Preservation of digital records

Concerns, approaches, efforts

Tefko Saracevic, Ph.D.

Tefko Saracevic

creative commons

ToC

- Introductory musings: definitions & problem statement
- Library involvement over time
- Preservation in digital environments
- Technological problems, solutions
- Preservation projects
- Preservation standards
- Concluding musings: issues, questions

Preservation – general definitions

- To preserve:
  "To keep alive, keep from perishing (*arch.*); to keep in existence, keep from decay, make lasting (a material thing, a name, a memory)" (OED, 2nd ed.)

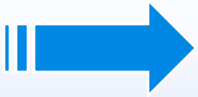
Tefko Saracevic

Problem

Unlike physical materials, which can remain in their current state for decades, even centuries

Content stored in digital formats is easily altered or even lost

While we are still able to read our written heritage from several thousand years ago

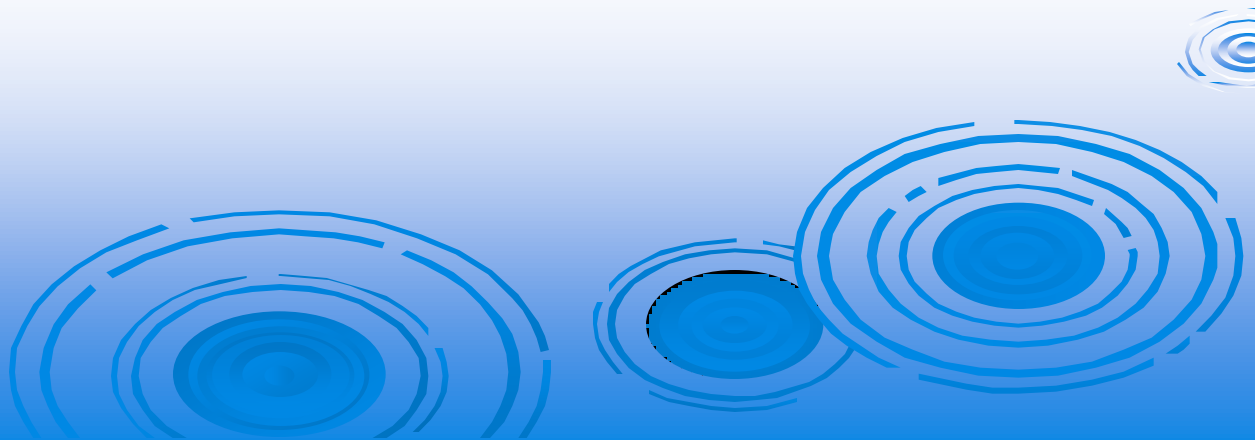


The digital information created merely a decade ago is in serious danger of being lost

Tefko Saracevic

Solutions

- **Preservation**
  - but digital preservation is only half the battle,
- **Permanence**
  - is needed for and closely linked with preservation of digital resources



Tefko Saracevic

Historically in libraries: long time involvement

- **Preservation**: maintaining or restoring access to artifacts, documents and records through the study, diagnosis, treatment and prevention of decay and damage

- **Conservation:** the treatment and repair of individual items to slow decay or restore them to a usable state. "Conservation" is occasionally used interchangeably with "preservation"

Tefko Saracevic

Paper degradation problems

- Paper embrittlement from acid decay; brittle books, newsprint ("slow fire")
  - mass deacidification efforts;
  - promotion of acid-free paper
  - many library projects globally
  - degradation affects any media, not just paper
- Infestation (mold, fungi, bacteria):
  - The use of gamma rays in book conservation
- Book conservation

Another solution: reformatting

- Use of other media to store documents
- Microfilm became popular & widespread
  - many advantages – easy storage
  - many disadvantages
    - no such thing as cuddly microfilm reader
    - searching not possible
- Followed by other media: various tapes, cartridges, optical disks, CD-ROMs …
- Finally, digital media

One of traditional concerns: preparedness

- Natural disasters & libraries – many experiences
- The Flood of the River Arno in Florence, Italy, (1966) damaged or destroyed great many rare books & art
  - led to establishing restoration laboratories in many places
- Recognizing importance of having a disaster preparedness & preservation plans
  - e.g U Delaware Library Disaster Response plan
  - ALA Disaster Preparedness and Recovery

