DHQ: Digital Humanities Quarterly
2024
Volume 18 Number 2

Towards a National Data Architecture for Cultural Collections: Designing the Australian Cultural Data Engine

Rachel Fensham <rachel_dot_fensham_at_unimelb_dot_edu_dot_au>, University of Melbourne; Australian Cultural Data Engine https://orcid.org/0000-0002-0339-4848
Tyne Daile Sumner , Australian National University; Australian Cultural Data Engine https://orcid.org/0000-0003-1717-7147
Nat Cutter , University of Melbourne; Australian Cultural Data Engine https://orcid.org/0000-0002-3699-0881
George Buchanan , RMIT University https://orcid.org/0000-0001-9044-6644
Rui Liu , University of Melbourne https://orcid.org/0000-0001-5631-8702
Justin Munoz , Independent Scholar https://orcid.org/0000-0002-5339-2904
James Smithies , Australian National University https://orcid.org/0000-0003-4801-0366
Ivy Zheng , University of Newcastle https://orcid.org/0009-0006-7028-9980
David Carlin , RMIT University https://orcid.org/0000-0002-8902-0918
Erik Champion , University of South Australia https://orcid.org/0000-0002-5362-6176
Hugh Craig , University of Newcastle https://orcid.org/0000-0002-9336-1678
Scott East , University of New South Wales https://orcid.org/0000-0001-7367-7707
Chris Hay , Flinders University https://orcid.org/0000-0001-8507-3556
Lisa M. Given , RMIT University https://orcid.org/0000-0003-1840-6175
John Macarthur , University of Queensland https://orcid.org/0000-0002-8787-1305
David McMeekin , Curtin University https://orcid.org/0000-0001-6445-1183
Joanna Mendelssohn , University of Melbourne https://orcid.org/0000-0001-7562-668X
Deborah van der Plaat , University of Queensland https://orcid.org/0000-0001-5379-9395

Abstract

This article summarises the aims, methods, information architecture, outputs, and innovations of the Australian Cultural Data Engine (ACD-Engine), a project that harnesses leading cultural databases to build bridges to research, industry, and government. The project investigates digital heritage collections, data ontologies, and interoperability, building an information architecture to enhance the open sharing of Australian cultural data. Working with a cross-disciplinary team, the ACD-Engine establishes conceptual and technical frameworks for better understanding the platforms and uses of cultural data across a range of national and international contexts. This new cyber-infrastructure advances cultural data aggregation and interoperability whilst prioritising data quality and domain distinctiveness to answer new research questions across disciplines. As such, the ACD-Engine provides a novel approach to data management and data modelling in the arts and humanities that has significant implications for digital collections, digital humanities, and data analytics.

Introduction

The Australian Cultural Data Engine (ACD-Engine) is a multidisciplinary project that harnesses leading Australian cultural databases to analyse cultural production, artistic networks, and the socio-economic implications and uses of arts and cultural data in the Australian context and internationally. Commencing in August 2021, the project received two years' funding through the Australian Research Council (ARC) Linkage, Equipment, Infrastructure and Facilities (LIEF) program. Based at the University of Melbourne, the ACD-Engine gathered a collaborative team of humanities 1 and social science researchers, data scientists, data engineers, and data visualisation specialists from the University of Melbourne, the University of Queensland, the University of New South Wales, Swinburne University of Technology, Curtin University, RMIT University, the University of Newcastle, Flinders University, and King's College London.

The core ACD-Engine team in Melbourne worked closely with institutional partners and domain experts from diverse cultural data backgrounds (architecture, design, performing arts, and visual arts) to foster improved data analytics and cross-disciplinary data sharing within the Australian arts and cultural sector. In line with approaches emerging in Europe and the UK, as well as other research and data-intensive contexts in Australia, we investigated the data quality and data ontologies that inform Australia's cultural data landscape. To coordinate an interoperable model for linking, interpreting, and using heterogeneous digital heritage data, the project team built a Research Software Engineering (RSE) capacity that could interrogate the potential of more robust, larger-scale understandings of arts and cultural data in future. The ACD-Engine has also examined how data analytics can provide meaningful insights to issues of national significance including employment, heritage, and cultural policy. 2

This article contributes to a growing genre of scholarship about collaborative research design within the digital humanities, including [Ahnert et al. 2023] [McGillivray et al. 2020]. Writing from a cross-disciplinary perspective, in this article we describe the project's objectives, design, and multi-stage approach to the extraction, aggregation, and analysis of a wide range of cultural data. Reporting on the organisation of research design and activity has been standard practice in government and the commercial sector for decades and helps to establish standardised and stable frameworks for research [Desfray and Raymond 2014]. Continuing this developing best practice, we outline our project's aims and objectives, contextualise these among some major national and international cultural data aggregation projects, and explain how we assembled our team. It was essential to establish principles for open data sharing among our project partners, so this process is also explained. As the central outcome of the project, we then provide a detailed overview of the ACD-Engine Architecture Construction Workflow that has resulted in the ACD-Engine information architecture prototype, as a significant contribution towards a national data architecture for Australian artistic and cultural collections. Developing from our work in data exploration and enrichment, we outline some cultural data outputs emerging from the ACD-Engine that illustrate how the discrete datasets developed by the project offer productive applications in humanities research, creative industries, and cultural policy. Finally, we summarise the project's key outcomes and suggest possible future directions for the generation, interoperability, and application of Australian cultural data that might build upon the ACD-Engine's work. 3

Aims and Work Programs

Arts and cultural data accumulate in the catalogues of collecting institutions (e.g., museums, art galleries, and member organisations), bespoke disciplinary datasets (e.g., biographical indexes or catalogues), curated digital repositories and archives (e.g., specialist libraries or websites), and in a range of smaller scale heritage, community, and creative industry organisations [Whitelaw 2015]. It also exists in government reporting, census and statistical modelling of culture, and online digital platforms providing links to cultural events. These disparate systems represent a “cyber-infrastructure” of humanities knowledges and expertise, but researchers and policymakers often lack the data literacies that would enable this abundance of cultural data to substantively contribute to global social and policy transformations [Smithies 2017]. 4

The ACD-Engine emerged from recognition of the fragmented nature of cultural data collections and the different computational models upon which they have been built, as well as the distinctiveness, sensitivities, and vulnerabilities of specialist data repositories. In response to this, the project's central aims were: 5

1. Improve the quality of existing Australian cultural data across a range of partner databases and cultural disciplines;
2. Support an expansion in the accessibility, use, and interoperability of cultural data in research, industry, and government contexts; and
3. Demonstrate how new insights into arts and cultural production can emerge by improving cultural data interoperability across platforms, systems, and regions.

To achieve these aims, we established the following overlapping programs of work. These programs generally moved from processes concerned with data integration and design to questions of capability, analysis, access, and public impact, but each program was iterative and informed the others as the ACD-Engine project developed. 6

Program A: Project Design

- Establish an approach for effectively working with cultural data across multi-disciplinary teams from diverse disciplines, backgrounds, and skill-sets; 7
- Shape a flexible and diverse humanities-trained workforce with data science skills, who can manage the social and ethical considerations of cultural data; and
- Develop a shared understanding of the process of cultural data sharing, which honours the labour of content creators as well as those transforming, enriching, and analysing data.

Program B: Data Exploration and Transformation

- Using a range of scraping and data wrangling methods, transform existing cultural collections into datasets for immediate use in computational contexts; 8
- Identify and resolve inconsistencies, redundancies, errors, and gaps in the data;
- Liaise closely with subject matter experts to augment and refine datasets in accordance with industry knowledge, cultural histories, and personal experience;
- Develop conceptual frameworks for the creation of qualitative subsets for use in comparative analysis; and
- Explore new adjacent data sources to complement existing datasets for innovative analytics.

Program C: Information Architecture Development

- Investigate and interrogate named entities and ontologies within cultural data collections; 9
- Engage with existing standards in the digital humanities to develop a prototype information architecture for cultural data;
- Interconnect different cultural data collections without dissolving their distinctive discipline-specific and ontological structures.

Program D: Analysis and Impact

- Design innovative models for revealing patterns in Australian arts production, cultural consumption, and tangible and intangible cultural heritage; 10
- Create new resources for policy formation, cultural production, and research infrastructure development; and
- Extend access to cultural data to facilitate new cross-collection, cross-disciplinary, and cross-institutional research.

Program A.1: Project Context

Cultural Data Aggregation in North America, Europe, and the United Kingdom

National and international efforts to aggregate humanities databases are reshaping concepts of heritage, creativity, and cultural inclusion in the global sphere [Bettivia and Stainforth 2017] [Smithies et al. 2023]. However, as Lisa M. Given, Sarah Polkinghorne and Joann Cattlin point out, “while cultural data initiatives are growing in number, globally, they lack a cohesive, sustainable, and healthy ecosystem to enable collaboration and sharing across related contexts” [Given, Polkinghorne, and Cattlin 2023]. Several international models for the expansion and interoperability of cultural data warrant consideration. The DARIAH-DE repository is a digital archive for the long-term preservation of humanities and cultural research data, aggregated across various services and applications (https://repository.de.dariah.eu/search/). Europeana PRO makes millions of cultural data assets from European galleries, libraries, archives, and museums available for searching and downloading (https://pro.europeana.eu/page/datasets). Similarly, the Digital Repository of 11

Ireland (https://dri.ie/about-us/) was launched in 2015 to facilitate the “preservation, curation, and dissemination of Ireland's humanities, social sciences, and cultural heritage data”. In 2020, the United Kingdom launched Towards a National Collection (TaNC), an £18.9 million, five-year program of investment aimed at “creating a unified virtual national collection” of the UK's museums, libraries, galleries, and archives, “dissolving barriers between different collections”, and “opening UK heritage to the world” [UKRI 2022].

