SMALL DATA & HAND-CRAFTED INFRASTRUCTURES

Alternative models for digital research in the humanities

Tim Sherratt ⌘ @wragge

Speaker notes

WHAT IS (E)RESEARCH INFRASTRUCTURE?

Speaker notes

Something of a fraud -- my inclination is to build, rather than provoke...

But, strangely enough, I do get quite passionate about the nature of research infrastructure, so who knows??

BIG MACHINES!

> “ *Raijin, named after the Shinto God of thunder, lightning and storms*

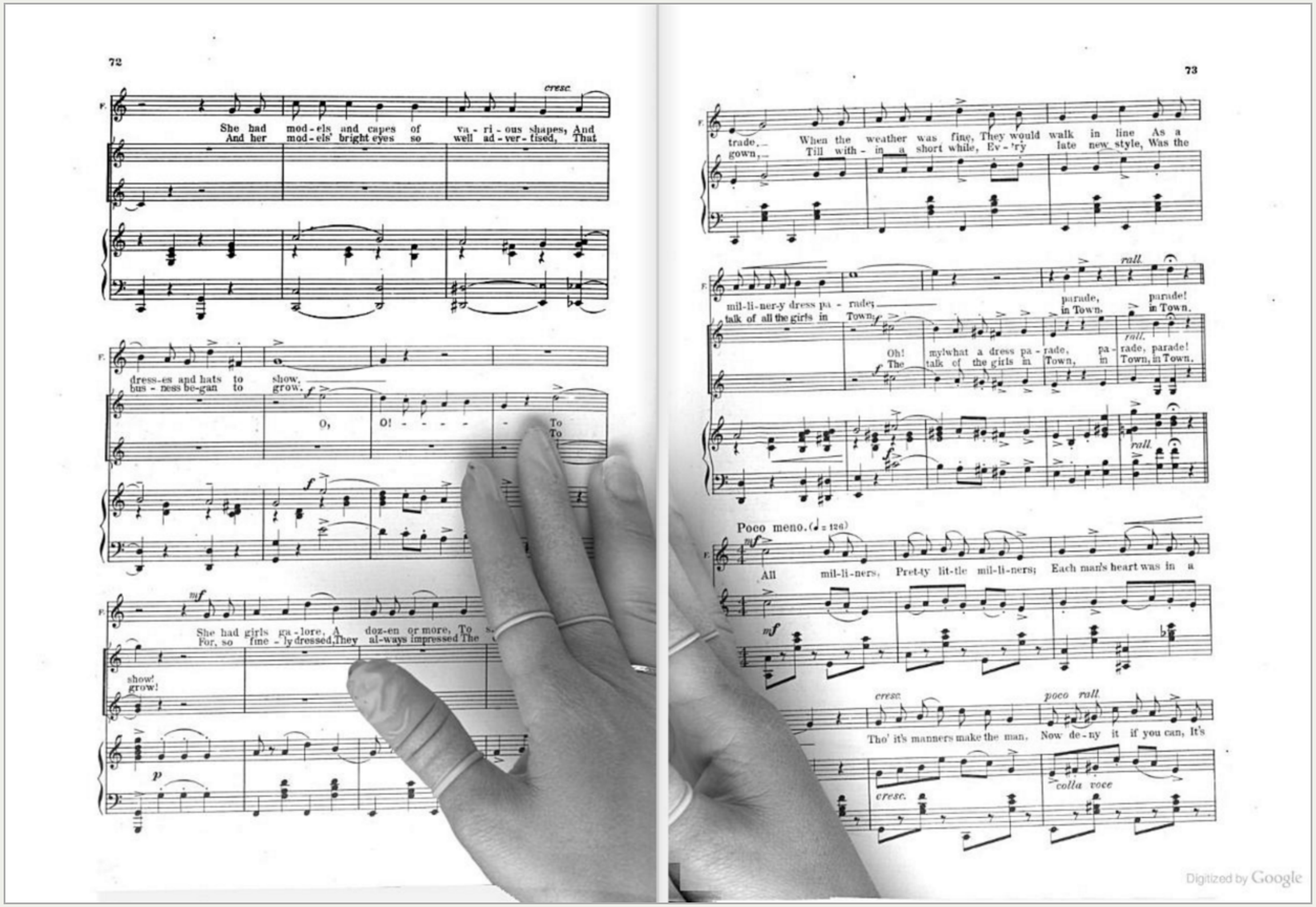
Speaker notes

Perhaps something like this?

Big machines!

NCI at ANU

THE ART OF GOOGLE BOOKS


72
73
She had mod-els and capes of va-ri-ous shapes, And
And her mod-els' bright eyes so well ad-ver-tised, That
dress-es and hats to show,
bus-ness be-gan to grow.
O, O!
To
She had girls ga-lore, A doz-en or more, To
For, so fine-ly dressed, They al-ways impressed The
show!
grow!
When the weather was fine, They would walk in line As a
Till with-in a short while, Ev-'ry late new style, Was the
mil-li-ner-y dress pa-rade;
talk of all the girls in Town;
parade, in Town,
parade! in Town.
Oh! my! what a dress pa-rade, pa-rade, parade!
The talk of the girls in Town, in Town, in Town.
Poco meno.
All mil-li-ners, Pret-ty lit-tle mil-li-ners; Each man's heart was in a
Tho' it's manners make the man, Now de-ny it if you can, It's
cresc.
poco rall.
colla voce
Digitized by Google

http://theartofgooglebooks.tumblr.com/post/154586958155/employees-hand-from-p-72-73-of-red-feather-a