Tefko Saracevic

Libraries not alone

- Sharing preservation concerns with
  - archives
  - historical institutions
  - museums
  - antiquarian practices
  - archeology
- National Center for Preservation Technology & Training

Preservation in digital environments

- Management of digital information over time
- Constant **effort** & **expenditures** to handle rapid technological and organizational advances
  - main stumbling block for preserving digital information beyond a couple of years

Terminology not fixed – here is a list of definitions from Digital Preservation Coalition

Digital preservation goals

- Ensuring the continued access to information and all kinds of records, scientific and cultural heritage existing in digital formats
- Long-term, error-free storage of digital information, with means for retrieval and interpretation, for all the time span that the information is required for

From preservation to permanence

The second half of the battle: permanence

1. **Identifier validity:** the extent to which the given name or identifier will always provide access to same resource
2. **Resource availability:** the extent to which given resource is guaranteed to remain available in electronic form
3. **Content invariability:** the extent to which the content of resource could change

from presentation Digital Archives at NLM

Technological problems: obsolescence

- Format obsolescence
  - when the software required to read the content or data is no longer available or is unable to understand the format of the data
- Requires
  - copying of content onto newer format
  - converting content from one format to another
  - avoiding loss of fidelity

- Technology obsolescence
  - when the hardware required to read the data is no longer available or new hardware or media emerges
- Requires
  - transfer from one technology or media to another
  - e.g. from one kind of tapes, disks to another
  - from microfilm to digital

Technological obsolescence solutions

- Transferring of content or data to newer systems
- Conversion from one format to another or one operating system to another or one programming language to another

- **Emulation** - content is both preserved and presented to readers in the original format
- **Migration** - content is presented in a current format; it may be preserved in a succession of current formats

Emulation problems: BBC Domesday project - 1986

- Attempt to re-do the survey of England done in 1085
- Unfortunately published on 12-inch laser disc in a format that died quickly in the marketplace
- Leeds Univ. & U. Michigan now trying to emulate original hardware

From Michael Lesk

Digital preservation projects

- Digital Preservation Coalition (DPC) (UK)
  - "to secure the preservation of digital resources in the UK and to work with others internationally to secure our global digital memory and knowledge base."
- Sound Direction (U Indiana)
  - "digital preservation & access for global audio heritage "
- Portico
  launched by JSTOR
  - "preserve scholarly literature published in electronic form"
  - many publishers & libraries participating
  - so far over 30 mill. units
- MetaArchive cooperative
  over 50 institutions

"**The Greatest Threat** to digital assets is not fire, flood or theft. It's the assumption that cultural memory organizations have taken the requisite steps to preserve them."

General international resource

- Preserving Access to Digital Information (PADI) (Australia)
  - "gateway to international digital preservation resources"
  - aims "to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access"
  - wide coverage:
    - policies, topics, projects, legal deposits …

Technology & services support: Duplication and sharing

- Protection against loss via multiple copies
- Individual backups are traditional, but do not protect against organizational disappearance
  - if not regularly exercised might not be dependable

- LOCKSS (Lots of Copies Keep Stuff Safe) at Stanford
  - "provides tools and support so libraries can easily and cost-effectively preserve today's web-published materials for tomorrow's readers"
  - free, open source software
  - and a YouTube video

Basic approach – from a quote

"...let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident."

— Thomas Jefferson, February 18, 1791

"Darwin's tortoise" dies, age 176" (June 26, 2006).

The LOCKSS logo is a tortoise; tortoises live a very long time

- LOCKSS Alliance is a library membership organization
  - they have LOCKSS Boxes
  - libraries around the globe participate
- Large number of publishers participate also
  - most open access

- LOCKSS software turns a PC into a digital preservation appliance (a LOCKSS Box)
  - collects newly published content
  - compares it with other LOCKSS Boxes
  - acts as a web proxy or cache
  - provides a web-based administrative interface

LOCKSS: time and consensus (Michael Lesk)

- The LOCKSS project is particularly interesting for two reasons:
  - A. running slowly
  - B. relying on a combination of consensus and reputation

A. By making it fast to find one copy of something, but slow to find all copies, it becomes difficult for a vandal to find and destroy all copies of a file

B. By relying on a weighted polling system in which a site can gain weight only by agreeing with many prior decisions, it is difficult even for an insider to insist on installing bad versions of files.

