Author/s:
Thieberger, N

Title:
The tip of the iceberg: Public good and the curation of humanities research records

Date:
2025

Citation:
Thieberger, N. (2025). The tip of the iceberg: Public good and the curation of humanities research records. Arena Journal

Persistent Link:
https://hdl.handle.net/11343/359696

ARENA
CRITICAL * RADICAL * AUSTRALIA * EARTH

THE TIP OF THE ICEBERG: PUBLIC GOOD AND THE CURATION OF HUMANITIES RESEARCH RECORDS

ARENA ONLINE
NICK THIEBERGER
25 JUL 2025

We stand at a point in history in which we can expect to see more humanities research data lost than is preserved. Research data is the primary material created in the course

of research: transcribed manuscripts, photographic images, media recordings and so on (hereafter referred to as primary records). This is typically publicly funded material, and, in the humanities it is of interest to the general public – if they could access it. Why can't they? Because the effort that goes into funding research is not matched by an effort to ensure that primary records created in that research are curated for future access and use. This is for two basic reasons: first, in most countries there is no national service that guarantees long-term curation and access to primary humanities records, and, second, most humanities researchers do not prepare primary records for reuse, even if they would subscribe to Jerome McGann's suggestion (below) that we all want our cultural records to be accessible. Humanities researchers are not trained in managing the primary records they produce and, as a result, this material is at risk of loss. In part both the lack of training and the risk of loss result from our disciplines needing to build methods for valuing the curation of primary data as a research output[1].

Humanities research data is intrinsically interesting to the broader public. Based on the discovery of manuscripts[2], we transcribe them and add new cataloguing information to increase findability of those records. We create what may be the only records in some of the world's small languages. We record otherwise ephemeral performances that can one day be used to relearn lost knowledge. While the academy has focussed on published outputs, articles and books, and provides longterm repositories for those outputs, it has not done so well at preserving this primary data.

All research projects need to plan for their endings. Given the uncertainty of funding, even established large collections need to consider how to survive a break in funds. The *Endings Project* in Canada[3] explored strategies for building endings into the initial planning of projects, and produced a special issue of *Digital Humanities Quarterly* on 'Project Resiliency'[4] which involves considerations, among others, such as: what are the key components of a project that need to be kept, what licences should be put on them, how should they be described in order to be findable in future.

As Jerome McGann noted in 2010, "Whether we work with digital or paper-based resources, or both, our basic needs are the same. We all want our cultural record to be comprehensive, stable, and accessible. And we all want to be able to augment that record with our own contributions."[5] The 'stable' and 'accessible' nature of this record is what permits augmentation, in what Evans[6] calls 'dialogic archiving', that is, the response today to material created in the past. But, between this intention and the reality lies a chasm, despite a number of statements and polices that should indicate progress towards curation of digital research materials (for example the Barcelona Declaration[7]), or works like Johnston 2017[8].

The digital, online archive affords new ways of interacting with primary records. It also, critically, permits the restitution of records that were unavailable for too long in analogue form, to the people recorded and to their descendants. With access to these

records, new interpretations can be provided to enrich the metadata, or indeed to transcribe manuscripts or recordings, which could then all go back into the archive to augment the record.

From an academic perspective, primary data is the warrant for claims made in research papers, and should therefore be accessible for readers who need to verify or to re-use that data for new research questions. So, for example, I arranged for two linguists, who had written a grammar and texts for the Nguna language from Vanuatu in the 1970s, to give permission for their books to be scanned and made available (with appropriate licences for re-use)[9]. I also arranged for their audio recordings to be digitised and archived. This created a corpus for this language that has since been used by several students who have written minor theses examining aspects of the language not considered in the original analytical work. Similarly, a rich collection of recordings made in Australian Indigenous languages from the Daly region of the Northern Territory formed the basis of fieldwork for a student with a small child. She was able to begin analysis of the archived materials, and later was able to travel to confirm her work with speakers.[10]

To make a theoretical claim based on the observation of some phenomena, good research practice requires some way for others to verify that same observation, with reference to a citable form of the phenomenon. More concretely, if a linguist claims that a certain sentence can be produced in a language, they should be able to link from their example sentence to the recording in which that sentence occurred. If an historian discovers a new fact in a manuscript that they found in an archive, they need to be able to point readers to an accessible (ideally online) form of that manuscript that they digitised in the course of their research.

