A Conceptual Architecture for Reproducible On-demand Data Integration for Complex Diseases

Ramkiran Gouripeddi[1,2], Karen Eilbeck[1,2], Mollie Cummins[1,2,3], Katherine Sward[1,2,3], Bernie LaSalle[1,2], Kathryn Peterson[4], Randy Madsen[2], Phillip Warner[2], Willard Dere[2,5], Julio C. Facelli[1,2]

[1]Department of Biomedical Informatics, [2]Center for Clinical and Translational Science, [3]College of Nursing, [4]Department of Gastroenterology, [5]Department of Endocrinology, University of Utah, Salt Lake City, Utah, USA

Introduction

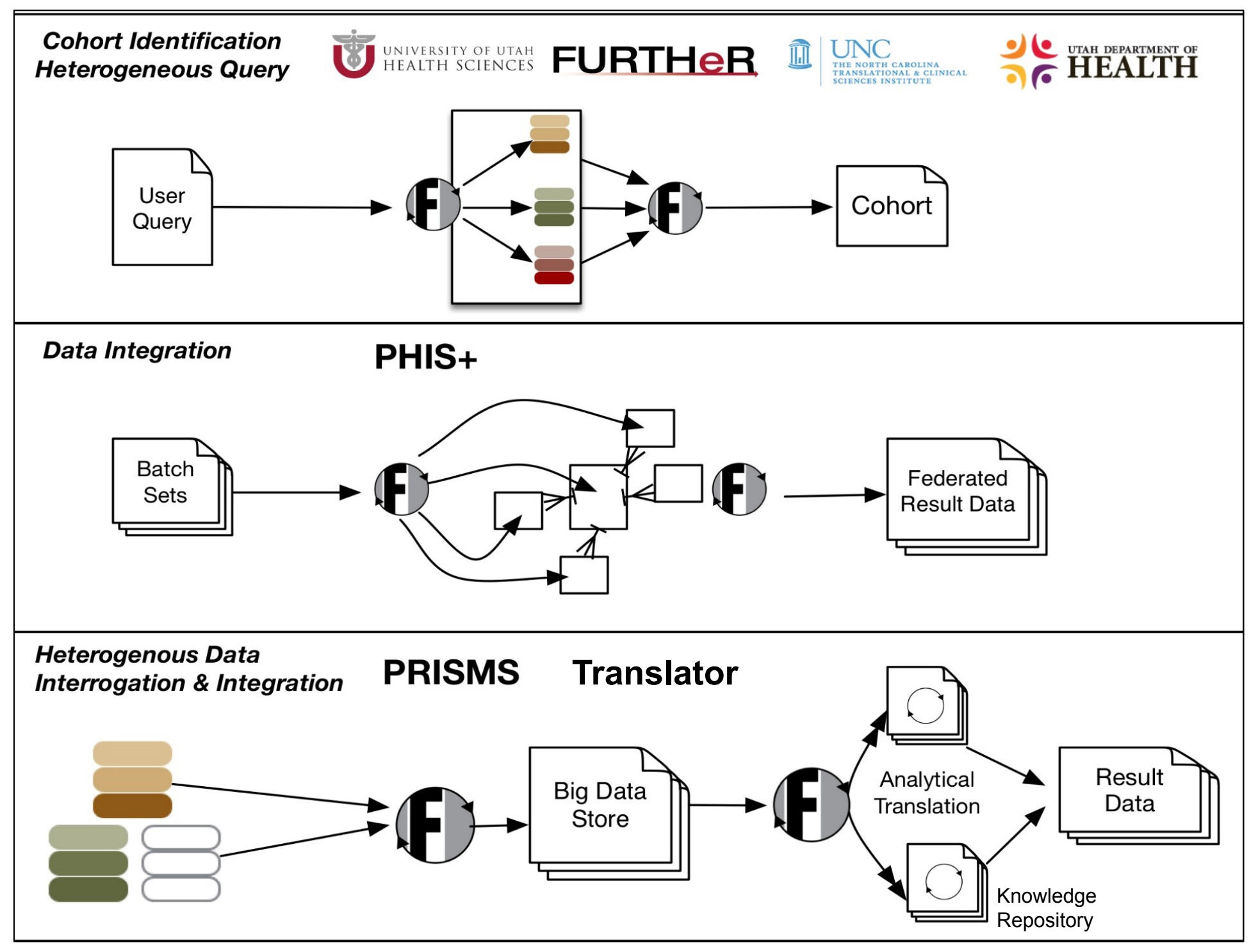
- Eosinophilic Esophagitis is a complex and emerging condition characterized by poorly defined phenotypes, and associated with both genetic and environmental conditions.
- Understanding such diseases requires researchers to seamlessly navigate across:
  - Multiple scales: Metabolome, Proteome, Genome, Phenome, Exposome
  - Models: Sources using different stores, formats, and semantics
  - Interrogate existing knowledge bases
  - Provision data in formats of choice to answer different types of research questions.
- Ensuring reproducibility requires sharability of methods used for:
  - Selecting data sources
  - Designing research queries
  - Executing data queries
  - Interpreting results and their quality.

