**MidWest Data Libs - Data Curation Discussion Outline**

Outcomes:

Group members feel comfortable describing data curation and why it is important when talking with a researcher.

Realization that while public access policies may require data to be shared, it won't necessarily be worth curating or retaining after the stipulated time. In some cases, data documentation will be enough,

1. Data Scenario Group activity - Put a sticker on the scenarios you think should be saved. (10 minutes)

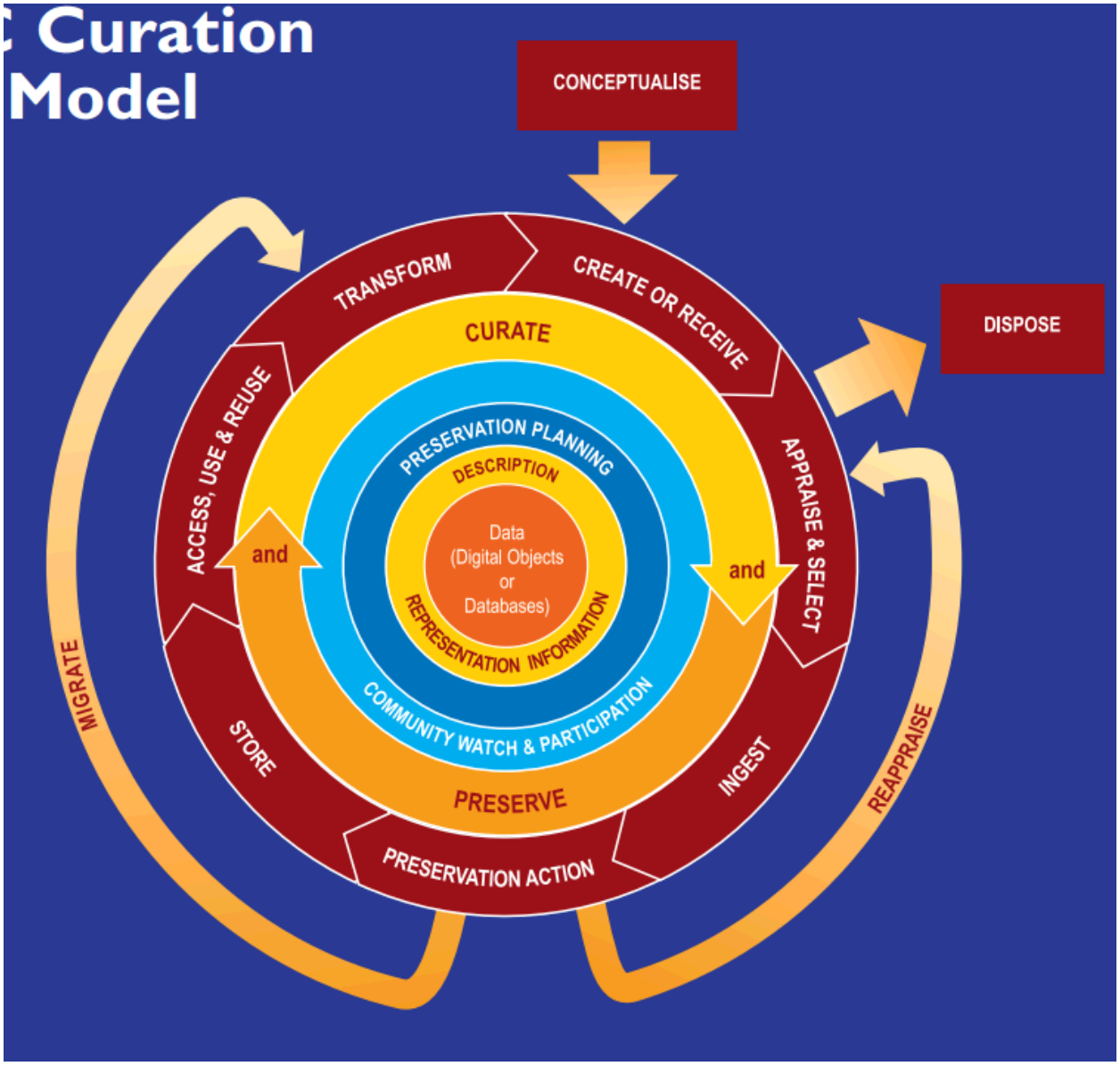
- 10 notebooks from Professor X, Nobel Prize winner
- csv files for survey conducted by a grad student for PhD. Survey topic: favorite apps for students. Funded by NSF.
- spreadsheet of arrival dates of migratory birds for last 20 years, partially NSF funded, professor is retiring
- DNA sequences from gut microbiome project.
- Collection of family letters from Library trustee/big donor..
- Interview recordings from ethnographic study in Liberia
- Satellite improvement project, NASA grant (ITAR Sensitive data scenario)
- Archeological find photographs and maps of survey area
- Recordings of focus groups of diabetes patients discussing where they find health invormation.
- Colection of glazed certamic tiles, numbered, with correcponding formulation spreadsheet. Used for teaching, but program has stopped.
- 3-D printer files of arrowhead that is said to have killed the first white settler in Jamestown (artifact still on display in museum).

2. Data Curation Defined. (5 min)

Wikipedia: **Data curation** is a term used to indicate management activities related to organization and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation. https://en.wikipedia.org/wiki/Data_curation

from Carlson  “focus on the core elements of data curation: planned management over time, availability for discovery and re- use, and adding value to enable or further usage,” http://www.emeraldinsight.com/doi/full/10.1108/00907321211203603

from The iSchool at Illinois: “Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.” http://www.lis.illinois.edu/academics/degrees/specializations/data_curation




DCC lifecycle http://www.dcc.ac.uk/resources/curation-lifecycle-model

3. Source of Data (relates to Uniqueness and Non-Reproducibility in 4)(5 min)

Observational

- Captured in real-time, typically outside the lab
- Usually irreplaceable and therefore the most important to safeguard
- Examples: Sensor readings, telemetry, survey results, images

Experimental

- Typically generated in the lab or under controlled conditions
- Often reproducible, but can be expensive or time-consuming
- Examples: gene sequences, chromatograms, magnetic field readings

Simulation

- Machine generated from test models
- Likely to be reproducible if the model and inputs are preserved
- Examples: climate models, economic models

Derived / Compiled

- Generated from existing datasets
- Reproducible, but can be very expensive and time-consuming
- Examples: text and data mining, compiled database, 3D models

4. Appraisal of Data for Curation and Long-Term Preservation (5 min)

a. Relevance to Mission

Does the item content fit the organization's remit and priorities, including any legal requirement to retain the item beyond its immediate use?

b. Scientific, Social, Cultural, Historical Value

Is the item scientifically, socially, or culturally significant? You need to anticipate future use, from evidence of current value.

c. Uniqueness

To what extent is the item the only or most complete source of the information it contains?

d. Potential for Redistribution

Is the item authentic (i.e. what it says it is), with integrity (unchanged), and usable? Does it meet IPR and ethical requirements?

e. Non-Replicability

Would it be feasible, or financially viable, to replicate the item?

f. Economic Case

What is the cost of managing and preserving the item, and does the value of the item justify this cost?

g. Full Documentation

Is the metadata and contextual information needed to find, access and reuse the item comprehensive and correct?

from NECDMC, Module 7 activity, http://library.umassmed.edu/necdmc/modules based on Whyte and Wilson http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

5. Small group discussion: Questions for discussion. You can stick with one area if the group is agreeable, or switch when 10 minutes is up. (30 minutes)

What are the legal, institutional, or funder requirements for curation and/or preservation?
How will Public Access Policies, as per OSTP, impact data curation?

Does your institution/library have a data curation policy?
What do you think should be included in a data curation policy?

What kind of documentation do you need for good curation?
What information do you need from the data creator/data steward?

6. Report Back (starting about 10 am)

What legal or funder issues must be considered?

- Data ownership policies are not the same as data curation policies.
- Whose responsibility - University, funder, or owner?
    - money is often the driver for policy
    - or a legal snafu creates awareness of need for policy
- IP complications--sometimes there are IP reasons that you don't have to/want to share, but you still need to curate
- Sensitive data
    - patient data, classified, commercial
    - should something be archived but not made available
- Starting to see requirements from journals as well
- Role of the library as steward
- We need more reward mechanisms for data sharing
    - need more carrots (have stick)
- Differing stakeholder requirements--who wins when the funder, journal, and institution have different policies?
- Who ensures compliance?
- IRB approval
- Dealing with older datasets
- State Board of Regents with their own requirements
- Embargo
- Other repository requirements
- Co-researchers with other mandates
- publisher requirements

What needs to be in a good data curation policy?

- how much data

- how long stored
- who is responsible for managing
- Input and cooperation from people throughout the library and campus
- campus buy-in
- Digital projects policy could be adapted
- There has to be a threshold of “we will not take ‘x’”
- Description
- Retention and discard policy
- Review process
- Talk to your archivists!! They’re already thinking about this.
- Creator right to withdraw? (from a repository)
- Needs to come back to mission of repository
- Library needs to have authority to make decisions about the data (similar to donor agreement) Trying to get other/supporting/supplementary files
- format preferences
- Determining value of data
- DOIs
- Requiring creators to tell us when data changes
- Withdrawal policies
- Ensuring depositors understand policy

What do you need to properly curate a dataset?

- time!
- planning
- documentation
    - readme, data dictionaries, codebooks, lab notebooks
- money
- storage space
- IRs are not always a great place for data because of discoverability and existing infrastructure
    - other organizations could be the curators
    - disciplinary repositories
- How to make sure depositor has the rights to actually deposit
- What is the significance of the data?
- Determining use and reproducibility
- Policy to determine retention after initial specified time
    - involve liaisons who have collection development skills
- Provenance of the materials
- What system and version was used