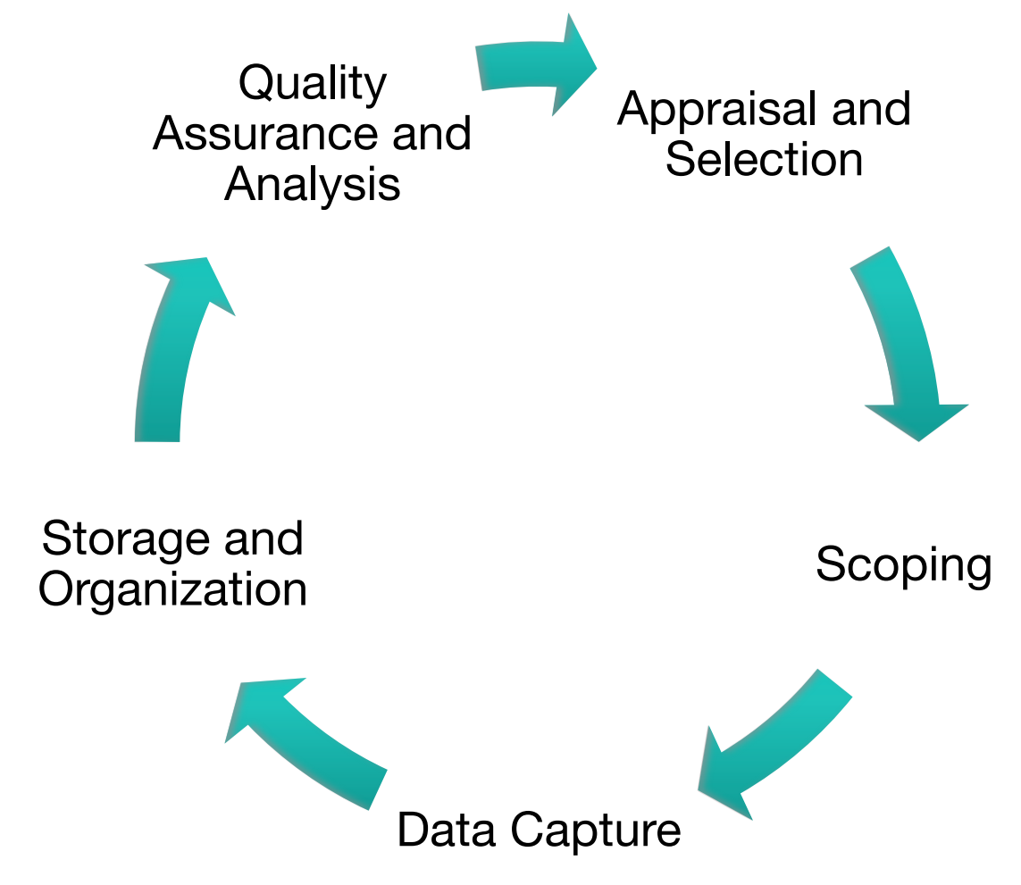
Crawling UCLA’s Web Archives: Capturing Issues & Making Recommendations that “WARC”

