutions present were concerned about back-up strategies for content in the cloud, especially when there is no institutional relationship with the service provider. Even when there is an institutional contract, some expressed a lack of trust and confidence in the commercial services. For content where they do have control, some libraries are using the non-commercial Digital Preservation Network (DPN) service to ensure long-term preservation.

In particular, for those striving for curation and preservation of data, fixity in large storage aggregations is important and institutions need the ability to test this in a trust-but-verify spirit. Vendors and service suppliers are totally ignoring this need.

Individuals representing universities, liberal arts colleges, national libraries and archives, and service providers from the US and Canada participated in the roundtables. Some of the trends, issues, and concerns which surfaced during the conversations are noted below.

Institutional Perspectives

- We did not identify any institution that had a clear governance regime for use of these sorts of cloud services. One campus stated that it was on its third attempt to develop a campus governance cloud committee to develop best practices and create an inventory of services being used.

- Some institutions have policies to move as much as possible to the cloud, though they haven't necessarily considered the full scope of the potential application of these policies; others have hybrid local/central models that include use of cloud services, and others do not have a campus policy.

- Both in the US and Canada, some institutions were sensitive to or legally bound to keep at least some types of data stored within their country. Geopolitical diversification is a genuine issue for many reasons. Other constraints involve US regulations such as HIPPA (the Health Insurance Portability and Accountability Act) or ITAR (the International Traffic in Arms Regulations). An interesting case in point: Box, greatly favored by most CIO's over Dropbox, does not have servers in Canada, so there are major issues with transborder flow of data, particularly student data, for Canadian institutions.

- Contracts with cloud storage services may include strong security provisions, particularly specifying what data the vendor is sharing with outside parties. Some campuses go through a risk management process when they are negotiating significant licenses.

- Institutions with state, regional, or national responsibilities for curation and preservation of some types of digital content may be particularly vigilant about having back-up systems, multiple storage locations, and disaster recovery plans in place.

- When institutions finally do recognize a need for an institutional arrangement and put it in place, early faculty adopters who have set up private accounts on a service (often using things like their personal Gmail accounts rather than institutional email) are very hard to identify so that they can be shifted into the institutional arrangement; often this can only be done if a faculty member takes explicit action to merge his or her account into the institutional offering, and it is frequently quite inconvenient.

- Institutions expressed concern about having exit strategies in place for moving from one platform to another. They recognize that they must formulate a plan, but they also realize that they don't necessarily do a very good job of it.

- In some cases, institutionally licensed cloud services may be available only to institutional IT developers (for example, a GitHub license) and this may leave faculty, students and others with no option but to find resources on their own.

- Some campuses noted that they are developing strategies in silos for various types of content (e.g., data, digital humanities projects, and video) and that these strategies are uncoordinated.

- Containerization and related strategies for software preservation, reproducibility and related reasons are creating an entirely new genre of materials that need to be both shared and archived. There are also second-order phenomena here such as libraries of 3D objects that might be reproduced in accretive (3D) printing facilities.

- While many of the participants were focused on services that catered to research data, participants also noted the growing use of cloud services for video storage. The need for privacy and intellectual property control surfaced in that context. One participant stated that he believed that vendors were actually more secure than higher education institutions. Besides the obvious players like YouTube and Vimeo, Kanopy was also mentioned frequently; it generally gained an initial foothold on campuses as a supplier of video material particularly useful for higher education, but also offers various storage packages. The Avalon system, developed at Indiana University and Northwestern University, but now seeing considerable take-up elsewhere, is also attracting interest as both a web-based and institutional tool for managing and providing access to audio and video materials.

- The question of GitHub in the cloud as opposed to institutional GitHub instances is emerging with growing frequency.

- Another service which was frequently mentioned by participants in the roundtables was Slack.

- Another type of content noted by a several institutions was the use of electronic lab notebooks as web based services. CNI believes there is going to be a great deal of growth is this area. Electronic lab notebooks have long been used widely in industry, but the costs of commercial software have historically made them unrealistic for most university settings. We are now seeing much lower cost solutions that are specifically targeted at university needs.

