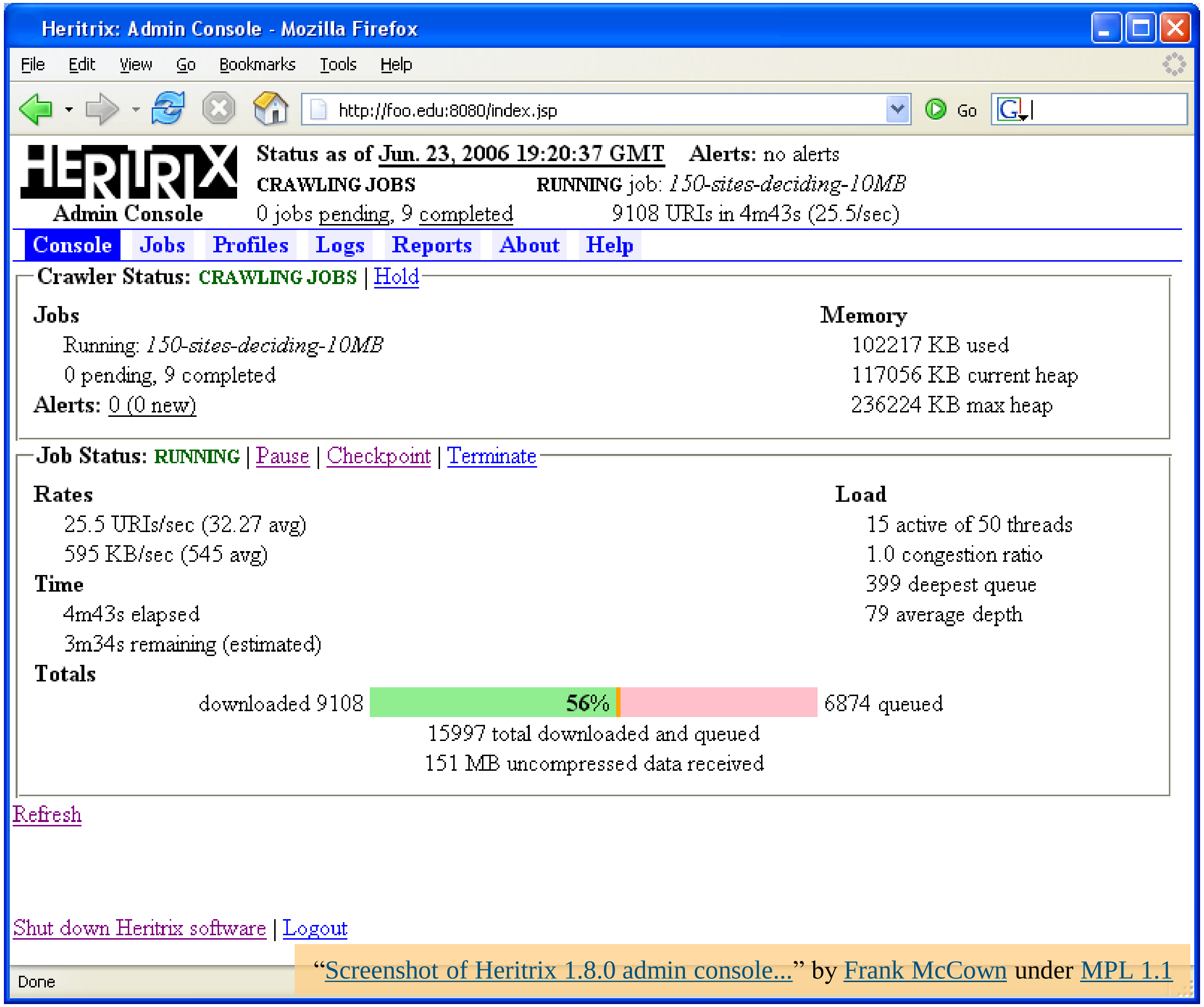
STANFORD UNIVERSITY LIBRARIES

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


Heritrix: Admin Console - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://foo.edu:8080/index.jsp
Go
HERITRIX
Admin Console
Console Jobs Profiles Logs Reports About Help
Done

