**Development and Applications of NCI SI Service Metadata Ontology (SSIMO)**

**Table of Contents / Wiki Pages**

* **Main Page / Introduction**
  + Introduction of SSIMO
  + Introduction of metadata and its uses for NCI SI project
  + Project purposes
  + Team
* **Services Metadata Ontology Development**
  + Brief Introduction
  + Source Code / License
  + Ontology Terminologies.
  + Intended Domains
    - List of Reference Ontologies / METAthesaurus
  + Upper Class Entities
    - Class Level Definitions
  + Object Relationship Properties
    - Object Level Relationships
  + Data Relationship Properties
    - Data Relationship Properties
* **Applications of NCI SI metadata Ontology**
  + Competence Questions
  + Mapping of SI-Resource Annotations
  + Mapping of Features in the Shared SI Service Query

**Main Page / Introduction**

**Introduction**

The NCI Center for Biomedical Informatics & Information Technology (CBIIT) accelerates cancer research by empowering scientists and clinicians with the data and tools needed to conduct productive research and meet the informatics and data needs of the community. Its fundamental objective is based in data accessibility, interoperability, integration, and sharing. To meet the objective, it is important to create a standards-based environment and integrative service that aligns with the larger clinical research ecosystem and supports seamless interoperability from basic research to translational bedside studies. The new Semantic Web technologies provide a powerful state-of-the-art way for developing such a standards-based shared semantic infrastructure service, leading to strong support of the NCI CBIIT mission.

Ontologies are widely used in biomedical data and metadata standardization, and robustly support data integration, sharing, and computer-assisted data analysis. Biological/biomedical ontologies are sets of computer- and human-interpretable terms for entities and relations in specific biological and biomedical domains. The Web Ontology Language (OWL) is a common language for ontology development. The contents of the OWL files can be expressed with RDF triples and stored in an RDF triple store database. The RDF data model makes statements about resources in the form of subject-predicate-object expressions (i.e., triples). The subject-predicate-object triple representation of data is flexible and powerful and can queried via the SPARQL RDF query language.

Funded by a subcontract proposal in response to Solicitation/RFP Number S22-071, our “CanSI” (which stands for “Cancer Semantic Infrastructure”) aims to develop and engineer a shared Semantic Infrastructure (SI) service. CanSI will support the integration and publication of terminologies and metadata developed by the NCI Cancer Data Standards Registry and Repository (caDSR) and Enterprise Vocabulary Service (EVS) programs. The cutting-edge Semantic Web (SW) technologies including the RDF (Resource Description Framework) triplestore and SPARQL (SPARQL Protocol and RDF Query Language) endpoints will be applied for the CanSI development.

The purpose of this document is to lay out our development of a metadata ontology for our CanSI system. We have named this metadata ontology the Semantic Infrastructure Service Metadata Ontology (SSIMO). SSIMO will specify resources, resource type, provenance, common queryable properties (e.g., definition, preferred terms), graph names, and various other entities as required to facilitate queries of multiple graphs using a common terminology. We have utilized terminology and framework from the previous SI schemas and as part of the Information Artifact Ontology (IAO). Note that Dr. He is also an active developer of both OMO and IAO.

We have developed and maintained SSIMO in the Protégé-OWL editor environment using Web-Protégé. We have deployed an internal version of our metadata ontology to the Web-Protégé in a way to facilitate the community review and discussion.

We have accounted for different non-interoperable representations may exist for the same metadata types. We will have done so thorough survey and discussion of different representations, and eventually find the best way to represent specific metadata types. We have also done the term mapping so that different metatype representations from different systems can be mapped to each other. We have also dealt with inconsistencies in terms of semantic relations through careful examination the internal semantic relations among different metadata types and represent such semantic relations in our metadata ontology.

**CANSI Metadata Ontology Development**

We provide a definitions for common related terms so that users who are unfamiliar with ontology can understand. Ontology is concerned with clear and shared definitions in order to enable data interoperability.

**Definitions of Related Ontology Terms**

**Term.** A word or phrase that is used to describe an object or thing. Terms can be composed of a single word or multiple words. If a term is listed as multiple words, it will be denoted in a paragraph with single quotation marks (‘’). Any definitions included within SSIMO will be denoted in italicized text.

