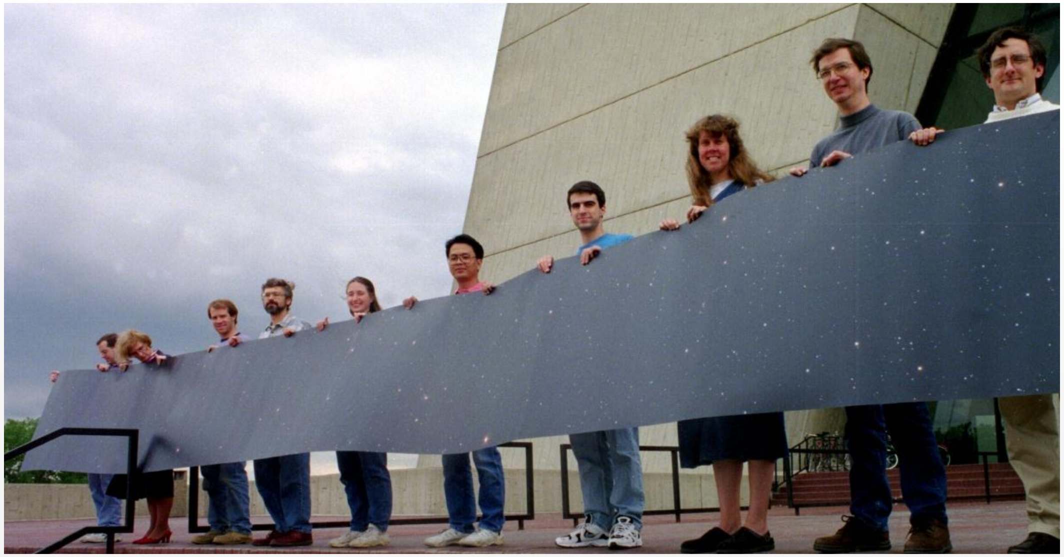
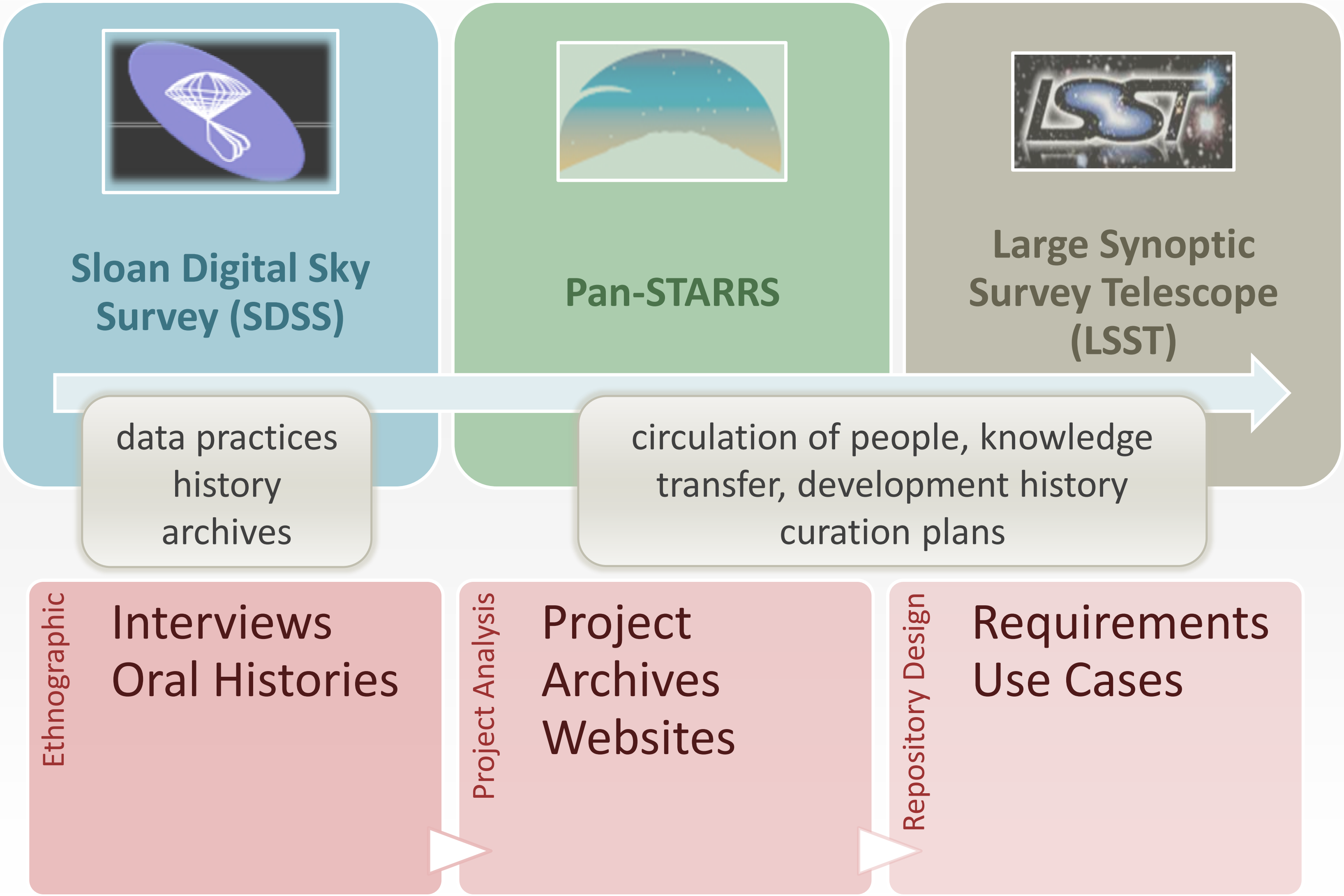


