EMBEDDING PRESERVABILITY: IFRAMES IN COMPLEX SCHOLARLY PUBLICATIONS

Karen Hanson
Portico
United States
karen.hanson@ithaka.org
https://orcid.org/0000-0002-9354-8328

Jonathan Greenberg
New York University Libraries
United States
jonathan.greenberg@nyu.edu
https://orcid.org/0000-0002-3429-4428

Thib Guicherd-Callin
LOCKSS
United States
thib@stanford.edu
https://orcid.org/0000-0002-6425-4072

Scott Witmer
University of Michigan Library
United States
switmer@umich.edu

Angela T. Spinazzè
ATSPIN consulting
United States
ats@atspin.com

Abstract – As part of a research project, a small team of preservation experts has been embedded within publisher workflows to analyze the challenges associated with preserving complex scholarly publications. As the project reaches the midway point, patterns are emerging regarding preservation-friendly practices that could potentially be incorporated into production processes and platforms to support preservation at scale. One common threat to the preservability of the analyzed publications is the inclusion of web pages that are hosted by a third party (e.g., YouTube videos, ArcGIS visualizations) within the text using *iframes*. The team is exploring methods to improve preservability in such instances while considering the constraints of the project partners and the requirement that preservation services can scale their processes across numerous publications.

Keywords – websites, publishing, publications

Conference Topics – From Theory to Practice; We're All in this Together.

I. BACKGROUND

A new generation of scholarly publications and publishing platforms are leveraging technology to support the integration of complex features, such as embedded streamed audio or video, interactive visualizations, and features for user feedback, into articles and monographs. These dynamic publications create challenges for preservation. The Embedding Preservability for New Forms of Scholarship project [1], which is funded by the Mellon Foundation and led by NYU Libraries, is investigating methods to make these publications more preservable at scale. They are doing this by embedding a team of preservation experts from NYU Libraries, University of Michigan Library, LOCKSS, and Portico into the publishing workflow. For three years, the *embedding team* will shadow the publication production process while engaging with the platforms used by those publishers, interviewing each, and providing feedback as they work on new publications. While an earlier research project developed a framework for understanding the scope of the challenges for preserving complex publications at scale and produced a set of guidelines that publishers could use to improve the preservability of them [2][3], the current project focuses on implementation of the guidelines. Where can the guidelines be integrated into the platform design, user documentation, and publisher workflows? What preservation-friendly practices are most effective, and most likely to be adopted by publishers? As the project approaches the midway point, early patterns indicate some common challenges for preservation and the team is exploring options to manage them. One of these challenges is the use of *iframes*.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/1210666

II. Iframes and Link Rot

In the publications analyzed, it is common to incorporate complex features from third party platforms using an iframe. An iframe is an HTML tag that allows the author to embed a view of a third-party web page into another web page - in this case, into the text of a publication. Typical examples include 3D ArcGIS visualizations and YouTube videos.

These embedded pages are rarely associated with a persistent URL and are at risk of link rot, where the page moves or is taken offline causing a broken link. This can happen at any time, sometimes before any preservation activity takes place, making these features vulnerable to permanent loss. In addition, the content embedded using an iframe tends to include the most complex and dynamic features in the publication, making them a significant piece of the work, but also sometimes technically difficult to replicate at high fidelity even with the latest web archiving techniques.

III. Need for a Standardized Approach

The embedding team is considering ways to minimize the loss of these resources where they are a core intellectual component of the publication. The team has observed that platforms rarely have sufficient standards or guidance around these integrations to ensure reliable scalable preservation. They are not uniformly managed within or between platforms and often have inadequate or missing captions and/or references. The team's initial suggestions for circumventing loss (e.g., implementing a local web archiving workflow, linking raw data and documentation) were not feasible in the short term for most publishers and platforms, so the team is considering ways to deconstruct these recommendations into smaller steps that require less effort and reduce the impact of this issue.

1) Use of Captions: The convention of adding context and rights information under a figure graphic has not been well adapted for embedding dynamic content. Even though these features are visually displayed, they are often missing captions, and so improving captions for third party content would be a positive step. The team is considering recommendations for appropriate caption content standards, and whether a tool to generate and format a caption, possibly to include machine-readable metadata in the HTML, might be useful. The goal is to include information useful for discovery and understanding of the missing material if the link breaks and leaves a gap in the publication.

2) Alt Text: The practice of adding alt-text to describe non-text features is helpful for accessibility and many publishers are already considering this in their workflows. Alt-text can provide information about embedded content even if it is no longer available. This recommendation would ask that publishers ensure that the iframes' *title* attribute is populated.

3) Data Citation, Archive References: The use of data citation practices to cite the published or archived data source for a visualization using a persistent link is a helpful practice to support preservation. In the case of a GIS visualization, a DOI link might point to a data repository containing the raw data and/or related software. For a non-persistent link to a website, if the resource is compatible with web archiving, this could be a persistent URL to an archived webpage.

4) License Tags: Publishers are used to clearing rights for re-use on graphics embedded in a published analogue work. The nature of the web, however, permits embedding of many resources using iframes without permission. This is convenient, but consequential if the preservation service for that content requires permission from copyright holders or an appropriate license to copy material to the archive. Determining the license of this content at scale requires machine-readable metadata for the object. One proposal is to use the *rel="license"* property in an appropriate HTML tag to designate a license to the embedded content [4]. This would enable the archive to take measured risks when copying the content by automatically reading the tag. It would allow publishers to tag content that should not be copied or be copied but kept in a dark archive until the copyright expires. This property could pair well with URIs from RightsStatements.org, whose purpose is to provide "standardized rights statements that can be used to communicate the copyright and re-use status of digital objects to the public." [5]

IV. CONCLUSION

The embedding team continues to engage with publishers and the developers of their platforms to explore ways to ensure third-party resources hosted outside of the publisher platform can be preserved. The second part of the project will determine which of these ideas for managing iframes are practical for publishers and can be easily implemented.

1. REFERENCES

[1] "The Andrew W. Mellon Foundation Awards NYU $502,400 For Libraries Project to Expand Capabilities for Preserving Digital Scholarship," 04-Aug-2021. [Online] Available: https://guides.nyu.edu/blog/The-Andrew-W-Mellon-Foundation-Awards-NYU-502400-For-Libraries-Project-to-Expand-Capabilities-F [Accessed 09-Mar-2023]

[2] J. Greenberg, D. Verhoff, and K. Hanson, "Guidelines for Preserving New Forms of Scholarship," 2021. [Online]. Available: https://doi.org/10.33682/221c-b2xj. [Accessed: 09-Mar-2023].

[3] J. Greenberg, D. Verhoff, and K. Hanson, "Report on Enhancing Services to Preserve New Forms of Scholarship," 2021. [Online]. Available: https://doi.org/10.33682/0dvh-dvr2. [Accessed: 09-Mar-2023].

[4] "CC REL by example." [Online]. Available: https://opensource.creativecommons.org/ccrel-guide/. [Accessed: 09-Mar-2023].

[5] Rightsstatements.org. [Online]. Available: https://rightsstatements.org/en/. [Accessed: 09-Mar-2023].