Rebecca Fordon, Dvorah Lewis, Niqui O’Neill

About the Archives

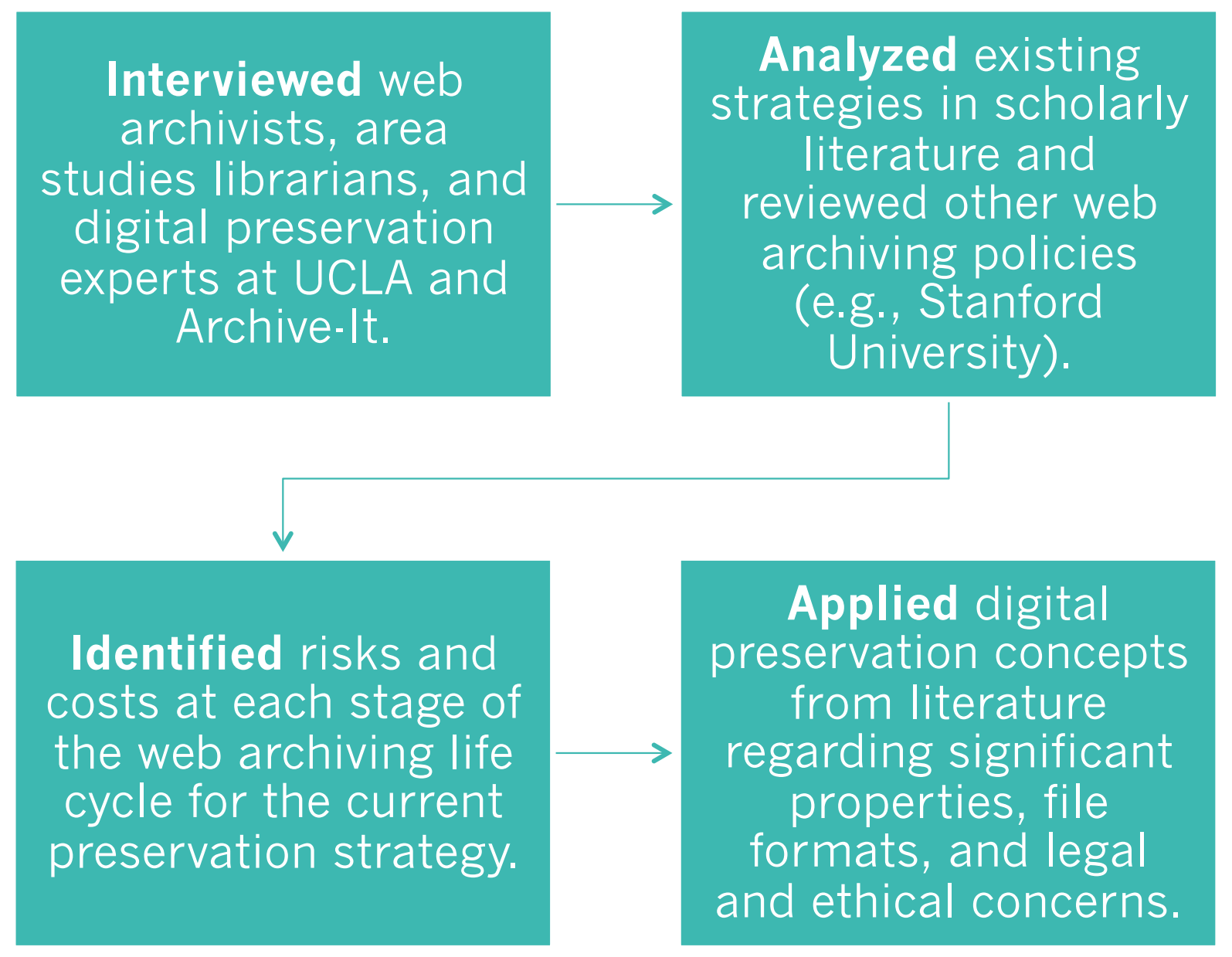
Some say the “web is forever,” but the reality is far from it. Due to censorship, spotty funding, and fluidity of content, the web and its contents are at risk of vanishing. UCLA’s Collections, Research, and Instructional Services (CRIS) works to document everything from political dissent in Russia to transient political movements or campaigns.

Our goal is to analyze current workflow and procedures to provide CRIS with a strategy to meet their objectives in preserving the web, making it accessible to users, and encouraging other departments across campus to join the initiative. Along with suggesting components of a digital preservation strategy, we aim to identify risks and costs of the current system and suggest recommendations to mitigate problem areas.



CRIS Web Archiving Work Flow

Information Gathering



Interviewed web archivists, area studies librarians, and digital preservation experts at UCLA and Archive-It.

