Other People's Data

Elizabeth Yakel
University of Michigan
ACA@UBC 4th Annual International Symposium
University of British Columbia
February 17th, 2012

Open Context

ICPSR A PARTNER IN SOCIAL SCIENCE RESEARCH | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
University of Michigan

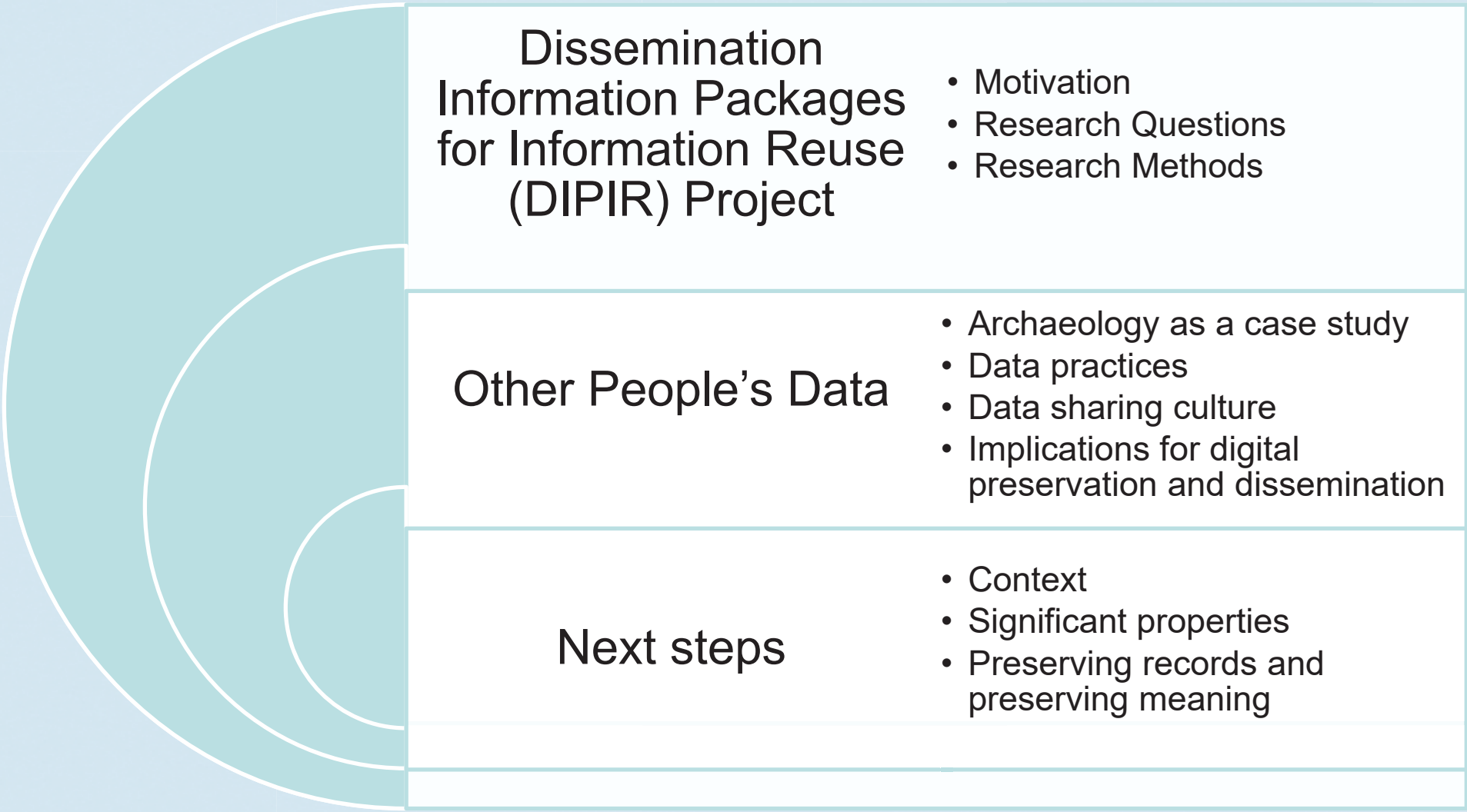
INSTITUTE of Museum and Library SERVICES

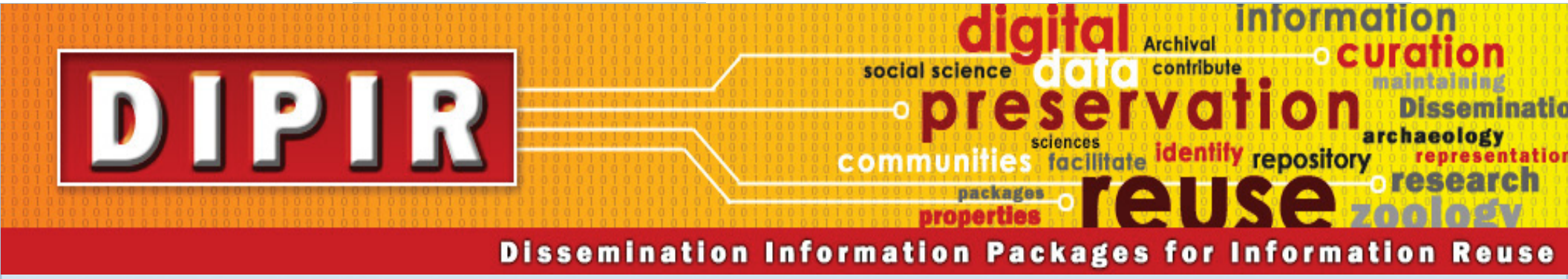
OCLC
The world's libraries. Connected.