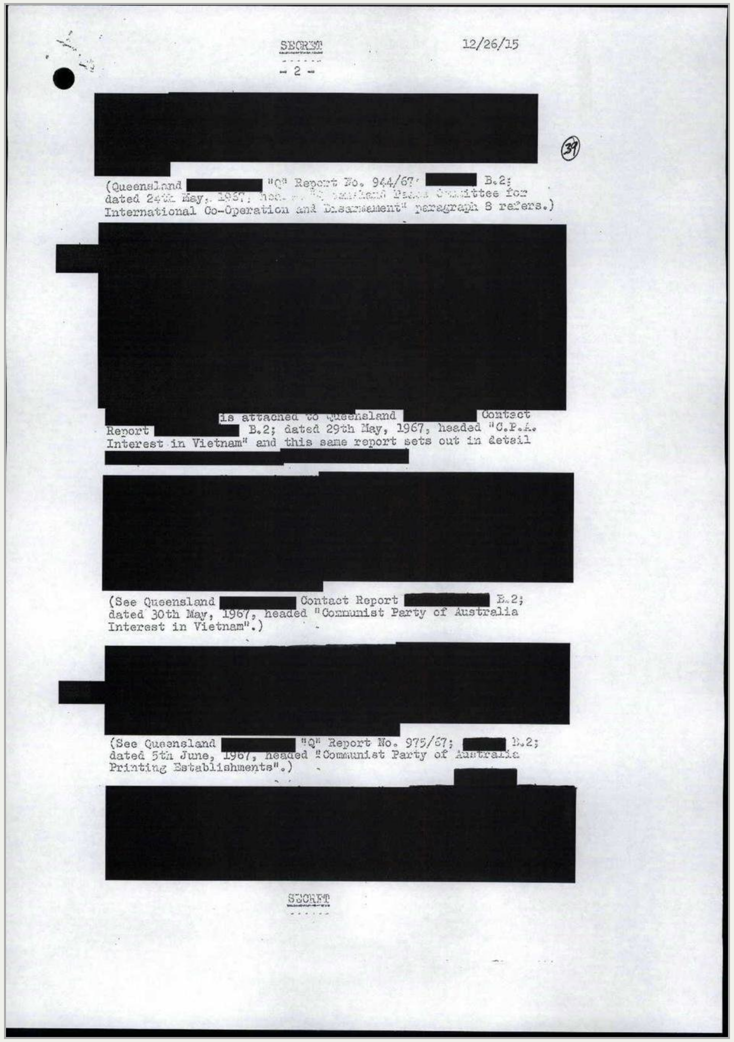
Speaker notes

Google Books project -- started with massive ambitions to digitise world's libraries.

But what now?

Reminder that even with large scale digital projects we can't just assume that they're going to stay around. Size isn't necessarily an indicator of security or stability.

ASIO FILES



SECRET 12/26/15

- 2 -

[redacted]

(Queensland [redacted] "Q" Report No. 944/67 [redacted] B.2; dated 24th May, 1967, headed "Queensland Peace Committee for International Co-Operation and Disarmament" paragraph 8 refers.)

[redacted] is attached to Queensland [redacted] Contact Report [redacted] B.2; dated 29th May, 1967, headed "C.P.A. Interest in Vietnam" and this same report sets out in detail [redacted]

[redacted]

(See Queensland [redacted] Contact Report [redacted] B.2; dated 30th May, 1967, headed "Communist Party of Australia Interest in Vietnam".)

[redacted]

(See Queensland [redacted] "Q" Report No. 975/67; [redacted] B.2; dated 5th June, 1967, headed "Communist Party of Australia Printing Establishments".)

[redacted]

SECRET

39

http://owebrowse.herokuapp.com/items/4025145/pages/32/

Speaker notes

Human dimensions of large digital undertakings is something that frequently pops up in my own research.

Might remember my interest in ASIO files in the NAA.

Downloaded more than 300,000 images -- extracted information about redactions.

#REDACTIONART

https://owebrowse.herokuapp.com/redactions/tags/art/

https://owebrowse.herokuapp.com/redactions/tags/art/

Speaker notes

A reminder that there are people in these machines...

SMALL PARTS LOOSELY COUPLED

Speaker notes

I want to focus not on the big machines — but examine research infrastructure as a web of interlocking tools, technologies, policies, and (most importantly) people

Small parts loosely coupled — a design philosophy, an approach to building digital systems

Theme for this talk

AS SIMPLE AS...

https://trove.nla.gov.au/newspaper/

https://trove.nla.gov.au/newspaper/

Speaker notes

Search the most ubiquitous and transformative research infrastructure — particularly coupled with OCR

Trove -- ability to search the contents of more than 200 million newspaper articles from 1803 onwards

Just so easy and so natural now, that we rarely think of how an individual search result is constructed.

MANUFACTURING A SEARCH RESULT

MANUFACTURING A SEARCH RESULT

- Preservation of print copies

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm
- Segmentation of pages

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm
- Segmentation of pages
- Recording article metadata

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm
- Segmentation of pages
- Recording article metadata
- Optical character recognition

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm
- Segmentation of pages
- Recording article metadata
- Optical character recognition
- Storage of images and metadata

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm
- Segmentation of pages
- Recording article metadata
- Optical character recognition
- Storage of images and metadata
- Indexing of OCR and metadata

MANUFACTURING A SEARCH RESULT

- Preservation of print copies
- Cataloguing
- Selection for microfilming
- Microfilming
- Selection for digitisation
- Create digital images from microfilm
- Segmentation of pages
- Recording article metadata
- Optical character recognition
- Storage of images and metadata
- Indexing of OCR and metadata
- Discovery interface

Speaker notes

Not automatic -- decisions are made at every point

Nothing natural or inevitable about the outcome

WHY THE PEAK?

https://plot.ly/~wragge/472.embed

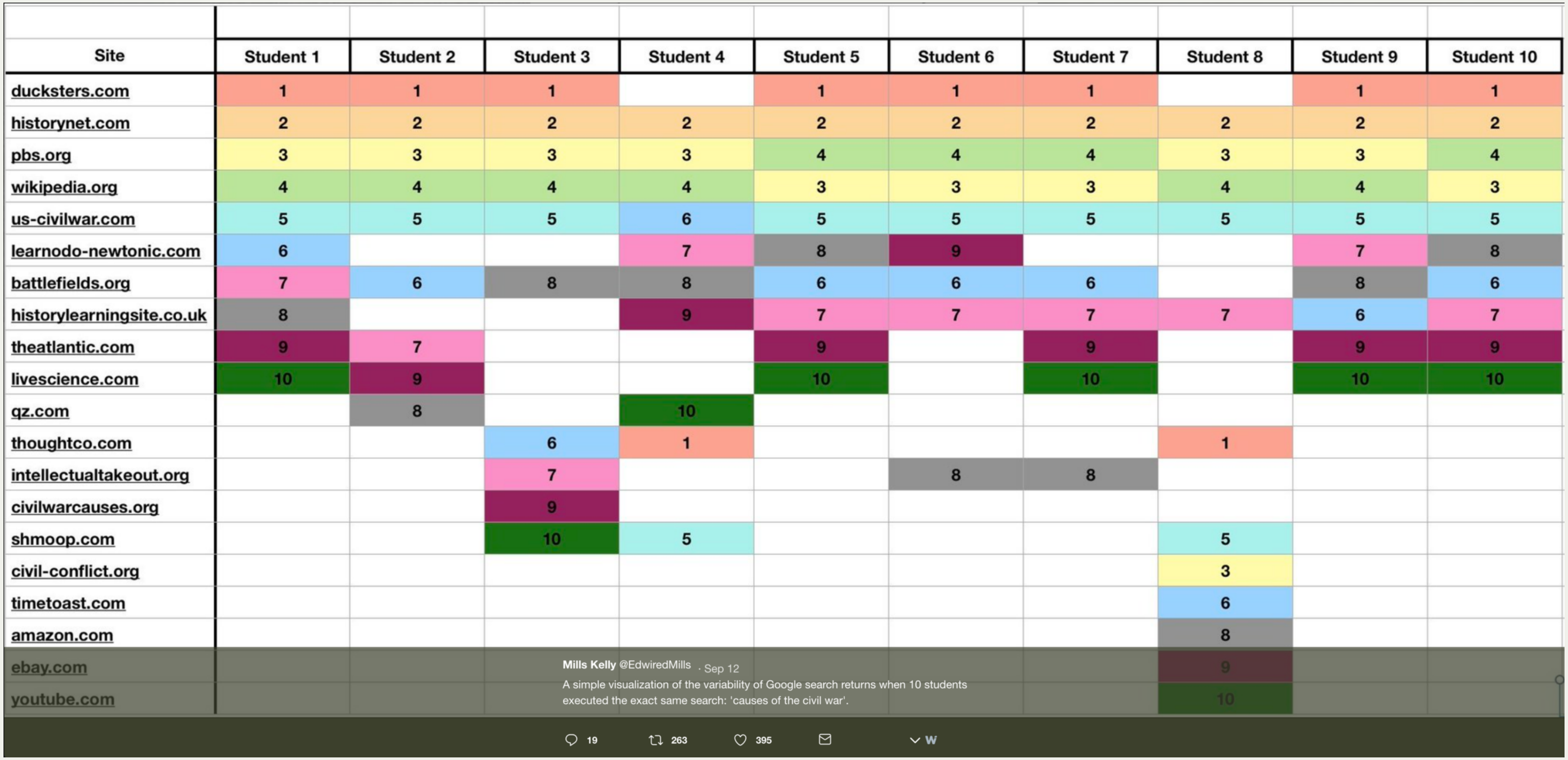
https://plot.ly/dashboard/wragge:494/present

