Should Web Archives Be Used For Research Data Preservation?

Todd Suomela
University of Alberta
Digital Initiatives - University Libraries
Edmonton, Canada T6G 2J8
Tel # 1-780.248.1952
todd.suomela@ualberta.ca

ABSTRACT

This poster describes some of the challenges for managing web archive collections when they intersect with research data preservation. A web archive collection at the University of Alberta inadvertently harvested large data files from another institution which resulted in overloading the subscription budget for Archive-IT. A discussion followed about the appropriate policy approaches for web archive programs when they encounter research data on the web. The poster presents some of the evaluation criteria used to make decisions about including or excluding research data from web archives. Existing web archive tools are ill-prepared to deal with research data. Furthermore, responsibility for preserving research data and web documents is difficult to determine. Finally, the role of third-parties outside of the original institutions where research data is created is still unclear. Future activity in this area should address these challenges.

General Terms

Infrastructure opportunities and challenges; Preservation strategies and workflows

Keywords

web archives, research data, preservation policies

1. INTRODUCTION

In 2011 the University of Alberta library created a Circumpolar web archive collection using the subscription service Archive-IT. The new collection supplemented an existing non-digital collection which had been in place for decades, expanding into the digital realm seemed a natural extension of already existing services. Since then the web archive collection has faced a number of challenges demonstrating how digital collections present new problems and opportunities for libraries and archives.

One of the major challenges is dealing with data files published by other institutions on the web. In the summer of 2014 a one-time crawl consumed almost one-third of the annual data budget for the Archive-IT subscription service. The root cause of this was the downloading of hundreds of zip files which contained digital geographic images in the form of TIFF files. This presented a challenge for the library as a whole because the data budget for Archive-IT subsumes 13 active collections across multiple disciplines and library sub-units. For a single collection to consume over a third of the data budget is unsustainable. The key question raised by this incident is what the role of web archives is in research data management. Should web archives be considered part of the research data preservation process? How should research data be managed across varying institutions? Who should be responsible for preserving research data?

2. DECISION MATRIX FOR WEB ARCHIVES AND RESEARCH DATA

This poster presents a decision matrix for evaluating the relationship between a web archive and research data. Which questions should be asked in order to make decisions about the inclusion or exclusion of research data within a web archive repository? Is a web archive repository the right tool for backing up research data?

The following questions are part of the decision matrix for research data in web archives.

- Where is the data being stored?
  - Is it available on the open web?
  - Is it available from a reliable institution? e.g. a university, government on non-profit
  - Is there a preservation plan in place for the data?
- Is there any immediate threat for the data to be lost or be no longer available? (This partly depends on the type of institution hosting the data.)
- Who is responsible for maintaining the data over time?
  - Are there any disciplinary repositories which may be preserving the data?
  - Is there a hierarchy for data responsibility?
- Who are the future and current audiences for sustaining a copy of the data in a web archive?

  - How is access currently being managed through the open-web? Are there provisions or licenses which may affect access to the data through other sources, such as a web archive?
  - What provisions or tools are there available for metadata management within the web archive toolset?

- Are there resources available for storing this data in the web archive?
  - Can the data be stored in the web archive given existing data budgets, metadata descriptive services, and staff budgets?

- What is the value timeline for this research data?
  - Can a web archive preserve enough information to make this data useful for researchers in the future?
  - Will this data be useful for researchers at your institution in 10 or 20 years?

3. CONCLUSION

The evaluation process for collecting research data through web archives is still underway at the University of Alberta libraries. Early conclusions from the process are presented below:

First, existing web archive software is ill prepared to preserve research data. Most web archiving tools are designed to capture the entirety of files published by a web domain or seed with minimal filtering. The controls for metadata creation are rudimentary, the presentation layer focuses on the fidelity of web browser presentation instead of information retrieval, and the collection management interfaces allow for very limited filtering or description of file types. Given the variety of file types associated with research data web archiving should not be the first choice for research data preservation. Specific data repositories, such as Dataverse, at local institutional or disciplinary levels, make much more sense.

Second, determining the responsibility for preserving the web, whatever the form of content which is posted, is currently quite difficult. Existing data repositories are covered by a diverse range of preservation policies. Marcial & Hemminger [1] conducted an online survey of 100 scientific data repositories and identified preservation policies for 62% of the sample, but the particular policy content was idiosyncratic. Preservation information or policies are even harder to identify for individual web sites or domain names. The only feasible way to preserve large volumes of the web and have a layered approach to preservation is to develop an automated description of preservation policies similar to the robots.txt files used to limit web crawling.

Third, the role of third-parties in preserving research data from institutions with which they do not currently share any explicit agreements is very challenging. Libraries may be willing to preserve research data from other institutions but these projects are often expensive, especially in the amount of personnel time needed to coordinate the activities between multiple institutions. Most web archiving programs do not have the resources to pursue such in-depth agreements for preserving research data. This means that decisions about collecting research data within web archives will continue to be a singular challenge.

4. ACKNOWLEDGEMENTS

Thanks to Chuck Humphrey and Umar Qasim for discussing this proposal.

5. REFERENCES

[1] L. H. Marcial and B. M. Hemminger. Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10):2029–2048, October 2010.