school of information
University of Michigan

OCLC
The world's libraries. Connected.™

Agenda


Dissemination Information Packages for Information Reuse (DIPIR) Project
• Motivation
• Research Questions
• Research Methods
Other People's Data
• Archaeology as a case study
• Data practices
• Data sharing culture
• Implications for digital preservation and dissemination
Next steps
• Context
• Significant properties
• Preserving records and preserving meaning



- Institute for Museum and Library Services (IMLS) funded project led by Dr. Ixchel Faniel & Dr. Elizabeth Yakel, 2010 - 2013.
- Studying data reuse to identify how contextual information about data that supports reuse can best be created, captured, and preserved.
- Three communities:
  - Quantitative social scientists, archaeologists, and zoologists.
- Intended audience
  - Researchers who use secondary data and the digital archivists, curators, repository managers, data center staff, and others who collect, manage, and preserve digital information.
- Significant properties

school of information
University of Michigan

OCLC
The world's libraries. Connected.™

Research Team

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
A PARTNER IN SOCIAL SCIENCE RESEARCH

OCLC

school of information
M

Open Context

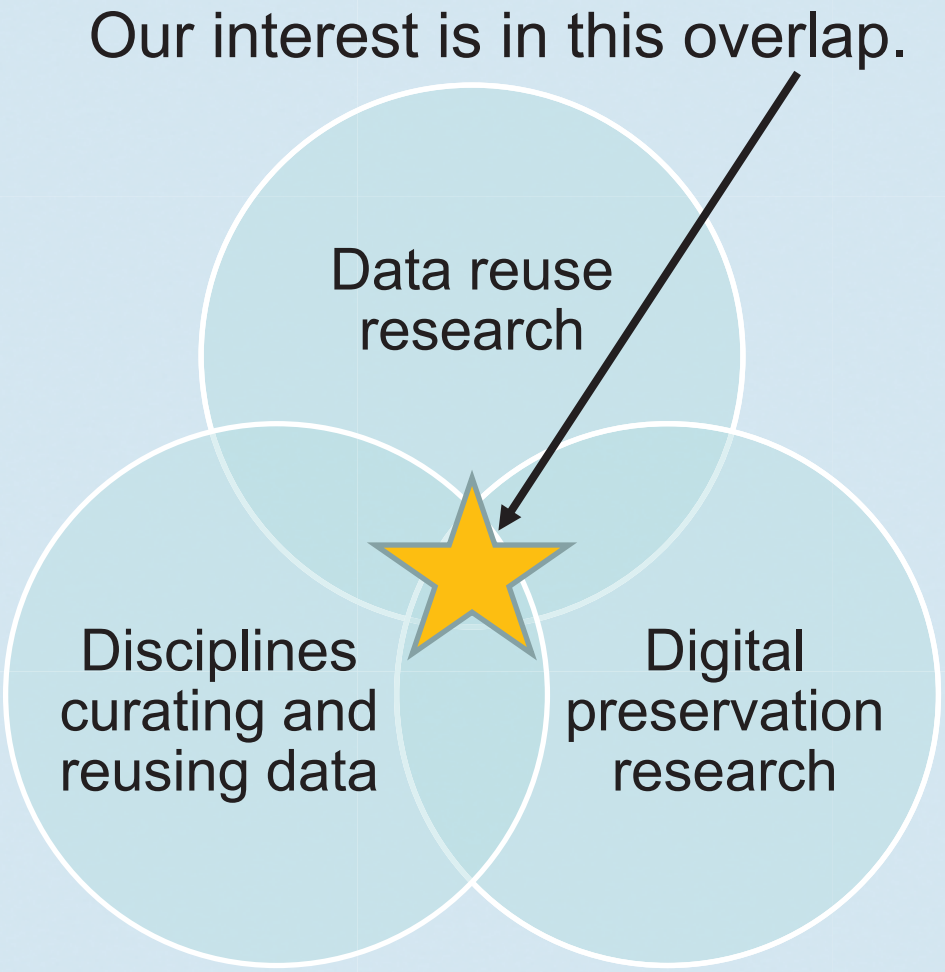
DIPIR Project

- **Nancy McGovern** — ICPSR/MIT
- **Elizabeth Yakel** — UM — School of Information
- **Ixchel Faniel** — OCLC
- **William Fink** — UM — Museum of Zoology
- **Eric Kansa** — Open Context

Research Motivation

Two Major Goals

1. Bridge gap between data reuse and digital preservation research
2. Determine whether reuse and preservation practices can be generalized across disciplines