**Set.** A collection of objects.

**Class.** A class represents a set of things that share a common feature. Each class contains axioms and annotations. The most prominent axioms used by a class is the ‘is a’ relationship. The ‘is a’ relationship represents the hierarchy relationship between classes. The highest term for class in this ontology is Thing, which represents all possible sets of things.

**Individual.** An individual represents a singular instance of a thing. All individuals belongs to some classes, but are focused either as individual instances. For an ontology, the use of individual can be used to represent specific version of an ontology. For example, ‘SSIMO version 1.07’ would be an individual, while the SSIMO would be a class.

**Property.** A property is a predicate which describes the relationship between two sets of entities. There are three kinds of properties: an annotation property links an entity to an annotation. An object property links a class or individual to another entity. A data property links a class or individual to a specific set of data. Properties have a specific Domain or Range which can restrict their usage.

**Annotation.** An annotation represents information associated with a specific class, individual, or property. Annotations often contain metadata, including definitions, synonyms, commentary, sources, and editor. For all classes, individuals, and properties shown, italicized text denotes the description used in the ontology.

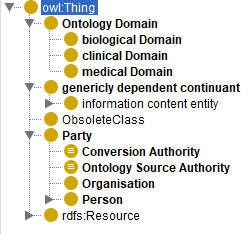
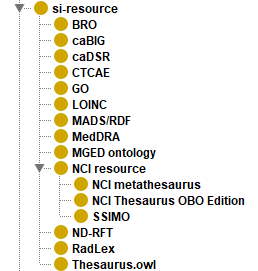
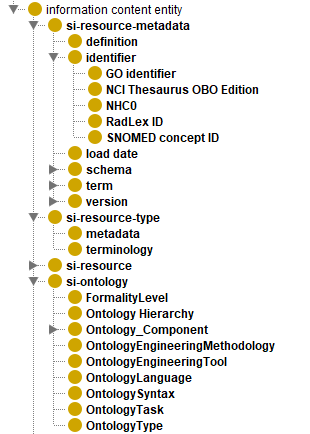
**Domain.** A set of valid inputs. In ontology, a domain is the set of valid classes, individuals, or annotations which are specified as the valid subjects of a property when it is used for an RDF triple.

**Range.** A set of valid inputs. In ontology, a domain is the set of valid classes, individuals, or annotations which are specified as the valid objects of a property when it is used for an RDF triple.

**RDF Triple.** An RDF TRIPLE is a triple statement comprised of three parts: subject, relationship, and object. The subject of a triple is bound by the domain of the relationship and is a class or individual. The object of a triple is bound by the range of the relationship.

**RDF Datasets.** A set of RDF Graphs that is made out of of one default graph and potentially multiple zero named graphs. The named graphs correspond to additional graphs which may or may not reuse terms from the default graphs.

