Three Perspectives on Data Reuse: Producers, Curators, and Reusers

Ixchel M. Faniel, Ph.D.

Associate Research Scientist
OCLC Research
fanieli@oclc.org

Elizabeth Yakel, Ph.D.

Professor
University of Michigan
yakel@umich.edu

Twitter @DIPIR_Project

OCLC® The world's libraries. Connected.

M | SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Archaeological Practice

- Changing nature of research questions

- More reliance on documentation as artifacts can often not be removed from sites
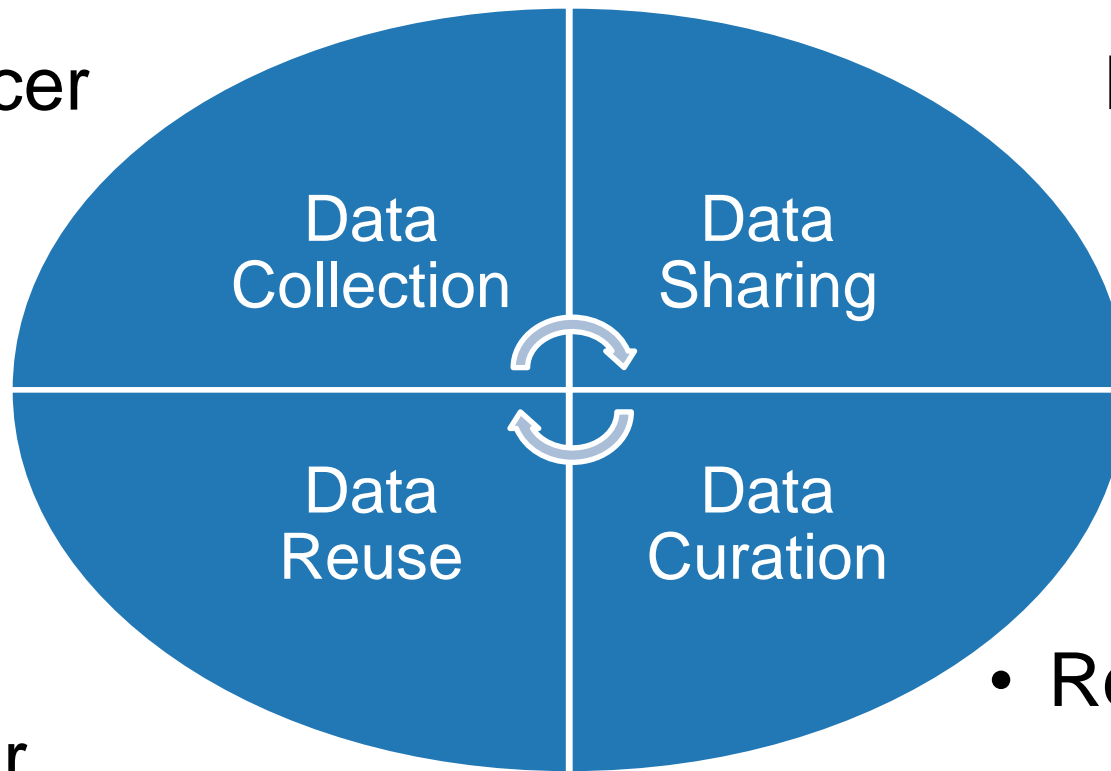
- Data reuse tradition mixed

http://cosmiclog.nbcnews.com/_news/2013/04/25/17914746-where-did-maya-culture-come-from-archaeologists-dig-into-tangled-roots?lite

