




Collections as data: what's next?

Margaret Warren
Director, Content Management
State Library of Queensland
margaret.warren@slq.qld.gov.au
 <https://orcid.org/0000-0002-4248-300X>

Abstract:

Collections as data is an approach to facilitating research with cultural heritage digital collections from galleries, libraries, archives and museums (GLAM) using computational methods. The Vancouver Statement on Collections as Data (2023) sets out principles for working with collections as data, which can be used by GLAM institutions across the world to make aspirations for ethical, responsible and sustainable use of collections as data a reality. This paper looks at the journey to the Vancouver statement, the iteration of the principles from the earlier Santa Barbara Statement on Collections as Data (2017) and proposes using the principles to progress this work across Australian GLAM institutions.

First published 9 July 2024



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International Licence](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Introduction

The Vancouver Statement of Collections as Data was released in 2023 following an international two-day meeting of practitioners and strategists, researchers and academics, who asked critical questions about how the work of collections as data has progressed since 2017, and what is its future. The Statement is an update to the Santa Barbara Statement on Collections as Data (2017) (Padilla et al. 2018) and reflects the changed environment in the period since the original statement was released, especially in relation to heightened concerns about data sovereignty, the rapid proliferation of artificial intelligence technologies being used with collections data, the environmental impacts of large-scale computing, and known historic and contemporary inequities represented in GLAM collections (Padilla et al. 2023).

This paper will look at what's next for collections as data in the Australian context following the release of the Vancouver Statement, and challenge Australian GLAM organisations to take up the opportunities and challenges of ethical, sustainable, accessible and responsible computational use of collections data.

The journey to the Vancouver Statement

Researchers have always worked with collections from institutions such as libraries, museums and archives, as data to inform their study, research and the development of new knowledge and stories. What changed from the beginning of wide-spread access to the internet and development of tools to store, manage, provide access to and interrogate increasingly large amounts of data, was the ability to consider and then use computational methods for research. At the same time cultural institutions began serious work on digitisation, including mass digitisation projects to increase online access to collections in a wide range of formats – text materials such as newspapers, books and manuscript collections, photographic collections, and audio and video materials such as oral history collections and digitised film. An early example is the British Library, which in 2004 began to digitise newspaper collections with a 2-million-pound grant from the Joint Information Systems Committee (JISC) in the United Kingdom, with a goal of making 2 million pages available online (Beals and Bell 2020). In Australia, the National Library commenced digitisation of collections in the early 2000s, with more than 42 million pictures, maps, manuscripts, books, serials, and newspapers digitised by June 2022 (National Library of Australia 2024). The environment, content, technologies and capacity have become more accessible for using computational methods in research with the rich digital resources of more GLAM institutions.

Always Already Computational: Collections as Data

The introduction to the final report of the *Always Already Computational: Collections as Data* project stated that, “While cultural heritage practitioners have broad experience replicating the analog experience of watching, viewing, and reading in a digital environment, they less commonly share the experience of supporting users who want to work with collections as data - a conceptual orientation to collections

that renders them as ordered information, stored digitally, that are inherently amenable to computation” (Padilla et al. 2018, p.7). This desire to bring together cultural heritage practitioners and the research community around the opportunities and challenges inherent in working with collections as data was the impetus for the *Always Already Computational: Collections as Data* project.

From 2016-2018 the project documented and shared with interested stakeholders what were then current approaches and potential innovation opportunities to develop cultural heritage collections to support computationally driven research. The project was funded by the Institute of Museum and Library Services. *Always Already Computational: Collections as Data* held two national forums in the United States, and generated a range of deliverables intended to guide institutions as they considered development of collections as data. These included a report, a tool designed for practitioners and organisations seeking to get started with collections as data named 50 things, Collections as Data Facets, documenting collections as data implementations, Collections as Data Personas, representing a set of high-level roles associated with collections as data activity, and the Santa Barbara Statement on Collections as Data. (Padilla et al. 2018)

The Santa Barbara Statement on Collections as Data

The Santa Barbara Statement on Collections as Data was a set of principles developed to ‘guide practitioners through the practical, theoretical and ethical dimensions of collections as data work’ (Padilla et al. 2018). Framed to raise questions rather than outline solutions, the outworking of the principles was intended to be in local contexts. Source

