Humanities Arts and Social Sciences – Data Enhanced Virtual Laboratory

Data Curation Framework (DCF)

Reference Document

Version 2.1
May 2019

CC BY

Contents

1. Document purpose

This is a public reference document created by and for the HASS DEVL project, made available under an open Creative Commons licence, and is a project deliverable which contributes to ongoing work in the curation space. The contents of the document are designed to support future project activities, such as the development of training and online learning modules, published case studies, and other material.

2. The Data Curation Framework (DCF)

2.1. Overview

The Data Curation Framework (DCF) provides a structure and common language for working through the challenges associated with data curation. It is a conceptual model that can be used to understand how tools, standards, systems, and processes (many of which exist in various forms already) can be brought together to improve data curation, and to help further develop data curation capability across the diverse HASS research sector.

Using clear language, the framework will help users identify the tools, standards, systems, and processes they may need to consider in order to curate research data within a particular context. Modelling data curation in this way offers researchers and data custodians a structure through which they can develop data curation capabilities, select appropriate tools and standards, and approach the management of sustainable, interoperable research data with a clear understanding of what is required.

A common language helps promote shared understanding between partners, collaborators, and custodians, and provides a clear and consistent means for describing what curation processes have been applied to particular datasets. This will help to make HASS data more accessible, understandable, and usable over time.

2.2. Content

The Data Curation Framework includes:

- definitions of *Key Terms*, e.g.: data curation, research dataset, and reference dataset;
- an overview of the *Key Principles*, requirements, actors, processes, and events which make up the curation process; and,
- *Cross References* to other standards and policies, showing how the DCF relates to existing resources relevant to data curation and preservation practice; and,
- information on *State Changes* which alter the custody, form, contents, location, or accessibility of data.

2.3. What is a 'framework'?

The DCF is a framework rather than a standard, policy, procedural document, or singular approach to data curation. Frameworks are not singular or linear (though they can incorporate linear process flows). They provide a structure which can be used to investigate and understand tools, systems, standards, policies, and procedures, and to communicate with colleagues and related professionals based on a shared understanding of key concepts.

2.4. What is the intention of the DCF?

The DCF is intended to assist HASS researchers working with datasets that need to be maintained, preserved, and made accessible over time. In doing so, it looks to go beyond

research data management requirements at a researcher or project level and the limitations of self-contained lifecycle models, providing a conceptual framework which allows researchers and their collaborators to think through key issues regarding data curation and interoperability in the broader HASS research domain.

The DCF is primarily intended to support the curation of reference datasets (defined below), though in doing so the DCF also provides a useful tool with which researchers and data custodians can develop research datasets into reference datasets.

2.5. Who is it for?

The DCF is intended for:

- Individual HASS researchers and research teams responsible for datasets with current or ongoing value to a research community.
- Other data custodians in universities, the academic sector, and GLAM.[1]

2.6. What is it for?

The framework is a shared conceptual lens which can be used by the HASS research community and related institutions to explore gaps in the current HASS data curation landscape. These gaps include:

- a defined place for researchers in which data curation happens;
- storage options which support interoperability;
- support for 'end of life' activities when projects wind down;
- effective support for distributed, decentralised collections;
- advice on curating social media data;
- agreement around responsibilities;
- clear data curation capability development pathways;
- support for understanding, choosing, and implementing standards; and,
- HASS-specific advice on data curation.[2]

2.7. What are the benefits of the DCF?

Using the DCF as part of HASS data curation activities will benefit:

- *data creators*, by making their data more sustainable, discoverable, accessible, and usable over time;
- *data custodians*, by supporting ongoing requirements around data preservation and access;
- *HASS researchers*, by helping to develop more effective collections and making them available to the broader research community; and,
- *HASS scholarship*, by developing persistent, citable, interoperable datasets.

---

[1] Galleries, libraries, archives, and museums, often collectively referred to as the GLAM sector.

[2] The gaps listed here were identified in consultation with researchers and other data custodians at the roundtable discussions held in Melbourne, Brisbane, and Adelaide in late 2018.

