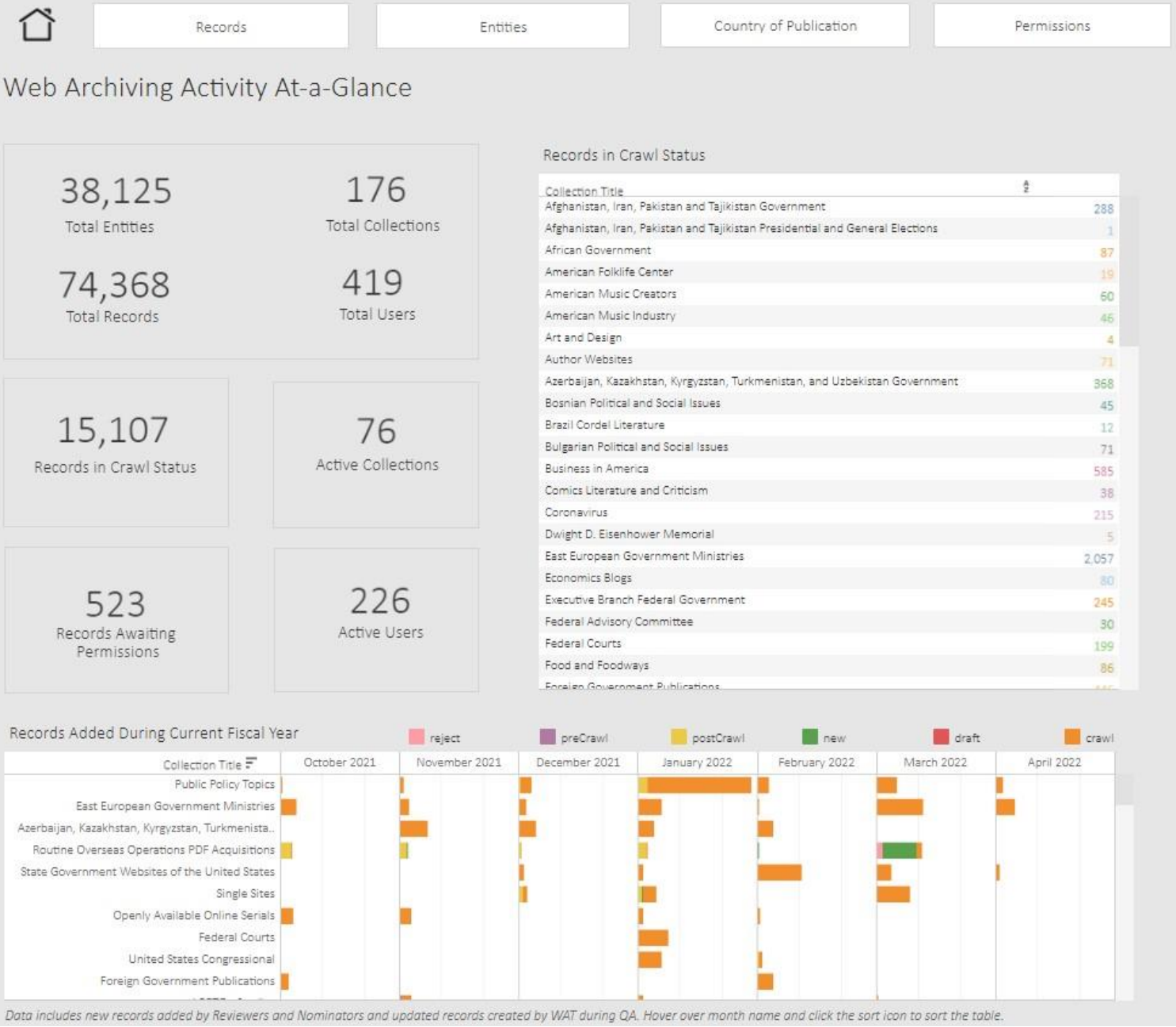
LIBRARY OF CONGRESS

May 2022

Building a Sustainable Quality Assurance Lifecycle at the Library of Congress

Presented by Grace Thomas and Meghan Lyon for the 2022 IIPC Web Archiving Conference


Records
Entities
Country of Publication
Permissions
Web Archiving Activity At-a-Glance
38,125
Total Entities
176
Total Collections
74,368
Total Records
419
Total Users
15,107
Records in Crawl Status
76
Active Collections
523
Records Awaiting Permissions
226
Active Users
Records in Crawl Status
Collection Title
Afghanistan, Iran, Pakistan and Tajikistan Government 288
Afghanistan, Iran, Pakistan and Tajikistan Presidential and General Elections 1
African Government 87
American Folklife Center 19
American Music Creators 60
American Music Industry 46
Art and Design 4
Author Websites 71
Azerbaijan, Kazakhstan, Kyrgyzstan, Turkmenistan, and Uzbekistan Government 368
Bosnian Political and Social Issues 45
Brazil Cordel Literature 12
Bulgarian Political and Social Issues 71
Business in America 585
Comics Literature and Criticism 38
Coronavirus 215
Dwight D. Eisenhower Memorial 5
East European Government Ministries 2,057
Economics Blogs 80
Executive Branch Federal Government 245
Federal Advisory Committee 30
Federal Courts 199
Food and Foodways 86
Records Added During Current Fiscal Year
reject
preCrawl
postCrawl
new
draft
crawl
Collection Title
October 2021
November 2021
December 2021
January 2022
February 2022
March 2022
April 2022
Public Policy Topics
East European Government Ministries
Azerbaijan, Kazakhstan, Kyrgyzstan, Turkmenista..
Routine Overseas Operations PDF Acquisitions
State Government Websites of the United States
Single Sites
Openly Available Online Serials
Federal Courts
United States Congressional
Foreign Government Publications
Data includes new records added by Reviewers and Nominators and updated records created by WAT during QA. Hover over month name and click the sort icon to sort the table.