The principles of the Santa Barbara Statement on Collections as Data are:

1. Collections as data development aims to encourage computational use of digitised and born digital collections.
2. Collections as data stewards are guided by ongoing ethical commitments.
3. Collections as data stewards aim to lower barriers to use.
4. Collections as data designed for everyone serve no one.
5. Shared documentation helps others find a path to doing the work.
6. Collections as Data should be made openly accessible by default, except in cases where ethical or legal obligations preclude it.
7. Collections as data development values interoperability.
8. Collections as data stewards work transparently in order to develop trustworthy, long-lived collections.
9. Data as well as the data that describe those data are considered in scope.
10. The development of collections as data is an ongoing process and does not necessarily conclude with a final version. (Padilla et al. 2018)

Collections as Data: Part to Whole

Always Already Computational: Collections as Data was succeeded by *Collections as Data: Part to Whole*, to advance collections as data implementation and use. *Collections as Data: Part to Whole* funded and supported 12 project teams, jointly led by librarians and disciplinary scholars in academic institutions in the United States. Their remit was to develop models and approaches to support collections as

data implementations. The projects aspired to “exhibit high research value, demonstrate the capacity to serve underrepresented communities, represent a diversity of content types, languages, and descriptive practices, and arise from a range of institutional contexts.” (Padilla 2019) The scope and range is illustrated in a sample of the identified projects: *Surfacing hidden water data: Water, people, displacement in Southern California*, from Claremont Colleges, *LGBTQ+ Audio Archive Mining Project* from the University of Wisconsin Milwaukee, and *Uncovering Health History: Transcribing and Publishing Early Twentieth-Century Tuberculosis Patient Records as Data* from the University of Denver.

Vancouver statement

The development and release of the Vancouver Statement on Collections-as-Data, as an update to the Santa Barbara Statement on Collections as Data, was the result of an international working event, *Collections as Data: State of the Field and Future Directions*, held in Vancouver, Canada in April 2023, as well as community feedback, and was a further deliverable of the *Collections as Data: Part to Whole project*.

The introduction states that, “The Vancouver Statement was created to be approached both by people who are new to the idea of supporting responsible development and computational use of collections as data, as well as people who are well-versed in long established approaches to gallery, library, archive, and museum practice.” (Padilla et al. 2023, p.2)

The principles are:

1. Collections-as-data development supports responsible computational use of digitised and born digital collections.
2. Collections-as-data stewards are committed to working against historic and contemporary inequities represented in collection acquisition, scope, description, access, and use.
3. Collections as data should be widely accessible, within the bounds of ethical, legal, and community expectations.
4. Collections-as-data benefit from participatory design.
5. Shared documentation provides context and helps others find a path to doing the work.
6. Collections-as-data development values social and technical interoperability.
7. Collections-as-data stewards work transparently in order to maintain integrity for long-term access to collections.
8. The progression of collections-as-data work depends on organisational commitments to sustainable infrastructure and processes.
9. The work of developing collections as data should balance benefits with concern for climate impact.
10. The work of developing collections as data should balance benefits with concern for exploitative labor.
11. Collections-as-data stewards should design access with appropriate consideration of data consumption by artificial intelligence or other technologies. (Padilla et al. 2023)

Iteration of thinking and practice with collections as data

A comparison of the principles in the Santa Barbara and Vancouver Statements of Collections as Data reveals some clear themes demonstrating the iteration of both thinking and practice in relation to collections as data, for both cultural heritage institutions and researchers and practitioners.

Internationally there have been heightened concerns about the intersection of collections as data work and data sovereignty, including issues related to Indigenous Cultural and Intellectual Property (ICIP) and the rights of First Nations peoples in the United Nations Declaration on the Rights of Indigenous Peoples, particularly the “right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions” (United Nations 2007). This can be seen in the inclusion of principle 3 in the Vancouver Statement, “Collections as data should be widely accessible, within the bounds of ethical, legal, and community expectations” (Padilla et al. 2023, p.3) in contrast to principle 6 in the Santa Barbara Statement, “Collections as Data should be made openly accessible by default, except in cases where ethical or legal obligations preclude it” (Padilla et al. 2018, p.4).

