Exploring collections through the GLAM Workbench

SLIDE

These slides — https://slides.com/wragge/2021-explore-collection-data/

SLIDE

GLAM Workbench — my attempt to grapple with some of the challenges of ‘collections as data’?

What do I mean by ‘collections as data’? I mean the opportunities that arise when we move beyond the web interface as our means of exploring digital collections and instead directly manipulate the data underneath those interfaces.

Here’s an example…

SLIDE

We’re all familiar with searching collections. We go somewhere like Trove, which provides more than 200 million digitised articles from Australian newspapers, and we type queries in the search box.

We then browse through the results, clicking on ones that look interesting. Through this process we gradually assemble a list of sources that seem most relevant to our interests.

SLIDE

My question is — what about the other 3 million?

How do we make sense of that?

Digital collections raise questions of scale and meaning that we’re still trying to grapple with.

SLIDE

Here’s the same search, but this time, instead of seeing just the first 20 results, we’re seeing everything — the total number of matching results per year.

Seeing the search data presented this way means that we can start to formulate different types of questions, we can explore change over time, shifts in language and meaning, the impact of historical events.

SLIDE

We can also compare results in a way that is impossible through a conventional search interface.

This opening up of possibilities, this enlargement of our scope for formulating and pursuing questions is to me what makes the idea of collections as data so exciting and important.

It offers new ways of *seeing* the collections of libraries, archives and museums that aren't constrained by carefully designed web interfaces.

This example is possible because Trove makes its data available through an API (an Application Programming Interface).

SLIDE

To support this sort of research, GLAM organisations are making more and more of their data available — through APIs, through data dumps, csv files, text or image collections, and other ways. This is fabulous!

But once you've created your API, or shared your CSV, then what?

There's still a gap between the data, which you've invested precious resources in documenting and sharing, and its use by researchers.

We might think that this gap can be filled by better documentation by GLAM organisations, and by more digital skills training for researchers.

Certainly these can help answer the 'how' questions — how can I query this API? How can I calculate the the frequencies of words in this text?

But what's still missing is the 'why'? Why would a researcher be interested in your data? Why should they invest time and effort in developing their digital skills, in understanding the contents of your CSVs, in learning the idiosyncrasies of your API? What's in it for them?

SLIDE

This is where I hope that things like the GLAM Workbench can help — by providing collections of examples, reservoirs of possibility. Not just descriptions of how to use the data, but glimpses of the possibilities that might emerge from the data. Those moments of excitement when you see your research questions from a totally different perspective.

GLAM Workbench

SLIDE

The GLAM Workbench provides a wide range of entry points into the digital collections of galleries, libraries, archives and museums. Mostly focused on Aust & NZ collections at the moment, but with some examples that are more broadly applicable.

SLIDE

These entry points take many forms:

- self-contained tools or apps
- tutorials
- examples of working code

- quick hacks or workarounds for problems

This diversity is deliberate. It takes time to develop digital skills and confidence, and people are going to be arriving at the GLAM Workbench at different points along this journey.

Someone who really hasn't worked with digital data before, could fire up an app. Get out something useful. Start to appreciate the possibilities.

Someone who wants to understand the technology behind an app might work through a tutorial that explains how the data is structured and queried.

Someone who has coding skills might just drop in for snippet of code that helps them make use of a collection API.

The GLAM Workbench is *LIVE* – all of the examples can be run live in the cloud, or a variety of other environments. The aim is to encourage experimentation. Don't just read about what's possible with collection data, **try it, now!**

The GLAM Workbench is meant to be edited, re-used, and hacked. All the code is accessible and openly licensed. Not intended to be the complete guide to collections as data **–it's a starting point** to encourage your own explorations

SLIDE

The GLAM Workbench is not:

- coding 101 — not a 'learn to code' site, but by working through examples, you will gain an understanding of what code can do, and you will start to identify patterns, and gain some confidence in seeing what happens when you change things.
- finished or perfect — always a work in progress

SLIDE

GLAM Workbench tries to address these sorts of questions:

- possibilities – why should I be interested?
- starting points – can you give me an example I can use?
- pathways – where do I go next?

Jupyter

As I've said, the GLAM Workbench contains **apps** and **tutorials**, it's **live** and **hackable**.

This is all possible because the GLAM Workbench tools and examples are all shared as Jupyter notebooks.

SLIDE

This is a Jupyter notebook:

- code, outputs, explanations
- step you through the possibilities

SLIDE

Why is Jupyter so useful?

- runs live code in your browser
- blurs the line between tutorial and tool — can learn while you're actually doing something useful

SLIDE

Jupyter in GLAMs

- more and more institutions using Jupyter to help users explore their collections

SLIDE

We're going to look at some examples now, and as you'll see, Jupyter notebooks lend themselves to a diversity of uses — the way they look can be changed by extensions or themes. They can be tailored to different audiences. Run on different platforms.

Finding GLAM data

SLIDE

Trove newspaper harvester

SLIDE

- Get newspaper articles in bulk — metadata, OCRd text, images
- Hundreds of thousands of articles
- Create your own dataset that can be used with a text analysis tool

SLIDE

- example — half a million articles including the word 'immigrants'
- extracted words before 'immigrants' - limited to adjectives describing place of origin

RecordSearch harvester

SLIDE

- Not every collection has an API
- NAA online database, RecordSearch, has lots of rich data but no API
- Screenscraper to extract structured data
- Also images from digitised files

SLIDE

- Harvested metadata of surveillance files
- downloaded more than 200,000 page images
- found redactions
- watch the video for more!