Speaker notes

Why the peak?

Decisions have been made. Collections are constructed.

IT JUST WORKS...?



Site	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7	Student 8	Student 9	Student 10
ducksters.com	1	1	1		1	1	1		1	1
historynet.com	2	2	2	2	2	2	2	2	2	2
pbs.org	3	3	3	3	4	4	4	3	3	4
wikipedia.org	4	4	4	4	3	3	3	4	4	3
us-civilwar.com	5	5	5	6	5	5	5	5	5	5
learnodo-newtonic.com	6			7	8	9			7	8
battlefields.org	7	6	8	8	6	6	6		8	6
historylearningsite.co.uk	8			9	7	7	7	7	6	7
theatlantic.com	9	7			9		9		9	9
livescience.com	10	9			10		10		10	10
qz.com		8		10						
thoughtco.com			6	1				1		
intellectualtakeout.org			7			8	8			
civilwarcauses.org			9							
shmoop.com			10	5				5		
civil-conflict.org								3		
timetoast.com								6		
amazon.com								8		
ebay.com								9		
youtube.com								10		

Mills Kelly @EdwiredMills · Sep 12

A simple visualization of the variability of Google search returns when 10 students executed the exact same search: 'causes of the civil war'.

19 263 395

https://twitter.com/EdwiredMills/status/1039629962813366275

Speaker notes

The power of Google encourages us to put a lot of trust in search boxes -- they just work...

But what about when it doesn't?

From US historian Mills Kelly -- asked students to search Google for 'causes of the Civil War'

TROVE MAKES RESEARCH TOO EASY

Speaker notes

Trove makes research too easy

I’ve heard arguments like this from a number of senior historians, and I understand their point — it’s so easy to find things of relevance that you might not ask **why** you are finding these particular things, and what do they mean in context.

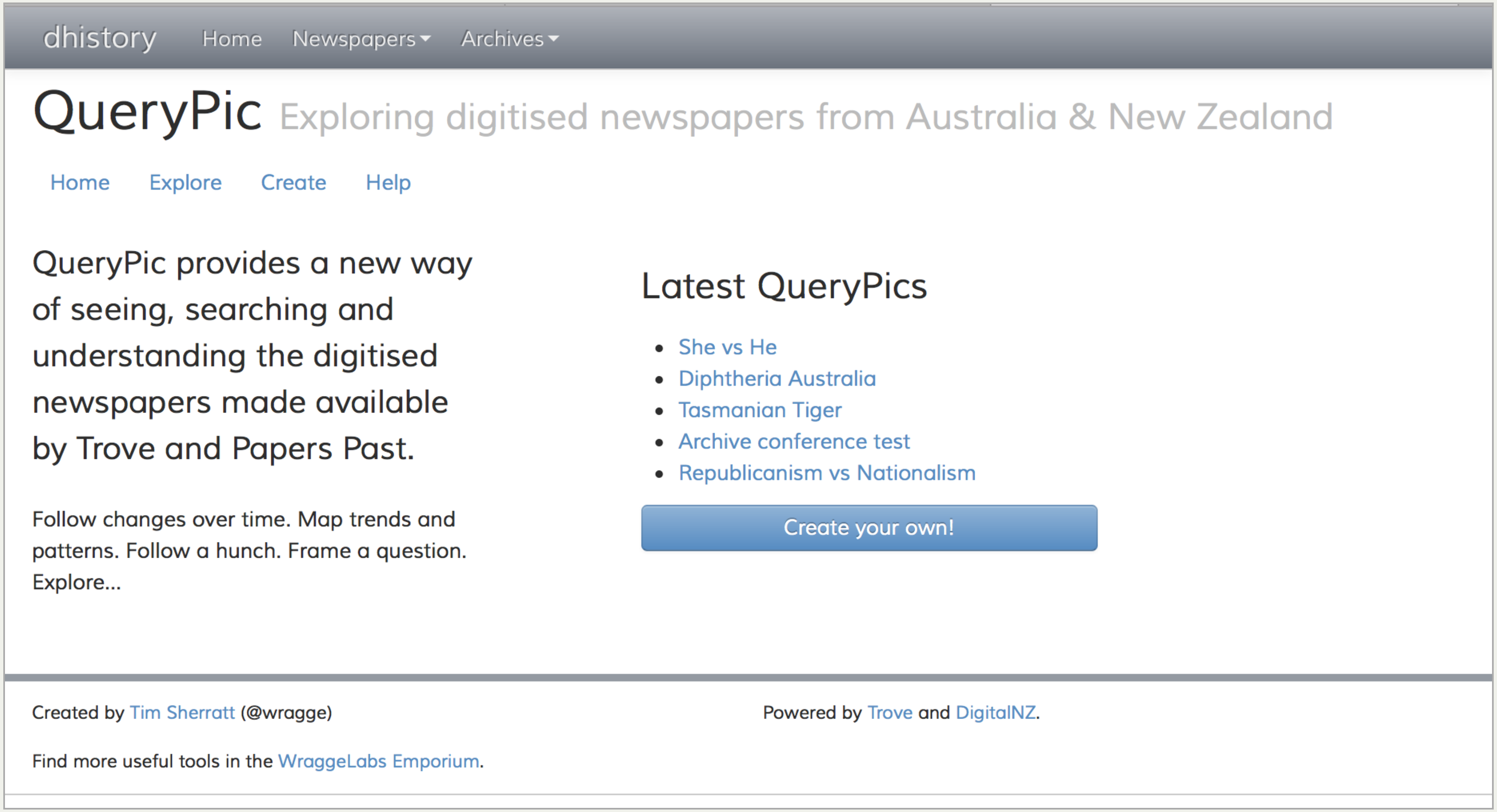
TRANSFORMING RESEARCH?

