Digital Preservation Network (DPN): An ARL Library Perspective

Brenda L. Johnson
Ruth Lilly Dean of Libraries
Indiana University

May 2, 2012

Association of Research Libraries Membership Meeting
Outline

• Why DPN?
  • Library models and initiatives
• Why DPN for your institution?
  • Key considerations
• Why DPN for IU?
  • An institution’s perspective
WHY DPN?
What James said...

Existing Library Digital Preservation Models

- HathiTrust Digital Library
  - Mirrors at University of Michigan and Indiana University
- Chronopolis
  - San Diego Supercomputer Center, National Center for Atmospheric Research and University of Maryland Institute for Advanced Computer Studies
- Publisher journal content
  - CLOCKSS
  - PORTICO
Data Curation Initiatives & Libraries

- CLIR/DLF Data Curation Fellowships
- ARL E-Science Institute
- UK Digital Curation Centre (DCC)
- Library of Congress National Digital Stewardship Alliance (NDSA)
- NSF DataNet Cooperative Agreements
DPN and Your Institution

• What types of born-digital data are your institution creating?
  • Lecture casts
  • Music and arts performances
  • Scientific and social science data sets
  • Specialty scientific instrument outputs (sequencers, telescopes, sensor networks)

• How much of that data needs long-term digital preservation?
  • Uniqueness
  • Can it be regenerated and at what cost?
  • Can it be re-used for new knowledge creation?
DPN and Your Institution

- What is your local capacity to hold that type of data?
  - Are your purchases measured in terabytes or petabytes?
  - Are you prepared for exabytes and zettabytes?
- Is your local data utility available across high-speed networks that enable mass-data transfer?
  - *New Internet2 100GB network*
DPN and Your Institution

• How could your local data curation and preservation resources be used more effectively if you had a national resource for digital preservation like DPN?
  • Capture of arts, humanities, social-sciences, and science content closer to the source or even directly from the experiment and with more original provenance
  • Could libraries target digital collection specific content in a federated fashion ala shared print repositories i.e. WEST, CIC-SPR, ASERL
DPN and Indiana University

• Why is it important for IU to be a member of DPN?
• What kinds of data does IU produce and what does IU want to preserve?
• What is the local infrastructure for storage at IU?
• Why does IU need DPN to serve as a long-term preservation utility?
Why DPN for IU?

- **Big Data is Exploding at IU**
  - Digital Library collections (40TB)
  - Media collections (10 PB)
  - WIYN Telescope (2 GB per image)
  - HTRC Derived Data Collections (500 TB)
  - Sustainability Data from SEAD (100 TB)
  - National Center for Genome Analysis Support (petascale data)
  - Muslim Voices Project (Center for the Study of Global Change)

- **Local Infrastructure for Storage at IU**
  - Research file system (48TB)
  - Scholarly Data Archive (15 PB)/Institutional Repository
    - 2 copies (Bloomington Data Center/Indianapolis Data Center)
IU Bloomington is home to at least 3 million sound and moving image recordings, photos, documents and artifacts. Well over half a million of these special holdings are part of audio, video, and film collections and a large number of them are one of a kind. These invaluable cultural and historical gems, and many more may soon be lost. Forever.

Meeting the Challenge of Media Preservation: Strategies and Solutions

September 2011
Why DPN for IU?

• **Big Data is Exploding at IU**
  - Digital Library collections (40TB)
  - Media collections (10 PB)
  - WIYN Telescope (2 GB per image)
  - HTRC Derived Data Collections (500 TB)
  - Sustainability Data from SEAD (100 TB)
  - National Center for Genome Analysis Support (petascale data)
  - Muslim Voices Project (Center for the Study of Global Change)

• **Local Infrastructure for Storage at IU**
  - Research file system (48TB)
  - Scholarly Data Archive (15 PB)/Institutional Repository
    - 2 copies (Bloomington Data Center/Indianapolis Data Center)
May 2, 2012

ARL Meeting Spring 2012: DPN: A Librarian’s Perspective
DPN and Indiana University

• Long-term network centric data preservation such as DPN enables local data preservation resources for:
  • Web-crawling
    – Local Web content (state politics or cultural events)
    – Short-term Web content (campus events or culture)
  • Data capture at source with provenance
  • Regional or collection specific content
  • Institutional data publishing
    – Multimedia content for dissertations
    – Multimedia content for published articles
Unfortunately, researchers are not always very good at the long-term preservation and curation of data. The question then becomes, to whom should we entrust our data if we want to be able to access it in another forty, fifty, or one hundred years?... I believe that the best option for the long-term preservation and curation of digital data—and by “long-term” I mean not just decades but centuries—will be universities.

Michael A. McRobbie, Indiana University President

EDUCAUSE Review, vol. 45, no. 3 (May/June 2010): 8-9
Questions?

Digital Preservation Network Website

http://d-p-n.org/