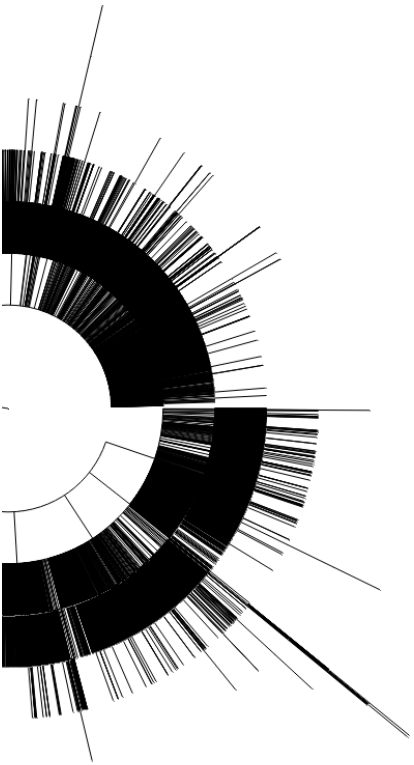
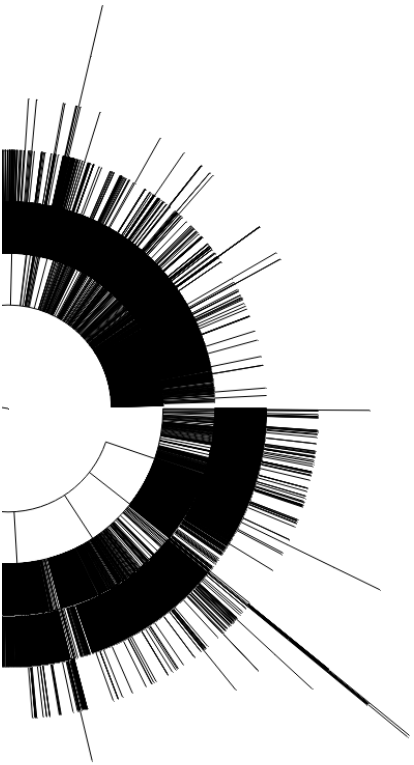


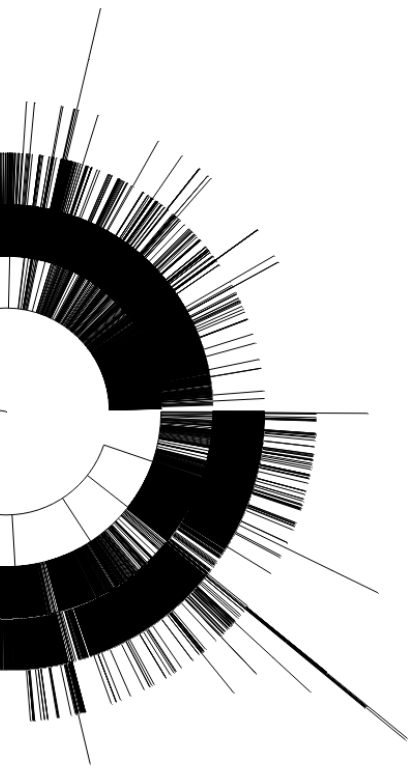
JUPYTER NOTEBOOKS FOR WEB ARCHIVES

Tim Sherratt • @wragge • #glamworkbench



PLAY ALONG

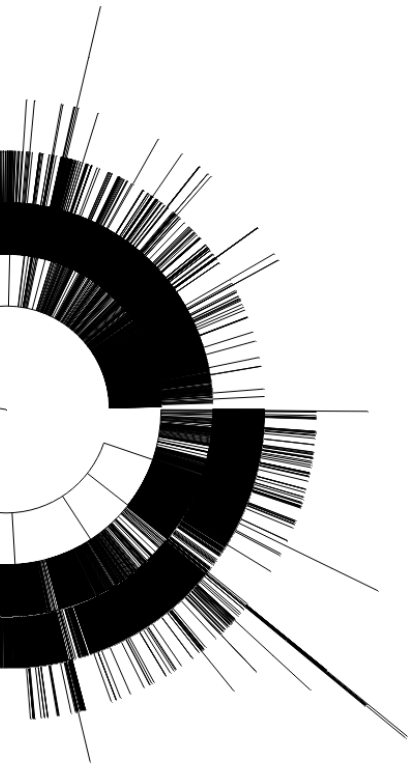
https://slides.com/wragge/iipc-jupyter



PROJECT

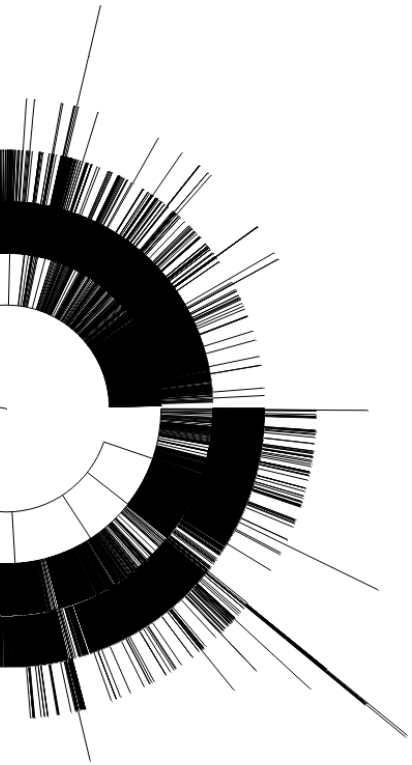
Asking questions with web archives – introductory notebooks for historians

- IIPC Discretionary Funding Program 2019–2020
- Project description
- Blog post
- Zenodo repository



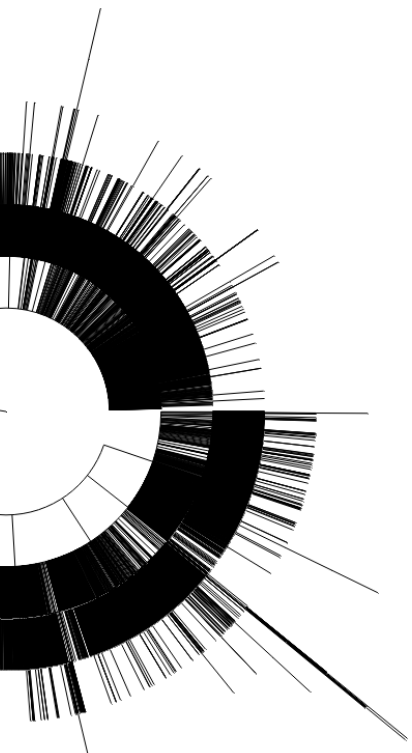
AIMS

- a starting point for researchers
- use existing APIs (Memento & CDX)
- no special tools
- complimenting projects like Archives Unleashed



JUPYTER

- combines text and live code
- use in your browser
- run in the cloud (no software to install)
- both tool and tutorial



JUPYTER

TEXT

CODE

Get the archived version of a page closest to a particular date

New to Jupyter notebooks? Try **Using Jupyter notebooks** for a quick introduction.

To get the archived version of a page closest to a particular date we can use the Memento API. Variations in the way Memento is implemented across repositories are documented in Getting data from web archives using Memento. The functions below smooth out these variations to provide a (mostly) consistent interface to the UK Web Archive, Australian Web Archive, New Zealand Web Archive, and the Internet Archive. They could be easily modified to work with other Memento-compliant repositories.

To get information about available Mementos:

```
query_timegate([timegate], [url], [date], [timezone])
```

To get a single Memento closest to your target date:

```
get_memento([timegate], [url], [date], [timezone])
```

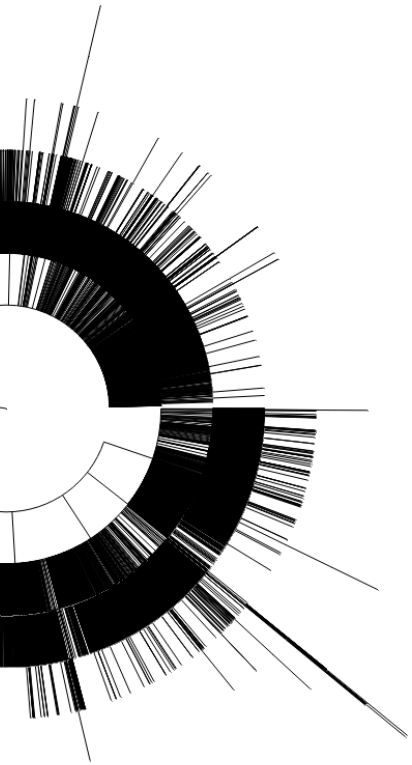
Parameters:

- `timegate` – one of 'ukwa' (UK), 'awa' (Australia), 'nzwa' (New Zealand), or 'ia' (Internet Archive)
- `url` – the url you want to look for in the archive
- `date` – the target date in ISO format, 'YYYY-MM-DD' (optional, will default to most recent date)
- `tz` – a timezone string for your local timezone (optional)