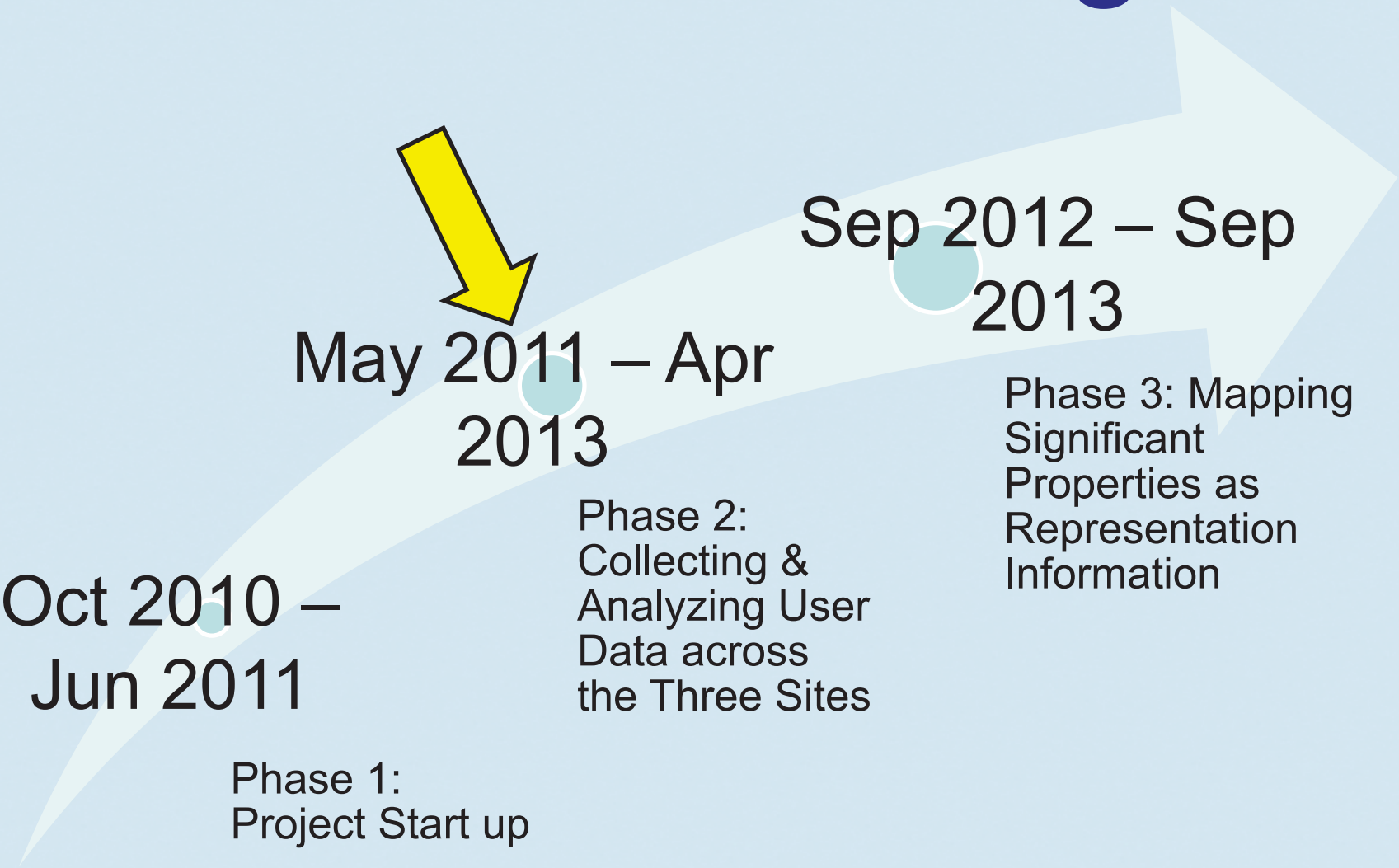


Research Questions

1. What are the significant properties of data that facilitate reuse by the designated communities at the three sites?

2. How can these significant properties be expressed as representation information to ensure the preservation of meaning and enable data reuse?

Research Design



Research Methods

- ICPSR
  - Staff interviews
  - Interviews with data reusers
  - Survey
- UMMZ
  - Staff interviews
  - Interviews with data reusers
  - Observations
- Open Context
  - Staff interviews
  - **Interviews with data reusers**
  - Web analytics

