SI MDR Infrastructure Summary Requirements Outline

Semantic Infrastructure (SI) MDR Requirements Summary

CBIIT's mission is to provide and advocate for the appropriate use of data science, informatics, and information technology (IT) to support and accelerate the NCI Mission to prevent cancer, treat cancer, and improve cancer outcomes. An important role of the NCI Semantic Infrastructure (SI) is to support the NCI research mission through community definition and collection of metadata. Data that have well defined linked metadata can improve the use, interpretation, and reuse of data and the extraction of information and knowledge from these data. Supporting both human readable and machine-readable definitions and metadata has been an important driver for the NCI Semantic Infrastructure. These general metadata characteristics are also among the key principles for data citation and are noted to enable data access, verifiability, and discoverability.

The primary goals for updating the metadata services are to:

- Simplify and streamline community creation, curation, maintenance, discovery, and reuse;
- Support content harmonization leveraging automated means for identification of overlapping content;
- Support interoperability and integration of data elements, modules of elements, and semantics into existing and novel workflows; and
- Support knowledge extraction.

The requirements outlined below are described in more detail in the DRAFT Extended MDR RequirementsV10.12.docx (the Extended MDR Requirements is a more descriptive document, and therefore lengthy. We are still working to make a more stakeholder readable version of it). This document was frozen in June 2016, and provided input for the EPLC documents and the SOW.

MDR Roadmap - Next Steps:

Through an RFI process and research, we identified over 13 potential solutions and partial solutions. We created a requirements spreadsheet and engaged with each of the potential providers to request a self-assessment of their capabilities versus NCI requirements. We mapped all the results into one document and created a quantitative score for each one based on the number of requirements they could meet. We then organized the requirements by highest priority and re-scored the responses, narrowing the possible choices to the top 5. We requested product demonstration and information sessions with each of the top 5. The two possible solutions who held position 4 and 5 respectively turned out not to be commercially available and thus dropped from the list. We have selected the top 3, scoring 97%, 75%, and 58%, for a 45-day software evaluation for usability testing, and also decided to test the NLM system for comparative purposes. We have developed an "Evaluation Scorecard" to be used in recording results from testing the features that are the most commonly used. The requirements matrix and evaluation scorecard are available upon request from the government sponsor.

We are currently developing two possible solutions for current and future customers to supplement and leverage metadata and models.

- Data Mapping and Transformation Tool: Ptolemy.V - Ordinal Data
  - A translation engine that leverages NCI CDEs will create a new file by copying source data you have registered in the tool, and converting it to the common standard format using CDEs. Through a mapping step you can create one or more translations for your data, or by reusing translations that have already been created by other users, you can create new composite, integrated data tables using data from a variety of source data. You can access your tables directly in the repository or you can download/export the data as a CSV file.
  - The project is being managed by Leidos.
- Metadata Template Builder: CEDAR - Stanford
  - A web based tool that uses NCI CDEs and leverages open terminologies in BioPortal to create a metadata template to collect and validate data.
  - The project is being managed by Leidos.

SI Data Semantics Outline

1 Executive Summary

1.1 Mission and Goals

2 Overview, Background, Alternatives

2.1 Background

2.2 Lessons Learned

2.3 Purpose and Scope

2.4 Stakeholders

2.4.1 Personas

2.4.1.1 Mary Metadata Curation Specialist

2.4.1.2 Danny Data Manager

2.4.1.3 Alice Application developer

2.4.1.4 Ralph Researcher/Analyst

2.4.1.5 Harry Harmonization Specialist

2.4.1.6 Pete Principal Investigator

2.5 Alternatives and Analysis

- We have been analyzing the capabilities of several existing metadata repositories and those providing repository capabilities, comparing this information with the NCI requirements. These solutions are partial solutions. A prioritization task will help us determine which solutions are the best match for NCI.

2.5.1 NLM CDE Repository

2.5.2 Semantics Manager - SOA Software - Akana

2.5.3 OneData – Software AG

2.5.4 Constellation – DOD

2.5.5 SALUS

2.5.6 Colectica

2.5.7 cTAKES, YTEX, MetaMap, UNIM (deferred further investigation, these are capabilities not full-blow repositories)

2.5.8 TransMart (deferred further investigation)

2.5.9 DataType Registry – CNRI

2.5.10 Oxford University Metadata Registry

2.5.11 DataOne- USGS (Earth)

2.5.12 openMDR – Ohio State

2.5.13 USHIK

2.5.14 Mayo

2.5.15 IMOS Consulting

2.5.16 Elsevier

2.5.17 CEDAR

2.5.18 NIH BRICS

3 Requirements and High Level Use Cases

3.1 Functional Requirements

3.1.1 Core Services

3.1.1.1 Search

- provide the ability for end users to find content based on user search criteria.

