Notes from IDCC 2010 CURATEcamp Preconference Workshop
Facilitators: Declan Fleming (UCSD), Mike Giarlo (Penn State)
Note taker: Patricia Hswe (Penn State)
6 December 2010

I. Introductions by attendees

- Alex Ball - UKOLN/University of Bath
- Charles Blair - U. of Chicago
- Eric Chen - Cornell
- Elizabeth Coburn - University of Illinois at Urbana-Champaign (GSLIS)
- Betsy Coles - California Institute of Technology
- Joann Croucher - U. of New South Wales
- Robin Dale - LYRASIS
- Mark Evans - Tessella
- Emily Gore - Clemson U.
- Neil Grindley - JISC
- Thomas Habing - U. of Illinois at Urbana-Champaign Library
- Steven Holloway - U. of Illinois at Urbana-Champaign (GSLIS)
- Huda Khan - Cornell
- John King - Ursinus College
- Katherine Kott - Stanford
- David Levinson - Lake Forest College
- John Mark Ockerbloom - U. of Pennsylvania
- Denise J. Massa - U. of Notre Dame
- Mairead Martin – Penn State
- Erin O'Meara - U. of North Carolina-Chapel Hill
- Belinda Ramnauth - Carnegie Public Library
- Linda Reib - Arizona State Library (Archives and Public Records)
- Dorothea Salo - U. of Wisconsin-Madison
- Sarah Shreeves - U. of Illinois at Urbana-Champaign (Library)
- C.M. Sperberg-McQueen - Black Mesa Technologies
- Peter Van Garderen - Artefactual Systems
- David Walls - U.S. Government Printing Office

II. Proposed Topics

(Asterisks indicate high level of interest)

- Documenting and sharing digital curation use cases ***
- Leveraging the NSF's demand for data plans **
    - Roles libraries are playing to assist researchers with this requirement?
    - Also, do others see this as an opportunity to further campus wide data management or are requirements like these more of a hindrance?
- "Quick wins" in digital curation ***

    - Which resources (tools, models, frameworks) can be used by people who have limited experience of curation tools to deliver clearly beneficial info management outcomes?
- How we can best work as groups over time to curate and preserve content?
    - Surviving attrition
    - How LOCKSS is helping at UPenn
- Moving from software development projects into managing the ongoing software dev and the live services side of running an active repository
    - Transition from "traditional" institutional repository support to broader and deeper systematic digital curation
    - From software intended for repositories of papers to _____?
    - Migration and validation tools
- DEMOS
    - Archivematica - http://archivematica.org/wiki/index.php?title=Main_Page
    - Curators' Workbench - http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/
    - Tessella's SDB platform - http://www.tessella.com/2010/08/press-release-tessella-unveils-world-leading-multi-tenant-digital-archiving-platform-2/
- Curation Microservices ****
- Sharing?
- Local vs clouds
- User communities to put content in
- Use case development ***
- From IR to DC **
- Data Models
- RDF/Semantic Web
- Accessioning
- Tracking agents through time
- Knowledge representation
- Departmental Collaboration
    - De-Siloization
- Failure
- Tools and Tech***
- Faculty Profiling

III. Curation microservices

- What are microservices?
    - How do we make sure that those of us doing microservices aren't doing 20 different things that aren't compatible?
    - Modular services - in a generic way
    - At Stanford - Hydra (https://wiki.duraspace.org/display/hydra/The+Hydra+Project) is considered a

