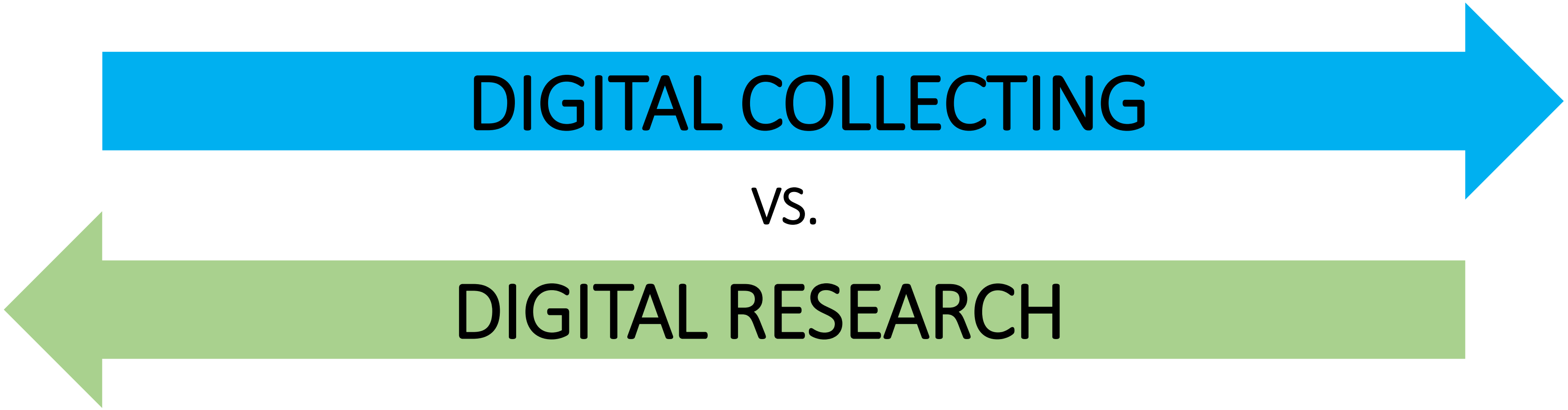


DIGITAL COLLECTING

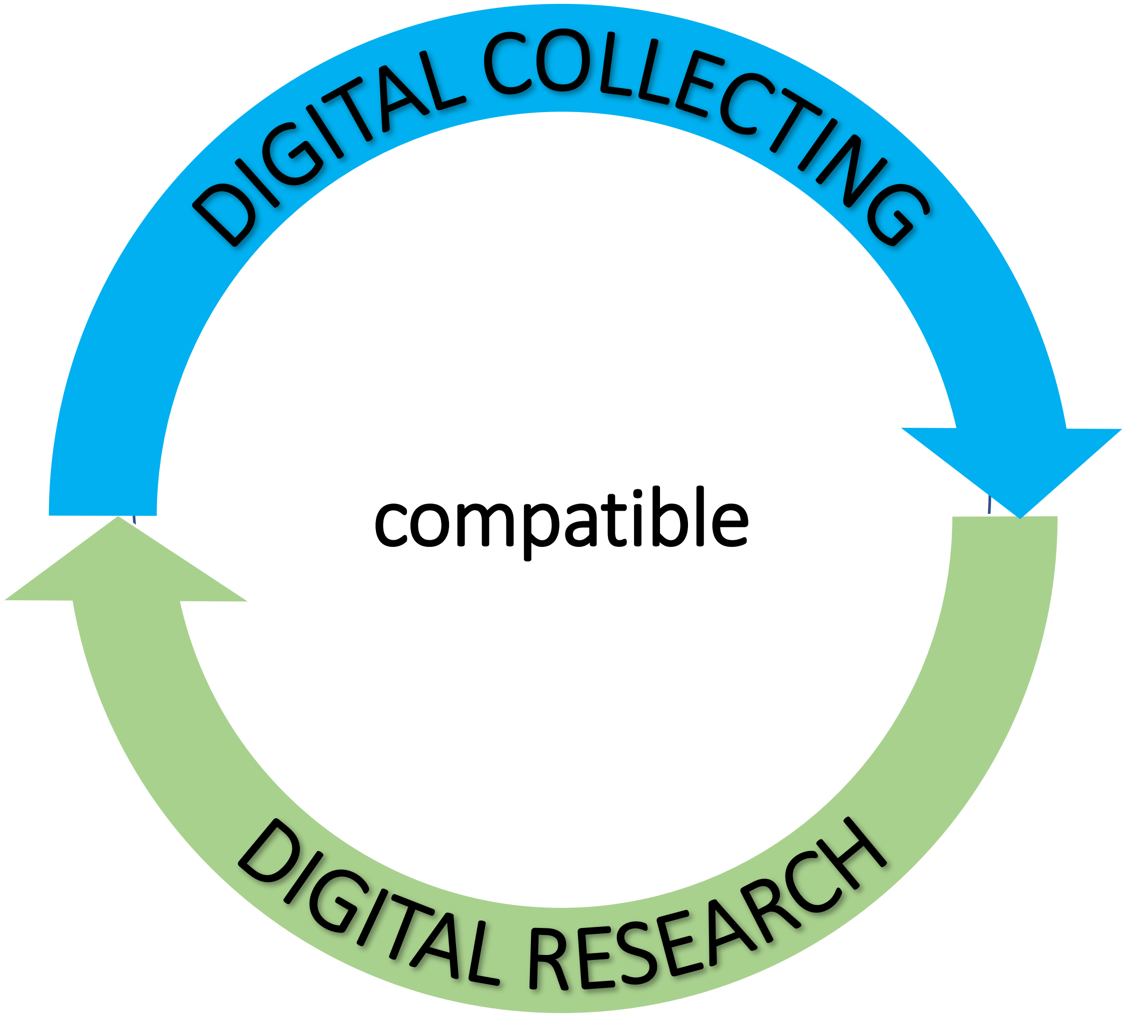
vs.

