

## Introducing Two Datasets: Alexander Turnbull Library's new offering of descriptive and name authority records

Nicola Frean

This paper outlines the contexts under which the Alexander Turnbull Library (ATL) is releasing two new datasets into the research environment, and invites their use.

These datasets are titled 'Turnbull Unpublished', and 'Turnbull Names'. The first contains descriptive metadata for all the Library's unpublished collections, and the second includes metadata for all the names indexed for those collections. Both datasets meet appropriate international encoding standards and at the time of writing (February 2017) will be into their fourth monthly iteration. This paper builds on a National Digital Forum (NDF) presentation at the time of their first release,<sup>1</sup> discusses the reasons, challenges and expectations for these datasets, and the manner of their release.<sup>2</sup>

Enabling discovery by all is a key Library goal because discovery and use justify resources put into acquisition and preservation. The creation of these datasets continues an approach since the Library began in 1920, following the bequest by Alexander Turnbull of his 55,000+ items, namely, that collection and preservation of primary collections is only of value if they are discoverable and useable by those interested. The 'items' now number in the millions and the Library has kept pace with technological advances for their communication – from typewritten cards to lists in *Archifacts* to topical publications to the provision of searchable databases, through to online and federated equivalents and tools. This latest offering further extends the accessibility and potential of the collections. It also integrates with broader efforts to make government information more readily available and transparent, for example the Cabinet encouragement that "all Public Service departments ... commit to releasing high value public data actively for re-use, in accordance with the Declaration and Principles, and in accordance with the NZGOAL Review and Release process."<sup>3</sup> The National Library of New Zealand (NLNZ) published its Use and Reuse Policy<sup>4</sup> in 2014, which included a statement on metadata.

It is relatively straightforward to load a dataset to data.govt.nz, once the dataset has an address.<sup>5</sup> In computer terms these are relatively small zipped files, even though they contain metadata records for c240,000 Names and c80,000 Finding Aids, cover huge and/or various collections, and include content and context information added by staff through generations of donation, acquisition, and use.

However, to merely throw more metadata online would not satisfy the needs of the new types of researcher who will make use of these datasets. For example, two digital scholars commented recently: “feedback from digital researchers on collection delivery .... [includes] calls to achieve greater transparency in the nature and extent of the digital offering by clearly indicating which metadata is made publically available and where, assess the data offered, be aware of advances in digital humanities research in topic modelling and how these relate to archival bias and outline the nature and scope of various data sets so that they can be combined. In all instances metadata delivery needs to take into account both human and computational use cases.”<sup>6</sup> The Turnbull metadata has recently been re-shaped behind the scenes to form datasets complying with four key international standards, and the channels through which it will flow are sustainable and can be assessed against what is already known or can be discovered about researchers’ needs.

The four key standards with which the ATL metadata complies are: EAD or Encoded Archival Description (2002); EAC-CPF or Encoded Archival Context – Corporate Bodies, Persons, and Families (2010), the Unicode character set, and the Creative Commons licensing model. In each case the decision on a standard has also affected the Library’s workflows, and these organisational aspects will continue to play out over time. Hopefully, they will facilitate more and better arrangement and description, leading to more and better metadata, and therefore richer and more inclusive engagement with the collections. This engagement may provide further community information which will add value to the collections. The remainder of this article deals in more detail with these four standards in relation to the Turnbull’s two new datasets.

‘Turnbull Unpublished’, the name given to the dataset for descriptive metadata on the collections, complies with the Encoded Archival Description standard. EAD is an encoding standard developed initially by the Society of American Archivists (SAA) and maintained now by the SAA and The Library of Congress (LC). It built upon the International Council of Archives’ General International Standard for Archival Description (ISAD(G)), but quickly transcended guidelines about the creation of printed finding aids and developed into an xml encoding standard against

which finding aids for the internet could be measured. For the Alexander Turnbull Library, the creation of an Arrangement and Description team in 2011 and its adoption of this standard significantly assisted harmonisation of practice. The previous TAPUHI system (1991-2015) had formalised differences in approach based on format – at the most extreme, requiring staff to re-accession parts of multi-format collections by type, and requiring researchers to follow complex textual information in order to reconstruct them.<sup>7</sup> The requirement to map metadata to EAD, and the realisation that this could be done, was the beginning of a reassessment of how common descriptive requirements could be brought together again. For example <extent> in EAD may include cm, mm, time in minutes, linear metres, numbers of reels of tape (of different sizes), digital size, and more. Tags concerning language usage may incorporate English, Māori, Danish, French and so on, but also no linguistic content. The values may be linked to other international encoding agreements or controlled vocabularies, which will make the resulting finding aids easier for the Library to validate, for computer-based instructions to share and understand, and for researchers to navigate. Efforts to harmonise descriptive practices which were formerly spread across separate TAPUHI ‘accounts’<sup>8</sup> clarified the Library’s own metadata and allowed it to be checked and relocated if necessary. ‘General Notes’, for example, had been used differently at different times – in the course of the data mapping project which led to these datasets, the Library was able to assign metadata to its appropriate EAD tag, which meant that it would automatically be published in the correct part of a finding aid. Data migration also enabled the Library to begin key thinking about its multitudinous permission statements, which feed into tags for Access Restrictions, Use Restrictions, and Copyright. From these, delivery processes could be streamlined. Other international standards are referenced within the metadata – for example repository, language, and language code. It is one of the advantages of EAD that this referencing can be wrapped into each finding aid and therefore retained within the dataset.