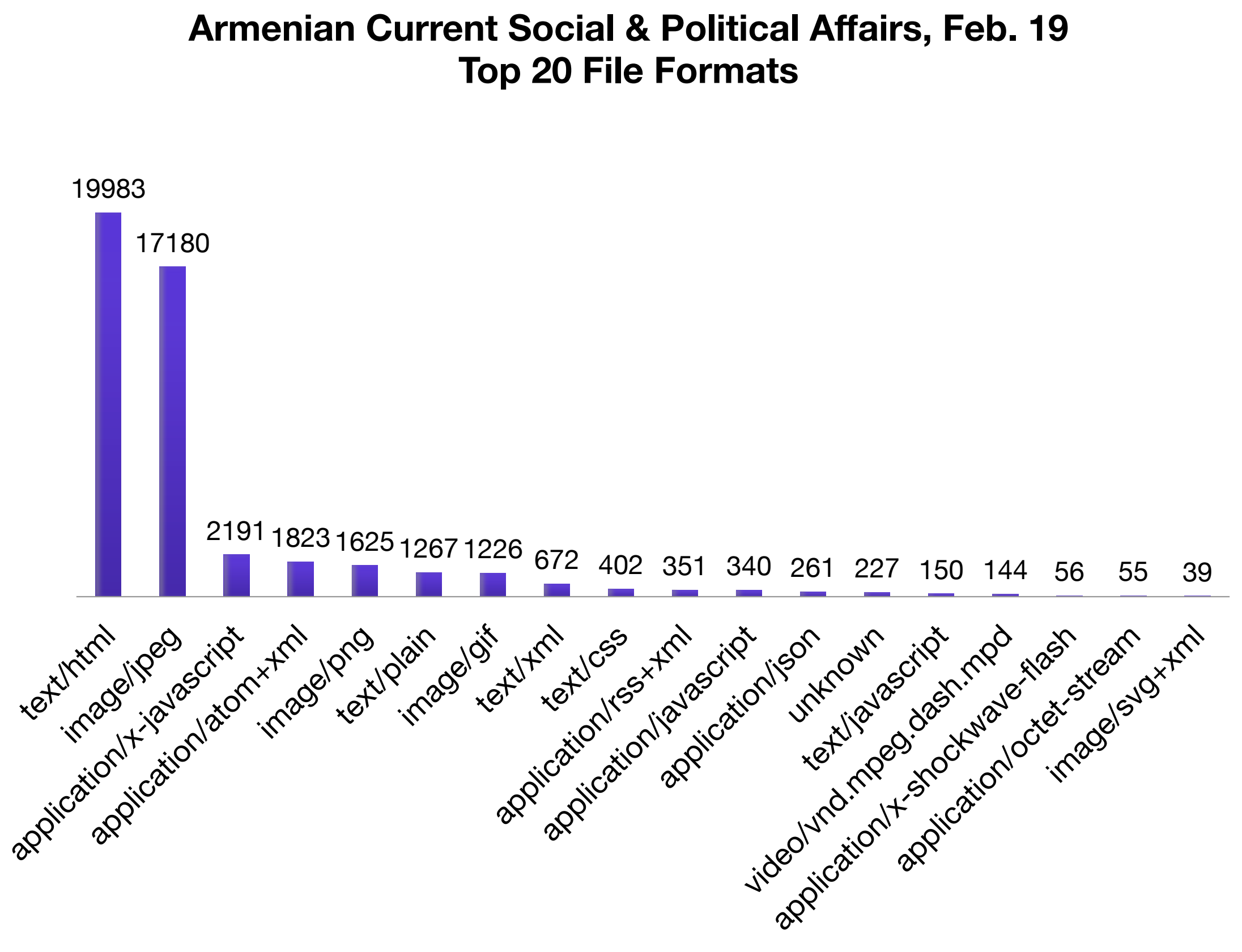
Analyzed existing strategies in scholarly literature and reviewed other web archiving policies (e.g., Stanford University).

Identified risks and costs at each stage of the web archiving life cycle for the current preservation strategy.

Applied digital preservation concepts from literature regarding significant properties, file formats, and legal and ethical concerns.