Transforming research

This to me points to one of the major failings of discussion around research infrastructure — the focus on **transformation**, the idea that these new tools and resources change the nature of research.

Most of us are never going to regard ourselves as ‘digital’ researchers (whatever that might mean), we just want to

THE CHALLENGE OF ABUNDANCE



QueryPic Exploring digitised newspapers from Australia & New Zealand

QueryPic provides a new way of seeing, searching and understanding the digitised newspapers made available by Trove and Papers Past.

Follow changes over time. Map trends and patterns. Follow a hunch. Frame a question. Explore...

Latest QueryPics

- She vs He
- Diphtheria Australia
- Tasmanian Tiger
- Archive conference test
- Republicanism vs Nationalism

Created by Tim Sherratt (@wragge)

Powered by Trove and DigitalNZ.

Find more useful tools in the WraggeLabs Emporium.

http://dhistory.org/querypic/

Speaker notes

One set of questions/problems revolve around the challenge of abundance — there’s so much stuff

I’ve developed tools like QueryPic and the Trove Newspaper harvester to help researchers zoom out the typical results page and get a bigger picture.

But every tool brings with it a new set of constraints.

WARNING LIVE CODE AHEAD!

launch binder

Speaker notes

This is probably where my talk get’s most provocative as I’m actually going to ask you to look at some code!

But let’s try asking some questions of Trove

SWITCH TO NOTEBOOK

Speaker notes

What have I been using?

Jupyter notebooks

Can be hosted on MyBinder

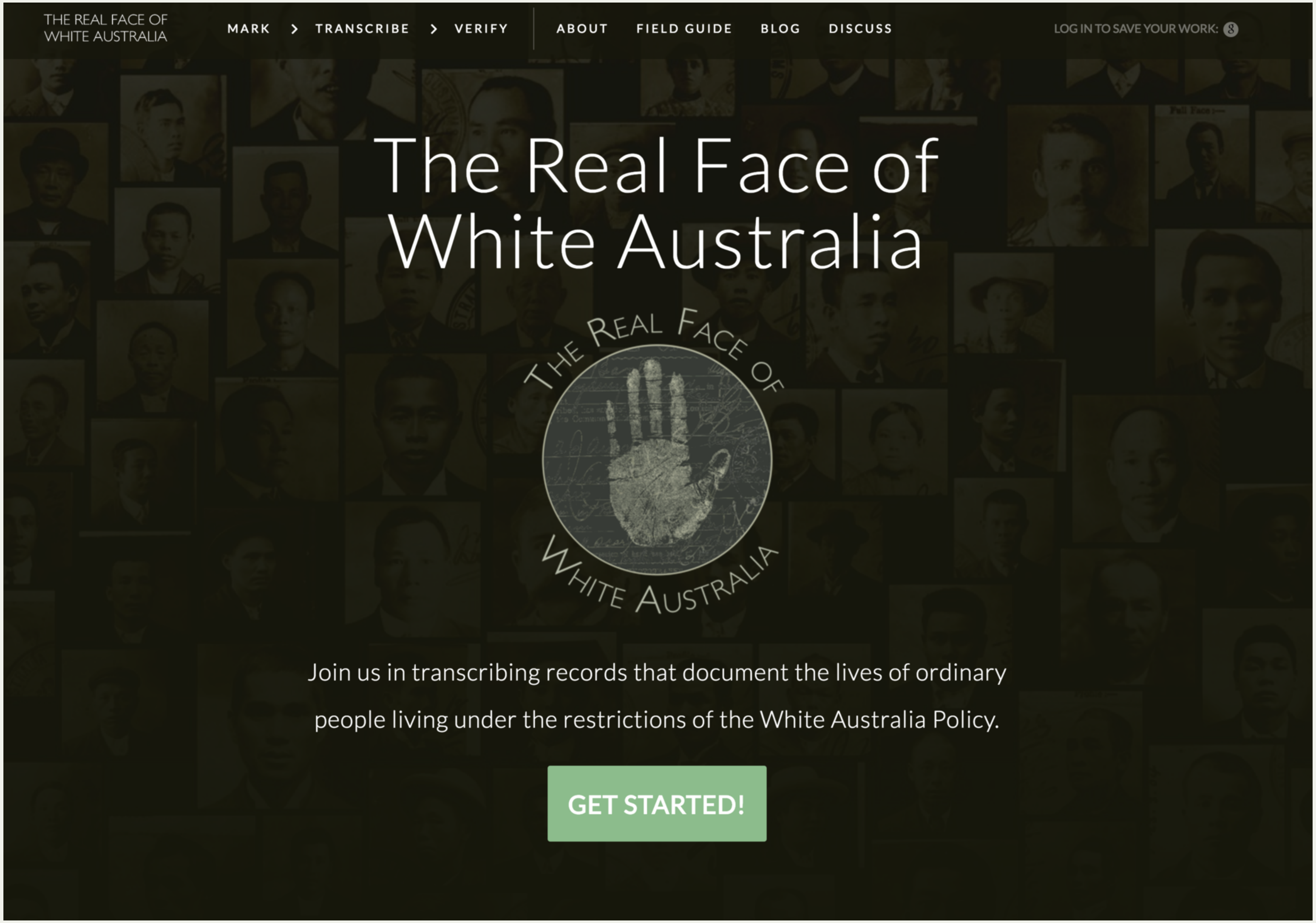
Speaker notes

Asking questions of Te Papa’s API (what’s an API??)

Not claiming great insights here — this is a notebook, exploring techniques, challenging assumptions.

But my notes also serve as an example of how others might use the API, as a resource for others to build upon.

DIY CROWDSOURCED TRANSCRIPTION



THE REAL FACE OF WHITE AUSTRALIA
MARK
TRANSCRIBE
VERIFY
ABOUT
FIELD GUIDE
BLOG
DISCUSS
LOG IN TO SAVE YOUR WORK:
THE REAL FACE OF WHITE AUSTRALIA
GET STARTED!

The Real Face of White Australia

Join us in transcribing records that document the lives of ordinary people living under the restrictions of the White Australia Policy.

http://transcribe.realfaceofwhiteaustralia.net/

Speaker notes

Other ways we can use existing tools in different ways

Real Face of White Australia

DIY Crowdsourced Transcription

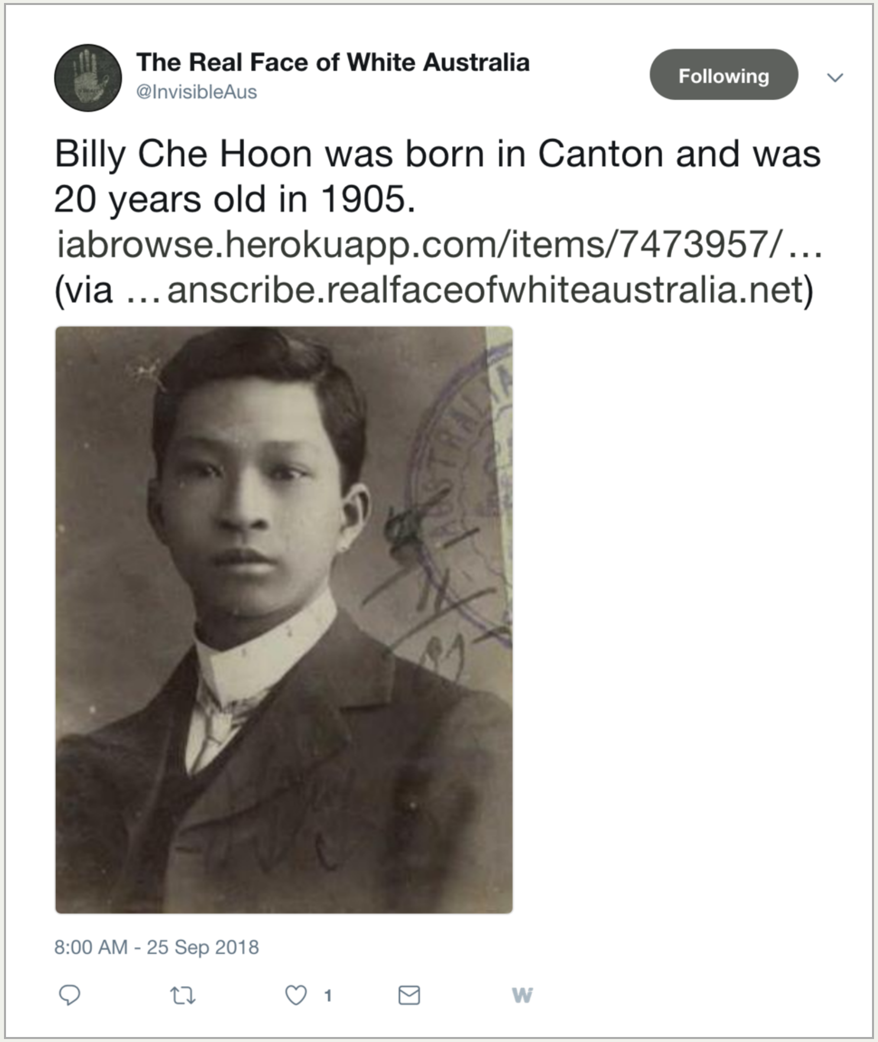
SMALL PARTS LOOSELY JOINED

- RecordSearch Harvester
- Scribe
- Amazon/Heroku
- GitHub
- Twitter

Speaker notes

* RS harvester — now as a notebook!
* Scribe — open source
* Amazon/Heroku
* Github — sharing data
* Twitter — mini stories twice daily

MINI STORIES TWICE DAILY



The Real Face of White Australia
@InvisibleAus

Billy Che Hoon was born in Canton and was 20 years old in 1905. iabrowse.herokuapp.com/items/7473957/… (via …anscribe.realfaceofwhiteaustralia.net)

8:00 AM - 25 Sep 2018

https://twitter.com/InvisibleAus

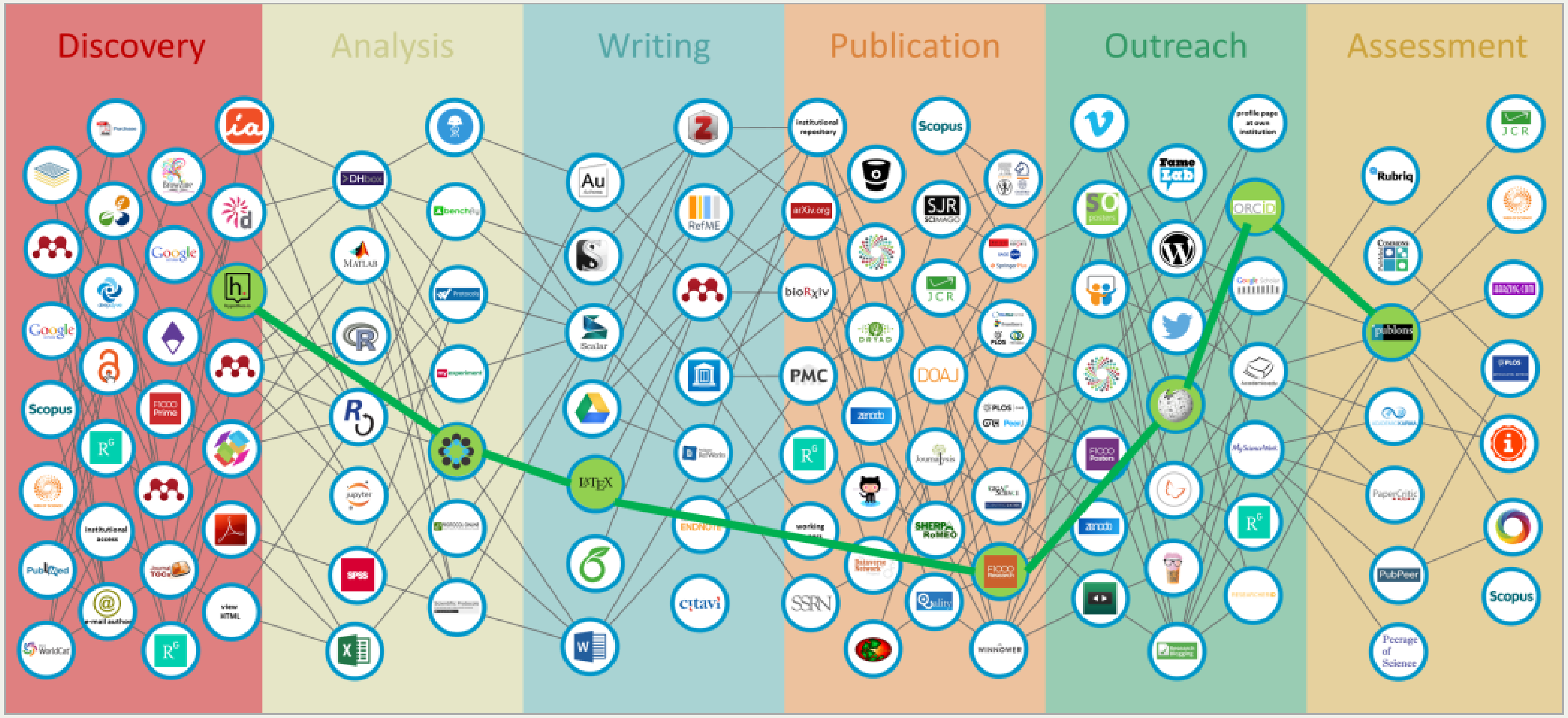
Speaker notes

Twitter stories — taking the transcribed data feeding it through a bot

Small parts loosely coupled — taking advantage of what's already there instead of thinking you have to invent something new.

Can do big things for small amounts of money. Don't have to wait for a grant (eek!).

OPEN SCIENCE TOOLS ECOSYSTEM



https://jrost.org/2018/09/13/workshop.html

Speaker notes

This is true across the web.

Interesting to look at how much research activity is made possible by discrete, but interlocking projects.

This is exactly what the Joint Roadmap for Open Science Tools is starting to examine. They’re investigating the ecosystem of open research tools to find where the gaps and synergies are.

FROM WEBSITE TO LABORATORY

https://historichansard.net/

http://historichansard.net/

Speaker notes

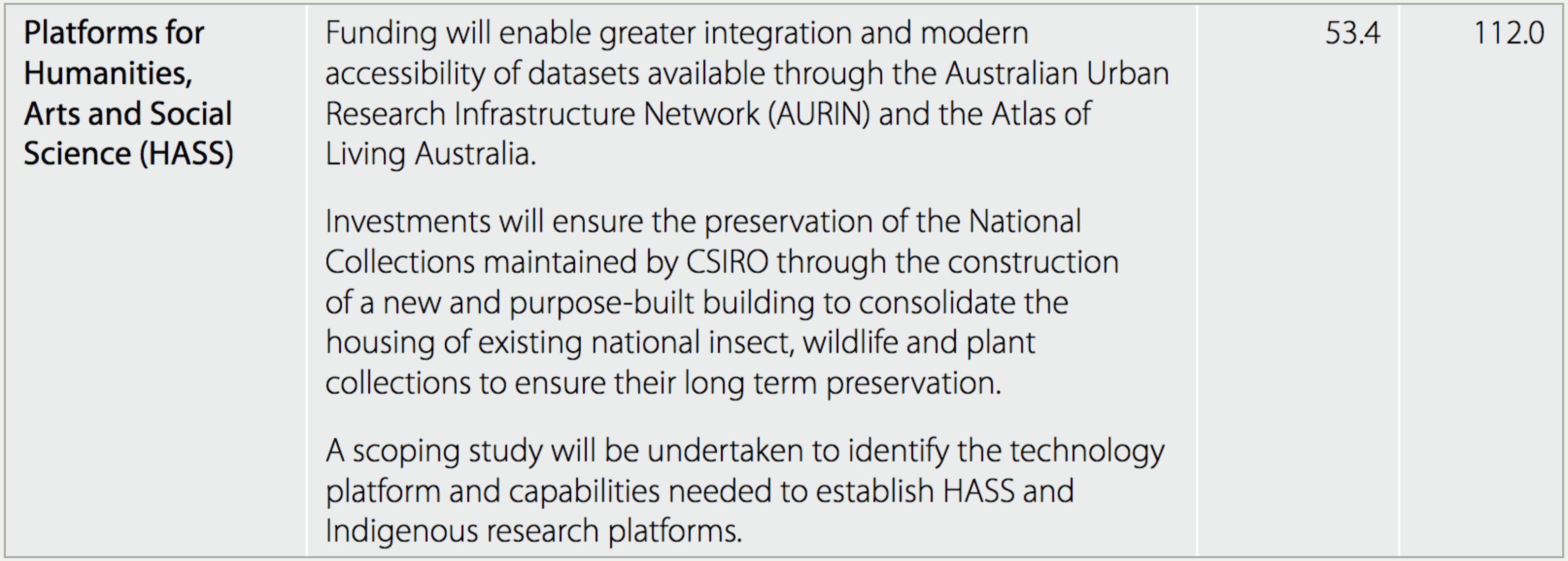
Historic Hansard

Link to XML on GitHub

Hypothes.is for annotation

FACILITIES FOR THE FUTURE!

RESEARCH INFRASTRUCTURE INVESTMENT PLAN



Platforms for Humanities, Arts and Social Science (HASS)	Funding will enable greater integration and modern accessibility of datasets available through the Australian Urban Research Infrastructure Network (AURIN) and the Atlas of Living Australia. Investments will ensure the preservation of the National Collections maintained by CSIRO through the construction of a new and purpose-built building to consolidate the housing of existing national insect, wildlife and plant collections to ensure their long term preservation. A scoping study will be undertaken to identify the technology platform and capabilities needed to establish HASS and Indigenous research platforms.	53.4	112.0

https://docs.education.gov.au/node/50601

Speaker notes

How do fund/develop these sorts of ecosystems of tools and data?

How do we support their creation, maintenance, and use?

In Australia — NCRIS and LIEF (maybe Linkage)

A MODEST PROPOSAL

Speaker notes

A modest proposal

No more big machines, digitised collections, or shiny interfaces without the tools and encouragement to critique, discuss, explore, and extend them.

Large infrastructure grants matched by small, rapid turnaround, low documentation grants that encourage

NEH OFFICE FOR DIGITAL HUMANITIES

> “ *Digital Humanities Advancement Grants (DHAG) support digital projects throughout their lifecycles, from early start-up phases through implementation and long-term sustainability.* ***Experimentation, reuse, and extensibility*** *are hallmarks of this program, leading to innovative work that can scale to enhance scholarly research, teaching, and public programming in the humanities.*

https://www.neh.gov/grants/odh/digital-humanities-advancement-grants

Speaker notes

NEH ODH example: ‘experimentation, reuse, and extensibility’

Build on past projects and make it easy for future projects to build on your work.

Also ILMS — where’s the funding for experimental infrastructure development in GLAMs?

AHRC-SMITHSONIAN FELLOWSHIPS IN DIGITAL SCHOLARSHIP

> “ *build capacity through the innovative application of digital methods and technologies to research in museums and cultural/heritage institutions*

https://ahrc.ukri.org/funding/apply-for-funding/current-opportunities/fellowships-in-digital-scholarship/

Speaker notes

AHRC digital scholarship grant

Aimed at encouraging innovative use of collections

DIFFERENT TYPES OF LEARNING?

- Survival skills
- What's possible?
- How do I...?