3.1.1.2 Faceted browsing

- provide users with the ability to see related content, without tacit knowledge of the underlying 11179 metamodel or specific information model or content in the repository

3.1.1.3 Curation and Maintenance

- provide a way to record and maintain structured content with minimal effort and training for curators

3.1.1.4 Download

- provide a human accessible mechanism for getting content out of the repository to meet a variety of end user stories

3.1.1.5 Password administration

- provide user password maintenance services consistent with NIH policies.

3.1.1.6 Compare

- provide customers with the ability to identify and select similar items and view features and attributes of the items side-by-side.

3.1.1.7 Application Programming Interfaces

- provide internal and external customers with a programmatic access to repository contents.

3.1.1.8 Subscription and Notification

- provide users of NCI metadata with information about important changes.

3.1.1.9 Administration

- provide the ability to setup and configure the repository for its use; user accounts and privileges; customizing lookup tables to reflect preferred values, naming conventions and business rules.

3.1.2 New Metadata Services

3.1.2.1 Standards Collaboration

- provide the ability for users to discuss existing content and evolve the content to meet community needs.

3.1.2.2 Harmonization activities

- provide the ability to help detect similar items; support reviewing, comparing, adding “mapping” metadata, recording decisions.

3.1.2.3 Registration, Submission, and Governance

- support a procedure wherein content from various areas in NIH can be developed and maintained as needed to meet specific use cases, while at the same time submitted for central harmonization and elevation to preferred standard for the community. The process, roles and responsibilities are described in ISO 11179-6 Registration.

3.1.2.4 Reproducibility of Results

- provide a mechanism for researchers’ to record structured information about research studies so that the published data and conclusions can be reproduced.

3.1.2.5 Data Discovery Metadata, HPC, GDC, and Cloud (Big Data Initiatives) (*Ian Fore, Eric Stahlberg)*

- provide metadata to support emerging and existing national and international standards to share, discover, interpret and use data.

3.1.2.6 Team Science Data Management (*Kara Hall*)

- providing support for the unique characteristics when science is conducted as a collaboration.

3.1.2.7 Metadata Driven Software Development (*TBD*)

- facilitate reuse of the metadata to automate or semi-automate user interface design (drop downs, screen labels, etc), data structure creation, data validation, data transmission, data transformations

3.1.2.8 Reporting and Content Quality Metrics (*Dianne Reeves*)

- provide customizable reporting and statistics to support metadata curators and content administrators to improve registry content and best practices.

3.1.2.9 Semantic Web Metadata Services (*Gilberto Fragoso)*

- The goal of these services are to provide the ability to leverage the semantics of CDEs to explore the meaning of the CDE, the meaning of directly related data, discovery of related data elements, data, and information using semantic web technologies.

3.1.2.10 Community Portals

- Provide the ability to organize content by communities, where CDEs, measures, CRFs and other related information about how to use the standards can be found

3.1.3 New API Services

3.1.3.1 Conformance Testing

- provide a means to test conformance to a specific item’s metadata specification

3.1.3.2 CRUD

- provide dynamic registration and maintenance of metadata via application programming interfaces.

3.1.3.3 Mapping and Transformation

- leverage the infrastructure's semantic and representational metadata to enable automated aggregation of similar or related data.

3.1.3.4 Standard Interfaces

- develop interfaces for a number of national standards for exchanging forms and data elements are emerging through ONC SDC, FHIR and IHE, transformations to other popular systems such as REDCap.

3.2 Non-Functional Requirements

3.2.1 Usability

- simplify and streamline creation and reuse of metadata

3.2.2 Constraints and Dependencies

3.2.3 Extensibility and Customization

- ability to easily extend and customize the architecture to include new kinds of content beyond those as 11179 administered items

3.3 Technical Requirements- NIH, CBIIT

3.3.1 Security

3.3.2 External API Integration

The repository needs the ability to retrieve content from other entities offering similar services.

4 Assumptions and Dependencies

5 Architecture and Constraints

5.1 Architecture Review and Refactor

5.2 Existing Metadata Migration

6 NCI Extensions

7 Performance Metrics

Annex A User prioritized Search Features

Annex B HPC & Cloud Computing

Annex C Advisory Group and Charter

Annex D Interview guide for requirements elicitation

References

Appendix A – Reporting and content quality