Concluding Thoughts

Overall, this is a fast-moving arena in which grassroots developments are quickly outpacing institutional policy and strategy. As commercial solutions are increasingly adopted by faculty and students, and favored products change in short timeframes, institutional responsibility for content becomes difficult if not impossible to manage in some cases. As the number of platforms proliferate, including those used for social media, it will become increasingly difficult to develop a coherent strategy for curation of content from these systems. One example offered was that research teams used to

communicate mainly through email, which some universities had procedures for archiving, but now much of that interaction has moved to channels like Slack, where the institution will have no trail of evidence if something goes wrong (and, indeed, little clue about the shift in the first place). One representative suggested that libraries need to think more about medium term, rather than long-term preservation, and have to become more agile.

One participant noted that the library's mission is to provide infrastructure for researchers to properly develop and document their work at every part of the process of creation of the evolving scholarly record. However, when the library creates infrastructure that is sounder but less convenient than what is available on the commercial market, convenience will trump quality every time. He concluded that if we want to impose any structure, libraries must make it as easy as possible.

There are numerous complex trends that are influencing developments. Commercial players such as Digital Science and Elsevier are attempting to create what might be characterized as "verticals" – that is, they don't necessarily interoperate gracefully with offerings from other vendors, but link together very well at various niches that the vendor addresses. Another confusing recognition is that a great deal of research data management is about small objects, like spreadsheets, rather than multi-terabyte objects, which is causing a lot of recalibration with respect to research data management and sharing systems. These smaller objects are amenable to very different and much lighter-weight solutions than terabyte or petabyte datasets.

We may see a move back to campus storage as costs for storage in the cloud can become higher than some anticipated, particularly if the institution insists on regular, high-level, fixity check audits rather than simply relying on commercially redundant, geographically distributed storage. An increasing number of institutions are examining where they put what type of content and at what cost. However, migration is also a thorny and complex process sensitive to storage, access, and network traffic charges. There are some particular sore points: many institutions emphasized the difficulties involved in contracting for services and accounts on Amazon Web Services (AWS) at the institutional or library level. A detailed examination of the situation here might be very helpful.

As time passes, new issues are arising, such as faculty members retiring but wishing for the institution to maintain stewardship of his or her materials on GitHub, Flicker, or wherever, or earlier-career faculty changing institutions and having to determine on whose site or under whose license their data will persist. There's very little policy to help institutions with this, and it's clear that a broad inter-institutional adoption of a best practice would be a great help to the community. Also, it's very clear that there is an ongoing tension between institutionally migrant faculty (assistant professors, adjuncts, postdocs) or graduate students who want individual control over materials rather than institutional arrangements and the interests of the institutions themselves. Ultimately, we suspect that there is going to need to be some sort of broad agreement about best practices for scholars moving from one institution to another with regard to the stewardship of their scholarly work.

Some areas are virtually unexplored, such as discovery of materials across this wide range of platforms, and integration with emerging campus discovery environments, with the exception of some modest work on research datasets.

We are a little concerned that there is a continuing thread of focus by libraries on more traditional institutional offerings like Omeka and institutional repositories. While it is theoretically possible for them to offer environments to manage a good deal of the content that is being housed on various external platforms, they are much less convenient and functional for what faculty and students want, and at best might be places for institutions to import and manage stewardship *copies* of material that the institutional community placed on these external platforms.

It is clear that these are relatively early days in addressing these developments. Perhaps the only really concrete best practice – which we believe deserves very wide adoption – is monitoring traffic patterns on campus border routers in order to get a sense of what external services are being frequently used, and some supplementing of this practice by asking faculty what they are actually doing day to day. It's also clear that institutional risk assessment staff has not yet recognized the thicket of issues here; as this slowly occurs, it will add additional dimensions to the challenge.

CNI Executive Roundtables, held at CNI's semi-annual membership meetings, bring together a group of campus partners, usually senior library and information technology leaders, to discuss a key digital information topic and its strategic implications. The roundtables build on the theme of collaboration that is at the foundation of the Coalition; they serve as a forum for frank, unattributed intra and inter-institutional dialogue on digital information issues and their organizational and strategic implications. In addition, CNI uses roundtable discussions to inform our ongoing program planning process.

The Coalition for Networked Information (CNI) is a joint program of the Association of Research Libraries (ARL) and EDUCAUSE that promotes the use of information technology to advance scholarship and education. Some 240 institutions representing higher education, publishing, information technology, scholarly and professional organizations, foundations, and libraries and library organizations, make up CNI's members. Learn more at cni.org.