British Columbia Library Association

BCLA Browser: Linking the Library Landscape Volume 3, no. 1 (2011)

Sayeed Choudhury on establishing a university data management program

By Joy Kirchner.

In conjunction with the International 2010 Open Access Week (October Oct. 18-24th,), the BC Research Libraries Group invited G. Sayeed Choudhury, Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of Johns Hopkins University, to speak on the *Case for Open Data and eScience – Establishing a University Data Management Program at Johns Hopkins*.

Sayeed Choudhury discussed John Hopkins University (JHU) work developing a university data management program and a service model to support data curation as part of an evolving cyberinfrastructure featuring open, modular components in support of JHU faculty associated with community-wide eScience projects. In addition to developing a technological framework for data conservancy at JHU, they are also developing new roles and relationships between the library and the academic community, most notably through the development of "data scientists" or "data humanists." Within these developments, Choudhury concluded that institutional repositories is the first step in a longer journey towards data conservation and that for institutional efforts to be successful, they must be integrated into a larger landscape of repositories that serve a distributed and diverse academic community.

Developing an R&D, grant funding program at Johns Hopkins

Choudhury began the lecture by describing JHU's ten year journey to developing the Digital Curation Center. Initially known as the Digital Knowledge Center, Choudhury began his affiliation with the Library when he was a Graduate student in the JHU engineering program working at the Center on a one-year contract under the direction of the University Librarian.

The centre was uniquely conceptualized as a Research & Development centre embedded within the Library as a way of aligning Library development activities with research practices at JHU. Choudhury was charged with conceptualizing how this new unit could support the impact of technologies on Libraries.

Choudhury described the elements that shaped the development of the centre. He approached the development of Center's activities by identifying problems that needed solving on campus and initiated R&D support by applying for research grants to solve these problems. Successful grant applications not only gave external validation to the work the unit was doing, but also aligned the work of the centre with faculty research practices. He also ensured that the indirect costs of research went back to the Library; he also networked with a number of researchers on campus and credited their influence in helping him develop some early principles going forward including focusing on *automated systems* as opposed to automation; the former being concerned with making existing processes faster and thinking about when machines (algorithms /code) are most effective and where people are most effective. Later as the Center's funding opportunities diversified and they received corporate financing this also influenced the Center's approach to move more to a project management approach involving specialists, most notably software developers, who were hired for specific projects.

Both establishing R&D projects within a library setting and writing research grants was very new to the Library. In the beginning Choudhury acknowledged some cultural dissonance between the Center's approach and traditional library practices. Even so, Library cultural practices also influenced a unique form of R&D that would not have happened if the Center wasn't placed within a Library setting. He explained that although there were many elements of their R&D that were similar to other research units on campus, he also spoke of developing an R&D model to uniquely support a service model - in this case, new library services – rather than producing a research paper ultimately, for example.

Initial projects that led to digital curation projects

The initial projects in the late 90's and early 2000's were based on converting analog to digital formats.

ISSN 1918-6118

One project included building a robot to digitize sheet music that could also produce musical sound from the sheet music. They began to engage computer scientists, engineers and others to assist with projects and they built name authority tools and metadata tools.

Later he discovered faculty were digitizing content and building discovery tools to access content but no one was thinking about the preservation aspect. He began to make connections with specific disciplines on his campus that were doing interesting digital projects and asked them how they were preserving their work. One of these was the SLOAN Digital Sky survey (http://www.sdss.org/), an ambitious project to digitize astronomical observatory images and release these publically. He learned preservation was not a question they were considering. It was at this point Choudhury saw a significant role for the Library and which eventually led to a partnership to work with the SLOAN group on the digital conservancy piece.

Approaches to data conservancy

From the SLOAN researchers he learned more about how astronomy researchers and scientists look at data releases in terms of levels of data flow. He explained data flow 0 is the actual telescope data that goes through various calibration processes that not too many people can actually read; level 3 of the data flow are essentially databases and level 4 of a data release are typically the kind of data that gets cited and where research queries can be constructed to run against the data.

Choudhury opted to focus on level 4 data as the most useful place the Library could bring expertise. His work involved archiving and linking research data to the publications that resulted from the research. The project was very successful and led to the Library gaining a better understanding of the research and publishing practises within the astronomy discipline. In turn, Choudhury's group obviously impressed the astronomers and they asked his team to begin to work on preservation aspects of the other data levels.

From this experience, his team also branched out to other disciplines which included working on humanities data curation projects involving the digitization of medieval manuscripts and archiving the research data connected to this project.

