Future of Web Archiving

Stephen Abrams  
California Digital Library

Martin Klein  
Los Alamos National Laboratory

Jimmy Lin  
University of Maryland

Michael Nelson  
Old Dominion University
Agenda

- Web archiving problems and opportunities
- Memento tools
- WarcBase platform
- Assessing quality of archives
- Discussion
Web archiving is important but (really) hard

Why web archiving?

Continuation of longstanding mission to collect, preserve, and provide access to the scholarly record and our cultural heritage

Publishing/dissemination platform of choice

But ... the web isn’t the web anymore
Web in transition

A “web” of notes with links (like references) between them…”
– Tim Berners-Lee, March 1989

Document retrieval ➔ Programming environment
Document viewer ➔ Virtual machine
HTML ➔ JavaScript
Common ➔ Personalized
Desktop ➔ Mobile/handheld/wearable
Information ➔ Things

www.flickr.com/photos/swamibu/2223726960
www.flickr.com/photos/sharples/79222765
(Some) other issues

- Crawlers don’t act like browsers
  - Need robots that act more like people
(Some) other issues

- Crawlers don’t act like browsers
- Responsiveness to time-sensitive content
  - Need to bypass very deliberate collection development procedures
(Some) other issues

- Crawlers don’t act like browsers
- Responsiveness to time-sensitive content
- Policies, rights, and permissions
  - Need to overcome legal barriers that follow the monetization of content

U Can’t Touch This

-M.C. Hammer
(Some) other issues

- Crawlers don’t act like browsers
- Responsiveness to time-sensitive content
- Policies, rights, and permissions
- Difficult integration into traditional management and discovery services

Leading to ...
(Some) other issues

- Crawlers don’t act like browsers
- Responsiveness to time-sensitive content
- Policies, rights, and permissions
- Difficult integration into traditional management and discovery services
- Siloed collections

www.flickr.com/photos/54159370@N08/7148880783
(Some) other issues

- Crawlers don’t act like browsers
- Responsiveness to time-sensitive content
- Policies, rights, and permissions
- Difficult integration into traditional management and discovery services
- Siloed collections

Scale

- Storage capacity
- Full-text indexing
- De-duplication
- Resources
Supporting research

- Little awareness in the scholarly community
- Poorly understood use cases
- Few tools
- Traditional find → download → manipulate locally workflows may not be feasible at web scale

► Need APIs and business models for in situ analysis
Technological opportunities

Better capture mechanisms

- Headless browsers
- API harvesters

Better discovery modalities

- Browsing the past should be as simple and intuitive as the now
Cooperative opportunities

- Complementary collection development
- Coordinated infrastructure support and operation
  - Or perhaps centralized – a HathiTrust for web archives?
- Crowd sourcing selection, description, quality assurance
And now ...