Digital Stewardship in the 21st-Century Library
The Penn State Experience

Michael Giarlo and Patricia Hswe
ALCTS Midwinter Symposium
January 7, 2011
Framing digital stewardship at PSU

• Content Stewardship Program at Penn State
• Deploying a Digital Collections Curator and Digital Library Architect
• Organizational contexts for curation in practice
• Current status and prospects
But first: some context

Why Penn State launched the Content Stewardship Program
It started with scientific data management

http://www.flickr.com/photos/ethanhein/2272885283/
At the same time . . .

There were other considerations
Got silos?
We’ve got these 4.

But we did steer clear of IRs

* The Libraries have experimented with Fedora, but there’s been no formal instance of it as an institutional repository at PSU.
Emphasis on digitization/production
(which we do very well)

Too well?

In mounting collections online, we weren’t considering the whole picture of lifecycle management.

http://www.dcc.ac.uk/resources/curation-lifecycle-model
About
The Content Stewardship program is an institutional initiative to address digital content and data management needs in areas such as digital library collections, scholarly communications, electronic records archiving, and e-science/e-research data management. Building on existing services and infrastructure, the program is developing an interoperable and extensible suite of discovery, preservation, curation, archival, and storage services.

RECENT ENTRIES
What we've been up to
This has been a busy year for us and we have taken a multi-pronged approach to meet our programmatic...
By MAIREAD MARTIN | Comments (0)

Welcome to the Content Stewardship Blog!
Welcome to the blog for the University Libraries and ITS' joint Content Stewardship program. In April 2009, I blogged about...
By MAIREAD MARTIN | Comments (0)

Home
This blog is licensed under a Creative Commons License.
Program’s First Year

• National search for Digital Collections Curator and Digital Library Architect
  – Year-long effort

• In parallel, investigation into storage issues
  – eXtensible Access Method (XAM) prototype
Digital Library Architect – job posting

• **Reports to ITS Senior Director of Digital Library Technologies**

• Leads design and development of technical architecture for DL applications, middleware, systems, and services

• Develops, publishes, and maintains roadmap for service and infrastructure implementation

• Collaborates on and contributes to strategic tactical planning and implementation in content stewardship domain

• With ITS and University Libraries (UL) leadership, develops a Content Stewardship governance model inclusive of stakeholder and user needs

• Engages in national and consortial efforts in the digital library domain, seeking mutually beneficial partnerships with peer institutions and in the private sector
The job of a digital library architect is roughly equal parts strategy, advocacy, evangelism, project management, and Art Vandelay jokes.
Digital Collections Curator – job posting

• **Reports to the Assistant Dean for Scholarly Communications**

• Leads development of **inclusive user-focused agenda** for digital scholarly content stewardship

• Investigates, recommends, and develops plans for **user-focused and repository-based services**, towards effective **management** of the creation, collection, and distribution of high-value digital scholarly content

• **Manages** broad set of existing digital collections and repository content

• Researches and develops in-depth knowledge of **emerging technologies**, **relevant national standards**, and **best practices**

• Communicates effectively with **internal stakeholders**
Contact Information

Office Address
Penn State University Libraries
W-311 Pattee Library
University Park, PA 16802

Office Phone
814-867-3702

Email Address
phswe AT psu DOT edu

Chat
pmh22@chat.psu.edu

Twitter
@pmhswe

Professional Links
LinkedIn Profile

Currently . . .

I am Digital Collections Curator at the Penn State University Libraries,
University Park Campus.

Digital collection curation marks a key component of the Libraries' approach
toward developing a cyberinfrastructure, e-content, and data stewardship
program, as stated in the Libraries' strategic plan.

As Digital Collections Curator, I concentrate mainly on three areas:

- **Assessment** - usability/usage of digital collection content
- **Data/Content Management** - development of policy, standards, and
  best practices toward improved functionality and discoverability of our
digital collections
- **Repository-based Services Planning** - planning and coordination of
  repository development, based on user studies and exploration of areas
  for potential services and collections

In the Libraries my affiliation is with Scholarly Communications, which
comprises Digitization and Preservation, the department of Scholarly
Communications Services, and the Office of Digital Scholarly Publishing.
In parallel, investigation & prototype of XAM

http://www.snia.org/forums/xam/

The work on XAM confirmed the need for a unified technical architecture
Deploying a Digital Collections Curator and Digital Library Architect

Activities in the first year
Got silos?
We’ve got these 4.

