Important Issues in Preservation & Access of Digital Content for Children

Howard Besser
NYU Moving Image Archiving & Preservation Program
http://besser.tsoa.nyu.edu/howard/Talks
http://www.nyu.edu/tisch/preservation/

Interesting & Challenging Projects
– International Children's Digital Library
– Internet Archive Children's Library
– Archive-It K-12 Web Archiving

Digital Preservation Issues
– Paradigm shifts needed for digital preservation
– The problems with preserving digital works, and the myths that digitizing is the same thing as preservation
– Re-Formatting (and lack of control over technology changes) are things that Conservators already know

International Children's Digital Library (ICDL)

• freely accessible online collection
• over 4,500 digitized children’s books
• in more than 60 different languages from more than 200 countries
• target audience: 3-13 years old
• All books are presented as published in their original language, complete with illustrations

International Children's Digital Library (ICDL)

• launched in 2002 by the U Md as LIS research project to better understand children’s online habits as well as to encourage a love of reading
• in early years, the Internet Archive’s Childrens’ Library was an important collaborator. ICDL is now run by the International Children's Digital Library Foundation (ICDL).
ICDL Interface

• User interfaces in more than a dozen languages
• designed to be simple for children to search for books by color, book length, intended age group, fiction or non-fiction, type of book, language of book (which may be different than the language of the user interface), etc.
• No e-Reader software needed

ICDL Connectivity

• books not downloaded, instead read through standard Internet browser (Firefox, Internet Explorer)
  – this means maintaining Internet connection throughout the reading experience.
• copyright holders gave ICDL permission
• books cannot be downloaded, copied, or printed

ICDLF mission

• "the mission of the International Children's Digital Library Foundation is to prepare children for life in an ethnically and culturally diverse world by building the world's largest online multicultural repository of children's literature. We have set out to change the world, book by book, byte by byte. We offer free access to exemplary works from more than 42 countries. Our online library can be accessed for free in 11 languages with innovative software that was developed by hearing from young people about their needs, interests, and capacities."
  - Executive Director Tim Browne

ICDL Goals

• “To create a collection of more than 10,000 books in at least 100 languages that is freely available to children, teachers, librarians, parents, and scholars throughout the world via the Internet. The materials included in the collection reflect similarities and differences in cultures, societies, interests, lifestyles, and priorities of peoples around the world. The collection's focus is on identifying materials that help children to understand the world around them and the global society in which they live. It is hoped that through a greater understanding of one another that tolerance and acceptance can be achieved.”

Additional Goals address children of immigrants

• “A fundamental principle of the Foundation is that children and their families deserve to have access to the books of their culture, as well as the majority culture, regardless of where they live.”
ICDL Volunteers

- over 1,000 from around the world
- identify books for the collection
- secure rights
- package and send the books either physically or digitally
- tell others about the collection
- recruit new users

Librarian Volunteers

- ICDL’s “goal is to build a collection of books that represents outstanding historical and contemporary books from throughout the world.”
- librarians can assist the collection development activities by helping select quality literature

Children as Collaborators

- project co-founder Alison Druin specializes in user interfaces for children
- one initial project goal: discover best types of user interfaces to support children in searching, browsing, reading, and sharing
- sets of children become “design partners” in the development of optimum user interfaces

The aggregation lets them offer other types of interesting services

- ICDL Exhibitions presents “collections of books from around the world with similar themes. The exhibitions include an overview of the theme, links to the exhibition books, descriptions of why the books were included, and related activities.”
- Exhibition themes include: strong women, overcoming difficulties, water as a natural resource, how seasons change in different parts of the world, cultural differences, etc.
ICDL Activities
http://en.childrenslibrary.org/books/activities/
- scavenger hunts
- creative writing
- digital storytime
- learning languages
- ...
- + training manual outlining the sorts of activities that teachers might use in conjunction with ICDL

ICDL Community Forum
- for educators, librarians, parents, and others to discuss common issues. Topics have included:
  - suggestions of activities to accompany the reading of a particular book
  - discussion of resources on bibliotherapy
  - requests for identification of a book when someone only remembers fragments of the plot

StoryKit
- free iPhone/iPad App for children to create and illustrate their own electronic storybooks, and share these stories with others
- Kindergarten and first grade teachers have used this to help their students learn to compose