Laura Wynholds
David S. Fearon Jr
Christine L. Borgman
Sharon Traweek
University of California, Los Angeles

WHEN USE CASES ARE NOT USEFUL: DATA PRACTICES, ASTRONOMY, AND DIGITAL LIBRARIES

Methods and Scope



Sloan Digital Sky Survey (SDSS)

Pan-STARRS

Large Synoptic Survey Telescope (LSST)

data practices
history
archives

circulation of people, knowledge transfer, development history
curation plans

Ethnographic

Interviews
Oral Histories

Project Analysis

Project
Archives
Websites

Repository Design

Requirements
Use Cases

Research Questions

1. Data practices

What are the data management, curation, and sharing practices?

2. Social networks

Who uses what data when, with whom, and why?

3. Curation

What data are most important to curate, how, and for whom?

Astronomy Data Practices

- Astronomy data
  - Heterogeneous
  - Highly distributed data collections
  - many important datasets have no home
- Data-intensive terabyte and petabyte scale projects favor large national and international collaborations
- Long tail of smaller-scale investigations with more modest data footprint

Complex Data Transformations

"Even defining what is the most basic thing, like what's the flux coming from a galaxy?...there are tens or hundreds of ways to do that. And if you don't know what you're looking at, you cannot do precision work."

- Observations are handled with a constant awareness of the constraints on their evidential value.
- Use requires expertise, knowledge and judgment.
- Astronomers rely on personal negotiations with data experts.
  - NASA data archives provide experts in addition to data
  - Astronomers reported contacting other researchers and being contacted by others regarding their data

Known Digital Library Design Challenges

- Heterogeneity
  - of data as a concept
  - of data uses and repurposing
  - of identity and identifiers
  - of approaches to data in scholarly publishing

- Use impacted by
  - tacit knowledge
  - origin
  - audience
  - trust
  - reliability
  - validity
  - dependencies
  - description
  - documentation

Top 5 Non-Technical Challenges

- Focus on use cases was failing to represent the predominance of social and socio-technical challenges:
  - I. Trust
  - II. Documentation
  - III. Value
  - IV. Funding and curation environments
  - V. Data Integration and Interoperability

I. Trust in Sources

- Extensively tested and vetted data products are generally considered well documented and trustworthy
  - e.g. SDSS, NASA satellite missions
  - canonical vs secondary data
- Trust is related to:
  - expertise & reputation of the data producers
  - adequacy of documentation
  - fears of misinterpretation,
  - how the data were processed

“the further you are from that expert, the less trust.”

II. Documentation

“Whenever you have not generated the data, then the documentation must be incredibly detailed and incredibly well thought through. Because if I produced the data, then I know exactly what's happening, but if somebody else did, it's so easy to miss essential details or misinterpret things.”

- Lack of clear provenance makes it easier to rerun secondary data manipulations than to figure out what was done.
- Astronomers were hesitant to share their own secondary data products citing cost of documentation and concerns regarding misinterpretation

III. Funding and Curation Environments

- curation impacted by
  - methods & approaches
  - funding sources,
  - project size
  - type of instrument
  - size of collaboration
  - space based vs. ground based
  - parent institution type
  - domain
    - Radio
    - Optical
    - Infrared
    - Ultraviolet and X-Ray
    - Theoretical

IV. Assessing Value

“Show some major new results that came about as a result of data mining, and then explain those tools in the language of astronomers, and then we will start to see a sea change.”

- Astronomers evaluate data based on anticipated use & quality of results
- Future use is difficult to anticipate.
  - Contested within and between subdomains

V. Data Integration and Interoperability

- Astronomers regularly draw observations from multiple data archives and projects
- Poor interoperability among archives and a steep learning curve has resulted in low adoption rates of tools in the community
- Perception of a lack of significant discoveries associated with data reuse

“There’s some off the shelf tools, they're not terribly scalable, most people don't even know how to use them, not even such as they are. The learning curves are very steep, their penetration in communities is very low."

“you risk having people grab data from these distributed archives and not really understand the data. I think there actually are papers that are having that problem now.”

Conclusions

- Trusted data products are central to astronomy research, especially when well calibrated, vetted, and publicly accessible.
    - Secondary data products are currently socially marginal
- Considerable resources, especially in the form of human expertise are required to curate canonical datasets
- Data digital libraries need to address issues of trust, documentation, interoperability, and appraising value as a part of design.

Acknowledgements

Comments & Feedback

- UCLA CENS Data Research Group: Matthew Mayernik, Katie Shilton, Jillian Wallis.

Research funding

- National Science Foundation
  - Data Conservancy: OCI0830976, Sayeed Choudhury, PI, Johns Hopkins University.
- Microsoft External Research