There may be many varieties of EAD globally, so three particular characteristics of the Alexander Turnbull Library’s dataset deserve mention. Firstly, the dataset complies with EAD 2002, as this was available ‘out of the box’ with the system purchased<sup>9</sup>. The metadata therefore contains some elements which have been deprecated in the recent revision of EAD3. Secondly, at present, the Turnbull dataset does not include the online addresses for the many born-digital objects or digital surrogates which are surfaced through the Library’s National

Digital Heritage Archive (NDHA) viewing channel. This is because the Arrangement and Description on which our metadata is based, is carried out on the original objects, not on their later online presentation or performance. The relevant EAD tags 'digital archival group' and 'digital archival object' therefore contain the PID reference which refers to the unique identifier within the digital repository. Future development may address this within the metadata, or future users may develop a tool which maps the dataset's PID to an online version.<sup>10</sup> Thirdly, the dataset reflects several legacy systems knitted together, and there has been limited effort to re-integrate format-based separations. For example, if a collection originally contained manuscript and photographic and ephemeral and map components (in analogue or digital form), each would have been described to the in-house conventions, and there should be textual links allowing their relationship to be traced. These textual links are still present; there are no live internal links as yet.

'Turnbull Names' is the title given to the dataset for records covering the names referenced in metadata on the Turnbull's unpublished collections. These records comply with the international Encoded Archival Context – Corporate Persons and Families (EAC-CPF) standard. This dataset contains information about all the people indexed in some way in relationship to the unpublished collections, the nature of that relationship (e.g. as Creator, Donor, or Subject). It also provides links to the descriptive records, or the other entity records, to which the term relates. The Library considers that this dataset has great potential for digital researchers, not only across the Turnbull's collections but also in potential relationships to other research tools.

It has taken considerable work to ensure that this Turnbull Names dataset complies with EAC-CPF, because the standard requires definition as a 'Person', a 'Corporate Body', or a 'Family'. Previous 'Name' records had not included such a field, and had indeed grown over the years to include names of horses, ships, electric stoves, works of fiction, and other 'entities' incapable of creating or donating records. Data migration rules coped with the bulk of this (e.g. at their simplest – 'Firm' = corporate body; 'Mrs' = person, etc. ) but it was a considerable data clean-up project to clarify around 40,000 names which had still to be migrated as 'Other'. In early 2017 Arrangement and Description staff are still working through some 4,000 remaining 'Other' records, which are not included in the dataset. Most of them will probably be re-cast as Subject terms. This sort of re-structuring of metadata behind the scenes has little crowd appeal, but is essential if ATL metadata is to be available for tools which enable comparison, disambiguation, or add value in the

wider online research environment. At the same time, potential users of this dataset must remember that the metadata is derived very specifically from the Alexander Turnbull Library's core business and resources. Thus, the dataset does not include every name in the unpublished collections' content (because the Library has not had resources to index everything). Nor is it a biographical dictionary, because staff are concerned primarily to contextualise the material held at the Library – not to research and write a whole-of-life essay. Nor do the records contain all personal information, because the Library is concerned with issues of sensitivity and privacy.

The third standard to which these datasets comply is the Unicode character set. Previously, the unpublished collections' computer system TAPUHI did not have the capability to communicate (i.e. enter as text and reliably store, search, print etc.) characters other than a-z/ A-Z and basic punctuation. This did not enable respect either for the content of the unpublished collections (which cover many languages and their diacritics), or for their donors and researchers. The software behind these two datasets has recently been upgraded to enable staff to reliably enter Unicode characters, and enables the public to receive consistent results whether they search with or without diacritics. The Library is using this new capability for new records, but has not applied diacritics retrospectively across legacy metadata because this requires considerable thinking about the implications and documentation around such a change (as well as the resource to undertake it). The first steps, however, will be to add macrons over 2017 to the system versions of Ngā Upoko Tukutuku / the Māori Subject Headings list, and of the Iwi / Hapū list. This will align these terms with other instances of their use online. Additionally, staff will consider how to bring macrons in legacy Names into line with current practice.

