HASS DEVL – Data Curation Framework

Roundtable Discussions – Key Findings
Prepared by: Mike Jones, 10 December 2018

Introduction

In November 2018, the Humanities and Social Sciences Data Enhanced Virtual Laboratory (HASS DEVL) project hosted three roundtable discussions, 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.

The following document draws together key findings on each of these areas from the three workshops.

What is 'data curation'?

Participants were asked what the term 'data curation' meant to them. While two specific models – the Digital Curation Centre (DCC) Curation Lifecycle Model,[1] and the FAIR data principles[2] – were mentioned, generally participants drew on their own experiences and contexts.

Key points from across the three workshops included:

- an emphasis on data curation as a means for supporting the access, use, and reuse of data, not just its preservation and storage;
- the inclusion of human expertise in conjunction with technology;
- the need for effective, well-structured description, including through the use of existing metadata schemas, standards, and conceptual frameworks;
- the importance of connections, relationships, and interoperability as a driver for, and outcome of, the curation process; and,
- a consideration for ethics and rights management as part of data curation.

[1] http://www.dcc.ac.uk/resources/curation-lifecycle-model/
[2] https://www.go-fair.org/fair-principles/

What level of awareness is there of existing frameworks and practices?

The following list of frameworks, principles, and policies was presented to all participants:

- Codes and guidelines for the conduct of ethical research (9/6)
- Institutional research data management framework (7/5)
- ANDS guidance on data publishing and reuse (10/6)
- DPMC public data policy (0/0)
- ATSILIRN Protocols for Libraries, Archives and Information Services (7/2)
- OAIS reference model (6/1)
- FAIR data principles (8/5)
- DCC Data Curation Lifecycle model (5/2)
- Santa Barbara Statement on Collections as Data (1/0)

In Melbourne, a discussion revealed that there was awareness of institutional data management frameworks and FAIR, among several participants, with more limited awareness of DCC and OAIS, and little to no awareness of the Santa Barbara Statement.

The two numbers after each item are combined figures for Adelaide and Brisbane, the first showing how many participants (out of a total of 11) knew of the item, and how many had actually used that framework or principle. As these figures show, there is generally good awareness of institutional guidelines, ANDS, ATSILIRN Protocols, and FAIR data principles, but these don't always result in implementation (particularly in the case of ATSILIRN and OAIS).

Additional frameworks, such as Tim Berners-Lee's 5* Open Data principles, the W3C Provenance Model Primer, and the Data.SA Open Data Toolkit were mentioned by participants as potential additional inclusions.

What is the data curation gap?

The discussion of the gap in existing data curation frameworks and practices developed a different focus in each session. In Melbourne, the focus was primarily on infrastructure, the need for a place for data curation to happen, the challenge for storage options which also support interoperability, and the problems for 'end of life' activities when projects wind down.

Brisbane focused on specific issues, including: the need to support distributed, decentralised collections; problems associated with curating social media data; the lack of agreement around responsibilities; and the effect on data curation of inward-facing researchers, particularly if curation is not built in from early in the process.

For Adelaide, the primary gap was capability: the need to build data curation capability for HASS researchers, and in the GLAM sector; the need to provide better support for understanding, choosing, and implementing standards and frameworks; and the need to develop HASS-specific frameworks and advice for data curation.

The Data Curation Framework

Overall, the three discussion sessions highlighted the need for clear, practical language and terminology, well-structured diagrams, and the use of illustrative examples to help people engage with and understand the data curation concepts developed by the HASS DEVL project, and to build data curation capability in the HASS and GLAM sectors.

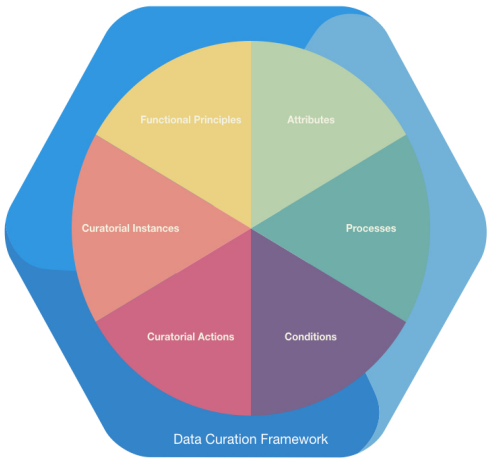
State Changes

Discussing the four state changes outlined – Custody, Location, Form, and Access – participants noted the need to disambiguate these elements and their coverage, and remove overlaps (for example, the use of 'form' under Access, and the use of 'access' under Custody).

The discussions also highlighted areas for inclusion, either within these categories or as separate categories:

- Augmentation – including the addition of more data, annotation, enhancement, etc.
- Connection – creating relationships, links, or connections with other datasets, or to other entities (people, organisations, places, events, references, etc.).

Elements



All three discussion sessions confirmed that the current visual representation of the DCF needs rethinking, to better reflect its structure and the relationship between elements, and to help provide guidance on how to apply the framework to particular scenarios.

At segment level the roundtables also confirmed the need to provide clearer terminology, and to review what is included under each section. The high-level findings from these discussions are provided below.

- **Functional Principles**
  - should be renamed ‘Principles’ and draw more explicitly on existing models (e.g. FAIR/FAIRER)
  - need to be represented as over-arching principles that sit above or around the rest of the framework
  - should include more explicit reference to ethics/rights
- **Attributes**
  - generally agreed these are elements, entities, or components rather than attributes, and need to be presented in this way
  - some elements in this category are processes rather than entities, and should be moved
- **Processes**
  - these are expressed in technical language which may not be accessible for HASS researchers
  - other processes need to be included, such as Transformation, Deidentification, Revision, Annotation, Interpretation, etc.
- **Conditions**
  - it is unclear what these are the ‘conditions’ for – the processes, or the whole framework?
  - the term used suggests access conditions, copyright, cultural considerations, and similar (particularly for HASS researchers) and may need to be clarified accordingly
- **Curatorial Actions**
  - widely agreed that these should be called Actors, Agents, or similar
  - other actors need to be included, such as Organisations and Communities, and there may be a need to include roles
- **Curatorial Instances**
  - widely agreed that these should be labelled Events or similar
  - other event categories may need to be made explicit, such as Described, Linked, Analysed, Stored, Preserved, or Destroyed

Based on this feedback, the following restructured Data Curation Framework was proposed at the conclusion of the Adelaide workshop.