DIGITAL RESEARCH

Are they compatible?

Andrea Goethals
eResearch 2021 Wellington

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND



DIGITAL COLLECTING

Reusable research-ready data

compatible

DIGITAL RESEARCH

Inform collecting, research outputs

Part 1 - Overview from the perspective of a digital collector

- The Library’s digital collections
- Current challenges
- Activities to provide them to researchers for data analysis

Part 2 - Discussion

- What can digital collecting institutions like the Library do so that these collections are more useful to researchers?
- How can we increase collaboration between collectors and researchers?

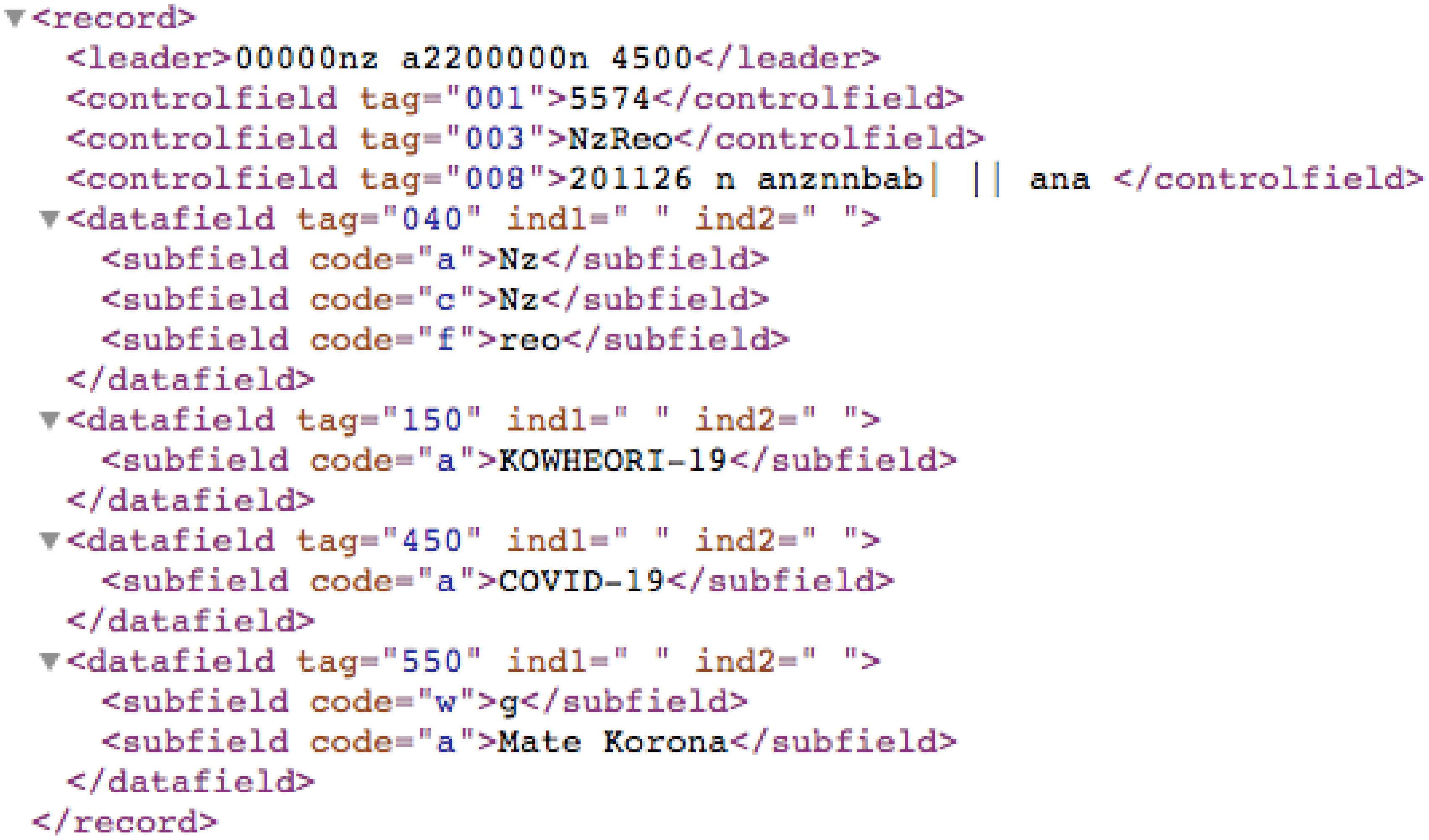
Purpose of the National Library is ... collecting, preserving and protecting documents, particularly those relating to New Zealand, and making them accessible for all the people of New Zealand, in a manner consistent with their status as documentary heritage and taonga …

- National Library of New Zealand Act
(Te Puna Maatauranga o Aotearoa) 2003

Open metadata datasets

- Te Puna Web Directory metadata (CC BY 3.0 NZ)
- Publications New Zealand metadata (CC BY 3.0 NZ)
- Index New Zealand metadata (CC BY 3.0 NZ)
- DigitalNZ API metadata aggregator
- Turnbull unpublished collections metadata (CC BY 4.0)
- Turnbull names (CC BY 4.0)
- Ngā Upoko Tukutuko metadata (CC BY-NC-ND 3.0 NZ)