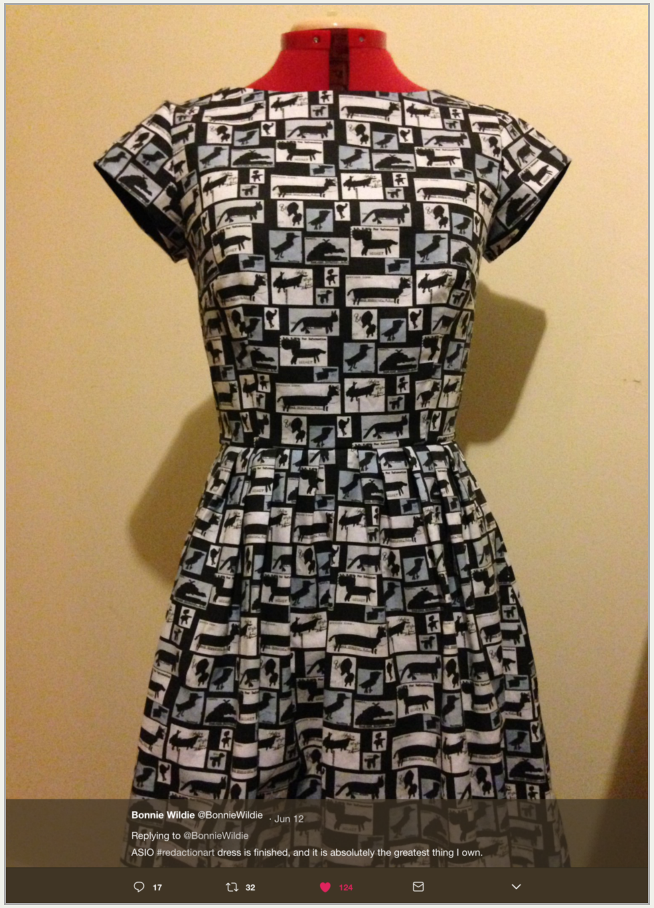
Speaker notes

Building infrastructure, also means supporting users of that infrastructure.

And recognising that the boundaries between builders/maintainers/users are permeable.

We need to support researchers to engage in the community, to connect to the ecosystem, not simply to learn how to use a particular piece of software.

THE JOYS OF SHARING



Bonnie Wildie @BonnieWildie · Jun 12

Replying to @BonnieWildie

ASIO #redactionart dress is finished, and it is absolutely the greatest thing I own.

https://bonniewildie.wordpress.com/2018/03/25/wearing-access-part-1-the-text/

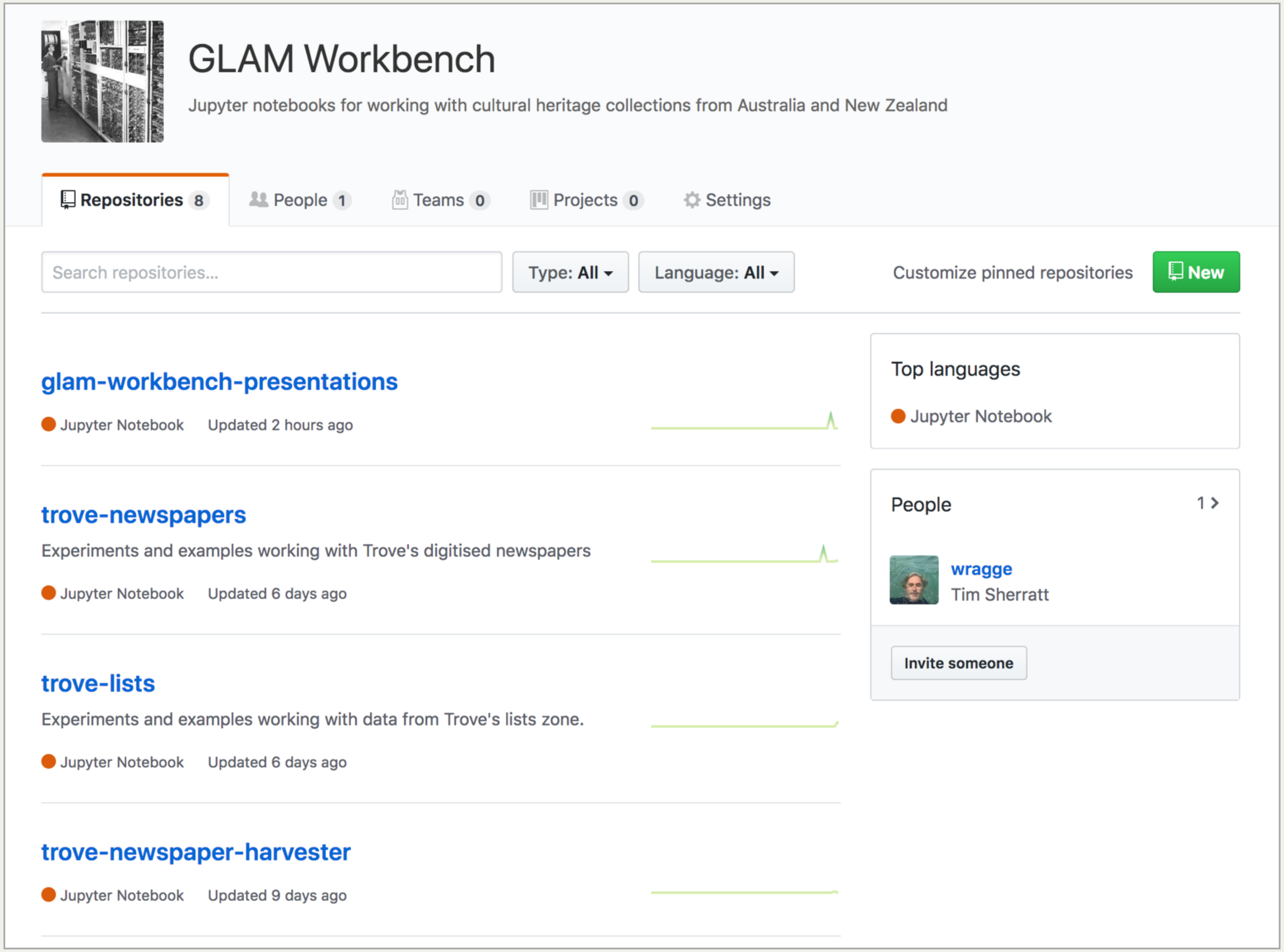
Speaker notes

We all have a role to play in the ‘How do I?’ space.

Not just about tutorials and workshops, it’s about sharing what we do, and how we do it.

As with Te Papa example, you rough, exploratory notes could be the jumping off point for someone else.

HOW DO I...?



GLAM Workbench

Jupyter notebooks for working with cultural heritage collections from Australia and New Zealand

glam-workbench-presentations

● Jupyter Notebook Updated 2 hours ago

trove-newspapers

Experiments and examples working with Trove's digitised newspapers

● Jupyter Notebook Updated 6 days ago

trove-lists

Experiments and examples working with data from Trove's lists zone.

● Jupyter Notebook Updated 6 days ago

trove-newspaper-harvester

● Jupyter Notebook Updated 9 days ago

Top languages

● Jupyter Notebook

People 1 >

wragge
Tim Sherratt

https://github.com/GLAM-Workbench

Speaker notes

GLAM Workbench — my new collection of how tos

Different ways to use and learn — multiple pathways

Examples of shared datasets — basis for collaboration & exploration

WHY BOTHER?