Most of these platforms have elected either to combine data into a single unified search engine or to provide a collection of external links. However, seeking to develop best practices for data sharing and data flows, some international aggregation projects and research groups, such as DHARPA (https://dharpa.org/), have begun to grapple with how to productively interconnect diverse data sources without flattening the distinctiveness of domain-specific collection ontologies. Another dimension to data aggregation and interoperability is the export and generation of cultural heritage datasets, which until recently has been dominated by the Open Archives Initiative Protocol for Metadata Harvesting (https://www.openarchives.org/pmh/), discussed in the section “Program B: Technical Design and Methods” below. 12

Cultural Data Aggregation and Interoperability in Australia

In Australia, the most important digital aggregator of cultural data is Trove, the National Library of Australia (NLA)'s search engine, which over the past several decades has evolved into a powerful platform offering access to the NLA's own collections, as well as other registered collections. While not specifically designed as research cyber-infrastructure, Trove has nevertheless attracted a wide local and international user base, ranging from professional researchers to the general public [Stainforth 2019]. Under its “People and Organisations” portal (https://trove.nla.gov.au/help/categories/people-and-organisations-category), Trove currently links to several cultural collections associated with the ACD-Engine, and NLA software developers as well as independent digital humanities experts have produced a range of tools that enable the searching and downloading of curated research datasets.[1] Historically, however, it has suffered from inadequate resourcing to maintain and renovate core infrastructure and develop research capability [Jones and Verhoeven 2022].[2] 13

Another important project has been the Humanities Networked Infrastructure (HuNI, https://huni.net.au/), a platform developed as part of the Australian government's National e-Research Collaboration Tools and Researchers programme. The design framework of HuNI recognises humanities data as consisting of “the various annotations, tags, links, associations, ratings, reviews, and comments produced during the humanities research process, together with the semantic ‘entities’ to which these annotations refer: concepts, persons, places, and events” [Burrows 2011]. HuNI's data-centred workflow, in a similar way to Trove, focuses on a “discovery environment” that enables browsing and searching across disparate data sources, though it does not house the data itself [Burrows 2013] [Burrows and Verhoeven 2016]. Australia also has a major centralised data repository, the Australian Data Archive (ADA). The ADA provides a central searchable catalogue of more than 5,000 social science datasets. It holds data from surveys, opinion polls, and censuses from Australia and the Asia-Pacific region (https://ada.edu.au/about-ada/). However, the ADA hosts very little material on arts and culture, and its uncurated, user-submitted datasets are not necessarily immediately suitable for computational analysis. A more encouraging development in the Australian cultural data landscape has been the formation of an Indigenous Data Network in 2018 (https://mspgh.unimelb.edu.au/centres-institutes/onemda/research-group/indigenous-data-network), and, more recently, national investment in a HASS and Indigenous Research Data Commons within the Australian Research Data Commons (ARDC). These projects justly champion the importance of Aboriginal and Torres Straits Islander custodianship of rich cultural traditions as they are expressed in data and housed in cultural collections. 14

In Australia, as much as in Europe, North America, and the United Kingdom, the demand for quality data and online access to data-rich sources of information has grown exponentially, with more stakeholders entering the field each year. As such, the cultural data landscape in Australia is broad, yet fragmented, and has been supported for decades by fixed-term national investment grants made to individual humanities, arts, and social sciences (HASS)-based digital tools, platforms, working groups, and labs (such as the ARDC's HASS Community Data Lab,: https://ardc.edu.au/project/hass-community-data-lab/). As Mike Jones and Alana Piper point out, the “funding models set up for traditional projects normally rely on them having a clear end date for financial support, after which digital 15 resource may not even be accessible, let alone updated” [Jones and Piper 2023]. Moreover, many individual research collections and digital databases have been built using bespoke technical approaches (often outsourced to private companies), which have led to institution-specific issues relating to design, access, maintenance, and use of data. Inevitably, the diversity of cultural data collections reflects different histories of data collection, management, curation, and maintenance, and can lead to intergenerational and sometimes competitive misalignments of the potentiality in cultural data. Similarly, and in spite of national and international calls for open data policies (see e.g., https://inke.ca/), the economic, cultural, and social data managed by governments is often limited by idiosyncratic data structures, short-term political support, or survey parameters (e.g., https://www.arts.gov.au/cultural-data-online).

Nonetheless, the appetite for accessing, connecting, and telling stories with Australian cultural data in the creative industries continues apace, as evidenced by a reinvigorated national cultural policy, REVIVE, announced in 2023 [Australian Government 2023]. Researchers across the humanities are also increasingly interested in the research possibilities to be realised by using structured cultural data and large digital collections. However, as Ahnert et al. rightly point out in their analysis of collaborative historical work using big data, relatively little research considers “the practical steps in getting hold of such data, and the restrictions with which it may come” [Ahnert et al. 2023]. 16

Program A.2: Project Structure and Team

Rather than trying to “solve” national collections in a single effort, we took a targeted, bottom-up approach to examine the larger problem space in a way that did not centre on acquiring more data, developing more tools, or building another aggregation platform. Instead, the project undertook a detailed exploration of heterogeneous cultural data collections, each with unique histories and infrastructures. We thus began with the same broad intent as the “Foundation” TaNC projects (https://www.nationalcollection.org.uk/Foundation-Projects), which sought to interconnect selected collections in vast, historic institutions, including the National Archives, the National Gallery, the Victoria and Albert Museum, the British Library, and the Tate Gallery. The ACD-Engine benefitted from the agility afforded by working with smaller, discipline-based databases with fewer institutional or historical barriers to cooperation (e.g., those committed to an open API). In so doing, the project focused on assembling discrete, cleaned, augmented, and merged datasets that could become usable in a variety of computational contexts and yet retain each collection's disciplinary distinctiveness and place within the national cultural data ecosystem [Zheng and Munoz 2023a]. 17

The ACD-Engine is, distinctively, a cyber-infrastructure that strategically built upon pre-existing projects, investments, and data models without erasing their unique affordances, histories, and investments. Centring on a skilled research software engineering and digital humanities team at the University of Melbourne, which experimented with innovative technical methods and theoretical interrogation, we established a multi-institutional network of discipline experts and digital humanities practitioners across Australia and the UK.[3] Harnessing this new capacity at the intersection of computation and cultural interpretation, we sought to enhance the contribution and impact of cultural data through focused interdisciplinary projects and themes.[4] 18

The following diagram explains how the contributing data sources were transformed through the ACD-Engine into new resources for a range of users and beneficiaries. 19

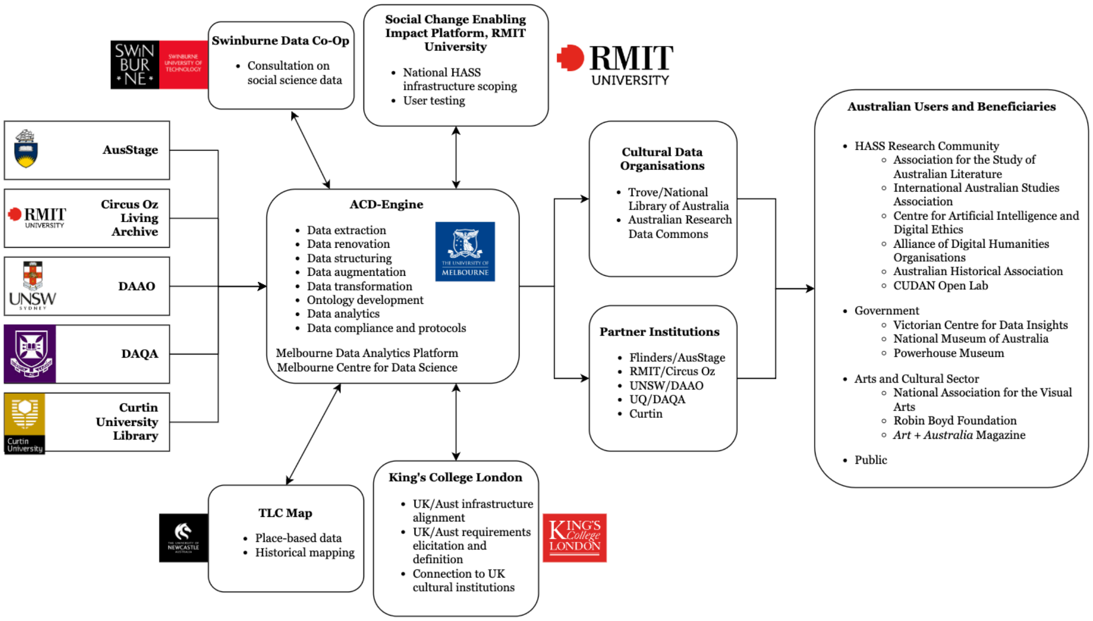

Swinburne Data Co-Op
Consultation on social science data
Social Change Enabling Impact Platform, RMIT University
National HASS infrastructure scoping
User testing
AusStage
Circus Oz Living Archive
DAAO
DAQA
Curtin University Library
ACD-Engine
Data extraction
Data renovation
Data structuring
Data augmentation
Data transformation
Ontology development
Data analytics
Data compliance and protocols
Melbourne Data Analytics Platform
Melbourne Centre for Data Science
Cultural Data Organisations
Trove/National Library of Australia
Australian Research Data Commons
Partner Institutions
Flinders/AusStage
RMIT/Circus Oz
UNSW/DAAO
UQ/DAQA
Curtin
Australian Users and Beneficiaries
HASS Research Community
Association for the Study of Australian Literature
International Australian Studies Association
Centre for Artificial Intelligence and Digital Ethics
Alliance of Digital Humanities Organisations
Australian Historical Association
CUDAN Open Lab
Government
Victorian Centre for Data Insights
National Museum of Australia
Powerhouse Museum
Arts and Cultural Sector
National Association for the Visual Arts
Robin Boyd Foundation
Art + Australia Magazine
Public
TLC Map
Place-based data
Historical mapping
King's College London
UK/Aust infrastructure alignment
UK/Aust requirements elicitation and definition
Connection to UK cultural institutions

Figure 1. The ACD-Engine's institutional structure (October 2023). Our five partner databases (AusStage, the Circus Oz Living Archive, DAAO, DAQA, and the Curtin University Library) appear on the left-hand side, flanked by our technical and advisory partners, the Swinburne Data Co-Op, RMIT's Social Change Enabling Impact Platform, the University of Newcastle's Time-Layered Cultural Map, and the King's Digital Lab. The ACD-Engine core team "ingested" these databases, working in partnership with internal Melbourne platforms, before distributing the results to our Partner Institutions and national Cultural Data Organisations. In turn, this data would reach a range of Australian Users and Beneficiaries across the HASS Research Community, Government, Arts and Cultural Sector, and the public.