Book metadata for kids
Browsing through the book

Can translate the book

Translated story

Location Search

Advanced Search

Advanced Search
iPhone & iPad

Internet Archive’s Children’s Library (IACL)
http://archive.org/details/iacl

- Project of Internet Archive
- Originally partnered with ICDL-

IACL/ICDL Original Partnership

- IACL
  - repository for high-quality digitized scans
  - took a major responsibility for metadata development
- ICDL primarily focused on
  - developing a children-friendly user interface and presenting the digitized works within a normal web browser

IACL/ICDL differences

- IACL
  - as part of the Internet Archive, has always been concerned with providing to users the highest quality scans
  - preferred downloadable PDFs or high-quality web-based readers that allow for zooming in at very high resolutions
- ICDL has been most interested in building
  - interfaces for children’s navigation through content
  - as large a corpus of digitized material as possible.
My Own Speculation

• © Rightsholders will let ICDL distribute their works if they restrict downloading, high resolution zooms, and printing
• IACL insists that higher qualities are necessary for scholarly access (zooming) and for preservation
• Results…

Implications for Preservation

• ICDL
  – sees itself as access service, therefore doesn’t assume responsibility for long-term archiving or preservation
  – likely presumes that the paper copies of all the material in its corpus will be collectively saved by analog libraries across the world
• IACL
  – has the word “Archive” in its name
  – assumes serious responsibility for long-term stewardship of its digital content.

Further problems w/ICDL Preservation

• the scanned images (at least the ones publicly available) are not high resolution
• the user interface would not allow a library’s archiving web crawler to easily ingest an entire book because viewing a book requires stepping through it one or two pages at a time
Archive-It Web Archiving

- Internet Archive paid service
- Curated collections

Archive-It K-12 Web Archiving

http://www.archive-it.org/k12/

- Developed in 2008 w/LC
- 3rd to 12th grade students (8-18 years old)
- Select and save websites for future generations
- This year: 7 schools in 7 US states

Archive-It K-12 Web Archiving

- “The innovative program provides a new perspective on saving history and culture, allowing students to actively participate and make decisions about what content will be saved, that might not be archived in the traditional archive environment. The decisions the students make help them develop an awareness of how the Web content they choose will become primary sources for future historians studying our lives. The program teaches critical thinking, collaboration and problem solving - essential life skills that students need today in order to be successful.”

Archive-It K-12 Web Archiving

WHAT TEACHERS AND STUDENTS ARE SAYING

"The program was not about just saving the websites and links but archiving those for future usage and preventing from losing data."
- Student

"I loved how this project changed my students view of history. They felt empowered to speak for their generation about what was important at their school, in their community, and in the larger American culture. It brought real world learning into the classroom."
- Jayne StanNielsen, Ames Middle School

LC Video

America's Young Archivists: The K-12 Web Archiving Program
Digital Preservation -

- Paradigm shifts needed for digital preservation
- The problems with preserving digital works, and the myths that Digitizing is the same thing as Preservation
- Re-Formatting (and lack of control over technology changes) are things that Conservators already know
- Key necessary function: Standards, Metadata, Best Practices, file mgmt, Digital Preservation Repositories

Traditional Preservation (analog)

Hymn to Ninkasi (recipe for Sumerian Beer)

Paper Conservation at Museu Murilo Mendes

Analog Preservation/Conservation
Casa Rui Barbosa

Kungl Biblioteket-Conservation storage

Very different concerns than for Digital Storage

Problems with digital Preservation-

- The Viewing Problem
- The Scrambling Problem
- The Inter-relation Problem
- The Custodial Problem
- The Translation Problem

Obsolescence-the Viewing Problem

- Digital Info requires a whole infrastructure to view it
- Each piece of that infrastructure is changing at an incredibly rapid rate
- How can we ever hope to deal with all the permutations and combinations
Old Digital Formats

The Web: The Inter-relation Problem

- Info is increasingly inter-related to other info
- How do we make our own info persist when it points to and integrates with info owned by others?
- What is the boundary of a set of information (or even of a digital object)?

UFJF, Jan 1997

UFJF, May 1998

UFJF, Apr 2000

UFJF, Feb 2003
The Custodial Problem

- In the past, much of survival was due to redundancy
- How do we decide what to save?
- Who should save it?
- How should they save it?

