WRANGLING DIGITAL CHAOS: CHARACTERIZATION & INGEST

Dr. Helen R. Tibbo
School of Information and Library Science
University of North Carolina at Chapel Hill
tibbo@ils.unc.edu
Some Streams of Activity Related to Digital Curation

- Care & properties of physical media
- Digital forensics & data recovery
- Hardware & software interoperability
- Institutional & manuscripts archives
- Social science data archives
- Physical science data archives
- Librarianship (esp. digital libraries)
- Art & museum curation
- Medical information (e.g. imaging, informatics, health records)
- Lawyers & auditors
- Computer-supported cooperative work (CSCW)
- Management of information systems (MIS)
- Research on documents & document-centric computing
- Standards development
Some Examples of Related Professional Organizations/Associations
• American Association of Law Libraries (AALL)
• American Association for History & Computing (AAHC)
• American Association for State & Local History (AASLH)
• American Council of Learned Societies (ACLS)
• American Library Association (ALA)
• American Health Information Management Association (AHIMA)
• American Medical Informatics Association (AMIA)
• American Society for Information Science & Technology (ASIST)
• Association of American Publishers (AAP)
• Association of Canadian Archivists (ACA)
• Association of College & Research Libraries (ACRL)
• Association for Computers & the Humanities (ACH)
• Association for Computing Machinery (ACM)
• Association for Information & Image Management (AIIM)
• Association for Information Management Professionals (ARMA)
• Association for Library & Information Science Education (ALIS)
• Association of Moving Image Archivists (AMIA)
• Association of Public Data Users (APDU)
• Association of Research Libraries (ARL)
• The Association for Work Process Improvement (TAWPI)
• Audio Engineering Society (AES)
• Australian Society of Archivists (ASA)
• Business Forms Management Association (BFMA)
• CENDI Meeting (interagency working group of senior Scientific & Technical Information managers from nine U.S. federal agencies)
• Coalition for Networked Information (CNI)
• Collaborative Electronic Notebook Association (CENSA)
• Committee on Data for Science & Technology (CODATA)
• Corporation for National Research Initiatives (CNRI)
• Council on Library & Information Resources (CLIR)
• Computer Professionals for Social Responsibility (CPSR)
• Dance Heritage Coalition (DHC)
• Data Management Association (DAMA)
• Depository Library Council (DLC)
• Digital Curation Centre (DCC) Associates Network
• Digital Library Federation (DLF)
• Digital Preservation Coalition (DPC)
• DLM Forum
• European Commission on Preservation & Access (ECPA)
• Information Resources Management Association (IRMA)
• Institute of Certified Records Managers (ICRM)
• Institute of Electrical and Electronics Engineers (IEEE) – including IEEE Computer Society
• International Association for Social Science Information Service & Technology (IASSIST)
• International Committee on Technical Interchange for Space Mission Operations & Ground Data Systems (SpaceOps)
• International Council on Archives (ICA)
• International Federation for Information Processing (IFIP)
• International Internet Preservation Consortium (IIPC)
• Medical Library Association (MLA)
• Midwest Archives Conference (MAC)
• Museum Computer Network (MCN)
• National Association of State Chief Information Officers (NASCIO)
• National Association of Government Archivists & Records Administrators (NAGARA)
• Nuclear Information & Records Management Association (NIRMA)
• Online Computer Library Center (OCLC) – including Research Libraries Group (RLG)
• Open Grid Forum (OGF)
• Society of American Archivists (SAA)
• Society for Information Management (SIM)
Note: Attending all of their conferences could violate current travel budget restrictions!
KEEP CALM AND CARRY ON
Approaches to Wrangling Chaos

- You are not alone! (or even first at this point)
- Everything looks overwhelming until you have a plan
- Organizing Principles & Models
- Community Documents & Standards
- Institutional Requirements & Context
- Institutional Strategies & Policies
Building Digital Curation Programs

- Are you in the:
  - Planning stage?
  - Planning and pilot stage?
  - Implementation stage?
Organizing Principles

DCC Digital Curation Lifecycle Model

http://www.dcc.ac.uk/resources/curation-lifecycle-model
OAIS

Chaos Happens

Chaos Happens
Producer-Archive Interface Methodology Abstract Standard

Phase objectives
- Define the information to be archived
- Develop agreement (data to be delivered, complementary elements, schedule)
- Actual transfer of the objects
- Validate the transferred objects

