The New Digital Landscape of Government Information: Implications for Research Libraries

Charley Barth
Office of the Federal Register
National Archives and Records Administration
The Digital Landscape

- 1 byte = one character
  1 kilobyte = one page of text
  1 megabyte = four 200-page books
  1 gigabyte = 4,473 books
  1 terabyte = 4,581,298 books

- NARA has used over 500 terabytes of storage thus far for our electronic records holdings. So that would mean about 2,290,649,000 books of 200 pages.

- For comparison, we have 12 billion sheets of paper across all our facilities

- Exponentially growing electronic records requirements

- Classified materials and Special Access restrictions only complicate an already complicated environment
Exponential Growth

Growth in Volume of Electronic Records
Transferred to NARA since 1970

- 7 MB in 1970
- 8 GB in 1980
- 15 GB in 1990
- 256 GB in 2000
- 83 TB in 2010
- 564 TB in 2020
Paper vs. Digital snapshot

- 12,000,000,000 Textual Records
- 225,771,902 Digital Copies Online
- 220,000,000 Partner Digital Copies
  - 3,700,000 1940 Census Digital Copies
  - 2,071,902 Digital Copies in OPA
What about email?

- As for emails in the Presidential Library system:
  - 2.5 Million emails from Reagan and GHW Bush administration
  - 20 Million emails from Clinton administration
  - 210 Million emails from GW Bush administration
  - Over 1 billion emails/tweets from the Obama administration
Why we do what we do

• Legislated requirements for archiving electronic government information
  – Presidential Records Act (44 U.S.C. Chapter 22)
  – Federal Records Act (44 U.S.C. Chapter 31)
  – Presidential memorandum relating to electronic records
    • THIS IS A MAJOR CHANGE for how Federal Agencies conduct digital management of their records
    • Managing Government Records Directive (M-12-18)
Timeline for Memorandum

2019
- Federal agencies manage all permanent records electronically

2017
- NARA completes overhaul of GRS

2016
- Agencies manage all email records in electronic format
- SAOs ensure records Schedules submitted to NARA

2015
- NARA issues improved Request for Records Disposition Authority process
Chief Records Officer

• In 2011, NARA appointed the first Chief Records Officer of the United States, Paul Wester
• 2 weeks ago, the CRO released the Automated electronic records mgmt plan
  – **Improve scheduling approaches** (big bucket and less permanent) and **training** of all Federal employees to better capture, maintain, and manage electronic Federal records.
  – Long term vision of bringing the Managing Government Records Directive, the **Open Data Policy and Plan**, and the Strategic Plan of the National Archives and Records Administration together. **Holistic approach** will address business analysis, eDiscovery, FOIA, security, privacy and other information management disciplines at the same time.
  – NARA will continue to work with its stakeholders to identify milestones and tasks that will move Federal records management toward digital government, including **increased automation, reduced burden on end users, and more consistent and affordable compliance** with recordkeeping requirements.
NARA's custodial unit for born-digital electronic records

- NARA's Electronic Records Division, Research Services (17 permanent staff), accessions, processes, arranges for preservation, describes, and provides access to the born-digital federal records scheduled for permanent retention in the National Archives.
- Current holdings consist of 932 record series from over 100 federal agencies; over 750 million unique files and over 320 Terabytes.
- Online access to select series is provided via direct download of full files and associated documentation via NARA's Online Public Access (OPA) catalog; record-level retrieval for 64 series and over 112 million records is available through the Access to Archival Databases (AAD) utility.
- One recent accession of note are the files of the EDGAR system, 1992-1996, from the Securities and Exchange Commission. The accession consists of over 700,000 XML-format files. The files are in the process of being processed and preserved in NARA's Electronic Records Archives (ERA).
Implications for Electronic Records

• Location independent. Moving to the Cloud

• Very high level requirements for security
  – Classified and unclassified records
  – Network and bandwidth issues are very real

• Libraries getting on-board!
  – GPO just announced the first ever, 100% digital, Federal Depository Library Program (FDLP) in North Dakota (Sitting Bull College)
  – Other all-digital, bookless libraries opening up across the nation (Biblio-Tech in Texas)

• Transfer of electronic records
  – Main area of effort is in Washington DC and Agency headquarters
  – However, NARA worked with DOD and Joint Staff to safeguard over 200 TB of digital war records in Iraq and Afghanistan
How are people finding out about our electronic holdings?

- Wikipedia: 1,274,493,206*
- Facebook: 39,491,534
- Archives.gov: 29,351,833
- Flickr: 11,258,709
- YouTube: 872,926
- OPA: 543,067
Biggest challenges to preserving all of our Records

• Paper version well preserved since late 1700’s (thru today)
  – Paper is well cared for in compliant, archival centers (NARA standard 36 CFR Part 1228, Subpart K)
  – Theft and Fire are biggest challenge

• From 1985 to 1994, we received a bunch of microfiche
  – Silver Halide process can last 500 years
  – Virtually impossible to mutilate

• From 1995 to current, the electronic version has become the official record copy in most cases
  – Storage (expense and trusted providers/cloud)
  – Compromised content (via internal/external hacking)
  – Best formats 25, 50, 100 years from now? (risk of migrating data from format to format)
    • How do you preserve 3D Holograms messages for example?
  – Hardware failure (IRS emails)
  – Human error (accidental or purposely deleting electronic content)
Conclusions

- Multiple examples of electronic records all needing different methods of maintenance and access
- Exponential growth across all agencies
- We are tracking over 1600 separate agency electronic series, all with divergent needs
- Next generation technologies/formats needed now.
  - Open source solutions like Alfresco? IBM content manager and SharePoint are king right now but use proprietary software code. Full 100% server hosted auto categorization next?
  - PDF/A standard is king right now (device independent/locked down)
    - Audio and video content is forbidden
  - We have to ask ourselves what is more important…the data or the format?
- Public/Private partnerships critical to effort –
  - We need your help!
  - OFR partnering with local law libraries, NARA partnering with Ancestry.com and Fold3
  - You Tube (video), Instagram (photos), Amazon (compliant cloud storage) to the rescue?