Platform Review
<table>
<thead>
<tr>
<th>ELEMENT</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adoption</td>
<td>How widely is software adopted and used?</td>
</tr>
<tr>
<td>Developer ecosystem</td>
<td>How extensible is software? How strong is developer support?</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Number of standards supported by software, in conjunction with other standards?</td>
</tr>
<tr>
<td>Metadata standards</td>
<td>What metadata standards are native supported?</td>
</tr>
<tr>
<td>Migration</td>
<td>What are migration capabilities of the system?</td>
</tr>
<tr>
<td>Object format support</td>
<td>What file formats does the software support?</td>
</tr>
<tr>
<td>Search</td>
<td>What kind of search functionality is exposed to the user? Evaluation of functionality?</td>
</tr>
<tr>
<td>Storage abstraction</td>
<td>What are the ways that data can be restored?</td>
</tr>
<tr>
<td>Upgrade</td>
<td>Ease and reliability of software upgrades?</td>
</tr>
<tr>
<td>Versioning</td>
<td>Does software preserve original contents when versions are made?</td>
</tr>
</tbody>
</table>

Curation Microservices Pilot Project

- Independent, interoperable services
- Small, self-contained
- Swap out/replace as needed
- Complex functionality arises from "strategic combination" of microservices
A few words about use cases

• No use cases = no services

• Challenges
  – What is a use case?
  – Constraints of working with particular systems
  – Lack of documentation on use cases

• Helpful tip: use case ~ workflow
Example of a use case

<table>
<thead>
<tr>
<th>Use case narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The curator has to manage the migration of CDs and DVDs to server storage. In doing so, he tests to verify that the data have transferred properly at the time of transfer; then he periodically needs to test that data in order to ensure they are still viable.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Use case process (workflow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;CURATOR MIGRATES AN OBJECT&quot; WORKFLOW</td>
</tr>
<tr>
<td>1. Curator receives optical media disk.</td>
</tr>
<tr>
<td>2. Curator determines what data files are contained in the disk.</td>
</tr>
<tr>
<td>3. Curator determines extent of the files (including checksum).</td>
</tr>
<tr>
<td>4. Curator uploads all data files from the disk to the server.</td>
</tr>
<tr>
<td>5. Curator receives confirmation that all files have transferred properly (i.e., verifies number of files, checksums, etc.).</td>
</tr>
<tr>
<td>6. Curator opens the files to verify readability.</td>
</tr>
<tr>
<td>7. Curator verifies readability of all data files.</td>
</tr>
</tbody>
</table>
Building a digital curation community - 1

WELCOME TO CURATECAMP!

Howdy, Campers! We're glad you're coming to CURATEcamp!

We've based CURATEcamp on the BarCamp or “unconference” model which may be very different from other conferences you've attended. This post is meant to provide some orientation around what you can expect of CURATEcamp, what will be expected of Campers, the overall theme for discussion topics, and some next steps for you.

HOW TO PREPARE FOR CURATECAMP

Be prepared to participate: come with an idea or two for sessions you can lead. Even better, add that idea to the agenda in advance! If you're not prepared to lead a session, no problem; you don't have to be an expert at your topic. Find a topic that interests you and contribute to the conversation however you can. We all bring different contexts and points of view: you can add data points to advance the discussion; you can ask questions that others might not have considered; you can demonstrate something. These are great ways to participate, since they spread knowledge and provoke nuanced group discussions instead of unidirectional lectures.

THE CURATECAMP AGENDA

While CURATEcamp is an unconference, we do have an overarching theme of digital curation. We intend for the Camp to be equally interesting to all practitioners, whether you develop software or do actual curation. We hope that all Campers come with an interest in digital curation, if not expertise with any one approach. Since this is a relatively new approach, we expect that there will be much more interest than experience among Campers. And that's a-okay; we're "building" the community around this approach.

