Legacy of Our Future History: 100-year Digital Archiving

H. M. Gladney
IBM Almaden Research Center
San Jose, California 95070

gladney@almaden.ibm.com

Most **new** content will originate in digital form.

Some content will be fundamentally digital.

What's needed to preserve the bitstreams?

What will be needed to interpret saved bitstreams?

http://www.almaden.ibm.com/u/gladney/100Year.pdf

What's the Problem?

- **Digital infrastructure is in its infancy[1] -- compare that for paper**
  - Infrastructure for handling paper has been refined over 3000 years
  - E.g., largest civilian employer in the U.S. is the U.S. Post Office with 765,000 employees[2]
- **Libraries today have little commitment to selecting and preserving digital documents**
- **Neither the U.S. National Archives and Records Administration[3] nor the Library of Congress fills the gap**
  - LoC American Memories program charitably funded, and funding is exhausted
  - American Memories is retrospective and narrowly selective, not prospective and broad
  - NARA does not archive broadly, and never will
  - NARA mission is to preserve records essential to continued functioning of the government
  - NARA publicly admits to difficulty with digital information (starting with e-mail)[4]

1. Expressed here is the U.S. subset, but the needs and the problem is worldwide.
2. See http://www.usps.gov/history/pfact98.htm
3. NARA is cited for illustration; the situation is similar in other nations.
4. John W. Carlin, Archivist of the United States, on the Report of the Electronic Records Work Group of NARA, 1998 http://www.nara.gov/nara/pressrelease/nr98-148.html

Problem Summary

We are spending immense energies and immense funds to create and disseminate digital information.

These expenditures are not appropriately complemented by efforts to save whatever might be worth saving.

What Are the Challenges?

Administrative: mission and funding

Synergetic: research libraries ambivalent on sharing content

Ability: infrastructure and staff skills

Cultural: libraries need to change

Selection: what's generated is >10^5 larger than 200 years ago

Legal: liability risk -- contributory copyright infringement

Technical:
- **Save the bit streams**
- **Standards for metadata**
- **Interpret streams in 100 years and later**

See *Books, Bricks, & Bytes,* Daedulus, J. Am. Acad. of Arts and Sciences, Fall 1966
LC21: A Digital Strategy for the Library of Congress, Nat'l. Academy Press, 2000.

Comments on Some Challenges

Media longevity	Rejuvenate from time to time
Device lifetime	Rejuvenate from time to time
Suppliers' guarantees	Dubious on 100-year scale
Interpretability of bitstreams	**A technical challenge**[1]
Natural disasters	Replicate in distant sites
Government misbehavior	Replicate in other nations
Destruction in war	Avoid war!! Replicate in "iron mountain" and in other nations

1. But see Raymond A. Lorie, *Long-Term Archiving of Digital Information, IBM Research Report RJ10185, March 2000.*

Computer Science Challenge: Interpreting Bitstreams in 100+ years

- All the other problems are "merely" engineering and operations
- Maintaining obsolete technology and architecture will not work
- Emulating obsolete machines and software will not work
  - Too many machine types, programs, and versions
  - Documentation often missing, usually poor, and seldom "complete"
  - Natural language poor for documentation
  - Formal methods difficult to verify or teach
- Save in terms of simple virtual machines/languages (ref. Turing)
  - Level 1: very simple data, e.g., raster graphics, ASCII text
  - Level 2: static documents, e.g., what SGML can represent today
  - Level 3: behavior of programs, with a universal virtual machine

Inform Students, Colleagues & the Public!

30 minute videotape, PBS broadcast Feb. 1998

T. Sanders, *Into the Future: On the Preservation of Human Knowledge in the Electronic Age,* Commission on Preservation and Access, Washington D.C., 1997.

Examples:

- Books printed on acid paper (approx. 1870 to 1930)
- Sarajevo burning of the national library
- LandSat digital tapes
- Government records of toxic waste sites
- NARA's holdings essential to ensuring govt. accountability (approx. 3,000,000,000 pieces of paper)

J. Garrett et al., *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access report, (August, 1995).

M. Hedstrom and S. Montgomery, *Digital Preservation Needs and Requirements in RLG Member Institutions*, (December 1998).

Jeff Rothenberg, *Ensuring the Longevity of Digital Documents,* Scientific American 272(1), 42-47, (1995).

UPF home page, proposed *Universal Preservation Format (UPF) for the archiving of media assets*.

The Digital Dilemma: Intellectual Property in the Information Age, National Academy Press, Feb. 2000. http://www.dlib.org/dlib/december99/12gladney.html

H.M. Gladney, *Are Intellectual Property Rights a Digital Dilemma?* iMP Magazine, Feb. 2000. See also http://www.almaden.ibm.com/u/gladney/Columbia.pdf

R. Lorie, *Long-Term Archiving of Digital Information,* IBM Research Report RJ 10185, March 2000. See also http://www.almaden.ibm.com/u/gladney/Lorie.pdf

Digital Preservation Archiving and Copyright http://www.almaden.ibm.com/u/gladney/ArchCopy.pdf

This talk: http://www.almaden.ibm.com/u/gladney/100Year.pdf