IIPC DISCRETIONARY FUNDING PROGRAM 2019-2020

FINAL REPORT

ASKING QUESTIONS WITH WEB ARCHIVES – INTRODUCTORY NOTEBOOKS FOR HISTORIANS

2020-06-04

LEAD IIPC INSTITUTION: The British Library

2ND IIPC INSTITUTION: National Library of Australia, National Library of New Zealand

OTHER INSTITUTIONS OR CONSULTANTS: Tim Sherratt

PROJECT TEAM MEMBERS: Tim Sherratt, University of Canberra. Andrew Jackson, The British Library

AWARDED FUNDING (IN USD)**: 3,500**

PROJECT WEBSITE: http://netpreserve.org/projects/jupyter-notebooks-for-historians/

BRIEF ABSTRACT OF THE PROJECT:

This project aimed to create a set of Jupyter notebooks that demonstrate how specific historical research questions can be explored by analysing data from web archives. The notebooks are targeted at researchers who have limited understanding of, or interest in, the technology of web archives, but want to do more than simply browse snapshots.

To avoid overwhelming researchers with the scale and scope of web archives, the notebooks created for this project work with data from IIPC members available through Wayback, Memento, and CDX APIs. They introduce tools and technologies gradually – building the understanding and confidence of researchers. These notebooks focus on four particular web archives: the UK Web Archive, the Australian Web Archive (National Library of Australia), the New Zealand Web Archive (National Library of New Zealand), and the Internet Archive. However, the tools and approaches here could be easily extended to other web archives.

This is not just another set of tutorials. By using Jupyter notebooks, the project provides live code and practical examples that yield immediate research benefits, while also bringing together distributed documentation in a form that can be understood by researchers with limited digital skills. While the project was deliberately focused on helping historians understand how their research might be enriched by web archives, the information and examples provided will be useful to any researcher seeking to develop their knowledge and skills.

PROJECT OUTPUTS / OUTCOMES:

The Web Archives section of the GLAM workbench contains the 16 notebooks created during this project:

- Types of data
  - Timegates, Timemaps, and Mementos
  - Exploring the Internet Archive's CDX API
  - Comparing CDX APIs
  - Timemaps vs CDX APIs
- Harvesting data and creating datasets
  - Get the archived version of a page closest to a particular date
  - Find all the archived versions of a web page
  - Harvesting collections of text from archived web pages
  - Harvesting data about a domain using the IA CDX API
  - Find and explore Powerpoint presentations from a specific domain
  - Exploring subdomains in the whole of gov.au
- Exploring change over time
  - Compare two versions of an archived web page
  - Observing change in a web page over time
  - Create and compare full page screenshots from archived web pages
  - Using screenshots to visualise change in a page over time
  - Display changes in the text of an archived web page over time
  - Find when a piece of text appears in an archived web page

Within the three main themes, the notebooks include a mixture of tutorial material, examples, quick hacks and standalone tools, each built around exploring particular historical questions.

The notebooks are CC-BY licensed to encourage reuse and adaptation, and stored in a public GitHub repository. The repository includes a 'requirements.txt' file to document the required software environment, and to make it easy for the notebooks to be run live on Binder without any need for researchers to install software on their own systems.

The results are embedded within the broader GLAM Workbench suite, helping to establish web archives within the wider cultural heritage sector, and helping a wider audience engage with the archived web.

ANECDOTAL INFORMATION:

The variation between APIs between different institutions and implementations caused some initial problems, but the dedicated Slack channel helped ensure Tim could contact and collaborate with individuals from each web archive in order to understand these differences and work around them.

This also provided an opportunity for the participating archives to reflect on the current state of the services they offer, and where improvements might be made. For example, some questions could only be answered using the APIs provided by the Internet Archive, which may prompt other archives to consider deploying a similar service.

In particular, the UK Web Archive attempts to offer a CDX API, but the fact that one data field is not present (compressed WARC record length) means that certain analyses are impossible. This prompted the UK Web Archive to improve their service and they will start including this field in their service in the near future.

BEST PRACTICES:

The notebooks provided in the Web Archives section of the GLAM Workbench are open access, and built on open standard and open source. This means anyone can use them, and we can take advantage of how easy they are to deploy on public services like MyBinder. Their portability and interoperability also means they can be easily deployed behind an institutional firewall, meaning the same analyses can be run on non-public collections (as long as the same, standard access APIs are available over an organisations intranet).

PROJECT CONTINUITY:

There are a number of possible follow-up projects:

- Funding of outreach and/or workshop events to encourage use of notebooks.
- Developing advice on running the notebooks locally and/or integrating them with existing services.
- Maintaining and extending the notebooks to include more web archives, or explore more questions.
- Funding improvements to pywb/OpenWayback/SolrWayback so IIPC members can expose APIs that are more consistent and more powerful.