**Metadata Ontology design**



A

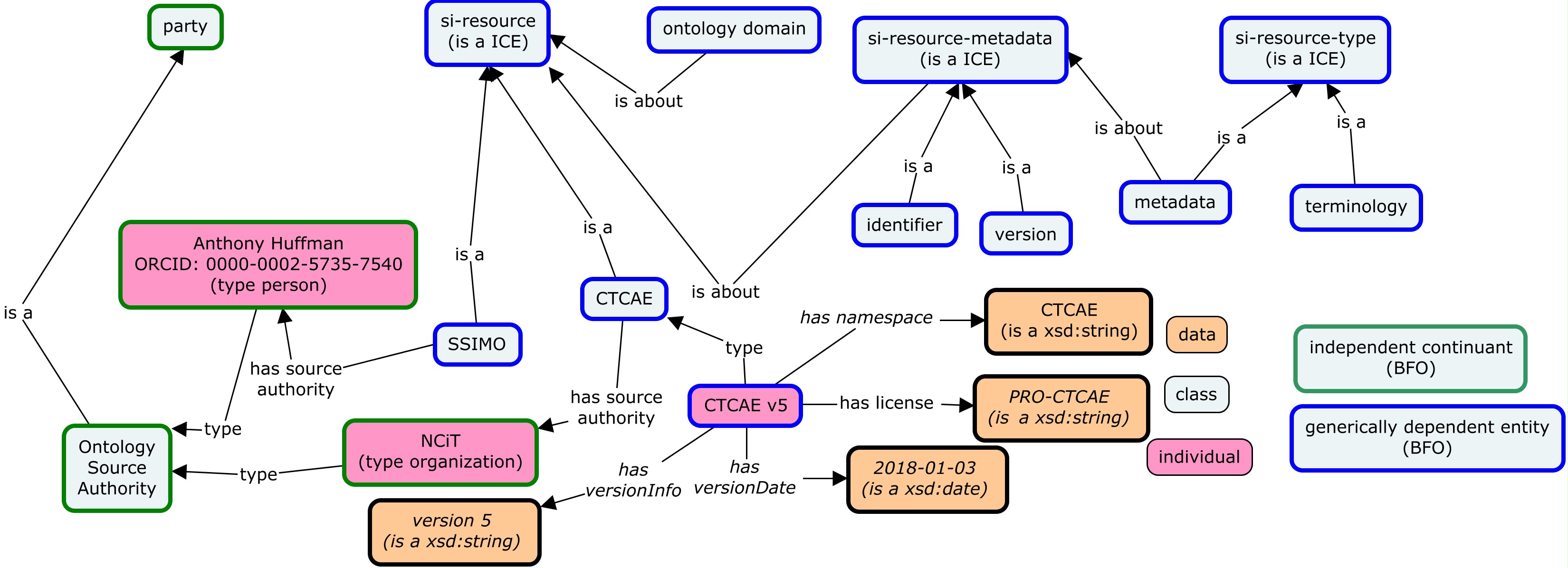
B

**Figure 1. High Level Representation of CANSI Metadata Ontology Tool.** A) Highest level terms for SSIMO include terms that have been mapped to BFO (generically dependent continuant) and those that have not. B) New terms were primarily to si-ontology, si-resource, si-resource-metadata. Red arrows denote locations of terms based of the new graph.

We generated the initial ontology using the prior metonto.rdf central hierarchy for the original base file. Afterwards, we assessed multiple different ontologies to identify terms that can be reused or implemented into SSIMO. Currently, SSIMO imports terms from the Ontology Metadata Vocabulary (OMV), OBO Metadata Ontology (OMO), and the Changes and Annotations Knowledge Base (ChAO KB). **Figure 1** shows the upper level classes that are not imported from other ontologies within the SI Service Meta Ontology. Afterwards, we consulted the above 12 categories to identify common metadata features that are occur within multiple knowledge sources and ontologies. The ontologies and knowledge bases we have queried, along with their definitions, are listed below:

1. **NCI Metathesaurus**. *NCI Metathesaurus (NCIm) is a wide-ranging biomedical terminology database that covers most terminologies used by NCI for clinical care, translational and basic research, and public information and administrative activities.* This is the resource that is most complete and imports several knowledge sources together.
2. **caDSR.** *Cancer Data Standards Registry and Repository. (*caDSR) *is a metadata repository based on the ISO/IEC 11179 Metadata Registry standard*.
3. **MGED ontology**. *MGED Ontology is an older ontology that is used for representation of Microarray Gene Expression Data (MGED).*
4. **NDF-RT**. *National Drug File – Reference Terminology is a knowledge schema used by FDA and the FedMed collaboration to code these essential pharmacologic properties of medications.*
5. **CTCAE**. *Common Terminology Criteria for Adverse Events (CTCAE) Dictionary is a schema used to location appropriate adverse event terms.* CTCAE creates a simple ontology but omits describing formal adverse events that cause death.
6. **LOINC**. *LOINC provides a set of universal names and ID codes for identifying laboratory and clinical test results.* The entire term contained in the LOINC represents specific information related to test components.
7. **SNOMED**. SNOMED CT or SNOMED Clinical Terms is a systematically organized computer-processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED functions an ontology.
8. **RadLex**. *RadLex radiology lexicon represents a schema of radiology terms. It is utilized by LOINC.*
9. **MedDRA**. *Medical Dictionary for Regulatory Activities (MedDRA) is a schema that standardized medical terminology for medicinal chemical entities.*
10. **GO**. *Gene Ontology (GO) is an ontology that represents functionality of genes, namely the role of specific genes when involved in cellular components, biological processes, or molecular functions.*
11. **BRO**. *Biomedical Resource Ontology (BRO) is an ontology used to enable semantic annotation and discovery of biomedical resources.*
12. ***SSIMO.*** *A metadata ontology that represents metadata for various NIH projects.* This is the internal abbreviation for the SI Service Metadata Ontology.