Keep in mind that records made by university employees technically belong to the university that employed the researchers, and that universities enter into agreements with funding bodies that include providing longterm curation of primary records. For example, this is the wording of my current Australian Research Council grant on the issue of data retention (my emphasis):

> The university undertakes to collect and *maintain research data*in accordance with the Australian Code for the Responsible Conduct of Research (2018) (the 2018 Code).
> [...]
> (c) We strongly encourage that data arising from the project is *deposited in an appropriate publicly accessible discipline and/or institutional repository.*

But there is no repository for primary records at our university, nor at most universities. They do have publication repositories, and we are given file *storage* by our universities, (like a hard disk, Onedrive, Sharepoint, Dropbox, googledrive, Figshare,

etc), but that is not enough. Instead, what we need is *curated* storage that has the following characteristics that distinguish it from simple storage:

- Accessibility: curated storage must be publicly available (outside of academia) via mediated access
- Metadata: curated records have metadata to aid discovery and access
- An Application Programming Interface (API) that can be found by search tools on the web to locate items.
- Use of standard language identifiers in the metadata like the three letter ISO-639-3[11] codes, or Austlang[12] in Australia to make records findable
- Licence conditions that determine what can be done with the item
- Updating data formats over time. Mere storage of bitstreams without curation is insufficient
- Data conversion to delivery formats, like wav to mp3, tif to jpg, plus integrity checking, enforcing filenaming conventions and so on

In order to ensure the longevity of priceless cultural material, in 2003 we established a model that has implemented these criteria for curating primary records, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[13]. We saw many analogue tapes, made by researchers since the 1960s, that were at risk of loss. As they were recorded in Papua New Guinea, or in countries of the Pacific, there was no institution in Australia that was required to preserve them. We felt a responsibility to ensure that these recordings were returned, at least to local museums or cultural centres if not to the location where they were recorded. For each item in the collection, the depositor specifies what access conditions to apply. We use a take-down principle should any user find material that they consider should not be made public.

For linguists and musicologists, recordings from the mid-twentieth century until the late twentieth century were made on analogue tape, reels, cassettes, and later on minidisks, DAT tape, CDs and so on. All of this media is now at risk of being unplayable. We began digitising these items, following relevant standards as advised by our National Film and Sound Archive and the National Library of Australia, with a one-year infrastructure grant from the Australian Research Council. We developed the naming convention we needed for the files, and the metadata terms needed to describe them. This collection now represents 1,380 languages, many of which have no other presence on the web, in 280 terabytes, with over 18,500 hours of audio recordings, plus 3,700 hours of video, plus text, and images. Having built the repository for legacy materials, PARADISEC increasingly receives material from current fieldwork, so that approximately half of the audio files are now born digital. It needs to be emphasised that the nine thousand hours of analogue media that we have digitised would probably have been lost were it not for our efforts. And, yes, those files on the language of Nguna that were mentioned earlier are all in PARADISEC for future re-use.

Rather than expecting a collection like ours to survive on a server for 100 years, we keep migrating it to the most appropriate platform as it emerges. Currently stored in Amazon[14] S3 Object storage, it is also backed up daily to a different city. Our automated ingestion and file-checking systems mean that we can operate in periods of limited funding with few staff. An interesting development recently has been the use of Research Object Crate (RO-Crate)[15] to encode metadata together with the files for each item in a standards-compliant jsonLD format. This means we can create items in our catalogue as we have done for some time, but that, each time an item's metadata is saved, it now writes a file (an RO-Crate) to the collection so that the item is self-describing. This allows us to copy any items from the collection and not risk disconnecting them from the metadata description that provides necessary contextual information. It also allows us, for the first time, to search both the metadata and the data (such as text files and transcripts) in the collection, providing a richer way of finding material.

Once we can extract all items in the collection that come from a particular village, for example – even if they were recorded by different researchers at different times – we can then create a catalogue of just that subset of files based on the RO-Crates in the subset. We have delivered these sub-collections on small portable devices[16] with wifi transmitters so that people using mobile phones can access the catalogue and download files.

