### Data Lifecycle Models

A Data Lifecycle provides a structured approach of organising activities required for a digital object to pass though the different stages in its lifecycle, in order to achieve effective usage, preservation and sharing of the object. Several types of data lifecycle models exist for different domains, but two research data lifecycles will be discussed, namely DCC Curation Lifecycle and DataONE Lifecycle models.

The DCC Curation Lifecycle Model by Data Curation Centre in the UK “provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle” (DCC, 2019).

The DataONE Data Lifecycle was developed by the Data Observation Network for Earth Leadership Team in the USA in collaboration with the broader DataONE community, and built upon the National Science Foundation’s life cycle model in order to achieve in successful management and preservation of data for use and reuse. (DataONE, 2019)

Chart

Description automatically generatedGraphical user interface, application

Description automatically generated

**Figure 1. DCC Curation Lifecycle Model Figure 2: DataONE Lifecycle Model**

### Common Features

DCC Curation Lifecycle model and DataONE Lifecycle model have some features in common.

To begin with, both data lifecycle models support well the growing data sharing culture in research, business and industry. In other words, both DCC and DataONE data lifecycles map well into typical research projects’ timelines and beyond; where each stage of the lifecycle produces specific data products that are not only are they essential in the next stage, but also well documented and using standards, tools and guidelines, which enable smoother data curation, discovery and re-use after the research project ends.

In addition, both data lifecycles have sequential stages; with each stage requiring a variety of considerations, responsibilities and activities. Five of the eight stages are similar for both lifecycles (See Appendix A). Although the typical starting point for both lifecycles is the planning stage, it is possible to use only part of the lifecycle for both lifecycles. For instance, a research project that is focused in reusing data from another research might focus on the Discover, Combine and Process stages. In the planning stage, all aspects relating to the subsequent stages of the lifecycle which covers data creation, management, maintenance, retention, sharing and publication are laid out in a formal document called Data Management Plan (DMP).

### Key Differences

The most outstanding difference between the two models is that DataONE is a simpler data lifecycle model that comprises of 8 sequential stages, whereas DCC model is more complex and comprehensive data lifecycle model which has full lifecycle activities and occasional activities, in addition to its 8 sequential activities (see Appendix A). Furthermore, DataONE lifecycle has a different focus from that of DCC lifecycles.

### Differences Between DCC Curation Lifecycle Model and DataONE Data Lifecycle Model.

With data at the centre of DCC Curation lifecycle model, we can see the various full lifecycle data curation activities as we move outside the rings, 9 sequential stages as we go outside and finally occasional activities on the outermost side of the model (see Figure 1). Full lifecycle activities are considered at all stages of the lifecycle and occasional activities may holdup or reorder sequential activities. On the other hand, DataONE Data lifecycle model is a simpler model (see Figure 2). This can be seen in the Preservation stage where in DataONE’s model it is one sequential stage, while in the DCC’s it is split into Preservation Planning, as a full lifecycle activity, and Preservation Action, as a sequential stage. What is more is that in DCC lifecycle, the Describe phase is a full lifecycle activity as opposed to one sequential stage in DataONE’s lifecycle.

Furthermore, DataONE Lifecycle’s emphasis is on data management for use and re-use (DataONE, 2019). For example, DataONE’s model includes the Assure stage, which is essential as it provides quality assurance and control for research data; and the Integrate stages which are essential for combining data from different datasets. These stages are not in DCC’s model. In contrast DCC Curation Lifecycle has a focus in curation lifecycle (DCC, 2019) and to complement this, the “curate and preserve” is a full lifecycle activity. The Appraise & Select and the Reappraise activities identifies data that should either disposed of or selected for reuse and long-term preservation. This process is done on either new data or preserved data.

Overall, DCC Curation lifecycle provides comprehensive activities which will enable successful data curation. While DataONE Data lifecycle model is a simpler model to use and have some essential stages missing in DCC Curation lifecycle, I consider the DCC Curation lifecycle as a better model for the purposes of managing research data.

### Data Lifecycle Stages

For the purpose of this assignment, I have chosen to DataONE Data Lifecycle as a model for outlining how the ideas, techniques, tools and formats learned from the course could be used at each stage of the lifecycle.

#### Plan:

Planning is typically the initial data management activity because it helps organise all the subsequent stages of each digital object throughout its lifecycle. It is important to consider domain specific and funder’s standards and policies at each stage of the model.

Plans should answer questions about what data will be created, how it will be described and stored, publication details and how it will be shared, and all these should be documented in a formal document called Data Management Plan (DMP). The DMP is becoming a research proposal requirement for most funders as a way of demonstrating and ensuring data safety, preservation and sharing during and after the research is complete.