Open Context

Web-based research data publication

BETA

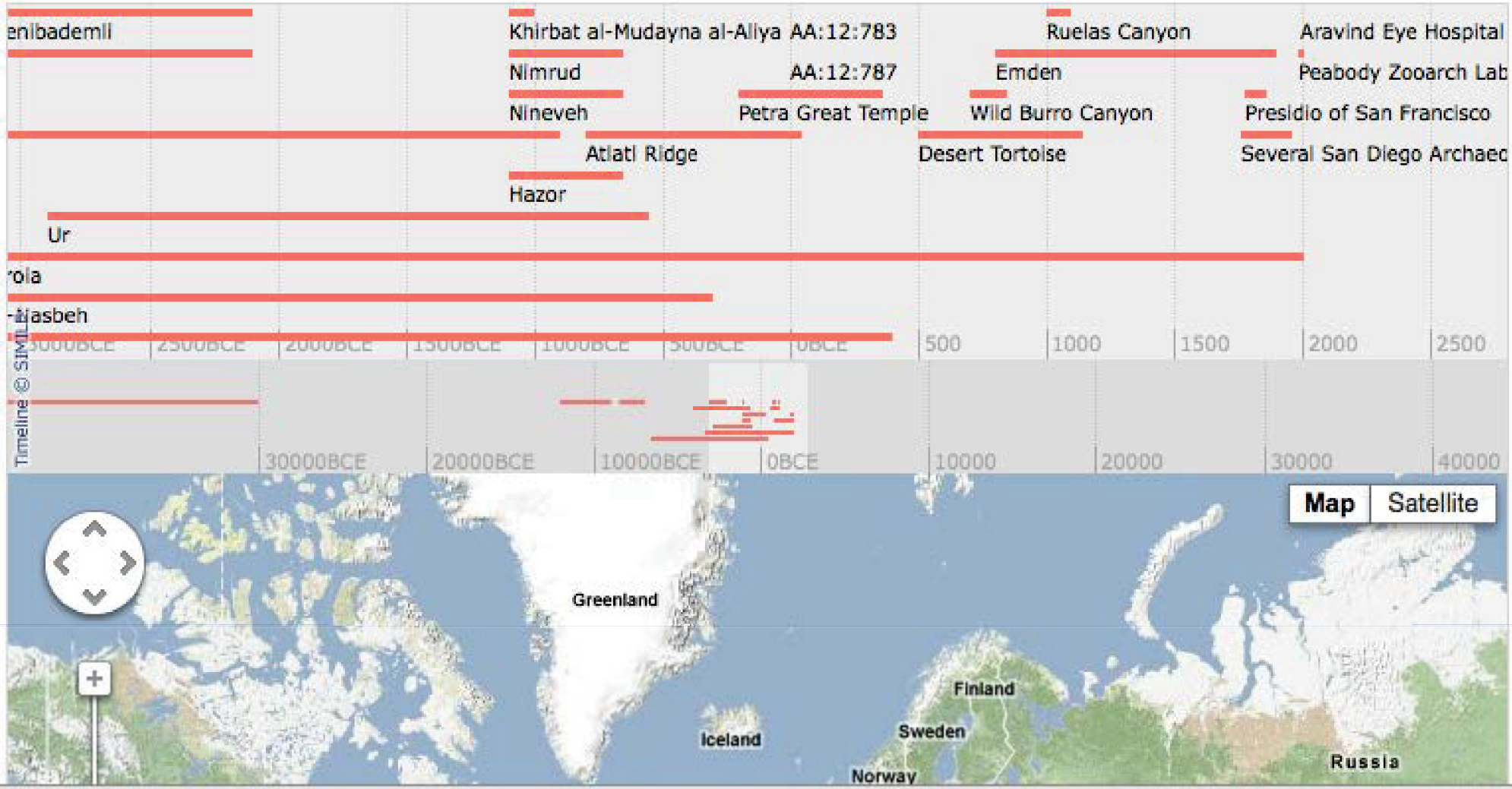
Search

GO

Home | About | Projects | Browse | Lightbox | Tables | Search | My Account

Welcome to Open Context

Use Open Context to discover, reference, and publish primary data and documentation collected in archaeology and other field sciences by professional researchers. *Scroll* through the timeline and map below to begin exploring data, images, reports and field documentation.



Open Context Interviews (n=20)

	Data Reuser	Curators	Expert
Male	11	4	11
Female	9	2	7

Semi-structured Interviews

- Background
- Data reuse
  - Actual experience, Critical incident
  - Aspirational
- Digital data repositories
- Data sharing

Data Collection Practices

- Interdisciplinary teams
  - Zooarchaeologist
  - Anthropologist
  - Metallurgy specialist
- Reliance on other team members to collect data for you
- Study duration long

To address large scale questions not just at one site but in the whole region that I'm now working with a lot more people. …in order to do anything useful or that's publishable in a prestigious journal…I'm sort of forced to cooperate with more and more people and share more data. This is sort of standard protocol, particularly, the data that's standardized as multiple researchers that's where you can do it, the most easily. (CCU13)

Data Sharing

- Emerging data sharing culture (proto-sharing)
- Lack of data sharing despite scholars sitting on decades of unpublished research
- Generational shift