What are the goals of QA?

1. improve the quality of the captures

2. provide a reasonable expectation of the usability of the archive

Grounded Theory for QA in Three Dimensions

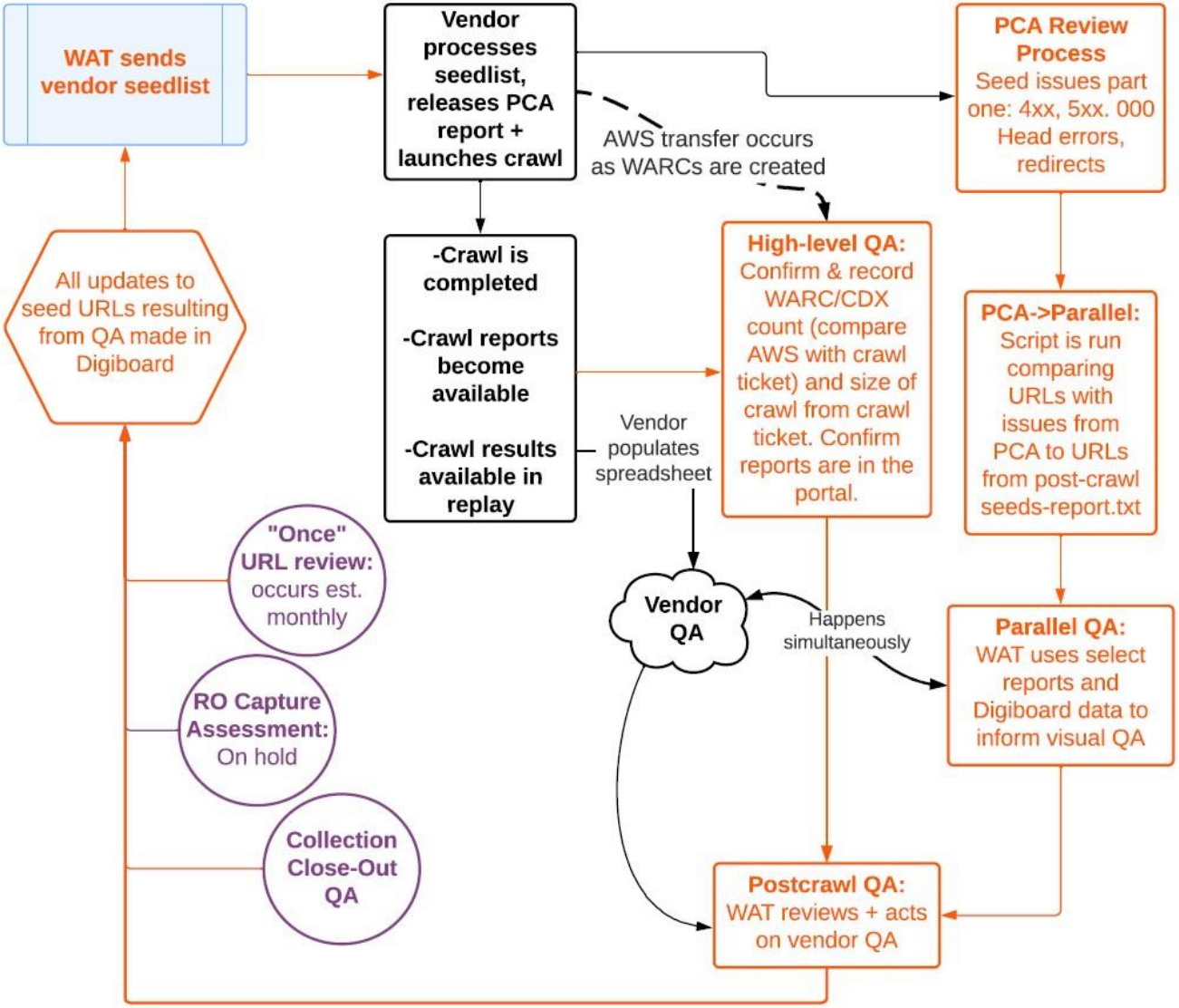
Archivability: degree to which the intrinsic properties of a website make it easier or more difficult to archive

Relevance: pertinence of the contents of an archived website to the original website

Correspondence: degree of similarity, or resemblance, between the original website and the archived website

