A Framework for Automated Metadata Discovery and its Management for Heterogeneous Data Integration

Ramkiran Gouripeddi[1, 2], Peter Mo[2], Randy Madsen[2], Phillip Warner[2], Ryan Butcher[2], Jingran Wen[1], Jianyin Shao[1], Nicole Burnett[1], Naresh Sundar Rajan[1, 2], Bernie LaSalle[1, 2], Julio C. Facelli[1, 2]

[1]Department of Biomedical Informatics, [2]Center for Clinical and Translational Science, University of Utah, Salt Lake City, Utah, USA

Introduction

- Modern biomedical research requires reuse and federation and/or integration of multiple disparate and heterogeneous data.
- Data embeds within itself different meanings (semantic) and structural (syntactic) descriptions either explicitly or implicitly.
- Metadata as described by the FAIR[1] (Findable, Accessible, Interoperable, and Reusable) principles is a requirement for reproducible research.
- This requires discovery of these metadata and its understanding to facilitate proper use of data.
- Current state of the art requires a great deal of human manual curation, which renders these procedures non-scalable and consequently of limited practical value in the emerging big data biomedical science paradigm.

Next Steps: Develop and evaluate this framework using workflow platforms (e.g. Swift, Pegasus).

References

1. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
2. Gouripeddi R, Facelli JC, et al. FURTHeR: An Infrastructure for Clinical, Translational and Comparative Effectiveness Research. AMIA Annual Fall Symposium. 2013; Wash, DC.
3. WG3 Members. (2015). WG3-MetadataSpecifications: NIH BD2K bioCADDIE Data Discovery Index WG3 Metadata Specification v1. Zenodo. 10.5281/zenodo.28019

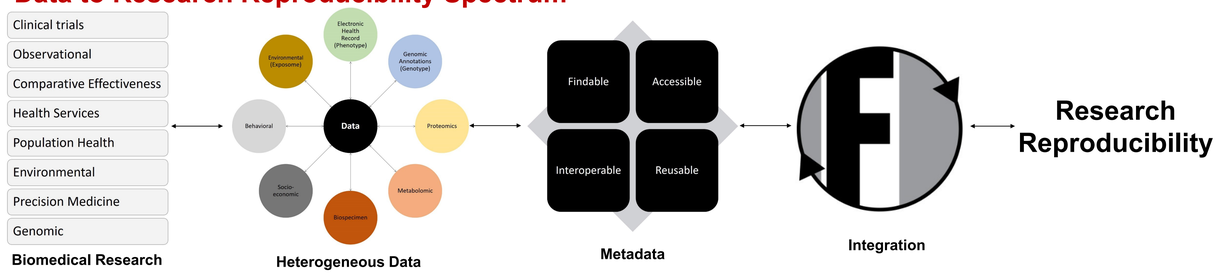
Acknowledgements

bioCADDIE is supported by the National Institutes of Health (NIH) through the NIH Big Data to Knowledge, Grant 1U24AI117966-01. OpenFurther is support NCRR/NCATS UL1RR025764, 3UL1RR025764-02S2, AHRQ R01 HS019862, DHHS 1D1BRH20425, U54EB021973, UU Research Foundation, NIBIB, NIH U54EB021973

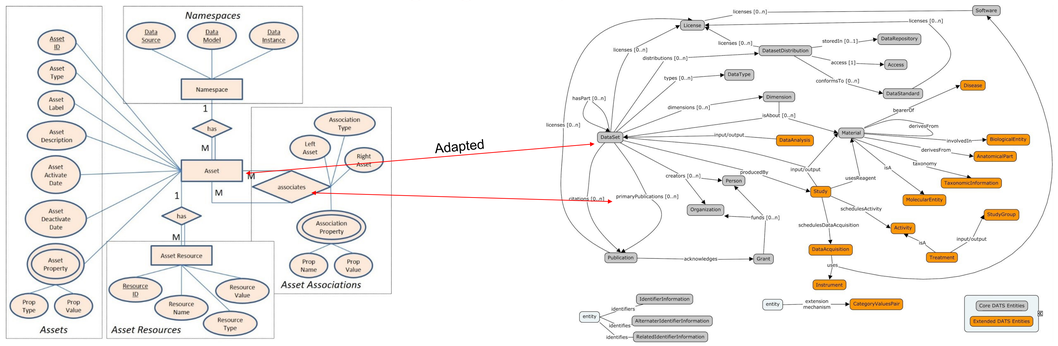
Contact: Ram Gouripeddi
ram.gouripeddi@utah.edu