Status as of Jun. 23, 2006 19:20:37 GMT Alerts: no alerts

CRAWLING JOBS RUNNING job: *150-sites-deciding-10MB*

0 jobs pending, 9 completed 9108 URIs in 4m43s (25.5/sec)

Crawler Status: CRAWLING JOBS | Hold

Jobs

Running: *150-sites-deciding-10MB*

0 pending, 9 completed

Alerts: 0 (0 new)

Memory

102217 KB used

117056 KB current heap

236224 KB max heap

Job Status: RUNNING | Pause | Checkpoint | Terminate

Rates

25.5 URIs/sec (32.27 avg)

595 KB/sec (545 avg)

Time

4m43s elapsed

3m34s remaining (estimated)

Totals

downloaded 9108 56% 6874 queued

15997 total downloaded and queued

151 MB uncompressed data received

Load

15 active of 50 threads

1.0 congestion ratio

399 deepest queue

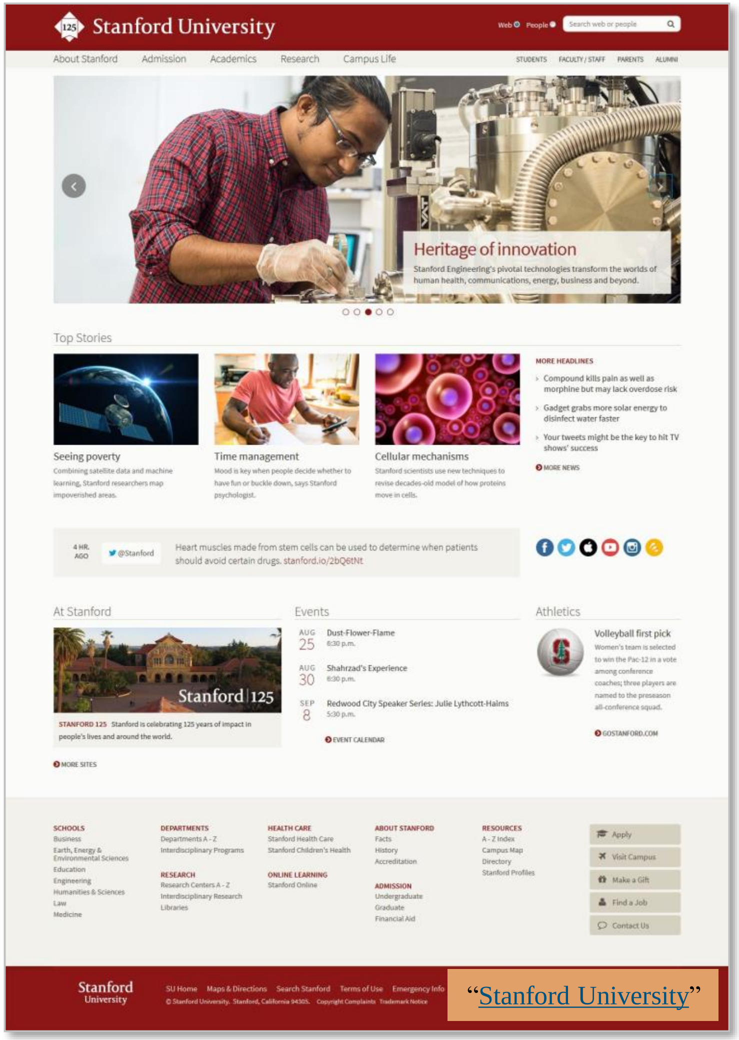
79 average depth

Refresh

Shut down Heritrix software | Logout

“Screenshot of Heritrix 1.8.0 admin console...” by Frank McCown under MPL 1.1

Heritrix is great for archiving the Web



Heritage of innovation

Stanford Engineering's pivotal technologies transform the worlds of human health, communications, energy, business and beyond.

Top Stories

Seeing poverty

Combining satellite data and machine learning, Stanford researchers map impoverished areas.

Time management

Mood is key when people decide whether to have fun or buckle down, says Stanford psychologist.

Cellular mechanisms

Stanford scientists use new techniques to revise decades-old model of how proteins move in cells.