3. Background and context

The Data Curation Framework (DCF) is designed to create an accessible, shared language and structure through which researchers and custodians can approach the curation and management of research data.

Fragmentation of high value data, tools, and services remains a significant problem for research and information infrastructure. This is largely due to ad hoc data handling and curation practices as data moves between systems and is processed or enhanced. In response to these challenges, the DCF offers a model for moving understandings of HASS data curation forward in the Australian research landscape.

3.1. HASS Data Enhanced Virtual Laboratory

The DCF emerges from the broader HASS Data Enhanced Virtual Laboratory project (HASS DEVL). HASS DEVL is a collaboration between a range of partner organisations, including AARNet, Alveo, AURIN, Australian Data Archive, eResearch SA Limited, Griffith University, The National Library of Australia, and The University of Melbourne.

The project is focused on supporting researchers working in the humanities, arts, and social sciences. The HASS research community is the largest portion of the Australian research landscape by any measure, commonly understood to comprise more than 40% of current funded research. HASS is a multidisciplinary grouping that represents significant domain specialisation, interdisciplinarity, and transdisciplinarity.

Since late 2017, the distributed HASS DEVL team has been working on a number of initiatives for HASS researchers, including Tinker (https://tinker.edu.au/), national Digital Humanities Pathways Forums, and the training of Digital Champions.

The DCF is a key component of this work.

3.2. Development process

The HASS DEVL DCF was first developed by Ingrid Mason in the first half of 2018, in consultation with other members of the project, and informed by the processing of three reference datasets (including the transfer of those data between platforms and parties).[3] An initial draft was produced (9 July 2018) and circulated within the distributed project team for discussion and feedback. Following this, a second draft was produced (24 August 2018) and circulated to the project team, and colleagues at the University of Sydney Library. The final version of Version 1 of the DCF was circulated to the project Steering Committee on 15 October 2018.

[3] The reference datasets were: Australian Government Gazette (1832-1968); Australian Census (1981, 1986, 1991); and Australian Bigamy Prosecutions (1849-1999). For more information on these datasets, see: https://tinker.edu.au/data/available-datasets/

Following completion of this version, Mike Jones arranged three roundtable discussions on the DCF, in Melbourne (12 November), Brisbane (26 November) and Adelaide (27 November). There were 25 participants, including HASS researchers, data custodians, professionals from the GLAM (galleries, libraries, archives, museums) sector, and members of the HASS DEVL project team.

Each discussion focused on the following areas:

- the definition and scope of the term 'data curation';
- the awareness of existing data curation frameworks and models;
- defining the gap (if any) that exists in the current data curation landscape for HASS researchers and data custodians; and,
- obtaining feedback on the structure, terminology, and scope of the high-level data curation framework developed by the HASS DEVL project.

What follows is Version 2 of the DCF, developed in response to the feedback from those roundtable discussions. Further consultation is now required to ensure that the needs of the HASS community are being met.

4. Definitions

4.1. Data

All researchers have data. Here, the definition used is drawn from the University of Melbourne's Management of Research Data and Records Policy (MPF1242):

> *Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational.*[4]

There are many different definitions of data available, some of which (like the University of Melbourne's) are very broad, others which are quite narrow. However, such variations have little bearing on the framework which follows.

The DCF is focused on the process of curation, rather than detailed analysis of the thing being curated.

4.2. Data curation

Data curation is an ongoing process by which data is documented, managed, and preserved over time. Curation also supports the access, use, and reuse of data, not just its storage.

Effective data curation combines human expertise with tools and technologies. It requires effective, well-structured description, including through the use of existing metadata schemas, standards, and conceptual frameworks. Consideration of ethics and rights management is also essential.

The purpose of data curation is to:

- support data preservation (where possible and desirable);
- ensure that data remains usable and understandable;
- enable access open (or controlled) based on clearly-articulated criteria;
- capture data provenance and decisions that are made in relation to that data are documented and accessible; and,
- manage data so that it remains connected, relational, and interoperable.