The University of Utah Biomedical Informatics

Data to Research Reproducibility Spectrum



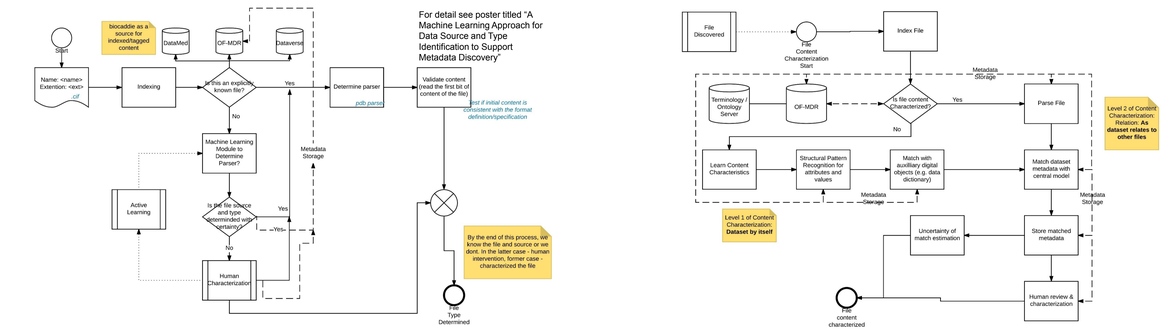
Metadata Specifications for Describing Digital Objects



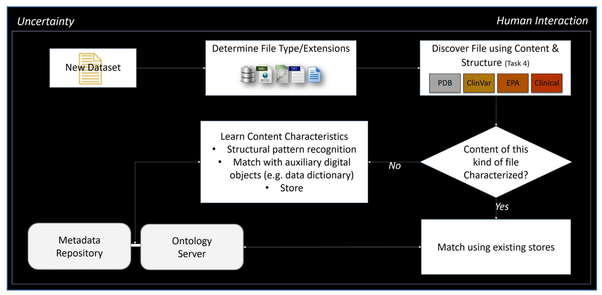
Conceptual Model of Metadata Repository

Schematic Overview of the bioCADDIE Metadata Specification

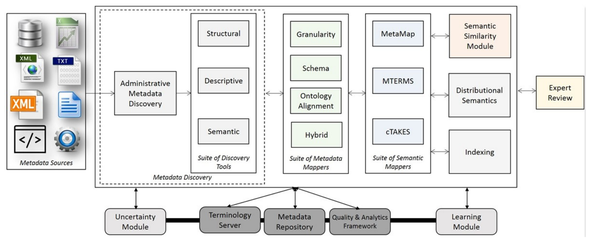
Workflows for Dataset Type & Source Discovery and Dataset Content Characterization



Metadata and Semantics Discovery and Mapping Service



High-level Workflow for Metadata Discovery



Architectural Overview of Metadata Discovery Service

Methods

- To overcome these limitations, we are prototyping a computational infrastructure that supports automated and semi-automated discovery mapping of metadata artifacts.
- Advanced OpenFurther's Metadata Repository[2] (MDR) to adapt DatA Tag Suite (DATS) metadata specifications developed by the bioCADDIE consortium[3] as assets for scalable interoperability between systems for creating, managing and using data.
- This method supports:
  - Harmonization of metadata of individual datasets (e.g. different protein files) for data integration.
  - Provides a flexible data resource metadata storage system that supports versioning metadata (e.g. DATS 1.0 to 2.1) and data files mapped to different versions.
  - Enhance descriptors of resources (DATS) with descriptions of content within resources
  - Translate to other metadata specifications (e.g. schema.org).
- MDR stored metadata is available for various data services including data integration.
- Designed a process workflow for Metadata Discovery and Mapping Service, for automated metadata discovery consisting of a 3-step process:
  1. Identification of data file source and format
  2. Detailed metadata characterization based on (1)
  3. Characterization of the file in relation to other files to support harmonization of content as needed for data integration.