CAMPING 101
Building a digital curation community - 2

• Listserv – Digital Curation (Google Group)
  – http://groups.google.com/group/digital-curation
• Blog – Content Stewardship @ Penn State
  – http://stewardship.psu.edu/
• Facebook – CURATEcamp Group
• Conferences
  – DLF Fall 2010 Forum, IDCC 2010
  – Code4Lib 2011
Organizational Contexts for Curation in Practice
Our organizational context
Curation in practice

Penn State Content Stewardship Program

Partnerships

User Assessment and Needs

Infrastructure planning and assessment

Policy and process

Service Development

- Consulting & Education
  - Intellectual property
  - Training
- Data management planning
  - Managed data services
    - E-Records
  - Online authoring environments
  - Publishing services
    - Acquired/licensed content
    - Theses
- OAIAS Reference Model
  - Service framework development
  - Platform reviews & exit strategies
  - Preservation networks
    - Caltech Trust
    - Penn State storage
    - Front-end aggregators/Interaces
- LOCKSS
  - MetaArchive
- Records retention schedules
  - Format standards
  - Record storage access
  - Data integrity assurance
  - Resource management
  - Copyright/license compliance
  - Accessibility guidelines
  - Collection guidelines

12/2/2010 - rslmalough@psu.edu
Current Status and Prospects
<table>
<thead>
<tr>
<th>Microservices</th>
<th>DCC Lifecycle Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authenticate</td>
<td>Access, Use &amp; Reuse</td>
</tr>
<tr>
<td>Migrate</td>
<td>Preservation Action</td>
</tr>
<tr>
<td></td>
<td>Store</td>
</tr>
<tr>
<td></td>
<td>Transform</td>
</tr>
<tr>
<td></td>
<td>Access, Use &amp; Reuse</td>
</tr>
<tr>
<td>Fixity</td>
<td>Preservation Action</td>
</tr>
<tr>
<td>Store</td>
<td>Store (part of the Migration process)</td>
</tr>
<tr>
<td>Identity</td>
<td>Preservation Action</td>
</tr>
<tr>
<td></td>
<td>Ingest</td>
</tr>
<tr>
<td>Annotate</td>
<td>Description and Representation Information</td>
</tr>
<tr>
<td>Version</td>
<td>Ingest</td>
</tr>
<tr>
<td></td>
<td>Preservation Action</td>
</tr>
<tr>
<td></td>
<td>Store</td>
</tr>
<tr>
<td></td>
<td>Access, Use &amp; Reuse</td>
</tr>
<tr>
<td>Characterize</td>
<td>Description and Representation Information</td>
</tr>
<tr>
<td>Authenticity/Audit</td>
<td>Ingest</td>
</tr>
<tr>
<td></td>
<td>Preservation Action</td>
</tr>
<tr>
<td></td>
<td>Store</td>
</tr>
<tr>
<td></td>
<td>Access, Use &amp; Reuse</td>
</tr>
</tbody>
</table>

These actions make up the Migration process in the lifecycle model.
Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences

Jeremy Goecks¹, Anton Nekrutenko² *, James Taylor¹ * and The Galaxy Team

* Corresponding authors: Anton Nekrutenko anton@bx.psu.edu - James Taylor james.taylor@emory.edu

For all author emails, please log on.

Published: 25 August 2010

Abstract

Increased reliance on computational approaches in the life sciences has revealed grave concerns about how accessible and reproducible computation-reliant results truly are. Galaxy http://usegalaxy.org [website], an open web-based platform for genomic research, addresses these problems. Galaxy automatically tracks and manages data provenance and provides support for capturing the context and intent of computational methods. Galaxy Pages are interactive, web-based documents that provide users with a medium to communicate a complete computational analysis.
For more information about Nekrutenko Lab, see:
http://www.bx.psu.edu/~anton/ and http://www.rps.psu.edu/indepth/galaxy.html
Full-circle: doing scientific data management

http://www.flickr.com/photos/ethanhein/2272885283/
Thank you!

michael@psu.edu

phswe@psu.edu

Any questions?