In [16]:
```
import requests
import arrow
import re
import json
```

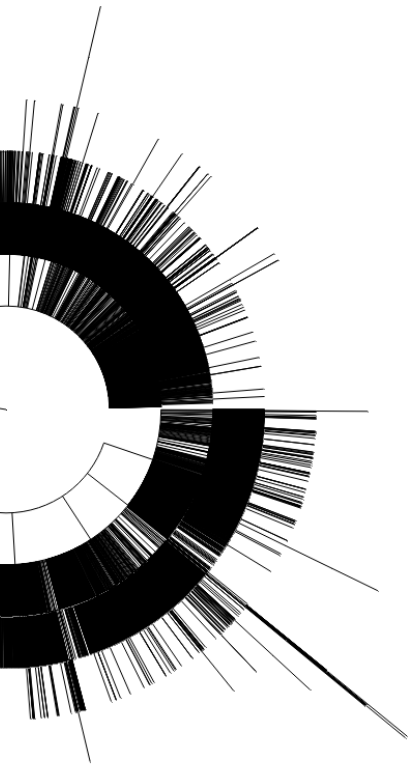
In [29]:
```
# These are the repositories we'll be using
TIMEGATES = {
    'awa': 'https://web.archive.org.au/awa/',
    'nzwa': 'https://ndhadeliver.natlib.govt.nz/webarchive/wayback/',
    'ukwa': 'https://www.webarchive.org.uk/wayback/en/archive/',
    'ia': 'https://web.archive.org/web/'
}

def format_date_for_headers(iso_date, tz):
    '''
    Convert an ISO date (YYYY-MM-DD) to a datetime at noon in the specified timezone.
    Convert the datetime to UTC and format as required by Accet-Datetime headers:
    eg Fri, 23 Mar 2007 01:00:00 GMT
    '''
    local = arrow.get(f'{iso_date} 12:00:00 {tz}', 'YYYY-MM-DD HH:mm:ss ZZZ')
    gmt = local.to('utc')
    return f'{gmt.format("ddd, DD MMM YYYY HH:mm:ss")} GMT'
```

RUNNING CODE



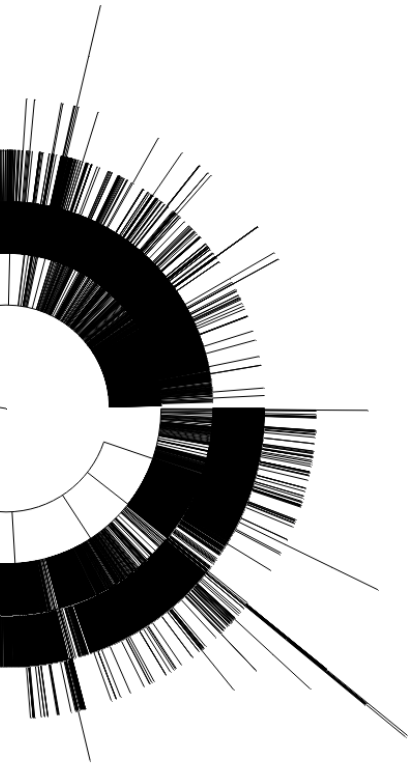
- click on a cell
- hit **Shift+Enter**

EDITING CODE



- click on a cell
- edit the contents
- hit **Shift+Enter** to run

GLAM WORKBENCH



GLAM Workbench

Search

GLAM-Workbench
30 Repositories

GLAM Workbench

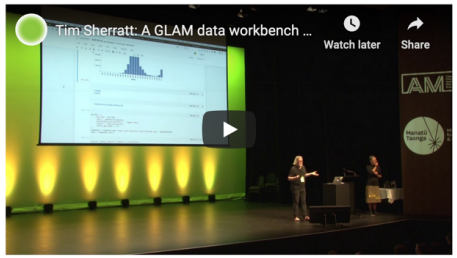
- Home
- About
- Help
- Data sources
- GLAM Labs
- Trove
- DigitalNZ
- Archives
- Libraries
- Web Archives
- Museums
- Government
- Suggest a topic

chat on gitter

Welcome to the wonderful world of GLAM data!

Here you'll find a collection of tools, tutorials, examples, and hacks to help you work with data from galleries, libraries, archives, and museums (the GLAM sector). The primary focus is Australia and New Zealand, but new collections are being added all the time. Let me know if there's some GLAM data you'd like me to explore – suggestions are always welcome!

Here's a brief explanation of what I'm trying to do, courtesy of the 2018 National Digital Forum in New Zealand.



There's more presentations here.

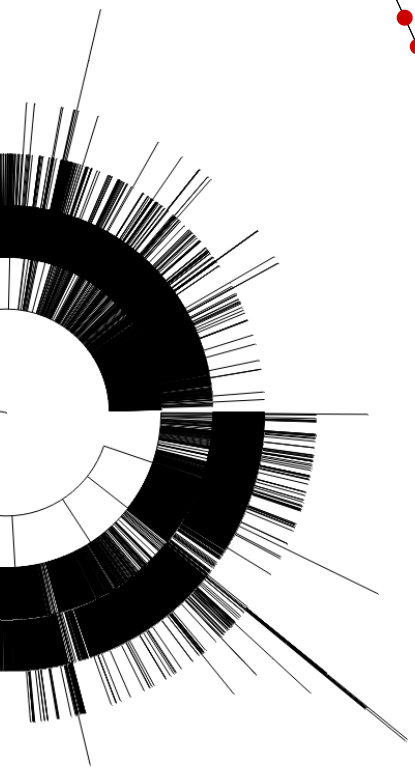
What is GLAM data?

When we talk about GLAM data we're usually referring to the collections held by cultural institutions – books, manuscripts, photographs, objects, and much more. We're used to exploring these collections through online search interfaces or finding aids, but sometimes we want to do more – instead of a list of search results on a web page, we want access to the underlying collection data for analysis, enrichment, or visualisation. We want collections as data.

Table of contents

https://glam-workbench.github.io/

GLAM WORKBENCH



GLAM Workbench

Search

GLAM-Workbench
30 Repositories

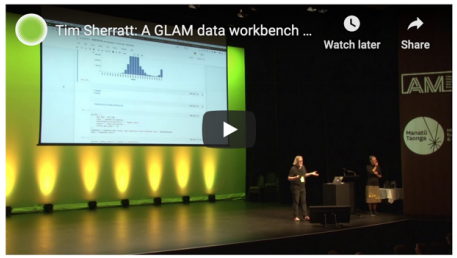
GLAM Workbench
Home
About
Help
Data sources
GLAM Labs
Trove
DigitalNZ
Web Archives
Museums
Government
Suggest a topic

chat on gitter

Welcome to the wonderful world of GLAM data!

Here you'll find a collection of tools, tutorials, examples, and hacks to help you work with data from galleries, libraries, archives, and museums (the GLAM sector). The primary focus is Australia and New Zealand, but new collections are being added all the time. Let me know if there's some GLAM data you'd like me to explore – suggestions are always welcome!

Here's a brief explanation of what I'm trying to do, courtesy of the 2018 National Digital Forum in New Zealand.



There's more presentations here.

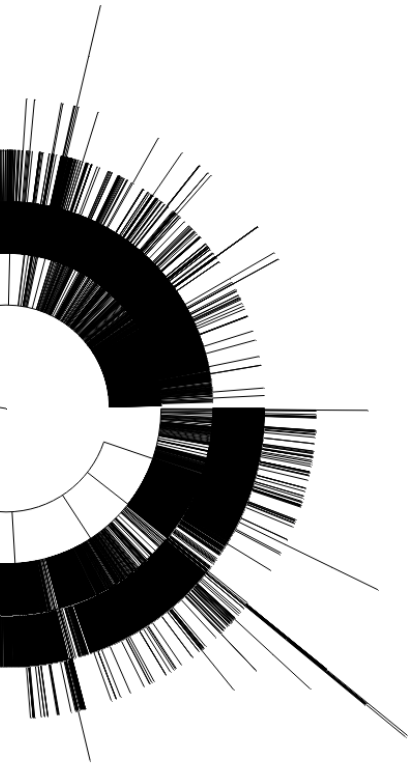
What is GLAM data?

When we talk about GLAM data we're usually referring to the collections held by cultural institutions – books, manuscripts, photographs, objects, and much more. We're used to exploring these collections through online search interfaces or finding aids, but sometimes we want to do more – instead of a list of search results on a web page, we want access to the underlying collection data for analysis, enrichment, or visualisation. We want collections as data.