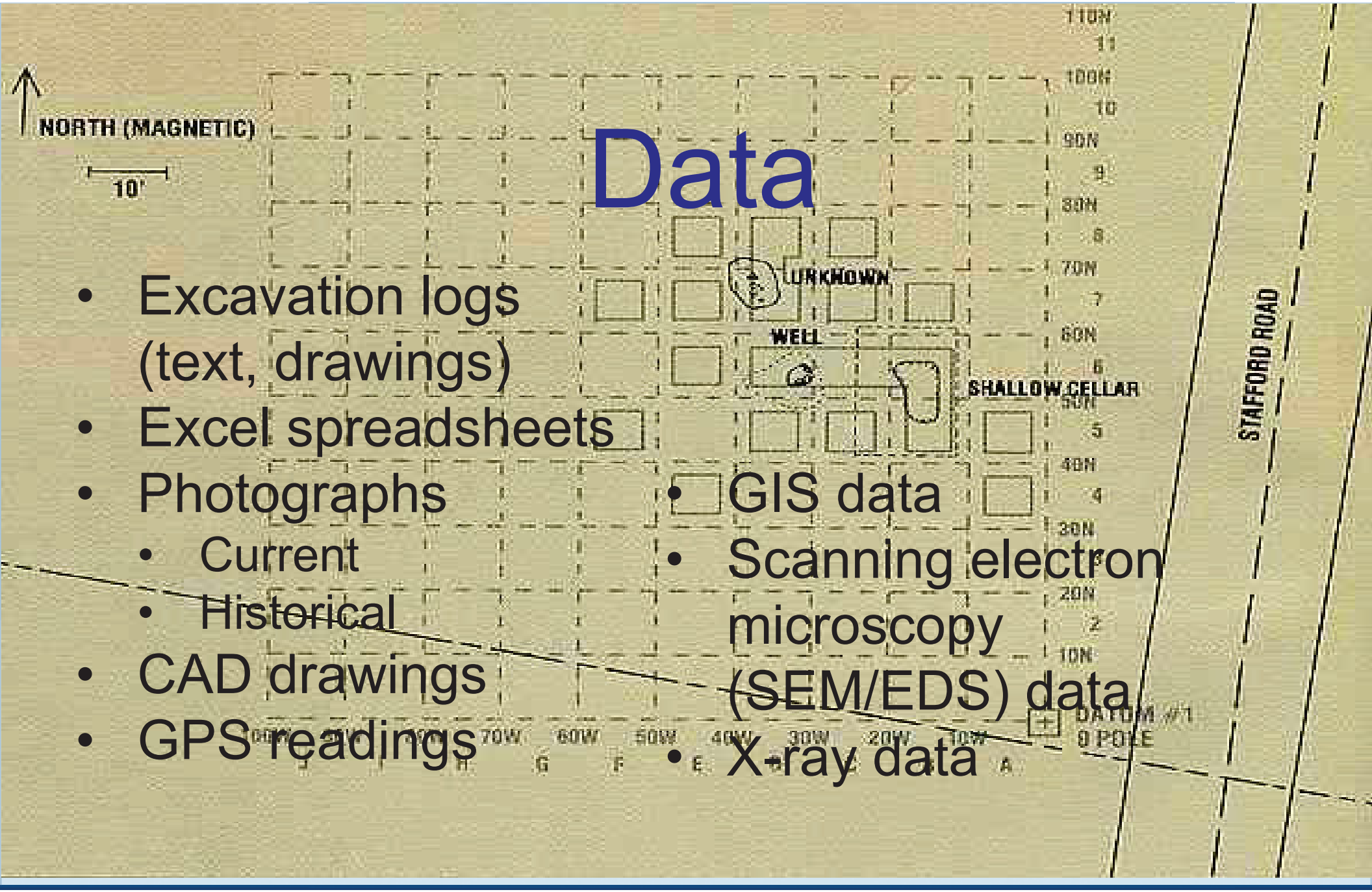
Data Documentation Practices

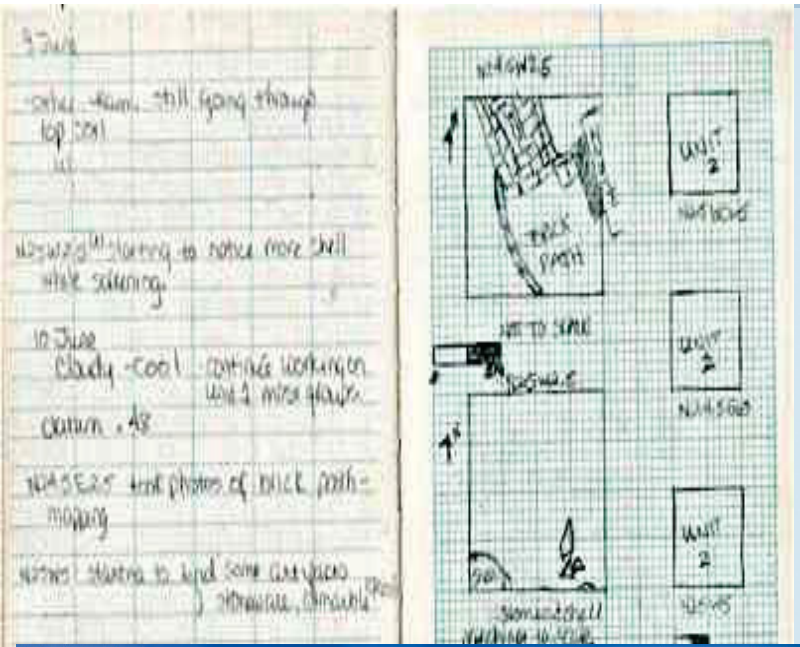
- More data collection; less collection of objects
- Individual systems of data collection
  - Developed over time
  - Handed down from mentors
- Some technological adoption
  - Excel over relational databases