microservices-based application
    - Another characteristic: You're trying to keep data/content long-term - preserve integrity of data and their description; building microservices is largely about making sure they operate in ways that data integrity is understood and respected.
    - Also means creating a dumb as bricks infrastructure - in trying to preserve data, you don't want to have to keep track of the infrastructure as well.
- How do you keep a modularization platform honest?
- Further definition: Microservices are tightly focused tools, yet loosely joined.
    - Another take on microservices from developer's point of view: It's about becoming professional software developers as opposed to boys in the basement
    - At DLF there was discussion about agile methodologies to get working with stakeholders, so can have enough information to go on – this is also so you don't get stuck in a technology trench and you can't get where you want to go very easily. Microservices can enable, or at least inform, this kind of approach.
    - UW-M has really good software development shop (agile before agile was popular) – but, as at many libraries, it's understaffed.
- Some techie thoughts about microservices:
    - Unix tools assume a data model/file system underneath - fit together because of common understanding of data they're dealing with.
    - Tricky thing with curation microservices - have to have understanding of applications (BagIT, OAI-ORE), assumptions of underlying data, and realize that one tool needs to understand what the other tool is doing/about
- Additional definition of microservices: Verifying checksum, creating a BagIt package - granular services that you can stitch together in comprehensive system . . .
- Technology dependencies?
    - CDL has specs and a reference model, also a particular technology stack - what are other technology solutions?
    - Spec will stay the same, service agreement will stay the same but curation microservices approach allows you to just swap out what you need to change . . .
- Fedora in microservices approach - how do you deal with that?
    - Does this microservice assume certain objects when it runs?
    - At U. of Chicago – we're pursuing Fedora and microservices together.
        - "When doing data curation: heterogeneity is your friend, homogeneity is your enemy."
    - At U. Wisconsin-Madison, we're building a Fedora infrastructure but also constraining it – so there are concerns on this point.
    - At Stanford, also taking a microservices approach using Fedora and Hydra.
        - One application created something called "Everyday Electronic Materials" - workflow for subject librarians (selectors) to process digital materials (http://lib.stanford.edu/eems)

IV. Data Management Requirement

- U. Wisconsin - did a lot of training, fielded a lot of phone calls, key audience are grant administrators (these are people you have to get to)
    - Strategic planning process - fought tooth and nail to have libraries included
    - Website: http://dataplan.wisc.edu/
    - Cyberinfrastructure day events
    - Central IT org - provide storage $1-2/GB, but not backing up data
- Other people who care are folks who run IT for colleges on campus
- UNC-Chapel Hill
    - Had training courses, tapped into existing resource (ODOM institute, mailing list of 7000 people), quite a few people from Health Sciences came even though it was for social science data
    - Key: get plan on the record. If you submit a DMP, you must have consulted with repository staff
- Key question: who pays for storage? Look into what Princeton is proposing via DataSpace (http://dataspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf)
- If NSF is requiring DMPs in grant proposals – there should be money for managing that data over time that is accounted for in budget proposed for the grant project.
- UIUC's experience
    - Met with Grants and Contracts folks, their take: NSF is going to be looking carefully at what institutions are going to be looking at, including indirect cost recovery.
    - Argument is that you want to build out collections through research data.
    - Probably going to have to make some arguments about what has been done to support researchers . . .
- At Cornell - recognition that there needs to be some coordination, library metadata services will do part of the project (i.e., help in DMP).
- Library system becomes broker for data
- Question: How are people thinking of handling size of data?
    - Paying for it? Planning to charge?
    - University has to realize digital stuff is being created – and that faculty is interested – but who is going to pay?
- Spaces for faculty collaboration and sharing of information (e.g., simulations, course materials)
    - HUBzero (http://hubzero.org/) – e-science gateway
        - Whole idea was to share computing resources to perform simulations
        - Package of existing software (written with Joomla!)
        - Uses VIVO (http://www.vivoweb.org/) for profiling, make it interoperable
        - Example of HUBzero instance: OpenParks Grid (http://www.clemson.edu/administration/president/report/documents/open_parks_poster.pdf) - dealing with millions of objects in parks

