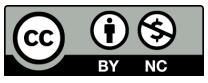
Recommendations for Harvesting Web Archives into Calisphere

Prepared by Matthew McKinley
California Digital Library
September 2020



https://bit.ly/csphere-web-archive-recs

Executive Summary **3**

Introduction **5**

Scope and Methodology **5**

Scope and Definitions 5

Methodology 6

Data Analysis and Findings **6**

Approaches of Other Regional/National Aggregators 6

Environmental Scan of Web Archiving Tools 7

Uptake of Archive-It by California Institutions 8

Archive-It Contextual Model, User Interface and OAI-PMH Feed 8

Conclusion 9

Implementation Recommendations **9**

1. Focus on Archive-It as Harvest Source 9

2. Represent Each Archive-It Collection as an Individual Calisphere Object 10

3. The isShownAt link Should Point Back to the Archive-It Collection Landing Page 10

4. The isShownBy Image Should be a Generic Icon Representing a Web Archive 12

5. The Default DCMI Type Metadata Value for Web Archives should be Interactive Resource 12

Proposed Web Archives Harvesting Pilot **13**

Pilot Tasks 13

Estimated Resources and Level of Effort 14

Estimated Timeframe 14

Future Exploration **15**

Appendix A: Assessment of Web Archive Harvesting and Display Strategies by Regional/National Aggregators **16**

Appendix B: Environmental Scan and Comparison of Web Archiving Tools **17**

Appendix C: Inventory of California Institutions with Web Archive Collections in Archive-It **20**

Existing Contributors 20

Potential New Contributors 21

Appendix D: Exploration of Archive-It Contextual Model, Interface and OAI-PMH Feed **22**

Archive-It Contextual Model and Interface 22

Archive-It OAI-PMH Feed 25

Collection-Level OAI-PMH Metadata 26

Seed-Level OAI-PMH Metadata 28

Executive Summary

Calisphere, a service provided by the California Digital Library, has to date aggregated over 1.7 million digital objects from institutions across California into a single platform for streamlined searching and browsing. These objects include images, text, audiovisual recordings and more--but they do not yet include an object which represents a growing corpus among current and future institutional Calisphere contributors: web archives.

A web archive is created by capturing a snapshot of a website in order to preserve and recreate the experience of interacting with that site. Using web archiving services such as Archive-It, many California institutions are capturing websites pertaining to their particular collecting focuses -- similar to how they are intentionally gathering and stewarding other unique archival resources -- and making them freely available for end users. Web archives fundamentally augment and relate to other unique digital collections already shared via aggregations such as Calisphere, and should be considered for inclusion alongside these existing collections. However, web archives are a different and more complex format than a text or image file, and this complexity has implications for how web archives might be sourced, organized, and displayed. This report is in many ways a feasibility study to see if web archives can be incorporated into Calisphere. To this end, four complementary areas of investigation were pursued:

- An assessment of web archive harvesting and display strategies by other regional/national aggregators
- An environmental scan and comparison of web archiving tools and related protocols for sharing metadata
- An inventory of California-based libraries, archives, and museums with web archives in the web archiving service Archive-It
- An exploration of the Archive-It contextual model, interface, and OAI-PMH feed

This investigation did not seek to gauge the feasibility of the complex work involved in rendering a web archive as an interactive object, but rather focused on the relatively more tractable task of presenting associated metadata and a representative image with a URL link pointing back to the web archive, in line with other harvested objects in Calisphere. Inventories of web archiving tools as well as web archives in other large-scale aggregators were created in order to understand their usage by California institutions and to determine their interoperability with the Calisphere harvesting infrastructure.

Due to its large California user base and sophisticated interface, Archive-It surfaced as the recommended source for harvesting web archives into Calisphere. That primary finding led to another key recommendation: due to the way web archives and "seed" URLs are displayed within Archive-it, each Archive-It web archive "collection" should be represented as a single Calisphere object, in order to provide the best experience for Calisphere end users. Additional recommendations focus on the details of incorporating web archives into Calisphere, including

the use of a generic icon to visually represent them, application of a default form/genre term ("Type" metadata), and specifications for the source URL link.

Having determined that harvesting web archives into Calisphere appears to be feasible, we then mapped out a potential pilot project for harvesting web archives collections. This plan identifies pilot partner criteria; project planning, design, and technical tasks; and provides estimates of staff time commitments and overall duration. We end the report with areas of potential future exploration and possible collaboration with the Archive-It team.

Introduction

This document presents results of an investigation into options for harvesting web archives and presenting them alongside other digital primary objects. This work was undertaken as part of the California Digital Library's (CDL) "Harvesting California's Bounty" 2019-2020 project, an initiative supported by the US Institute of Museum and Library Services under the provisions of the Library Services and Technology Act (LSTA), administered in California by the State Librarian. Our 2019-2020 project activities included an objective to explore and evaluate options for including web archives in Calisphere, an aggregation service provided by the CDL, that brings together cultural heritage materials from across California.

Web archiving involves capturing a particular website at a specific moment in time (along with any component files, as well as technical and descriptive metadata) in order to preserve the content and, ideally, allow future users to recreate the experience of accessing the website via specialized software. Recognizing the research value of the website as a medium for communication and publishing, many libraries, archives, and museums have begun capturing and collecting web archives as primary research objects.

These web archives collections should therefore be considered for inclusion alongside institutional research collections already captured and displayed via Calisphere and other aggregators. However, the interactive nature of web archives makes the process of capture and display substantially different from traditional research objects such as text, image, audio and video. Investigation was thus needed to determine if harvesting metadata records for web archives into Calisphere is even feasible and, if so, what form this harvesting might take.

To that end, we conducted an analysis of current web archiving tools and related literature, as well as an evaluation of the representation of web archives in other aggregations of cultural heritage objects. Out of this work, we have identified a set of changes needed within the current Calisphere harvesting infrastructure and user interface in order to best capture and provide access to web archives within the site.

Scope and Methodology

Scope and Definitions

This document seeks to identify potential web archive repositories, as well as the work needed on CDL's part to harvest and display descriptive metadata and a representative thumbnail for each web archive object--in much the same way Calisphere currently displays image, text, and most audiovisual objects from harvested data sources. It does not address what might be needed to interactively render and display web archives within the Calisphere itself, which is out of scope.

For this report, a web archive is defined as the object or set of objects resulting from capturing a web page's complete content, including any source code, stylesheets and embedded media, for the purpose of future research and/or re-enactment.

Methodology

This investigation included the following methodology:

- An assessment of web archive harvesting and display strategies by other regional/national aggregators
- An environmental scan and comparison of web archiving tools and related protocols for sharing metadata
- An inventory of California-based libraries, archives, and museums with web archives in Archive-It
- An exploration of the Archive-It contextual model, interface, and OAI-PMH feed

Data Analysis and Findings

Presented below are brief summaries of the data gathered and analyzed for each area listed in the previously described Methodology section. This analysis, along with data presented in the corresponding appendices, supports the implementation recommendations in this report.

Approaches of Other Regional/National Aggregators

We evaluated how other US regional and national aggregators of cultural heritage objects currently harvest and display web archives in order to surface any best practices or strategies which could be adopted by Calisphere. In order to be included for assessment, aggregators needed to include web archives that were 1) publicly accessible, 2) finalized, and 3) interactive.