Implementation

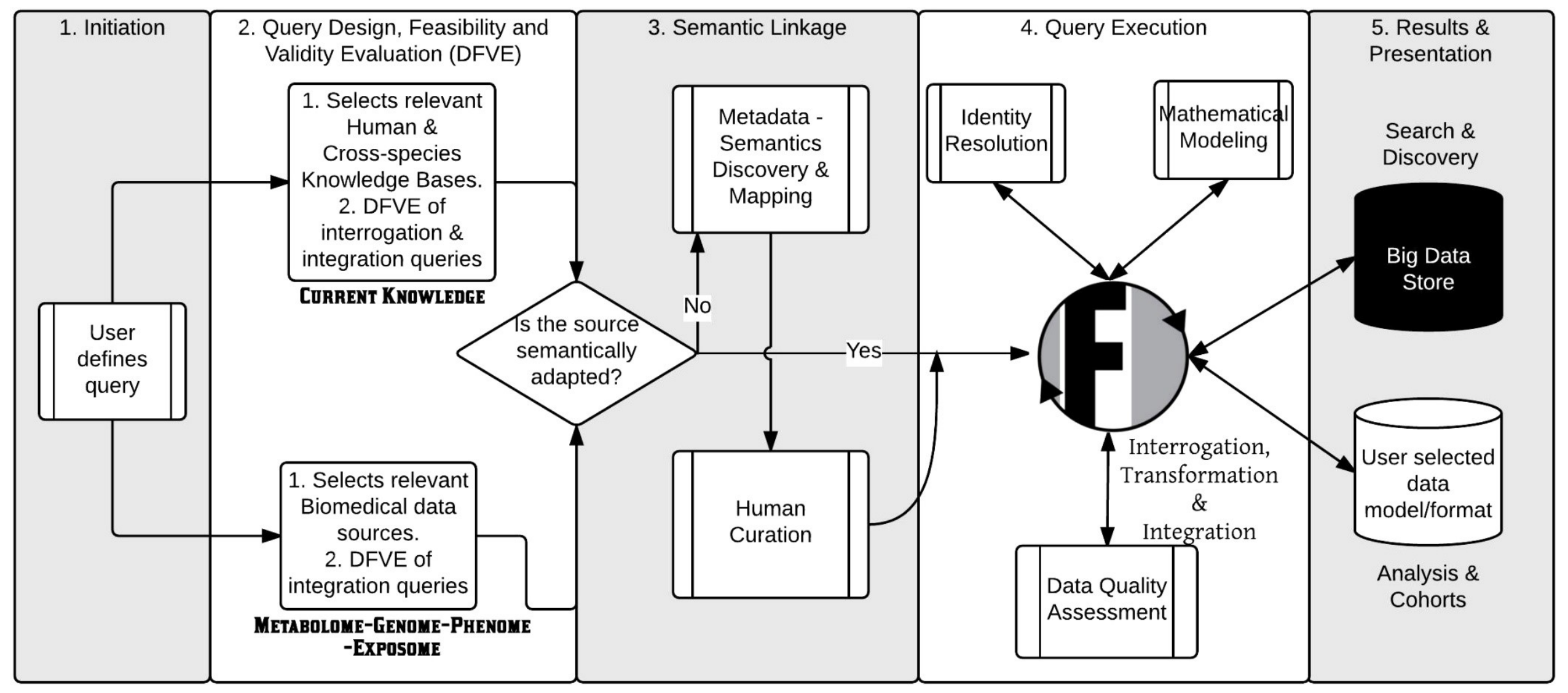
- Uses process exchange formats: DAX (Description of an abstract workflow in XML format), BPMN (Business Process Model and Notation) and Scientific workflow systems: Pegasus[2], Apache Taverna[3].
- Sharable content (PC, rules, and workflows), assembling, and executing mechanism.

