Theoretical Foundations for Updating DCC Curation Lifecycle Model

Caihong Huang, MLIS Class of 2019, huangc34@uw.edu Advisors: Sayeed Choudhury, Johns Hopkins University; Carole Palmer, University of Washington

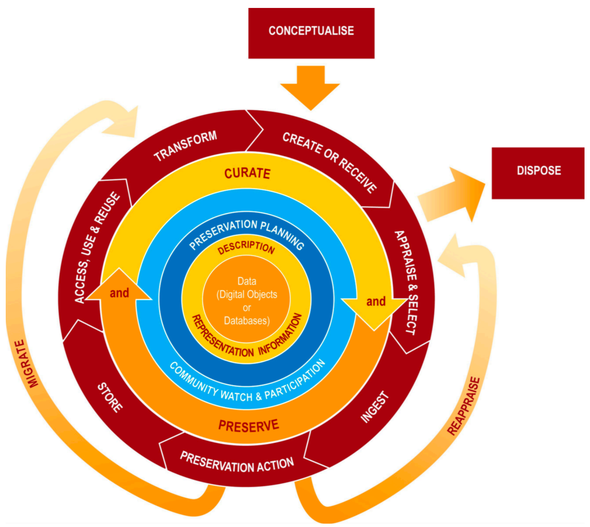
Information School UNIVERSITY of WASHINGTON

Alfred P. Sloan FOUNDATION

OVERVIEW

The DCC Curation Lifecycle model is a conceptual model which provides a high-level overview of the lifecycle stages required for successful curation. In this project, I studied the needs and possible fields for updating this model through close examination of literature about the original model, relevant research and data lifecycle models, and interview materials. Based on implications of four most relevant lifecycle models studied, recommendations on transformation of model structure and content are made.

WHAT IS DCC CURATION LIFECYCLE MODEL?



Reference: Digital Curation Centre. (N.D). DCC Curation Lifecycle Model. Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle-model

- Created by Digital Curation Centre (DCC), UK, in **2009**
- Intended to provide a **high-level overview** of the lifecycle stages for successful curation
- Designed to facilitate the organization and planning of curation and preservation activities **within an organization**

WHY DOES IT NEED AN UPDATE?

- The DCC model is one of the most **important** and widely used models for data curation
- The model is **not definitive** and evolving
- The key element of the model, **data**, has changed dramatically since its creation in 2009
- **Little research** has been done in updating or extending the model
- The model is **implementation-agnostic** and its visualization is **conceptual**

> *"The next stage of the project is the development of domain-specific variations to help further contextualise training and resources."*
>
> -- Sarah Higgins, creator of DCC model, 2009

QUESTIONS & METHODOLOGY

- How to optimize the original DCC in consideration of the change of context and environment related to data?
- Given the impact of big data, how to develop domain-specific model variations for the emerging fields such as Machine Learning and Artificial Intelligence?
- How to restructure the model visualization for easier understanding and better support in implementation?

WHAT DOCUMENTS DID I STUDY?

- Literature of lifecycle models
- Documentation including the original paper and subsequent DCC reports
- Interview materials with DCC staff (collected by the advisor during his field trip to DCC)

FINDINGS

50+ research & disciplinary data lifecycle models

4

4 most relevant models:

- Records Continuum Model (RCM) *
- Data Life Cycle Laboratories (DLCLs) *
- Big Data Lifecycle Model
- Comprehensive Data Lifecycle Model (COSA-DLC)

Records Continuum Model (RCM)

Revised from Upward, F. (1996). Structuring the records continuum (Series of two parts) Part 1: post custodial principles and properties. Archives and manuscripts, 24(2), 268.

Data Life Cycle Laboratories (DLCLs) Model

Refence: van Wezel, J., Streit, A., Jung, C., Stotzka, R., Halstenberg, S., Rigoll, F., ... & Giesler, A. (2012). Data life cycle labs, a new concept to support data-intensive science. *arXiv preprint arXiv:1212.5596*.

RECOMMENDATIONS

Transformation in Structure

- Object-oriented -> **Process-oriented approach**
- Towards a research lifecycle model: including **pre-research** (e.g. planning) and **post-research** (e.g. publishing, teaching) phases

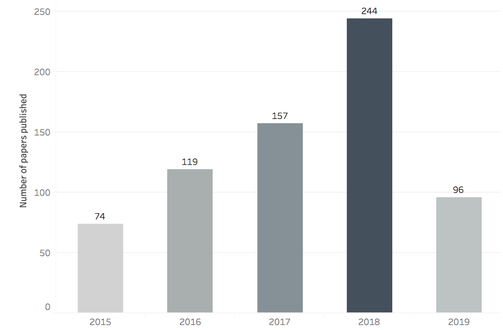
Change in Content

- Incorporating **societal, technical, disciplinary culture** besides organizational environment

NEXT STEP

- Investigating the possibility of incorporating **data science principles** especially **fairness and transparency in machine learning**

Numbers of papers about algorithm fairness published since 2015



Statistics collected from Web of Science on May 1, 2019 with the following search query: TS= (machine learning AND data) AND TS=(fairness OR bias)

ACKNOWLEDGEMENT

This work is part of the "DCC Lifecycle Curation Model 2.0" research project supported by Alfred P. Sloan Foundation Grant under Grant Number G-2018-11214. Sincere thanks to my advisors, Sayeed Choudhury and Prof. Carole Palmer, for their continuous support throughout the Capstone project.