In terms of personnel, the ACD-Engine was comprised of a core team at the University of Melbourne and collaborators at institutions around Australia, summarised as follows: 20

Melbourne Team

The interdisciplinary Melbourne team, in consultation with project partners, undertook the day-to-day work of the ACD-Engine project as outlined in the project workflow below. The team comprised the following six part-time posts, totalling approximately three full-time positions. 21

- Two lead researchers (in ARC parlance, Chief Investigators or CIs; professorial level, expertise in digital humanities, theatre studies, and information systems)
- One cultural data research fellow (postdoctoral level, expertise in digital humanities and literary studies)
- One project manager/cultural data research fellow (postdoctoral level, expertise in digital humanities and historical studies)
- One data scientist and visualisation specialist (PhD/postdoctoral level)
- One data engineer (graduate student level)

In additional, we benefited from the expertise of students affiliated to, but not funded by the project: 22

- One doctoral student in open linked data (graduate student level)
- Three short-term student interns (undergraduate/graduate student level)

Project Partners (Cultural Data)

Cultural researchers from our project partners were involved in the ACD-Engine's reflexive exploration, transformation, and analysis of the cultural data extracted from their databases. Each chief investigator was tasked to the project at either half a day or one day per week, with research assistants on short-term part-time contracts. They provided five primary sources of cultural data across five different artistic fields, many of which developed from ARC investment in 23 research and/or database construction, as follows:

- AusStage: The Australian Live Performance Database (https://www.ausstage.edu.au/). Funded by seven ARC LIEF grants, AusStage centres on events of live performance (more than 133,000 at the time of writing), linked to the contributors, works, venues, and organisations involved. AusStage is now one of the largest databases of performing arts in the world. AusStage is headquartered at Flinders University, which provided one chief investigator to the ACD-Engine project as well as local support staff.
- Circus Oz Living Archive (https://circusozlivingarchive.com/). Funded initially by an ARC LIEF grant, it houses archival footage from decades of Circus Oz performances, densely tagged to delineate acts, skills, and performers. Each event is also indexed through AusStage. The Circus Oz Living Archive originated at RMIT, which provided one chief investigator.
- Digital Archive of Queensland Architecture (DAQA) (https://qldarch.net/). Developing from an ARC Discovery Project on modernist architecture, this database centres on oral histories, connected with pages for architects, architectural firms, building projects, and published articles. DAQA is headquartered at the University of Queensland, which provided two chief investigators and hosted one research assistant.
- Design and Art Australia Online (DAAO) (https://www.daao.org.au/). Initially designed as a digital repository for several print-based biographical dictionaries of Australian artists, the DAAO has been funded by several ARC LIEF grants. The database centres on more than 17,000 artist biographies,[5] with associated artistic works, events, recognitions, art collections, and groups such as awarding bodies, galleries, and artist collectives. DAAO is hosted by the University of New South Wales, which provided two chief investigators and hosted one research assistant.
- The Summerhayes Collection at the Curtin University Library (https://www.curtin.edu.au/library/collections/special-collections/architecture/). This archival collection includes mainly architectural drawings from the West Australian Summerhayes architectural firm, described with library metadata. Curtin University provided two chief investigators and hosted one research assistant.

Each database presented distinctive opportunities to interrogate prior assumptions about data entry, coding protocols, and relationality. Taken together, they provided a manageable volume of data for an integrated information architecture to emerge. This under-the-hood approach, secured by the generous commitment of partners to the exploration of their existing databases, facilitated an iterative knowledge-sharing approach between data engineers, subject domain experts, and digital humanities scholars. 24

Project Partners (Analysis and Policy) and Advisory Board

Our partner institutions also included teams at Swinburne University's Data Co-Op (one chief investigator), specialising in government data analysis, and the Time-Layered Cultural Map of Australia (TLC Map) (https://tlcmap.org/, see [Arthur et al. 2020], one chief investigator), which provided a platform for geospatial analysis and detailed interrogation of demographic contexts. Researchers from the Social Change Impact Enabling Platform at RMIT University and the King's Digital Lab and Department of Digital Humanities at King's College London (one chief investigator each, along with supporting local staff), provided understanding of and alignment with national and global infrastructure priorities, as well as connections to user groups in Australia and the UK. The ACD-Engine was also supported by an international advisory board comprised of researchers and infrastructure specialists from the University of Melbourne, Flinders University, the National University of Singapore, the University of Ghent, and the Victorian government's Centre for Data Insights, which met approximately three times a year across the project to provide strategic advice and oversight. 25

Program A.3: Open Data Sharing

With its emphasis on the active construction and investigation of a shared cultural data framework or model, this project sits alongside major initiatives that support open data in research. Before beginning technical work, we needed to agree on how to share data between the project partners and the ACD-Engine central team as well as how to attribute authorship to individual and co-authored datasets and data-informed publications. A central concern was achieving a balance between energising new research by working across multiple data sources and making raw data available for 26 wider users, while also honouring the intellectual labour embedded in and the ongoing development of each partner database. Each partner database was built on unique data that had been collected by domain experts and, as such, carried within it their labour, both conceptual and quotidian (who among us has not dedicated extraordinary amounts of time to transcribing, annotating, and generating metadata!). Conversely, since each database was at least partially funded by national research agencies and often supplemented by volunteer contributors, the data was in a sense already public and designed to be shared with arts and humanities researchers as well as the public at large. Therefore, much like other data aggregation and interdisciplinary projects, the ACD-Engine faced both the potentials and pitfalls of sharing data [Smithies, Millar, and Thomson 2015] [Millar et al. 2018].

27 With these concerns in mind, we sought to develop a data-sharing agreement that outlined a pipeline for recognition of contributors to ACD-Engine datasets and collaborative papers, as well as a model for citation of data sources, that would adhere to recently released recommendations from the Australian Research Council for authorship of collaborative publications. After surveying the existing literature and participating in an Australian Research Data Commons laboratory on this topic, it became clear that most projects share general principles, but there was no well-established data sharing policy standard that would meet our needs. Our final data sharing policy, one component of which is represented in the visual summary below, has enabled us to rapidly determine responsibilities for work outputs and to appropriately assign authorship and acknowledge additional contributors.

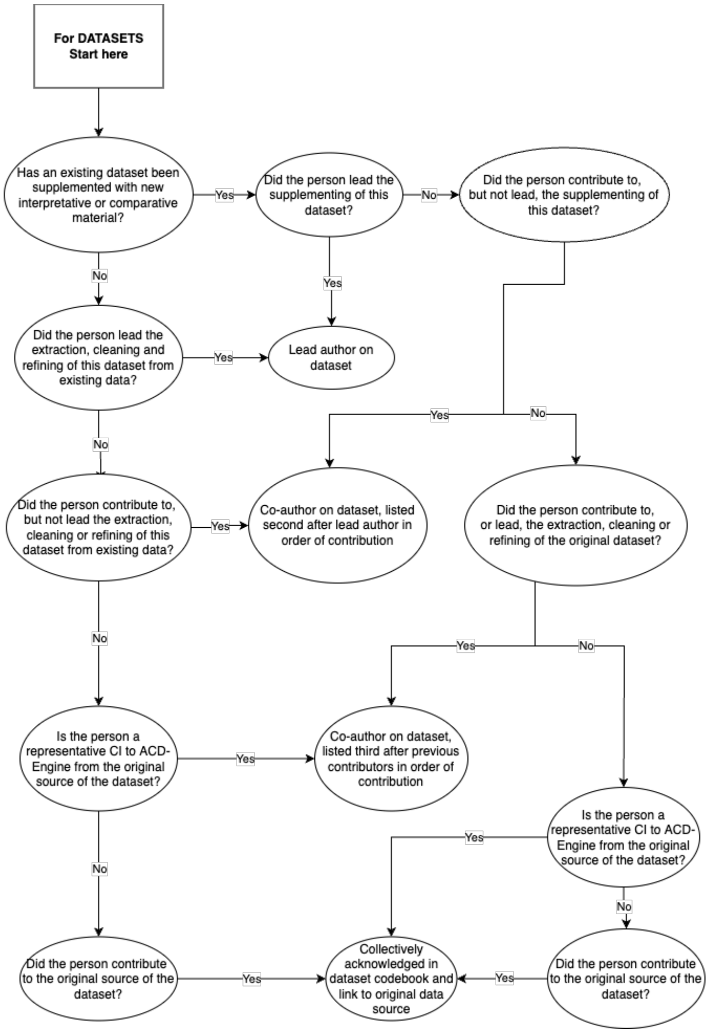

For DATASETS Start here
Has an existing dataset been supplemented with new interpretative or comparative material?
Yes
Did the person lead the supplementing of this dataset?
No
Did the person contribute to, but not lead, the supplementing of this dataset?
No
Yes
Did the person lead the extraction, cleaning and refining of this dataset from existing data?
Yes
Lead author on dataset
Yes
No
No
Did the person contribute to, but not lead the extraction, cleaning or refining of this dataset from existing data?
Yes
Co-author on dataset, listed second after lead author in order of contribution
Did the person contribute to, or lead, the extraction, cleaning or refining of the original dataset?
No
Yes
No
Is the person a representative CI to ACD-Engine from the original source of the dataset?
Yes
Co-author on dataset, listed third after previous contributors in order of contribution
Is the person a representative CI to ACD-Engine from the original source of the dataset?
Yes
No
No
Did the person contribute to the original source of the dataset?
Yes
Collectively acknowledged in dataset codebook and link to original data source
Yes
Did the person contribute to the original source of the dataset?