Analysis

- **File format obsolescence**: After a period of time, file formats will be unreadable by modern technology as the software and hardware for reading these files becomes obsolete. The large number of file types that go into creating a WebARChive file further increase the risk of obsolescence.



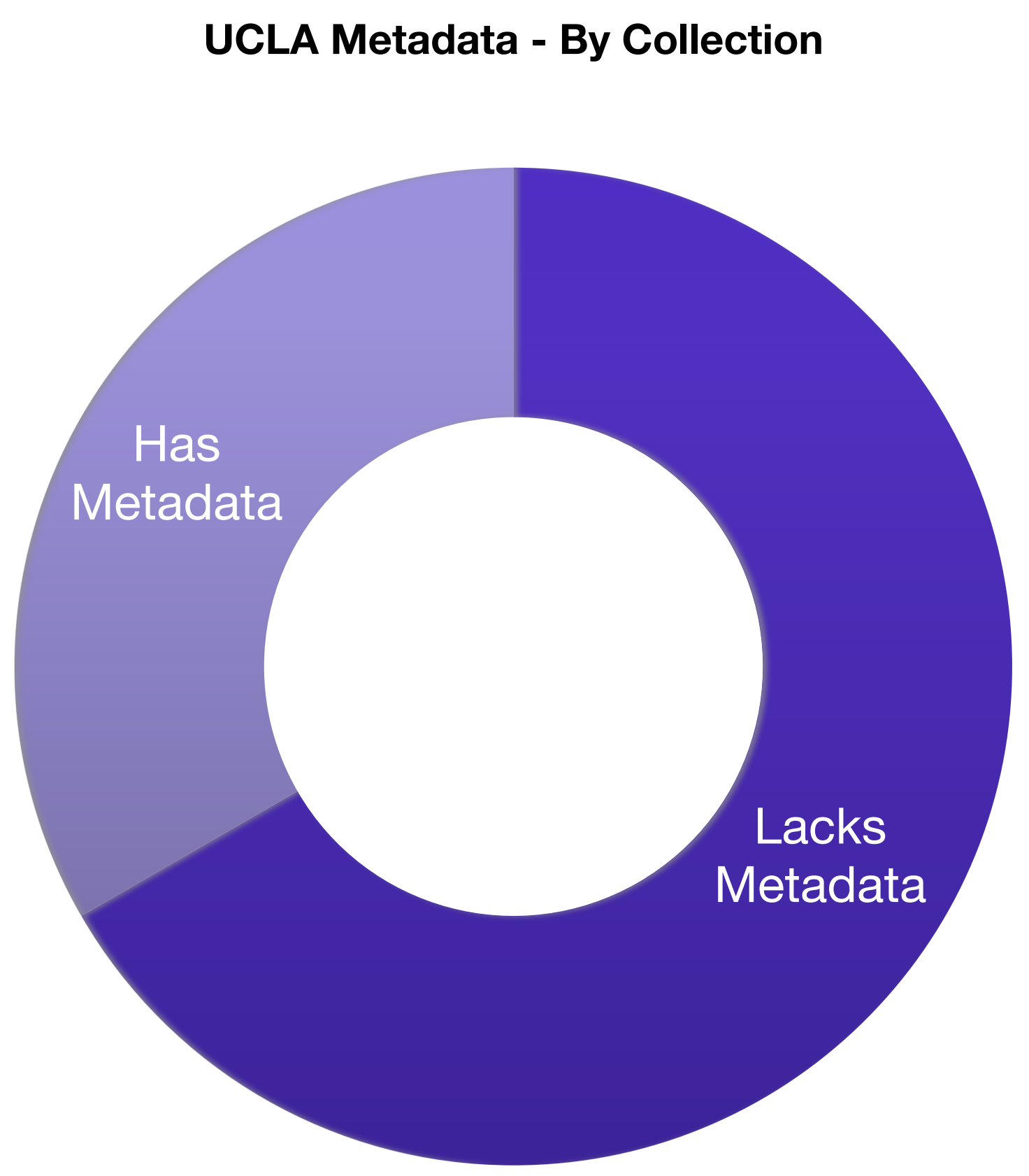
- **Data loss risk**: Currently CRIS stores all of its data on Archive-It’s servers, with the exception of older collections migrated from WAS. Archive-It stores all of its online backups in the Bay Area. Without backing up to a geographically distributed system, CRIS is vulnerable to data loss.

```
User-agent: *
Allow: /ads/public/
Allow: /svc/news/v3/all/pshb.rss
Disallow: /ads/
Disallow: /adx/bin/
Disallow: /archives/
Disallow: /auth/
Disallow: /cnet/
Disallow: /college/
Disallow: /external/
Disallow: /financialtimes/
Disallow: /idg/
Disallow: /indexes/
Disallow: /library/
Disallow: /nytimes-partners/
Disallow: /packages/flash/
multimedia/TEMPLATES/
```