Table of contents

https://glam-workbench.github.io/

GLAM WORKBENCH



Timegates, Timemaps, and Mementos

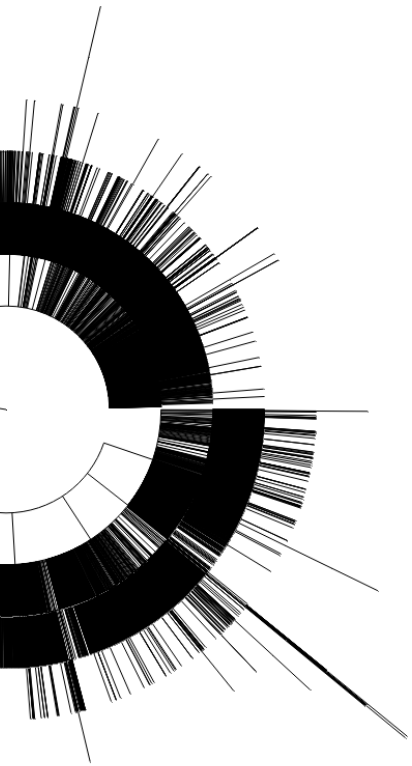
Works with AWA, NZWA, IA, & UKWA

Systems supporting the Memento protocol provide machine-readable information about web archive captures, even if other APIs are not available. In this notebook we'll look at the way the Memento protocol is supported across four web archive repositories – the UK Web Archive, the National Library of Australia, the National Library of New Zealand, and the Internet Archive.

- Download from GitHub
- View using NBViewer
- Run live on Binder

https://glam-workbench.github.io/web-archives/

GLAM WORKBENCH



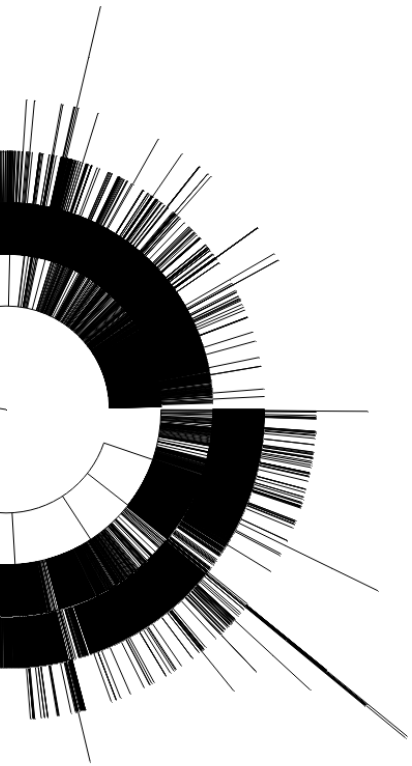
Timegates, Timemaps, and Mementos

Works with AWA, NZWA, IA, & UKWA

Systems supporting the Memento protocol provide machine-readable information about web archive captures, even if other APIs are not available. In this notebook we'll look at the way the Memento protocol is supported across four web archive repositories – the UK Web Archive, the National Library of Australia, the National Library of New Zealand, and the Internet Archive.

- Download from GitHub
- View using NBViewer
- Run live on Binder

WORKS WITH THESE ARCHIVES

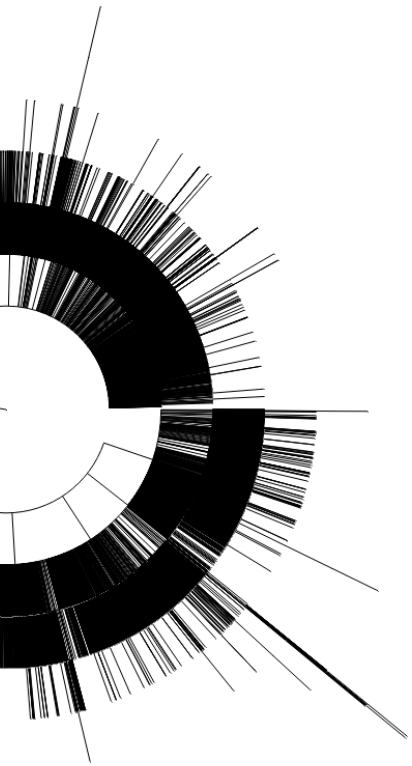


REPOSITORIES USED

- Australian Web Archive
- New Zealand Web Archive
- UK Web Archive
- Wayback Machine (Internet Archive)

But notebooks can be adapted to work with other Pywb, Open Wayback, and Memento compliant systems.

GLAM WORKBENCH



Timegates, Timemaps, and Mementos

Works with AWA, NZWA, IA, & UKWA

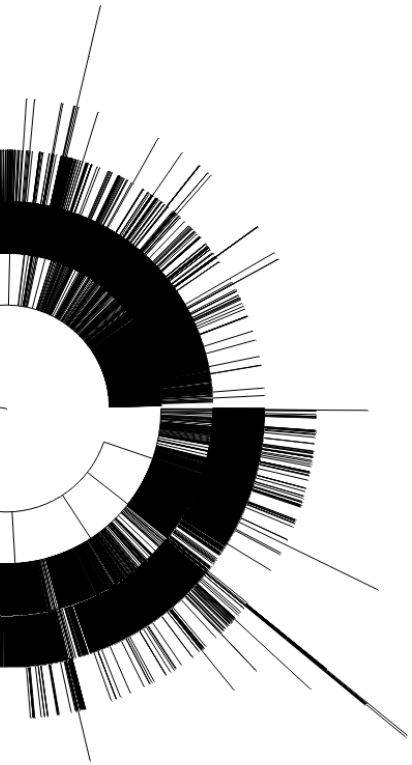
Systems supporting the Memento protocol provide machine-readable information about web archive captures, even if other APIs are not available. In this notebook we'll look at the way the Memento protocol is supported across four web archive repositories – the UK Web Archive, the National Library of Australia, the National Library of New Zealand, and the Internet Archive.

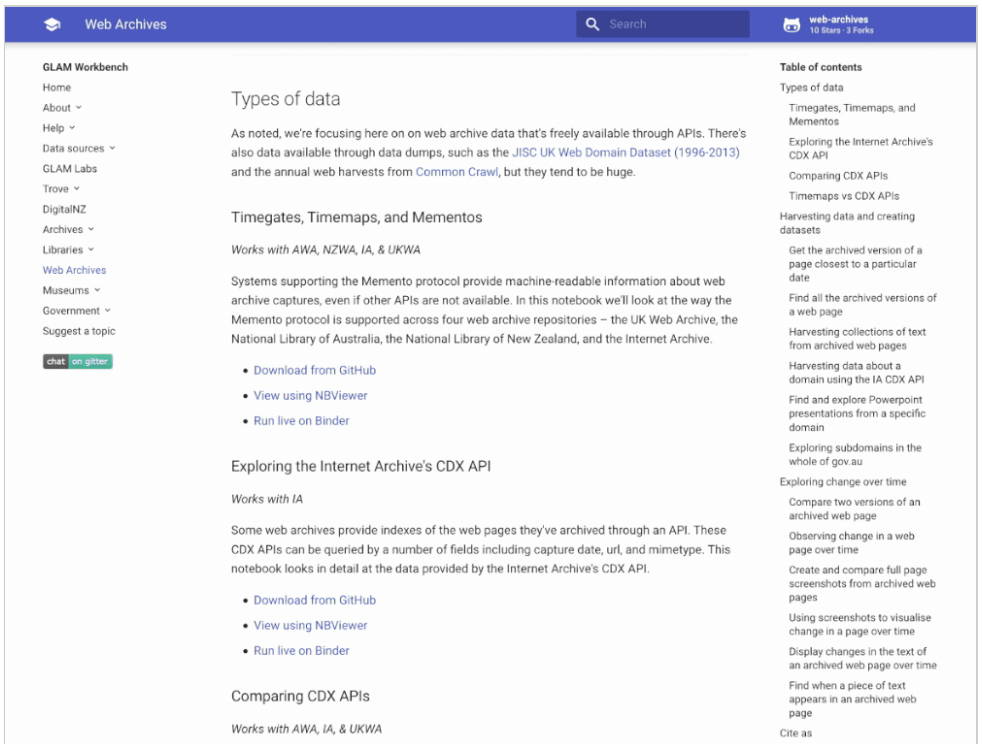
- Download from GitHub
- View using NBViewer
- Run live on Binder

RUN LIVE!

VIEW & DOWNLOAD

BINDER





Web Archives

Search

web-archives
10 Stars · 3 Forks

GLAM Workbench
Home
About
Help
Data sources
GLAM Labs
Trove
DigitalNZ
Archives
Libraries
Web Archives
Museums
Government
Suggest a topic

chat on gitter

Types of data

As noted, we're focusing here on on web archive data that's freely available through APIs. There's also data available through data dumps, such as the JISC UK Web Domain Dataset (1996-2013) and the annual web harvests from Common Crawl, but they tend to be huge.

Timegates, Timemaps, and Mementos

Works with AWA, NZWA, IA, & UKWA

Systems supporting the Memento protocol provide machine-readable information about web archive captures, even if other APIs are not available. In this notebook we'll look at the way the Memento protocol is supported across four web archive repositories – the UK Web Archive, the National Library of Australia, the National Library of New Zealand, and the Internet Archive.

- Download from GitHub
- View using NBViewer
- Run live on Binder

Exploring the Internet Archive's CDX API

Works with IA

Some web archives provide indexes of the web pages they've archived through an API. These CDX APIs can be queried by a number of fields including capture date, url, and mimetype. This notebook looks in detail at the data provided by the Internet Archive's CDX API.