By *publicly accessible*, we mean those web archives that are openly available for searching, browsing, and viewing, without restrictions. By *finalized*, we mean captured and discrete web

archives; this excludes records linking to live web URLs[1]. By *interactive*, we mean the end user must be able to click through to a web-based interface where they can explore the web archives; this excludes static "placeholder" records, or packaged files containing raw code[2]. Aggregators that met these three criteria were then evaluated by their required/optional record-level metadata, level of web archives representation (collection vs. seed URL), and existence of a representative thumbnail, as reflected in Appendix A.

Unlike CDL, the majority of DPLA service hubs choose to surface their aggregated content exclusively through DPLA and do not have their own dedicated user interface. The two hubs with dedicated interfaces that also contained records for websites, Mountain West Digital Library [3] and Portal to Texas History[4], link to live web URLs, not web archives, and are thus excluded from this comparison. This left only national-level aggregators to be assessed: DPLA (US), DigitalNZ (New Zealand) and Trove (Australia). Europeana does not host web archive objects, so could not be included. Even within this limited comparison set, substantial variance across all factors exists, including required vs. optional metadata, representing web archives at the collection or object/seed level, and strategies for visually marking web archives via a thumbnail. In short, there is no overarching consensus on how to best represent harvested web archives in aggregations.

Environmental Scan of Web Archiving Tools

We continued with a comparison of the existing tools for creating web archives, evaluating their features and usage to determine their feasibility as potential sources for harvesting web archives into Calisphere (see Appendix B). The list of compared tools was sourced from OCLC Web Archiving Metadata Working Group's 2018 "Review of Harvesting Tools" report[5]. Information from this report was augmented by further analysis of each tool's interface and documentation. Next, we conducted an environmental scan to determine if any newer tools had become available since 2018. None, however, had sufficient infrastructure for Calisphere harvesting.

Each candidate tool was analyzed according to two criteria essential for Calisphere harvesting: 1) the existence of a publicly accessible digital collection environment, including an object view page with sufficient context for each object as well as search and browse functionality, and 2) the ability to support harvesting at a collection and/or object level. Additionally, data was gathered on any California-based library, archive, or museum (LAM) using the tool for web archiving, along with a narrative description of the tool and how it is deployed. Taken together, these factors determine the feasibility of harvesting web archives from a given tool into Calisphere, and inform the implementation recommendations in this report.

---

[1] Example from the Digital Library of Georgia https://dlg.usg.edu/record/geh_ohvhp
[2] Example from the Connecticut Digital Archive https://ctdigitalarchive.org/islandora/object/30002%3A22204981
[3] https://mwdl.org/
[4] https://texashistory.unt.edu/
[5] https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata/harvesting-tools.html

The analysis revealed that not all web archiving tools prioritize end user interaction or third party harvesting. Of the eight tools evaluated, only one tool, Archive-It met both of Calisphere's requirements and, as discussed more fully in the next section, Archive-It is also more extensively used by a significant number of California-based LAMs (see Appendix C).

Uptake of Archive-It by California Institutions

While Archive-It meets Calisphere's initial core requirements, its value as a harvesting source is dependent upon its uptake by California institutions for the capturing web archives. To this end, an inventory of California institutions with content in Archive-It (see Appendix C) was assembled, using Archive-It's complete List of Collecting Organizations. For every California-based institution, we noted details such as number of collections/seeds, level of description, any potentially unique or interesting collections, and whether the institution was an existing Calisphere contributor.

Of the 37 institutions surveyed, 25 are existing Calisphere contributors. Although there are a few outliers, including UCLA with 58 web archive collections and San Francisco Public Library with 356, most institutions had somewhere between one and twenty web archive collections in Archive-It, with the majority having fewer than ten. The level of description varied; some collections and seeds were very well described while others had almost no metadata, and some collections at the same institution had very different levels of description.

Archive-It Contextual Model, User Interface and OAI-PMH Feed

The final assessment area focused on the Archive-It contextual model and user interface, comparing metadata displayed there to what is made available via OAI-PMH -- the web protocol most often used for harvesting content into Calisphere.

The Archive-It contextual model and user interface presents web archives within four hierarchical layers:

1. A top-level institution landing page, containing all of an Institution's public collections of web archives
2. A collection landing page, containing (usually multiple) entries consisting of a URL link and metadata for each captured web archives, otherwise known as a "seed"
3. A seed landing page within the Internet Archive's Wayback Machine web archive access service[6], containing all captures of a particular seed arranged by date
4. An interactive display of each particular capture of a seed, where the end user may interact with the object as if they were browsing the page in real time.

Each of these layers is described and explored in Appendix D, with corresponding screen captures for illustration.

[6] https://archive.org/web/

Archive-it's OAI-PMH feed was evaluated against the Archive-It display. In general, the OAI-PMH output matches what appears in the Archive-It interface, although some metadata is normalized and other potentially valuable technical metadata generated by Archive-It is left out.[7] Two other important findings also influenced our implementation recommendations for the *isShownAt* record URL link and representative thumbnail, respectively:

- The OAI-PMH results for seed level Archive-It objects provide a URL link back to the *seed landing page*, which lacks important seed-level metadata and context, instead of linking back to the seed metadata on the "Sites" tab of the *collection landing page*
- The OAI-PMH results for collection level Archive-It objects *do not* contain a URL pointing to a representative thumbnail characterizing the web archive collection

Implementation Recommendations

Based on analysis of the web archiving tools and web archives display interfaces gathered above, representing web archives in Calisphere does appear to be feasible. A logical next step would be to launch a pilot project, both to test this assessment in practice and to support a focus on the harvesting mechanics -- which service to harvest web archives from, how best to organize and display them in the Calisphere interface, etc. The following recommendations are offered as guidelines for a potential pilot project and, should the project be successful, any future harvesting of web archives into Calisphere. The first two recommendations are key, as they will determine how web archives are harvested into Calisphere and how they are organized for users in the Calisphere interface. The remaining recommendations follow from these and focus on how best to implement technical and descriptive measures in order to align web archives with the other cultural heritage objects gathered under Calisphere.

1. Focus on Archive-It as Harvest Source

As discussed in the data analysis section above, though many tools exist for capturing and displaying web archives, only one platform emerged as an appropriate harvesting endpoint: the Internet Archive's Archive-It service. There are two key reasons:

- Use: The overwhelming majority of current and potential Calisphere contributors engaged in capturing web archives solely use Archive-It for capture and display. Further, Archive-It is heavily utilized by US libraries, archives, and museums.
- Feature Requirements: As captured in Appendix B, other web archiving tools do not have the publicly searchable, collection-based access interface necessary for Calisphere harvesting and object-level linking

Archive-It is thus the best harvesting source for Calisphere, even taking into account the challenges it presents for metadata capture and display, including custom metadata fields, non-standardized levels of description, and a lack of descriptive metadata at the seed landing

[7] This is discussed more in both Appendix D and the Future Exploration section

page. The remaining recommendations focus on the key decision points associated with using Archive-It.