MORE HEADLINES

- Compound kills pain as well as morphine but may lack overdose risk
- Gadget grabs more solar energy to disinfect water faster
- Your tweets might be the key to hit TV shows' success

MORE NEWS

4 HR. AGO @Stanford Heart muscles made from stem cells can be used to determine when patients should avoid certain drugs. stanford.io/2bQ6tNt

At Stanford

STANFORD 125 Stanford is celebrating 125 years of impact in people's lives and around the world.

MORE SITES

Events

AUG 25 Dust-Flower-Flame 6:30 p.m.

AUG 30 Shahrzad's Experience 6:30 p.m.

SEP 8 Redwood City Speaker Series: Julie Lythcott-Haims 5:30 p.m.

EVENT CALENDAR

Athletics

Volleyball first pick

Women's team is selected to win the Pac-12 in a vote among conference coaches; three players are named to the preseason all-conference squad.

GOSTANFORD.COM

SCHOOLS
Business
Earth, Energy & Environmental Sciences
Education
Engineering
Humanities & Sciences
Law
Medicine

DEPARTMENTS
Departments A - Z
Interdisciplinary Programs

RESEARCH
Research Centers A - Z
Interdisciplinary Research
Libraries

HEALTH CARE
Stanford Health Care
Stanford Children's Health

ONLINE LEARNING
Stanford Online

ABOUT STANFORD
Facts
History
Accreditation

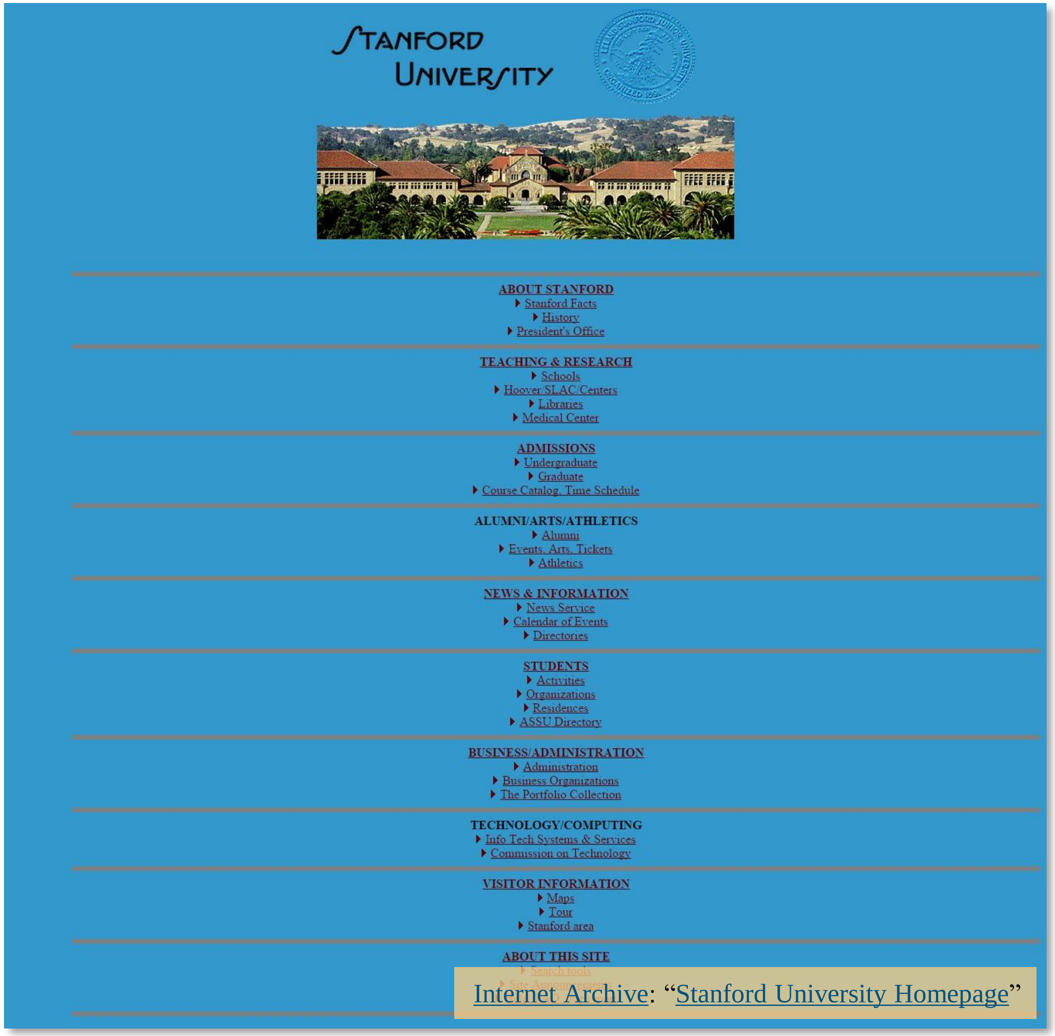
ADMISSION
Undergraduate
Graduate
Financial Aid