Having developed relationships with Pacific cultural centres by returning these recordings, they began requesting assistance with digitising their collections of tapes. We do this by applying for small grants, for example, with the Vanuatu Cultural Centre (450 tapes), the Solomon Islands National Museum (275 tapes), or the Yap National Archives (120 tapes). We hold copies in trust in PARADISEC that can refresh their collections should the need arise.

We are part of an international network of archives that focus on curating records of cultural performance, the Digital Endangered Languages and Musics Archives Network (DELAMAN). This allows us to share innovations and to collaborate on development of methods or tools. In working with this group, the point being made in this paper is continually reinforced: there are no national systems for curating humanities research data in most countries. The single example appears to be the French service known as Huma-Num[17] which guarantees the longevity of our French sister language archives, Pangloss[18] and Cocoon[19]. We hope that a similar system will be developed in Australia to care for the countless digital projects that have ended or will soon end. Equivalent systems are taken for granted in the Sciences. No-one would consider disposing of astro-physical data that is, in Australia, accumulated and archived at the rate of approximately 4PB per year[20], or 16 times the size of the collection we have built over 23 years.

The work we have done is really the tip of the iceberg, it has rescued valuable primary records of humanities research, but it is obvious from our work that much more has been created and lost than is being curated by projects like ours.

Acknowledgments

Thanks to David Nash, Amanda Harris, and Linda Barwick for helpful comments.

---

[1] Nick Thieberger, Anna Margetts, Stephen Morey, Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics*. Vol 36: 1, 1-21

[2] For example, I have discovered information in manuscripts that were previously on paper and in a single location simply by digitising them and making their contents searchable. A further failing of our current research methods is that there is generally no way of knowing if someone else has already digitised the same records in the past, since holding institutions (libraries, archives, and so on) often do not have the capacity to accept the digitised version of the records they hold on paper, or to add a note to their catalogue to indicate that this work has been done.

[3] https://endings.uvic.ca/

[4] https://digitalhumanities.org/dhq/vol/17/1/index.

[5]McGann, Jerome. 2010. "Sustainability: The Elephant in the Room." Conference presentation at Online Humanities Scholarship: The Shape of Things to Come. Ms. http://shapeofthings.org/papers

[6] Evans, Nicholas. 2024. Keeping time: how the digital repatriation of western Arnhem Land song traditions deepens their meaning. In Thieberger, Nick, Amanda Harris, Sally Treloyn and Myfany Turpin. 2024. *Keeping Time: Dialogues on Music and Archives in Honour of Linda Barwick*. Sydney: SUP. Pp. 23-42

[7] https://barcelona-declaration.org/

[8] Johnston, Lisa R. 2017 *Curating Research Data Volume One: Practical Strategies for Your Digital Repository*. Chicago, Illinois: Association of College and Research Libraries

[9] These items are located in two collections:
https://catalog.paradisec.org.au/collections/EF1,
https://catalog.paradisec.org.au/collections/AS1

[10] https://www.paradisec.org.au/blog/2023/04/grammars-from-archival-records/

[11] https://iso639-3.sil.org/code_tables/639/data

[12] https://aiatsis.gov.au/austlang/data

[13] PARADISEC has since won awards for its practice, most recently the Digital Preservation Award from the Research Data Alliance (2024) and the Canadian Social Knowledge Institute's Open Scholarship Award (2025)

[14] We chose to opt out of any use by Amazon of our material, and the servers are located in Australia

[15] https://www.researchobject.org/ro-crate/ supported by the Language Data Commons of Australia (https://www.ldaca.edu.au)

[16] Raspberry Pi computers, see https://www.paradisec.org.au/blog/2024/02/using-raspberry-pi-in-ranongga/

[17] https://www.huma-num.fr/

[18] https://pangloss-dev.huma-num.fr/ Note that this is still in development so that not all the functionality of the PANGLOSS site is available in the Huma-Num version.

[19] https://cocoon.huma-num.fr/exist/crdo/

[20] https://www.csiro.au/en/about/facilities-collections/ATNF/ATNF-data-archives

ABOUT THE AUTHOR

NICK THIEBERGER

Assoc/Prof Nick Thieberger is a linguist at the University of Melbourne and is the Director of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). He works with speakers of languages of Vanuatu and is supporting the preservation of language records with a number of agencies in the Pacific.

More articles by Nick Thieberger

CATEGORISED: ARENA ONLINE

ADD COMMENT