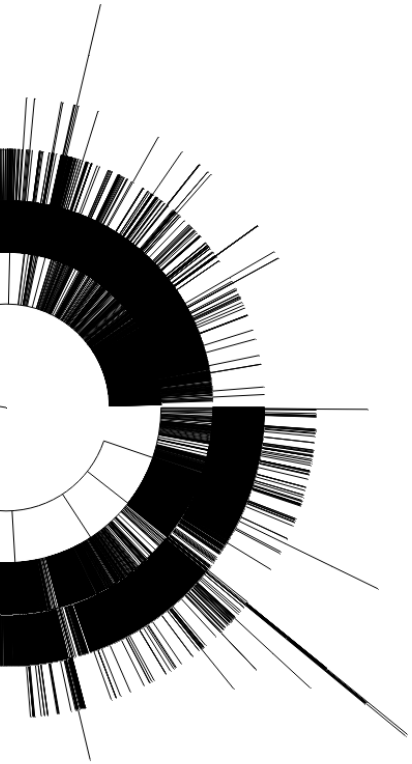
- Download from GitHub
- View using NBViewer
- Run live on Binder

Comparing CDX APIs

Works with AWA, IA, & UKWA

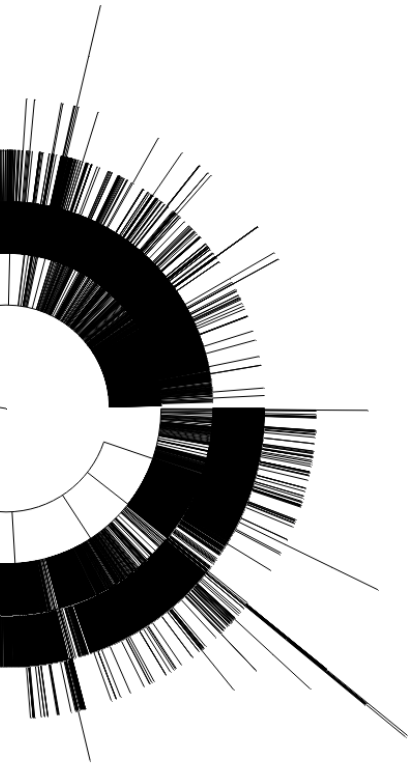
Table of contents

- builds a customised computing environment
- opens notebook ready-to-run



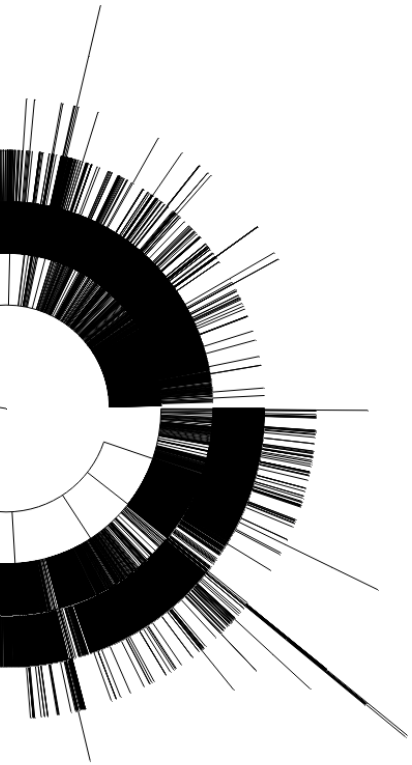
BINDER LIMITS

- inactive notebooks are closed
- notebooks and data are not saved!
- use download links in notebooks



MAIN THEMES

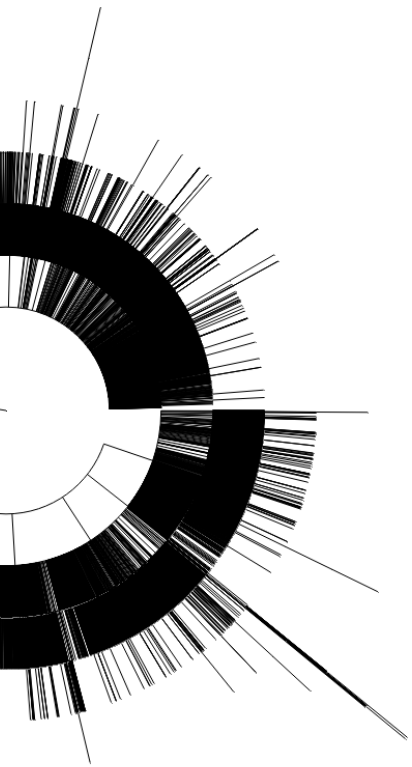
- Types of data
- Harvesting data & creating datasets
- Change over time



TYPES OF DATA

- Timegates, Timemaps, and Mementos
- Exploring the Internet Archive's CDX API
- Comparing CDX APIs
- Timemaps vs CDX APIs

TIMEMAPS & MEMENTOS



Australian Web Archive

A `HEAD` request that follows redirects returns no results

```
In [5]: query_timegate('awa', 'http://www.nla.gov.au')
```

```
https://web.archive.org.au/awa/http://www.nla.gov.au/

Out[5]: {}
```

A `HEAD` request that doesn't follow redirects returns results as expected

```
In [6]: query_timegate('awa', 'http://www.nla.gov.au', allow_redirects=False)
```

```
https://web.archive.org.au/awa/http://www.nla.gov.au/

Out[6]: {'original': 'http://pandora.nla.gov.au/pan/161756/20200306-0200/www.nla.gov.au/index.html',
         'timegate': 'https://web.archive.org.au/awa/http://pandora.nla.gov.au/pan/161756/20200306-0200/www.nla.gov.au/ind
         'timemap': 'https://web.archive.org.au/awa/timemap/link/http://pandora.nla.gov.au/pan/161756/20200306-0200/www.nl
         'memento': 'https://web.archive.org.au/awa/20200305172547mp_/http://pandora.nla.gov.au/pan/161756/20200306-0200/w
```

A query without an

```
In [7]: query_timegate
```

```
https://web.ar

Out[7]: {'original': '
         'timegate': '
         'timemap': 'h
         'memento': 'h
```

Summarising the differences

As you can see above, there are a couple of significant differences in the way that Timegates behave across the four repositories.

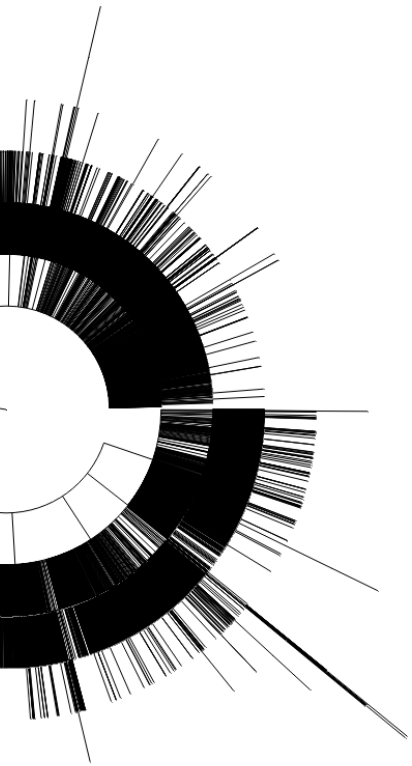
- The Wayback systems (IA and NZWA) provide more information than the Pywb systems (`first memento` , `last memento` , `prev memento` , and `last memento`)
- The UKWA and NZWA don't return a `memento` unless you include a date in the `Accept-Datetime` header. The NLA and IA return a recently captured `memento` as a default. (Though no necessarily the *most* recent?)
- You can use either `HEAD` or `GET` with UKWA and NZWA, but IA and AWA behave different depending on the type of request and whether redirects are followed. To get results from either a `HEAD` or `GET` request, AWA requests should not follow redirects. To get results from a `HEAD` requests, IA requests should follow redirects. `GET` requests to IA will return results whether or not redirects are allowed, however, those results differ.

Normalising Timegate responses and queries

Here's some code to smooth out the differences between systems, and return Memento data as a Python dictionary. Specifically it:

- Inserts the current date into requests from the UKWA or NLNZ if no date is specified. This means they behave like the other repositories that return a recent Memento.
- Follows redirects for requests to the IA.
- If there is no `memento` value in the response (as sometimes happens with NLNZ), it looks for a `first` , `last` , `prev` or `next` value instead.