The team also used their R&D and grant funding experience as a vehicle to support analysis and assessment of software for library projects. Choudhury describes successfully working with others to acquire funding from the Mellon foundation to support an assessment project to research the selection of repository software. In this example, Choudhury described how his team began with use-studies, identified disciplinary requirements and mapped the requirements to a particular system as a way of defining best choice for data conservancy.

All of these projects later evolved into thinking about infrastructure development for data management. Again, Choudhury applied for a grant with other research entities and was successful in receiving a National Science Foundation grant to explore infrastructure for data conservancy. Choudhury ensured the Library was the project lead.

On building data infrastructure for data management

In talking about data infrastructure or *cyberinfrastructure*, Choudhury explained that this is very early days of this development and described it in terms of an "invention stage." For this reason his initial approach to infrastructure development was to employ a set of navigation principles rather than developing a rigid road map that might result in a path-dependent approach that would be difficult to fix later on should they need to. Secondly, he felt it was important to start with a definition of data curation that embraces a shared vision with his partners. The definition they came up with is: "data curation is a means to collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society."

In working through an infrastructure model, he spoke to his belief that you can't build services on things you don't have and that you need to have the data first before considering preservation. He also described how he spent a lot of time thinking about what a flexible architecture would look like and looked to community reference models, such as those models developed through the Open Archives Initiative (http://www.openarchives.org/)and the PLANETS Project framework for long term preservation of digital content (http://www.planets-project.eu/about/)

Questions to ask before starting on a data management project

Choudhury recommended a set of questions one should ask before embarking on a data management project:

ISSN 1918-6118

- what data are you preserving;
- will you consider preserving the context that data exists in;
- how much depth and breadth do you want to preserve.
- Are you considering algorithms used to process data?
- Are you considering the storage frame work.

Finally, he strongly suggested that one cannot truly understand data preservation until you have had the experience of handling data first. He also strongly recommended that the framework that is developed should not disrupt existing disciplinary workflows. He also acknowledged that there is no theoretical framework for data conservancy currently but this will be needed over time.

Who is an ideal data manager?

One audience participant asked Choudhury what he is looking for in a data manager. Choudhury responded with a description of an ideal candidate: deep disciplinary knowledge, can converse with scientists in a meaningful and respectful way and will be viewed by a scientist with credibility. Also they should have significant metadata knowledge and software developments skills or understanding of architecture. He was quick to say this was an ideal picture and one is not likely to get all these attributes in one person. He mentioned there are a number of education opportunities that are worth pursuing such as the summer institution on data curation offered at the University of Illinois, Graduate School of Library and Information Science. http://www.lis.illinois.edu/articles/2010/05/gslis-host-2010-summer-institute-data-curation

Practical advice and lessons learned

Choudhury concluded the session with some more practical advice and lessons learned drawn from his own experience. These include:

- The importance of drawing up agreements with your data partners. At JHU they utilize Memorandum of Understanding (MOU) - five year explicit agreements that itemize costs, payment plan, support needed, considers what happens if something goes wrong and how to address this.
- He also recommended tapping into the business school as they did at JHU. Students were hired to construct storage analysis costs, market analysis, disaster recovery plans.
- He recommended enlisting expertise. For instance, with the new NSF requirement to construct data management plans, his team responded by suggesting they could help build the NSF template for a data management plan. They enlisted the help of a company that builds decision support systems and life science researchers at JHU.
- He spoke to the importance of fostering a culture of innovation within your organization. He explained innovation can't be enforced but comes from a few individuals. Organizations need to find ways to foster and encourage these individuals. Develop an organizational mantra that allows a willingness to fail. Innovation is often born out of failure.
- Know when it is important to lead and when it is better to defer to an expert.
- Recognize grant funding can't sustain a project long-term - build in sustainability measures.
- Know when to let go of a cultural practice that no longer benefits an activity. It is more important (and more difficult) to unlearn a cultural practice that no longer supports an activity then it is to learn something new.
- He described data as the "new special collections" and suggested it should be thought of that way.
- It is essential to engage faculty or researcher champions who can become your ambassadors. Relationships and partnerships with others are critical to success.
- Choose projects carefully. Scope is very important.

For a viewing of Sayeed Choudhury's archived talk and other BC Research Libraries Lectures, see: http://blogs.ubc.ca/bcrlglectures/

Joy Kirchner is a Librarian at the University of British Columbia and one of the BCRLG Lecture Series program Coordinators.

ISSN 1918-6118