2. Represent Each Archive-It Collection as an Individual Calisphere Object

After examining how Archive-It organizes crawled seeds and collections, as well as how web archives are represented in other repositories and aggregators, two candidate models for representing web archives in Calisphere emerged:

- Model 1: represent an Archive-It collection as an individual Calisphere object
  - *Pros*: there is typically more metadata at the Archive-It collection level, to support discovery and interpretation by end users.
  - *Cons*: we would not be able to incorporate Archive-It seed-level information into the Calisphere object; the seed-level information -- which in many cases may simply be a URL -- can also convey important information to end users (who may be interested in searching for a web archive, by a specific seed URL).

- Model 2: represent an Archive-It seed as an individual Calisphere object
  - *Pros*: this more closely mirrors Calisphere's model for collections and objects, where the Archive-It seed would be analogous to a Calisphere object.
  - *Cons*: there is typically less metadata at the Archive-It seed level (only a seed URL to harvest from is required). Archive-It collection metadata doesn't directly map to the Calisphere collection metadata model (some metadata may have to be subsumed within a narrative description).

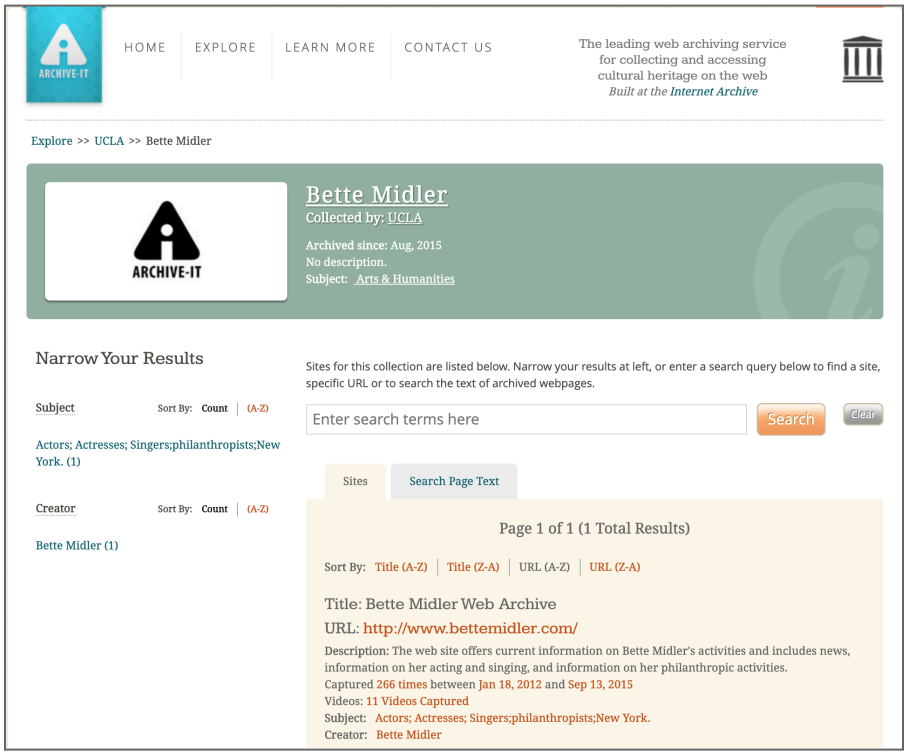
As shown, both models have strengths and weaknesses. Ultimately, **we recommend Model 1 due to the general lack of metadata and context at the seed landing page**, as discussed further in Recommendation #3 below. Thus, each Archive-It collection will constitute a Calisphere object, no matter the number of seeds, and will be gathered under a single 'Web Archives' collection in Calisphere for each institution.

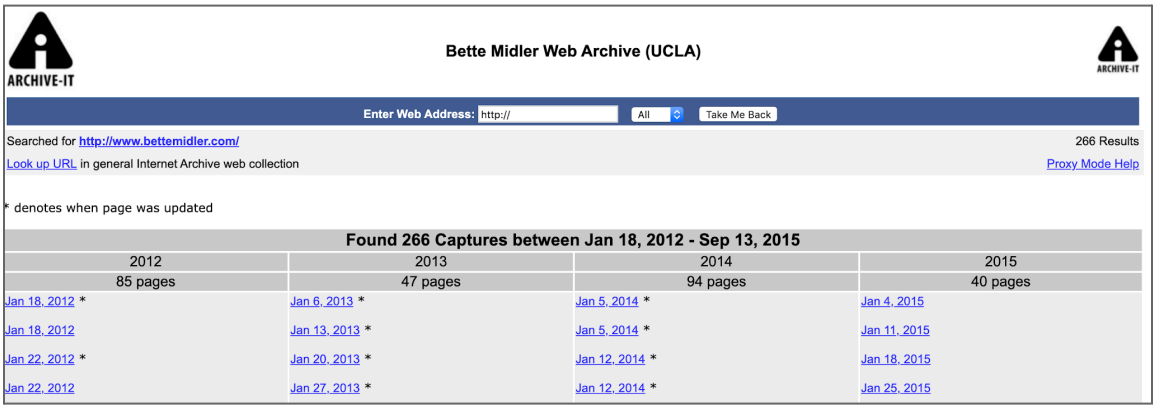
3. The *isShownAt* link Should Point Back to the Archive-It Collection Landing Page

As previously noted, any given Calisphere object holds a URL link within an *isShownAt* field, pointing users back to a representation of the object at the harvesting endpoint. For a web archives' *isShownAt* object page URL link, both the Archive-It collection landing page[8] and seed landing page[9] were considered. Ultimately, **the Archive-It collection landing page is recommended for the *isShownAt* URL link in Calisphere** for the following reasons, illustrated by the image comparison below:

[8] Example collection landing page (UCLA): https://archive-it.org/collections/4949
[9] Example seed URL capture page (UCLA): https://wayback.archive-it.org/5974/*/http://www.garrysouthgroup.com/

- The collection landing page contains context to help users make sense of the web archives that the seed capture page lacks, including content format, additional explanatory and contact information for Archive-It, etc.
- The seed landing page is sparse and potentially confusing for users and is devoid of context. It also lacks most of the seed-level descriptive metadata displayed on the collection landing page.





Comparison of collection landing page seed metadata and seed landing page, UCLA Bette Midler Web Archives https://archive-it.org/collections/6172

As a relatively new digital object, web archives are potentially unfamiliar to many users and there is no analog version to compare to like a text or image object.[10] Given the learning curve

[10] "Archived web data often is provided in a way that exceeds the limits of users' technical knowledge, constituting a widespread barrier to use" and "A need for user support services derives from the complexity of accessing and using web archives" are two conclusions drawn from the User Needs literature review performed in *Descriptive Metadata for Web Archiving* Report (OCLC 2018), 12.

required to understand these objects, it makes more sense to point Calisphere users toward the page with more clarifying context and description.

4. The *isShownBy* Image Should be a Generic Icon Representing a Web Archive

Each Calisphere object includes a thumbnail representation of the object at the harvesting endpoint; the thumbnail is displayed in Calisphere search and browse results, and also while viewing the specific object in Calisphere. The thumbnail is referenced through a URL, stored in an *isShownBy* field.