Brenda Reyes Ayala

https://link.springer.com/article/10.1007/s00799-021-00314-x


WAT sends vendor seedlist
Vendor processes seedlist, releases PCA report + launches crawl
AWS transfer occurs as WARCs are created
PCA Review Process
Seed issues part one: 4xx, 5xx. 000 Head errors, redirects
All updates to seed URLs resulting from QA made in Digiboard
-Crawl is completed
-Crawl reports become available
-Crawl results available in replay
High-level QA:
Confirm & record WARC/CDX count (compare AWS with crawl ticket) and size of crawl from crawl ticket. Confirm reports are in the portal.
PCA->Parallel:
Script is run comparing URLs with issues from PCA to URLs from post-crawl seeds-report.txt
Vendor populates spreadsheet
"Once" URL review:
occurs est. monthly
Vendor QA
Happens simultaneously
Parallel QA:
WAT uses select reports and Digiboard data to inform visual QA
RO Capture Assessment:
On hold
Collection Close-Out QA
Postcrawl QA:
WAT reviews + acts on vendor QA

Archivability

Archivability: degree to which the intrinsic properties of a website make it easier or more difficult to archive

Brenda Reyes Ayala
https://link.springer.com/article/10.1007/s00799-021-00314-x

Web Archiving Known Challenges

Created by Meghan Lyon just a moment ago

Certain platforms and websites, including social media, regularly present capture challenges for the web archiving community, particularly when using at scale web-crawlers (i.e. Heritrix). WAT has created this page for recommending officers and collection leaders to use as a reference for up-to-date guidance on difficult content. This page is created as technical guidance and is not a statement about collecting policy.

Platform	Capture status	Playback status	WAT Guidance	Time of update
Twitter	Twitter is currently not being captured.	Playback of older captures may be slow. Some captures will not replay embedded images.	Do not add Twitter seeds to the crawl.	July 8 2021
Facebook	The log in feature on Facebook is blocking the crawler from capturing content. Facebook changes its source code frequently, which impedes our ability to successfully capture it at any given time.	If a capture is successful, playback is usually limited to the first page without the option to scroll for more content.	Do not add Facebook seeds to the crawl.	July 8 2021
Instagram	All Instagram URLs redirect to the log in page on the live web, blocking the crawler from capturing content.	Playback of older captures may include missing images and functional issues (i.e. inability to "load more")	Do not add Instagram seeds to the crawl.	July 8 2021
YouTube	We are currently unable to capture content hosted on YouTube.com, including embedded YouTube videos.	Successfully harvested legacy YouTube content will not replay in our current OpenWayback environment.	Do not add YouTube seeds to the crawl.	July 8 2021
Soundcloud	Soundcloud audio and Soundcloud embeds are not being captured.	Older SoundCloud captures will not replay in our current OpenWayback environment.	Do not add SoundCloud seeds to the crawl. See note below for guidance on Soundcloud files embedded in podcasts.	December 15 2021
Vimeo	We are currently unable to capture content hosted on Vimeo.com	Captured Vimeo content will not replay in our current OpenWayback environment.	Do not add Vimeo seeds to the crawl.	July 8 2021
Medium	Medium content can be reliably captured	Playback issues, such as images failing	Continue to add and leave Medium	July 8 2021

Relevance

Relevance: pertinence of the contents of an archived website to the original website

(a)Topic relevance: degree to which an archived website (or a web archive) includes only content that is closely related to that of the original website or the topic of the larger web archive

(b)Size relevance: the similarity in size of the archived website to the original website

Brenda Reyes Ayala

https://link.springer.com/article/10.1007/s00799-021-00314-x

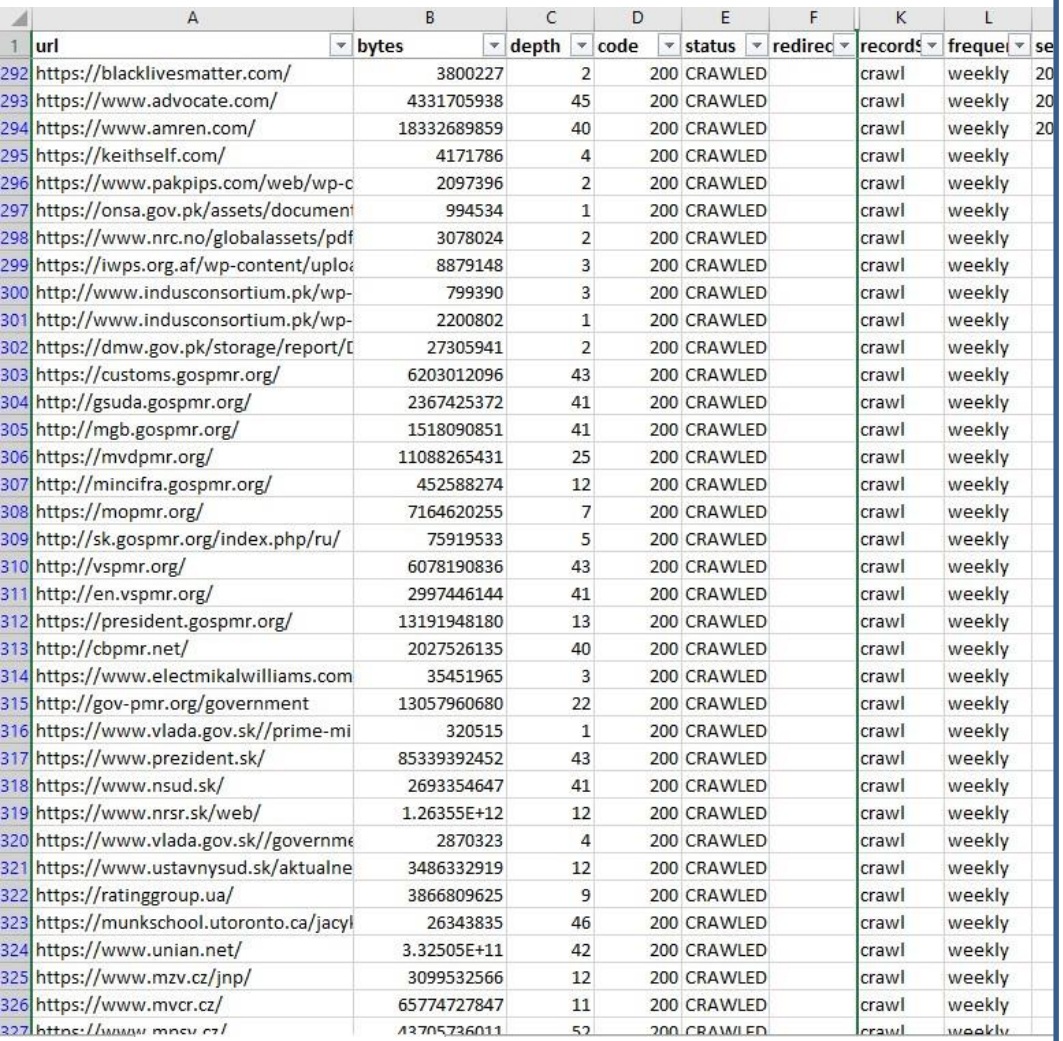
Size Relevance, cont.

