# QUANTIFYING THE VALUE OF METADATA CAPITAL FOR DATA SCIENCE

Adrian T. Ogletree, Herbie Huang, Austin C. Mathews
School of Information & Library Science
University of North Carolina at Chapel Hill

## OVERVIEW

- Advances innovative, nascent work on "metadata capital" in support of metadata reuse and its value for data science
- Study utilizes a multi-method approach including case studies, collaborative workflow modeling, and content analysis
- Incentivizes metadata operations and support for data-intensive science

## BACKGROUND

"Capital" refers to a tangible asset with a measurable value that can fluctuate over time.

Reuse of good-quality metadata may have greater value than its original net worth.

"Metadata capital" describes the economics of metadata generation and propagation through net gains or losses, quantifying metadata re-use, and applying quality metrics.

Modified Capital-Sigma Summation

$$R + \sum_{i=1}^{N} a_i = R + a_1 + a_2 + a_3 + \ldots a_n$$

R = value of the metadata record

*i* = number of usages

a = incremental increase in value

n = maximum number of reuse

## RESEARCH QUESTIONS

- What are the economic benefits of metadata generation and propagation?
- How do the values of these investments fluctuate over time?

## CASE STUDIES

### Research Triangle Institute (RTI) Center for the Advancement of Health Information

- Goals: Development of a SGHI Exchange (SGHIx) Market to connect individuals to health researchers
- Self-generated health information (SGHI): information created, recorded, gathered, or inferred by or from individuals through applications and devices, e.g. activity tracking sensors, mobile health apps
- What is the worth of an individual's health information? Are certain subsets of metadata more valuable than others? How much more valuable is de-anonymized health information?

### Office of Scientific Information Management at the National Institute of Environmental Health Sciences (NIEHS)

- Goals: Enable environmental health data sharing through the development of an ontology to guide searches of massive, discipline-specific data repositories
- An "ontology" refers to the formulation of definitions, classifications, and relationships, using the tools of logic and formal semantics
- Next-generation sequencing technologies are drastically increasing the amount of data produced
- What is the long-term economic cost of not investing in a standardized ontology?

## WORK PLAN

| TASK | MONTH 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Environmental scan/refinement | ■ | ■ | | | | | | | | | | |
| 2. Develop use cases | ■ | ■ | | | | | | | | | | |
| 3. Pilot study (design and implement) | | | ■ | ■ | | | | | | | | |
| 4. Data analysis for pilot | | | | ■ | ■ | | | | | | | |
| 5. Revise research plan | | | | | | ■ | | | | | | |
| 6. Formal, full study | | | | | | | ■ | ■ | ■ | | | |
| 7. Half year report and advisory assessment | | | | | | ■ | | | | | | |
| 8. Explore successive growth rates equation | | | | | | | | | ■ | ■ | | |
| 9. Data analysis and draft reporting | | | | | | | | | ■ | ■ | ■ | |
| 10. Scientific/scholarly output | | | | | | | | | | ■ | ■ | ■ |
| 11. Final report and final assessment | | | | | | | | | | | | ■ |

Acknowledgements | NCDS support, Rebecca Boyles, Tom Caruso, Jane Greenberg (PI), and Angela Murillo

MetaDataCAPT'L