Authentication of digital resources

- **Authenticity:** The digital material is what it purports to be
  - refers to the trustworthiness of the electronic record as a record
- Confidence in the authenticity of digital materials over time is particularly crucial owing to the ease with which alterations can be made.
- In the case of "born digital" and digitized materials:
  - the fact that whatever is being cited is the same as it was when it was first created unless the accompanying metadata indicates any changes
- A number of mechanisms used to establish the authenticity of digital materials

Archiving

- OCLC's Digital Archive
  software & services for
  - **Web archiving**: Item-by-item
  - **Batch archiving**: For collections
  - available to users in multiple ways - through OCLC's FirstSearch, Connexion, library own OPAC or a Web portal

- Dutch National Library agreement with publishers
  - digital archive for scientific research – some dozen publishers deposit journals

  "which will be made available in perpetuity to the research community including authors, researchers, historians and librarians "

Really BIG project

- National Digital Information Infrastructure and Preservation Program (NDIIPP)
  - "... implementing a national strategy to collect, preserve and make available significant digital content, especially information that is created in digital form only, for current and future generations. "
- Collaborative approach – building a national network – a number of institutions involved
- Congress: $100 mill. + a lot from private sources
  - Description in Wikipedia

NDIIPP …

- Digital files selected for preservation:
  - Geospatial data
  - Web sites
  - Television
  - Social science datasets
  - E-Journals
  - Historical materials
  - Provides suggestions for Personal archiving – your own records

NDIIPP …

- Some of the NDIIPP-NSF Digital Preservation research projects
  - U California Libraries: Tools for Web archiving Digital Preservation Repository - a YouTube video
  - Preserving Digital Public Television (PBS)
  - National Geospatial Digital Archive (NGDA)
  - North Carolina Geospatial Data Archiving Project (NCGDAP)
  - Sustainability of digital formats (LoC)

and then there is

- Cybercemetery
  - Maintained by U North Texas Libraries
  - "**archive of government websites** that have ceased operation (usually websites of defunct government agencies and commissions that have issued a final report)."

And in Europe: EU projects

- Open Planets Foundation

"A community hub for digital preservation

  - "to provide practical solutions and expertise in digital preservation"
  - about 20 members internationally
  - open sources

- CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access & Retrieval (EU project)
  - "... research, implement, and disseminate innovative solutions for digital preservation "
  - a community of members
- DigitalPreservationEurope (DPE)
  - "to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials"

Developing standards for approaches to preservation

- Standards needed to deal with
  - impacts of changing technologies, including support for new media & data formats
  - changing user communities
- Open Archival Information System (OAIS) Reference Model
  - conceptualization of a system that addresses digital preservation - provides a general framework
  - model describes components and services required to develop and maintain archives