53	1078620984652	https://www.alternet.org/
53	11195091392	https://www.colorlines.com/
53	66116285268	https://techcrunch.com/
53	45386066348	http://historynewsnetwork.org/
53	15765050515	https://www.americanthinker.com/
53	8174570883	https://www.takimag.com/
53	10110633328	https://www.sarawakrose.org/
53	12890110745	http://www.drudgereport.com/
53	116457230794	https://www.upworthy.com/
53	812640183209	http://duma.gov.ru/
53	8832688190	https://www.hoodedutilitarian.com/
53	306333240990	https://er.ru/
53	24517253298	https://ips-dc.org/
53	26134713387	https://science.thewire.in/
53	12809537565	https://www.rightwingwatch.org/topics/cap
52	103383987719	https://www.axios.com/
52	44912415745	https://www.mpsv.cz/
47	27228881878	https://www.planetizen.com/
47	5595086152914	https://blavity.com/
46	24101429	https://munkschool.utoronto.ca/jacyk/
46	18453326965	https://www.freep.com/
46	53873065596	https://talkingpointsmemo.com/
46	4708531984	https://www.culture.gov.sk/
45	17832465527	http://www.back2stonewall.com/
45	2464399795	https://dse.md/
44	2232209421	https://democrats.org/
44	9255811074	https://www.larouchepac.com/
44	51569049088	https://www.akparti.org.tr/
44	17123065191	https://www.malaymail.com/
44	38216029555	https://www.chalkbeat.org/
44	75877161261	https://www.propublica.org/
44	187563510	https://www.wvwvaf.org/
44	43723023382	https://www.nerc.gov.ua/
44	6736657972	https://www.theamericanconservative.com/
44	25342948229	https://www.mpo.cz/
44	12074159288	https://www.pbshawaii.org/
44	1251841466	https://cherta.media/
44	1058389921682	https://spravedlivo.ru/
43	33545212610	https://revealnews.org/
43	1149218094	https://endcitizensunited.org/
43	1758080382	https://www.rga.org/
43	19207761566	https://www.mil.gov.ua/
43	70529480847	https://dsp.gov.ua/
43	24002447071	https://www.davr.gov.ua/
43	10926847039	https://moz.gov.ua/
43	25880556136	https://www.irrawaddy.com/
43	9056119473	https://www.foodpolitics.com/
43	6076096348	https://torrentfreak.com/
43	15796396628	https://memohrc.org/ru
43	4101683153	https://www.teapartypatriots.org/
43	3700640833	https://reproductiverights.org/