4.3. Datasets

Datasets can have different functions and forms, can be static and ongoing, and can have single, dual, and multiple usage types. Datasets may be formalised (i.e. they are curated, documented, and codified through individual determinations or community agreements,

[4] The University of Melbourne, 'Management of Research Data and Records Policy (MPF1242)', Melbourne Policy Library, 20 November 2013, https://policy.unimelb.edu.au/MPF1242#section-3.1.

with published processes and governance that indicate how the data is collected, managed, and is to be interpreted) or may be lacking in some or all of these characteristics at various times throughout their generation, use and retention. As datasets become formalised, and move from undocumented and unmanaged to documented and managed, the potential for reuse increases.

4.4. Research Datasets

Research datasets can be inputs to research or outputs from research. The research questions guiding data selection, capture, processing, analysis and interpretation through the research process, generate research value. Research datasets may be developed to address a single research question, or a set of closely related research questions. Here, the term is used to specify datasets which remain largely within this context, and which have not been documented, generalised, or made accessible to researchers working in contexts separate from the research questions which led to the dataset being assembled. Where this occurs, a research dataset may become a reference dataset.

4.5. Reference Datasets

Typically, reference datasets will be versioned and will have well-documented governance and change processes. Reference datasets form part of data infrastructure that enables ongoing research use, i.e. the datasets may be cited, integrated, interpreted and contested, and utilised to serve or inform a wide range of related or unrelated research questions, as part of the scholarly generation of knowledge in multiple research domains. As such, reference datasets have a special status in a data ecosystem, in that these datasets are intended to be a reliable data infrastructure component that has been curated for wide reuse. Reference datasets are intentionally created through editorial processes created by data and research domain specialists, so that multiple research processes and outputs can be built upon them.

5. Existing standards

There are numerous existing standards and policies in Australian and internationally which relate to elements of data curation. Many of these reference documents provide useful guidance for specific scenarios – aiming for open data; digital preservation; working with Aboriginal and Torres Strait Islander material – but this level of focus can be a barrier to a broader, conceptual understanding of the requirements of data curation. The DCF aims to step above this sphere of individual standards, requirements, processes, and models to look at the broader ecosystem and what is required in this space.

Using the DCF in this way will help researchers and data custodians understand, evaluate, and apply these existing standards and policies more effectively. To assist with this, and to show how the concepts outlined in the framework relate to existing standards and policies, cross-references are included to five of these works:

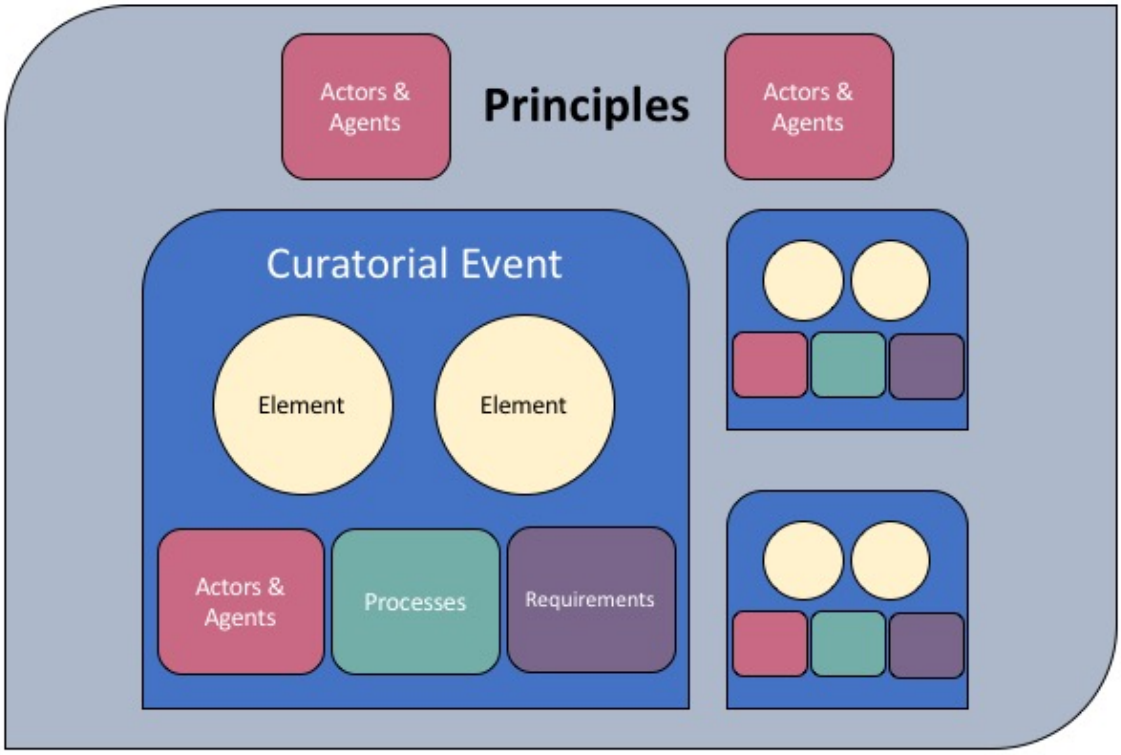
- ATSILIRN Protocols for Libraries, Archives and Information Services: http://atsilirn.aiatsis.gov.au/protocols.php
- Australian Code for the Responsible Conduct of Research (ACRCR): https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018
- DCC Curation Lifecycle model: http://www.dcc.ac.uk/resources/curation-lifecycle-model/
- FAIR/FAIRER data principles: https://www.force11.org/group/fairgroup/fairprinciples
- OAIS Reference Model (ISO 14721): http://www.oais.info/