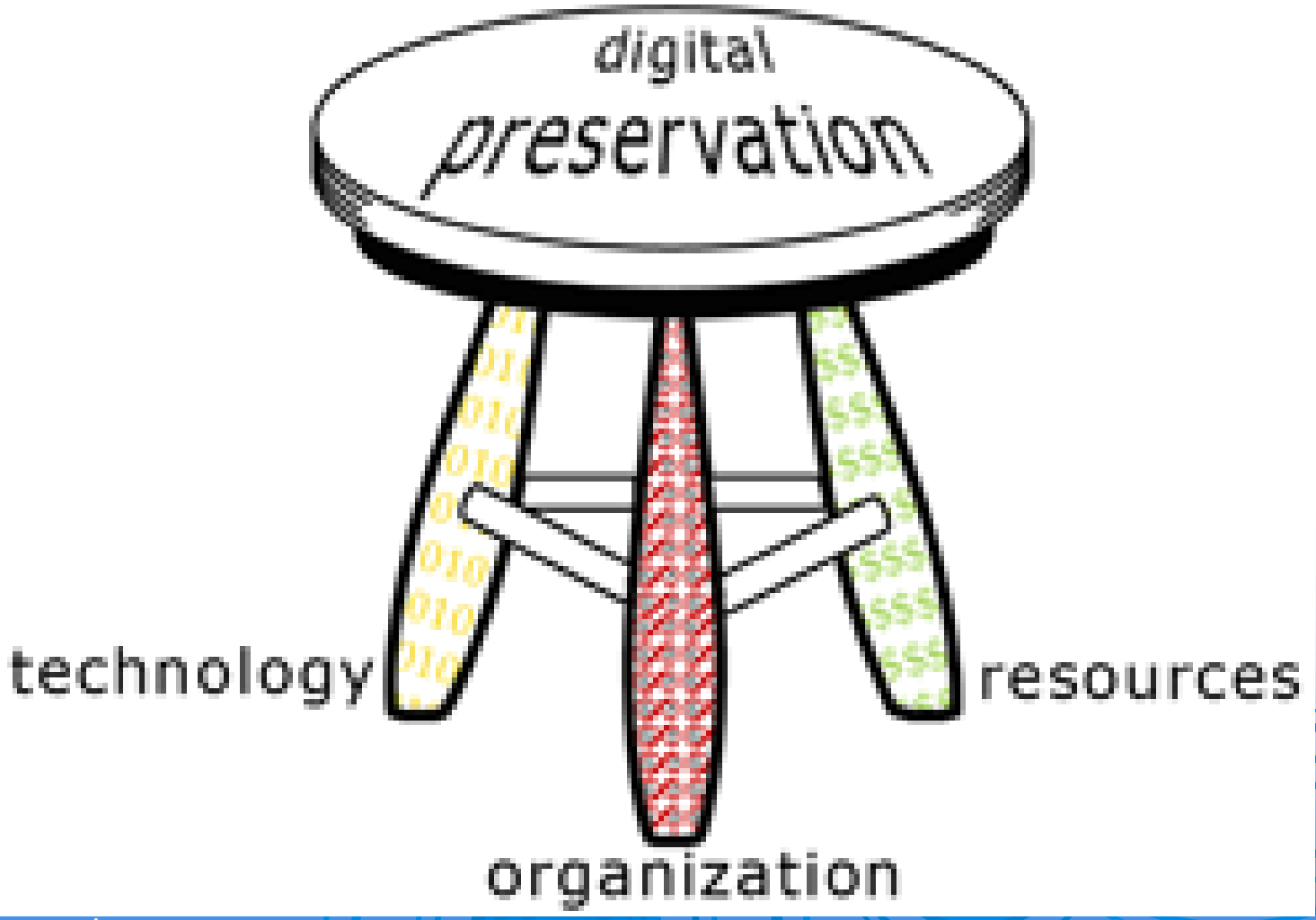
Learn more about preservation

Cornel University Tutorial – moved to MIT Digital Preservation Management – Implementing Short-term Strategies for Long-term Problems

- Well done & exhaustive treatment with quizzes, explanations, and questions.

Summary of preservation requirements

(from Cornell/MIT tutorial)



Summary: preservation requires

- Organizational Infrastructure:
  - *What* are the requirements and parameters for the organization's digital preservation program?
- Technological Infrastructure
  - *How* will the organization meet defined digital preservation requirements?
- Resources Framework - $$$$
  - *What* resources will it take to develop and maintain the organization’s digital preservation program?