What types of works should we save?

- Not only “classic” works
- Many marginal or everyday works later become very important to history or culture
  - Photographs on the street
  - Home Movies
  - Educational, training, and industrial films
  - Television commercials
  - News out-takes
  - Twitter and Facebook postings
- Orphan Works
Many years later we realize that we should have saved everyday works.

Childhood photo of Barry O.

Edward Snowden Facebook photos.

First US Gay Marriage.

Everyday photos become historically important.

Everyday Facebook pages become historically important ("We are all Khaled Saeed").
History of Passe-Livre Movement
Nov 2010

For Understanding how quickly large number of Brazilians went into the streets, we need to understand how they found out about police violence

Occupy—not just websites

The Custodial Problem: How to save information?

Conceptual Approaches to Digital Preservation

- **Refreshing** (move to a new support) always necessary due to volatility of physical strata
  - Impact on evidential value
- **Migration** -- advantages & disadvantages
- **Emulation** -- advantages & disadvantages
- And will need a long-term managed environment

Migration

- Wordstar to Word 1 to Word 3, …
- Tables and complex features often get corrupted
- Need to repeat every 4-5 years (maybe forever)
- We know how to do this ourselves
- If there’s a problem, we can catch it soon
**Emulation**

- Keep the Wordstar file format, but write emulators to make it work in newer environments
- +A better chance of carrying over complexity
- +Many more features can survive
- -Problems may not be caught until it’s too late
- -Specialists and a whole infrastructure of emulators required
- -Serious © problems (reverse engineering?)

---

**Responding to serious Longevity Problems**

- Previous formats required little ongoing intervention (remote storage facilities, Iron Mtn); digital formats require intense ongoing management
- Need for:
  - Regular re-formatting (Refreshing)
  - Preservation Repositories
  - Preservation Metadata

---

**Truth or Myth:**

**A digital work will last forever**

- Truth: You can repeatedly make exact copies of the file onto newer supports
- Myth: You will probably not be able to read that file 10 years from now (because the software to read it won’t work)
- So Yes, we can make a digital file last forever, but we have to do a lot of work to make sure that we will be able to read it!

---

**Paradigms Shifts needed**

<table>
<thead>
<tr>
<th>Physical preservation</th>
<th>Analog</th>
<th>Digital</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to preserve?</td>
<td>Repair original</td>
<td>ongoing mgmt</td>
</tr>
<tr>
<td></td>
<td>atmospheric ctrl</td>
<td>infinite copying</td>
</tr>
</tbody>
</table>

---

**After Digitization, you need Preservation**

- When you digitize a work, you spend the most time and money making the digitization
- If you have spent so many resources on digitization, you don’t want to repeat that later
- So you really need to preserve the results of any digitization that you do

---

**Born-digital works are both easier and harder to preserve than analog works**

- +With a born-digital work, we don’t have to worry so much about the “original artifact” (there really isn’t one)
- -We know that digital works face a whole range of obsolescence problems, so they must be reformatted at least once per decade
Back-up is not Preservation

- Back-up may save the file, but it doesn’t change it to make it readable in the future
- Many back-up services do not use checksums to assure that the file doesn’t change over time

Managed Environment

- More than temperature & humidity control
- Periodic monitoring of the works
- Periodic monitoring of the technical environment for viewing the works (software, systems, hardware)
- Trusted repositories

Storage Media

- Removable media (like CDs) is not a long-term answer
- The long-term answer requires ongoing management, and involves regular migration or emulation. This solution is only viable with storage on spinning disks.

Storing on CDs becomes a big problem over time

Consumers replace their CDs with a hard disk (& so should you)

We can’t rely on commercial services to preserve our works
A Web resource may be taken down (censored) quickly after it’s put up

Página do Brasil de Fato no Facebook está bloqueada há 4 dias

Besser-IFLA Children, Bangkok, 14/8/2013

Arab Spring Internet communication
(violation of user agreements)

Turkey blocks Twitter, Facebook, Tumblr (earlier this month)

Turkey Twitter (this week)

Continued accessibility of Works--Youtomb

- monitoring 435,900 YouTube videos
- 23,712 videos taken down for alleged copyright violation
- 109,428 videos taken down for other reasons
- http://youtomb.mit.edu/
YouTomb

Copyright is a Serious Issue

People in certain countries couldn’t see the one-time stream of restored Metropolis

Who is preserving today’s “born-digital” works?