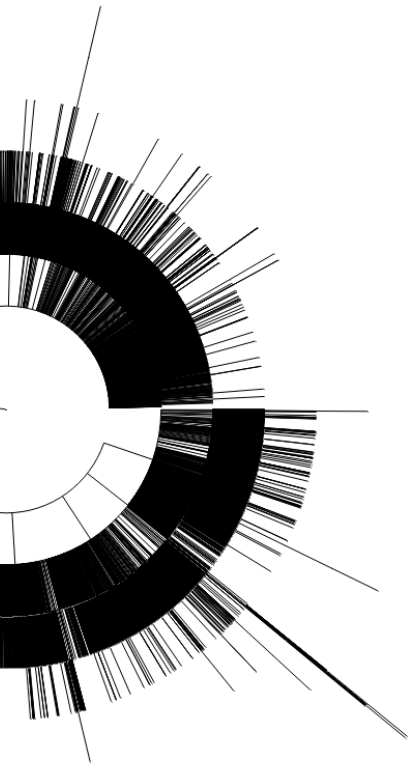
https://glam-workbench.github.io/web-archives/#timegates-timemaps-and-mementos



HARVESTING DATA

- 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

HARVESTING DATA



Harvesting collections of text from archived web pages

New to Jupyter notebooks? Try **Using Jupyter notebooks** for a quick introduction.

This notebook helps you assemble datasets of text extracted from all available captures of archived web pages. You can then feed these datasets to the text analysis tool of your choice to analyse changes over time.

Harvest sources

- Timemaps – harvest text from a single url, or list of urls, using the repository of your choice
- CDX API – harvest text from the results of a query to the Internet Archive's CDX API

Options

- `filter_text=False` (default) – save all of the human visible text on the page, this includes boilerplat
- `filter_text=True` – save only the significant text on the page, excluding recurring items like boilerp
  Trafilatura.

Usage

Using Timemaps

```
get_texts_for_url([timegate], [url], filter_text=[True or False])
```

The `timegate` value should be one of:

- `nla` – National Library of Australia
- `nlnz` – National Library of New Zealand
- `bl` – UK Web Archive
- `ia` – Internet Archive

Using the Internet Archive's CDX API

Use a CDX query to find all urls that include the specified keyword in their url.

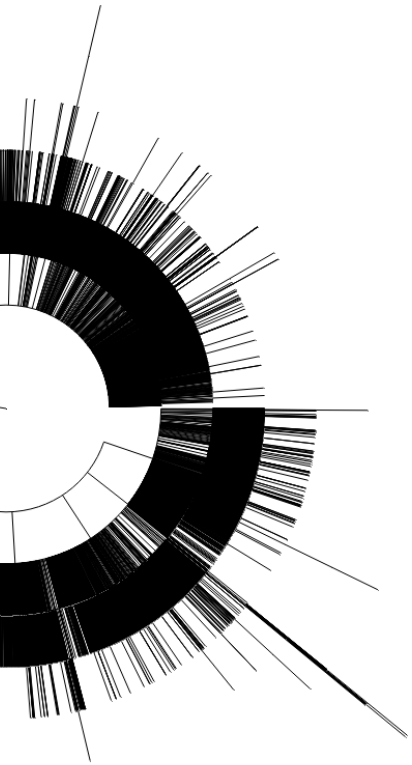
```
get_texts_for_cdx_query([url], filter_text=[True or False], fil
'statuscode:200', 'mimetype:text/html'])
```

```
1 DFAT's Procurement Policy provides for the efficient, effective, economical and ethical delivery of purchasing and
  procurement activities. Contractual arrangements entered into by the Department are conducted in accordance with the
  Commonwealth Procurement Rules & other Departmental Policies.
2 Value for Money
3 Achieving value for money is a critical consideration for the achievement of DFAT's strategic objectives. It is a
  requirement under the Public Governance, Performance and Accountability Act (2013) and the Commonwealth Procurement
  Rules. Building on these requirements DFAT has developed eight Value for Money Principles to guide decision making and
  maximise the impact of its investments.
4 List of laws, guidelines and policies
5 This document provides a List of laws and guidelines of the Commonwealth of Australia that may apply to the delivery of
  goods and services and a list of guidelines and policies contractors undertaking activities for DFAT must comply:
6 Aid Adviser Remuneration Framework
7 The Aid Adviser Remuneration Framework defines DFAT's policies and procedures for determining the remuneration of
  commercially contracted international aid advisers and outlines requirements for implementing and monitoring these
  policies.
8 The Framework ensures that adviser remuneration represents value for money. DFAT staff (and Managing Contractors
  engaging advisers on DFAT's behalf) must apply the Framework. An Adviser Remuneration Calculator is also available to
  assist.
9 Contractor and Adviser Performance Assessment for aid program agreements
10 The Aid Contractor and Adviser Performance Assessment guideline sets out DFAT's approach to managing Contractor and
   Adviser Performance Assessments for aid program agreements, and how this information will be utilised to inform future
   procurement outcomes. Templates for the assessments are also provided.
11 Eligibility criteria for contracts
12 All procurement of goods and services is subject to the
13 Commonwealth Procurement Rules, and other relevant legislation and Commonwealth Government policies.
```

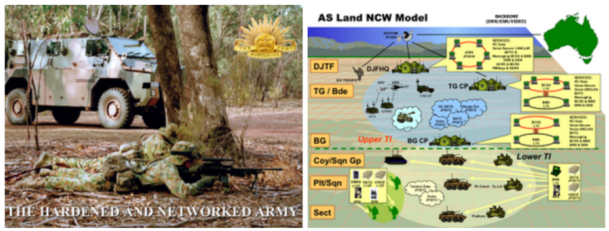

```
▾ root:
    timegate: "https://web.archive.org/web/"
    url: "http://dfat.gov.au:80/about-us/business-opportunities/tenders/Pages/dfat-procurement-policy.aspx"
    filter_text: true
    date: "2020-08-05 10:43:36"
  ▾ mementos: [] 27 items
    ▾ 0:
        url: "https://web.archive.org/web/20160511023254id_/http://dfat.gov.au:80/about-us/business-opportunities/tenders/Pages/dfat-procurement-policy.aspx"
        text_file: "text/au-gov-dfat-about-us-business-opportunities-tender/au-gov-dfat-about-us-business-opportunities-tender-20160511023254.txt"
    ▾ 1:
        url: "https://web.archive.org/web/20160602195117id_/http://dfat.gov.au/about-us/business-opportunities/tenders/Pages/dfat-procurement-policy.aspx"
        text_file: "text/au-gov-dfat-about-us-business-opportunities-tender/au-gov-dfat-about-us-business-opportunities-tender-20160602195117.txt"
    ▾ 2:
        url: "https://web.archive.org/web/20160714201451id_/http://dfat.gov.au:80/about-us/business-opportunities/tenders/Pages/dfat-procurement-policy.aspx"
        text_file: "text/au-gov-dfat-about-us-business-opportunities-tender/au-gov-dfat-about-us-business-opportunities-tender-20160714201451.txt"
    ▸ 3:
```


https://glam-workbench.github.io/web-archives/#harvesting-collections-of-text-from-archived-web-pages

HARVESTING DATA



Find and explore Powerpoint presentations from a specific domain



New to Jupyter notebooks? Try **Using Jupyter notebooks** for a quick introduction.

Web archives don't just contain HTML pages! Using the `filter` parameter in CDX queries we can limit our results to particular types of files, for example Powerpoint presentations.

This notebook helps you find, download, and explore all the presentation files captured from a particular domain, like `defence.gov.au`. It uses the Internet Archive by default, as their CDX API allows domain level queries and pagination, however, you could try using the UKWA or the National Library of Australia (prefix queries only).

This notebook includes a series of processing steps:

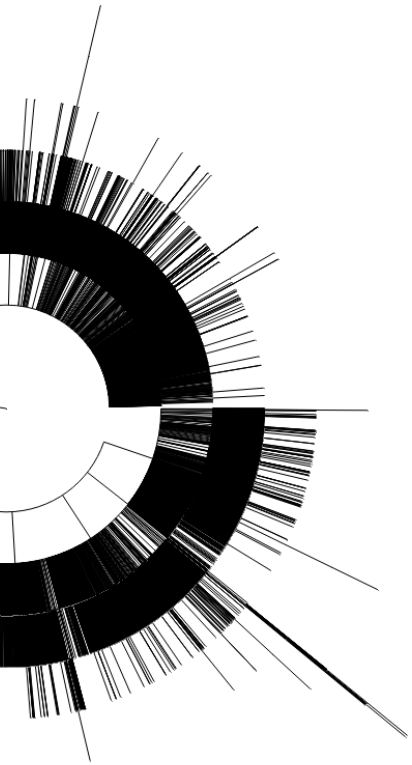
1. Harvest capture data
2. Remove duplicates from capture data and download files
3. Convert Powerpoint files to PDFs
4. Extract screenshots and text from the PDFs
5. Save metadata, screenshots, and text into an SQLite database for exploration
6. Open the SQLite db in Datasette for exploration

Here's an example of the SQLite database created by harvesting Powerpoint files from the `defence.gov.au` domain, running in Datasette on Glitch.

Moving large files around and extracting useful data from proprietary formats is not a straightforward process. While this notebook has been tested and will work running on Binder, you'll probably want to shift across to a local machine if you're doing any large-scale harvesting. That'll make it easier for you to deal with corrupted files, broken downloads etc.

https://glam-workbench.github.io/web-archives/#find-and-explore-powerpoint-presentations-from-a-specific-domain

HARVESTING DATA



Exploring subdomains in the whole of gov.au

New to Jupyter notebooks? Try **Using Jupyter notebooks** for a quick introduction.

