Collaborating to Address a Digital Backlog of Web Archives

Jesse Johnston, Grace Thomas, Rick Fitzgerald
Library of Congress
Case Study: Applied Research

- How can we reduce the amount of underdescribed Web materials?
  - While: building on organizational strengths, maintaining standard descriptive practice, and achieving higher throughput to describe large-scale digital collections?

- Where can we apply inspirations or take lessons from archival description practices?
LC Web Archives

- Program since 2000
- Thematic & Event
- Permission to crawl required
- Much available: loc.gov/websites/collection

@jesseajohnston
Slides: https://goo.gl/ACpTbH
Digital collections: the challenge of scale

Extent of Web Archives Described

- Total Entities (approx)
- Entities described

- 2000-2017: 58%
- 2018: 66%
New Processes & Products

Collaborative Inputs
- Recommenders - nominate collections
- Cataloger - developed MODS framework
- Web Archivist - develop python scripts
- Web Services - push to loc.gov site
Minimal Description

Pulled information from crawl database into minimal MODS framework

More details [here](https://goo.gl/ACpTbH)

Example (**cute overload! ;)**: [loc.gov](https://loc.gov), [MODS](https://cataloging.libraryofcongress.gov/standards/mods)

Relational Series
Implications

- More web archives available!
- Alternative to hierarchical description

Longer version available on The Signal:

http://blogs.loc.gov/thesignal/2018/08/more-web-archives-less-process/

Slides: https://goo.gl/ACpTbH
Contact

Jesse Johnston
Grace Thomas
Rick Fitzgerald

webcapture@loc.gov

http://www.loc.gov/websites/collections
Library of Congress Digital Inventory

- Web Archives: 33.3%
- Other collections: 29.9%
- Copyright Card Catalog: 4.5%
- NDNP: 7.5%
- VHP: 24.9%