- In the past, we knew about history by finding written documents:
  - Changes between different drafts of a scientific or literary paper
  - Letters and correspondence between a scientist (or literary figure) and colleagues (that both help contextualize the work, and lets us see changes in thought processes or discovery)
- But today, these documents are not on paper! They are in the form of:
  - Email correspondence
  - Word processing files that do not show changes between drafts/versions

Who will take responsibility to save these works for future study?

Tools for Preserving Websites: Archive-It

Standards, Metadata, & Best Practices to follow-

- Risk Management
- Best Practices for Reformatting
- Preservation Repositories & Metadata
- Other Metadata & Standards

Risk Management

- We can’t say definitively that we can make every digital work persist
- What we CAN say is that the more a digital work conforms to standards and best practices, the greater the likelihood that we can assure persistence
- Our preservation repositories can even accept deposits of non-conforming works, but the less they conform, the less likely that they’ll be salvageable
- Persistence is most likely for works that share standards, metadata, and best practices
Reformatting Best Practices

- Think about users (and potential users), uses, and type of material/collection.
- Scan at the highest quality that does not exceed the likely potential users/mater.
- Do not let today’s delivery limitations influence your scanning file sizes; understand the difference between digital masters and derivative files used for delivery.
- Many documents which appear to be bitonal actually are better represented with grayscale scans.

- Include color bar and ruler in the scan.
- Use objective measurements to determine scanner settings (do not attempt to make the image good on your particular monitor or use image processing to color correct).
- Don’t use lossy compression.
- Store in a common (standardized) file format.
- Capture as much metadata as is reasonably possible (including metadata about the scanning process itself).

Digital Preservation Players

- Collection staff (need to reach agreement on SIP/DIP and acceptable AIP transformations)
  - preservation/conservation staff
  - metadata staff
  - access staff
- Repository staff
- Agreement negotiators

Digital Preservation Repository Attributes

- Administrative responsibility
- Organizational viability
- Financial sustainability
- Technological suitability
- System security
- Procedural accountability

Preservation Repositories: Open Archival Info System Model

Producer

Repository Administration

Management

Consumer

Preservation Repositories: too difficult for small institutions

- Too complex for small institutions to manage.
- Will be done through partnering (small archive with large Library or University, or national organization), service bureau (OCLC), or through consortia (archive association, state-wide organization, …)
- Archive or museum will direct what is needed, but digital repository will carry out the actual work (as defined in SIP/DIP/AIP)

Metadata

- Technical Imaging Metadata Standards (Z39.87)
- Containers/Packaging (METS)
- Preservation (PREMIS)
PREMIS

- Designed to track the changes you make to a work in order to preserve it
- Every time you move a file to a new support (refresh) or migrate to a new file format, this generates a new PREMIS record
- The PREMIS record becomes a history of changes that you made to a digital work

OCLC/RLG Efforts
PREMIS Data Model

PREMIS Data Dictionary Example

Conclusions for preserving all types of complex works:

From the technological point of view

Standards offer the best hope of overcoming Impediments

- Easier to maintain a single set of standards over long periods of time
- Puts your institution in the same group with lots of other institutions who will face obsolescence and migration problems periodically throughout the future

What do we need to do?

- When we receive a born-digital work (or when we digitize a work):
  - Try to make sure that it is encoded in a widely-used non-proprietary format
  - Have a plan for refreshing the storage medium and migrating the format over time (every 5-10 years)
  - Keep records of every change, edit, migration, refreshing process
  - Make sure that processes or media decay (bit-flipping) don’t change the file (check-sums)

Paradigms Shifts needed
So, with digitalworks, the focus should be less on stable temperature (Helsinki underground vaults).

And less on the construction of Vaults (Helsinki underground vaults).

But more on ongoing management of a work without worrying so much about physical embodiment.

Important Issues in Preservation & Access of Digital Content for Children
Howard Besser, NYU Moving Image Archiving & Preservation Program

- http://besser.tsoa.nyu.edu/howard/Talks/
- http://www.nyu.edu/tisch/preservation
- http://www.digitalpreservation.gov/
- http://www.oclc.org/research/projects/pmwg/
- http://www.interpares.org