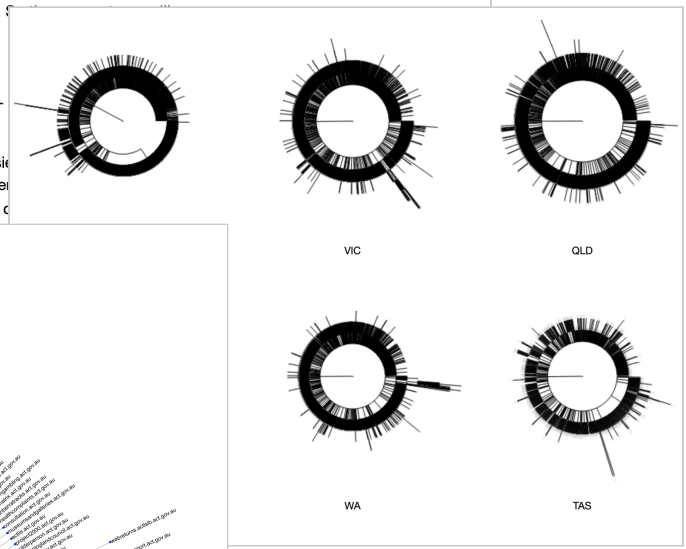
Most of the notebooks in this repository work with small slices of web archive data. In this notebook we'll scale things up a bit to try and find all of the subdomains that have existed in the `gov.au` domain. As in other notebooks, we'll obtain the data by querying the Internet Archive's CDX API. The only real difference is that it will take some hours to harvest all the data.

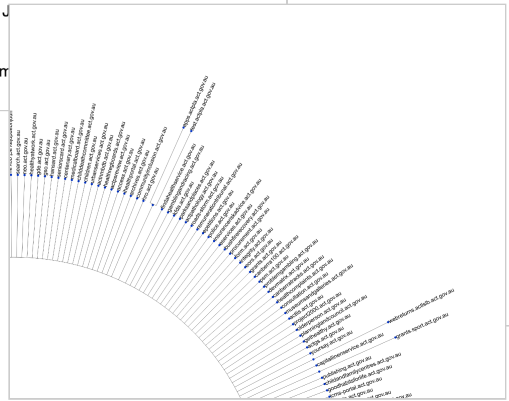
All we're interested in this time are unique domain names, so to minimise the amount of data we'll be harvesting we can make use of the CDX API's `collapse` parameter. By setting `collapse=urlkey` we can tell the CDX API to drop records with duplicate `urlkey` values – this should mean we only get one capture per page. However, this only works if the capture records are in adjacent rows, so there probably will still be some duplicates. We'll also use the `fl` to limit the fields returned, and the `filter` parameter to limit results by `statuscode` and `mimetype`.

- `url=*.gov.au` – all of the pages in all of the subdomains under `gov.au`
- `collapse=urlkey` – as few captures per page as possible
- `filter=statuscode:200,mimetype:text/html` – only successful captures of HTML
- `fl=urlkey,timestamp,original` – only these fields

Even with these limits, the query will retrieve a LOT of data. To make the harvesting process easie
`requests-cache` module. This will capture the results of all requests, so that if things get inte
requests from the cache without downloading them again. We'll also write the harvested results
The file format will be the NDJSON (Newline Delineated J
is received.

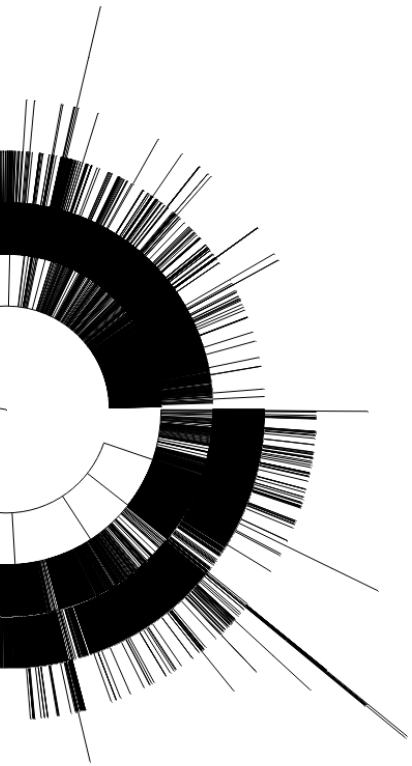
For a general approach to harvesting domain-level inform

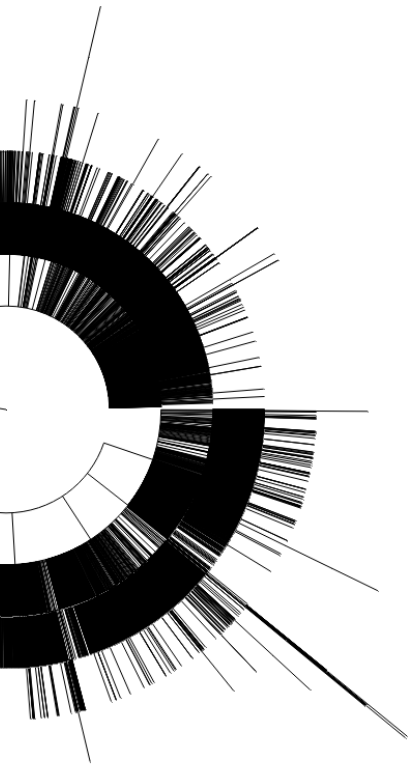




https://glam-workbench.github.io/web-archives/#exploring-subdomains-in-the-whole-of-govau

HARVESTING DATA

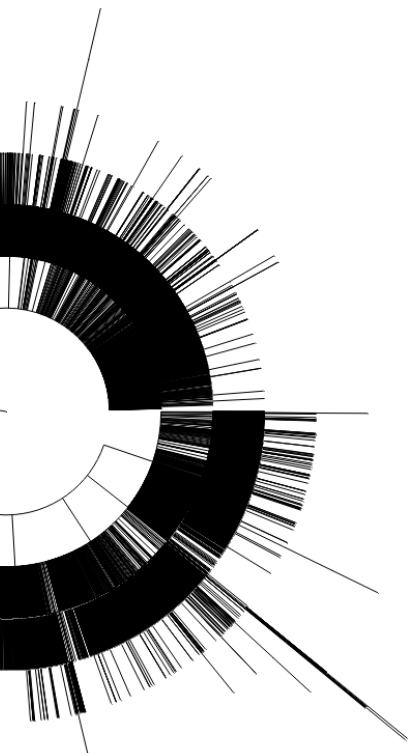




CHANGE OVER TIME

- 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

CHANGE OVER TIME



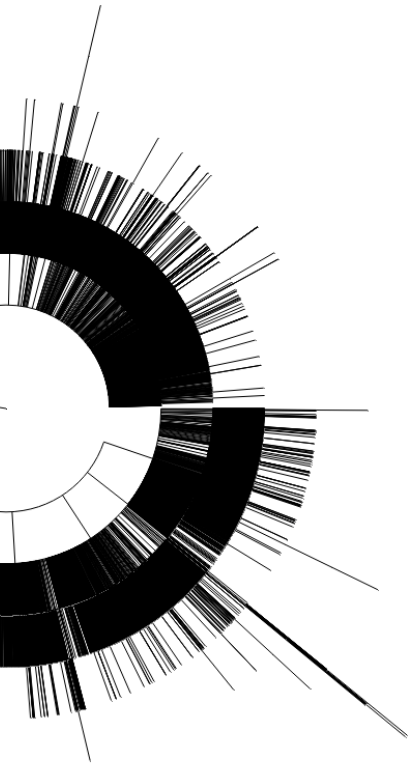
Display changes in the text of an archived web page over time ¶

Works with AWA, NZWA, IA, & UKWA

The word "Anzac" has been a part of Australian thought, language, and life since 25 April 1915. Devised by a signaller in Egypt as a useful acronym for "Australian and New Zealand Army Corps', it quickly became a word with many uses and meanings.		75	The word "Anzac" has been a part of Australian thought, language, and life since 25 April 1915. Devised by a signaler in Egypt as a useful acronym for "Australian and New Zealand Army Corps," it quickly became a word with many uses and meanings.
places: notably "Anzac area" on Gallipoli and "Anzac Cover" itself	n	78	places: notably "Anzac area" on Gallipoli and "Anzac Cove" itself
~~The word generated many slang terms in the first Australian Imperial Force (AIF) and has become a part of the Australian language.~~	n		
The term became popular largely due to the work of the official correspondent and historian Charles Bean. While still on Gallipoli he edited The Anzac book, which sold tens of thousands of copies and was reproduced with additional material in 2010. The title of the first two volumes of his official history (The story of Anzac) confirmed the word's place in Australian language.		80	The term became popular largely due to the work of the official correspondent and historian Charles Bean. While still at Gallipoli Bean edited The Anzac book, which sold tens of thousands of copies and was reproduced with additional material in 2010. Later, the title of the first two volumes of the Official History of Australia in the War of 1914-1918 (The story of Anzac) confirmed the word's place in Australian language.
The use of the word "Anzac" in Australia has been governed by federal legislation since 1920 under the Protection of Word "Anzac" Regulations.		81	The use of the word "Anzac" in Australia has been governed by federal legislation since 1921 under the Protection of Word "Anzac" Regulations.
		82	Rising Sun collar badge found at Gallipoli - https://www.awm.gov.au/collection/C1245615
Historians examining the importance of Anzac to Australia coined the phrase "Anzac legend" (or, more critically, "Anzac myth"). This refers to the representation of Australians in war, how they think, speak, and write of their war experience (which is not always the same thing as how they experienced it).	n	84	Historians examining the importance of Anzac to Australia coined the phrase "Anzac legend" (or, more critically, "Anzac myth"), referring to the representation of Australians in war; how they think, speak, and write of their war experience (which is not always the same as how they experienced it).
~~Though aspects of the legend have been criticised, there is general consensus on what is regarded as the Anzac spirit. Anzac came to stand for the qualities which Australians have seen their forces show in war. These qualities collectively make up the Anzac spirit and include endurance, courage, ingenuity, good humour, and mateship.~~		85	Though aspects of the legend have been criticised, there is general consensus on what is regarded as the "Anzac spirit". Anzac came to stand for the positive qualities which Australians have seen their forces show in war. These qualities are generally accepted to include endurance, courage, ingenuity, good humour, and mateship.
Perhaps the best (and most widely misquoted) reflection of the meaning of Anzac is to be found in Charles Bean's ~~one-volume history of Australia in the Great War,~~ Anzac to Amiens. In describing the evacuation of the Anzac area, Bean wrote:		86	Perhaps the best (and most widely misquoted) reflection of the meaning of Anzac is to be found in Charles Bean's Anzac to Amiens. In describing the evacuation of the Anzac Cove area, Bean wrote:
~~Open gallery~~	n	91	Fundraising badge: South Australia Anzac Day 'Help Those Who Helped You' - https://www.awm.gov.au/collection/C1231574
~~Photo-mechanical colour portrait entitled "The spirit of Anzac" or "The Digger". The model was Pat Hanna, who served in the New Zealand forces during the First World War, tried to recreate the "look of something between fear and defiance which we have all seen so often, and which will always remain in my memory as typical of our gallant old cobbers 'the Diggers'". P02591.001~~			

This notebook displays changes in the text content of a web page over time. It retrieves a list of available captures from a Memento Timemap, then compares each capture with its predecessor, displaying changes side-by-side.

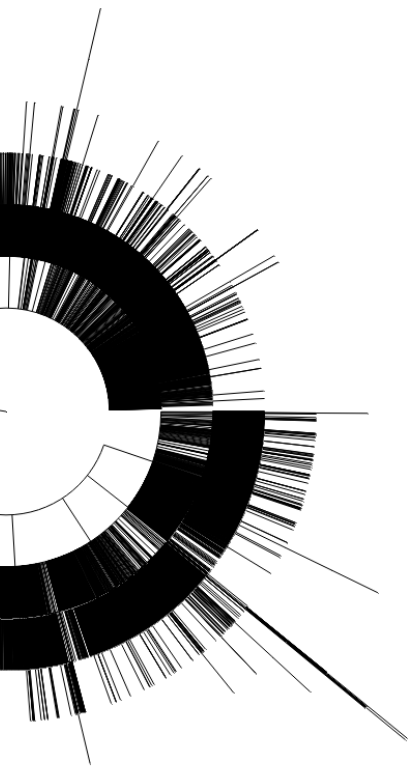
- Download from GitHub
- View using NBViewer
- Run live in Appmode on Binder



APPMODE

- hides all code cells
- runs all code cells automatically
- turns a notebook into an app

CHANGE OVER TIME



Display changes in the text of an archived web page over time

This notebook displays changes in the text content of a web page over time. It retrieves a list of available captures from a Memento Timemap, then compares each capture with its predecessor, displaying changes side-by-side.

By default, the notebook only displays lines that have *changed*. If you want to see more context, you can adjust the parameters in the `show_all_differences()` function to show lines around each change, or the complete text content.

Archive: UK Web Archive URL: https://www.gov.uk/coronavirus

Show text changes Clear all

Key

- ~~deleted text~~
- changed text
- added text

		21 March 2020			22 March 2020
t		No Differences Found	t		No Differences Found

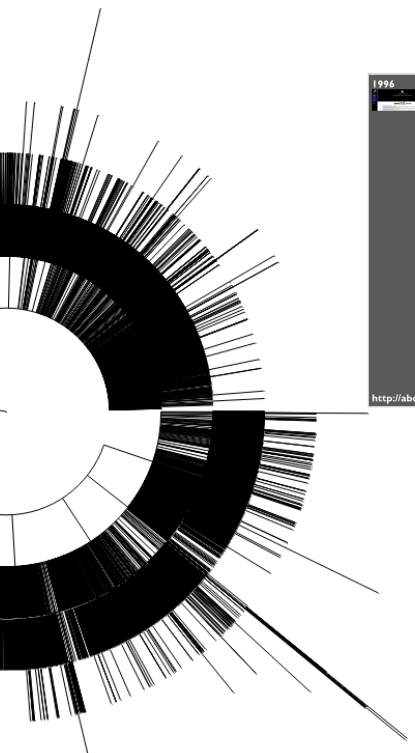
		22 March 2020			22 March 2020
t		No Differences Found	t		No Differences Found

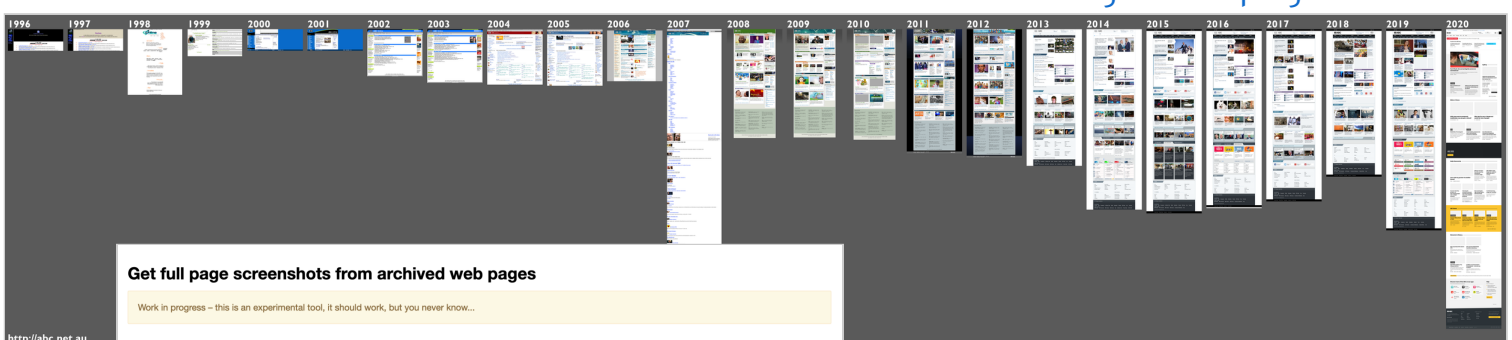
		22 March 2020			23 March 2020
n			n	16	Announcements
				17	Schools and childcare are only open to children of key workers
				18	Pubs, restaurants, and leisure venues must stay closed or they may be fined
				19	Food shops and pharmacies will remain open
n			n	27	How to protect extremely vulnerable people (shielding)
n			n	34	School closures, education, and childcare
				35	What parents and carers need to know about closures

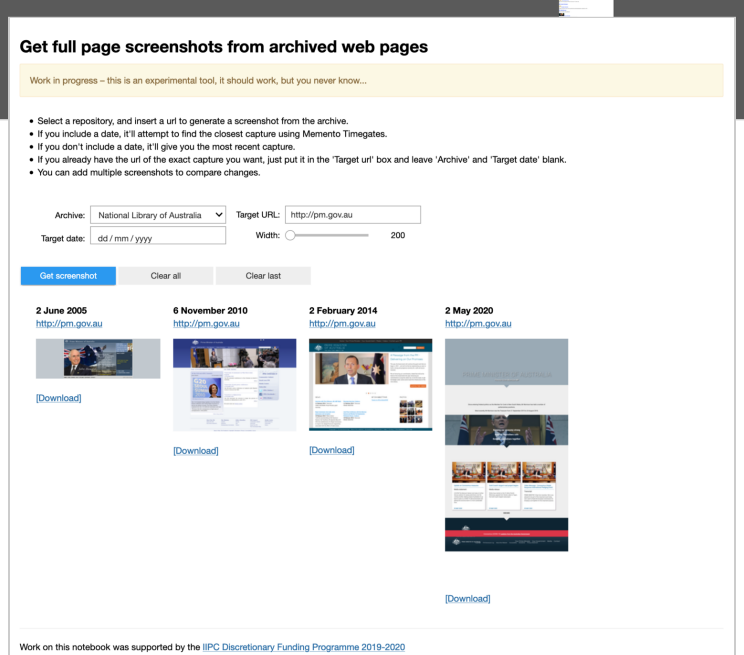
https://glam-workbench.github.io/web-archives/#display-changes-in-the-text-of-an-archived-web-page-over-time

CHANGE OVER TIME

https://glam-workbench.github.io/web-archives/#using-screenshots-to-visualise-change-in-a-page-over-time

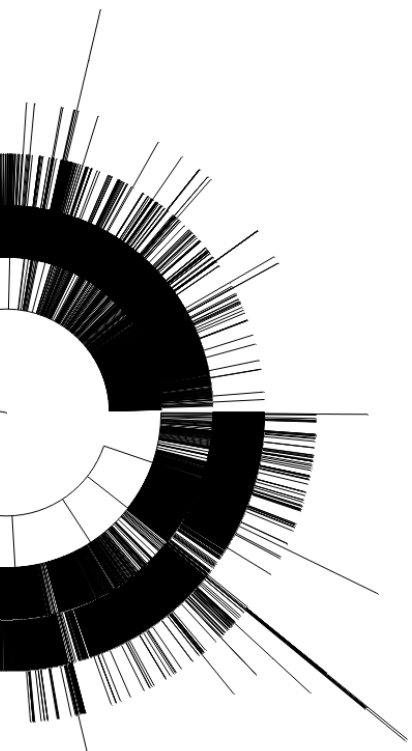






https://glam-workbench.github.io/web-archives/#create-and-compare-full-page-screenshots-from-archived-web-pages

CHANGE OVER TIME



Find when a piece of text appears in an archived web page

This notebook helps you find when a particular piece of text appears in, or disappears from, a web page. Using Memento Timemaps, it gets a list of available captures from the selected web archive. It then searches each capture for the desired text, displaying the results.

You can select the direction in which the notebook searches:

- **First occurrence** – find the first capture in which the text appears (start from the first capture and come forward in time)
- **Last occurrence** – find the last capture in which the text appears (start from present and go backwards in time)
- **All occurrences** – find all matches (start from the first capture and continue until the last)

If you select 'All occurrences' the notebook will generate a simple chart showing how the number of matches changes over time.

By default, the notebook displays possible or 'fuzzy' matches as well as exact matches, but these are not counted in the totals.

Work in progress – this is an experimental tool...

Archive: ---

URL: p://discontents.com.au/about-me

Search text: Trove

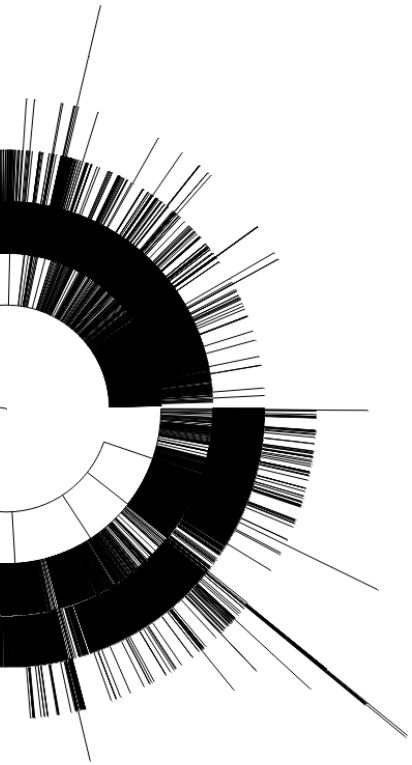
Find: First occurrence

Find text | Clear all

Work on this notebook was supported by the IIPC Discretionary Funding Programme 2019-2020

https://glam-workbench.github.io/web-archives/#find-when-a-piece-of-text-appears-in-an-archived-web-page

SUGGESTIONS? PROBLEMS?



GLAM Workbench

Search

web-archives
10 Stars · 3 Forks

GLAM Workbench

- Home
- About
- Help
- Data sources
- GLAM Labs
- Trove
- DigitalNZ
- Archives
- Libraries
- Web Archives
- Museums
- Government
- Suggest a topic

chat on gitter

Web Archives

We tend to think of a web archive as a site we go to when links are broken – a useful fallback, rather than a source of new research data. But web archives don't just store old web pages, they capture multiple versions of web resources over time. Using web archives we can observe change – we can ask historical questions. This collection of notebooks is intended to help historians, and other researchers, frame those questions by revealing what sort of data is available, how to get it, and what you can do with it.

Web Archives share systems and standards, making it much easier for researchers wanting to get their hands on useful data. 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.

Web archives are huge, and access is often limited for legal reasons. These notebooks focus on data that is readily accessible and able to be used without the need for special equipment. They use existing APIs to get data in manageable chunks. But many of the examples demonstrated can also be scaled up to build substantial datasets for analysis – you just have to be patient!

These notebooks are a starting point that I hope will encourage researchers to investigate the possibilities of web archives in more detail. They're intended to compliment the fabulous work being by projects such as Archives Unleashed to open web archives to new research uses.

The development of these notebooks was supported by the International Internet Preservation Consortium's Discretionary Funding Programme 2019-2020, with the participation of the British Library, the National Library of Australia, and the National Library of New Zealand. Thanks all!

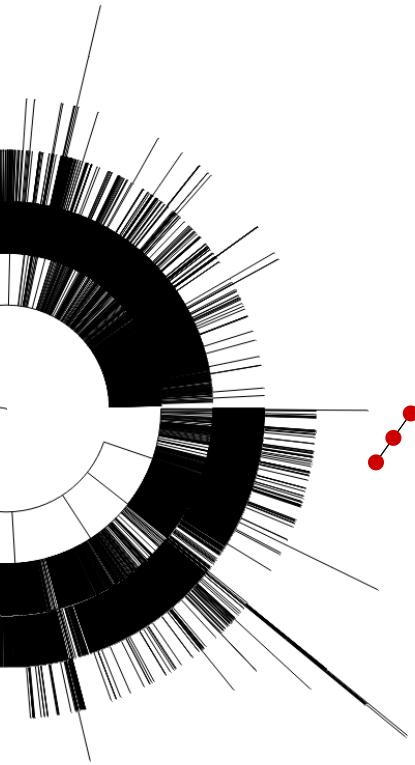
For more information on web archives projects, training, technologies, and standards see the Awesome Web Archiving list.

Explore live on Binder

Table of contents

https://glam-workbench.github.io/web-archives/

SUGGESTIONS? PROBLEMS?



GLAM-Workbench / web-archives

Watch 3 | Star 10 | Fork 3

Code | Issues | Pull requests 2 | Actions | Projects | Security | Insights

master | 3 branches | 2 tags | Go to file | Code

wragge Add Zenodo link 989edbe on Jun 15 125 commits

domains	Add govau notebook	2 months ago
getting-started	Merge commit 'c4a113a5cac82414c7c3991ee7a56b5423334e26' as...	2 months ago
images	Fix image	2 months ago
plugins	Update notebooks	3 months ago
.gitignore	Moving back to pip	3 months ago
LICENSE	Create LICENSE	2 months ago
README.md	Add Zenodo link	2 months ago
apt.txt	Add jupyter-archive	3 months ago
change_in_a_page_over_time.ipynb	Try to embed charts	2 months ago
comparing_cdx_apis.ipynb	Add getting started links	2 months ago
display-text-changes-from-timema...	Add getting started links	2 months ago
explore_presentations.ipynb	Add getting started links	2 months ago
exploring_cdx_api.ipynb	Add getting started links	2 months ago
find-text-in-page-from-timemap.ip...	Finalise and check	2 months ago
find_all_captures.ipynb	Add getting started links	2 months ago

About

No description, website, or topics provided.

Readme

MIT License

Releases 2

Bump release version for ... Latest
on Jun 15

+ 1 release

Contributors 2

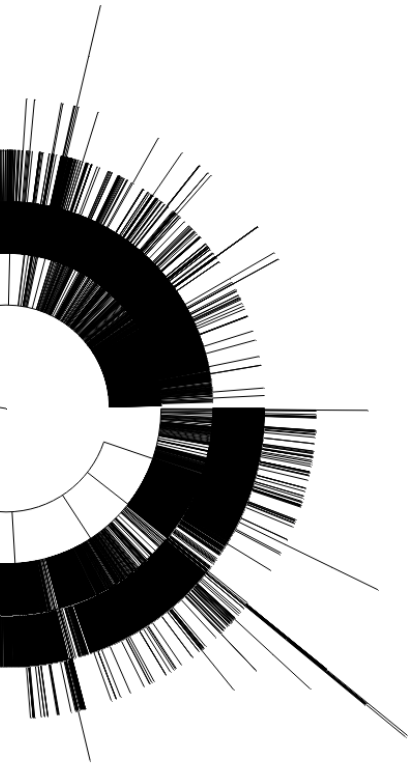
wragge wragge

anjackson anjackson

Languages

Jupyter Notebook 100.0%

https://github.com/GLAM-Workbench/web-archives



Tim Sherratt • @wragge • #glamworkbench