Currently, OMO is used for annotations and object relationships that involved metadata or ontologies. OMV is used to represent class entities related to organizations and individuals. The ChAO annotation is used for classes related to components of an ontology. We have standardized the different relationships in accordance to an Ontology Design Pattern (ODP). An example of how a specific ontology or class following the ODP is shown as part of **Figure 2.**

**Figure 2. Utilization for Ontology Design Pattern for SSIMO.** The ODP provides the layout for how classes and individuals will be annotated with object and data properties within SSIMO. Each box represents a class, data, or individual that are connected by different properties. The arrows represent an RDF Triple that is formed between boxes and the property listed in the middle of the arrow. The arrowhead points to the object of this RDF Triple. Italized property names indicate that this relationship is not included in SSIMO but should be the result of a successful output of the SSIMO API tool.

The primary definitions for these terms are included for rest of the document.

**Upper Level Classes.**

**ObsoleteClass.** A category for terms no longer supported and replace with a different term in the system. Legacy IRIs are included for reference of earlier mistakes.

**Party.** *A party is a person or an organisation.* This class is used primarily to identify the releveant authority for an information resource. Party is intended to be linked to data that represent contact information, such as mail address, emails, phone numbers, etc.

**Ontology Domain.** *The subject of knowledge that is covered by an information content entity.* The domain of an ontology can cover material entities, events, processes, and metadata. Currently, we have broad domains included in SSIMO but more will be added once we have decided how to divide the fields of knowledge covered in the various NCI knowledge bases and ontologies.

**Information Content Entity.** *A generically dependent continuant that is about some thing.* An information content entity refers to the data or information stored within documents, figures, or databases. This is a term is imported from OMO. All following classes are considered Information Content Entities.

**SI-Ontology.***Metadata that describes the components of an ontology.*

**Ontology.** *An information content entity that is a knowledge graph which contains a central hierarchy and connected via RDF triplets.*

**Formality Level.** *The level of formality of an ontology.* This term is used to describe for differences between informal definitons and more stringently defined meanings.

**OntologySyntax.** *The format an ontology utilizes in its construction.* The structure of RDF and OWL are examples of Ontology Syntax.

**Ontology\_Component.** *The entities and relationships that are used in an ontology.* The three major components of that are defined in an ontology are class, individual, and property.

**OntologyTask.** *A task that an ontology is intended to utilize.* This metadata describes the role that an ontology is intended to participate in.

**OntologyType.** *A descriptor of the type of an ontology.*

**Ontology Hierarchy.** *A component of an ontology that is the hierarchical knowledge graph that stores knowledge in the form of triplets.* Figure 1 contains the Ontology Hierarchy for SSIMO using Protégé. Other terms can be used instead of ‘is a’ to construct this hiearchy, such as ‘type’.

**SI-Resource.** *An information content entity that is used as a resource for an ontology.* The usage refers to other ontologies, thesauruses, and schemas that are used by other databases. More specifically, all si-resources were consulted during the development of this project for mappings.

**SI-Resource-Metadata.**

*An information content entity that contain a resource to describe resource-metadata. The usages refers to the components found for the data of studies.* All resources utilize this metadata, either as part of the schema or as part of individual entities. These are common features that are common or saved with all ontologies stored.

**Definition.** *The metadata that explains the units or data points measured as part of a study.* The typical use of definition is to explain types of studies, experimental types, or data from experiments.

**Identifier.** *A set of symbols which is used to designate an individual information content entity.* These specifically referred to unique terms for a specific study or data set. Examples of identifiers include specific IRI.

**Term.** *A common name or synonym that describes a dataset or data entry.* This will be used as one input to find multiple possible entries within the system. Dataset or entries can have multiple terms.

**Preferred Term.** *A term that is the one that is used as the default term id.* This will be the default name used. Additional terms are linked through use of different annotations.

**Schema.** *The format that the resource is set up as. This format is done for both databases or studies.* The primary examples of schema include LOINC data codes, OBO formating, MeDRA hierarchy.

**Reference.** *The original source of information that is used to justify the creation of an entity.* This could be to identify either a party that is responsible for information or another knowledge graph or another information content entity that contained this information.

