The Georgia Institute of Technology Example

Data Curation Program Development in U.S. Universities

Tyler O. Walters,
Associate Director, Technology & Resource Services,
Library & Information Center,
Georgia Institute of Technology

University of British Columbia, Vancouver, March 2010

A Story of DC Program Development…

- **Incremental Progress…**
- Individuals reaching out to faculty/labs/research centers has made the difference
- <u>Top-Down measures</u>
  - NIH, NSF policies
- <u>Bottom-Up entrepreneurs</u>
  - Johns Hopkins, University of California-San Diego, University of Illinois at Urbana-Champaign, Michigan, MIT, others
  - Progress without benefit of national mandates and high-level university policies to build programs

A Story of DC Program Development…

- Georgia Tech is typical of US research universities:
  - devoid of top-level mandates, incentives
  - rich with independent, "bottom-up" action
  - $525 million in research. #10 in US universities w/o med. school
- Address:
  - program antecedents & context
  - library's related inter-institutional partnerships (dig. pres.)
  - Library organizational developments
  - Partnerships with research communities on campus
  - Model for DC program development

Data Curation Antecedents & Context

- SMARTech, GT's IR
  - IRs have become the "catch-all" for a diversity of scholarly and research output at universities
- New lifecycle management opportunities:
  - Digital scientific research data
  - Libraries participate in reports:
    - *i.e. "To Stand the Test of Time: Long-term Stewardship of Data Sets in Science and Engineering" (ARL/NSF, 2006)*

Related Inter-Institutional Partnerships

Digital libraries/archives & digital preservation:

- NDIIPP / MetaArchive Cooperative / LOCKSS
- Chronopolis and Data-PASS

- GT: example of how libraries leverage existing activities to generate data curation knowledge, skills, and cyberinfrastructures

NSF Office of Cyberinfrastructure

- NSF DataNet:
  - DataONE and the Data Conservancy
    - Many libraries, library-related organizations involved
- GT: partner on a proposal under consideration
  - GT attracts partners, begins resource allocation for data curation

GT Milestones

Summer 2008: Library Data Curation Work Group

- Library technologists, digital initiatives librarians, subject librarians in: biosciences; physics, earth & atmospheric sciences; civil/ environmental engineering; chemical/biomolecular engineering, polymer, fiber, & textile engineering; materials science; chemistry.

- Devised interview questions about researchers' data practices & needs, began interview process
- Collected subjective data about researchers' data retention, sharing needs & storage practices

***Findings*:**

- Preserve final datasets. Research community may question findings. May need to re-examine datasets. Need data access in support of published papers

GT Milestones *continued*

Research Data Project Librarian

- Gained from Digital Library Development unit
- Flattened organization, required more efficiency in IR initiatives, e-publishing, digital collections project management & technology expertise

- RDPL leads, coordinates research data project group
- Reaches out, builds relationships with GT faculty. Assesses & learns about faculty data practices
- Reviewing Data Audit Framework for use in further domain interviewing

Technology Planning/Development

Library's Digital Development Team

- Comprised of network, storage, programming, & digital library/ archives specialists

- Beginning to assess & implement a technology infrastructure for data curation

- Core library systems for data curation include:
  - Sun StorageTek 2540 disk array
  - SL 500 Tape Library
  - managed by Sun's SAM server software & ZFS
  - Current storage capacity of these two units combined is 529 TB

Partnering with Research Communities

Neuroscientists at GT/GSU Center for Advanced Brain Imaging (CABI)

- CABI = 27 faculty + 35-40 researchers
- Each faculty's lab holds min. 4-5 TB data , Center total ca. 120 TB
- Grad students responsible for data & its retrieval
- No domain-wide ontology, thesaurus, or metadata scheme, despite past national-level attempts at creating a national data center
- Neuroscience may be a leading example of a scientific domain that will curate its data in a diffused fashion; hence, university-level solutions for data curation will become significant.

GT Center for Advanced Brain Imaging (CABI)

Functional Magnetic Resonance Imaging (fMRI) to conduct brain studies

Data Formats:

- Digital Imaging & Communications in Medicine (DICOM)
- Neuroimaging Informatics Technology Initiative (NIfTI)
- Electroencephalographic (EEG) data as well, stored as numeric data in spreadsheets

CABI *continued*

- Both raw & "finished" datasets need preservation to verify research & reproduce past studies
- Leading data management problems:
  - long-term storage & preservation
  - identification & retrieval of research data sets
- Concerned about retrieval & use of datasets from past studies to verify former research
- Presentation of data in published journals:
  - Publishers rules vary greatly
  - limit how many tables & graphs can be shown; therefore, some researchers publish URLs to data that reside elsewhere (repositories)
  - Desire linking its e-publishing activities with its digital research data, however, struggles with how best to enact the primary-secondary source relationship

GT Department of Biomedical Engineering

- Five bioscientists – disparate research projects
- Fields of study:
  - genetic expressions found in social insects
  - motor functions of invertebrate animals
  - bacterial gene mapping
  - computational modeling of intracellular metabolic & signaling pathways
  - studying a variety of biological structures
- Scientific methods producing the digital research data:
  - genetic sequencing
  - fluorescent imagery in fluid mechanics studies
  - electron microscopy & crystallography
  - mass spectronomy
  - DNA microarray studies

GT Bioscientists *continued*

- Data formats:
  - .csfasta, .qual, .BMP, .RAW, CCP4, MRC, .sfd, JPEG, & a number of spreadsheet file formats. 65-80 TB total from the five faculty.
- Data storage practices range from:
  - maintaining data on hard drives disconnected from CPU
  - local server data storage
  - outside IT storage firm, manages tens of TBs of data
- Repository services: e.g. NCBI, EM Data Bank (cryo-microscopy)
  - services cannot accommodate every data format used, nor manage all data these bioscientists generate

