# Curation

## Introduction defining context and scope

“Digital curation is the selection, preservation, maintenance, collection and archiving of digital assets. Digital curation establishes, maintains and adds value to repositories of digital data for present and future use. This is often accomplished by archivists, librarians, scientists, historians, and scholars”(Wikipedia).

It should be noted that Cataloguing, Curation and Provenance are commonly grouped together since the metadata, workflow, processes and legalistics associated with each have >70% intersection and therefore rather than generating independent systems a common approach is preferable. Moreover, there are strong interdependencies with identification and citation, with AAAI, with processing, with optimisation, with modelling and with architecture.

## Change history and amendment procedure

The review of this topic will be organised by Keith Jeffery in consultation with the following volunteers: . They will partition the exploration and gathering of information and collaborate on the analysis and formulation of the initial report. Record details of the major steps in the change history table below.For further details of the complete procedure see item 4 on the Getting Started page.

Note: Do not record editorial / typographical changes. Only record significant changes of content.

| Date | Name | Institution | Nature of the information added / changed |
|---|---|---|---|
| 11 Jan 2016 | Keith Jeffery | | Draft content provided by Keith Jeffery |
| | | | |

## Sources of information used

Relevant sources are DCC, OAIS (both discussed below) and RDA which has several relevant groups notably preservation[1] but also active data management plans[2] and reproducibility[3].

---

[1] https://rd-alliance.org/groups/preservation-e-infrastructure-ig.html
[2] https://rd-alliance.org/groups/active-data-management-plans.html
[3] https://rd-alliance.org/groups/reproducibility-ig.html

## Two-to-five year analysis

The ideal curation state is aimed to ensure availability of digital assets through media migration to ensure physical readability, redundant copies to ensure availability, appropriate security and privacy measures to ensure reliability and appropriate metadata to allow discovery, contextualisation and use including information on provenance and rights. The current practice commonly falls far short of this with preservation commonly linked with backup/recovery (usually limited to the physical preservation of the digital asset) and lacking the steps of curation (selection, ingestion, preservation, archiving (including metadata) and maintenance. Furthermore in the current state while datasets may be curated it is rare for software or operational environments to be curated.

**Lifecycle**

The desirable lifecycle is represented by a DCC (Digital Curation Centre) diagram [1] available in Fig. 1.

**Data Management Plan**

Increasingly research funders are demanding a DMP (Data Management Plan). Different organisations have proposed different templates and tools for plans but that of DCC is used widely[2] as is the US equivalent[3]. A DMP is defined (Wikipedia) “A data management plan or DMP is a formal document that outlines how you will handle your data both during your research, and after the project is completed”.

**OAIS Reference Model**

OAIS (Open Archival Information Systems Reference Model — ISO 14721:2003) provides a generic conceptual framework for building a complete archival repository, and identifies the responsibilities and interactions of Producers, Consumers and Managers of both paper and digital records. The standard defines the processes required for effective long-term preservation and access to information objects, while establishing a common language to describe these. It does not specify an implementation, but provides the framework to make a successful implementation possible, through describing the basic functionality required for a preservation archive. It identifies mandatory responsibilities, and provides a standardised method to describe repository functionality by providing detailed models of archival information and archival functions[4]. A set of metadata elements in a structure has been proposed.[5]

:

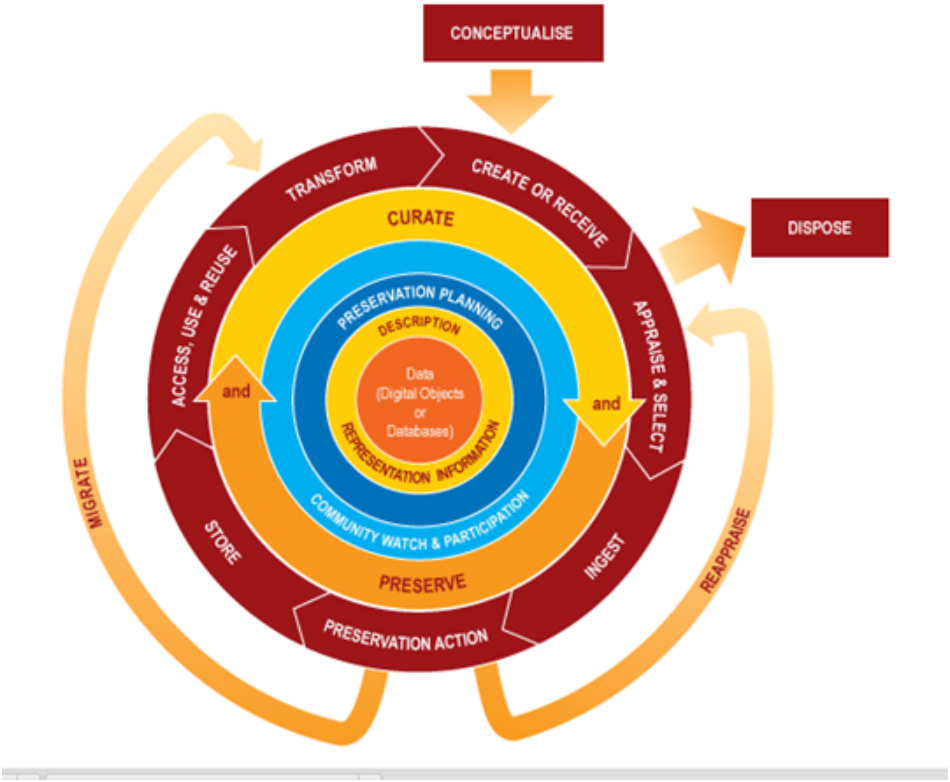


Fig. 1: The Curation Lifecycle Model

**Problems to be Overcome**

The following are some important problems which need to be addressed for curation:

1. **Motivation**: There is little motivation for a researcher to curate her digital assets. At present curation activity obtains no 'reward' such as career preferment based on citations. In some organisations curation of digital assets is regarded as a librarian function but without the detailed knowledge of the researcher the associated metadata is likely to be substandard. Increasingly funding agencies are demanding curation of digital assets produced by publicly-funded research.
2. **Business model**: Curation involves deciding what assets to curate and of those, for how long they should be kept. The duration is the problem, economics and business models do not manage well the concept of infinite time. First a business justification is needed in that (a) the asset cannot be collected again (i.e. it is a unique observation, experiment); (b) the cost of collecting again (by the same or another researcher) is greater than the cost of curation.
3. **Metadata**: Metadata collection is expensive unless it is automated or at least part-automated along the lifecycle re-using information already collected. Commonly metadata is generated separately for discovery, contextualisation, curation, provenance when in fact much of the metadata element content is shared across these functions. A comprehensive but incrementally completed metadata element set is required covering the required functions of the lifecycle.
4. **Process**: The lifecycle is well understood and needs process support. The incremental metadata collection aspect is critically important to success. Workflow models – if adapted to such incremental metadata collection with appropriate validation –are likely to be valuable here[6].
5. **Curation of data**: It may be considered that curation of data is straightforward –but it is not. First the dataset may not be static (by analogy with a type-specimen in a museum); both streamed data and updateable databases are dynamic thus leaving management decisions to be made on frequency of curation and management of versions with obvious links to provenance. Issues related to security and privacy change with time and the various licences for data use each have different complexities. The data may change ownership or stewardship. Derivatives may be generated and require management including relationships to the original dataset and all its attendant metadata.
6. **Curation of software**: Software written 50 years ago is unlikely to compile (let alone compose with software libraries and execute) today. Since many research propositions are based on the combination of the software (algorithm) and dataset(s) then the preservation and curation of the software becomes important. It is likely that in future it will be necessary to curate not only the software but also a specification of the software in a canonical representation so that the same software process or algorithm can be reconstructed (and ideally generated) from the specification. This leaves the question of whether associated software libraries are considered part of the software to be curated or part of the operating environment (see below).
7. **Curation of operational environments**: It is necessary to record the operational environment of the software and dataset(s). The hardware used – whether instrumentation for collection or computation devices – has characteristics relating to accuracy, precision, operational speed, capacity and many more. The operating system has defined characteristics and includes device drivers – i.e. a software library used by the application. It is a moot point whether software libraries belong to the application software or to the operational environment for the purposes of curation. Finally the management ethos of the operational environment normally represented as policies requires curation.

---

[1] http://www.dcc.ac.uk/resources/curation-lifecycle-model
[2] https://dmponline.dcc.ac.uk/
[3] https://dmp.cdlib.org/
***[4]*** *Higgins, S. (2006). "Using OAIS for Curation". DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3354. Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation*
[5] http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf

[6] Keith G Jeffery, Anne Asserson: ‘Supporting the Research Process with a CRIS’ in Anne Gams Steine Asserson, Eduard J Simons (Eds) ‘Enabling Interaction and Quality: Beyond the Hanseatic League’; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press ISBN 978 90 5867 536 1

# Sketch of a longer-term horizon

There is some cause for optimism:

1. Media costs are decreasing – so more can be preserved for less;
2. Awareness of the need for curation is increasing; partly through policies of funding organisations and partly through increased responsibility of some researchers;
3. Research projects in ICT are starting to produce autonomic systems that could be used to assist with curation.

However, the major problem is the cost of collecting metadata for curation. Firstly, incremental collection along the workflow with re-use of existing information should assist and workflow systems should be evolved to accomplish this. Secondly, improving techniques of automated metadata extraction from digital objects may reach production status in this timeframe[1].

---

[1] http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction

# Relationships with requirements and use cases

# Summary of analysis highlighting implications and issues

1. Commonality of metadata elements across curation, provenance, cataloguing (and more) so a common metadata scheme should be used;
2. Metadata collection is expensive so incremental collection along the workflow is required: workflow systems should be evolved to accomplish this;
3. Automated metadata extraction from digital objects shows promise but production system readiness is some years away
4. ENVRIplus should adopt the DCC recommendations
5. ENVRIplus should track the relevant RDA groups and – ideally –participate

# Bibliography and references to sources