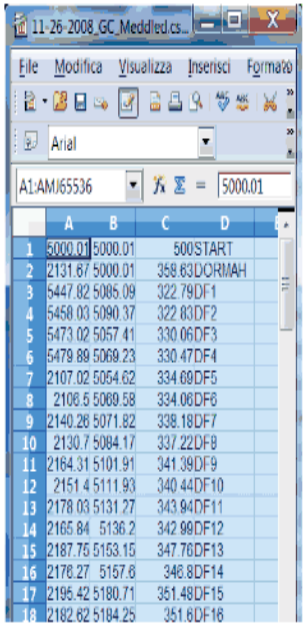
“I use an Excel spreadsheet…which I have inherited from my research advisers. Actually…my dissertation advisor was still recording data for each specimen on paper when I was in graduate school so that's what I started out doing and then quickly, I was like, "This is ridiculous.“… I just started using an Excel spreadsheet that has sort of slowly gotten bigger and bigger over time with more variables or columns…I've added …color coding…I also use…a very sort of primitive numerical coding system, again, that I inherited from my research advisers…So, this little book that goes with me of codes which is sort of odd, but for every one of my generation and earlier, it's like, who sort of learned under the American system of Near Eastern zooarchaeology, we, and all my fellow graduates, we all know that a 14 is a sheep.” (CCU13)

Data

- Excavation logs (text, drawings)
- Excel spreadsheets
- Photographs
  - Current
  - Historical
- CAD drawings
- GPS readings
- GIS data
- Scanning electron microscopy (SEM/EDS) data
- X-ray data