Bulletin cartoons

SLIDE

- GLAM Workbench provides some pre-harvested datasets, as well as the methods used to create them
- Trove digitised journals — different system to newspapers
- High res images — not accessible through API
- Bulletin editorial cartoons
- Different entry points — dataset of images, compiled into PDFs

Asking different questions

SLIDE

Trove newspapers over time — app and notebook

SLIDE

- Examples we started with — visualising a complete set of search results
- Possibilities for exploring change over time — language, technology, events
- Simple app, and notebook tutorial

DigitalNZ open collections

SLIDE

- Digital NZ is an aggregator (like Trove and Europeana)
- Look across the whole of the aggregated collection to find pockets of material that’s openly licensed or out of copyright
- Fireworks!

Counting words (Hansard)

SLIDE

- Words spoken in the Australian Parliament are recorded in Hansard, available online through a database
- Well structured XML underneath, but difficult to aggregate — created a repository that brought it all together

- Examples of what you can do with that aggregated content

Web page screenshots

SLIDE

- Web archives document change over time, will be vital for historical research
- But the amount of data can be intimidating
- Web archives section provides examples to get researchers started — what sorts of questions can you ask
- Not just Aust and NZ — IA and UKWA, can also be extended further
- Here looking at how the design of a page changes over time

Tracking text

SLIDE

- Another web archive example
- Looking when particular pieces of text appear or disappear from a page
- Runs as an app

Hacking heritage

Random DigitalNZ

SLIDE

- On my wishlist for collection APIs is a random sort feature
- Encourages play and serendipitous discovery
- When it’s not built in, we can create workarounds such as this for DigitalNZ

Non-English newspapers

SLIDE

- Trove contains a growing number of non-English language community newspapers – Chinese, French, German, Croatian & more
- But there’s no easy way of finding these through the interface, and keyword searches are unlikely to surface them
- Notebook uses a language detection library and sampling of articles to identify newspapers with non-English content
- Created a list of newspapers

Scissors and paste

SLIDE

- The GLAM Workbench also provides some examples of how you can get creative with collections
- Embedded with the HTML pages that display newspaper articles in Trove are the coordinates of each OCRd word
- Can use this to snip out images of words from a search
- Build up and download a message!

Bringing documentation alive

Random Museum Vic

SLIDE

- Museums Victoria **does** provide a random sort option in the collection API
- A quick example to show how little code is necessary to create a random object viewer using the API

Explore NMA API

SLIDE

- NMA has an API that provides rich collection data
- Notebooks work through both how to formulate API requests, but also **the structure of the data itself**
- Temporal and spatial analysis

Memento

SLIDE

- Most web archives comply with the Memento protocol, which means they have a baked in API for delivering web page captures
- one near a particular data, a Memento
- a list of captures of a page – a Timemap
- Notebooks work through what's expected according to the Memento protocol, and also what's actually delivered by different systems
- Again — documentation with live practical examples that can be adapted and extended

Pathways

SLIDE

Hopefully you can start to see from these examples the ways that I’m trying to build a range of different pathways through online collections.

Building confidence

SLIDE

- One aim of all the notebooks is really to build confidence
- From getting started notebook — real examples from NMA
- Have a go, change things, see what happens

Apps and notebooks

SLIDE

- But also provide alternative interfaces to particular tasks
- Apps and notebook views
- While its aimed at developing skills and confidence, there may be Workbench users who have no interest in the code and just make use of the apps — that’s fine too!

Running notebooks

SLIDE

The existence of multiple pathways is also reflected in ways users can actually use and run the notebooks.

Binder

SLIDE

- Every section in the GLAM Workbench has a button to open the notebooks in Binder
- Every notebook has a link to open in Binder
- Use notebooks live in the cloud — no waiting?

SLIDE

- What does Binder do?
- Read requirements, spins up a customised computing environment
- All the software needed to run the notebooks
- Great for experimentation and exploration
- **Click and go!**

In the context of the GLAM Workbench, Binder is critical because it encourages exploration by making it easy for users to ‘just try it’.

One click and they have a live, but safe and structured environment where they can experiment. No software to install, no command line to navigate. Just click, then do things. Real things.

Going back the Workbench's aims to expose possibilities, to make the 'why' visible, being able to do real work in a learning environment is really important.

- But there are limits...

Reclaim Cloud

SLIDE

- Still under construction! Rolling out to repositories.
- One click install — create your own persistent environment
- Easy to manage costs

Docker

SLIDE

- Run locally but without having to manage a complete environment
- Free, but requires more technical knowledge

Building these sorts of pathways and possibilities to meet people where they are in terms of their skills and experience is something I spending a lot of time on at the moment.

GLAM Workbench is open

SLIDE

- Code on GitHub
- Data on GitHub or other free cloud services
- Openly licensed
- Repositories preserved in Zenodo (citable via DOI)
- Please copy, edit, develop, re-use!