Concluding issues – Lesk's questions

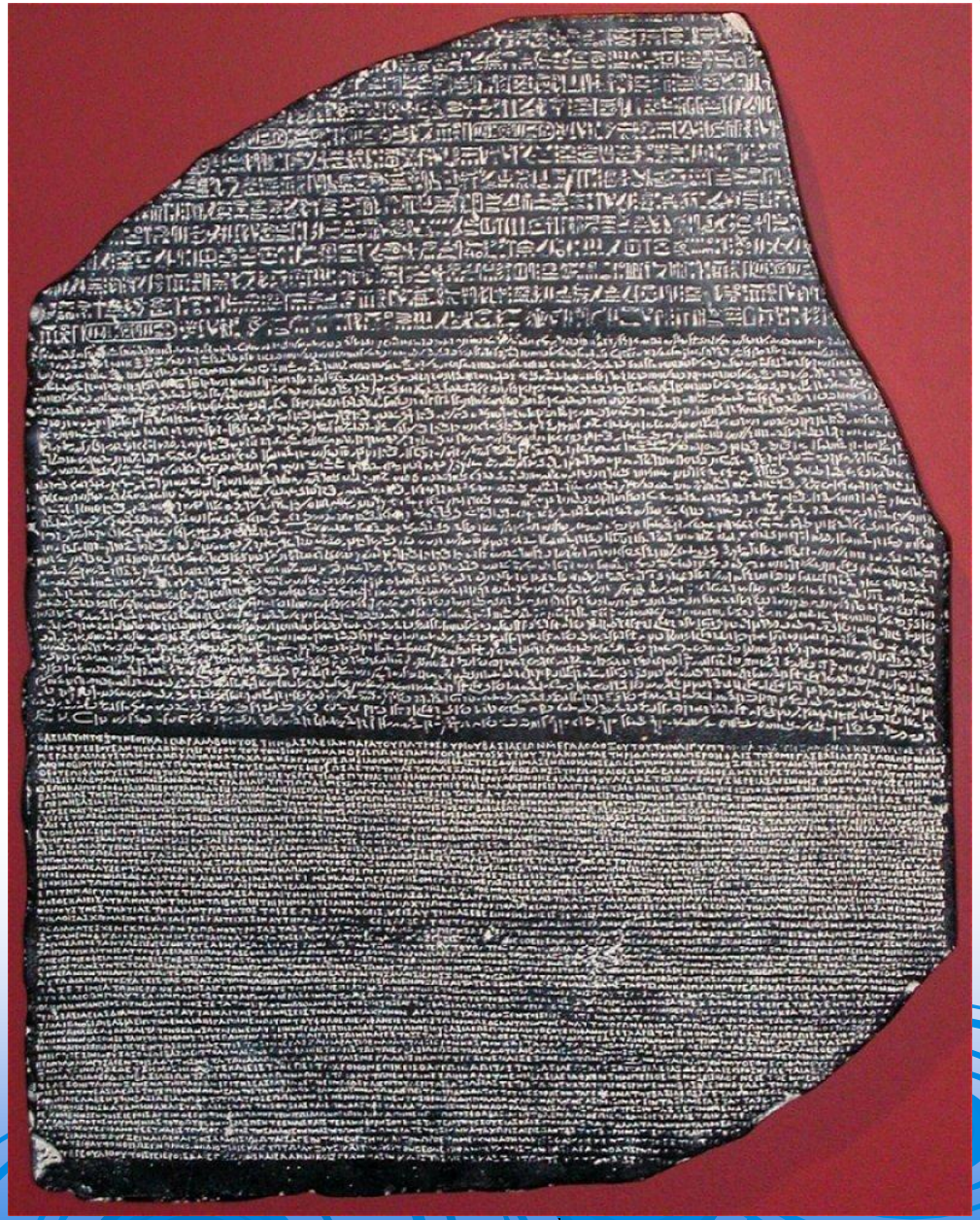
- Should there be compulsory clear-text deposit of electronic resources?
- How should digital preservation be funded?
- Should we select or just keep everything?
- Whose responsibility?
  - What if the publisher will not let subscribers do archiving and copying?
  - What if the publisher provides temporary access only to encrypted files, and then goes bankrupt?

Concluding warnings

- Current approaches to digital preservation are **still limited**
- They are **labor intensive**
- And very **costly**
- And require **institutional commitment & organization**
- But: **sustainable digital libraries depend upon the availability of preservation tools, services & efforts**


EDISON
HOME
PHONOGRAPH

We need many [Rosetta stones]



Tefko Saracevic


METADATA
HERITAGE
READ
DIFFICULT
NATIONAL
REQUIREMENTS
FIND
LIBRARY
YOUTUBE
ARCHIVING
CONCLUDING
ARCHIVAL
AVAILABLE
ISSUES
ORGANIZATION
FORM
LOST
SERVICES
ANOTHER
RESEARCH
SLOW
ELECTRONIC
EXTENT
PAPER
TECHNOLOGICAL
RECORDS
INFORMATION
DATA
SCIENTIFIC
LIBRARY
MEDIA
SYSTEM
REQUIRED
PUBLISHED
WEB
CURRENT
OBSOLESCENCE
DIGITAL
CONTENT
MATERIALS
PRESERVE
FORMAT
MANY
CULTURAL
LOCKSS
FORMATS
ACCESS
HARDWARE
NUMBER
INFORMATION
FILES
ACCESS
PREPAREDNESS
FUNCTIONS
ONE
KEEP
LIBRARIES
PROJECTS
TOOLS
TIME
SUPPORT
PERSONAL
PROJECT
RESOURCE
RESOURCES
PROVIDES
CREATED
NDIIPP
GEOSPATIAL
TREATMENT
PROVIDE
PRESERVATION