There are many other references which may be relevant to data curation in some contexts, but which are not covered here, including:

- AIATSIS Guidelines for Ethical Research in Australian Indigenous Studies: https://aiatsis.gov.au/research/ethical-research/guidelines-ethical-research-australian-indigenous-studies
- ANDS guidelines on Publishing and Reusing Data: https://www.ands.org.au/working-with-data/publishing-and-reusing-data
- Data.SA Open Data Toolkit: https://data.sa.gov.au/toolkit
- DPMC Public Data Policy: https://pmc.gov.au/public-data/public-data-policy
- 5 Star Open Data Principles: https://www.w3.org/DesignIssues/LinkedData.html
- Institutional research data management frameworks: https://www.ands.org.au/guides/creating-a-data-management-framework
- Santa Barbara Statement on Collections as Data: https://collectionsasdata.github.io/statement/
- W3C PROV (Provenance) Model Primer: https://www.w3.org/TR/prov-primer/

6. Framework components

This section introduces the primary components which together form the Data Curation Framework.



Any representation of the DCF is by necessity more singular and segmented than the framework will be in practice. As noted earlier, the DCF is not a process or standard. Data curation consists of numerous curatorial events where people and other agents use processes and specify requirements to work with elements. All these curatorial events take place in the context of broader principles.

For example, a data custodian (**Actor**) interested in making their datasets (**Elements**) more accessible and interoperable (**Principles**) decides to migrate the format (**Processes**) of their current dataset using a repeatable, reversible process (**Requirements**). These components together constitute an act of data curation (**Curatorial Event**).

The following sections outline each component in more detail, starting with the underlying principles, then curatorial events, and then each component of an event. Cross-references are provided to the five reference documents noted above. These references should be treated as indicative rather than comprehensive. References to the OAIS are drawn from the August 2009 'Pink Book,' which is publicly accessible at no cost, rather than the full ISO standard (which is behind a paywall).[5]

[5] https://digital.library.unt.edu/ark:/67531/metadc123534/m2/1/high_res_d/650x0p11.pdf

6.1. Principles

All data curation activities should be governed by an underlying set of principles. Possible examples include that data curation aim to make data and datasets:

- accessible — *ref: FAIR (A1. to A2.); ATSILIRN (4.); OAIS (3.2.6)*
- discoverable — ref: FAIR (F3.)
- ethical — *ref: ACRCR (P1, P4, P5, P6, P7, P8)*
- findable — *ref: FAIR (F1. to F4.)*
- interoperable — *ref: FAIR (I1. to I3.); OAIS (6.)*
- persistent — ref: FAIR (F1.)
- reusable — *ref: FAIR (R1. to R1.3.)*
- revisable
- understandable — *ref: OAIS (3.2.4); ACRCR (P3)*
- useful

None of these should be considered mandatory. *For example*, while it may be suitable to operate under the FAIR principles and curate historical data about the manufacturing industry in the 1930s so as to make it accessible and reusable, it may not be appropriate to do so for Aboriginal population data from the 1970s.