The environmental impacts of large-scale computing are well documented, and exponentially increasing. Anthropologist Steven Gonzalez Monserrate’s case study *The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage*, notes, “the Cloud now has a greater carbon footprint than the airline industry. A single data center can consume the equivalent electricity of 50,000 homes. At 200 terawatt hours (TWh) annually, data centers collectively devour more energy than some nation-states” (Monserrate, 2022). This provides tension for researchers and institutions working with collections as data and is reflected in principle 9 of the Vancouver Statement, “The work of developing collections as data should balance benefits with concern for climate impact” (Padilla et al. 2023, p.4), and in that it did not appear as an issue in the Santa Barbara Statement is significant.

In a similar way, while machine learning and large-scale text mining were technologies well used in 2017, the rise of other artificial intelligence technologies, particularly generative AI, and concerns about indiscriminate use of content in generative AI applications, is reflected in principle 11 in the Vancouver Statement, “Collections-as-data stewards should design access with appropriate consideration of data consumption by artificial intelligence or other technologies” (Padilla et al. 2023, p.4).

The known historic and contemporary inequities represented in GLAM collections were not acknowledged in the Santa Barbara Statement, and the intentional inclusion of a principle in the Vancouver Statement that addresses the requirement for action as principle 2, “Collections-as-data stewards are committed to working against historic and contemporary inequities represented in collection acquisition, scope, description, access, and use” (Padilla et al. 2023, p.2) is perhaps the most challenging iteration in the field, but also the area where intentional action could potentially make the most difference.

So, what shall we do next?

What will Australian GLAM organisations do to take up the opportunities and address the challenges of ethical, sustainable, accessible and responsible computational use of collections as data? The landscape is by no means a blank slate. There are sophisticated and innovative projects and initiatives being undertaken in the academic and GLAM sectors. Three examples illustrate different approaches to progress opportunities with collections as data.

Work is being undertaken by Dr. Sam Hames, post-doctoral Research Fellow, and Dr. Naomi Barnes on the transcribed and digitised Proceedings of Federal Parliament, consisting of nearly a billion words recorded since 1901. They are examining how computational methods on the proceedings can create more generous search interfaces and bring to the surface new ways of looking at policy evolution and decision making by government, and presented some of their findings at the *Making Meaning: collections as data symposium* held at State Library of Queensland in March 2024 (State Library of Queensland 2024). Hames is a Research Fellow in Computational Humanities, and Barnes is interested in how crisis influences education politics, with a specific focus on moral panics. Their work is an example of the principle of participatory design, where specialists from different fields collaborate on collections as data projects.

The GLAM workbench is a collection of tools, tutorials, and examples, compiled by Tim Sherratt to help practitioners and researchers work with data from galleries, libraries, archives, and museums with a focus on Australia and New Zealand. The GLAM Workbench focuses on tools and training materials to support researchers to create research datasets from a variety of GLAM collections (Sherratt 2021). Sherratt's work is an example of sharing documentation and expertise to provide context to researchers to find a path to doing their own collections as data work.

One of the questions posed in the creation of the *Virtual Veteran*, a chat bot taking on the persona of a First World War soldier, was how to use generative AI with collections data, and to limit the scope to trusted sources of information and to include citations to the data sources used by the chat bot to create the answers. The application, developed by State Library of Queensland, used digitised First World War diaries and manuscripts from the library's collection, digitised newspaper articles from the collections on Trove, and the text of C.W. Bean's First World War Histories as the data sources, and created the application using only these data sources in a closed system, not able to be used by other generative AI applications (State Library of Queensland 2024). This project is an example of designing access to collections as data with appropriate consideration of data consumption by artificial intelligence or other technologies to drive a creative output.

However, there remain challenges to fully embracing the Vancouver Statement principles as a benchmark for GLAM institutions' work with collections as data.