GT Bioscientists *continued*

Findings:

- Desire to search their data more effectively
- Share online with research team, with colleagues at other institutions once initial studies were documented & results published
- Current state of practice is simple approaches to storage
- Storage has been significant challenge; not had opportunity to consider more robust data discovery & retrieval tools such as domain-based ontological terms, metadata schemas, or search interfaces. No staff to implement these.
- Data preservation needs recognized, for final datasets used in articulating the published research findings
- Problems of ensuring availability of final datasets. Recognize need to verify earlier research results & connect published findings to supporting data

Neuroimaging and GT Project Team

MIT: Martinos Imaging Center / GT: Ctr. for Advanced Brain Imaging

Synergies in data curation to advance science through data sharing, publishing, and preservation

The GT Team:

- Library: data curator, storage/network manager, programmer, repository librarian, psychology librarian, AD for technology (Walters)
- OIT: director of infrastructure and architecture (Chen)
- CABI: Prof. Corballis, graduate student
- Advisors: Prof. David Bader, Exec. Director, High-Performance Computing
  Dr. Bill Underwood (GTRI), digital archives research
  Prof. Leo Mark (Computing), atmospheric science data curation

Stages of Preservation Implementation

- Stage 3: Select collection(s) for ingest. Document:
  - **Content**
  - **Formats**
  - **Metadata requirements** (fixity, provenance, context, reference)
    - Key people to involve:
      - Administrator (collection identification)
      - Archivist/curator (collection identification)
      - Metadata librarian (research metadata requirements)

From Skinner / Walters: Implementing a Preservation Strategy

Data Curation Strategy

- Data deposition/acquisition/ingest
  - SIPs prepared by CABI graduate student / GT Research Data Librarian
- Data curation and metadata management
  - Collaborate on metadata guidelines, policies on access, retention, formats , etc.
- Data protection (policies, tools, procedures)
  - Chen (OIT), Baines (OIT Info. Security), Helms and Walters (Library), Corballis (CABI)
- Data discovery, access, use, dissemination
  - Collaborate on portal design, descriptive metadata for expert and citizen use
- Data interoperability, standards, integration
  - Identify, develop, and use in-common ontologies, semantic frameworks, data transfer and integration protocols between partners
- Data evaluation, analysis, and visualization
  - Build technical framework to incorporate researcher's tools

Modeling for DC Program Development

- Lack models for data curation program development to guide through pre-program activities, program initiation, & growth

  Basic Model Components:

  1. **Assess faculty data practices**
  2. **Design & build initial technology platforms**
  3. **Create & pilot service models**
  4. **Develop data curation policies**

- Yield common understandings for developing programs at individual universities & lay groundwork for inter-institutional collaborations

#1: Assess faculty data practices

- Informs all other curation program components & is fundamental to the creation of a data curation program
- Assessment Tools:
  - Data Audit Framework
  - Aspects of Risk Assessment:
    - Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
    - Trustworthy Repositories Audit & Certification (TRAC)
  - Faculty interviews / surveys / profiles as done by MIT, Purdue University, University of Illinois at Urbana-Champaign
- All are methods that help us understand how researchers create, store, manage, use, & share data in their research
- Assessment data influences technology, service, policy design

#2: Design & build initial technology platforms

- Understand data practices, aspirations. Then select technologies
- Digital Curation models, e.g.:
  - Open Archives Information System (OAIS)
  - Digital Curation Centre (DCC) Lifecycle Model
- Steps in the DCC lifecycle process, e.g.:
  - "select & appraise," "ingest," "describe," "store," "access," "share," "reuse," "preserve," & "transform"
  - may be core to any data curation system & will require software designed to support & execute them effectively
- Determine which lifecycle steps are most critical to an institution's scientists, then assess & test certain curation software components
- Georgia Tech is utilizing the information its gathering on faculty data practices to build a data repository addressing these lifecycle steps

#3: Create & pilot service models

GT- initial view from faculty needs assessments:

- Storage
- Receipt of & augmentation of metadata
- Search function to locate existing datasets
- Preserve datasets identified as critical to verifying research
- Piloting:
  - Islandora (Drupal / Fedora)
  - MIT DataSpace curation tools *(under development)*
  - Designing business & service models for the DC service
  - Storage service models *(library / OIT / cloud)*

#4 Policy development

Further develop initial DC policies as experience is gained from the previously program components

- A critical area: selection of datasets for preservation
  - MACRO: which research projects are the most significant & should have its data preserved?
  - MICRO: which datasets from a given project are most significant & require long-term retention?
- Other:
  - minimally required metadata & acceptable data formats
  - use & reuse parameters, & access regulations
  - adherence to gov't policies on data access & mgmt., e.g. NSF, NIH

General Conclusions

- Gathering resources for developing data curation programs at the institutional level is proving to be a challenge
- Program development is incremental & characterized by the reallocation of existing library resources
- Grant funds to initiate programs is very significant & needed
- Identify researchers to explore data curation approaches is critical
- Model-building shapes programs to meet university needs & prepares it to collaborate & leverage inter-institutional efforts

THANK YOU!

Tyler Walters
Georgia Institute of Technology Libraries
tyler@gatech.edu

Skype / ooVoo / Google Talk: TyWalters1

Article by the same name as this presentation:

http://www.ijdc.net/index.php/ijdc/article/view/136