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

Ramkiran Gouripeddi1,2, Karen Eilbeck1,2, Mollie Cummins1,2,3, Katherine Sward1,2,3, Bernie LaSalle1,2, Kathryn Peterson4, Randy Madsen2, Phillip Warner2, Willard Dere2,5, Julio C. Facelli1,2

1Department of Biomedical Informatics, 2Center for Clinical and Translational Science, 3College of Nursing, 4Department of Gastroenterology, 5Department of Endocrinology, University of Utah, Salt Lake City, 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: Pegasus2, Apache Taverna3.
• Sharable content (PC, rules, and workflows), assembling, and executing mechanism.

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 interroation 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 (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) that performs complex interogation and integration queries requiring consideration of interdependencies and precedence across the selected sources.

References

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

Process-Workflow Module (PWM) integrates with OpenFurther to provide functionalities for translational research 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.