Phases:
- Preliminary Phase
  - Preliminary Agreement
- Formal Definition Phase
  - Dictionary Formal model Submission Agreement
- Transfer Phase
  - Transferred object files
  - Anomalies
- Validation Phase
  - Validation agreement

AIP creation

http://public.ccsds.org/publications/archive/651x0m1.pdf
Submission Agreement: The agreement reached between an OAIS and the Producer that specifies a data model for the Data Submission Session. This data model identifies format/contents and the logical constructs used by the Producer and how they are represented on each media delivery or in a telecommunication session.

In the framework of this abstract methodology, the Submission Agreement will also deal with other aspects such as validation, change management and schedule.

-CCSDS, 2004
Submission agreements can cover a single ingest project or serial ingest project. A **standing** submission agreement covers a serial ingest project while a **regular** submission agreement covers a single ingest. A standing submission agreement would cover an agreement between a producer and an archive to transfer records on a recurring basis. Normally, records covered by a standing agreement are going to be relatively homogeneous records that the producer creates on a regular basis.

- TAPER, 2009
Characterize

- “To describe or delineate the character or peculiar qualities of (a person or thing)” (Oxford English Dictionary, Second Edition, 1989)
Characterizing Digital Objects

- Assertions about specific aspects of the content or character of a digital object
- Can be:
  - Very specific and machine-readable (e.g. file = 25 bytes; character-encoding = UTF-8, image is 48 pixels wide)
  - Relatively vague and human-readable (e.g. “This is a video from YouTube”; “I was able to view this on a Macintosh”)
  - Or somewhere in between
Characterization

- Characterizations = Surrogate Representations or Metadata
- What to capture in characterization?
  - International standards and frameworks
  - Future use
  - Local needs and priorities
- Stems from policies and resources
- Characterization model is formalized in submission agreements and other policies
Repository Policies

- Mission & Governance
- Service Policy
- Collection Policy
- Access Policy
- Privacy policy
- Non-Exclusive Deposit License
- Format Support Policy
- Withdrawal Policy
- Metadata Policy
- Contributor Policies
- Preservation Policy
Considerations

- **Organizational**: e.g., Statutory body; Institutional mandates
- **IPR**: e.g., Copyright; Publisher agreements
- **Privacy**: e.g., Personal identifiers, Culturally sensitive information
- **Services**: e.g., Authentication and Authorization
- **Language**: e.g., Understandability; Consistency
- **Workflow**: e.g., Transfer scenarios; PAIMAS stages
Policy Planning

- Establish a policy planning group
- Identify regulatory landscape
- Devise policy framework
- Draft, vet and revise
- Make available
- Establish regular reviews
Ten Principles

- In January 2007 representatives of four preservation organizations convened at the Center for Research Libraries in Chicago to seek consensus on core criteria for digital preservation repositories, to guide further international efforts on auditing and certifying repositories. The organizations were:
  - The Digital Curation Center (U.K)
  - DigitalPreservationEurope
  - NESTOR (Germany)
  - Center for Research Libraries (North America)
Ten Principles for Digital Preservation Repositories

- The repository commits to continuing maintenance of digital objects for identified community/communities.
- Demonstrates organizational fitness (including financial, staffing, and processes) to fulfill its commitment.
- Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.
- Has an effective and efficient policy framework.
- Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
Ten Principles for Digital Preservation Repositories

- Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.
- Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.
- Fulfills requisite dissemination requirements.
- Has a strategic program for preservation planning and action.
- Has technical infrastructure adequate to continuing maintenance and security of its digital objects.
Know Thyself & Thy Users!

- “The key premise underlying the core requirements is that for repositories of all types and sizes preservation activities must be scaled to the needs and means of the defined community or set of communities.”
  - CRL, 2007
Thank You!

Helen R. Tibbo
School of Information and Library Science
University of North Carolina at Chapel Hill

tibbo@ils.unc.edu

Thanks also to IMLS for funding
DigCCurr I (RE-05-06-0044) and
DigCCurr II (RE-05-08-0060-08)