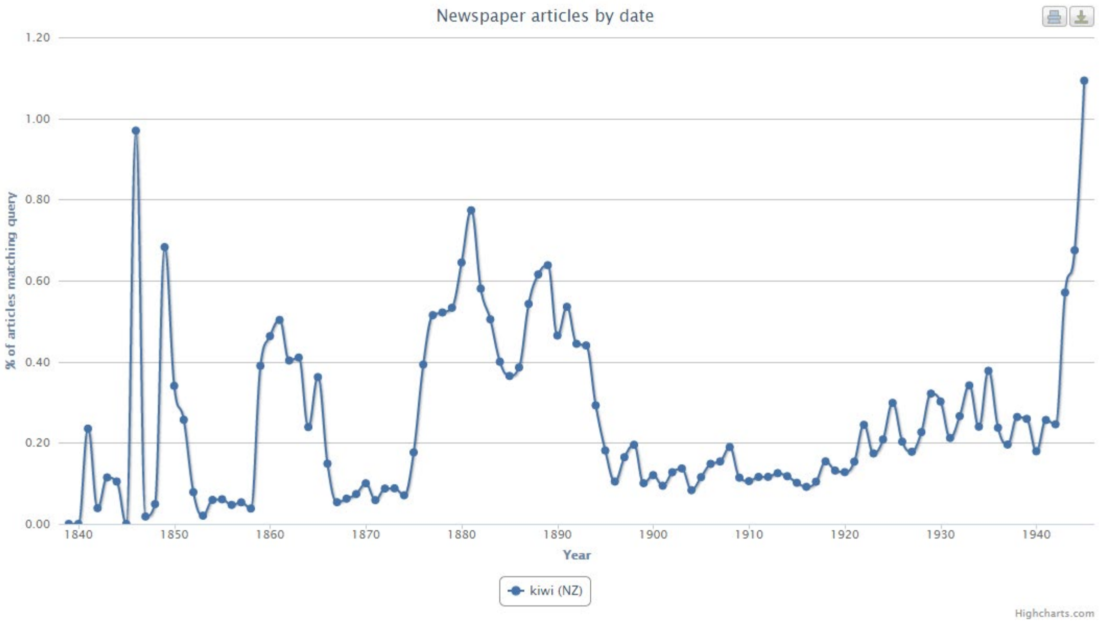
Download from https://natlib.govt.nz/about-us/open-data



```
▼<record>
   <leader>00000nz a2200000n 4500</leader>
   <controlfield tag="001">5574</controlfield>
   <controlfield tag="003">NzReo</controlfield>
   <controlfield tag="008">201126 n anznnbab| || ana </controlfield>
 ▼<datafield tag="040" ind1=" " ind2=" ">
     <subfield code="a">Nz</subfield>
     <subfield code="c">Nz</subfield>
     <subfield code="f">reo</subfield>
   </datafield>
 ▼<datafield tag="150" ind1=" " ind2=" ">
     <subfield code="a">KOWHEORI-19</subfield>
   </datafield>
 ▼<datafield tag="450" ind1=" " ind2=" ">
     <subfield code="a">COVID-19</subfield>
   </datafield>
 ▼<datafield tag="550" ind1=" " ind2=" ">
     <subfield code="w">g</subfield>
     <subfield code="a">Mate Korona</subfield>
   </datafield>
 </record>
```

Dataset — Papers Past newspaper open data pilot

- Downloadable datasets of digitised newspapers (1839-1899)
  - Individual titles or script batch downloads
  - METS/ALTO XML files (includes full text)
- Starter Kit
  - All the issues for 7 small newspapers (335 MB)
- International examples of how others have used newspaper data
- Feedback? paperspast@natlib.govt.nz



“kiwi” occurrence using Tim Sherratt’s QueryPic tool. It uses the DigitalNZ API to search Papers Past data for the frequency of terms and graphs the results (http://dhistory.org/querypic/)