**Properties.**

**Object Properties.**

**is a.** A property that is used to link a class to a class or a property to a property. ‘Is a’ refers to a relationship where the subject is contained within the set of the object. Protégé uses the term to establish the hierarchy of the ontology.

**Is about.** *A (currently) primitive relation that relates an information artifact to an entity.*

**type.** A property that is used to link an individual to a class where the individual is contained within the set of the class.

**Data Properties.**

**has authority**. *A data property that links a resource or ontology to a party. The party that ‘has authority’ over some information content entity has a role that is realized in approving changes or policy over the second entity.*

**has conversion authority.** *A data property that links a party to an information content entity where the resource or ontology is converted from one data type or structure to another.* This is used to account for which authority converted the resource into an RDF, either directly or after flattening the representation of data within the resource.

**has source authority.** *A data property that links a resource or ontology to a party where the party is a source for some data.* This is used to identify the party is responsible for the creation and maintenance of the ontology.

**has contact data.** *A data property that links a party to an information content entity where the information content identity has metadata or data to describe the procedure to contact the party.* This property should correspond to a string or set of integers, depending on the type of contact data.

**has license.** *A data property that is used to link an entity to a license such that the license governs the terms of use permitted for that entity.* This property is intended to be used to link an SI-Resource to a license name, but could have a different entity for the domain.

**has name space.** *A data property that designates the origin of the name space for a class, individual or property.* This property is used to account for situations where a class, individual, or property is imported or used in multiple ontologies. This object property is intended to be used to guide SPRAQL queries.

**has metadata.** A data property that link an SI resource to metadata.

**has namespace.** *A data property that designates the origin of the name space for a class, individual or property.* The use of this property is either identify the original graph, or of information stored within the schema of a system. T

**has version Date.** *A data property that points to the IRI property that identifies the date of version release.* The object for this relationship should point to the dc:date annotation. This annotation property should be utilized in the API tool but not saved in SSIMO.

**has version Info.** *A data property that points to the IRI property that identifies the instance of version release.* The object for this relationship should point to the owl:version. This annotation property should be utilized in the API tool to identify a single instance of an SI resource. However, this should not be saved in SSIMO.

**Applications of NCI SI Metadata Ontology**

**Competence Questions.**

Below is a list of competence questions to be addressed using SSIMO. Currently, each competence question is answered with the procedure used by the API tool to answer the competence question.

Further use of these questions will be used to guide the use of the API tool:

* How do we get all related information about a term?
  + All related information for a term is found as part of any annotations or axioms that use the term as a subject.
* How do we get the list of all terms in the service?
  + All terms in the service are those that are listed in an RDF triple with ‘have name space’ for any terms in the service.
* What resources are found in the service?
  + All terms found in the service are listed as children of SI-Resources.
* What authority created the resource?
  + All terms that are listed in an RDF triple with the specific resource and ‘has source authority’.
* What authority converted the resource to RDF?
  + All terms that are listed in an RDF triple with the specific resource and ‘has conversion authority’.
* What is the version of the resource?
  + All annotations that are listed in an RDF triple with the specific resource and ‘has version info’.
* What are the names of the "common" properties (i.e. preferred term, term, textual definition, identifier) used in the "globals" graph?
  + These common properties are listed in Mapping of Si-Resource Annotations (see below).
* What is the type of the resource?
  + The resource is either a child of si-resource-type. This will be linked by an axiom to the class that represents each resource.
* What is the domain of the resource?
  + The domain is a child of Ontology domain. This will be linked by an axiom to the class that represents each resource.

**Mapping of SI-Resource Annotations.**

As part of the process for creating new terms, we went through NCIT resources to map annotations or relationships from each NCIT resource to the appropriate SSIMO terms. It is common for many SI-resources to have multiple terms that map to a single SSIMO class or vice versa. The origin of each term is listed in parenthesis afterwards. NCITMetathesaurus has many annotations which implicitly identify an Ontology Domain due to identifying another, more specific ontology. These terms, for the sake of brevity, are excluded.

**Identifier (SSIMO):** Loinc name (LOINC), IRI (GO), Concept ID (SNOMED), Radlex ID (Radlex), ID (ndrft), MeDRA code (MeDRA), CTCAE Term (CTCAE) UAN, UCN, UE, USN, NA (NCITMetathesaurus).