The data lifecycle from 3 perspectives

- Data Producer

- Data Producer

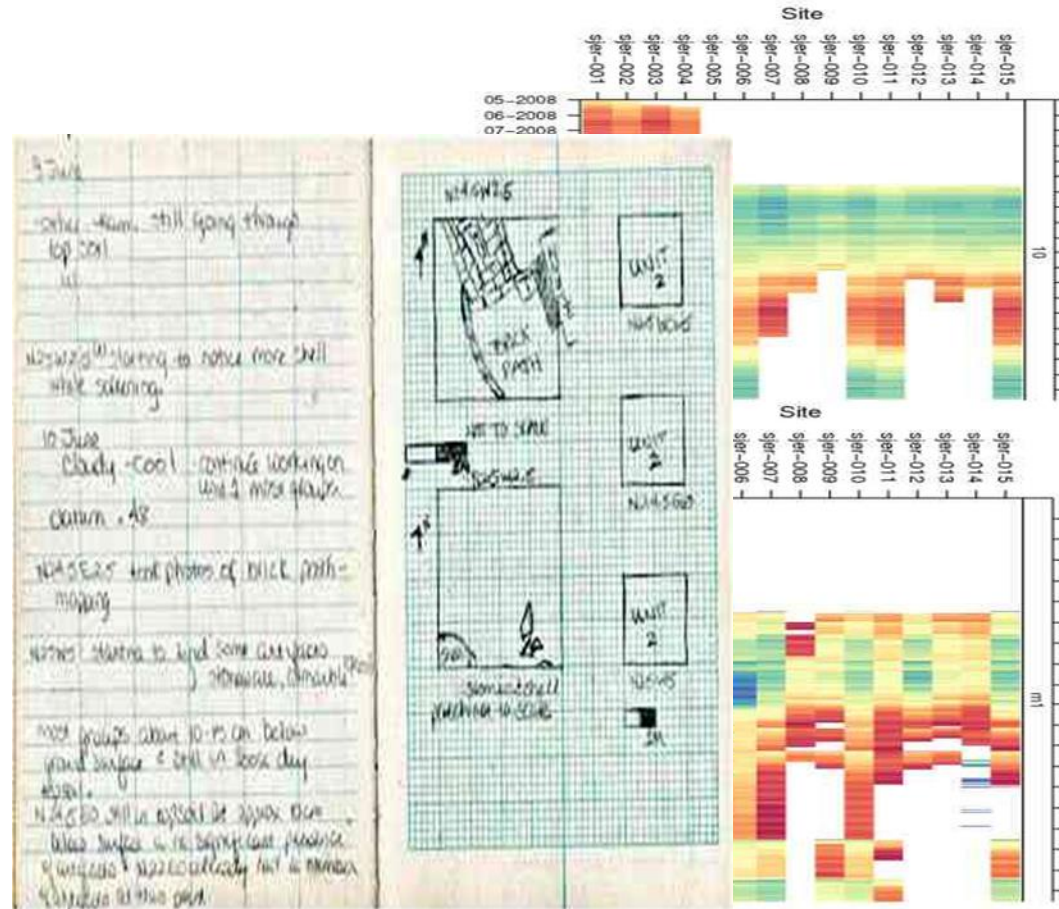Data Collection | Data Sharing

Data Reuse | Data Curation

- Data Reuser

- Repository Staff

Research Question

How do actions in one part of the data lifecycle create challenges or facilitate work at another point in the lifecycle?

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Underlying Case

Research Design

- Data collected over 1.5 years (2012 – 2014)