High-res images

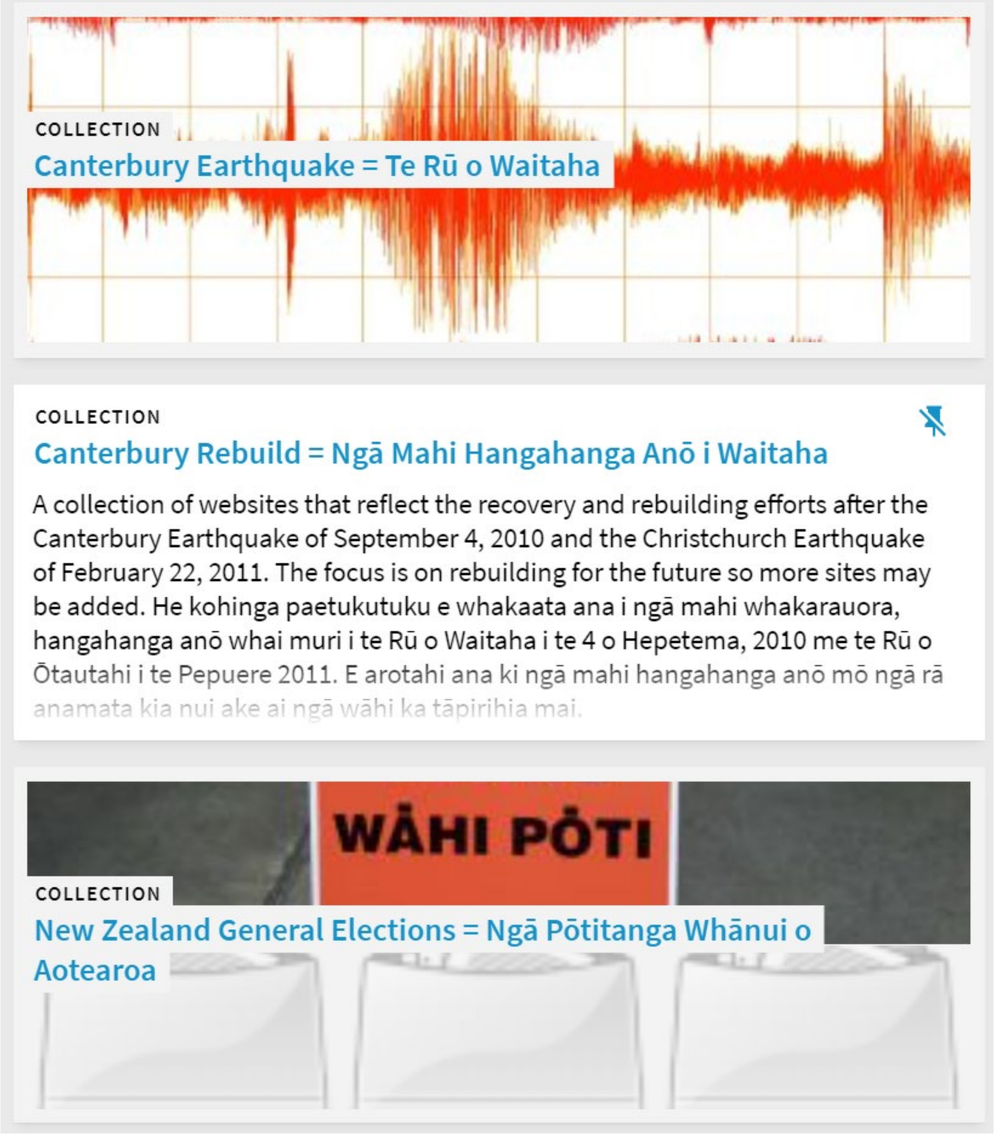
Free, CC By 4.0
54,600 online items

https://natlib.govt.nz/photos?il%5Batl_free_download%5D=true

Image: Wellington, Aerial photograph taken by Whites Aviation, 21 Jan 1959, Whites Aviation Collection, Alexander Turnbull Library

Selective web harvests (since 1999)

- 42,025 websites, 399 million files
- 3 web curators & a web archive engineer, in-house developers to maintain specialised tools
- Accessible through Library catalogue, working on a full-text search solution
- Harvest collections ->



COLLECTION

Canterbury Earthquake = Te Rū o Waitaha

COLLECTION

Canterbury Rebuild = Ngā Mahi Hangahanga Anō i Waitaha

A collection of websites that reflect the recovery and rebuilding efforts after the Canterbury Earthquake of September 4, 2010 and the Christchurch Earthquake of February 22, 2011. The focus is on rebuilding for the future so more sites may be added. He kohinga paetukutuku e whakaata ana i ngā mahi whakarauora, hangahanga anō whai muri i te Rū o Waitaha i te 4 o Hepetema, 2010 me te Rū o Ōtautahi i te Pepuere 2011. E arotahi ana ki ngā mahi hangahanga anō mō ngā rā anamata kia nui ake ai ngā wāhi ka tāpirihia mai.

COLLECTION

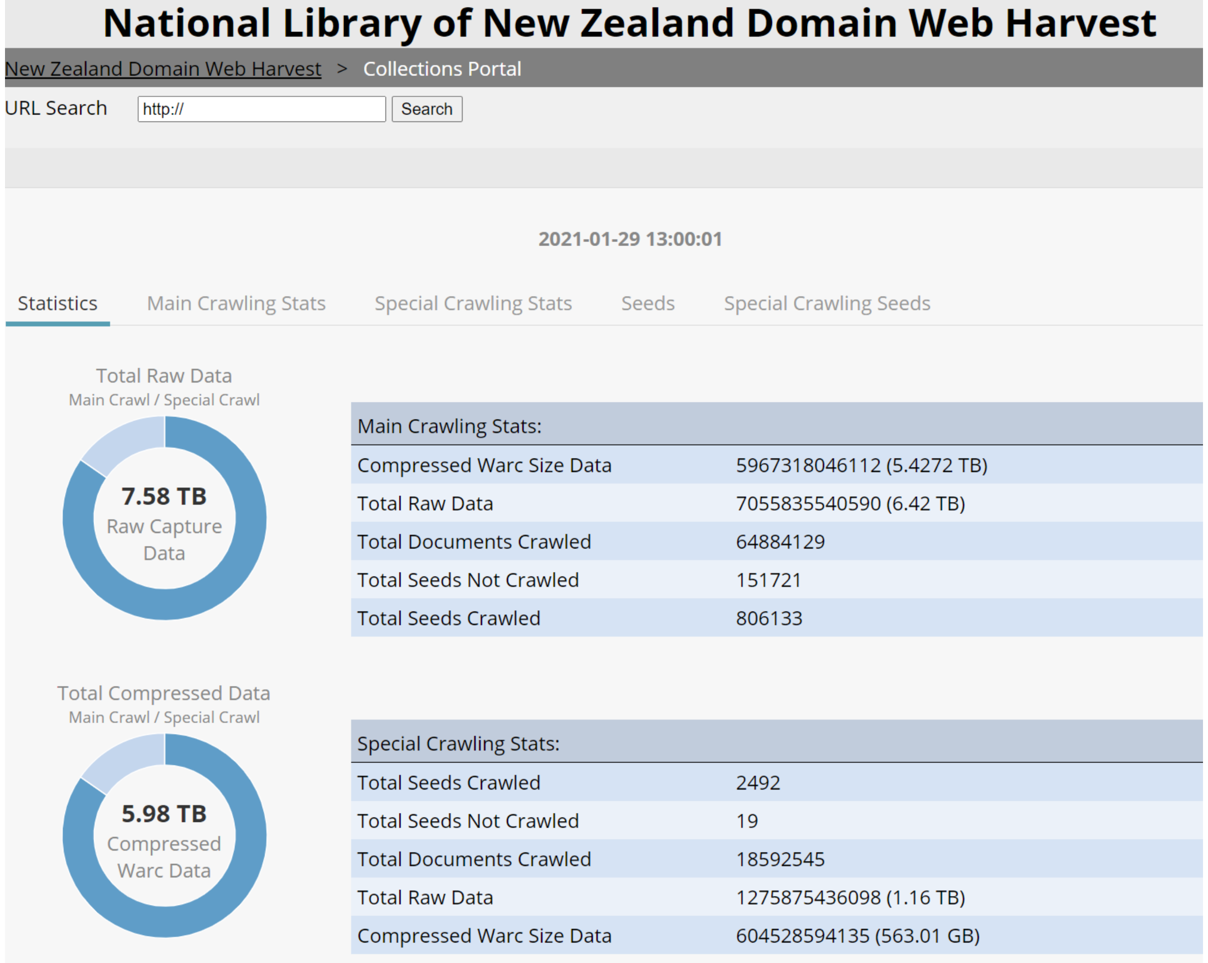
New Zealand General Elections = Ngā Pōtitanga Whānui o Aotearoa

Jupyter notebooks

To help protect your privacy, PowerPoint has blocked automatic download of this picture.

Whole of domain web harvests

- Collecting the .nz web space (2008, 2010, 2013, 2015, 2016, 2017, 2018, 2019, 2020, 2021)
- Over 1.6 billion files, 137 TB of data (compressed)



National Library of New Zealand Domain Web Harvest

New Zealand Domain Web Harvest > Collections Portal

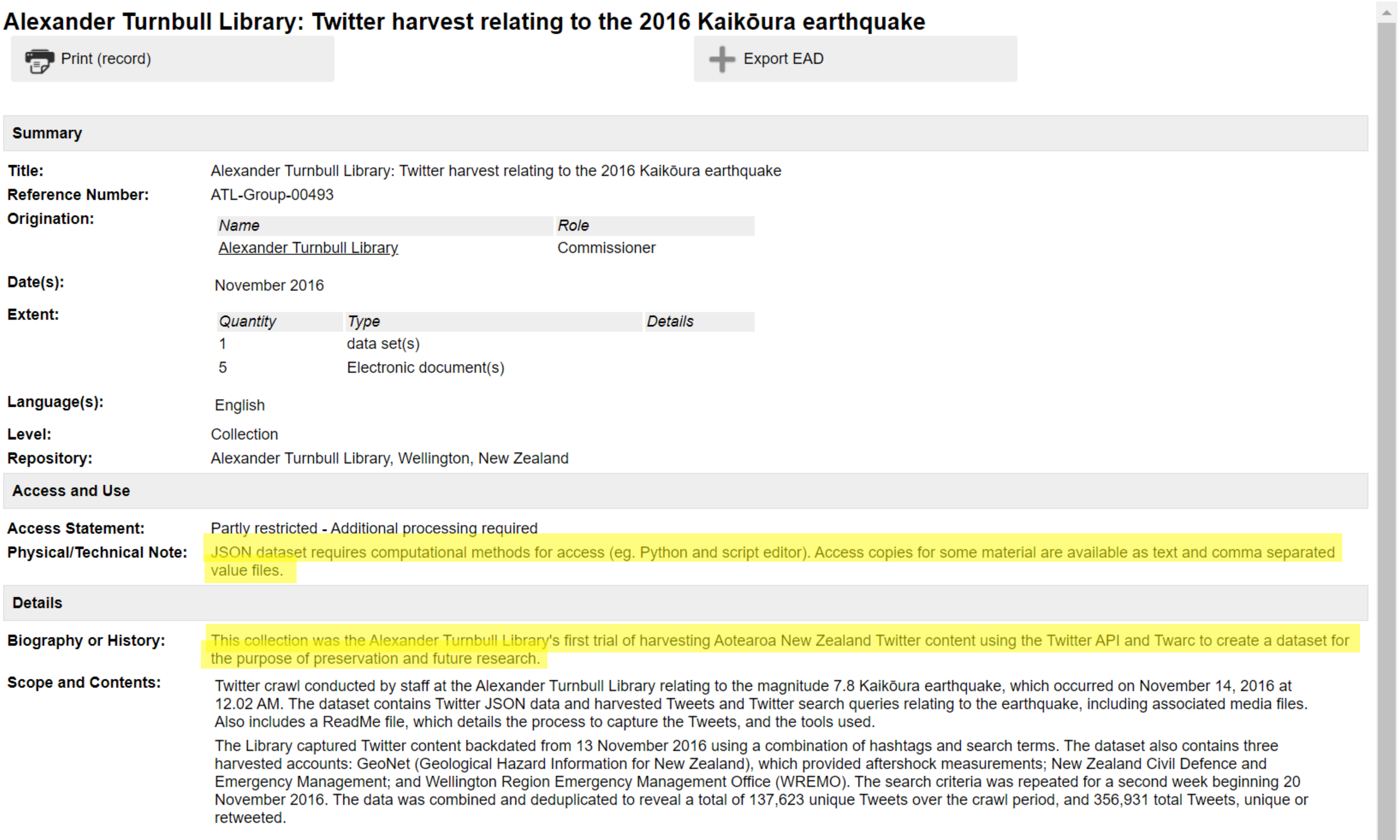
2021-01-29 13:00:01

Total Raw Data
Main Crawl / Special Crawl
7.58 TB Raw Capture Data

Main Crawling Stats:	
Compressed Warc Size Data	5967318046112 (5.4272 TB)
Total Raw Data	7055835540590 (6.42 TB)
Total Documents Crawled	64884129
Total Seeds Not Crawled	151721
Total Seeds Crawled	806133

Total Compressed Data
Main Crawl / Special Crawl
5.98 TB Compressed Warc Data

Special Crawling Stats:	
Total Seeds Crawled	2492
Total Seeds Not Crawled	19
Total Documents Crawled	18592545
Total Raw Data	1275875436098 (1.16 TB)
Compressed Warc Size Data	604528594135 (563.01 GB)



Alexander Turnbull Library: Twitter harvest relating to the 2016 Kaikōura earthquake

Summary

Title: Alexander Turnbull Library: Twitter harvest relating to the 2016 Kaikōura earthquake

Reference Number: ATL-Group-00493

Origination:

Name	*Role*
Alexander Turnbull Library	Commissioner

Date(s): November 2016

Extent:

Quantity	*Type*	*Details*
1	data set(s)	
5	Electronic document(s)	

Language(s): English

Level: Collection

Repository: Alexander Turnbull Library, Wellington, New Zealand

Access and Use

Access Statement: Partly restricted - Additional processing required

Physical/Technical Note: JSON dataset requires computational methods for access (eg. Python and script editor). Access copies for some material are available as text and comma separated value files.

Details

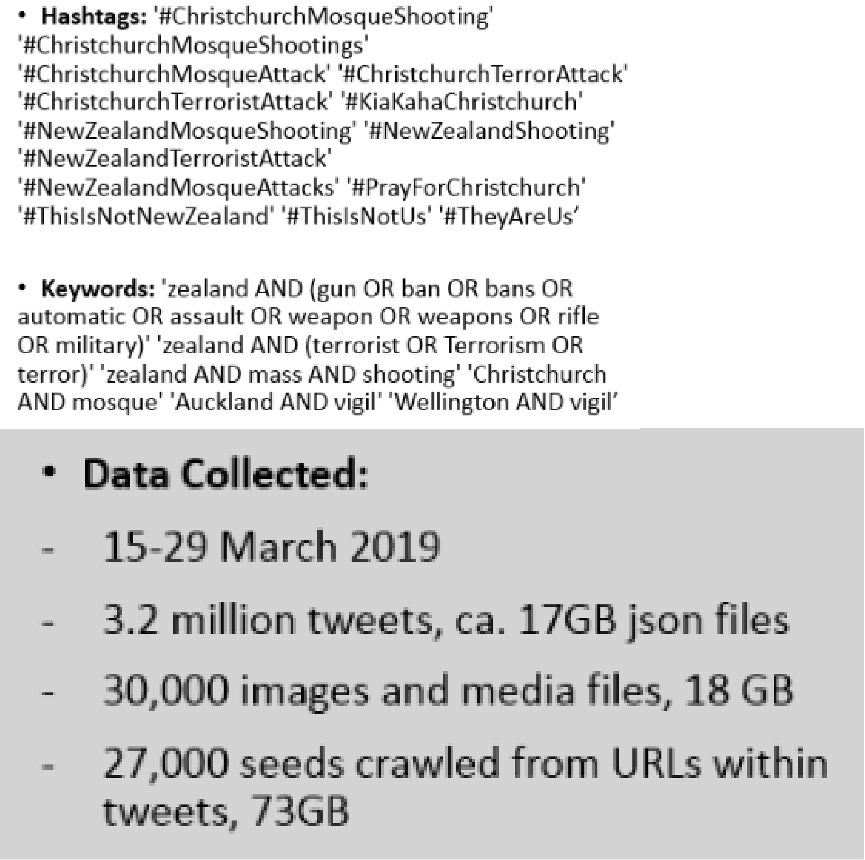
Biography or History: This collection was the Alexander Turnbull Library's first trial of harvesting Aotearoa New Zealand Twitter content using the Twitter API and Twarc to create a dataset for the purpose of preservation and future research.

Scope and Contents: Twitter crawl conducted by staff at the Alexander Turnbull Library relating to the magnitude 7.8 Kaikōura earthquake, which occurred on November 14, 2016 at 12.02 AM. The dataset contains Twitter JSON data and harvested Tweets and Twitter search queries relating to the earthquake, including associated media files. Also includes a ReadMe file, which details the process to capture the Tweets, and the tools used.

The Library captured Twitter content backdated from 13 November 2016 using a combination of hashtags and search terms. The dataset also contains three harvested accounts: GeoNet (Geological Hazard Information for New Zealand), which provided aftershock measurements; New Zealand Civil Defence and Emergency Management; and Wellington Region Emergency Management Office (WREMO). The search criteria was repeated for a second week beginning 20 November 2016. The data was combined and deduplicated to reveal a total of 137,623 unique Tweets over the crawl period, and 356,931 total Tweets, unique or retweeted.

Twitter datasets

- Related to significant events in recent history
  - Kaikoura Earthquake (2016)
  - General Election (2017, 2020)
  - Christchurch mosque shootings (2019)
  - Covid-19 (2020)



- **Hashtags:** '#ChristchurchMosqueShooting' '#ChristchurchMosqueShootings' '#ChristchurchMosqueAttack' '#ChristchurchTerrorAttack' '#ChristchurchTerroristAttack' '#KiaKahaChristchurch' '#NewZealandMosqueShooting' '#NewZealandShooting' '#NewZealandTerroristAttack' '#NewZealandMosqueAttacks' '#PrayForChristchurch' '#ThisIsNotNewZealand' '#ThisIsNotUs' '#TheyAreUs'

- **Keywords:** 'zealand AND (gun OR ban OR bans OR automatic OR assault OR weapon OR weapons OR rifle OR military)' 'zealand AND (terrorist OR Terrorism OR terror)' 'zealand AND mass AND shooting' 'Christchurch AND mosque' 'Auckland AND vigil' 'Wellington AND vigil'

- **Data Collected:**
  - 15-29 March 2019
  - 3.2 million tweets, ca. 17GB json files
  - 30,000 images and media files, 18 GB
  - 27,000 seeds crawled from URLs within tweets, 73GB

Example of the Christchurch mosque shootings data: collecting search terms used and key statistics

Challenges

- Pivoting to support computational access
- Legal challenges
- Social & ethical questions (of collecting & not collecting)
- Collecting from foreign sites & platforms (Facebook, etc.)
- Poor quality OCR – Papers Past
- Methodological questions – what do researchers want / expect?

What we’ve done so far

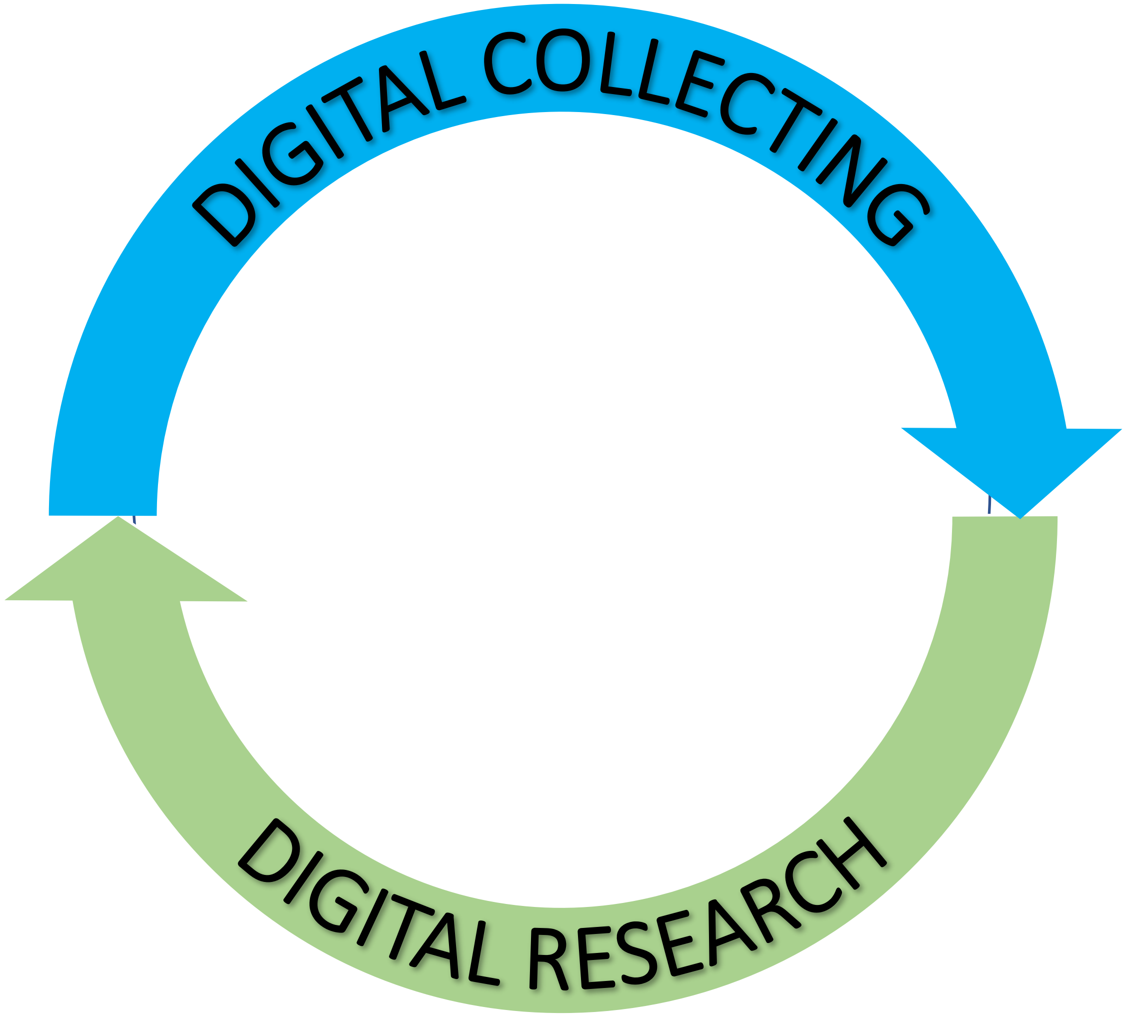
For Researchers

- Webpage with open data
- Direct provision on request
- Workshops, hackfests, competitions, exhibits using our data
- Interviews, surveys
- Creating derivative datasets with less perceived risk
- Exploring potential for pilot projects

Within the Library

- Digital Research Working Group
- Digital Research Coordinator (term)
- Staff professional development
- Developing new policies, e.g. takedown
- Exploring risk-based approaches