Last but not least, the datasets consist of metadata created by the Library, and are therefore Crown Copyright. (N.B. It is crucial that staff and users understand that this is completely separate from permission requirements covering access and use / reuse of the *content* of collections, which is of course still covered by the appropriate metadata about that collection/item). Crown Copyright means that this metadata can be made available under a Creative Commons licence, as is preferred for government information. The licence is CC4.0 BY, meaning that the datasets may be used with attribution, for commercial and non-commercial purposes.<sup>11</sup> To date, the chief expressions of interest have come from individual digital humanists or software developers<sup>12</sup>, however one interesting potential use may be through commercial services which

would re-surface Alexander Turnbull Library metadata about unpublished collections, alongside other resources.

Requiring that the Library's systems have the ability to deliver to these standards created a vigorous framework for the recent data migration and IT project. The chosen vendor already delivered a basic EAD 2002 report 'out of the box' and now delivers an EAC-CPF report also. The new system Tiaki will deliver datasets regularly (at this stage, monthly), and they can be validated against international tools<sup>13</sup>.

At this early stage, the datasets will be available through two delivery channels – the NLNZ's ftp site and the NLNZ's webpages. One pilot user has expressed a preference for OAI-PMH delivery, another has suggested GitLab – the Library will be exploring these options too.

These are interesting times, when 'authority' is increasingly challenged, and balanced, by individual or 'community' or 'popular' perspectives. In such circumstances to publish an 'authority set' requires clarity of purpose, and some contextualisation. As this article has indicated, the new datasets service caps five years' effort to modernise the Library's arrangement and description approach. Standardising descriptive metadata in order to generate these datasets has immediate benefits for digital researchers, and has positioned the Library well for networking, partnerships, and development. In the next stage, the Turnbull is very keen to see how clients will use the datasets and how the research and archival communities will benefit long-term, to have interesting new conversations and partnerships, and to keep reporting what is discovered.

*Nicola Frean, Leader Arrangement and Description  
Alexander Turnbull Library, National Library of New Zealand*

#### Endnotes

1. The datasets are available via [www.data.govt.nz](http://www.data.govt.nz)
2. Note: the Library uses several other controlled vocabulary indexing terms to increase the collections' discoverability. All of these are included in the EAD dataset as terms, however the Turnbull does not intend to share these as datasets in the same way. This is because they are either selections from externally managed authority lists, or do not have international standards against which they can be validated. The Library hopes to develop partnerships which will make linking to external datasets easier in future.
3. <https://www.ict.govt.nz/guidance-and-resources/open-government/declaration-open-and-transparent-government/>, retrieved 3 February 2017.
4. National Library of New Zealand's Use and Reuse Policy (<https://natlib.govt.nz/files/strategy/Useand-reuse-policy.pdf>).
5. Other steps, in this Turnbull Datasets instance, have included: covering Metadata in Policy statements, a Privacy Impact report, the addition of a LICENCE.txt file, advice on culturally sensitive metadata, text for the web pages, and measures to ensure regular refreshment and validation, and that users are now monitored.
6. Melinda Johnston and Thomas Kontges 'Increasing metadata usability for the New Zealand Cartoon Archive at the Alexander

- Turnbull Library', in *Alexandria: The Journal of National and International Library and Information Issues* 2016, Vol. 26(2) 91-106.
7. See John Sullivan's blog on the National Library of New Zealand's website for further background on the transition from TAPUHI, <https://natlib.govt.nz/blog/posts/tapuhi-takes-a-bow> retrieved 2 February 2017.
  8. TAPUHI 'accounts' consisted of Manuscripts (including Music), Photographic Archive, Drawings Paintings and Prints (including the Cartographic Archive, digital Cartoons, and Ephemera), and Oral History and Sound. There were additional tools shared across these accounts, for consistent Name, Place, Subject, Iwi/Hapu, and Genre records.
  9. Emu, from KE Software, now part of the Axiell Group.
  10. This is considered relatively straightforward as the PID is included in the address.
  11. Approval was confirmed as "Tiaki Metadata and Tiaki Datasets are released on a regular basis under a CC BY licence", internal document, 19 September 2016, National Library of New Zealand
  12. E.g. "The first dataset I intend to process is "People Australia"; a collection of biographical sources and published by the National Library of Australia's "People Australia". Hopefully soon after I will be able to add a related dataset from New Zealand." From Conal Tuohy's blog at <http://conaltuohy.com/blog/oceania/>, retrieved 2 February 2017.
  13. For example <https://validator.w3.org>.