Sample of robots.txt file from nytimes.com

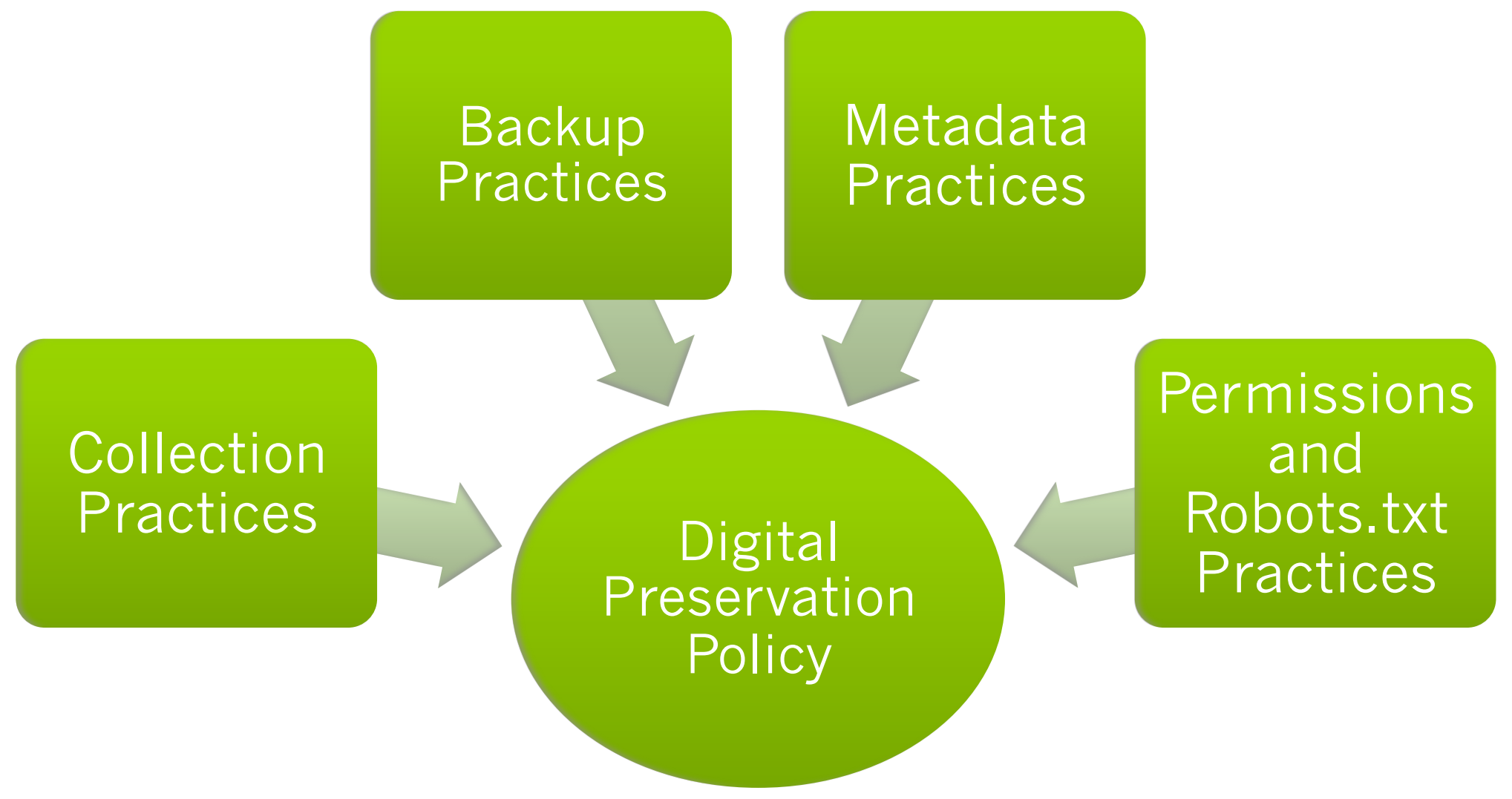
- **Copyright and permissions risk**: Although CRIS typically mitigates copyright risk by avoiding commercial sites and contacting the site owner before overriding robots.txt exclusions, it lacks formal policies. In addition to the liability risk, this presents potential for reputation harm with site owners who might not understand why their site is being crawled.