Depth, bytes, and seed

http response code, status, seed, and redirect

```
403 CRAWLED https://idph.iowa.gov/Emerging-Health-Issues/Novel-Coronaviru
302 CRAWLED https://www.pia.gov.ph/ https://pia.gov.ph/
302 CRAWLED http://www.pia.gov.ph/ https://www.pia.gov.ph/
302 CRAWLED https://gas.rk.gov.ru/ru/index https://gas.rk.gov.ru/check
302 CRAWLED https://gkz.rk.gov.ru/ru/index https://gkz.rk.gov.ru/check
302 CRAWLED https://kkp.rk.gov.ru/ru/index https://kkp.rk.gov.ru/check
302 CRAWLED https://kpk.rk.gov.ru/ru/index https://kpk.rk.gov.ru/check
302 CRAWLED https://msh.rk.gov.ru/ru/index https://msh.rk.gov.ru/check
302 CRAWLED https://power.lenobl.ru/ https://power.lenobl.ru/ru/
302 CRAWLED https://press.lenobl.ru/ https://press.lenobl.ru/ru/
302 CRAWLED https://set.rk.gov.ru/ru/index https://set.rk.gov.ru/check
302 CRAWLED https://sfn.rk.gov.ru/ru/index https://sfn.rk.gov.ru/check
302 CRAWLED http://www.senate.gov.ph/ http://legacy.senate.gov.ph/
302 CRAWLED https://crimea.lenobl.ru/ https://crimea.lenobl.ru/ru/
302 CRAWLED https://gkmn.rk.gov.ru/ru/index https://gkmn.rk.gov.ru/check
302 CRAWLED https://mchs.rk.gov.ru/ru/index https://mchs.rk.gov.ru/check
302 CRAWLED https://meco.rk.gov.ru/ru/index https://meco.rk.gov.ru/check
302 CRAWLED https://mgsn.rk.gov.ru/ru/index https://mgsn.rk.gov.ru/check
302 CRAWLED https://must.rk.gov.ru/ru/index https://must.rk.gov.ru/check
302 CRAWLED https://mzem.rk.gov.ru/ru/index https://mzem.rk.gov.ru/check
302 CRAWLED https://safety.lenobl.ru/ https://safety.lenobl.ru/ru/
302 CRAWLED https://gkreg.rk.gov.ru/ru/index https://gkreg.rk.gov.ru/chec
302 CRAWLED https://gkvet.rk.gov.ru/ru/index https://gkvet.rk.gov.ru/chec
302 CRAWLED https://gkvod.rk.gov.ru/ru/index https://gkvod.rk.gov.ru/chec
302 CRAWLED https://intsm.rk.gov.ru/ru/index https://intsm.rk.gov.ru/chec
302 CRAWLED https://mkult.rk.gov.ru/ru/index https://mkult.rk.gov.ru/chec
302 CRAWLED https://mprom.rk.gov.ru/ru/index https://mprom.rk.gov.ru/chec
302 CRAWLED https://mtrud.rk.gov.ru/ru/index https://mtrud.rk.gov.ru/chec
302 CRAWLED https://mzhkh.rk.gov.ru/ru/index https://mzhkh.rk.gov.ru/chec
302 CRAWLED https://smpgo.rk.gov.ru/ru/index https://smpgo.rk.gov.ru/chec
302 CRAWLED https://intrud.rk.gov.ru/ru/index https://intrud.rk.gov.ru/ch
302 CRAWLED https://minfin.rk.gov.ru/ru/index https://minfin.rk.gov.ru/ch
302 CRAWLED https://msport.rk.gov.ru/ru/index https://msport.rk.gov.ru/ch
302 CRAWLED https://mstroy.rk.gov.ru/ru/index https://mstroy.rk.gov.ru/ch
302 CRAWLED https://mtrans.rk.gov.ru/ru/index https://mtrans.rk.gov.ru/ch
302 CRAWLED https://mzdrav.rk.gov.ru/ru/index https://mzdrav.rk.gov.ru/ch
301 CRAWLED https://www.uda.ke/site/ https://uda.ke/site/
301 CRAWLED https://uda.ke/ http://www.uda.ke/site/
301 CRAWLED http://kaygranger.com/ https://kaygranger.com/
301 CRAWLED http://kontradaya.org/ https://kontradaya.org/
301 CRAWLED http://www.uda.ke/site/ https://www.uda.ke/site/
301 CRAWLED http://www.lorenlegarda.com.ph/ https://www.lorenlegarda.com.
301 CRAWLED http://namfrel.org.ph/home/index.html https://namfrel.org.ph/
301 CRAWLED https://partidofederalngpilipinas.com// https://partidofedera
200 CRAWLED https://er.ru/
200 CRAWLED https://dse.md/
200 CRAWLED https://eji.org/
200 CRAWLED https://ijr.com/
200 CRAWLED https://kprf.ru/
200 CRAWLED https://rost.ru/
200 CRAWLED https://wipr.pr/
200 CRAWLED http://cbldf.org/
200 CRAWLED http://cbpmr.net/
200 CRAWLED http://fleen.com/
200 CRAWLED http://gomag.com/
200 CRAWLED http://vspmr.org/
200 CRAWLED https://chej.org/
200 CRAWLED https://famm.org/
```