Within Archive-It, an institution may optionally display a custom image on either their institutional landing page and/or collection/s landing page in Archive-It[11]. However, the majority of institutions and collections surveyed for this report did not have custom images -- and also critically, URL links to the images are not surfaced through either Archive-It's OAI-PMH or API endpoint. We would thus need to work out another way to capture this image for each collection, which would be another layer of complexity in service of a feature only a minority of institutions actually use. Thus, **it is recommended to create and use a generic icon representing a web archive, whether or not a custom image is assigned to the Archive-It Institution or Collection**.

Calisphere currently utilizes generic icons based for particular content formats, such as audio recordings. For example, see https://calisphere.org/collections/27299/.

5. The Default DCMI Type Metadata Value for Web Archives should be *Interactive Resource*

To ensure a clear and uncluttered user experience when searching/browsing/otherwise interacting with the Calisphere site, all objects have "type" metadata that reflect their form/genre. The "type" metadata values are normalized to one of twelve standard type values from the Dublin Core Metadata Initiative's (DCMI) Type Vocabulary[12] before final publication to Calisphere.

Based on the descriptions given within the DCMI Type Vocabulary, there are three type values that could potentially be applied to web archives:

- *Collection*: an aggregation of resources
- *Service*: a system that provides one or more functions
- *Interactive Resource*: a resource requiring interaction from the user to be understood, executed, or experienced

[11] Example customized institutional image from San Francisco Public Library: https://archive-it.org/organizations/160
Example customized collection image from UCLA: https://archive-it.org/collections/4949
[12] https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/

It is recommended that *Interactive Resource* be applied as the default "type" metadata value for web archives, as it gets at the basic fundamental nature of a website as a resource that is *interacted* with by an end user, and allows for dynamic two-way communication between the user and web server.

Proposed Web Archives Harvesting Pilot

While harvesting and displaying web archives in Calisphere appears feasible according to our analysis, we propose a small-scale pilot project (to be prioritized and initiated, based on other CDL development work) to test this assessment and evaluate results with contributor engagement. A pilot project would offer an opportunity to work out any technical issues with harvesting and displaying web archives and, if successful, would serve as a model for future, production level Calisphere web archive harvesting. The work of this proposed pilot is broken down into tasks, as informed by the recommendations above, along with estimates of needed staff/time commitments and overall project timeline.

Pilot Tasks

1. Review Archive-It policies and best practices; identify any potential gaps or discrepancies between Archive-It and CDL policies and best practices, and determine approaches to address the discrepancies.
2. Identify three candidate institutions from Existing Contributors list in Appendix C, according to following criteria:
    - One University of California (UC) institution; two institutions from outside UC
    - Must already be established as Calisphere contributors
    - Must have less than 10 total published web archives collections
    - Must be well described at collection level, with ample metadata (ideally also well described at seed level, with ample metadata).
3. Perform user experience-based analysis and design
    - Investigate and analyze user experience requirements
    - Create wireframe mockup images of Calisphere interface to share with potential pilot institutions, including:
        - '[institution name] Web Archives' collection landing page
        - web archives record landing page
        - web archives displayed in search/browse results
4. Communicate with candidate institution:
    - Identify key contact
    - Send introductory message and arrange call to explain project
    - Confirm institution's interest and participation
5. Coordinate with Calisphere UX programmer and/or Front End Web Developer:
    - Design generic web archives thumbnail image for display in Calisphere interface
    - Determine where and how to integrate thumbnail into Calisphere search/browse results and record page
6. Develop code to harvest web archives into Calisphere:

    - Develop fetcher file to capture and save metadata records via Archive-It OAI-PMH interface
    - Develop mapper file to normalize metadata according to Calisphere Metadata object model
7. Iterate code and thumbnail display on Calisphere Test; refine as necessary
8. Publish pilot collections to Calisphere Production and notify institutions. Establish a re-harvest schedule with contributors, as appropriate
9. Publicize collections via contributor and end-user newsletter
10. Monitor and evaluate interest level from potential contributors:
    - Establish a mechanism to gauge broader interest from contributors on sharing web archive collections with Calisphere (e.g., contact us or post to Freshdesk community forum)
    - Track usage of pilot collections in Google Analytics for 3-6 months
    - Evaluate usage relative to other items in candidate contributors' collections
11. Re-evaluate pilot project and determine whether to expand harvesting beyond candidate sources

Estimated Resources and Level of Effort

- *OAC/Calisphere Service & Outreach Manager, including User Experience Design Service team support* - 60 hours
    - Coordinate and track project work
    - Investigate and analyze user experience requirements
    - Identify candidate institutions and collections for pilot harvesting
    - Communicate with and secure approval from candidate institutions
    - Develop content highlighting availability of collections for contributor and end-user newsletter
- *Metadata Harvest Programmer* - 10 hours
    - Develop code to harvest web archives via Archive-It OAI-PMH interface
- *UX programmer* or *Front-End Web Developer* - 36 hours
    - Design generic web archives thumbnail image
    - Implement changes based on user experience requirements analysis
    - Integrate generic web archives thumbnail image into Calisphere interface

Estimated Timeframe

Conservative estimate is a 6 month timeframe, from start to finish:

- Months 1-2:
    - Produce wireframe mockup images
    - Identify candidate institutions and collections
    - Secure candidate institution participation
- Months 3-4:
    - Design generic web archives thumbnail
    - Integrate generic web archives thumbnail into Calisphere interface
    - Develop code to harvest web archives via Archive-It OAI-PMH interface

- Month 5:
    - Iterate code and thumbnail on Calisphere Test interface; refine as necessary
- Month 6:
    - Publish pilot collections to Calisphere Production and notify institutions
    - Publicize collections via contributor and end-user newsletter

Future Exploration

These recommendations outline an initial plan for harvesting Archive-It web archives -- based on their current form and structure -- into Calisphere. However, many of the sources consulted for this report emphasize the importance of preserving the *context* of web archives creation and capture (e.g., collecting scope, rationale for capturing the web archive)[13]. Preserving context and provenance through descriptive metadata is especially important for an aggregation service such as Calisphere, where objects are often discovered and interacted with in a setting apart from their local repository and institutional context. In order to better capture the creation context and dynamic nature of web archives in Calisphere, we've identified several areas of possible future exploration and coordination with the Archive-It team:

- Advocate for enhancements to the overall Archive-It UI and navigation, based on recent user-based research.[14] This includes providing clearer pathways to terms of use and policies and help-related documentation. This also includes a consistent representative image for a given web archives collection, which could then be used as a representative thumbnail image for harvested web archives in aggregations such as Calisphere, DPLA, etc.
- A more informative Archive-It seed landing page[15], incorporating:
    - Any seed-level metadata. (Currently, seed-level metadata only appears on the Archive-It collection landing page, not on the seed landing page)
    - UX-focused testing and redesign to visually align the seed landing page with institution and collection landing pages
- More technical metadata exposed in Archive-It OAI-PMH results:
    - Include date of original capture at the collection level (Example: *Archived Since: Nov, 2007*) -- perhaps in a *dc:date.created* field
    - Include capture date range and frequency at the seed level (Example: *Captured 10 times between Apr 26, 2012 and Nov 8, 2019*) -- perhaps in a *dc:date.modified* field