**Preferred Name (SSIMO):** Full specified name (SNOMED), prefLabel (ndrft), Preferred Name (Radlex), DN, MTH\_PT, MTH\_GP, (NCITMetathesaurus).

**Definition (SSIMO):** Loinc Name (Loinc), Definition (GO), Defining Relationship(SNOMED), Definition (Radlex), Definition (CTCAE), DE (NCITMetathesaurus).

**Synonym (SSIMO):** short name, long common name, FSN name (LOINC), has\_exact\_synonym (GO), synonyms (SNOMED), EQ, ES, ETAL, Synonyms, MTH\_SYN (NCITMetathesaurus).

**Reference (SSIMO):** code source (LOINC), Definition (GO), Definition Source (NCITMetathesaurus).

**Ontology Domain (SSIMO):** Loinc name (SSIMO), has\_obo\_namespace (GO), MeDRA code AND MeDRA level (MeDRA), CDO (NCITMetathesaurus).

**Mapping of Features within the Shared SI Service Query**

Currently, there is a public SPARQL endpoint driven by open source Virtuoso Web Site Linking Policy and accessible from https://shared.semantics.cancer.gov/sparql. It is meant to provide an integrated view of the terminology and data elements produced and/or published by the Cancer Data Standards Registry and Repository (caDSR) and Enterprise Vocabulary Services (EVS) projects in the Semantic Infrastructure group. We anticipate that eventually it will include various other datasets of utility to the NCI. This endpoint will be serviced through SSIMO.

**Table 1. Common Prefixes Declared Within Shared SI Service.**

|  |  |
| --- | --- |
| **Prefix** | **IRI** |
| dc | http://purl.org/dc/elements/1.1/ |
| owl | http://www.w3.org/2002/07/owl# |
| rdf | http://www.w3.org/1999/02/22-rdf-syntax-ns# |
| rdfs | http://www.w3.org/2000/01/rdf-schema# |
| skos | http://www.w3.org/2004/02/skos/core# |
| void | http://rdfs.org/ns/void# |
| xml | http://www.w3.org/XML/1998/namespace |
| xsd | http://www.w3.org/2001/XMLSchema# |

The Shared SI Service SPARQL editor list a set of three non-internal graph names: http://cbiit.nci.nih.gov/caDSR, http://ncicb.nci.nih.gov/xml/owl/EVS/Thesarus.owl, and http://ncicb.nci.nih.gov/xml/owl/EVS/Thesarus.rdf. The caDSR list document name last two graphs are variations of each other that primarily differ in which ontology language is used; the Thesaurus.owl utilizes OWL and the Thesarus.rdf utilizes RDF. These graphs together share a common set of prefixes (Table 1). Additionally, the caDSR graph name contains a set of different namespaces which correspond to the original ontology that each term is located from (Table 2). SSIMO is capable of modeling these relationships through the use of the following data annotations.

‘CaDSR’ ‘has namespace’ ‘http://cbiit.nci.nih.gov/caDSR# ’

‘

#### Table 2. Namespaces referenced in caDSR RDF.

|  |  |
| --- | --- |
| Description | IRI |
| caDSR | http://cbiit.nci.nih.gov/caDSR# |
| ISO 11179 | http://www.iso.org/11179/MDR# |
| MADS/RDF | http://www.loc.gov/mads/rdf/v1# |
| NCI Thesaurus | http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl# |
| NCI Metathesaurus | http://ncim.nci.nih.gov/NCIMetathesaurus.owl# |
| MGED Ontology | http://mged.org/MGEDOntology# |
| NDF-RT | http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl# |
| CTCAE v5 | http://ncicb.nci.nih.gov/xml/owl/EVS/ctcae5.owl# |
| LOINC | https://loinc.org/code# |
| SNOMED | http://snomed.info/id# |
| RadLex | http://radlex.org/RID/ |
| caBIG | http://ncicb.nci.nih.gov/cabig# |
| MedDRA | https://identifiers.org/meddra# |
| Gene Ontology | http://purl.org/obo/owl/GO# |

Procedural Updates.

1. Utilize common yaml file.

2. Triplicate properties as into class, individual, and annotation.

3. Use SSIMO as upper level ontology primarily.