Library Council Meeting

Thursday, February 9, 2017

1. Introductions and Announcements

Attendees: Deborah Jakubs, Lisa Huettel, Naomi Nelson, Jeff Kosokoff, Tim McGeary, Tom Hadzor, Bob Byrd, Ann Elsner, Dominika Baran, Nivedhitha Subramanian, Manoj Mohanan, Victoria Szabo, Phil Stern, Jocelyn Olcott, Tom Witelski, Jimmy Roberts, William Johnson, Matthew Keller, Emily Daly, Charles Piot

Great turnout today and great agenda.

Our Library budget meeting is next Tuesday, Valentine’s day. Today’s agenda is short but very pithy. The collections budget is something that affects faculty and students. It goes a separate path from the Libraries budget. Bob and Jeff put together a detailed document about expected increases in cost and inflation, journal increases by publishers and a lot of other data.

1. The Dynamics of the Duke University Libraries’ Collections Budget - Jeff Kosokoff and Bob Byrd

The collections budget request was submitted last November for FY ‘18. We’d like to give you some information about how we develop that request, the data we use, and to talk about results. We welcome questions and comments at any point.

To begin with, we wanted to show you the fiscal year ‘17 budget. Subscriptions and books take up the majority of the budget.

Spending on Rubenstein and overhead from collections budget:

* + Rubenstein Library’s funding has increased to $500K from FY ‘08’s total of $88K.
  + 95% of FY 18 collections budget will pay for content through purchase, membership and licensing.
  + Remaining dollars support collection overhead: onboarding, hosting encouraging collaborative collections projects and open access subventions

Forming the Annual Collections Budget Request to the Provost

* Data sources: SAP/Aleph, industry analysis;/inflation estimates; contracted increases. Our justification is price inflation, and content inflation and new needs.
* From 2016 to 2017, we are expecting increases from 5.2 to 7l2% average 5.9% increase for journal.
* 60% of our journal subscriptions have a 4.2% increase in cost.

Strategies and flexibility

* Baseline fiscal discipline demands that we are careful stewards of our money. Rightsizing and leveraging opportunities.
* Careful discretionary budget management allows us to strategically reassign money to meet new needs.
* Spendable endowment balances are available to supplement purchasing and applied to support extraordinary one time purchases. If we are careful about curating what we have, it frees up money for us to buy new things. The better we do at managing, the more we have for new collections.
* New resources examples from FY ‘16 request: Sage Research Methods online - $101,000, NAACP papers, modules 4, 5 and 6, $112,000, early American newspapers, series 10 – 90K, Digital theater plus - $55K

In 2015, the Provost gave us all of the money we requested for new resources.

We requested a 4.2% increase and received a 2% increase. We received a -2.2% change from FY 17 to FY18.

1. #DataRescue – Preserving Threatened Federal Data Ensuring access to federal data – Naomi Nelson

We wanted to have a chance to bring everyone to speed about what’s going on with #datarescue and what the libraries’ role in that may be.

During the Presidential campaign, there started being more conversation around climate change, and so scientists got very concerned about the climate change data they currently use not being accessible in the future.

The End of Term web archive (<http://eotarchive.cdlib.org>) captures everything on a .gov website and partners with a number of universities and institutions to figure out how to do it and go out to do it. The Library of Congress, California Digital Library, University of North Texas Libraries, Internet Archive, George Washington University Libraries, Stanford University Libraries, and the U.S. Government Publishing Office have joined together for a collaborative project to preserve public United States Government web sites at the end of the current presidential administration ending January 20, 2017. This harvest is intended to document federal agencies' presence on the World Wide Web during the transition of Presidential administrations and to enhance the existing collections of the partner institutions.

In this collaboration, the partners will structure and execute a comprehensive harvest of the Federal Government .gov domain. The Internet Archive will crawl broadly across the entire .gov domain. The University of North Texas and the California Digital Library, will supplement and extend the broad comprehensive crawl with focused, in-depth crawls based on prioritized lists of URLs, including social media. This dual-edged approach seeks to capture a comprehensive snapshot of the Federal government on the Web at the close of the current administration. They will also accept nominations of urls.

The environmental data and governance initiative (EDGI) securely shares with us information about vulnerable programs or data via their email: [envirodgi@protonmail.com](mailto:envirodgi@protonmail.com). This email is encrypted and its servers are located in Switzerland.