See also: ATSILIRN (esp. 1., 6., 7., 10.); ACRCR (full code).

6.2. Curatorial Events

When an **Agent** uses **Processes** to curate **Elements**, the result is a **Curatorial Event**. These take place in time and space, producing outputs (such as new or altered **Elements**) and records (evidence of the **Curatorial Event**).

Possible categories for events include:

- document — *ref: ATSILIRN (5.); ACRCR (P3)*
- describe — *ref: FAIR (F2.); ATSILIRN (5.); DCC (Description /Representation Information); OAIS (4.2.1.4.2)*
- appraise — *ref: DCC (Appraise and Select; Reappraise)*
- preserve — *ref: OAIS (5.); DCC (Curate and Preserve)*
- destroy — *ref: DCC (Dispose)*
- analyse
- use
- link
- store — *ref: DCC (Store)*
- access — *ref: DCC (Access, Use and Reuse)*
- move — *ref: DCC (Migrate)*
- transform — *ref: DCC (Transform); OAIS (5.1.3.4)*

6.3. Elements

Data curation activities also involve a number of elements, which could include:

- research data *ref: DCC (Digital Objects, Databases)*
- research datasets *ref: DCC (Digital Objects, Databases)*
- reference datasets *ref: DCC (Digital Objects, Databases)*
- metadata *ref: FAIR (F2., A2.)*
- data documentation
- provenance documentation *ref: FAIR (R1.2.)*

6.4. Actors & Agents

The right actors and agents need to be involved in data curation activities, to establish the **Principles** for data curation, and to meet the requirements of specific **Curatorial Events**. Understanding the various actors and agents involved and their role in different parts of the data curation process will also assist with documenting provenance.

Actors and agents could include:

- researchers *ref: OAIS (2.1)*
- data creators *ref: OAIS (2.1)*
- data custodians
- projects
- programs *ref: OAIS (2.1)*
- systems *ref: OAIS (2.1)*
- software applications
- tools
- organisations
- repositories

6.5. Processes

Data curation processes are applied to elements by actors and agents. Examples of processes include:

- documentation *ref: DCC (Description/Representation Information)*
- augmentation *ref: DCC (Transform)*
- migration *ref: DCC (Migrate)*
- normalisation
- integration
- redefinition
- generalisation
- connection

6.6. Requirements

A set of requirements should be established as a basis for the development and implementation of consistent, sustainable, clearly-documented data curation events.

Requirements could include that all data curation activities be:

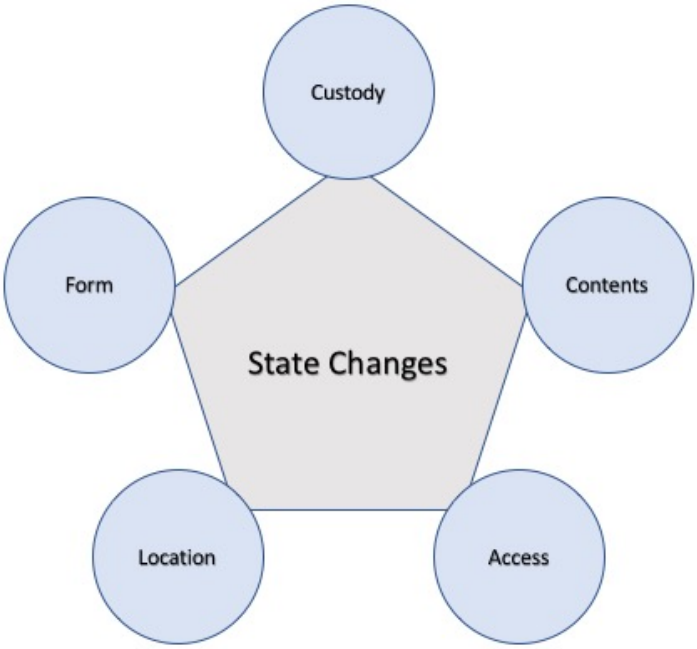
- predictable
- referenceable
- reliable *ref: DCC (Preservation Action)*
- repeatable *ref: OAIS (4.2.3)*
- retrievable *ref: FAIR (A1.)*
- reversible *ref: OAIS (5.1.3.4)*
- understandable *ref: OAIS (3.2.4)*

