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

CC BY

Agenda

- Web archiving problems and opportunities
- Memento tools
- WarcBase platform
- Assessing quality of archives
- Discussion

www.flickr.com/photos/adesigna/4090782772

Web archiving is important but (really) hard

- Why web archiving?

www.flickr.com/photos/alaig/3522953697

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

www.flickr.com/photos/hier_gibt_es_nichts_zu_sehen_bitte_gehen_sie_weiter/840587382

Web in transition

> A "web" of notes with links (like references) between them ..."
> – Tim Berners-Lee, March 1989

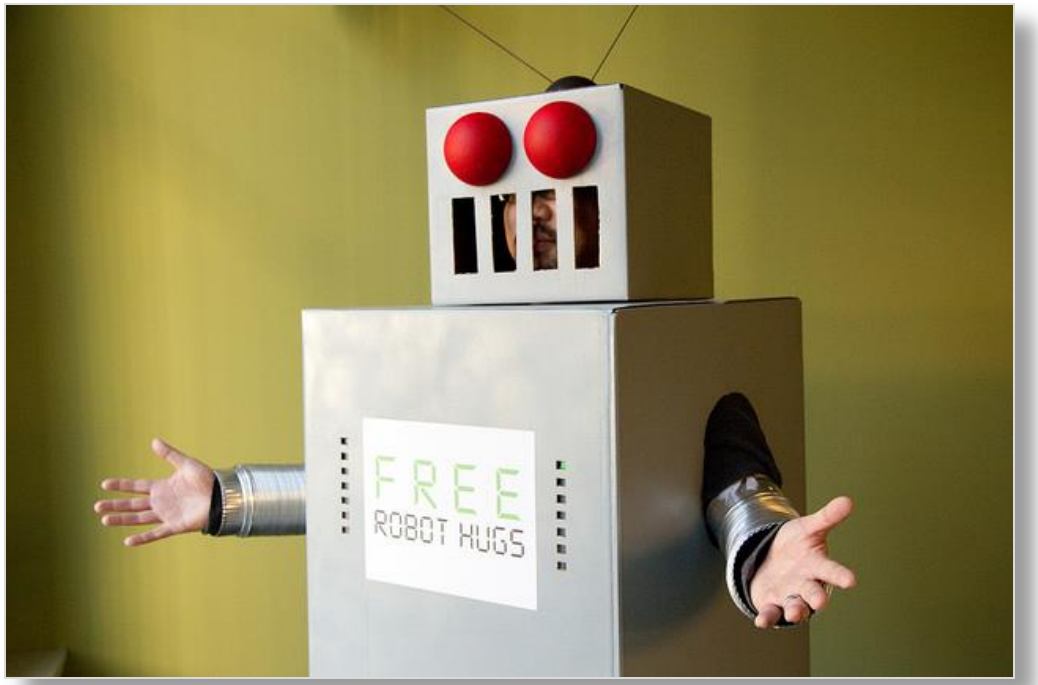
Document retrieval	➡	Programming environment
Document viewer	➡	Virtual machine
HTML	➡	JavaScript
Common	➡	Personalized
Desktop	➡	Mobile/handheld/wearable
Information	➡	Things

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*



www.flickr.com/photos/benhusmann/5126030385

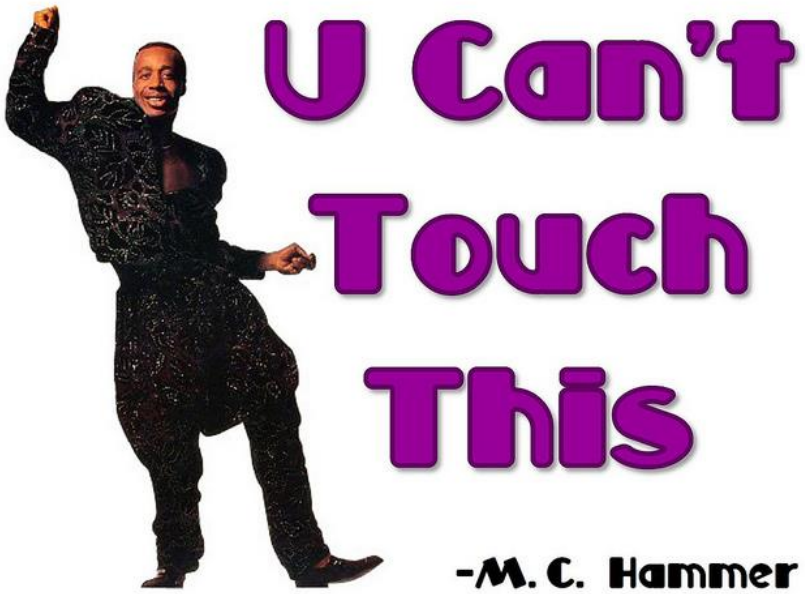
(Some) other issues

- Crawlers don't act like browsers
- Responsiveness to time-sensitive content
  - *Need to bypass v-e-r-y deliberate collection development procedures*

Gaurdian News and Media Limited

(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*



www.flickr.com/photos/vblibrary/7414544704

(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 …*

www.flickr.com/photos/21664580@N04/2095574414

(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

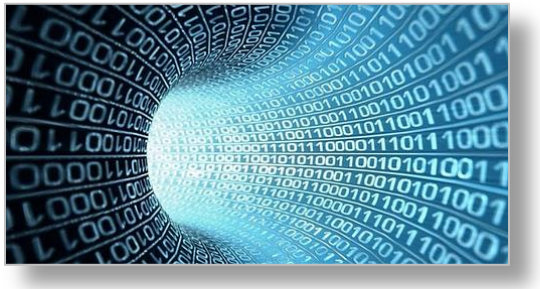
- Crawlers don't act like browsers
- Responsiveness to time-sensitive content
- Policies, rights, and permissions
- Difficult integration into traditional management and discovery services
- Siloed collections
- Scale
  - *Storage capacity*
  - *Full-text indexing*
  - *De-duplication*
  - *Resources*

Raiders of the Lost Ark © Paramount Pictures

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*

berkeley.edu/teach



www.flickr.com/photos/infocux/8450190120

Technological opportunities

- Better capture mechanisms
  - *Headless browsers*
  - *API harvesters*
  - …

www.flickr.com/photos/shebalso/6357626617

- Better discovery modalities
  - *Browsing the past should be as simple and intuitive as the now*
  - …

www.flickr.com/photos/bartelomeus/4184705426

Cooperative opportunities

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

www.flickr.com/photos/chiotsrun/4115059294

www.flickr.com/photos/sagesolar/9230445157

And now …

cdn.ws.citrix.com/wp-content/uploads/2012/05/iStock_000010348904XSmall.jpg