The University of Toronto had one of the first data rescue events. They had a hack-a-thon on December 17, 2016 to save climate data before Trump brings overwhelming reaction. Penn libraries started a program called DataRefuge. Project partners include: University of Michigan Libraries, Internet Archive (End of Term Archive), Temple University Libraries, Environmental Data Governance Initiative (EDGI), ProjectARCC, Union of Concerned Scientists and Climate Mirror. They could expand all kinds of federal data and want to create a reverse depository system. Not receiving data, but selecting and harvesting it. They are not taking data off federal websites—but ensuring access to this high quality data. IA has a chain of custody, making sure it has provenance. The Data Refuge process also has a chain of custody to also ensure the authenticity of uncrawlable data. The hope is that the data will continue to be available through federal sources or through Climate Mirror, the data harvested through this process can serve as an additional check and source of verification.

DataRescue events create trustworthy copies of federal climate and environmental data, while the Internet Archive, datarefuge.org, and a consortium of major research libraries holds these copies. All of the tools created and workflows created are open source and open access.

The DataRescue events are a more intensive way of participating. It works most effectively if done together. Get people together with a range of different skills to capture data. U of Michigan had 275 participants. They got a terabyte of data processed. A pretty good day’s work. There is a lot of logistical work needed, food, their biggest piece of advice is to buy twice as many power strips than we would need.

What do we want to save? Websites, social media, data, etc. Analyze an agency and divide it into offices, figure out what they do, determine which are most at risk, prioritize. EDGI’s primer structure helps to make sure that nothing is missed.

This is the widest path and requires a variety of skill levels. Consider this path if you are a coder, hacker, have front end web experience, or just have a great attention to details. Volunteers search the internet and use a Chrome extension to seed the Internet Archive (telling it what to save), using the EDGI primer structure.

Data refuge track

* Researchers review the urls received, search for the best source
* Harvesters figure out the best way to download the data
* Domain experts verify the data (complete, what it says it is?)

As they’re doing that, they’re picking up provenance info, picking up the data, info about where this thing came from. Authenticity and provenance work. A researcher 50 years from now and will be able to see that the data is what it says it is. Trying to find ways to have authenticity built in but not have all the tools up front. Verifiers need to be someone who has expertise in the data, they are registered and their names are associated with the files they verify.

* Prepare for preservation and access (describe and bag)

Enter the librarians. Ensure data preservation. Can store the data in Data Refuge. Don’t need a ton of people doing this. Baggers need command line skills. (Files, manifest listing the files, checksums). Participants need to be identified and vetted.

Describers will then put the data in Datarefuge.com. They will add basic metadata. Participants need to be identified and vetted. Data is currently being stored in internet archives End of Term, Penn’s DataRefuge repository, and at the institution that harvested it.

Recent developments

This is a fast moving project – it’s only been around since the middle of the Fall and really started picking up in December. ARL is thinking of taking this on as an initiative.

They are looking to expand DataRefuge beyond climate data. Michigan is now taking up arts data and figuring out how to create a road map for that.

Christine Ogilvie Hendren, Executive Director, Center for the Environmental Implications of NanoTechnology (CEINT) is finding collaborators, connecting with Hack Duke, the Innovation co-lab and others, possibly making this a wider triangle project, collaborating with RTI, NCSU and UNC colleagues.

That’s what we know today about this but we’re interested in hearing from you about how this intersect with the work you’re doing. Is this something you’re interested in participating in? What do you think about expanding this past the hard sciences?

Someone suggested getting the Law school involved and students and asked if we anticipated sending out an email to the entire duke university community.

We don’t want to just make it a library project. It has to be a whole campus project. What is the Libraries’ role in this? Would we be coordinating this? It doesn’t have to be that way. Are we providing expertise or infrastructure for this, but this has happened in a number of institutions. OIT has expressed some interest in this.

Datasets that could be searchable – good idea to put throughout the community. Interest in a broader conversation about this. Sounds like there should be some sort of survey on campus and find a way to give that back as well.

North Carolina data, having something locally relevant would be a real driver. This could be related to the strategic plan - local to global theme.

We are being proactive before the administration takes down more information. We appreciate thought interest, comments, and we will come back with a plan.

Meeting adjourns 1:22 pm.