resources
    - But HUBzero doesn't do the best job for data management and presentation
    - Open Parks Grid is using it for front end, Fedora on back end
- IRs and data management: How do we think we're doing it differently? Nobody wanted to give us stuff when we started IRs. What are we doing differently to obtain collections we're going to curate?
    - Penn - has IR, largely for papers. Had a full-time data/content wrangler. Then she left, wasn't replaced. Even in context of IR, people wanted to do things they hadn't originally thought of doing.
        - Taking a couple of steps - system for publicizing people's work (look at VIVO), building up a repository whose initial use cases are special collections, putting in a bunch of digital data and managing it (haphazardly). Transition to become IR for handling data.
    - UIUC - using IDEALS as place to collect data, not the best but it's what they have and can put things into it.
        - For ETDs - using Vireo, capturing supplemental files providing nice case studies for data curation that UIUC wants to do.
        - IRs get a bad rap - but as a starting place, can begin to collect some of the data and get it in.
        - UIUC also works with faculty to collect read me files.
    - What about systems – like BePress – that are optimized for papers, rather than data sets?
    - What are incentives for obtaining these at-risk materials? U. of Rochester - faculty said they weren't going to do anything that would require a whole separate set of steps. What faculty want is something already a part of their scholarly process.
        - But what are we doing to activate the use cases we're discovering? Are we actually interacting more with researchers within the digital curation model?
    - At Clemson - while they're building the system, libraries hook into faculty citation, then try to get the work cited into IR.
    - Maybe look at discipline-based repositories out there - and try to understand that sphere.
    - Could use something like VIVO to manage different locations of the same work we need to track. E.g., here's the publisher version we can point to, here's pre-print, here's post-print, here's data set that's associated with it.
- With data often there isn't yet an infrastructure in place . . .
- At University of Chicago - we have a policy, but it's aimed at selectors.
- At Data Conservancy – there's a collection policy for data, so making the effort to engage within a broader context (scientific community).
- Important to understand collection policy from content standpoint - but also what about use cases?

    - Use cases are going to drive microservices.
    - In order to be successful, you have to understand what the use cases are. It isn't easy to get to that point for some people.
- What are some challenges?
    - Library environments - hard to argue to money guys that you need to build in layers. Libraries think in silos.
    - Need to ask: who are stakeholders?
    - More often than not, response to data mangement planning is consulting – consultation services are what is offered.
        - Rather, need to have a process in place that you can plug people into.
    - What do best practices mean in this arena? (Dev, test, prod, QA, tools used for code repositories)
    - What is governance like in this space? How do you make governance decisions?

Note from Patricia: If anyone has more to add to this section, or the next on Quick Wins, or notes to contribute about the use case discussion we had, then that would be great. Email me at phswe@psu.edu.

V. Quick Wins

[Note: some of the points below may have more to do with collaboration and working together than with quick wins.]

- What are the things your institutions are doing to enable quick wins?
- What does the user mean by digital preservation?
    - Need to write down requirements
- Peter Van Garderen - how to get the pieces to work together; you can point people to 50 different tools, but are they going to know what to do with them?
    - People need something like Flickr.
- What is the right approach for involving end users?
    - Use cases - user stories - where to start collecting those?
    - Are there aspects to user stories that are generalizable?
- Dorothea - first son parable - give me all this knowledge and I will make use of it; second son - what is this to you?; third son - what is this? (For more, see BoT post: http://scientopia.org/blogs/bookoftrogool/2010/12/07/the-four-sons-of-digital-curation/)
    - Bring idea of partnering, but a lot of people think they can throw a smart grad student at it and the problem will be solved; or grad student puts it up but they're gone after two years and then it's not sustainable.
- What about PLANETS - why aren't we using this capacity? (http://www.planets-project.eu/)
    - There's a need for pre-education even before we can use that environment.

    - PLANETS - very specific tool, but doesn't solve immediate issue of access
    - How do you sell what PLANETS did to a broader community?
    - What's the use case for PLATO?
        - PLATO is already too much for archivists that Archivematica works with
        - Often, you have to work with what you have.
    - PLANETS is working on registry to supplement what Pronom is doing.
- How do you make policy for digital preservation formats?
- Leverage DropBox approach? Especially as far as getting a common file space among disparate people (like even 3rd and 4th sons)
    - uses the cloud
    - makes some assumptions about identity

VI. How we can best work as groups over time to curate and preserve content?

- Community-based challenges - how can we incorporate what we know is happening around us so we don't have to build everything ourselves?
- Another set of pressures that work against collaboration, to a certain degree - we work in competition with each other for soft funding.
    - On other hand granting agencies are encouraging collaboration.
- Library of Congress NDIIPP - a lot of funding in the beginning, people competed heavily for it, but tools weren't shared at first.
- Create and contribute to a shared use case repository
- What is DLF's role in the larger community?
- What about http://digitalcurationexchange.org/?