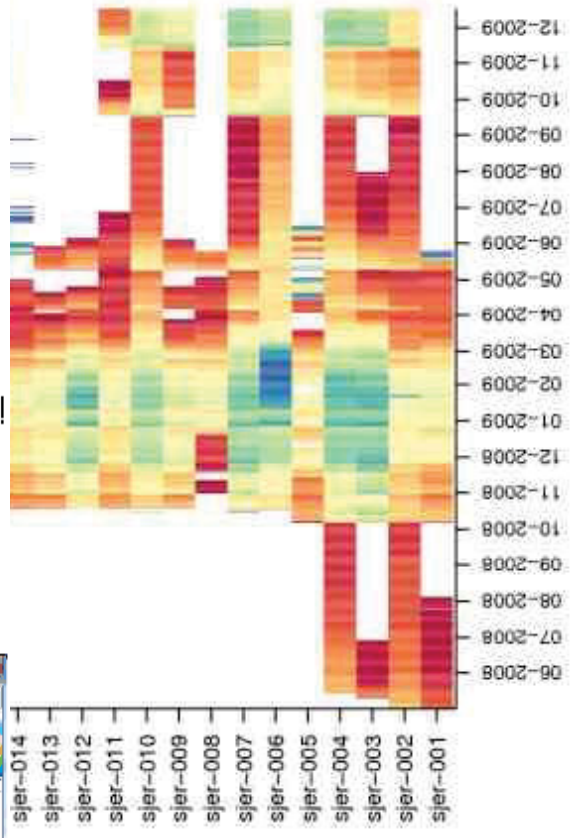


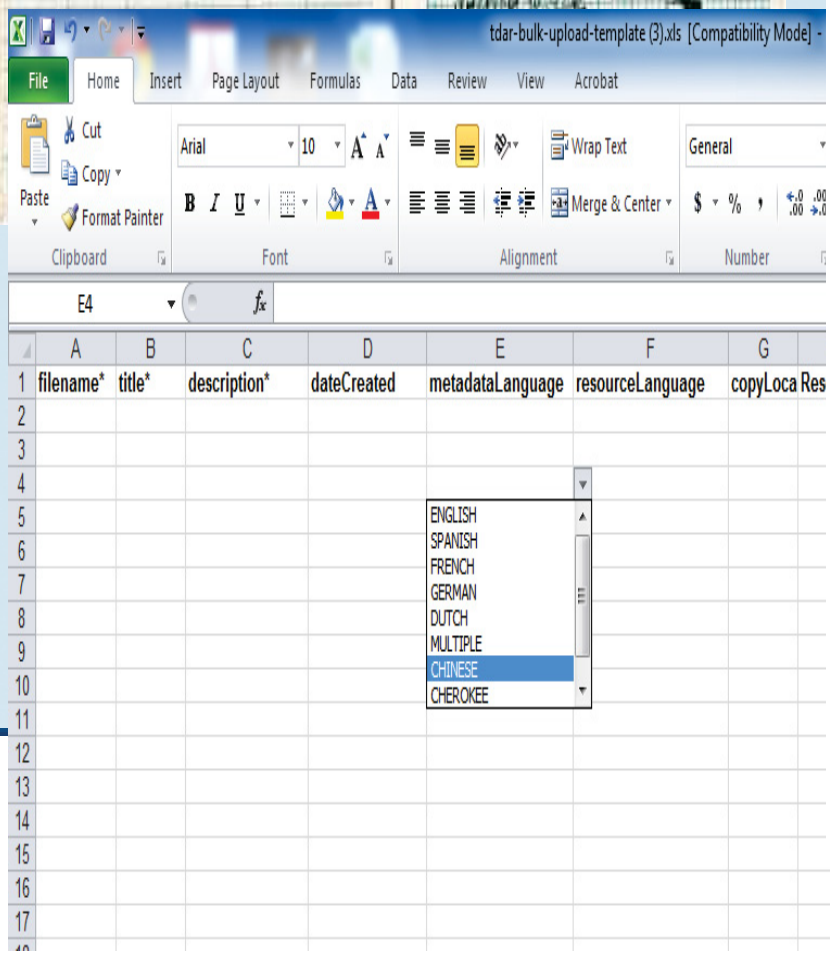




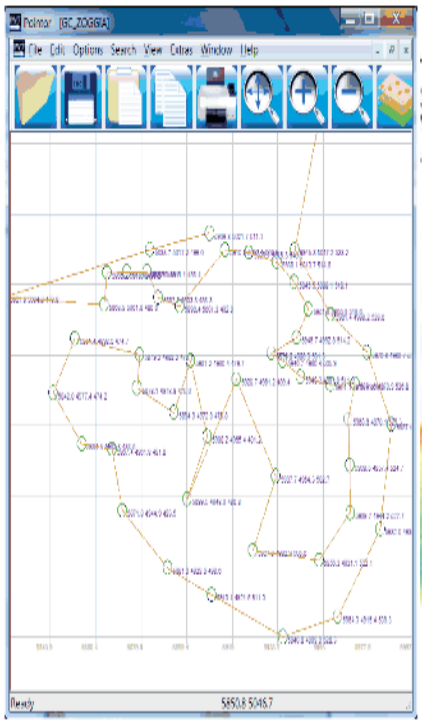
Spreadsheet
(Excel, OpenOffice etc.)

Use the Windows clipboard to transfer the point list to Pointor. Just Copy and Paste!

