Lessons from Engaging with Repositories and Cyberinfrastructure Initiatives

Gary Motz, Jon Dunn
Natural History Collections and Digital Repositories
Imago

- Repository for data from two IU natural history collections
- Generalized DarwinCore schema
- Prototype awaiting replacement repository
CollectiveAccess & Resource Space Tandem (CARST)

- Repository for geological collections
- Over 2 million object records from several dozen collections
- Emphasis on data preservation and accessibility to public
- ABCD-EFG, ISO 19115, DarwinCore, and EML schema
University Collections

- Novel upper-level administrative position to coordinate resources for all collections
- Goal to provide CMS, DAMS, and public-facing web accessibility for all
USGS National Digital Catalog

- Repository for US geological collections data
- Federated collections-level discoverability for state surveys and USGS science centers
- Currently manual entry of metadata; duplicative efforts
Cyberinfrastructure Needs for Digitized Collections

- **Imago / CARST / IUCollections / USGS NDC**
  - 2 servers, <0.5 FTE dev, <100GB storage for 100,000+ media
    - Automated, distributed backup of archived records and media
  - 2 servers, 3.5 FTE dev, >700TB storage for 2,000,000+ media
    - Manual backup of records and media to on-site tape
  - 2 servers, 5 FTE, <unknown> storage/media requirements
  - 1 server, 0.4 FTE dev, ....

- Modest requirements, substantial hurdles
IU Libraries
Repository Services
Imago

- Repository for natural history collections data
- Based on Sufia (Hyrax predecessor), Samvera, Fedora
- Goal to inform future work on digital collections and research repositories
Current IU Libraries Repository Environments

- Digital collections
  - Format-specific repository services
    - Image Collections Online (Fedora/local application)
    - Media Collections Online (Fedora/Samvera/Avalon)
    - Archives Online (Fedora/XTF/local application)
    - Etext platforms (XTF)
  - Online exhibit platforms (Omeka)

- University research
  - IUScholarWorks institutional repository (DSpace)
    - Research publications, presentations, and data
Digital Collections Repository Environment

- Focus on library special collections and archives
- Service and technical ownership in Library Technologies
- Collaboration with IUPUI, University Information Technology Services, IU Office of the Bicentennial (President’s Office) to expand to libraries/archives university-wide
- Reliant on standards to enable scalability of systems and support models: EAD, MODS, IIIF
- Heavy focus on audio/video driven by IU Media Digitization and Preservation Initiative (MDPI)
- Consolidation on Samvera/Hyrax
Research Repository Environment

- **IUScholarWorks**
  - Current DSpace-based repository for research publications, papers, presentations, data
  - Format-agnostic but standardized metadata (Qualified Dublin Core)

- **DataCORE**
  - Hyrax-based research data repository currently under development
  - Based on UM’s Deep Blue Data work
  - Goal: more flexible metadata structures, greater interoperability with other systems via APIs

- Service ownership in Scholarly Communication (Public Services); technical ownership in Library Technologies
Repository Storage Infrastructure

- Mandate to use central IT-provided storage resources
  - SAN or CAS for small files / immediate access (Enterprise Systems)
    - Hitachi SAN, Hitachi Content Platform
  - HSM for large files / archival (Research Technologies)
    - IBM HPSS, IBM enterprise tape

- Fedora managed by central IT; soon application hosting as well
Challenges for Libraries

- Staffing/prioritization
- Cross-institutional collaboration
- Scaling of service models
  - Beyond libraries/archives
  - Across disciplines
- Growing fuzziness around collections vs. research data
  - Multi-modal imaging
  - “Collections as data”
  - Researcher output, annotations
- Finding common goals, interests, and incentives across collections, researchers, library, cyberinfrastructure providers
Challenges for Natural History Collections

- Staffing/prioritization
- Cost of supporting cyberinfrastructure/storage
  - Centralized campus CI / local department resources
- Software development
  - Cost, turnover, specialization, etc.
- Metadata and data management training!
- Recognition of a need to collaborate and minimize reinvention of the wheel
Solutions for Natural History Collections, Libraries, and Institutional Resources

- Commitments to long-term synergy and collaboration
- Communication and recognition of shared priorities and needs
- Strategies for sustainable development
- Community-led movement for metadata schema alignment and adoption
- Reliance on national resources
- Well-documented practices for integration with communities of practice
- Workshops, like this, that engage diverse stakeholders from both large and small institutions
Thank you!

Gary Motz
garymotz@indiana.edu

Jon Dunn
jwd@indiana.edu