Discussion



DIGITAL COLLECTING

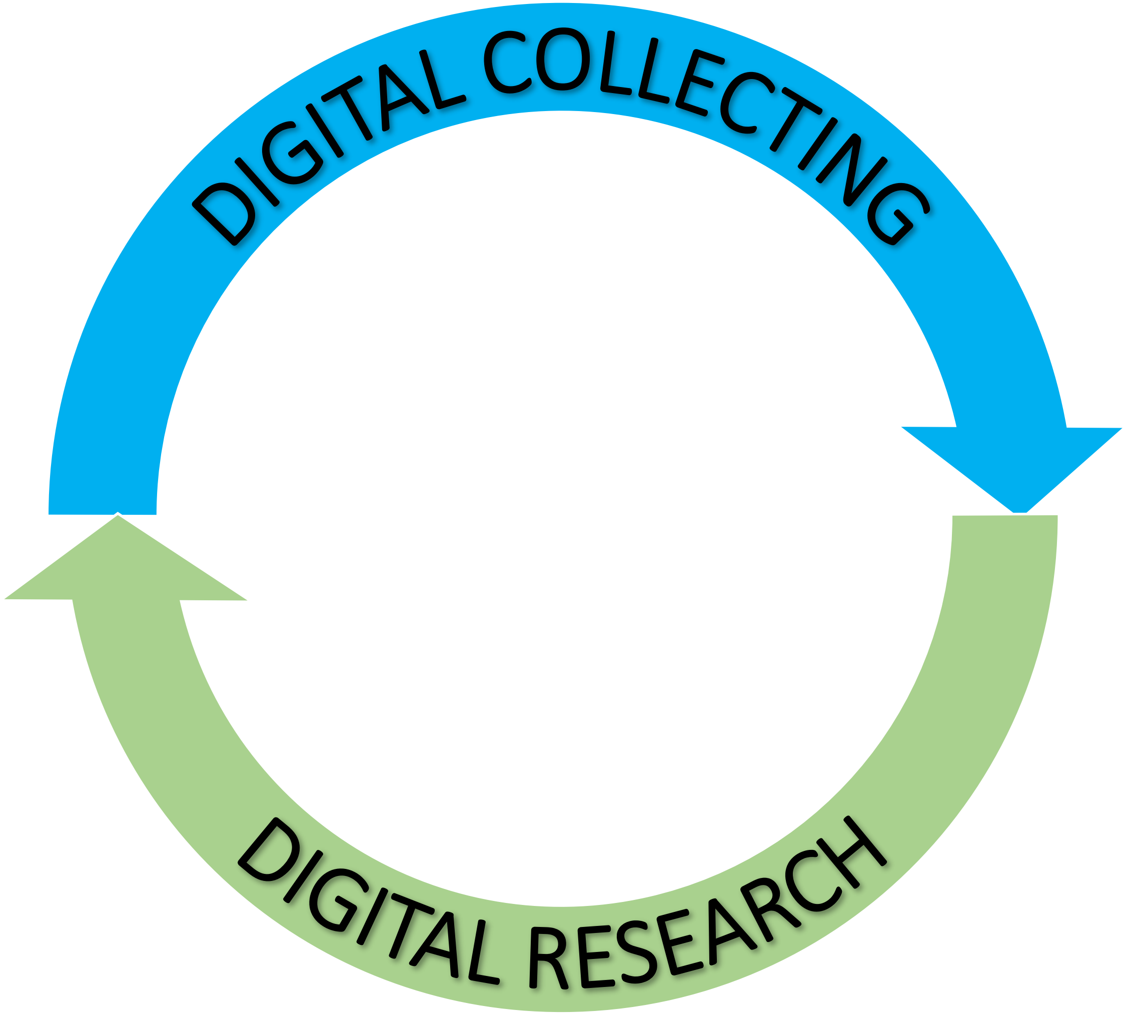
Reusable
research-ready
data

DIGITAL RESEARCH

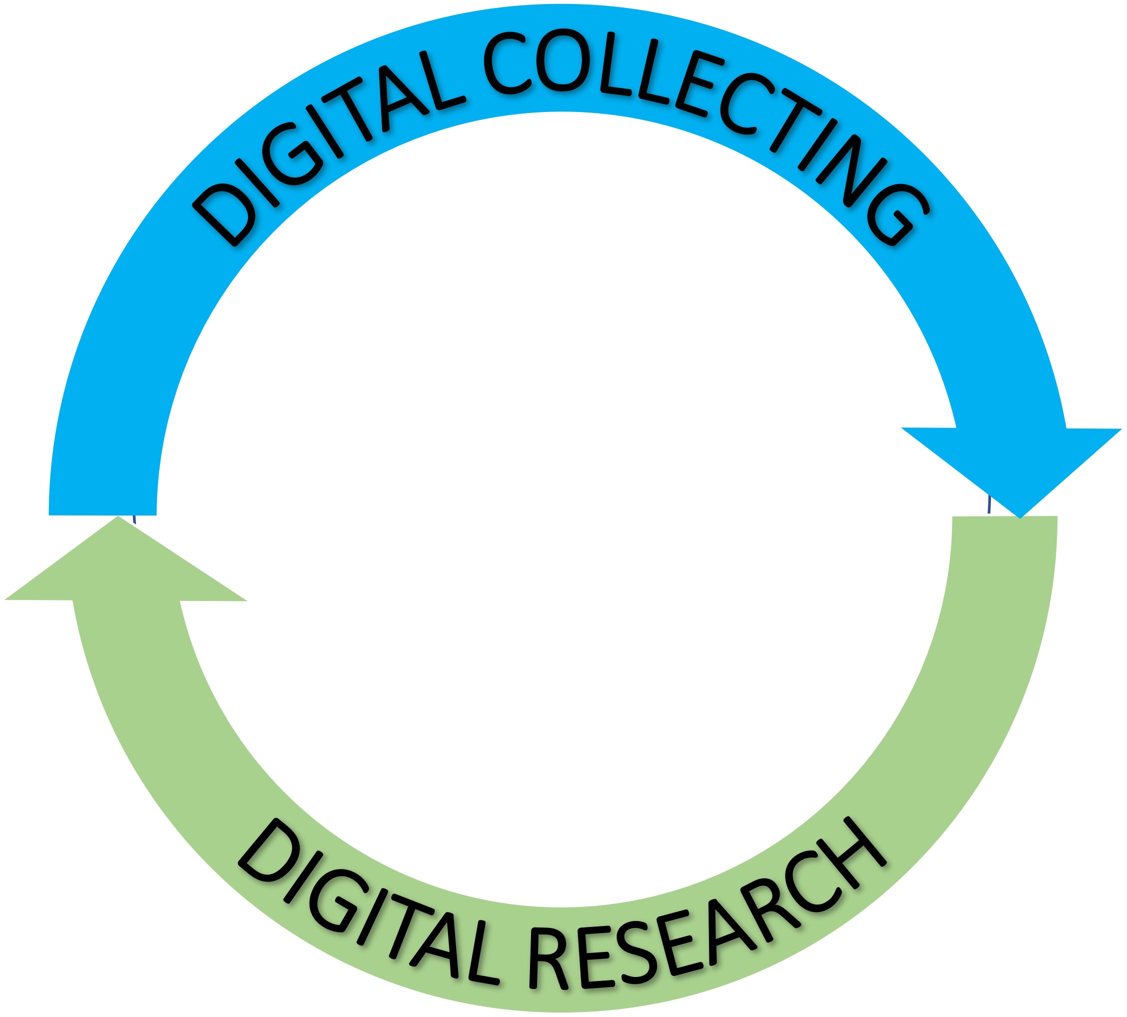
Inform collecting,
research outputs

What can collecting institutions do to make their data useable for research? What’s important to you as a researcher?

- How its provided?
- Documented?
- Formatted?
- Licensed?



Reusable research-ready data



Inform collecting,
research outputs

How do researchers want to collaborate with collecting institutions?

- Inform collecting / processing decisions?
- Submit your research outputs for preservation / reuse?

Thank you

Andrea.Goethals@dia.govt.nz

Acknowledgments:

Gillian Lee
Steve Knight
Digital Research Working Group