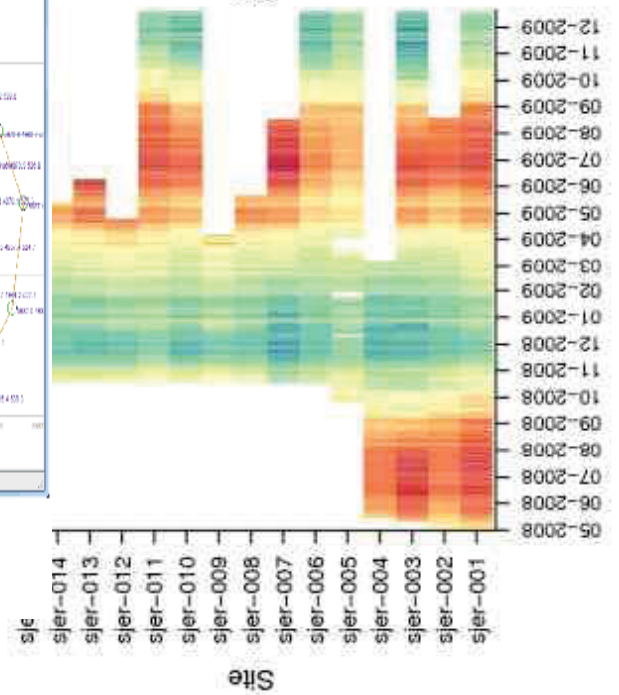




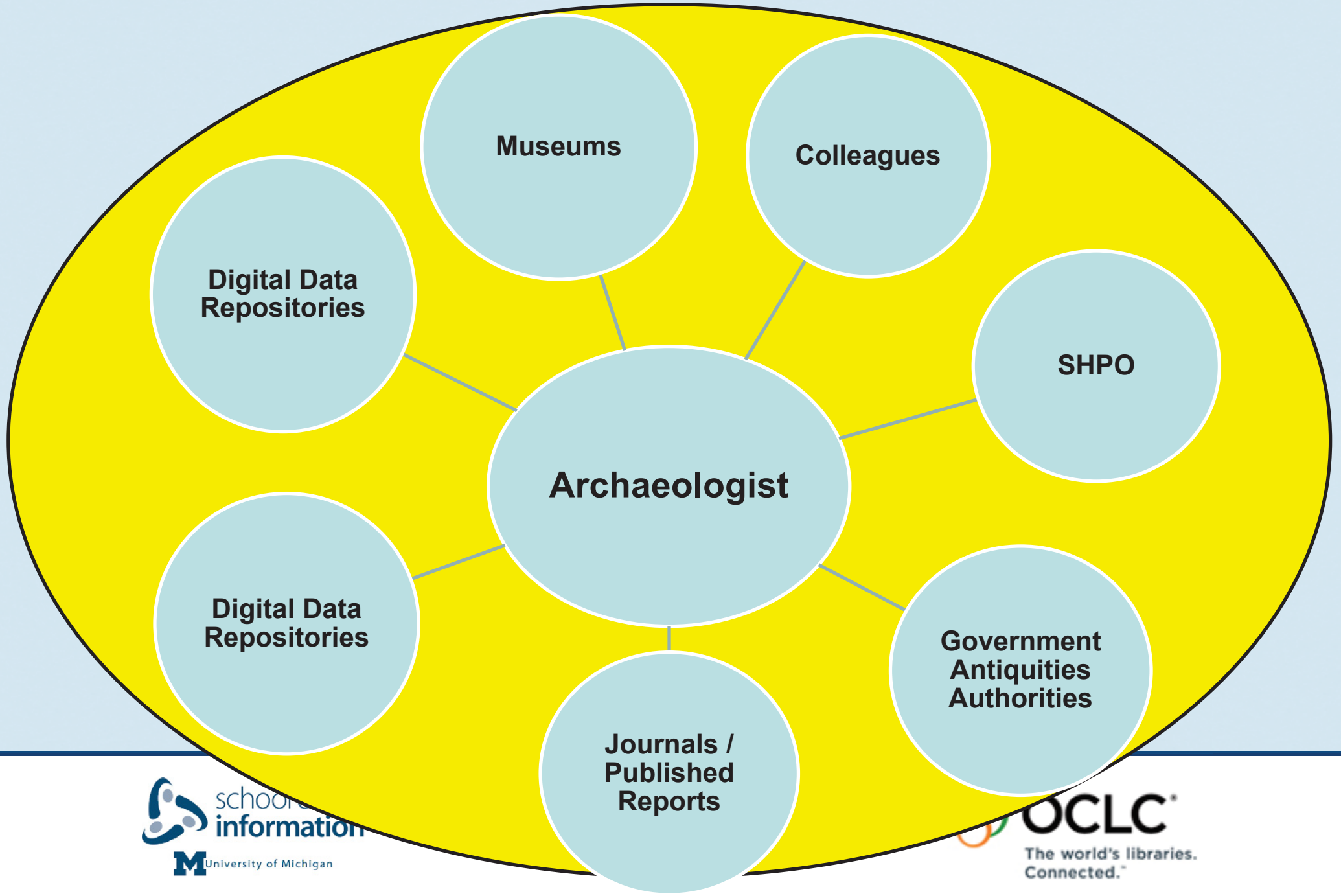
Pointor



Save the points (and lines and coordinates if you want) to a DXF file ready to be inserted into any CAD program.



Sources of Data



Data Reuse

- Qualities in the data
- Functionality of the data
- Preserving meaning
- Preserving evidence

Qualities in the Data

- Consistency
- Professionally stated (jargon)
- Depth of description
- Method transparent

Reuse Begins with Oneself

"I also transferred all the data, my own data, from paper to computer Excel that was my also first introduction to using worksheets. That was actually an eye-opener for me when back then for the first time because I didn't have any categorical approach there because it was very difficult to analyze that dataset. So that experience taught me to be very careful, taught me to be consistent and to be using standardized, systematic approaches to documenting my research and datasets." (CCU16)

"Recognizing the actual veracity of the data, it is, depending on the context, is something that has to be considered basically on the basis of professionalism in recording." (CCU12)

Functionalities of the Data

- Comperanda
- Reanalysis

Reanalysis

- “that was one of our most serious problems. So we did not have access to critical information, such as archeological contexts, excavation methods, sampling methods, even identification methods. We didn't know if the analysts actually used comparative collections or just publish manuals to identify specimens or how the samples.” (CCU16)

I had not used it so intensively. I was... I knew this in terms of data inaccuracy, it's one of the interesting things that... So I was doing a spatial analysis as well as a chronological one. I wasn't interested in some of the inaccuracies. I mean I had expected to go through and do some basic data cleaning, and I did and there were some obvious stuff that was easy to fix. But I was kind of surprised at the number of things where there was totally inaccurate data, like a site is not at all where it's supposed to be. The latitude and longitude are just completely off. And I thought that was interesting.

Preserving Meaning

- Strength of the preserved context
- Provenience

Contextualization

If haven't yet found anything where the data, …has actually been published, …my next options are to talk to the museums where these collections have been curated…That's often actually really hard to do because nobody seems to know. So, I talk to the museum... If this raw data was also curated at that museum, or if it's just a collection itself and then go from there. (CCU06)

Provenience

- What I do is I look at these collections, these collections of animal bones, identify them to skeletal elements, and to taxon as best as possible while keeping that linked to provenience information, where in the site it came from because that's what gives chronological control through time.” CCU06

Preserving Evidence

- Validity
- Reliability

Validity

“They'll present a slide with like eight or nine pieces of pottery and then say, "Well, you can see from the slide that this is Roman era stuff." And like I have no idea from this grainy black and white slide what you got there.” (CCU01)

Reliability

“The verification of whether or not the data are real is something that, and I don't want to say it's got a basis of trust, but it really has been. But it's frequently measured on the metadata about how everything was recovered and whether or not it ultimately corresponds with similar works that have been done and are done later.”(CCU12)

Disciplinary Repositories

- New
- Untested
- Sustainability
- Breaking down the culture of not sharing
- Fragmented
- No interoperability

Repositories Dependent on Individuals

"The ultimate success or failure of repositories is going to depend on the ability of archeologists to create databases of the information that they have been storing in different ways because of financial preservation concerns for decades." (CCU12)

Other People's Data

- Qualities in the data
- Functionality of the data
- Preserving meaning
- Preserving evidence

Thank-you

- Questions?