There are tools, templates and guides available to help develop a good Data Management Plan, such as DMPTool (DMPTool, 2019) and DCC’s DMP online (DMPOnline, 2019)

#### Collect:

At this stage, digital objects are generated, stored and organised in a selected data format using a specific file system. This stage helps in thinking about how best to organise these objects and choose the appropriate data format to be used for organising and storing the collected data for long-term preservation.

Standard open formats have established long-time support and for this reason they should be used to store data to ensure better portability and longevity, whereas proprietary formats are to be discouraged for long-term data preservation because they risk losing the possibility to interpret the data, for example, if the company closes for bankruptcy. There are a variety of standard formats to choose from, most are portable and have rich API libraries, some of them are general, such as CDF, XML, HDF5, PDF and CSV, and some are particular to a given domain, such as netCDF, a standard format for environmental science community. However, knowing which standards are used for the discipline of interest is essential for making right decisions when choosing a standard format.

#### Assure:

In this stage, it is important gather as much documentation of data and procedures as possible to ensure data integrity. Quality control and assurance procedures and measures are essential for this stage as they ensure that data collection goes smoothly, and that data can be validated at any time after the research project ends. Audit Trails for the database, algorithm/code testing data and calibration data are some of the important data to be collected at the time of data creation.

#### Describe:

Accurate and consistent description of data is vital for data users, especially for curation purpose, because it ensures that data are easily understood even when the owners are not available to explain things. This data about data is called metadata; and there are metadata standards that provide structure and consistency to data documentation which results in data being discovered, accessed, understood and re-used.

When creating metadata, it is imperative to use meaningful variable names, provide record units, provide a description for technical terms for research data and make the metadata accessible by storing it in the data location. This will facilitate for data discovery.

Dublin Core Metadata Initiative and Metadata Encoding and Transmission Standard (METS) (Loc.gov, 2019) is one of the many metadata standards available to document data. Each standard has a different focus, and it is not easy to make a choice. Organisational guidance, coupled with data types, field of practice and motivations for using medatada should help in choosing and justifying the appropriate metadata standard.

#### Preserve:

There are multiple ways in which data can be lost or corrupt; for example, storage device failure, natural disasters, accidental deletion and cyber-attacks. At this stage, it is crucial to think about backup and replication of the data to reduce the risk of losing them.

It is important to follow your institution’s policy in making sure all data are backed up and secure. Nowadays, most institutions have IT Disaster Recovery Plans (DRP) as part of their Business Continuity Planning (BCP).

Data backups should serve to restore data if needed, perform re-analysis for validation of data products or re-use in future.

Things to consider when backing up data are:

When choosing storage media, consider the capacity, obsolescence, susceptibility, cost and viability of the medium. For example, Liner Tape Open, DNA storage. Also consider backup frequency, strategies, media, locations.

Identify data sensitivity and determining the security and privacy to avoid the risk of losing data confidentiality and integrity. This will also ensure better decisions on data sharing (Mantra.edina.ac.uk, 2019).

Decide on which data to retain, share and preserve (raw data, clean databases, data products, documents, software and algorithms, results). Data that cannot be regenerated, for example are weather data, are of high value and should be preserved. Most institutions have a required length of time which data should be preserved. When preserving data, store them together with the metadata.

#### Discover:

Data sharing is one of the requirements for most funders and scientific research community. Data is shared through open repositories. Ensuring that the shared data is discoverable is key to data sharing. Persistent Identifiers add significant value to the data as they uniquely identify data, and make it addressable by facilitating discovery of data, citation of data and linking it with other data such as associated artefacts institutions and the researchers. PIDs enable originality of research and ensures that data objects to be traced back to your research data when your data file is combined, with either your own data files or other investigators’ data. There are different PID systems, for example, Handle system and Digital Object Identifier (DOI), and depending on the repository or service used to upload data, you will get a PID that is supported by the service.

A good description of the data provided in the Description stage will facilitate data to be searchable and understood.

#### Integrate:

Combining datasets to create new insights is done at this stage. Good metadata are essential as re good data integration tools. When combining data remember to check if there are licensing conditions restrictions. The integration of multiple data sets from different sources requires that they be compatible. Methods used to create the data should be considered early in the process, to avoid problems later during attempts to integrate data sets.

#### Analyse:

At this stage, new data is derived from existing ones by using computer software, such as Python and R, to apply the algorithms for deriving the data. This process may include analysing digital sensor data, simulation input, or re-analysing integrated third-party data. Documentation of the existing data, processes, tools, algorithms and any assumptions made in the analysis is essential to ensure validity and re-use of the results.