Working against historic and contemporary inequities represented in collection acquisition, scope, description, access, and use will require ongoing and intentional focus for all staff working across the lifecycle of collections. Integration of ICIP principles, consideration of data sovereignty, and consultation with data owners within these frameworks will be required when working with First Nations collections

as data. Work such as the *Guidelines for First Nations collection description* provide a strong starting point for descriptive practice (Raven, 2023). The National and State Libraries Australasia (NSLA) has endorsed the ATSLIRN¹ protocols, ALIA Statement on Copyright and Intellectual Property, IFLA Statement on Indigenous Traditional Knowledge and the United Nations Declaration on the Rights of Indigenous People in its *Position statement: Indigenous Cultural and Intellectual Property (ICIP)* (NSLA 2023).

Addressing sustainability and the environmental impact of storage and large-scale computing will require changes to procurement processes; investing in cloud providers who use renewable energy, making improvements in energy efficiency and collaborating with industry partners on more eco-friendly solutions. As Monserrate notes, “The ecological dynamics we find ourselves in are not entirely a consequence of design limits, but of human practices and choices - among individuals, communities, corporations, and governments - combined with a deficit of will and imagination to bring about a sustainable Cloud. The Cloud is both cultural and technological. Like any aspect of culture, the Cloud’s trajectory - and its ecological impacts — are not predetermined or unchangeable” (Monserrate 2022).

Finally, increasing the sharing of experiments, successes and failures, knowledge gained, and innovative solutions delivered across the GLAM sector will both encourage and inspire us to progress ethical, responsible and sustainable collections as data work, and to continue to meaningfully contribute to research, learning and storytelling using our digital collections.

References

- Allen L and Enderle L (2019) *Text mining*, Always Already Computational Project, accessed 25 June 2024. <https://collectionsasdata.github.io/methodsprofiles/>
- Beals, M H and Bell E (2020) *British Library 19th century newspapers*, The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges, accessed 25 June 2024. <https://doi.org/10.6084/m9.figshare.11560059>
- Monserrate, SG (2022) *The staggering ecological impacts of computation and the cloud*, MIT Press Reader, accessed 25 June 2024. <https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/>
- NLA (National Library of Australia) (2024) *Digitisation of library collections*, NLA, accessed 25 June 2024. <https://www.nla.gov.au/collections/building-our-collections/growing-our-digital-library/digitisation-library-collections>
- NSLA (National and State Libraries Australasia) (2023) *Position statement: Indigenous Cultural and Intellectual Property (ICIP)*, NSLA, accessed 25 June 2024. <https://www.nsla.org.au/resources/indigenous-cultural-and-intellectual-property-icip>
- Padilla T et al. (2018) *The Santa Barbara statement on collections as data*, Always Already Computational Project, accessed 25 June 2024. <https://zenodo.org/records/3066209>
- Padilla T et al. (2019) *Always already computational: collections as data final report*, Always Already Computational Project, accessed 25 June 2024. <https://zenodo.org/records/3152935>
- Padilla, T. (2019) *Call for Proposals*, Collections as data - part to whole, accessed 24 June 2024. <https://collectionsasdata.github.io/part2whole/cfp/>
- Padilla T et al. (2023) *Vancouver Statement on collections as data*, Collections as Data: Part to Whole Project, accessed 25 June 2024. <https://zenodo.org/records/8342166>
- Raven, T (2023) *Guidelines for First Nations Collection Description*, Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS), the Australian Library and Information Association (ALIA), the Council of Australian University Librarians (CAUL), CAVAL, National and State Libraries Australasia (NSLA). Accessed 25 June 2024. <https://nla.gov.au/nla.obj-3250767341/view>
- Sherratt T (2021) GLAM Workbench (version v1.0.0), accessed 25 June 2024. <https://doi.org/10.5281/zenodo.5603060>
- State Library of Queensland (2024) *Making meaning 2024*, State Library of Queensland, accessed 25 June 2024. <https://www.slq.qld.gov.au/making-meaning-2024>.
- State Library of Queensland (2024) *Virtual Veterans: your AI guide to a rich collection of World War I resources*, State Library of Queensland, accessed 25 June 2024. <https://www.anzacsquare.qld.gov.au/virtual-veterans>

UN (United Nations) (2007) *United Nations declaration on the rights of indigenous peoples*, UN, accessed 25 June 2024. <https://social.desa.un.org/issues/indigenous-peoples/united-nations-declaration-on-the-rights-of-indigenous-peoples>

Endnotes

ⁱ Aboriginal and Torres Strait Islander Library, Information and Resource Network Inc