Deciding on requirements in conjunction with selecting tools, engaging particular actors, and designing curatorial events will help to ensure data curation is managed effectively.

7. State Changes

Datasets are not static or fixed. Curation is thus an ongoing process rather than a singular activity. When the state of a dataset changes, effective curation activities become particularly important to ensure that the data remains (or becomes) accessible, understandable, usable, and interoperable.

Key changes in state include changes in: *Custody*, *Contents*, *Access*, *Location*, or *Form*.



7.1. Custody

Changes in custody refer to control of the dataset, including:

- who is responsible for maintaining the data;
- who determines (with appropriate consultation) what is accessible; or,
- who provides access to current or potential users.

For example, when a research dataset is transferred from a research project to a repository like the Australian Data Archive to become a reference dataset, its **custody** has changed.

7.2. Contents

A change in contents refers to alterations or additions to the content of a dataset, including:

- the addition of new items to a dataset; or,
- the augmentation of existing data through annotation, transcription, the addition of GIS data, or similar.

For example, when a social media dataset containing Tweets harvested via the Twitter API is expanded through the addition of the most recent month of tweets, the **contents** of the dataset has changed.

7.3. Access

Changes in access refer to alterations in the access conditions and protocols surrounding a dataset, such as the process for requesting access or the credentials required from potential data users before access is granted. Access also relates to conditions (what can be done with the data) and ethical responsibilities (what should be done with the data, or what is permitted).

For example, if a data custodian consults with representatives from an Aboriginal community as part of their curatorial responsibilities and determines that only women should be permitted to access a particular dataset, they should change their policies to reflect this cultural protocol. When this occurs, **access** to the data has changed.

Changes in location or custody often also result in changes to access; however, there are also many instances where access changes independently from these states.

7.4. Location

Changes in location refer to data being moved, such as when:

- data is moved to a different server or data centre;
- data is moved from a project drive to an institutional repository; or,
- data is moved from one type of repository to another.

For example, when the funding period for a research project finishes and the dataset is moved from a faculty server in an institution to the institutional repository, the **location** of the data has changed.

Though changes in custody often result in a change in location, data is often moved without a change in custody, as in the example given where the researchers responsible for the dataset may well maintain authority over who can augment or access the dataset.

7.5. Form

Change in form refers to the format in which data is held (as distinct from the contents of the dataset, which is covered below). Form can change as a result of research activity, through format migration to support digital preservation activities, or through the creation of dissemination copies for use.

For example, if a researcher decides to convert data stored in Excel into SPSS files to allow for alternative forms of analysis, the **form** of the data has changed. Similarly, preservation actions may include converting data stored in Excel into .csv files to move away from a proprietary format.

In most cases when form 'changes' the earlier form will remain. With reference to the example above, creating an SPSS file does not *replace* the Excel file that is the source. Therefore, most changes of form involve the creation of additional forms, rather than the transformation of one form into another.

7.6. Capturing state changes

State changes represent a change in a dataset which could affect the ways in which researchers understand, interpret, and analyse those data. It is therefore essential that the data curation process capture information on the state changes outlined above as these changes occur over time.

Documentation related to state changes should be:

- *cumulative*, rather than limited to the most recent change;
- *provenanced*, indicating why the change was made, by whom, and based on what authority;
- *persistent*, so it is preserved along with the dataset;
- *citable*, allowing researchers to reference information about the state change as part of their own datasets or research outputs;
- *discoverable*, so data users and custodians can access information about the dataset; and,
- *accessible*, to allow users and custodians to get access to relevant information as required.