- 9 data producers

- 2 repository staff

- 7 data reusers

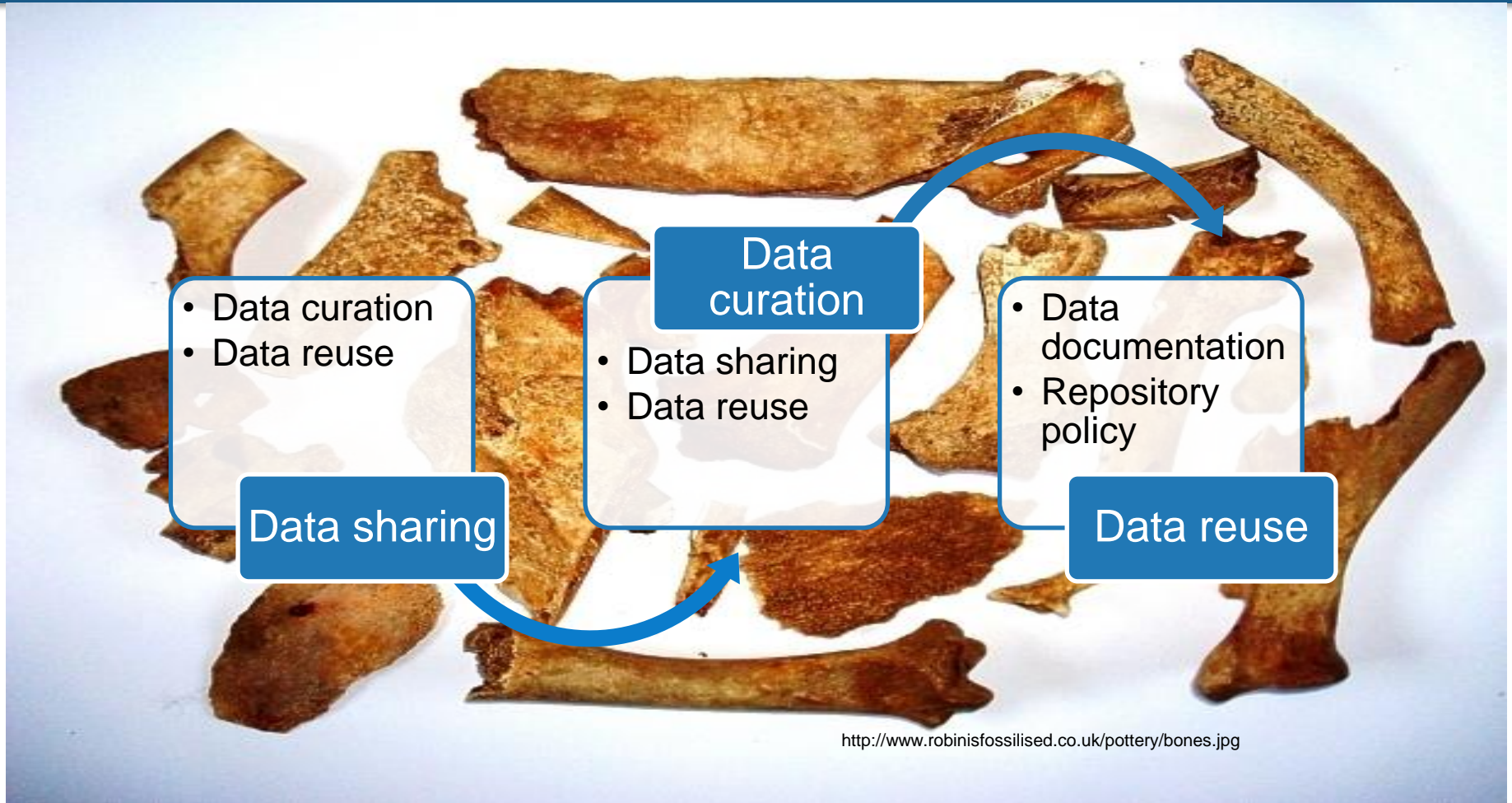- Culminated in several conference presentations and 1 publication

Data Collection

- Interviews

- Email exchanges with data producers, repository staff, and data reusers

- Focus group

- Observations at conference presentations

Data Analysis

- Code set developed and expanded from previous interview protocol

Findings

- Data curation
- Data reuse

Data sharing

Data curation

- Data sharing
- Data reuse

- Data documentation
- Repository policy

Data reuse

http://www.robinisfossilised.co.uk/pottery/bones.jpg

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Findings

STANDARDIZATION

Everything is turned into intentionally difficult codes... hundreds of lines …you have to translate, such as what is a 1 or 1.5, what's that mean? **It was really important to streamline that translation process** (Data Producer 3).