References

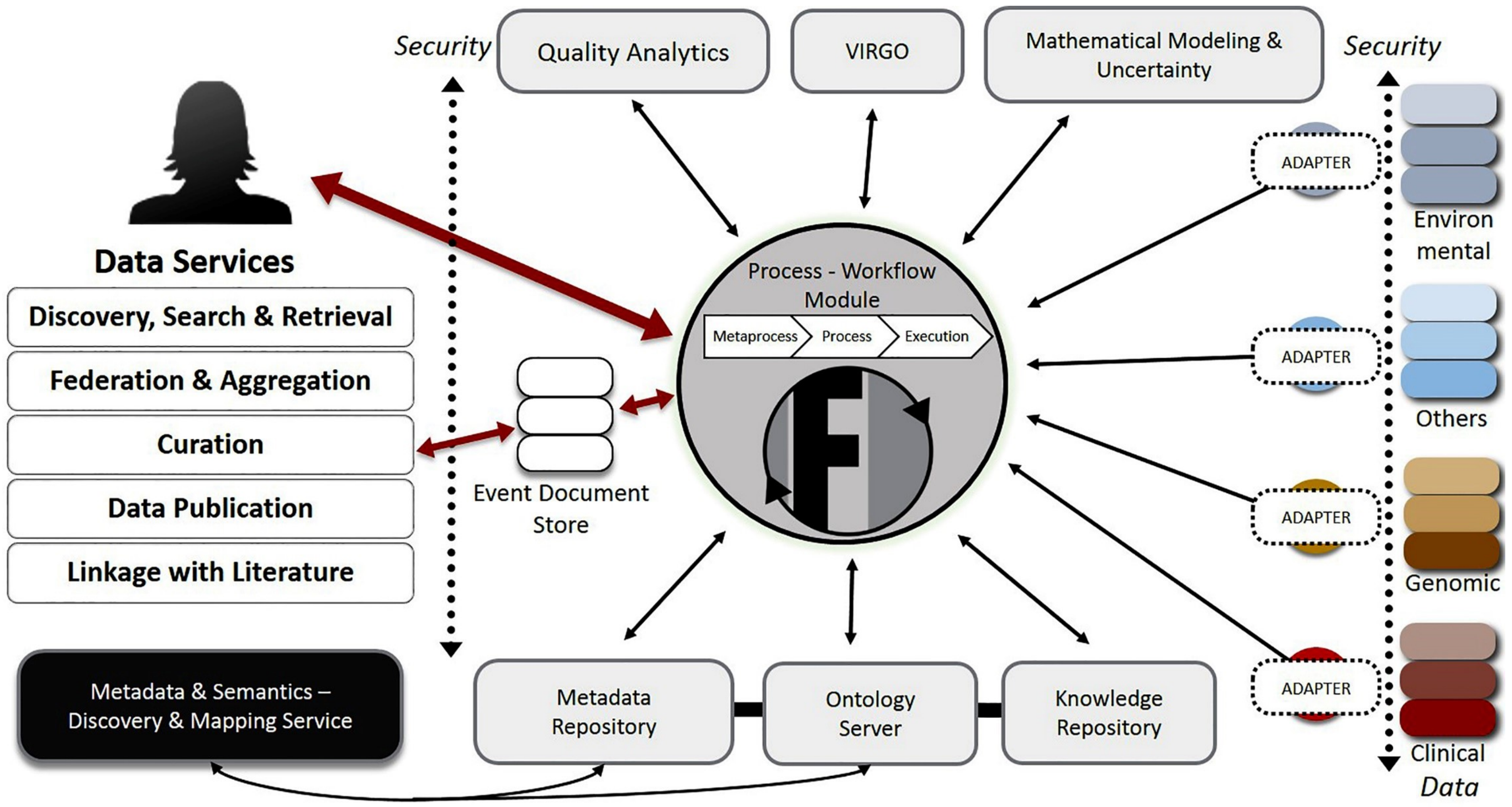
1. Gouripeddi R, Facelli JC, et al. FURTHeR: An Infrastructure for Clinical, Translational and Comparative Effectiveness Research. AMIA Annual Fall Symposium. 2013; Wash, DC.
2. Pegasus. The Pegasus Project. 2016; https://pegasus.isi.edu/.
3. Apache Software Foundation. Apache Taverna. 2016; https://taverna.incubator.apache.org/.