- **Collection documentation inconsistency**: UCLA and CRIS currently do not have any formal web archiving collection policy (with the exception of the UCLA Online Campaign Literature Web Archive). The lack of a formal policy diminishes context and authenticity, and makes it more difficult to plan for future collection development. These factors may produce further consequences, involving resource allocation, risk management, and potential for admissibility in legal proceedings.



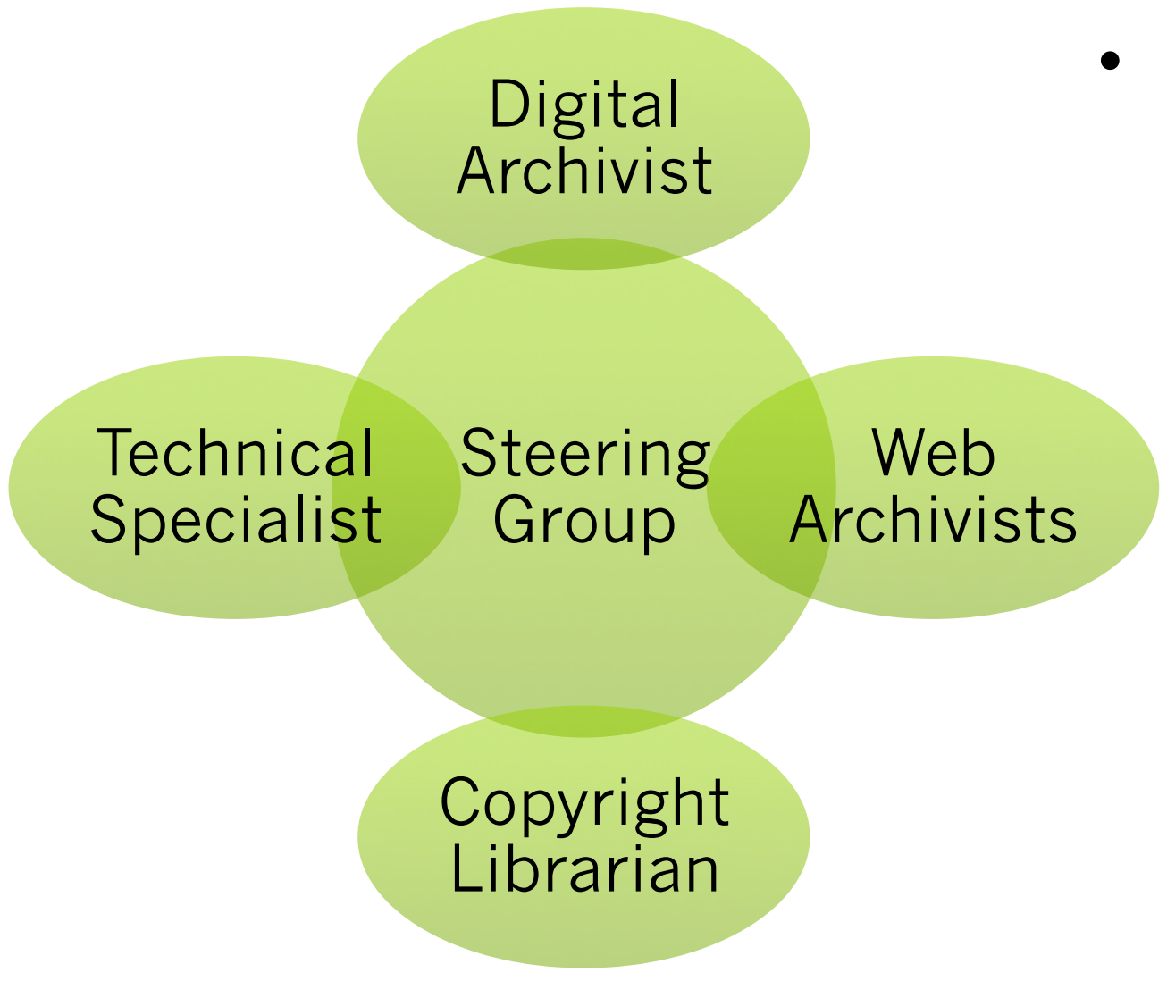
- **Metadata inconsistency:** UCLA web archivists across departments lack consistent metadata procedures, and lack archival principles informing workflow. Many of the collections lack consistent descriptions, and approximately two-thirds of the collections do not have any descriptions. As a result, users cannot fully access and contextualize the collections.