RESOURCES
A - Z Index
Campus Map
Directory
Stanford Profiles

© Stanford University. Stanford, California 94305. Copyright Complaints Trademark Notice

"[Stanford University](#)"

…of ten years ago



STANFORD UNIVERSITY

ABOUT STANFORD
- Stanford Facts
- History
- President's Office

TEACHING & RESEARCH
- Schools
- Hoover/SLAC/Centers
- Libraries
- Medical Center

ADMISSIONS
- Undergraduate
- Graduate
- Course Catalog, Time Schedule

ALUMNI/ARTS/ATHLETICS
- Alumni
- Events, Arts, Tickets
- Athletics

NEWS & INFORMATION
- News Service
- Calendar of Events
- Directories

STUDENTS
- Activities
- Organizations
- Residences
- ASSU Directory

BUSINESS/ADMINISTRATION
- Administration
- Business Organizations
- The Portfolio Collection

TECHNOLOGY/COMPUTING
- Info Tech Systems & Services
- Commission on Technology

VISITOR INFORMATION
- Maps
- Tour
- Stanford area

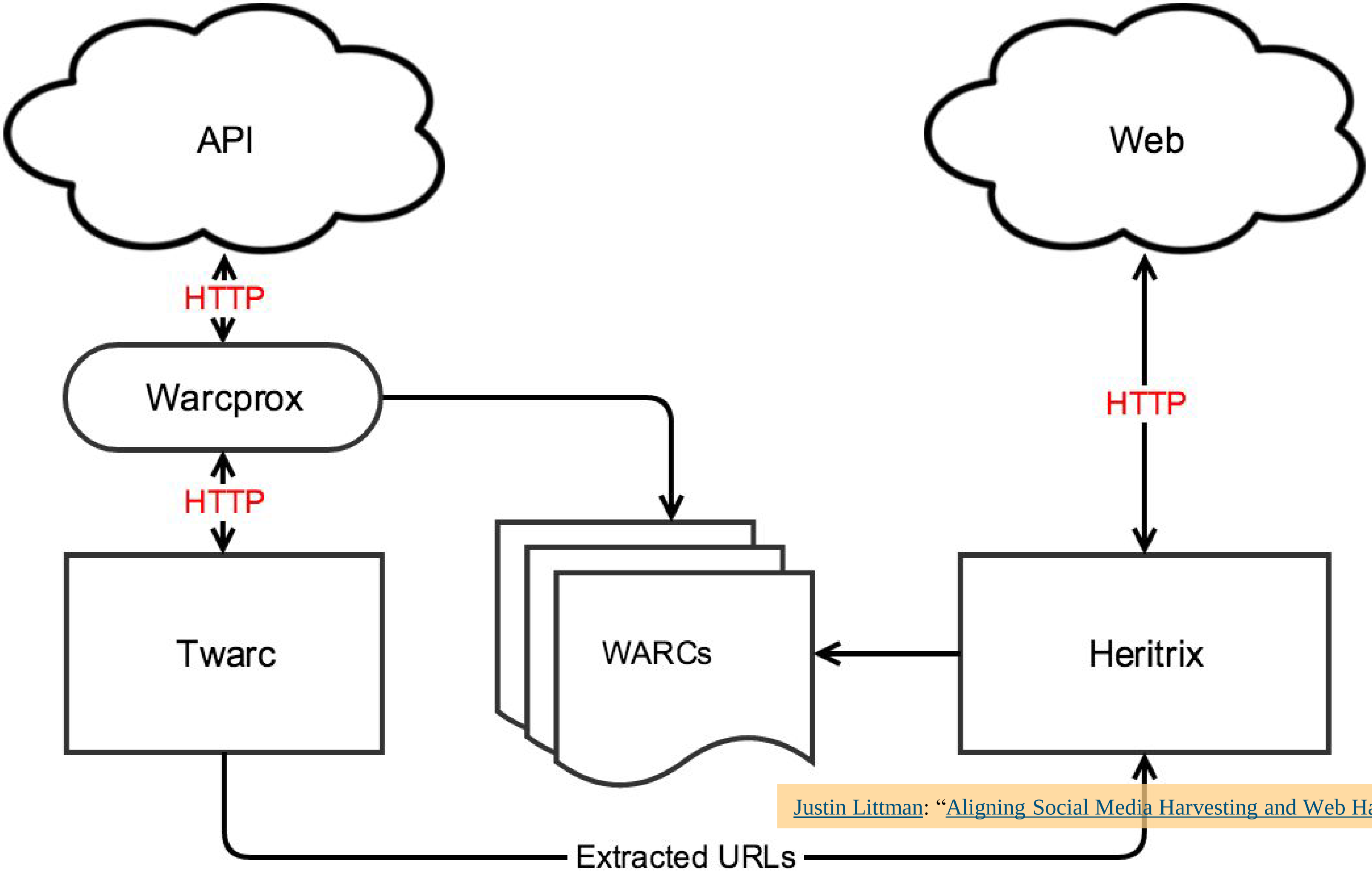
ABOUT THIS SITE
- Search tools
- Site Announcements
- [illegible]