Cohort Identification Heterogeneous Query
UNIVERSITY OF UTAH HEALTH SCIENCES
FURTHeR
UNC THE NORTH CAROLINA TRANSLATIONAL & CLINICAL SCIENCES INSTITUTE
UTAH DEPARTMENT OF HEALTH
User Query
Cohort
Data Integration
PHIS+
Batch Sets
Federated Result Data
Heterogenous Data Interrogation & Integration
PRISMS
Translator
Big Data Store
Analytical Translation
Result Data
Knowledge Repository

Flexible Data Integration


1. Initiation
2. Query Design, Feasibility and Validity Evaluation (DFVE)
3. Semantic Linkage
4. Query Execution
5. Results & Presentation
User defines query
1. Selects relevant Human & Cross-species Knowledge Bases. 2. DFVE of interrogation & integration queries
Current Knowledge
1. Selects relevant Biomedical data sources. 2. DFVE of integration queries
Metabolome-Genome-Phenome-Exposome
Is the source semantically adapted?
No
Yes
Metadata - Semantics Discovery & Mapping
Human Curation
Identity Resolution
Mathematical Modeling
Interrogation, Transformation & Integration
Data Quality Assessment
Search & Discovery
Big Data Store
User selected data model/format
Analysis & Cohorts

A conceptual metaprocess model leveraging OF for researcher-defined construction of translational research query processes.



Process-Workflow Module (PWM) integrates with OpenFurther to provide functionalities for translational researchers to perform complex queries. Through meta-process, process and integration layers the PWM provides generic concepts to construct on-demand query processes. Main components of OF include an Ontology/terminology Server (TS); a Metadata Repository (MDR); Software Services (SS) which can be consumed by various tools; Data Source Adapters (DSA); Administrative and Security Components (ACS); Virtual Identity Resolution on the GO (VIRGO); Quality and Analytics Framework (QAF); a Mathematical Modeling and Uncertainty Module (MM) and a Federated Query Engine (FQE) that orchestrates queries between the PWM, SS, MDR, TS, DSA, ACS, VIRGO, MM and QAF. In addition, the architecture consists of an Event Document Store (EDS) that provides an integrated Big Data store in which data is represented as events in time; a Metadata & Semantics - Discovery and Mapping Service (MDMS) that semi-automatically discovers and maps metadata and semantics in data; and a Knowledge Repository (KR) for storing knowledge facts.

Methods

- Content, design, and development of the framework is informed by user-centered design methodology.
- Consists of researcher and integration-centric components to provide robust and reproducible workflows.
- Develop higher level of formalisms for building multi-source data platforms on-demand.
- Formalisms based on the principles of meta-process modeling and provide reproducible and sharable data query and interrogation workflows and artifacts.
- A framework based on these formalisms consists of a layered abstraction of processes to support administrative and research end users:
  - **Meta-process** (top): An extendable library of computable generic process concepts (PC) stored in a metadata repository[1] (MDR) and describe steps/phases in translational research life cycle.
  - **Process** (middle): Methods to generate on-demand queries by assembling instantiated PC into query processes and rules. Researchers design query processes using PC, and evaluate their feasibility and validity by leveraging metadata content in the MDR.
  - **Execution** (bottom): Interaction with a hyper-generalized federation platform (e.g. OpenFurther[1]) that performs complex interrogation and integration queries requiring consideration of interdependencies and precedence across the selected sources.

Acknowledgements: OpenFurther is support NCRR/NCATS UL1TR001067, UL1RR025764, 3UL1RR025764-02S2, AHRQ R01 HS019862, DHHS 1D1BRH20425, U54EB021973, UU Research Foundation, NIBIB, NIH U54EB021973. **Contact:** Ram Gouripeddi ram.gouripeddi@utah.edu