---

[13] *Descriptive Metadata for Web Archiving* Report (OCLC 2018), 7. https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations.pdf; "Web Archives and You/ Web Archiving and Us" (Wickner, Code4Lib 2018), 9. https://osf.io/xp4mh/; "Why, what, when and how? Curatorial decisions and the importance of their documentation" (Royal Danish Library, IIPC 2019) http://netpreserve.org/ga2019/wp-content/uploads/2019/07/IIPCWAC2019-SABINE_SCHOSTAG-Curatorial_decisions_and_the_importance_of_their_documentation.pdf

[14] "Sowing the Seeds for More Usable Web Archives: A Usability Study of Archive-It" (The American Archivist, 2019) https://doi.org/10.17723/aarc-82-02-19

[15] "Sowing the Seeds for More Usable Web Archives"

- Include a URL link to custom thumbnails for web archives at the collection level, where present

Appendix A: Assessment of Web Archive Harvesting and Display Strategies by Regional/National Aggregators

Service	Required metadata	Other metadata	Example web archive	Collection and/or seed level representation	Thumbnail representation
DPLA	Title[16]	Combination of simple and qualified Dublin Core with Europeana Data Model for technical metadata[17]	https://dp.la/item/edd0c72fa3fe2411de0cde3bb82aaa5f?q=%20%09Federal%20Depository%20Library%20Program%20Web%20Archive	Collection	Thumbnail of collecting institution's logo
DigitalNZ	Category (type), collection, content partner, description, title, landing URL, thumb URL, usage rights	Combination of Dublin Core and DigitalNZ custom schema	https://natlib.govt.nz/records/38194357	Seed	Generic thumbnail icon representing a web archive
Trove	Auto-created: Title, Date Archived, URL	Description, etc. AACR2 Rules http://pandora.nla.gov.au/manual/kincat.html	https://webarchive.nla.gov.au/awa/20191107020951/https://www.myagedcare.gov.au/	Seed	Screenshot of latest capture
Europeana	*N/A (web archives not supported)*				

[16] DPLA Metadata Application Profile, Appendix B
https://drive.google.com/file/d/1fJEWhnYy5Ch7_ef_-V48-FAViA72OieG/view
[17] DPLA Metadata Application Profile, Page 4

Appendix B: Environmental Scan and Comparison of Web Archiving Tools

Tool	Description	Example collections	Example/s of California-based users	Supports indexing and public browse/search?	Supports harvesting at collection and/ or seed level?	How it's deployed
Archive-It	An administrative tool with a user-friendly interface that allows partners to collect, describe, manage and publish web archives collections for end users to search, browse, and interact with archived web resources. Archive-It captures web content using IA's open-source web crawler Heritrix; and preserves partners' web archives by storing (and having the ability to download) its WARC files that contain web-captured content in IA's digital repositories.	Collection: https://archive-it.org/collections/9302 Object (seed): https://wayback.archive-it.org/9302/*/http://kurdistantv.net/ku/news/kurdistan	Many. See Appendix C: Inventory of California Institutions with Web Archive Collections Available for Harvesting via Archive-It	Yes https://archive-it.org/	Yes	Institutions set up one or more 'collections' by theme or collection area. Each collection contains one or more 'seed' URLs, which contain all captures of the website hosted at that URL according to a frequency and time span set by the collection creator. Collections and seeds may both have descriptive metadata, and all metadata and collection/seed titles/URLs are indexed and searchable/browseable from Archive-It interface.
HTTrack	A capture tool that uses a browser utility to download a website to an off-line directory, saving content to a folder hierarchy in a way that mirrors the original website structure. HTTrack can also update an existing mirrored site.	N/A (no public online access interface)	None	No	No	Designed for offline preservation and/or custom site mirroring, HTTrack **has no mechanism for users to make their 'saved' sites publicly available** beyond direct 'mirroring'--i.e. replicating a site exactly on another server configuration.
Memento Suite	A suite of tools and protocol intended to facilitate access to archived versions of web content. The Memento 'Time Travel' service searches across an aggregation of web archives resources for an archived site at a specific point in time. The Memento protocol allows http content negotiation, such as the ability to navigate between different versions of a web resource based on the capture/crawl date. This suite of tools is more about the	Time Travel: http://timetravel.mementoweb.org/ Searches these web archives repositories: http://timetravel.mementoweb.org/about/	Stanford http://mementoweb.org/depot/native/stanfordwebarchive/	Yes	No	Memento web archives packages are in some ways more sophisticated than Archive-It captures. Such as the 'reconstruct' service, which attempts to source and display contemporaneous linked elements/components of an archived page in order to give a 'true' view of that page at a certain point in time http://timetravel.mementoweb.org/about/. However, as it is used by Archive-It and other services, **Memento is really**

	discovery of archived sites than crawling new ones.					**a protocol/ framework for accessing a web archive rather than an end-user facing tool**.
Netarchive Suite	a web archiving software package, developed primarily by and for several European National Libraries, that can be used to plan, schedule and run web harvests of parts of the Internet. NetarchiveSuite is divided into several modules, including a harvester module, an archive module, and an access module.	Royal Danish Library (not publicly accessible) http://netarkivet.dk/adgang/ Royal Austrian Library (not publicly accessible) https://webarchiv.onb.ac.at/	None	Yes	Unclear	Developed by Royal Danish Library to harvest Danish web since 2005; Developed and maintained by RDL along with National Libraries of France, Austria, Spain & Sweden. Uses Heretrix as a crawling mechanism (Internet Archive).
SiteStory	A transactional web archiving service that is installed on a web server and captures every version of a resource as it is being requested by a web browser. The resulting archive is effectively representative of a web server's entire history, although versions of resources that are never requested by a browser will also never be archived. As a browser requests a resource published by a server SiteStory is enabled on, that resource is delivered to the browser but also pushed into the archive--i.e. usage of a page by a web browser triggers the archiving.	N/A (no public online access interface)	Unclear--it's a capture tool that generates WARC files for inclusion in Wayback/Archive-It, so some CA institutions may be using it	No	No	Needs to be installed on the web server it is archiving, effectively a technical 'opt in' and not feasible for a majority of collecting institutions. Not necessarily designed for public-facing access; more for capturing all website activity/transactions as required by mandate. Access possible by exporting as WARC and uploading to Wayback https://mementoweb.github.io/SiteStory/getStarted.html
Web Archive Discovery	indexes and parses WARC/ARC files into JSON before posting resulting records to an Apache SOLR server to facilitate full-text indexing. The SOLR server can then be queried by a front-end GUI. **Web Archive Discovery is intended to be middleware**, and users can put their own interface on top of SOLR. This could be useful for end users with technical skills, but **it is not intended to be an end-user tool**.	https://www.webarchive.org.uk/wayback/en/archive/10000101000000/http://ansionnachfionn.com/	None	No	Yes, via SOLR query	Used by UK Web Archives https://www.webarchive.org.uk/en/ as well as a number of Canadian institutions via WARCLight (EX: https://utoronto.archivesunleashed.org/). Requires customized interface to interact with SOLR server.