	A	B	C	D	E	F	K	L	
1	url	bytes	depth	code	status	redirec	records	freque	se
292	https://blacklivesmatter.com/	3800227	2	200	CRAWLED		crawl	weekly	20
293	https://www.advocate.com/	4331705938	45	200	CRAWLED		crawl	weekly	20
294	https://www.amren.com/	18332689859	40	200	CRAWLED		crawl	weekly	20
295	https://keithself.com/	4171786	4	200	CRAWLED		crawl	weekly	
296	https://www.pakpips.com/web/wp-c	2097396	2	200	CRAWLED		crawl	weekly	
297	https://onsa.gov.pk/assets/document	994534	1	200	CRAWLED		crawl	weekly	
298	https://www.nrc.no/globalassets/pdf	3078024	2	200	CRAWLED		crawl	weekly	
299	https://iwps.org.af/wp-content/uploa	8879148	3	200	CRAWLED		crawl	weekly	
300	http://www.indusconsortium.pk/wp-	799390	3	200	CRAWLED		crawl	weekly	
301	http://www.indusconsortium.pk/wp-	2200802	1	200	CRAWLED		crawl	weekly	
302	https://dmw.gov.pk/storage/report/D	27305941	2	200	CRAWLED		crawl	weekly	
303	https://customs.gospmr.org/	6203012096	43	200	CRAWLED		crawl	weekly	
304	http://gsuda.gospmr.org/	2367425372	41	200	CRAWLED		crawl	weekly	
305	http://mgb.gospmr.org/	1518090851	41	200	CRAWLED		crawl	weekly	
306	https://mvdpmr.org/	11088265431	25	200	CRAWLED		crawl	weekly	
307	http://mincifra.gospmr.org/	452588274	12	200	CRAWLED		crawl	weekly	
308	https://mopmr.org/	7164620255	7	200	CRAWLED		crawl	weekly	
309	http://sk.gospmr.org/index.php/ru/	75919533	5	200	CRAWLED		crawl	weekly	
310	http://vspmr.org/	6078190836	43	200	CRAWLED		crawl	weekly	
311	http://en.vspmr.org/	2997446144	41	200	CRAWLED		crawl	weekly	
312	https://president.gospmr.org/	13191948180	13	200	CRAWLED		crawl	weekly	
313	http://cbpmr.net/	2027526135	40	200	CRAWLED		crawl	weekly	
314	https://www.electmikalwilliams.com	35451965	3	200	CRAWLED		crawl	weekly	
315	http://gov-pmr.org/government	13057960680	22	200	CRAWLED		crawl	weekly	
316	https://www.vlada.gov.sk//prime-mi	320515	1	200	CRAWLED		crawl	weekly	
317	https://www.prezident.sk/	85339392452	43	200	CRAWLED		crawl	weekly	
318	https://www.nsud.sk/	2693354647	41	200	CRAWLED		crawl	weekly	
319	https://www.nrsr.sk/web/	1.26355E+12	12	200	CRAWLED		crawl	weekly	
320	https://www.vlada.gov.sk//governme	2870323	4	200	CRAWLED		crawl	weekly	
321	https://www.ustavnysud.sk/aktualne	3486332919	12	200	CRAWLED		crawl	weekly	
322	https://ratinggroup.ua/	3866809625	9	200	CRAWLED		crawl	weekly	
323	https://munkschool.utoronto.ca/jacyl	26343835	46	200	CRAWLED		crawl	weekly	
324	https://www.unian.net/	3.32505E+11	42	200	CRAWLED		crawl	weekly	
325	https://www.mzv.cz/jnp/	3099532566	12	200	CRAWLED		crawl	weekly	
326	https://www.mvcr.cz/	65774727847	11	200	CRAWLED		crawl	weekly	
327	https://www.mpsv.cz/	43705736011	52	200	CRAWLED		crawl	weekly	

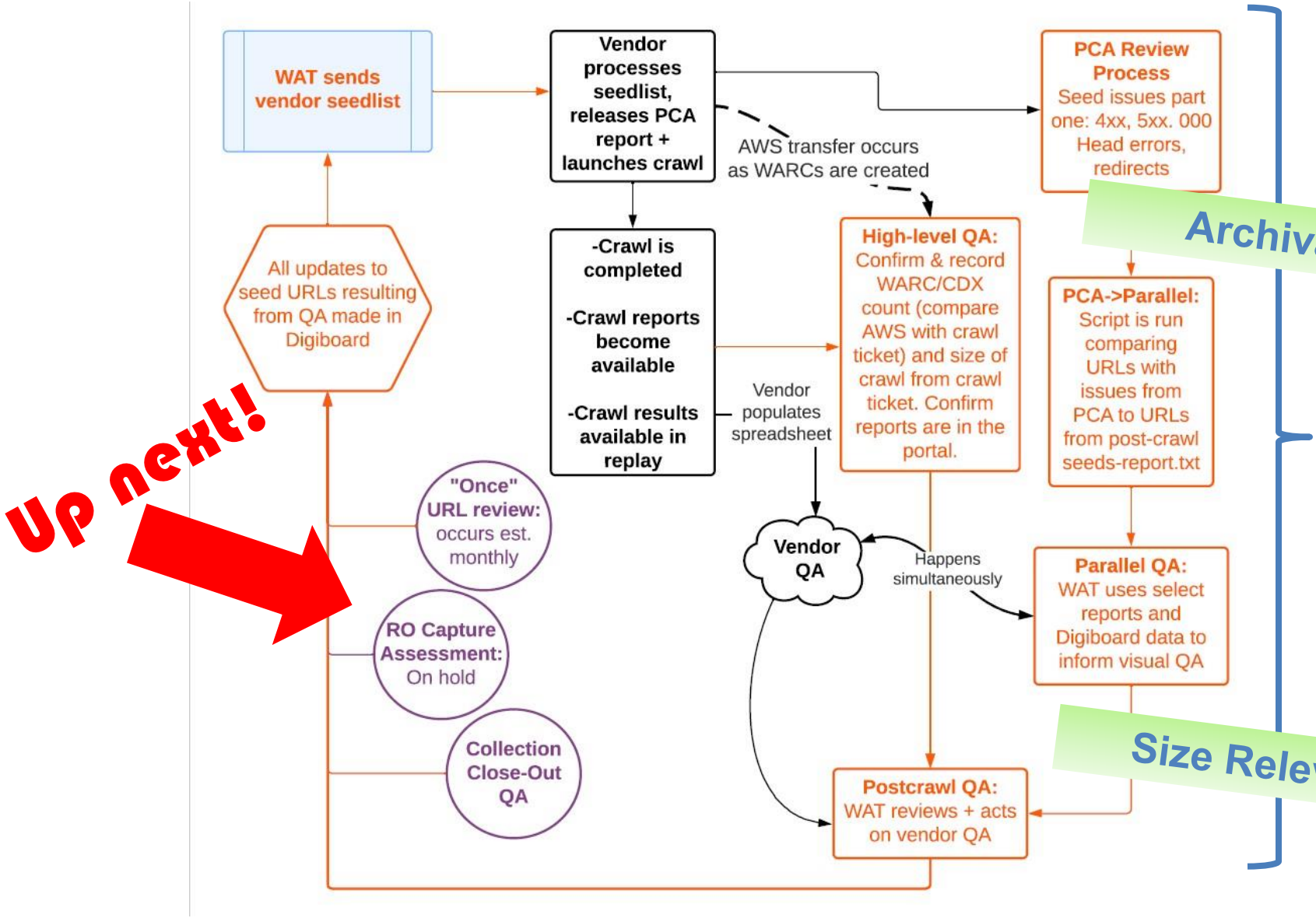
Targets:

1. Seeds with low capture bytes

2. Seeds with low capture depth

3. Newer or older records in Digiboard

Easier to: Sort by collection and find all response codes for a given domain.


WAT sends vendor seedlist
Vendor processes seedlist, releases PCA report + launches crawl
AWS transfer occurs as WARCs are created
PCA Review Process
Seed issues part one: 4xx, 5xx. 000 Head errors, redirects
Archivability
All updates to seed URLs resulting from QA made in Digiboard
-Crawl is completed
-Crawl reports become available
-Crawl results available in replay
High-level QA: Confirm & record WARC/CDX count (compare AWS with crawl ticket) and size of crawl from crawl ticket. Confirm reports are in the portal.
PCA->Parallel: Script is run comparing URLs with issues from PCA to URLs from post-crawl seeds-report.txt
Vendor populates spreadsheet
Up next!
"Once" URL review: occurs est. monthly
RO Capture Assessment: On hold
Collection Close-Out QA
Vendor QA
Happens simultaneously
Parallel QA: WAT uses select reports and Digiboard data to inform visual QA
Size Relevance
Postcrawl QA: WAT reviews + acts on vendor QA

Correspondence

Correspondence: degree of similarity, or resemblance, between the original website and the archived website

(a) Visual correspondence: similarity in appearance between the original website and the archived website

(b) Interactional correspondence: the degree to which a user's interaction with the archived website is similar to that of the original

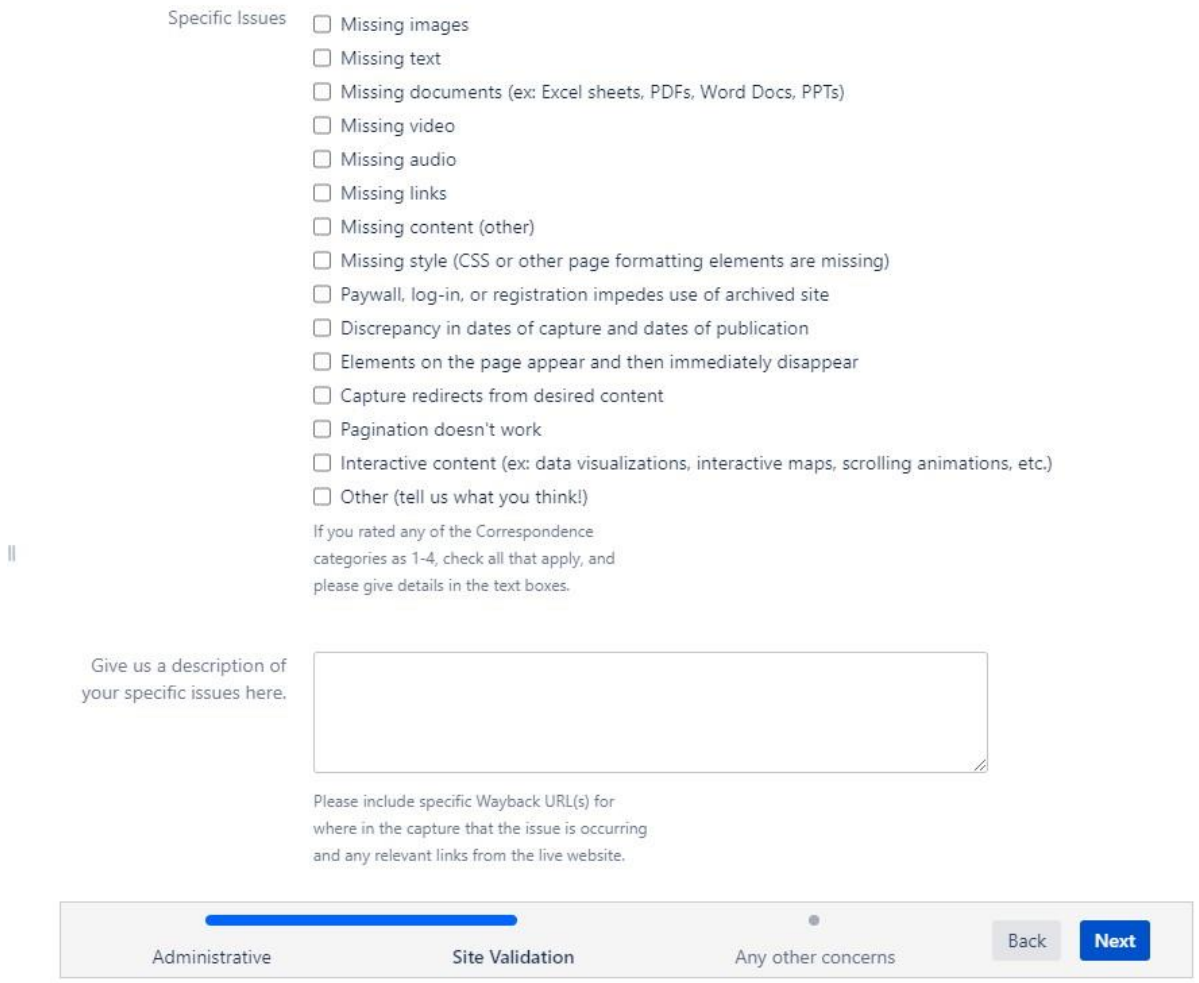
(c) Completeness: the degree to which the archived website contains all of the components of the original