Recommendations



Backup Practices
Metadata Practices
Collection Practices
Digital Preservation Policy
Permissions and Robots.txt Practices

- **Store locally:** Backing up Archive-It files on a local server along with metadata information will mitigate risk of loss. CRIS may be able to use Archivematica for processing, since UCLA Digital Library plans to make it available virtually.
- **Continue emulation:** Because of the massive file sizes of web archives, emulation will prove a more viable solution for addressing file format obsolescence than migration.
- **Identify significant properties**: Identifying properties of primary concern for preservation (e.g., content, context, and connections) and “acceptable loss,” as well as documenting the choices in a collection policy will assist in quality assurance procedures.
- **Standardize metadata:** Clarifying and expanding required fields will provide context and increase accessibility. Some of the fields can include the Dublin Core Metadata Element Set. Additionally, “creator” should refer to the creator of the seedURL and “contributor” should refer to the curator.



- **Create policy steering group:** Many of the risks here can be addressed with a consistent policy, developed by a steering group with relevant expertise.

References:

Bragg, Molly and Kristine Hanna. “The Web Archiving Life Cycle Model.” *Archive-It* (website). March 2008. https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

Hanna, Kristine. “Archive-It Storage and Preservation Policy.” *Archive-It* (website). June 22, 2015. https://webarchive.jira.com/wiki/display/ARIH/Archive-It+Storage+and+Preservation+Policy.

Hedstrom, M. & Lee, C.A. “Significant properties of digital objects: Definitions, applications, implications.” In *Proceedings of the DLM Forum 2002, Barcelona, May 6-8, 2002*. Luxembourg: Office for Official Publications of the European Communities, 2002.

Webb, Colin, David Pearson, and Paul Koerbin. “‘Oh, you wanted us to preserve that?!’ Statements of Preservation Intent for the National Library of Australia.” *DLib Magazine*. January/February 2013. doi:10.1045/january2013-webb.

Acknowledgments:

The authors would like to thank Oliver Mattheussens and the area studies librarians at CRIS for allowing us this opportunity to work with them. We would also like to thank the many people who provided their expertise and guidance: Shira Peltzman, Martin Klein, Karl Blumenthal, and Jean-François Blanchette.