VII. Demos

- Demo of Cuator's Workbench
    - haven't tested it for email
    - largely for pre-processing large undescribed collections
- Archivematica - Using Micro-Services and Open-Source Tools to Create a Comprehensive Digital Curation System
    - Looking at a lot of digital forensics tools
    - Archivematica is pipeline for creating AIP
    - Developed default normalization path.
        - Compiles the METS, creating a BagIt.
    - VM building
        - Build on Ubuntu
        - Do all developments, end-user deployments off same VM
        - Virtualization has made a huge differece
    - Created Debian packages for JHOVE, because they needed them
    - Would be nice if there were a universal standard for SIPs
- Mark Evans, SDB Digital Archive (Tessella): http://www.tessella.com/2010/08/press-release-tessella-unveils-world-leading-multi-tenant-digital-archiving-platform-2/
- Who's using METS? Who's not using METS?

- At Stanford - METS barrier of entry was so high, now not using it for current repository environment
- Another demo, by Tom Habing - Hub and Spoke Workflow Manager, http://dli.grainger.uiuc.edu/echodep/hands/index.html
  - Uses common METS profile for interoperability
  - Performs preservation actions and return to repository
  - H&S did checksum validations, versioning of METS pkgs, so every time you pull something out of repository, you got a Master METS file and one or more profiles conforming to the METS profile (METS pkg was privileged file, ingested along with other files)

VIII. Further Discussion about METS

- Stanford - pared down the data model, big issue for them was getting through-put (getting past pre-processing bottle neck); they're blending curatorial approach with large-scale data
- Metadata actually takes up a lot of space - can outstrip files, processing it takes a while
- If you can regenerate metadata, why save it?
- Can you do everything with bags that you can do with METS?
- Declan - METS can be complex but it's not directly actionable without a profile and lots of programming. RDF is inherently linkable.
  - METS is QA tool for UCSD.
- METS, because of its XML nature, makes some assumptions about serialization, while RDF doesn't have any of that.
- Elizabeth Cockburn (GSLIS): Institutions wanting to implement standards - have a hard time finding information (such as PREMIS profiles)
  - Just showing what elements you're using (or semantic units) would be valuable to share
  - GPO has a fully operational PREMIS implementation (but GPO's profile isn't registered in registry)
- One differentiation: METS is about the structure, while PREMIS gives information about what's being done to these objects
- Difference between METS and RDF: What is RDF? Something you can break down into a data model. Outside of libraries, better known than METS.

IX. Data modeling and helping faculty/researchers

- At Wisconsin, we're looking on much higher level at kinds of things we have
  - One example is have folk songs, which we're trying to figure out how to represent - part of it may be series of pages that represent the music, part of it could be oral history,
- Stanford article in D-lib on Fedora - reflection of this implementation and how they're changing to new system, lessons learned: http://www.dlib.org/dlib/september10/cramer/09cramer.html)

- Should we be thinking more along lines of FRBR kind of data model?
    - FRBR model can be built up incrementally - aggregate things with common properties
    - Traditional archival approaches may help here - groupings, rather than object-level access.
- U. of Chicago - how do you design a repository, including how you're going to record things about your objects so that people can build whatever silo they want, whatever catalog they want?
- Who's building the access systems? We can take stuff even if people don't know how to make it accessible – an example is the Preserving Virtual Worlds project (http://pvw.illinois.edu/pvw/), which archives computer and video games but doesn't necessarily make them accessible via a repository.
- Maybe the approach should be: Do enough to bring researcher to the right box or folder and let them sift through it. This way they can be drawn to annotate, do folksonomies, etc.
- At the same time, we're all using different models - why don't we get on the same page and *use the same models*? Troubleshooting would be easier.
- Researchers are asking for help at the start, not the end - managing data from the start of collection.
    - Islandora (http://islandora.ca/) does some of this.
- How do you get things from faculty so that they're not taking care of it - but how do you do this cleanly?