Migrating government information from CD-ROMs: scaling a pilot project

Federal Depository Library Conference, Fall 2006

Julie Linden, Yale University Library

Gretchen Gano, New York University Library

Outline

- The CD problem
- Details of the Yale library pilot project
- Towards a collaborative model
- Issues

Introduction: the CD problem

- Analyzed and summarized by John Hernandez and Tom Byrnes, Spring 2004 DLC meeting:
  - Degradation of physical media; obsolescence of software, file formats; proprietary software
  - Lack of backwards compatibility (Census GO)
- Is any given CD in *your* collection still usable?

Review of projects tackling the issue

- CD-ROM Analysis Project
- UCSD GPO Data Migration Project (5 ¼" floppy disks)
- CIC Floppy Disk Project (3 ½" floppy disks)

Yale Library pilot project: overview

- Approach
- Selection process
- Analysis and migration
- Metadata
- Cost and time analysis

Criterion	Rationale
Information not available elsewhere in digital format	Information is therefore more at risk; migrating it preserves it and makes it potentially more accessible
Information accessed with software that is proprietary, uncommon, or obsolete	Migration frees the data from software dependency and increases its accessibility
Information is free of copyright restrictions	Copyright restrictions may prohibit migration of data or placing data on a server
CDs have circulated	Indicates that the information has been used by our patrons
CDs come from all GDIC collections and a variety of agencies	Provides a broad, diverse sample, potentially exposing a range of migration issues
Information will be in a variety of file formats	Provides a broad, diverse sample, potentially exposing a range of migration issues

Analysis and migration

Student tasks:

1. Transfer files from CD-ROM to server.
2. Analyze the files on CD-ROM and document the analysis. (Example: Marriage & Divorce)
3. Migrate specific files from one format to another (e.g. SETS to ASCII; Excel to ASCII; Microsoft Word to plain text).
4. Document errors or problems encountered.
5. Document time spent on tasks.

Metadata

- We chose to create MARC21XML, MODS, DDI (Data Documentation Initiative: standard for social science numeric data), Dublin Core, and PREMIS metadata.

Cost and time analysis

- Analysis and migration:
  - average of 1.75 hours/CD
  - at $11/hour (student wage), that is $19.25/CD.
- Potential for scripting or programming batch migration of files.
- What efficiencies might be gained by migrating CDs from a single agency?

Conclusions from the pilot

- Selection: A “hunt and peck” approach is time-consuming - each disk may require a different workflow
- Migration: Access to programming expertise is necessary for normalizing some formats
- Metadata: Settle on metadata standards, required elements, and content rules as early as possible. If it can’t be automated, do you need it?
- Preservation: for data, ASCII files are preferable for long-term preservation over something that emulates the original CD application/functionality

Toward a collaborative model

- Data sources
  - Suggestion to approach agencies for original data sources
  - Investigate NARA holdings
- Organizing institution
- Decentralized processing
- Engage depository partners with appropriate infrastructure
- Open repository model

Facets of collaboration

- Inventory of CDs
- Upload original files
- Cataloging / metadata production
- Preservation
- Normalization
- Emulation
- Quality assurance
- Copyright clearance

Issues

- Beginning with data series, but how to migrate universe of CD content?
- Data formats and original documentation
- Accessibility to non-specialists
- Certifying authenticity
- Copyright
- Of course, funding

Contact

Julie Linden, Yale University Library
julie.linden@yale.edu

Gretchen Gano, New York University Library
gretchen.gano@nyu.edu