Rethinking Web Archiving Quality Assurance for Impact, Scalability, and Sustainability

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

Archives 2016
209 - Balancing Quality of Life and Quality Assurance
August 4, 2016
QA panelists

Lori Donovan
Internet Archive / Archive-It

Dory Bower
Government Publishing Office

Dallas Pillen
Bentley Historical Library

Alex Thurman
Columbia University Libraries

Nicholas Taylor
Stanford University Libraries
balancing QA + quality of life?
we set and forget; I’m just glad we’re doing something

steady, ongoing QA is challenging

occasionally I set aside a lunch hour to do some QA

did more QA at the beginning but, well, I don’t really look at the reports any more

my strategy right now is to let the big schools figure it out
2015 SAA WebArchRT discussion

• if you could only apply 3 QA practices to your web archives, which 3?
• do you apply different QA practices to web archives created for different use cases?
• how do you ensure that staff time allocated to QA is best spent?
quality assurance in the lifecycle
quality assurance, expansively

typical QA
• parsing robots.txt
• scoping rules
• object count limits
• test crawling
• inspecting archived site
• reviewing reports
• patch crawling

and more
• seed selection
• assessing live site
• capture tool selection
• crawl scheduling
• crawl duration limits
• monitoring crawl
• archivability advocacy
• training
3rd highest desired skill

NDSA: “2015 NDSA Web Archiving Survey”
low perceived programmatic progress

NDSA: “2015 NDSA Web Archiving Survey”
greatest collaboration interest

- Policy + Risk Management
- Capture Configuration
- Collaborative Collection Dev
- Input on APIs + Standards
- Metadata Standards
- QA Techniques + Strategies
- Tool Dev
- Other

NDSA: “2015 NDSA Web Archiving Survey”
web archiving at Stanford

• 7 Archive-It accounts
• Heritrix, Webrecorder
• local preservation, discovery, access
• program manager, curators, students
• tens of collections
• thousands of seeds
quality assurance goals

- maximize impact + efficiency of QA efforts
- enable diverse, distributed, + approachable contributions
- calibrate investments in quality based on tool capabilities
capture, behavior, appearance
capture, behavior, appearance
in practice

care more about...
• report data
• crawl finishing
• 4xx, 5xx, complete robots.txt block
• plausible duration
• plausible object counts
• scoping out extraneous content
• new seeds

care less about...
• visual inspection
• reviewing every capture
• appearance fidelity
• behavior fidelity
• partial content out of scope
• partial content blocked by robots.txt
• ongoing seeds
more next from Lori, Alex, Dallas, Dory