Brenda Reyes Ayala
https://link.springer.com/article/10.1007/s00799-021-00314-x

Correspondence measurement rubric

Category	Ranking 1-5
Visual correspondence: similarity in appearance between the original website and the archived website	☐ 1 - unrecognizable - there is no similarity between the original website and the archived website ☐ 2 - barely recognizable - there is some similarity between the original and the archived website ☐ 3 - recognizable, but not exactly the same - the archived site is similar to the original website ☐ 4 - appears nearly perfect - the archived site looks mostly the same as the original website ☐ 5 - appears perfect - the archived site looks like a perfect replica of the original website at the time of the crawl
Interactional correspondence: the degree to which a user can interact with the archived website" have examples, like user can click navigation buttons, click on a link and it opens, scrolling works - basic interactions.	☐ 1 - unable to interact with any features of the archived website ☐ 2 - unable to interact with most of the features of the archived website ☐ 3 - able to interact with about half of the features of the archived website ☐ 4 - able to interact with most of the features of the archived website ☐ 5 - able to interact with all of the features of the archived website
Completeness: the degree to which the archived website contains all of the components of the original	☐ 1 - all components of the original website are missing from the archived website (all content missing) ☐ 2 - most components of the original website are missing from the archived website (most content missing) ☐ 3 - half of the components of the original website are present on the archived website (half content missing) ☐ 4 - most of the components of the original website are present on the archived website (some content missing) ☐ 5 - all components of the original website are present on the archived website (no missing content)

Common QA issues



Capture Assessment Form Beta

Created by Meghan Lyon, last modified on Apr 22, 2022

INSTRUCTIONS: Please use the rubric below to assess your captures. Fill out a separate form for each capture date that you review. Use the "view archive" link in Digiboard to access on-site replay. Reviewing the most recent capture would be ideal, in case any adjustments need to be made in Digiboard that will be informed by the live website.

Please review the known challenges page for expected challenges and the current status of dynamic content and social media capture.

A note on our method: Levels of correspondence are derived entirely from Brenda Reyes Ayala's grounded theory study of QA in web archives, specifically Ayala's categories of correspondence, which are defined as the "degree of similarity, or resemblance, between the original website and the archived website." This form represents the Web Archiving Team's application of this concept, interpreted to suit Library needs. Our goal is not only to have our captures assessed by subject specialists, but to accumulate data about all archived content from those who know the content best.

To receive a copy of your submission, please sign into Confluence before completing the form.

Site validation rubric

Category	Ranking 1-5 (worst to best)
"***Visual correspondence:*** *similarity in appearance between the original website and the archived website" Reyes Ayala, B. (2021)*	1 - unrecognizable - there is no similarity between the original website and the archived website 2 - barely recognizable - there is some similarity between the original and the archived website 3 - recognizable, but not exactly the same - the archived site is similar to the original website 4 - appears nearly perfect - the archived site looks mostly the same as the original website 5 - appears perfect - the archived site looks like a perfect replica of the original website at the time of the crawl
"***Interactional correspondence:*** *the degree to which a user's interaction with the archived website is similar to that of the original" Reyes Ayala, B. (2021). For example, navigation buttons function, click on a link and it opens, scrolling works, etc. Basic interactions.*	1 - unable to interact with any features of the archived website 2 - unable to interact with most of the features of the archived website 3 - able to interact with about half of the features of the archived website 4 - able to interact with most of the features of the archived website 5 - able to interact with all of the features of the archived website
"***Completeness:*** *the degree to which the archived website contains all of the components of the original." Reyes Ayala, B. (2021)* *Reyes Ayala, B. Correspondence as the primary measure of information quality for web*	1 - all components of the original website are missing from the archived website (all content missing) 2 - most components of the original website are missing from the archived website (most content missing) 3 - half of the components of the original website are present on the archived website (half content missing) 4 - most of the components of the original website are present on the archived website (some content missing) 5 - all components of the original website are present on the archived website (no missing content)

Grace Thomas – grth@loc.gov
Meghan Lyon – mlyon@loc.gov