How Does Data Curation Impact Discovery?

Sara Lafia*, Andrea Thomer, Libby Hemphill, Dharma Akmon, David Bleckley, Amy Pienta, Elizabeth Yakel

(*Lead contact: slafia@umich.edu)

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?

About MICA

- MICA is a research initiative at ICPSR and the University of Michigan studying the impacts of curatorial actions on digital collections.

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

METADATA

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

Approach

Annotate curation activities in a corpus of work logs:
ICPSR keeps detailed records of curation in JIRA tickets. By annotating a subset we hope to train a classifier to automatically extract them.

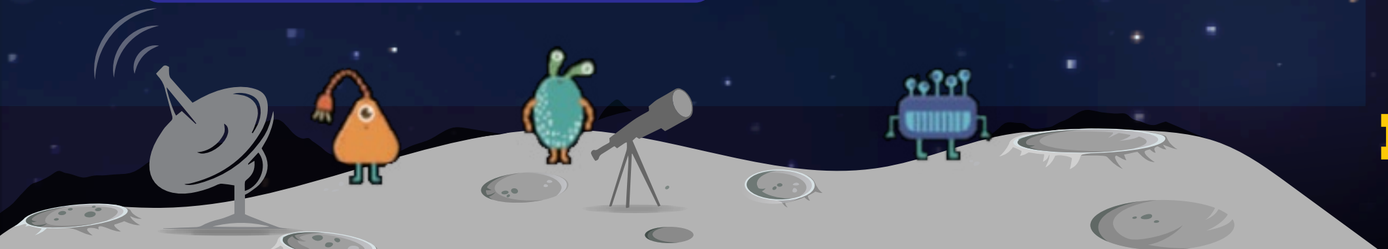
Examples:

- Data (Review-and-remediate-disclosure-risk): Compared new variables from aggregate dataset
- Documentation (Create-codebook): Re-run hermes to get the updated title on the setup files and the codebook cover

Track dataset citations in literature:
Query bibliometric APIs for formal and informal references.

Examples:

- Formal (DOI): https://doi.org/10.3886/ICPSR04254.v1
- Informal (study name): The land cover for the 1930s was derived from the "Population and Environment in the U.S. Great Plains" dataset (Gutmann 2005).



Warp Speed to the Future

- Associate time spent on curation activities with dataset features and reuse patterns
- Track diversity and secondary impact of dataset citations with the ICPSR bibliography

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