Web Curator Tool	An open-source workflow management application designed to allow non-technical users to manage the selective web archiving process, but with no native access interface. Harvesting workflow comprises a series of specialized tasks to support web acquisition and description including: permission/authorization, selection/scoping/scheduling, harvesting, quality review and archiving. **WCT does not attempt to be a digital repository, access tool, catalog or records management system**.	N/A (no public online access interface)	None	No	No	Meant as an administrative interface for collectors to manage web archiving processes; no native public user interface. Wayback recommended as access mechanism
Webrecorder	A human-centered tool to archive web content through interactive browsing, capturing the exact sequence of navigation through a series of web pages or digital objects and preserving the unique experience of an individual user at a moment in time. Webrecorder creates high-fidelity, interactive, contextual archives of social media and other dynamic content, such as embedded video and complex JavaScript.	https://webenact.rhizome.org/	Unclear--individual creators from 'webenact' collection may be California based, but no obvious institutional connection	No	No	"Webrecordings" can be exported as WARCs, but no searchable index of these type of web recordings exist--only a few browseable collections on the rhizome.org site (https://webenact.rhizome.org/).

Appendix C: Inventory of California Institutions with Web Archive Collections in Archive-It

(gathered from https://archive-it.org/explore)

Existing Contributors

- 100,000 Poets https://archive-it.org/organizations/565 (Stanford)
- 2014 Congressional Election Cycle https://archive-it.org/organizations/775 (Stanford, UCB, et al)
- Academy of Motion Pictures Margaret Herrick Library https://archive-it.org/organizations/824
    - 10 collections
- CalPoly SLO https://archive-it.org/home/calpoly
    - 8 collections, well described
- CSU Chico https://archive-it.org/organizations/1321
    - 1 collection, many seeds
- CSU Fullerton https://archive-it.org/organizations/1172
- CalTech https://archive-it.org/home/caltech
    - 1 collection, many seeds
- Hoover Institute https://archive-it.org/organizations/294
    - 2 collections, well described
- Humboldt State https://archive-it.org/organizations/1479
    - 3 collections, including Indigenous Communities of Northern California
- Los Angeles County Metro Transportation Library https://archive-it.org/organizations/1502
    - 4 collections
- Los Angeles Philharmonic Association https://archive-it.org/organizations/1359
    - 6 collections, well described
- Loyola Marymount https://archive-it.org/organizations/1361
    - 2 collections, minimal metadata
- Pepperdine University https://archive-it.org/organizations/598
    - 2 collections, only one with metadata/seeds
- San Francisco Public Library https://archive-it.org/organizations/160
    - 356 collections, very well described
- UC Berkeley
    - Nasa Wavelength https://archive-it.org/organizations/952
        - 2 collections, many seeds
- UCSF Industry Documents Library https://archive-it.org/home/industrydocuments
    - 6 collections, well described
- UCSF https://archive-it.org/organizations/986
    - 8 collections, well described
- UCLA https://archive-it.org/organizations/877
    - 58 collections, different levels of metadata

- UC Berkeley https://archive-it.org/organizations/985
    - 20 collections, different levels of metadata
- UC Merced https://archive-it.org/organizations/983
    - 2 collections, minimal metadata
- UC San Diego https://archive-it.org/organizations/982
    - 25 collections, well described
- UC Santa Cruz https://archive-it.org/organizations/984
    - 5 collections, minimal metadata
- UC Davis https://archive-it.org/organizations/950
    - 9 collections, minimal metadata
- UC Irvine https://archive-it.org/organizations/947
    - 8 collections, well described
- UC Santa Barbara, https://archive-it.org/organizations/948
    - 5 collections, minimal metadata

Potential New Contributors

- 99 Antenna (Sausalito, CA) https://archive-it.org/organizations/1491
    - 1 collection, minimal metadata
- CAPE Charter School https://archive-it.org/organizations/497
    - 17 collections, very well described
- City of San Francisco https://archive-it.org/organizations/571
    - 12 collections, minimal metadata
- Getty Research Institute https://archive-it.org/organizations/1157
    - 2 collections, active
- IT History Society https://archive-it.org/organizations/416
    - 1 collection, well described
- Los Angeles County Library https://archive-it.org/home/lacountylibrary
    - 2 collections, minimal metadata
- Miramonte High School https://archive-it.org/organizations/299
    - 6 collections, well described
- Mount San Antonio College https://archive-it.org/home/mtsac
    - 2 collections, minimal metadata
- Occidental College https://archive-it.org/organizations/425
    - 4 collections, well described
- Pacific Union College https://archive-it.org/organizations/1025
    - 5 collections, well described
- San Diego Public Library https://archive-it.org/organizations/1314
    - 3 collections, well described
- San Jose State University, School of Information https://archive-it.org/organizations/853
    - 55 collections, very well described
- Sonoma Academy (K-12) https://archive-it.org/organizations/486
    - 3 collections, minimal metadata
- Southern California Association of Law Libraries https://archive-it.org/organizations/1128
    - 1 collection, minimal metadata

- Stanford
    - Graduate School of Business https://archive-it.org/home/stanfordgsb
        - 1 collection, well described
    - University Archives https://archive-it.org/organizations/933
        - 56 collections, well described
    - Humanities Lab https://archive-it.org/organizations/198
        - 8 collections, well described
    - Law Library https://archive-it.org/organizations/1055
        - 9 collections, well described
    - EAL/PCS https://archive-it.org/organizations/925
        - 1 collection, minimal metadata
    - University Libraries https://archive-it.org/organizations/1041
        - 4 collections, minimal metadata
    - Social Sciences Research Group https://archive-it.org/organizations/159
        - 14 collections, well described
- UCOP Communications https://archive-it.org/organizations/949
    - 16 collections, minimal metadata
- UC Libraries https://archive-it.org/organizations/898
    - 11 collections, minimal metadata

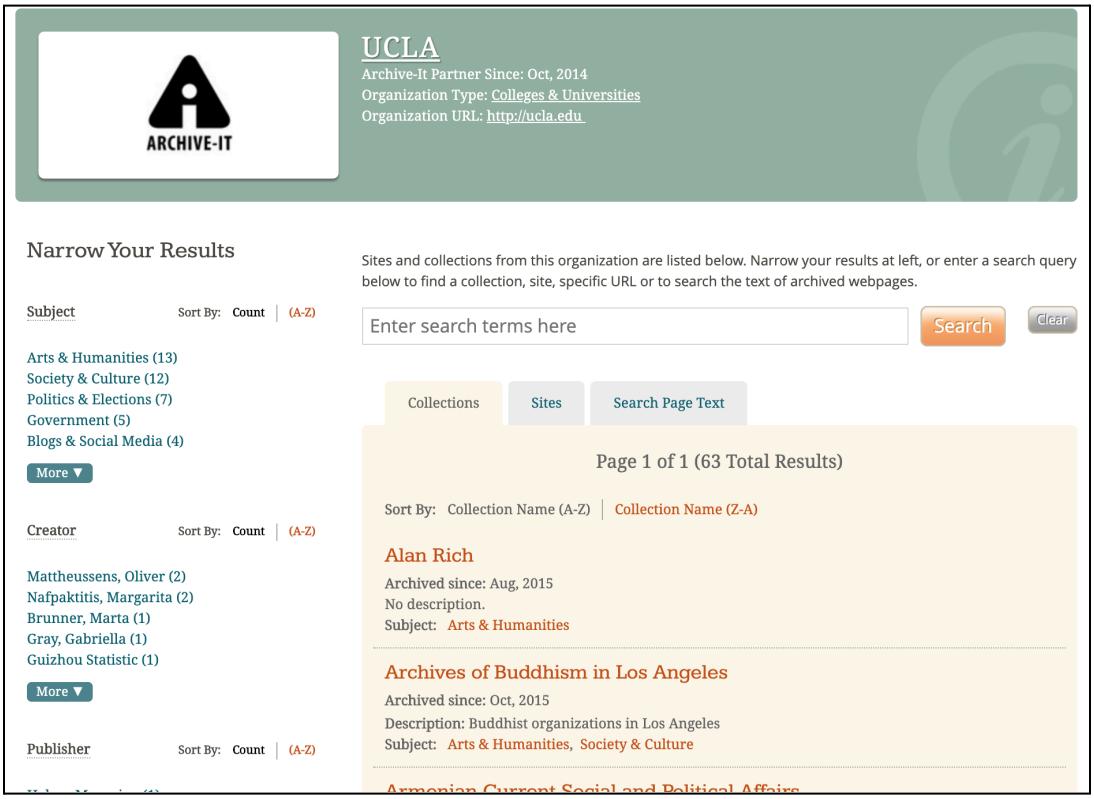
Appendix D: Exploration of Archive-It Contextual Model, Interface and OAI-PMH Feed