Speaker notes

Why would you bother sharing? Being open takes work and risk.

Where are the rewards? How does it help your career?

Better off with failed funding application?

GUIDELINES FOR THE PROFESSIONAL EVALUATION OF DIGITAL SCHOLARSHIP BY HISTORIANS

> “ *Historians who take a strong interest in digital media and information technology, or who choose to work exclusively in digital environments, should be evaluated in terms of their overall ability to use sustained, expressive, substantive, and institutional innovation to advance scholarship.*

https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians

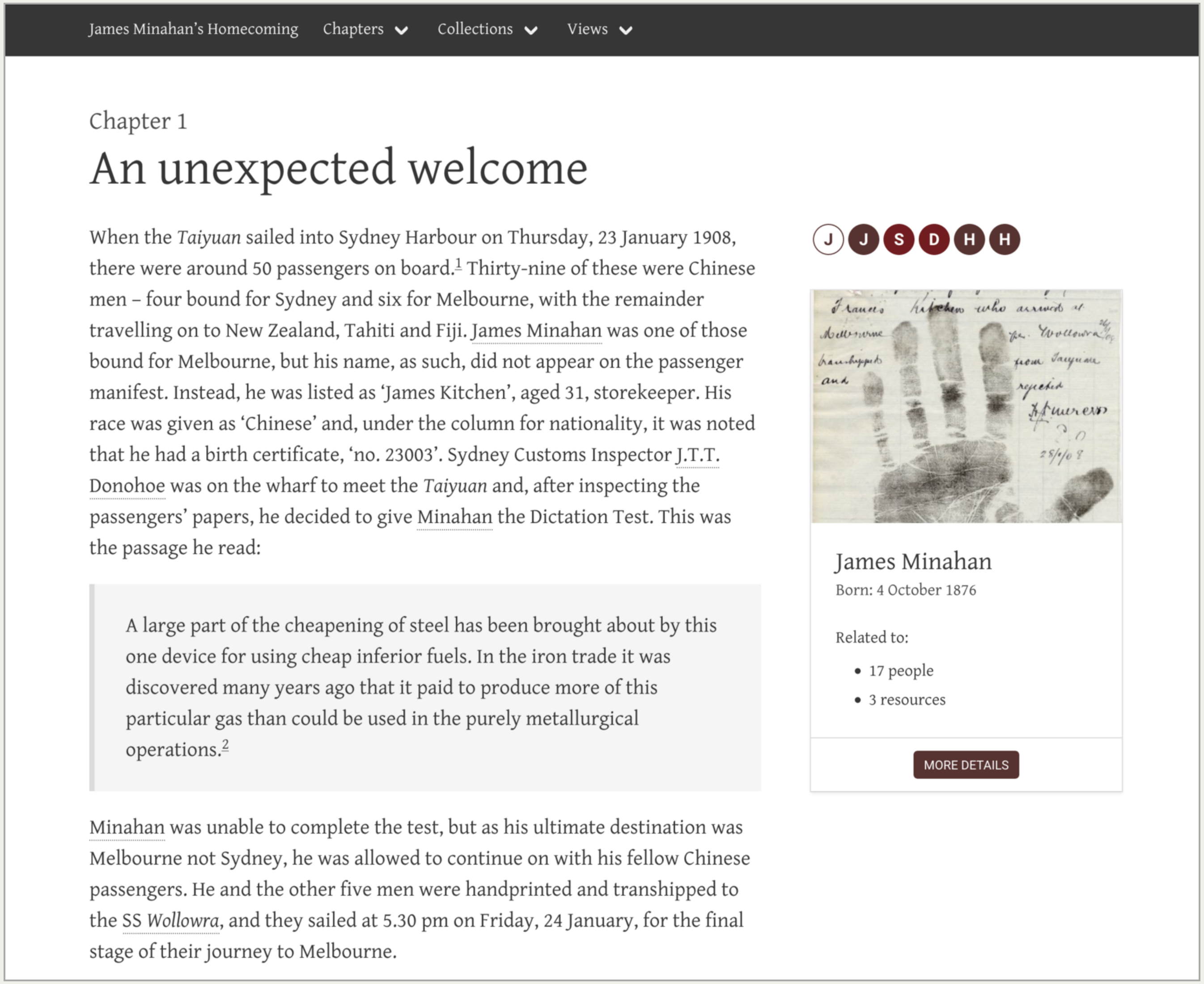
Speaker notes

AHA Guidelines

Work in the digital realm evaluated on its own terms, and not by trying to translate its methods and products back into conventional research outputs.

Promotion / Hiring / Professional development

LODBOOK


James Minahan's Homecoming
Chapters
Collections
Views
J
J
S
D
H
H
Francis Kitchen who arrived at Melbourne per Wollowra 26/1/08 transhipped and from Taiyuan rejected
28/1/08
MORE DETAILS

Chapter 1

An unexpected welcome

When the *Taiyuan* sailed into Sydney Harbour on Thursday, 23 January 1908, there were around 50 passengers on board.[1] Thirty-nine of these were Chinese men – four bound for Sydney and six for Melbourne, with the remainder travelling on to New Zealand, Tahiti and Fiji. James Minahan was one of those bound for Melbourne, but his name, as such, did not appear on the passenger manifest. Instead, he was listed as 'James Kitchen', aged 31, storekeeper. His race was given as 'Chinese' and, under the column for nationality, it was noted that he had a birth certificate, 'no. 23003'. Sydney Customs Inspector J.T.T. Donohoe was on the wharf to meet the *Taiyuan* and, after inspecting the passengers' papers, he decided to give Minahan the Dictation Test. This was the passage he read:

> A large part of the cheapening of steel has been brought about by this one device for using cheap inferior fuels. In the iron trade it was discovered many years ago that it paid to produce more of this particular gas than could be used in the purely metallurgical operations.[2]

Minahan was unable to complete the test, but as his ultimate destination was Melbourne not Sydney, he was allowed to continue on with his fellow Chinese passengers. He and the other five men were handprinted and transhipped to the SS *Wollowra*, and they sailed at 5.30 pm on Friday, 24 January, for the final stage of their journey to Melbourne.

James Minahan

Born: 4 October 1876

Related to:

- 17 people
- 3 resources

https://wragge.github.io/lodbook-james-minahan/

Speaker notes

Change the nature of publication itself

* open access
* playing with the form
* narrative and data — incentives to share

- What can I do?
- Where do I start?
- Who can I help?

Speaker notes

If we think beyond infrastructure as big machines, the questions facing us change from ‘What do we need?’, ‘What will be build?’, ‘How much money is available?’ to

What can I do?
Where do I start?
Who can I help?

https://timsherratt.org