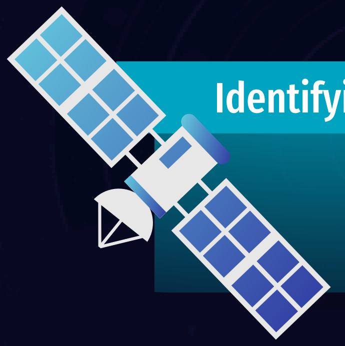
Measuring and Improving the Efficacy of Curation Activities in Data Archives

Dharma Akmon, Sara Lafia, Andrea Thomer, Libby Hemphill, Amy Pienta, Elizabeth Yakel, David Bleckley, and Allison Tyler

Research Questions

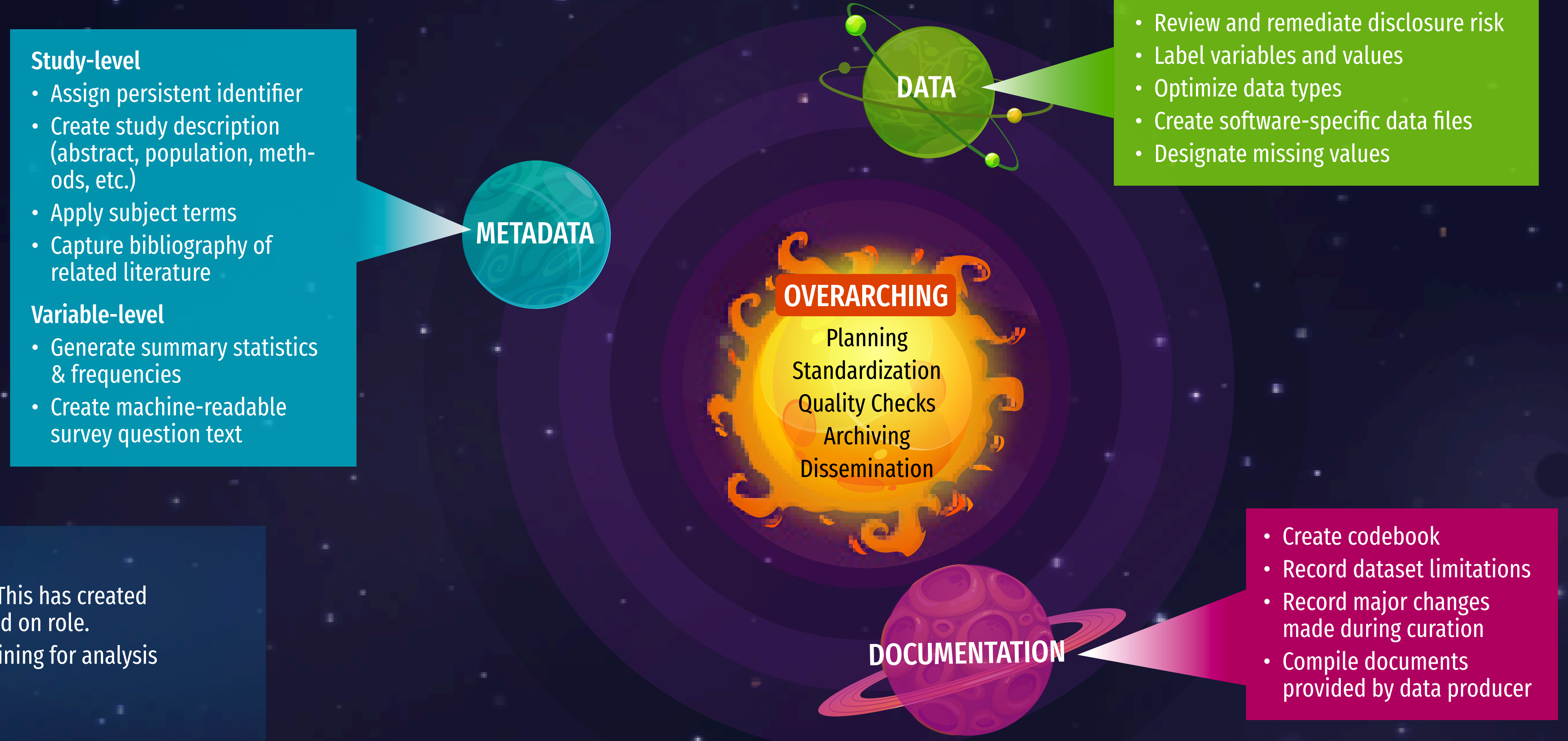
- What impacts do specific curatorial actions have on research data's impact or reuse?
- How should we prioritize curatorial actions to achieve impact and return on investment?



Identifying Curatorial Actions

- Interviews with ICPSR data curators, project managers, and project directors (n=37)
- Annotating JIRA curation work tickets (n=1,618)

Curatorial Activities at ICPSR Organized by Type



Study-level

- Assign persistent identifier
- Create study description (abstract, population, methods, etc.)
- Apply subject terms
- Capture bibliography of related literature

Variable-level

- Generate summary statistics & frequencies
- Create machine-readable survey question text

DATA

- Review and remediate disclosure risk
- Label variables and values
- Optimize data types
- Create software-specific data files
- Designate missing values

OVERARCHING

Planning
Standardization
Quality Checks
Archiving
Dissemination

DOCUMENTATION

- Create codebook
- Record dataset limitations
- Record major changes made during curation
- Compile documents provided by data producer

Early Findings

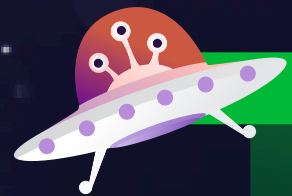
- Different stakeholders talk about curation and impact in different ways
- Interviews reveal substantially different perspectives on curation activities based on role. This has created challenges in achieving intercoder reliability, and as a result we are coding interviews based on role.
- Curation work tickets use consistent but generic language, presenting challenges to text mining for analysis

"Most of my decisions are like what level of curation can we afford ... Sometimes we get really complicated data but it's in such good shape when it gets here, you can ... subject it to the lowest level of curation and it still has as much information content as the highest level of curation."

"Funding will guide [curation], and appropriateness. Like some data just [aren't] really useful or appropriate for online analysis. Because of what the data are or whatever it may be, it just may not be appropriate for that kind of thing, for people to just go in and be able to explore and run some crosstabs that might not be the best idea for our customers or user group."

"After completing quality checks, I know that the data are ready to share. It's very important to make sure the data we release does not pose any disclosure risks and this is an important step in doing so."

"Even though it takes a lot of time to create, codebooks are really important resources for data reusers. They're especially valuable when they explain the full text of questions and variables used in the project."



Warp Speed to the Future

- Explore ability to train a text classifier to automatically identify curatorial actions in curation work tickets
- Associate time spent on curation activities with dataset features and reuse patterns
- Track diversity and secondary impact of dataset citations with the ICPSR bibliography

SCHOOL OF INFORMATION UNIVERSITY OF MICHIGAN · INSTITUTE of Museum and Library SERVICES · NSF · ICPSR

MICA is funded by the Institute of Museum and Library Services (LG-37-19-0134) and the National Science Foundation (#1930645).