Archive-It Contextual Model and Interface

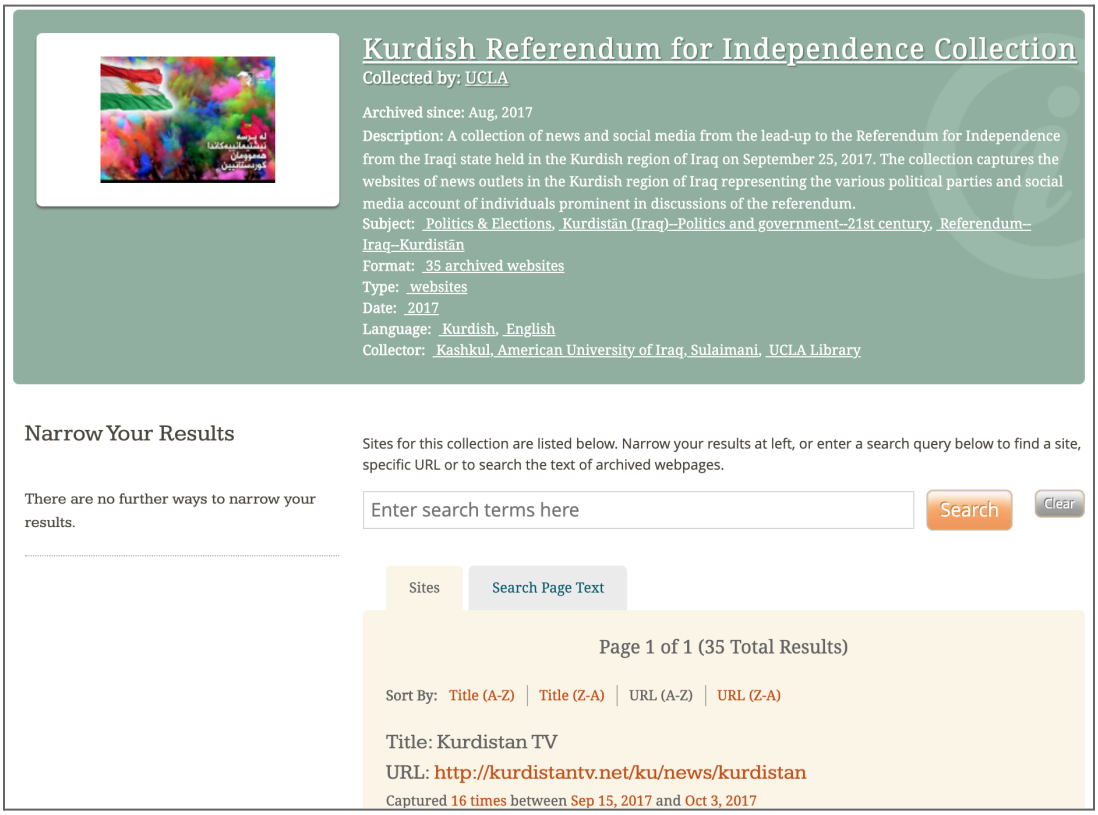
The Archive-It user interface presents web archives within four distinct layers of context:

1) **Institution landing page**
2) **Collection landing page**
3) **Seed landing page** in the Wayback machine
4) **Interactive display of the web archive** in the Wayback machine

The **institution landing page** displays information about the institution responsible for curating and capturing web archive collections in Archive-It. It contains optional descriptive information and an optional custom thumbnail for the institution, and a list of the web archive collections that the institution is curating:



Clicking on a collection on the "Collections" tab produces a **collection landing page,** which serves as a summary for related sets of web archives (e.g., related by topic, provenance) collected by the institution. The landing page displays descriptive information about the collecting area, with an optional custom thumbnail representing the collecting area. It also lists the individual web archives on a "Sites" tab; each web archive is represented by a particular "seed" URL and includes optional descriptive metadata for that web archive:



https://archive-it.org/collections/9302

Clicking on the seed URL for a particular web archive produces the **seed landing page.** The page again displays the seed URL for the captured web archive, along with the date of each capture. Note that the seed landing page does not include any metadata for the web archive:



https://wayback.archive-it.org/9302/*/http://kurdistantv.net/ku/news/kurdistan

Clicking on a capture date produces the **interactive display of the web archive** allowing users to dynamically interact with the archived web content saved on that particular date:



Some further notes on Archive-It metadata and harvesting, confirmed in an email from Internet Archive Web Curator Sylvie Rollason-Cass:

- The only required metadata fields are Title (at the collection level) and Seed URL (at the seed level)
- Collection and seed URLs within the Archive-It interface may be treated as persistent; they have not changed since Archive-It began and there are no plans to alter them.
- That said, if access to a document in Archive-It is removed at the behest of a collection holder or for any other reason, a standard “Not in Archive” message will appear, the same as if that page had never been crawled by Archive-It. There’s not currently any way to indicate that the document was removed.

Archive-It OAI-PMH Feed

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) allows a data provider such as Archive-It to expose standardized metadata for harvesting by a service provider such as CDL/Calisphere.[18] Although Archive-It provides several other methods for metadata capture, including an Application Program Interface (API), where available OAI-PMH is the preferred method for Calisphere harvesting as it is geared specifically toward transmitting structured metadata and is broadly utilized by libraries, archives, and museums.

OAI-PMH works by designating a specific OAI URL endpoint for a data provider, to which one of six OAI-PMH verbs must be affixed in order to retrieve specific information. Archive-It’s OAI endpoint may be accessed at *http://www.archive-it.org/oai/organizations/**[Organization***

[18] https://www.openarchives.org/pmh/

Number], where [Organization Number] is the organization's unique ID within the Archive-It system.

For example, a list of San Francisco Public Library's (SFPL) Archive-It collections or "Sets" available for harvesting via OAI-PMH may be retrieved by inserting SFPL's Organization Number (as found in their Archive-It institution landing page https://archive-it.org/organizations/160) and using the OAI-PMH verb **ListSets**:

https://www.archive-it.org/oai/organizations/160?verb=ListSets

```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd" xmlns:ex="http://exslt.org/dates-and-times"
xmlns:exslt="http://exslt.org/common" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2020-04-13T23:03:30Z</responseDate>
  <request verb="ListSets">http://archive-it.org/oai/organizations?verb=ListSets</request>
  <ListSets>
    <set>
      <setSpec>collection</setSpec>
      <setName>collection</setName>
    </set>
    <set>
      <setSpec>collection:3804</setSpec>
      <setName>Crash of Asiana Flight 214, San Francisco, CA, 7/6/2013</setName>
    </set>
    <set>
      <setSpec>collection:11866</setSpec>
      <setName>Chinese Cultural Institutions</setName>
    </set>
  </ListSets>
</OAI-PMH>
```

Archive-It supports harvesting of metadata at two key levels, which we will assess further:

1. Collection-level metadata (corresponding to metadata appearing on a collection landing page)
2. Seed-level metadata (corresponding to metadata appearing for each seed listed under the "Sites" tab of the collection landing page)

Collection-Level OAI-PMH Metadata

The OAI-PMH results faithfully convey the descriptive metadata shown on the collection landing page, including the required 'Title' field, with two exceptions:

- Some minor normalization, such as changing a Subject value from "Society & Culture" to "societyAndCulture"
- The OAI-PMH results are missing the contextually valuable "Archived since" field and value, auto-generated by Archive-It based on a collections' first web crawl and capture.

Note that the OAI-PMH results (<dc:identifier> element) also clearly provide a URL link back to the collection landing page.This URL link is important for Calisphere harvesting purposes: any given Calisphere object holds a URL link within an *isShownAt* field, pointing users back to a

representation of the object at the harvesting endpoint[19]. Users can then view and interact with the object at that endpoint. Note also that the OAI-PMH results do not contain a URL pointing to a representative thumbnail characterizing the web archive collection.

Collection-level OAI-PMH result:

https://archive-it.org/oai?verb=ListRecords&metadataPrefix=oai_dc&set=organization:160


```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd" xmlns:ex="http://exslt.org/dates-and-times" xmlns:exslt="http://exslt.org/common" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2020-05-22T20:51:20Z</responseDate>
  <request verb="ListRecords" set="organization:160">http://archive-it.org/oai?verb=ListRecords&amp;metadataPrefix=oai_dc&amp;set=organization:160</request>
  <ListRecords>
    <record>
      <header>
        <identifier>http://archive-it.org/collections/11726</identifier>
        <datestamp>2019-06-20T17:49:16.520Z</datestamp>
        <setSpec>organization:160</setSpec>
      </header>
      <metadata>
        <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:database="http://www.oclc.org/pears/">
          <dc:subject>societyAndCulture</dc:subject>
          <dc:title>Shuttered SF: Businesses &amp;amp; Institutions Out/Going Out of Business</dc:title>
          <dc:identifier>http://archive-it.org/collections/11726</dc:identifier>
        </oai_dc:dc>
      </metadata>
    </record>
    <record>
```


Associated collection landing page:

https://archive-it.org/collections/11866



[19] Calisphere users can access this URL link to the original object by clicking either the thumbnail image or 'View source [object type] on contributors website' text underneath. Example: https://calisphere.org/item/04b23cbda8429cb2c517805b3ba98b18/

Seed-Level OAI-PMH Metadata

The OAI-PMH results faithfully convey the descriptive seed metadata shown on the Collection Landing Page, including the required URL field, with two exceptions:

- Fields with multiple values such as Title and Description are split into multiple discrete values in the OAI-PMH (which actually makes them easier to work with for the purpose of Calisphere harvesting)
- The OAI-PMH results are missing the contextually valuable date span and number of captures--"Captured 8 times between Nov 2, 2012 and Oct 21, 2016"--auto-generated by Archive-It based on a seeds' capture frequency.

Note however that the OAI-PMH results (<dc:identifier> element) provide a URL link back to the *seed landing page*, which lacks seed-level metadata and context. It does ***not*** link back to the seed metadata on the "Sites" tab of the *collection landing page*, as one might expect.

Seed level OAI-PMH result

https://www.archive-it.org/oai/organizations/877?verb=ListRecords&metadataPrefix=oai_dc&set=collection:5903

```
<record>
  <header>
    <identifier>http://wayback.archive-it.org/5903/*/http://www.alforassembly.com/</identifier>
    <datestamp>2017-01-17T22:18:23Z</datestamp>
    <setSpec>collection:5903</setSpec>
  </header>
  <metadata>
    <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
    http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:database="http://www.oclc.org/pears/">
      <dc:coverage>66th Assembly District of California</dc:coverage>
      <dc:description>California General Election, November 6, 2012</dc:description>
      <dc:description>California Primary Election, June 3, 2014</dc:description>
      <dc:description>California General Election, November 4, 2014</dc:description>
      <dc:description>California Primary Election, June 7, 2016</dc:description>
      <dc:description>California General Election, November 8, 2016</dc:description>
      <dc:title>Muratsuchi, Al (2012-11-06)</dc:title>
      <dc:title>Muratsuchi, Al (2014-06-03)</dc:title>
      <dc:title>Muratsuchi, Al (2014-11-04)</dc:title>
      <dc:title>Muratsuchi, Al (2016-06-07)</dc:title>
      <dc:title>Muratsuchi, Al (2016-11-08)</dc:title>
      <dc:subject>General Election</dc:subject>
      <dc:subject>Democratic Party</dc:subject>
      <dc:subject>California State Assembly</dc:subject>
      <dc:subject>2012</dc:subject>
      <dc:subject>Primary Election</dc:subject>
      <dc:subject>2014</dc:subject>
      <dc:subject>2016</dc:subject>
      <dc:creator>Muratsuchi, Al</dc:creator>
      <dc:identifier>http://wayback.archive-it.org/5903/*/http://www.alforassembly.com/</dc:identifie
      r>
    </oai_dc:dc>
  </metadata>
</record>
```

Associated seed listed under the "Sites" tab of the collection landing page:

https://archive-it.org/collections/5903?q=muratsuchi&page=1&show=Sites

Page 1 of 1 (1 Total Results)

Sort By: Best Match | Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: **Muratsuchi**, Al (2012-11-06), Muratsuchi, Al (2014-06-03), Muratsuchi, Al (2014-11-04), Muratsuchi, Al (2016-06-07), Muratsuchi, Al (2016-11-08)

URL: http://www.alforassembly.com/

Description: California General Election, November 6, 2012 California Primary Election, June 3, 2014 California General Election, November 4, 2014 California Primary Election, June 7, 2016 California General Election, November 8, 2016

Captured 8 times between Nov 2, 2012 and Oct 21, 2016

Videos: 47 Videos Captured

Subject: General Election, Democratic Party, California State Assembly, 2012, Primary Election, 2014, 2016

Creator: Muratsuchi, Al

Coverage: 66th Assembly District of California

Date of Election: 2012-11-06, 2014-06-03, 2014-11-04, 2016-06-07, 2016-11-08

Page 1 of 1 (1 Total Results)