The OAIS Reference Model (ISO 14721) provides detailed guidance on the documentation of state changes as part of the preservation of digital data.

8. Applying the DCF

8.1. 'Two Digital Humanities Case Studies'

Over the same period as the development of Version 2 of the DCF, Dr Joanne Mihelcic researched two projects in Australian universities:

- the Theatre and Dance Platform (T&DP) at the University of Melbourne; and,
- Tracking Infrastructure for Social Media Analysis (TrISMA) at Queensland University of Technology.

Mihelcic's work helps to highlight the value of a shared framework for data curation activities in the HASS sector. This section draws on a selection of her findings.

A more detailed examination of these two case studies is available in the document titled 'Two Digital Humanities Case Studies'. Extracts from that report are provided here in italics.

> *this study highlighted the significance of planning and designing the infrastructure to support sharing and reuse of quality data*

Principles: planning and designing to support sharing and reuse is about undertaking data curation to make data accessible, discoverable, understandable, reusable, and useful.

> *Projects are the product of collaborations beyond traditional academic practices or research; they involve interdisciplinary partnerships within and often outside of university boundaries*

Actors & Agents: clearly defining who and what is involved in the collaboration, their role, and their purpose helps with understanding such collaborations.

State Changes: understanding the context of actors and agents (for example, university researchers working with data creators outside of university boundaries) can also help to define the sorts of processes and documentation which might be required when **Curatorial Events** result in changes to the **Custody**, **Access**, and **Location** of data and datasets.

> *There are limitations regarding the use of Twitter data for research and ethical research practice especially as not all Twitter users are aware of how public their data is.*

Principles: working within the boundaries of clearly articulated ethical principles will ensure such data is curated in a way which recognises these limitations.

The data curation framework as a whole can also contribute to two areas highlighted in Mihelcic's report:

- '*whole of life' project planning* and the *costs for operationalising the management and sharing of data* requires a detailed understanding of the **Principles** which will be maintained over time (e.g. is the aim to ensure the data remains accessible and usable over time, or only that the data remain discoverable?) and an understanding of the **Elements, Actors & Agents,** and **Processes** which are likely to be required as part of undertaking current and future **Curatorial Events**.
- *Development of research specific collections is undertaken using an ad hoc approach when instigated by academic researchers.* The DCF provides a means by which researchers and their collaborators can more clearly map out and articulate what is required, and bring together (or request) the resources and capabilities required to achieve their aims.

More broadly, the DCF as outlined here can operate as a framework for future research of this type, guiding areas of discussion, providing a structure through which to develop questions for in-depth interview questions and reviews, and developing a common language which can be used to describe findings and present results in a way which helps to facilitate comparisons across projects, programs, institutions, and domains.

8.2. State Changes

This example of a state change is drawn from Version 1 of the DCF. It demonstrates how concurrent changes in state often arise through data curation activities.

Existing State

Custody	ADA
Contents	Statistical metadata and value data, geospatial metadata and value data.
Access	ADA Platform
Location	ADA Platform
Form	Statistical data tables derived through digitisation of Census records discoverable with ADA descriptive metadata from ADA catalogue. Data made available as: csv or spatial formats (MapInfo, Geodatabase, GeoJSON).

New State

Custody	AURIN
Contents	Statistical metadata and value data, geospatial metadata and value data, historical boundaries data.

Access	AURIN Portal and API
Location	AURIN Platform
Form	Statistical data associated with Census metadata integrated with geospatial data in GDA94 format, is presented using Historical boundaries, and is discoverable within AURIN catalogue.

Data Curation

- Geospatial value data is normalised into GDA94.
- Historical boundary definitions (aka geographies) are integrated with statistical data.
- Statistical value data is generalised and presented within historical boundaries.
- Statistical metadata is normalised (from Census dictionary).
- ADA data description metadata is mapped across to AURIN schema.