Sustainable Web Archiving at Scale:
An Introduction
June 21-22, 2021

DPOE-N
Digital Preservation Outreach & Education Network

Decision Chart	Archive-It	Conifer	ArchiveWeb.page	Browsertrix Crawler
Software/Service Costs	Subscription	Free/Subscription Open Source	Free/Open Source	Free/Open Source
Skill Level	Novice-Intermediate	Novice-Intermediate	Novice-Intermediate	Advanced
Harvest/Capture	Automated	Manual	Manual	Automated
Harvest/Capture Time	Less time intensive	More time intensive	More time intensive	Less time intensive
Scale	Large scale (QA required)	Small scale	Small scale	Large scale
Quality Assurance	More time intensive for media-heavy sites	QA during capture, less time intensive	QA during capture, less time intensive	QA after capture, less time intensive
Digital Storage Infrastructure	Hosted cloud storage	Hosted cloud storage	Browser storage/hosting: not long term solution. ReplayWeb.can load via HTTP, S3, IPFS, and GoogleDrive storage.	Local storage
Embedded Media	Low amount/Low importance - Requires QA	Medium amount/High importance - Autoscroll & video auto-play w/ some QA	Medium amount/High importance - Autoscroll & video auto-play w/ some QA	High amount/High importance - Autoscroll & video auto-play w/ some QA

Emulated/Remote Browsers	NA	Yes	Chrome only	Chrome only
Social Media/Privacy	Low Importance- Lower fidelity/ no login access	High importance- Higher fidelity/login credentials required/can pose access obstacles	High importance - Higher fidelity/login credentials stored separately	High importance - Higher fidelity/credentials stored separately
Cataloging	Built-in template for Dublin Core fields (standard + custom fields) OAI-PMH crosswalk to OCLC WorldCat for collection-level records ArchivesSpace integration via WASAPI. Supports full text search of metadata and content.	Unstructured. Title, list, description fields. Embeddable in web pages. Supports full text search of title and URL fields of hosted collections.	Generates JSON indexing/ supports full-text search of all content in WACZ	Generates JSON indexing/ supports full-text search of all content in WACZ
Access	Online	Online	Online or Local	Local
Export Format	WARC	WARC	WARC or WACZ	WARC or WACZ