Interoperability and Technical Collaboration for Web and Social Media Archiving

Nicholas Taylor (@nullhandle)
Web Archiving Service Manager
Stanford University Libraries

DocNow Advisory Board Meeting
August 22, 2016
Heritrix archival crawler

![Screenshot of Heritrix 1.8.0 admin console...](https://foo.edu/0000/index.jsp) by Frank McCown under MPL 1.1
Heritrix is great for archiving the Web
...of ten years ago
newer capture approaches

- **headless browsers**
  - to prospect content only apparent by executing JavaScript

- **archiving proxies**
  - to enable more, and more specialized, capture tools to write to WARC

- **leveraging APIs**
  - to more reliably collect higher-fidelity data from major social media services
WARC in Social Feed Manager

Justin Littman: “Aligning Social Media Harvesting and Web Harvesting”
web archiving system APIs (WASAPI)

National Digital Platform Projects funded in August 2015

Systems Interoperability and Collaborative Development for Web Archiving (LG-71-15-0174-15): The Internet Archive, working with partner organizations University of North Texas, Rutgers University, and Stanford University Library will undertake a two-year research project to explore techniques that can expand national web archiving capacity in several areas.
technical architectures to facilitate contributions by a broad community?
community frameworks to enable broad participation in shaping technologies?
how to build more, and more distributed, capacity?
how to make web and social media archiving more inclusive?