National Library's Web and Twitter Collections for Research

eResearch NZ 2020 - Dunedin

By: Lana Alsabbagh, Digital Research Coordinator, National Library of New Zealand

February 13, 2019

Background

The National Library of New Zealand has performed a “whole-of-domain” harvest since 2008, acquiring publicly available web content from the New Zealand .nz, .net, .org and .com domains.

The Library also collects tweets and related media about specific events in NZ recent history (Kaikoura earthquakes, PM’s pregnancy and baby announcements, 2017 General election, etc)

NO ENTRY

Most of the collections aren’t readily accessible

How do we make these collections available to researchers for computational use?

ASK THE RESEARCHERS!

Two methods:

1. Phone interviews
2. Online survey by email

Criteria:

Interested and involved in Digital Humanities or ‘big data’

Discipline-neutral

Based in NZ

Things to consider ….

- Is this data relevant for researchers?
- What questions can the data help researchers answer?
- What do researchers need?
  - Training and awareness
  - Support and advice
  - Equipment (hardware, software, network access, processing power)
  - Data (raw data? tailored datasets? WARC file format?)
  - Metadata, documentation (what is included/excluded and why?)
  - A physical or virtual workspace?

Phone interview questions

- Are you or would you be interested in using the domain harvest dataset? Is it relevant to your current or future research?
- Could you give me an example of a research question you might be able to answer using the domain harvests?
- What would make it easier for you to use it? (Training, data cleansing, context, etc)

Metadata and documentation

“ the first thing would sort of clearly indicating the kinds of data to researchers that are available”

“having some kind of online resource that tells researchers the kinds of data that are available”

“I guess the first thing would be just to know what’s in there.”

“”just knowing the whole process of data analysis, and the documentation obviously is going to be really important ..a pathway of how all of this data was collected and when it was last updated and all of that kind of documentation. I know those questions come up as part of the analysis, as part of putting things together, so having those be accessible as needed as part of the dataset would be useful.”

“being able to get as much contextual data as possible while being able to limit what we don’t need.”

“documentation and publication of exactly what’s in it, like what metadata can you get for the different sites, at what time periods were different sites measured, that sort of thing, so that you could figure out what could be the sampling frame for a study”

Infrastructure

“making those data or some subset of those data readily available for download, and/or having somebody at the university act as a go-between between researcher and the data where they can put together specific datasets or data files for the researcher and then send it to them.”

“that would take a lot of burden off researchers if the data doesn’t come in technical archive formats”

“I know from a couple of other colleagues , they would never be able to use it because they do not have the technical capability to actually deal with those archiving formats and the amount of data. So for them, access to data is better to be facilitated in a way that they can actually use it almost directly, like let’s say they can specify the parameters that they want to extract from the metadata and get it as a [text-realised??] dump somewhere or something like that.”

“do we also provision that in a way that actually not everyone is doing the same thing over and over again?”

“making access flexible and open”

Training, help, and support

“definitely training would be the number one thing that I would put on the wishlist to help facilitate using the dataset.”

“Integration with institutional libraries would probably be a main one”

*“*Some examples of the types of questions people have asked would be helpful!”

Survey

- Demographics: Employee role (faculty, librarian, etc), career stage (early, mid, late)
- Familiarity with web archives and digital research methods
- Relevance (or not) of domain and Twitter harvests, sample research questions
- Obstacles; what would make it easier to start using the datasets; what do you need from the National Library

Findings

- Most don’t use the Library’s web archive, or web archives in general
- Most are familiar with digital research methods
- Most (>60%) feel the domain and Twitter harvests are relevant to their research
- General confusion re. content, format, and potential

Findings

- Most (95%) want datasets they can download and use in their own infrastructure
- Metadata, documentation (77%) | training & support (55%)
- Top 3 obstacles were complexity of tools, intellectual property issues, and lack of mass data storage (eg. .NZ domain = 140 TB compressed)
- The Library should provide raw datasets; respondents divided on whether Library should provide hardware, software, processing capability

What do you think?

Email web.archive@dia.govt.nz

or andrea.goethals@dia.govt.nz