Internet Archive: “Stanford University Homepage”

newer capture approaches

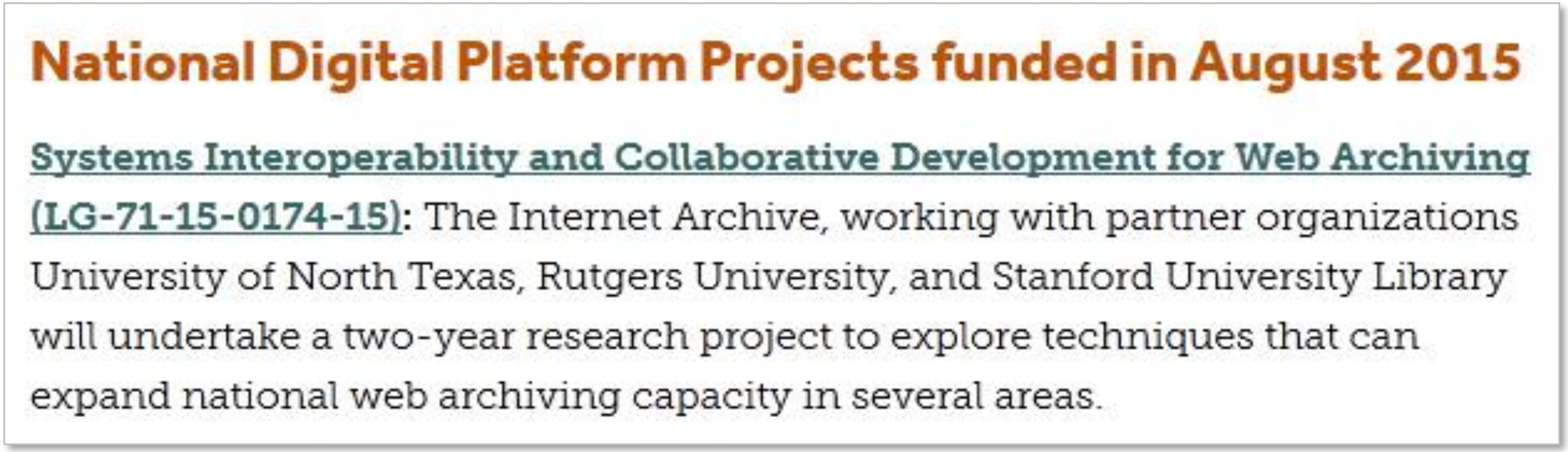
- headless browsers
  - to prospect content only apparent by executing JavaScript
- archiving proxies
  - to enable more, and more specialized, capture tools to write to WARC
- leveraging APIs
  - to more reliably collect higher-fidelity data from major social media services

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.

INTERNET ARCHIVE

INSTITUTE of Museum and Library SERVICES

RUTGERS UNIVERSITY

UNT®

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?

DocNow