Figure 2. Extract from ACD-Engine Data Sharing Agreement Diagram pertaining to datasets.

28 Following these paths in an open data landscape, the work of different kinds of contributors is made explicit, including work by volunteers who undertake substantial amounts of data entry, software engineers who clean and sort data into usable subsets, and those who provide domain expertise in the organisation of a data output. This method of distributed

attribution for the collective authors of an open data artefact will inevitably shape the national arts and humanities research environment, facilitating greater recognition of the intellectual work undertaken at different stages of cultural data projects by diverse contributors (for the full data sharing agreement, see https://www.acd-engine.org/policy-ethics-governance).

Program B: Technical Design and Methods

To combine and harmonize data from multiple sources, formats, and structures into a single, unified framework, the ACD-Engine was informed by the long-standing technical data warehousing approach pursued in many fields of the sciences (see https://www.ibm.com/topics/etl). For the humanities in Australia, this approach is relatively nascent, with the possible exception of PARADISEC (https://www.paradisec.org.au/), a consortium of language collections with a platform that warehouses a wide range of data inputs and outputs and provides tools to analyse across their content. 29

Work to enable the interoperability of digital collections in the arts and humanities is long-standing, initially based in the domain of digital libraries. The OAI-PMH protocol, one of the most widely deployed technologies for digital library interoperability aimed to provide a very simple method for harvesting data from primarily digital libraries, but it has several key shortcomings [van de Sompel et al. 2004]. As Carl Lagoze points out, OAI-PMH is “based on the premise that discovery occurs within the boundaries of the digital library”, thus “relying on a technology distinct from that in the mainstream web architecture”. OAI-PMH also presupposes that “structured metadata” are “necessary and fundamental to a digital library environment, and that the Dublin Core vocabulary is the key to ‘semantic’ interoperability” [Lagoze 2010, 216]. The tangible effect of these limitations is the prioritization of simplicity and efficiency at the cost of functionality and robustness, as well as limited ability to implement feedback mechanisms to correct unusable or poor-quality data. 30

Two core aims of the ACD-Engine were to improve the quality of cultural data and better support the interoperability of cultural data across new disciplinary bridges and in new contexts outside the traditional institutional repository domain. Our technique diverges from OAI-PMH in several ways. First, our approach is dynamic real-time access, working in real time rather than on harvested, static data, using a method similar to the more sophisticated Z39.50 protocol [Lynch 1997]. Second, the access module for each collection normalises the format of data items, which again OAI-PMH does not require but Z39.50 does. Neither difference, however, addresses the problems of data quality that pervade many collections accessible via OAI-PMH [Warwick et al. 2009]. The ACD-Engine's current access modules can be set to discount incomplete (meta-)data, and we anticipate that future iterations of the ACD-Engine platform will provide feedback mechanisms for data quality, in line with emerging best practice in digital library infrastructures. 31

Our approach to data management, while not unique, is pioneering in its application to Australian artistic and cultural data. Furthermore, our approach provides an exemplar for undertaking future projects in this field: 32

1. **Extracting** encoded digital data from a range of separate, heterogeneous sources using APIs and various forms of web scraping in collaboration with project partners;
2. **Identifying** and **resolving** inconsistencies, redundancies, errors, and gaps in the data, specifically where a datapoint definition (attribute) has specific meanings in one dataset that are incompatible with a like unit in another dataset, or where significant limitations exist in the coverage of available data;
3. Structuring the data into a uniform architecture adapting recognised metadata schema relevant to the arts and humanities without flattening disciplinary distinctiveness;
4. **Augmenting**, **expanding**, and **enriching** the data through a variety of methods including manual data entry (conducted by project partners and research assistants), data warehousing, and ETL (extract, transform, load) tools; and
5. Transforming the data into machine-readable datasets that can be easily understood and deployed by end users.

We spent much of the first year of the project intensively investigating the contents and structures of our partner databases to understand how their distinctive data attributes could be usefully extracted, processed, and enhanced by an integrated model. Working across multiple, heterogeneous data sources, we faced the challenges of bespoke and 33 discipline-specific metadata, varying levels of accessibility, and distinct knowledge transfer capabilities, including custom-made and obsolete websites, APIs, and database backend structures. We tackled these problems, to use Tim Sherratt's terminology, by “digging down through layers of technology, descriptive practice, and institutional history, to understand what is delivered so conveniently through our browsers” [Sherratt 2019]. Respectful questioning of the relative weight to be given to a specific datapoint, or how to identify the relationships between diverse concepts across multiple artistic disciplines, required close interrogation of the databases as knowledge domains that have been shaped by the affordances of their specific information architecture, as well as detailed consultation with discipline experts. Only through this exhaustive, iterative process, scrutinising data science methods in their application to artistic research and vice versa, could we develop the depth of understanding necessary to build a robust and intellectually sophisticated integrative architecture. Seeking to guide our diverse team through this process of investigation and consultation, we gradually developed an easy-to-understand data integration workflow diagram (Figure 3).

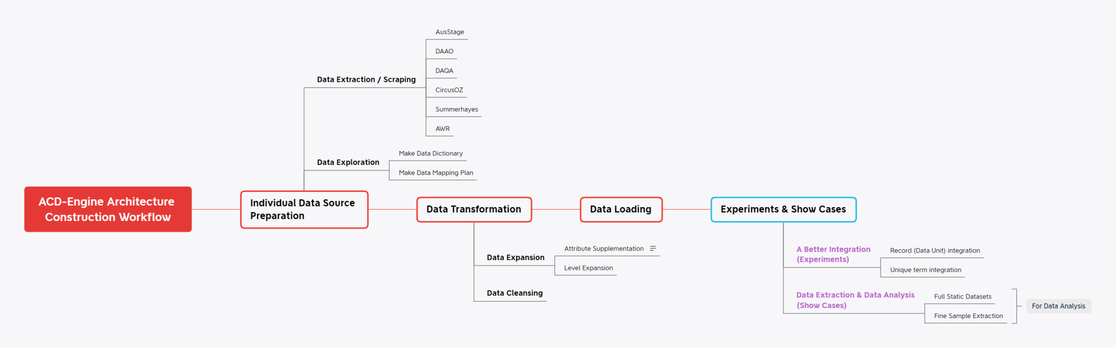


Figure 3. ACD-Engine Architecture Construction Workflow.

As the project progressed, we added and trained skilled software engineers and data scientists to the team and hosted several research students and interns who each helped to refine the technical competency of the model. Some members of the team investigated existing data ontologies for Open-Linked Data to align the data attributes selected for the ACD-Engine with best practice models in the arts and humanities [Liu, McKay, and Buchanan 2023] [Sumner et al. 2023]. Others designed the components of the ACD-Engine's architecture for cultural data (ACDEA) featuring core entities and key attributes [Zheng and Munoz 2023c]. The Data Transformation stage involved creating an aggregated data dictionary that explained the attributes and values for each data source, as well as a data mapping plan that outlined how the information from different sources combined into the ACDEA.[6] 34

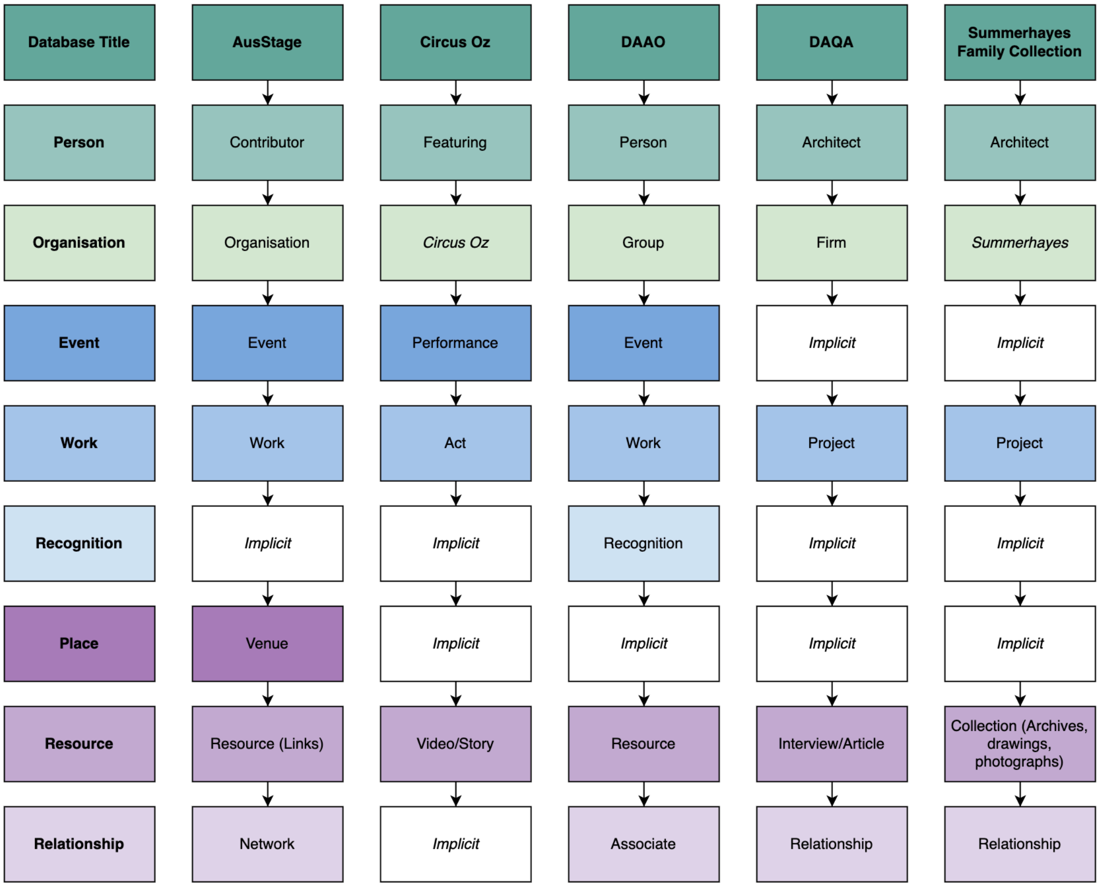
Having extracted data from our partner databases, the ACD-Engine team then enriched selected subsets of the data. First, we asked the discipline experts on our research team to enter new data for selected key attributes that had not been captured in the original databases but would facilitate more robust analyses across the ACD-Engine unified datasets. In some cases, as much as 90% of a single attribute was missing from a dataset. For instance, while career commencement dates and birth dates were commonly available in the “People” entity, very few career end dates or death dates were listed. 35

We also cleaned the data by identifying and correcting errors and inconsistencies. For example, due to the nature of integrating data from multiple sources, similar records with minor variations often appeared in multiple places. It was crucial to identify and handle duplicated records appropriately to avoid skewing analysis results. We used artificial intelligence and machine learning to accomplish these tasks, but for the most part undertook deduplication with a mix of simple automation and manual correction. Future development of AI in this space would be fruitful, particularly when working with larger volumes of data. Through this process, our data science team shared insights with our data partners and experts, who engaged in a generative back-and-forth process around definitions, gaps, opportunities, and limitations. 36

Program C: ACD-Engine Architecture

Qualitative data modelling is now indispensable to the creation of federated digital humanities databases [Rankin, Gordon, and Potter 2009]. As we have shown, building the ACDEA required a close interrogation of the data models and entities (or, in non-digital language, cultural concepts) that shaped the different cultural databases. A key finding was the extent to which each database was structured around a meaningful yet discrete entity[7]: AusStage around events, DAAO around artist biographies, DAQA around oral history interviews, and so on. Each database contained a wealth of front-facing data on people and organisations — indeed, both DAAO and AusStage contribute to the “People and Organisations” sub-platform in Trove, which also draws in other collections such as the Australian Women's Register and the Encyclopaedia of Australian Science — but the level of information and detail among these entities was markedly different across each database. We also found, on the other hand, that the public-facing structure of each database to a degree obscured the richness of data entities and relationships that existed embedded in their systems. DAQA, for instance, has front-end browse capacities for “Architects” (people), “Firms” (organisations) and “Projects” (works), but also contained substantial information that aligned with the events and places that appeared in the front ends of DAAO and AusStage. 37

A key aim for the ACDEA, therefore, was to ensure the people, organisations, works, and events entities across the databases were interoperable, which necessitated a fixed, consistent set of minimum attributes. For the person entity, for example, which invites biographical analysis, we recommend a minimum of first and last name, as well as date and place of birth and death. Inevitably, as mentioned above, many of the people listed in our partner databases did not have all of these basic attributes; as appropriate, we supplemented selected entries where new information was available (the “top 500” or “top 100”), and/or based our analyses on the most complete entries [Cutter, Fensham, and Sumner 2023]. Similarly, the interrogation of arts organisations and their histories required the collection of changing organisational names, establishment dates, and locations — again, assembling a baseline of accurate data points was essential for comparative extraction and analysis. Each entity inevitably generated a range of connections to other entities, which required their own set of attributes. Ultimately, we established standardised definitions and attributes for eight entities: Person, Organisation, Work, Recognition, Place, Resource, and Relationship. Each unified ACDEA entity corresponded to an entry atributed to a somewhat different name in our source datasets. 38



Database Title	AusStage	Circus Oz	DAAO	DAQA	Summerhayes Family Collection
Person	Contributor	Featuring	Person	Architect	Architect
Organisation	Organisation	*Circus Oz*	Group	Firm	*Summerhayes*
Event	Event	Performance	Event	*Implicit*	*Implicit*
Work	Work	Act	Work	Project	Project
Recognition	*Implicit*	*Implicit*	Recognition	*Implicit*	*Implicit*
Place	Venue	*Implicit*	*Implicit*	*Implicit*	*Implicit*
Resource	Resource (Links)	Video/Story	Resource	Interview/Article	Collection (Archives, drawings, photographs)
Relationship	Network	*Implicit*	Associate	Relationship	Relationship

Figure 4. A simplified visualisation of the ACDEA, showing how differing entity titles are translated into our harmonised data structure. White rectangles indicate that the database does not explicitly recognise the entity in question, but that information on this entity is nonetheless available within. Circus Oz and Summerhayes are also italicised, being the only organisations in their respective datasets.

ACDEA Construction Principles

We constructed the ACDEA based on the following principles [Liu, McKay, and Buchanan 2023]: 39

- **Refer to the Most Comprehensive Existing Data Models**: For each entity's basic metadata requirements, we looked to the database with most comprehensive orientation towards that entity. For example, to construct its own “Person” entity, ACDEA refers to the “person” data model from DAAO, which is a person-oriented data collection, whereas AusStage has a comprehensive approach to “Event” entity collection.
- **Keep Differences & Unify Commonalities**: Seeking then to integrate the central entity model with similar entities in the other datasets, ACDEA follows the principle of keeping differences between the data models that are significant and add value to the data, while integrating similar data attributes and unifying common elements to ensure consistency and usability. This approach helps to capture the unique characteristics of each data source and ensure that ACDEA includes all relevant information.
- **Fuzzy Mapping**: This method involved mapping data attributes from each data source to a common attribute in the unified data model, even if they were not an exact match.[8] For instance, the DAAO has an attribute called “other occupations” and AusStage has an attribute called “(contributor) functions”. These attributes can both be mapped to the “person” entity in ACDEA, and a common attribute called “career”. Reading across these multiple elements of career allows ACDEA to capture as much information as possible while still maintaining consistency and usability.

While remaining open to the inclusion of diverse or novel attributes, this model provided us with a view of the range and 40 complexity of data from cultural databases across multiple disciplines, without diminishing the significance of each entity in disciplinary histories. Each of the terms we use for these entities, we argue, has a firm basis in existing cultural research methodologies and an intuitive, difficult-to-dispute accessibility. We arrived at these terms following extensive debates about the meaning of various concepts (for the results of one discussion, see [Cutter, Fensham, and Sumner 2023]). Simultaneously, this data ontology facilitates the potential future integration of datasets from other sources or search engines of specific interest to a research project, whether they might be from urban design, music, or the art market. Each entity in ACDEA contains three essential identifying metadata:

- Record data, including the original (database) ID, URL, associated entity name, timestamps, and other important details. This metadata is crucial in maintaining the integrity and traceability of the data back to its original source.
- Content details of the record type, provided as description.
- Related records which refer to the one-dimensional connections to other records within the same entity. This feature provides easier access to the relationship network of a particular record.

ACDEA Entity Definitions

To facilitate use of ACDEA, each entity has been described in a data dictionary [Zheng and Munoz 2023b] and detailed visualisation [Zheng and Munoz 2023c][9] with a simplified version provided in the table below:

Entities	Attributes	Description
Person, representing individuals who are relevant to cultural data	Names	All versions of the individual's name, including display name, primary name, and alternative names
	Summary	The biography/descriptive summary
	Gender	Male, female, non-binary, unknown
	Birth	The date and place of birth
	Death	The date and place of death
	Roles	The roles an individual has held for an extended period of time, as defined by entry records
	Career	Features of an individual's career (e.g., persons and organisations worked for/with, career start/end date)
	Residences	The places an individual has resided and the time period of their residence
	Nationality	Identifying markers of nation
	Languages	Identifying language groups
	Related entities	To link to all other entities in the architecture
	National Library Archive ID	The ID assigned by the National Library of Australia to an individual
Organization, representing organizations that are relevant to the cultural data	Names	All versions of the organization's name, including display name, primary name, and alternative names.
	Description	The biography/descriptive summary

	Types	The type of the organization
	Long Term Roles	The roles that the organization has held for an extended period of time
	Locations	The places where the organisation has been located
	Operation	The features of the operation of the organisation, including periods and coverage
Event, representing cultural events	Title	
	Description	The brief description of the event
	Types	the type/genre of the events
	Time & Place	The specific date and location where the event took place
Work, representing cultural works such as books, architecture, performance, art, etc.	Titles	All titles of the work, including the primary title and alternative titles
	Summary	The summary of the work
	Medium	Unique to visual art and design
	Time & Place	The specific date and location where the work was created
Recognition, representing awards or recognition received by individuals or organizations	Title	Title of the award
	Summary	Descriptive summary of the award or recognition
	Types	The types of the recognition
	Time & Place	The specific date and location where the recognition was granted
Place, representing the geographic locations or facilities where the cultural events take place	Names	All names of the location, including the primary name and alternative names
	Address	The standardized physical address of the location, including country, state, suburb, street, and postcode
	Geo-Coordinates	The geographic coordinates (latitude and longitude) of the locations
	Type	The type of venue or building situated in a location
	Start & End Date (Optional)	The start and end date of a location, which is optional information for tracking the changes of a location with the type “venue” over time
Resource, representing resources that are relevant to the cultural data, such as books, images, paintings, oral histories, video and audio files. The description details refer to the metadata schema used in the	Titles	All titles of the resource, including the primary title and alternative titles

National Library of Australia and are more typically bibliographic		
	Description	Descriptive summary of the resource
	Type	The type or genre of the resource
	Authoring	The author or creator of the resource
	Source	The source of the resource
	Date	The associated dates of the resource, such as created date, published date, etc.
	Acquisition	The method or process used to acquire the resource
	Right	The rights associated with the resource, such as copyright or licensing information
	Format	The format information of the resource
	Identifier	Rhe identifiers associated with the resource
Relationship, representing the relationships between the various entities within the cultural data. The semantic structure plays a crucial role in establishing connections and dependencies between the other entities	Subject	The entity that initiates the relationship
	Object	The entity that is related to the subject
	Predicate	The type of relationship that exists between the subject and the object
	Time	The time period during which the relationship between the subject and the object existed

Table 1.

42 Some of the attributes listed above represent discrete attributes, while others (such as “career” under “person”) provide an entry point to further nested data attributes. In addition, the bottom-up construction of the ACDEA ensures that all relevant attributes from the originating data sources are included (even though some attributes do not map across all datasets) without losing sight of larger questions of scale, sustainability, and interoperability with other national and international collections.

Program D: Data Enrichment and Project Outputs

43 Developing from our work in data exploration and enrichment, the ACD-Engine set out to interrogate and test the interoperability of a range of cultural databases for use of cultural data in new analytic contexts. These included but were not limited to scholarly research, data analytics and public interest, statistical experimentation, and national infrastructure strategy and policy.

44 In order to demonstrate the value of data analytics and speak to live debates in cultural scholarship that would engage researchers, we began our analytical program by investigating four themes: Working in the Arts: Women, Careers, and Creativity; Cultural Hotspots and Local Scenes; Cultural Diffusion and Arts Diplomacy; and Environmental Politics and

the Arts. Each theme aligned with the high-level entities of our information architecture for the purposes of data extraction and analysis: for instance, questions about gender and careers were to be answered through interrogation of the “People” entity of the ACDEA. The first two themes in particular became fruitful sites for data experimentation, guided by specific research questions raised by our domain experts. The core team of data scientists transformed data throughout this process by producing data “experiments” that examined the robustness of specific data attributes within each dataset and showcasing them to research partners, who provided feedback and drove further investigation.

Key Project Outputs

The results are showcased in a number of tangible, open-access outputs, including but not limited to: 45

- The ACD-Engine Cultural Data Workbook. This public-facing Jupyter notebook [Zheng and Munoz 2023a], freely available through GitHub, provides new users in research, industry, and government with the framework to re-examine our data outputs and to identify further information required to use ACD-Engine data effectively. It also showcases a range of bespoke analytics on ACD-Engine datasets (https://acd-engine.github.io/jupyterbook/about.html), a prototype geotemporal map of our datasets (https://acd-engine.github.io/jupyterbook/Analysis_ACDE_Map.html), intersections with IMDB and AustLit (https://acd-engine.github.io/jupyterbook/Analysis_ExternalDatasets.html), and an exploration of Melbourne's evolving music scenes using setlist.fm, Discogs, and Spotify data (https://acd-engine.github.io/jupyterbook/Analysis_TLC_Workshop.html).
- Seven ACD-Engine Unified Datasets (Event, Organisation, Person, Place, Recognition, Resource, Work) using data from AusStage, Circus Oz, DAAO, DAQA, and the Summerhayes Collection) (https://www.acd-engine.org/datasets).
- Six Data Dictionaries (ACD-Engine Unified, AusStage, Circus Oz, DAAO, DAQA, and Summerhayes Collection) explaining the data structures and column meanings for all metadata fields across each data source (https://www.acd-engine.org/datasets).
- “ACD-Engine Data Entry Guidelines”, a basic guide to effectively and consistently entering data into cultural databases (https://www.acd-engine.org/policy-ethics-governance).
- “Mapping Cultural Data: The Basics”, a simple, easy-to-use guide to mapping cultural data (https://www.acd-engine.org/resources).
- “Australian Cultural Data in Trove, VIAF and Wikidata”, a prototype tool developed in collaboration with historian and hacker Tim Sherratt, which highlights the extensive and largely unexplored connections between Australian cultural databases and global data aggregators (see https://github.com/wragge/acd-engine and https://acde-links-yajhxrvxsa-ts.a.run.app/acde-links).

Partner-Specific Outputs and New Research Directions

In the process, we undertook investigations into each partner database, leading to a number of new insights and directions for future research. Each data partner explains some of their key findings below: 46

- DAAO/University of New South Wales
    - Interfacing with data scientists around the material captured by DAAO included undertaking new research to supplement information for 500 top artists (“DAAO 500”; see https://acd-engine.github.io/jupyterbook/Analysis_DAAO500.html) and helped us understand the DAAO's strengths and limitations. One application has been a statistical exploration of gendered name changes and their impact on historical visibility of women and non-binary artists [Sumner et al. 2023] [Cutter, Fensham, and Sumner 2023].
    - An analysis of data on visual arts events captured in DAAO enabled key features and trends in Australian cultural institutions to be identified quantitatively (e.g., the boom in cultural institutions, such as experimental art spaces, in the late 1980s) and analysed against known historic trends.
    - Looking across all the ACD-Engine's datasets helped us understand the interconnections (and

    duplications) of practitioners across different fields, offering insight into the well-known but little-analysed "portfolio career" in the arts. For example, Elaine Haxton, Margaret Olley, Sidney Nolan, Frank Hinder, and George Gittoes all worked in theatre (mainly stage design) as well as the visual arts, while Sally Morgan is both a visual artist and a writer. We intend to investigate these connections, as well as the "other occupations" field in DAAO, to consider new research questions. For example, what is the cultural impact of artists/actors/performers working as schoolteachers?

- AusStage/Flinders University and Circus Oz Living Archive/RMIT University
    - Collaborating on the "Data Dictionary" allowed the AusStage team to reconsider its data ontology, in particular redefining the category of "Event" to standardise its use and to recognise that a single performance at one venue is different from a season. Factoring in duration and significance allows for future interoperability with other databases that hold "Event" data.
    - Tracing gendered career pathways through the dataset (see https://acd-engine.github.io/jupyterbook/Analysis_AusStage.html) revealed data transcription errors in the gender field of "Contributor" (in ACDEA, "Person") records, compounded by a default to "Male" when adding a new "Contributor". The AusStage team developed a new gender policy, including a new "Non-Binary" category that can be used in defined circumstances and a gender field that defaults to "Unknown".
    - We prepared an "AusStage 250" (a dataset of 250 top performing arts practitioners analogous to the "DAAO 500"), which was enriched with basic biographical data, including dates and places of birth and death, as well as person attributes not ordinarily captured in the AusStage database. This process focussed attention on how community data entry shapes what is recorded in cultural databases. Some of the best documented "Contributors" were recognised not because of their general significance, but because particular AusStage authors were interested in supplementing their entries.
    - We also undertook a focused investigation of the Circus Living Archive and enriched AusStage data relating to Circus Oz, some results of which can be found here: https://acd-engine.github.io/jupyterbook/Analysis_CircusOz.html.
- DAQA/University of Queensland
    - While recording the achievements of women architects was an aim of the original DAQA research project, the original database lacked gender metadata. By adding this attribute, a key facet of "People" profiles across the other ACD-Engine datasets, we were able to investigate the proportions of women practicing architecture in Queensland, which firms employed women, and when and where historical shifts in participation and employment occurred. This also allowed us to consider the relationship to gender balance in the visual arts and design through comparisons with DAAO.
    - ACDEA enabled uses of DAQA that had not been envisaged in its design (see https://acd-engine.github.io/jupyterbook/Analysis_DAQA_Part1.html and https://acd-engine.github.io/jupyterbook/Analysis_DAQA_Part2.html). While DAQA recorded the names of architectural firms (in ACDEA, "Organisations") and their predecessors, the ACD-Engine's work enabled us to visually model and analyse the multiple progenitors and successors of key firms. By examining the top five longest lineages of firms in Queensland we gained a new understanding of the social and commercial structures of architectural cultures throughout the twentieth century.
- Summerhayes Collection/Curtin University
    - The team at Curtin generated a 360-degree panorama tour of the Subiaco Hotel in Perth, Western Australia, in order to investigate the potential of open linked data and panoramic tours to enhance the localised and contextual comprehension, virtual exploration, and spatial perception of architectural heritage locations. The Subiaco Hotel was chosen because the

    - Summerhayes architecture collection documents its architectural history and repeated renovation by the Summerhayes family firm, and because it is linked with other architectural datasets that were being integrated into the ACDEA.
    - This project revealed that when flexible, robust, and relational metadata is produced, exploration of datasets across national collections becomes feasible. By creating linked open datasets of varying media, the 360-panorama can enrich cultural heritage datasets. This allows for researchers as well as visitors to experience immersive, content-rich panoramic tours [Rahaman, Champion, and McMeekin 2023].
- Time-Layered Cultural Map/University of Newcastle
    - Funding from the ACD-Engine allowed us to build a new function in TLCMap to combine map layers in a single map while preserving the structure of the components. With this new capability, TLCMap users can create a “Multilayer” and see how patterns in a new layer fit with patterns in existing layers, or they can divide a new layer into multiple components to see how categories interact. One example of this functionality is a timeline map of Aboriginal Protection and Welfare Board sites in NSW overlaid with a historical map of railway stations, which reveals a correlation between the development of rail transport and the treatment of Aboriginal people in the late nineteenth and early twentieth centuries.
    - TLCMap also benefited from engagement with the ACD-Engine through a deeper understanding of data models. Individual “Person” entries were difficult to map on TLCMap since they require recognition of why place of birth or death — the most common mappable attributes — might be significant in artistic career paths, as a component of a larger cultural phenomenon with more self-evident geolocations. TLCMap now benefits from access to the ACD-Engine's unified data structure (ACDEA) and datasets, with unlimited potential for map layers to be created from the newly curated data, and also from the lived experience of structuring rich but loose cultural and historical materials on secure foundations and sound principles.

As this showcase highlights, with the capability of the ACD-Engine's robust datasets and outputs now readily available, members of the project team have begun to explore a range of novel use-cases for data analysis, particularly where data interrogation might be placed alongside scholarly and topical debates, or interfaced with other smaller datasets that invite new understandings of digital cultural heritage. Further applications might include: 47

- Open linked data annotation of digital models (e.g., reconciling entities with national and international data repositories such as Trove, Wikidata, International Standard Name Identifier, or the Virtual International Authority File[10])
- Bulk data fetching across different data sources
- Multilayer spatial visualisation
- Text-mining functionalities
- Multilayer network analysis
- Extension of data queries to national and international collections

We further anticipate that the datasets will now be used in data analytics pedagogy in the digital humanities, information science, and digital collections. Data scientists and graduate researchers could also experiment with the data coding and models in their own research. 48

Key Project Innovations

In summary, the ACD-Engine has made a distinctive and timely contribution to a complex and evolving national data ecosystem. In the two years designated for the project, a diverse range of intellectual, technical, and infrastructural advancements were made which addressed our original project aims. These include but are not limited to: 49

Aim 1: Improve the quality of existing cultural data in Australia across a range of partner databases and cultural disciplines. 50

- Undertook unprecedented and comprehensive investigation of databases: collecting histories, shortcomings, funding histories, and workforce constraints;
- Definitively established the coding logics, disciplinarity distinctives, and sensitivities and vulnerabilities of specialist collections as well as their software capabilities;
- Mapped and synthesised data ontologies critical to specific categories of cultural data across a range of fields, including history, cultural studies, literature, and art and design history; and
- Undertook systematic analyses of core categories of cultural data and documentation, identified the challenges for interpretation of cultural data, such as changing social norms (e.g., naming conventions, gender differences, career pathways, and organisational locations), and promoted solutions for localised aggregation methods.

Aim 2: Support an expansion in the accessibility, use, and interoperability of cultural data in research, industry, and government contexts. 51

- Built a sophisticated and integrated Research Software Engineering (RSE) team for wrangling, structuring, and interrogating large scale datasets;
- Produced an interoperable data model (ACDEA) for linking and interpreting arts and cultural data between different systems and platforms;
- Tested the suitability of ACDEA for interoperability with international cyber-infrastructures;
- Delivered robust, cultural datasets suitable for re-use and further testing with a range of research tools and platforms; and
- Produced dynamic visualisations and critical case studies exploring key research questions in arts and culture.

Aim 3: Demonstrate how new insights into arts and cultural production can emerge by improving cultural data interoperability across platforms, systems, and regions. 52

- Developed a novel policy framework for data-sharing, in line with open access principles, for the attribution of data-led research in the humanities; and
- Promoted innovative and multidisciplinary ways of working across cultural collections, including topical analysis featured in public media outlets.[11]

These ACD-Engine outcomes lay significant groundwork for the future development of a national data architecture for Australian cultural collections. The prototype information architecture, knowledge base, and RSE capabilities developed by the ACD-Engine pave the way for future work on Australian cultural data, including record unification (harmonising duplicated records, predicates in relationships, and place records across multiple data sources), identifying more implicit data, integrating Australian cultural data with more data sources, and consideration of persistent identifiers for “Person” records across national collections. 53

Discussion and Conclusion

National cultural collections continue to grow in significance as investment at government and inter-governmental levels extends expectations of greater utilisation of public resources. Greater integration should not, however, diminish the important role that distinctive databases and curated collections have to their source communities and disciplinary experts. Given the volume and complexity of cultural data in national collections, dissolving barriers between one collection and another, as this project has identified, requires careful understanding of how cultural data might be organised beyond each collection. 54

If the goal of national collections involves expanding and enhancing (trans)-national data resources, we would assert that more priority should be given to developing robust information architectures. As we have found, cultural data 55 sharing and interoperability is rich terrain for the digital humanities and will be more so as new cyber-infrastructures, including artificial intelligence, transform the ways we engage with, reflect upon, and utilise cultural data. If we are to grow research impact and public engagement, national collection projects ought to strenuously pursue best-practice data-sharing approaches, particularly responsible attribution, documentation, and data contextualisation by collection owners [Loukissas 2019] [Lee 2023] [Alkemade et al. 2023]. Without so doing, we risk developing an array of increasingly disparate data structures and infrastructures that will inhibit both disciplinary and interdisciplinary research. The intensive (and unavoidably expensive) investment of human labour and expertise is critical in this process to safeguard database quality and disciplinary distinctiveness against the oversimplification too often wrought by external metadata structures, especially as they are increasingly implemented by machines.

In this context, the challenges we face for future research in this space are principally structural. Research infrastructure for the Australian arts and humanities is maturing, but national and university-based funding remains largely tied to short-term projects like the ACD-Engine, rather than medium- or long-term workforce and infrastructure development. As longer term strategies slowly emerge, researchers (particularly those early in their careers) will continue to repeatedly enter and exit the field as they pursue employment and funding, making it difficult to build a consistent approach to infrastructure development for cultural data research. The ACD-Engine's exploratory work has already seeded new investment and opportunities for further research collaboration that recognises the central role of data curation and aggregation in cultural collections to further re-use, as well as knowledge exchange across industries, universities, and governments. 56

Considering the future use of the ACD-Engine within the Australian and international cultural data landscape, a number of possible scenarios present themselves. Though the ACDEA is based around text and metadata, the process through which it was developed could be fruitfully applied to other digitalised media (images, film, audio recordings, digital modelling). The ACD-Engine's federated model, in which nodes of disciplinary expertise and data collections interface with a central hub of data science, information science, and digital humanities experts, could facilitate ongoing partnerships across disciplines to further demonstrate the utility and analytical power of working across multiple collections in historical analysis of cultural production. Future funding applications might focus on the testing and application of software tools (artificial intelligence, machine learning, 3D/4D modelling, immersive VR) in the enhancement and analysis of Australian cultural data, or on new collaborative projects between researchers, industries, and governments to explore the quality of insights available from cultural data. At the same time, we hope that funders will invest in maintaining and improving Australia's ecosystem of cultural databases, many of which rest principally on the custodianship of individual researchers and the current goodwill of university- or community-based hosting providers. This project shows the potential of these databases, the necessity of human labour to extract and interpret it, and how much may be lost should these databases disappear. 57

In this critical moment where automation is rapidly pervading the arts and cultural space, we both welcome the utopian possibilities and caution against a dystopian erosion of the contexts, histories, and community values embedded in digital cultural heritage. The ACD-Engine represents a pragmatic, stress-tested, and feasible model for comparative data extraction and attribution towards future cultural and critical enquiry. 58

Notes

[1] A key innovator in this space is Tim Sherratt, whose GLAM Workbench (https://glam-workbench.net/) and Trove Data Guide (https://wragge.github.io/trove-data-guide/home.html, in-progress at the time of writing) have been transformative for researchers in this space.

[2] Completed in September 2023, the “Trove Enhancements” project managed by the Australian Research Data Commons improved the Trove pages for researchers and updated the public Trove API to provide better support for Australian HASS researchers (https://ardc.edu.au/project/trove-researcher-platform-for-advanced-research/). Trove development remains tied to fixed-term, outcome-focused projects, and ongoing funding is not assured.

[3] In addition to our partners, we have also engaged with a range of stakeholders in the sector including the Victorian Centre for Data Insights, the Powerhouse Museum, the National Association for the Visual Arts, the National Museum of Australia, the National Gallery of Australia, the Centre for Artificial Intelligence and Digital Ethics, the Robin Boyd Foundation, and *Art + Australia* magazine.

[4] These objectives align with those of the TaNC scheme in the UK and projects such as the British Library's “Living with Machines” project (https://www.bl.uk/projects/living-with-machines).

[5] For a full report on the data scope of each database, see https://acd-engine.github.io/jupyterbook/Integration_ACDEA_DataReport.html.

[6] For the full data map, see https://acd-engine.github.io/jupyterbook/Integration_ACDEA_Overview.html.

[7] Or, to use a humanities term, a disciplinary construct: a definable object, actor, or location around which a relational database can be constructed.

[8] For a different context of interest, see https://github.com/Living-with-machines/DeezyMatch.

[9] For a general discussion of data dictionaries, see [Data Dictionary, n.d.].

[10] For a proof-of-concept tool showcasing some potential connections between entities in our databases with Trove, Wikidata, and the Virtual International Authority File, built in collaboration with Tim Sherratt, see https://github.com/wragge/acd-engine.

[11] Several public-facing outputs of the Engine's work in this space include [Sumner and Fensham 2023] and [Sumner and Munoz 2023]. The ACD-Engine also contributed to a topical analysis of Australia's prestigious Archibald Prize in collaboration with the *Guardian* data journalism team [Nicholas and Touma 2023].

Works Cited

Ahnert et al. 2023 Ahnert, R. et al. (2023) *Collaborative historical research in the age of big data*. Cambridge, England: Cambridge University Press.

Alkemade et al. 2023 Alkemade, H. et al. (2023) “Datasheets for digital cultural heritage datasets”, *Zenodo*, 25 September. https://doi.org/10.5281/zenodo.8375034.

Arthur 2014 Arthur, P.L. (2014) “Biographical dictionaries in the digital era”, in Arthur, P.L. and Bode, K. (eds.) *Advancing digital humanities: Research, methods, theory*. New York: Palgrave Macmillan, pp. 83-92.

Arthur et al. 2020 Arthur, P.L. (2020) “Time-layered cultural map of Australia”, *Proceedings of the fifth annual digital humanities in the Nordic countries conference, 2020*. Riga, Latvia, 21-23 October. CEUR, pp. 184-191. Available at: https://hdl.handle.net/11541.2/143150.

Australian Government 2023 Australian Government Department of Infrastructure, Transport, Regional Development, Communications and the Arts (2023) *National cultural policy: Revive: A place for every story, a story for every place*. 9 February. Available at: https://www.arts.gov.au/publications/national-cultural-policy-revive-place-every-story-story-every-place.

Bettivia and Stainforth 2017 Bettivia, R.S. and Stainforth, E. (2017) “All and each: A socio-technical review of the Europeana project”, *Digital Humanities Quarterly*, 11(3).

Borgman 2007 Borgman, C. (2007) *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: The MIT Press.

Burrows 2011 Burrows, T. (2011) “Sharing humanities data for e-research: Conceptual and technical issues”, *Proceedings of the sustainable data from digital research conference, 2011*. Melbourne, Australia, 12-14 December. Available at: https://ses.library.usyd.edu.au/handle/2123/7938.

Burrows 2013 Burrows, T. (2013) “A data-centred ‘virtual laboratory’ for the humanities: Designing the Australian Networked Infrastructure (HuNI) service”, *Literary and Linguistic Computing*, 28(4), pp. 576-581. https://doi.org/10.1093/llc/fqt064.

Burrows and Verhoeven 2016 Burrows, T. and Verhoeven, D. (2016) “Aggregating data for social linking in the humanities and creative arts: The Humanities Networked Infrastructure (HuNI)”, *Proceedings of the ninth annual metadata and semantics research conference, 2015*. Manchester, England, 9-11 September. Springer Link, pp. 109-119. https://doi.org/10.1007/978-3-319-24129-6_36.

Cathro and Collier 2020 Cathro, W. and Collier, S. (2020) “‘Developing’ *Trove*: The policy and technical challenges”, *eLucidate*, 7(3), pp. 3-14. Available at: https://www.vala.org.au/vala2010/papers2010/VALA2010_127_Cathro_Final.pdf.
.

Cutter, Fensham, and Sumner 2023 Cutter, N., Fensham, R., and Sumner, T.D. (2023) “The slipperiness of name: Gender and biography in Australian cultural databases", *Gender & History*, 2023. https://doi.org/10.1111/1468-0424.12699.

Data Dictionary, n.d. "Data Dictionary" (n.d.) *Science Direct*. Available at: https://www.sciencedirect.com/topics/computer-science/data-dictionary.

Desfray and Raymond 2014 Desfray, P. and Raymond, G. (2014) *Modeling enterprise architecture with TOGAF: A practical guide using UML and BPMN*. Cambridge, MA: Morgan Kaufmann.

Given, Polkinghorne, and Cattlin 2023 Given, L. M., Polkinghorne, S., and Cattlin, J. (2023) "Structural elements and spheres of expertise: Creating a healthy ecosystem for cultural data initiatives", *Journal of the Association for Information Science and Technology*, 2023. https://doi.org/10.1002/asi.24849.

Hunter et al. 2012 Hunter, J. et al. (2012) "A Web 3.0 approach to building an online digital archive of architectural practice in post-war Queensland", *Proceedings of the international council on archives conference, 2012*. Brisbane, Australia, 20-24 August. Available at: http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00326.pdf.

Jones and Piper 2023 Jones, M. and Piper, A. (2023) "Digital history: State of the field review essay", *Australian Historical Studies*, 55(1), pp. 178-203. https://doi.org/10.1080/1031461X.2023.2267586.

Jones and Verhoeven 2022 Jones, M. and Verhoeven, D. (2022) " Trove's funding runs out in July 2023 – and the National Library is threatening to pull the plug. It's time for a radical overhauls", *The Conversation*, 23 December. https://theconversation.com/troves-funding-runs-out-in-july-2023-and-the-national-library-is-threatening-to-pull-the-plug-its-time-for-a-radical-overhaul-197025.

Lagoze 2010 Lagoze, C.J. (2010) *Lost identity: The assimilation of digital libraries into the Web*. PhD thesis. Cornell University. Available at: https://carllagoze.files.wordpress.com/2012/06/carllagoze.pdf.

Lee 2023 Lee, B.C.G. (2023) "The 'collections as ML data' checklist for machine learning and cultural heritage", *Journal of the Association for Information Science & Technology*, 2023. https://doi.org/10.1002/asi.24765.

Liu, McKay, and Buchanan 2023 Liu, R., McKay, D., and Buchanan, G. "Person-oriented ontologies analysis for digital humanities collections from a metadata crosswalk perspective", *Proceedings of the 86th annual meeting of the association for information science & technology, 2023*. London, England, 27-31 October. RMIT University, pp. 255-266. Available at: https://researchrepository.rmit.edu.au/esploro/outputs/conferenceProceeding/Person-Oriented-Ontologies-Analysis-for-Digital-Humanities/9922290625501341.

Loukissas 2019 Loukissas, Y.A. (2019) *All data are local: Thinking critically in a data-driven society*. Cambridge, MA: The MIT Press.

Lynch 1997 Lynch, C.A. (1997) "The Z39.50 information retrieval standard", *D-Lib Magazine*, 3(4).

McGillivray et al. 2020 McGillivray, B. et al. (2020) *The challenges and prospects of the intersection of humanities and data science: A white paper from the Alan Turing Institute*. Available at: https://doi.org/10.6084/m9.figshare.12732164.

Millar et al. 2018 Millar, P. et al. (2018) "The challenge, the project, and the politics: Lessons from six years of the UC CEISMIC Canterbury Earthquakes Digital Archive", in Bouterey, S. adnd Marceau, L.E. (eds.) *Crisis and disaster in Japan and New Zealand: Actors, victims, and ramifications*. New York: Palgrave Macmillan, pp. 159-179.

Nicholas and Touma 2023 Nicholas, J. and Touma, R. (2023) "How to win the Archibald prize: What 100 years of data tells us", *The Guardian*, 5 May. Available at: https://www.theguardian.com/artanddesign/2023/may/05/how-to-win-the-archibald-prize-what-100-years-of-data-tells-us.

Rahaman, Champion, and McMeekin 2023 Rahaman, H., Champion, E., and McMeekin, D. (2023) "Outside inn: Exploring the heritage of a historic hotel through 360-panoramas", *Heritage*, 6(5), pp. 4380-4410. https://doi.org/10.3390/heritage6050232.

Rankin, Gordon, and Potter 2009 Rankin, R., Gordon, M., and Potter, R. (2009) "Implementing a federated data archive with asynchronous data query, gathering and analysis capabilities", *AGU Fall Meeting Abstracts*, 33(SM33B-1570).

Ryan et al. 2022 Ryan, T. et al. (2022) "Codebooks", *Australian Cultural Data Engine*, June-November. Available at: https://www.acd-engine.org/datasets.

Sherratt 2019 Sherratt, T. (2019) "Hacking heritage: Understanding the limits of online access", in Lewi, H. et al. (eds.) *The Routledge international handbook of new digital practices in galleries, libraries, archives, museums and heritage sites*. New York: Routledge.

Smithies 2017 Smithies, J. (2017) *The digital humanities and the digital modern*. New York: Springer.

Smithies et al. 2023 Smithies, J. et al. (2023) “MaDiH (مديح): A transnational approach to building digital cultural heritage capacity”, *Journal on Computing and Cultural Heritage*, 15(4), pp. 1-14. https://doi.org/10.1145/3513261.

Smithies, Millar, and Thomson 2015 Smithies, J., Millar, P., and Thomson, C. (2015) “Open principles, open data: The design principles and architecture of the UC CEISMIC Canterbury Earthquakes Digital Archive”, *Journal of the Japanese Association for Digital Humanities*, 1(1), pp. 10-36. Available at: https://ir.canterbury.ac.nz/items/1add02b2-49f2-49b7-8c22-261fa35f7d7d.

Stainforth 2019 Stainforth, E. (2019) “Disruptive forms, persistent values: Negotiating digital heritage and ‘the memory of the world’”, in Carter, T. et al. (eds.) *Creating heritage: Unrecognised pasts and rejected futures*. New York: Routledge.

Sumner and Fensham 2023 Sumner, T.D. and Fensham, R. (2023) “The hidden stories in Australian cultural data”, *Pursuit*. Available at: https://pursuit.unimelb.edu.au/articles/the-hidden-stories-in-australia-s-cultural-data.

Sumner and Munoz 2023 Sumner, T.D. and Munoz, J. (2023) “Gendered labour in the Australian arts”, *Pursuit*. Available at: https://pursuit.unimelb.edu.au/articles/gendered-labour-in-the-australian-arts.

Sumner et al. 2023 Sumner, T.D. et al. (2023) “What's in a name? A cross-section of biography, gender & metadata in the *Design & Art Australia Online* database”, *Proceedings of the twentieth international conference on Dublin Core and metadata applications, 2022*. Virtual, 3-7 October. Available at: https://rest.neptune-prod.its.unimelb.edu.au/server/api/core/bitstreams/9cb75031-a202-4e98-9e7b-15ab8a5c6924/content.

UKRI 2022 UK Research and Innovation (2022) *Towards a national collection: Opening UK heritage to the world*. Available at: https://www.librarydevelopment.group.shef.ac.uk/referencing/harvard.html.

Warwick et al. 2009 Warwick, C. et al. (2009) “Documentation and the users of digital resources in the humanities”, *Journal of Documentation*, 65(1), pp. 33-57. https://doi.org/10.1108/00220410910926112.

Whitelaw 2015 Whitelaw, M. (2015) “Generous interfaces for digital cultural collections”, *Digital Humanities Quarterly*, 9(1). Available at: https://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.

Zheng and Munoz 2023a Zheng, I. and Munoz, J. (2023a) “Australian Cultural Data Engine workbook”, GitHub. Available at: https://acd-engine.github.io/jupyterbook/about.html.

Zheng and Munoz 2023b Zheng, I. and Munoz, J. (2023b) “ACD-ENgine unified data dictionary”, *Australian Cultural Data Engine*. Available at: https://acd-engine.org/datasets.

Zheng and Munoz 2023c Zheng, I. and Munoz, J. (2023c) “Overview of the Australian Cultural Data Engine architecture”, *Australian Cultural Data Engine*. Available at: https://acd-engine.github.io/jupyterbook/Integration_ACDEA_Overview.html.

van de Sompel et al. 2004 van de Sompel, H. et al. (2004) “Resource harvesting within the OAI-PMH framework”, *D-lib Magazine*, 10(12). Available at: https://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html.