INTEGRATION

[Repository staff] did a great job integrating it…I don't know what kind of format the datasets had before they got **integrated…Integrated to the extent that it's comprehensible, but I believe there was a lot of work because I know that different zooarchaeologists did things in a number of different ways** and coming from different traditions. (Data Reuser 9).

SAVING TIME AND EFFORT

It was great that **[repository staff] did a lot of the cleaning…you can't do that on your own**…you can do it,…you will have to change a lot to integrate it into yours [database] and **that will take a lot of time** (Data Producer 4).

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Findings

DATA COLLECTION PRACTICES

I just keep getting stuck on exactly what I am supposed to do with my excel spreadsheets and with issues like that fact that in some cases I have sampled assemblages for just caprine **specimens…so those data cannot be used** to calculate NISP [Number of Identified or Number of Individual Specimens] frequencies for the total site (Data Producer 3)

DATA DOCUMENTATION: UNDERSPECIFIED STANDARDS
We do have tooth wear data, but it just wasn't in a format that could be clearly integrated. Some sites have clear A, B, C phases, while others have number codes by tooth. **We could provide all of that to the analysts, but it will be a lot of columns of pretty disparate data** (Repository Staff 2).

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Findings

DATA FORMAT

In addition to the project info attached, I'll need your datasets, preferably as Excel tables. **If you export them from a database, please indicate the key(s) for us so we can stitch them back together again!** Since you've already published on these data, feel free to send the entire datasets (Repository Staff 2).

DATA CONDITION

It took 10 times longer to deal with those [coded] datasets but if it helped the researcher to get their stuff in … (Repository Staff 2)

http://visiblepast.net/see/wp-content/uploads/2011/03/fig3.jpg

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Findings

<u>DATA PRODUCERS' SELECTION</u>

I did think quite carefully about…those …big subjective descriptions we write about the units...before including that, but I decided to... I mean obviously different people write in different styles. It's not exactly like your personal diary entry, but it... Can be quite informal. I always write them in quite formal prose myself, but some people are a little bit less formal...**I couldn't really be sure if people …would necessarily want them out but they are an important part of the data set** (Data Producer 10).

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Findings

REPOSITORY PROCEDURES

There are some inherent issues with CSV as it is a very simple text-based format…the simplicity is why it is preferred for interoperability and longevity …**we need to give users a few tips on working with CSV**. I'm also looking into other open spreadsheet formats, but Excel…gets these wrong (Repository Staff 1).

DATA PRODUCERS' DOCUMENTATION

In my case **it changed already**…**I had a completely different recording system for [teeth]**…just using…Payne… (Data Reuser 6).

Two things that I would now do differently: One of them is writing down with my data what exactly all those criteria are…I always kind of had a few notes …**but…writing down more systematically exactly what those criteria have been**. And second one is **just dropping numeric codes**, not doing …numeric codes anymore (Data Reuser 10).

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Tightly vs. Loosely Coupled Activities

- The data lifecycle is a tightly coupled activity

- Archaeological data management is loosely coupled
  - Don't think about data sharing or reuse outside and sometimes inside the group

- Consider all stages of data lifecycle during data production and documentation

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Short and Long Term Benefits

- Individual rewards
    - Data reusers: Publication of the article
    - Data producers: Data archiving / data publication
    - Repository staff: Better data submitted
- Science: gaining new knowledge
    - Data producers
    - Data reusers
    - Designated community

OCLC The world's libraries. Connected.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Short and Long Term Benefits

- Persistence: Data now in repository that anyone can use

- Repository staff: Building repository reputation as a trusted source

- Designated community: Increasing the visibility of new archaeological data sharing and reuse practices

OCLC· The world's libraries. Connected.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Acknowledgements

- Post-doctoral researcher: Anthea Josias, Ph.D.

- Doctoral students: Rebecca Frank, Adam Kriesberg

- Study participants for allowing us to collect data for this research.

Questions?

Ixchel M. Faniel

fanieli@oclc.org

Elizabeth Yakel

yakel@umich.edu

OCLC· The world's libraries. Connected.

M | SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN