**Development and Applications of NCI SI Service Metadata Ontology**

**Table of Contents / Wiki Pages**

* **Main Page / Introduction**
  + Introduction of CANSI
  + 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
  + Use cases
  + Usage for GO API tool development
* **Discussion**
* **Funding / Acknowledgments**

**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. NCI CBIIT has the fundamental objective of supporting 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 an 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. The SPARQL RDF query language is used to retrieve data stored in a RDF triple store.

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. The metadata ontology 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 the metadata ontology 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.

**Source Code:**

Ontology is available for …

Ontology is available under GitHub website:

https://github.com/NCI-Semantic-Infrastructure/si-schemas/tree/main/metadata-SI

https://github.com/NCI-Semantic-Infrastructure/si-schemas/blob/main/metadata-SI/metonto2.rdf

note: we may need to rename it later.

**License:**

License is TBD.

**Funding/ Acknowledgments:**

This project is funded by NCI CanSI project, with the funding number: …

**Team:**

Our team includes the following members:

NCI members:

* Zhengwu Lu, technical project manager.
* Gilberto Fragoso, government project manager

He Lab in the University of Michigan (UM)

* Michael Cooke, RESTful API developlment
* Anthony Huffman, metadata ontology development, QA
* Jie Zheng, QA, Semantic infrastructure development and testing.
* Oliver He, UM side project manager and involves different tasks.

**CANSI Metadata Ontology Development**

**Definitions of Related Ontology Terms**

**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.

**Individuals.** 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, version 1.0.7 of Service Metadata Ontology would be an individual, while the Service Metadata Ontology would be a class.

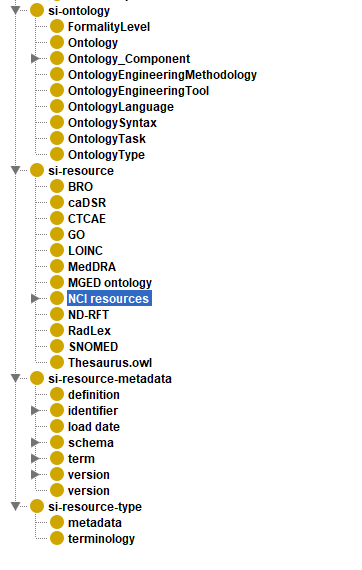
**Properties.** A property is a verb phrase which describes the relationship between two sets of entities. There are three kinds of properties: an annotation property links an entity to an annotation. An object property links a class or individual to another entity. A data property links a class or individual 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.

**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 and an be an annotation, class, data, or individual.

**RDF Datasets.** A set of RDF Graphs that is made out of of one default graph and potentially multiple zero named graphs.

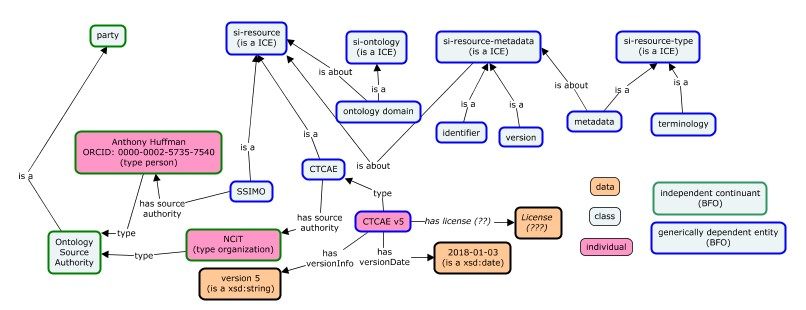
**Metadata Ontology design**

**Figure 1**. High Level Representation of CANSI Metadata Ontology Tool. New terms were added to si-ontology, si-resource, si-resource-metadata.

We generated the initial ontology using the pre-existing 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 ontolgies 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 knowledgebases we have queried, along with their defintions, 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. Ontology Design Pattern Example 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 ‘is a’ relationship is used to represent the child-parent relationship between classes. The ‘type’ relationship is used to represent child-parent relationship between individual-class. Additional child-parent relationships are included within each box. Finally, proposed mapping for each class and individual to existing Basic Formal Ontology (BFO) are included, but not explicitly integrated into SSIMO.

The primary definitions for these terms are included for rest of the document. Children for each upper level term is included in smaller text below each term.

**Upper Level Classes.**

**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.

**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.

**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*

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

are examples of Ontology Syntax.

**Ontology\_Component.** *The components of an ontology.* The three major components of an ontology are class, individual, and property.

**OntologyTask.** *A task that an ontology is intended to utilize.*

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

**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. All preferred terms are terms.

**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.

**Properties.**

**Object Properties.**

**has graph names.** A property that links data to a resource where that data is located from the database. This can also be done for SPRAQL usages.

**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 source authority.** A data property that links a resource or ontology to a party where the information content entity is a source for some data.

**has conversion authority.** A data property that links a resource or ontology to a party. A data property that links a party to an information content entity where the resource or ontology is converted from one datatype to another.

**Data Properties.**

**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.

**has license.** A data property that links an SI resource to a liscence.

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

**has namespace.** *A data property that desginates the origin of the name space for an 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.

**has version Date.** *A data property that points to the IRI property that identifies the date of version release*: this format can be stored as a string or set of integers. This annotation property should point to the annotation for owl versionInfo.

**has version Info.** *A data property that points to the IRI property that identifies the instance of version release*: this format can be stored as a string or set of integers. Each version info is mapped to a single instance of an SI resource. This should be pointing to version dc:date.

**Applications of NCI SI Metadata Ontology**

**NCI Shared Semantic Infrastructure Services**

#### [Service and Datasets](https://shared.semantics.cancer.gov/documentation.html" \l "datasets) [↰](https://shared.semantics.cancer.gov/documentation.html)

This is a public SPARQL endpoint driven by open source [Virtuoso](https://github.com/openlink/virtuoso-opensource) [](https://www.cancer.gov/policies/linking) 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.

The endpoint is fronted by this web-site which provides a [SPARQL query editor](https://github.com/TriplyDB/Yasgui) [](https://www.cancer.gov/policies/linking) for ad hoc queries to facilitate testing queries and data exploration, documentation (this page), sample queries, and various file downloads. This can be seen as part of Figure 2.

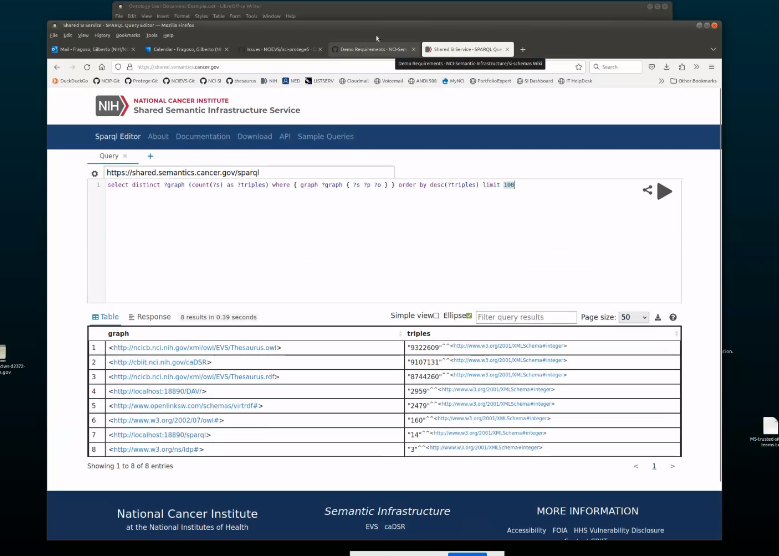


Figure 2.

The web server running this web site also examines queries sent to the endpoint for various issues. This has two consequences:

1. The rules might on some cases prevent valid queries from executing.
   * If you believe that your query is valid, please contact us at [NCIAppSupport@nih.gov](mailto:NCIAppSupport@nih.gov?subject=Request to Review SPARQL Query), we will review your query and either rewrite it or edit our rules to allow your query to run.
2. Add to the response time, which will depend on the query itself and the size of the resultset and can add between 20 and 500 msecs to the response.
   * If this is an issue for you, and your application runs from within the NCI network or externally on behalf of an NCI program, please contact us at [NCIAppSupport@nih.gov](mailto:NCIAppSupport@nih.gov?subject=Request to Safelist) to be safelisted. We will contact you for additional information.

The current datasets in the endpoint are:

* A subset of the data in caDSR, primarily the Data Elements *sans* some administrative information (no forms are included). Monthly XML exports from the caDSR are taken and converted to RDF and loaded into the quadstore; the monthly export taken for processing is from the first day of the month.
* The NCI Thesarus (NCIt) in a flat RDF representation. Monthly distributions of the NCIt in OWL DL are processed and the flat RDF generated. This RDF consists of a conversion of class expressions to simple assertions so that, with the exception of owl:Axioms, blank nodes are eliminated.

We are not currently making the RDF available but could periodically post a downloadable version to serve as an example of the dataset(s). Please let us know if this is of interest to you via email at [NCIAppSupport@nih.gov](mailto:NCIAppSupport@nih.gov?subject=Shared SI Service RDF Downloads).

**Table 1. Currently used graph names in SSI**

|  |  |
| --- | --- |
| 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# |

**Table 2. Common prefixes already declared in SSI**

|  |  |
| --- | --- |
| 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# |

**Table 3. Namespaces refenced 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# |

**Competence Questions:**

Below is a list of competence questions to be addressed using the NCI SI metadata ontology:

* How to get all related information about a term?

**Use Cases:**

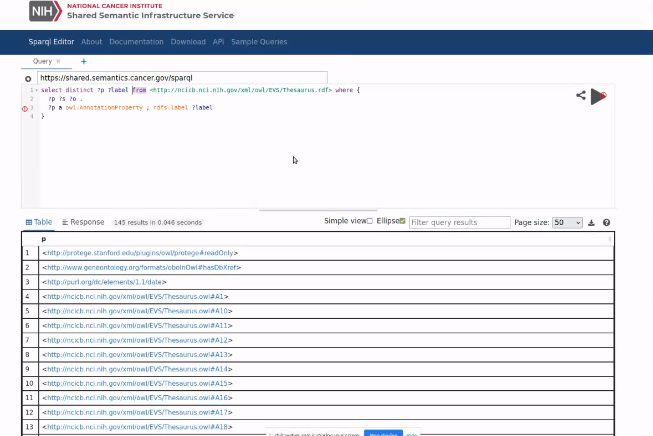
**Table A. Mapping of high-level annotation terms.** Terms that are underlined represent information that is accessible within a specific type of metadata, but is not an onto mapping of the CANSI term to the other source ontology. Italized terms represents object properties instead of annotations. GO annotations are stored as axioms within the internal documentation. NCITMetathesaurus annotations import from several ontologies and primarily use prefixes to identify these terms; the use of specific prefixes are used to identify the domain of each term. Only general terms or terms unique to NCITMetatheasus are shown in Table A.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CANSI** | **LOINC** | **GO** | **SNOMED** | **Radlex** | **ndrft** | **MeDRA** | **NCITMetahesasurus** |
| Unique Identifier | Loinc names | IRI | Concept ID | RadLex ID | ID | Medra code | UAN, UCN, UE, USN |
| Preferred Name |  |  | Full specified name |  | prefLabel |  | DN, MTH\_PTGB, MTH\_PT |
| Synonyms | Short name, long common name,  FSN names | has\_exact\_synonym | synonyms |  |  |  | Synonyms,  MTH\_SYN |
| Definition | *Loinc name* | Definition | Defining relationship | Definition |  |  | DE, MTH\_FN |
| Parent | class | *is\_a* |  | Parent |  | Parents | Parent |
| Child |  | is\_a |  |  |  | Children | Children, MTH\_LLT |
| Reference | Code source |  |  |  |  |  | Source, Definition Source |
| Domain | Loinc name | has\_obo\_namespace |  |  | *ID* | *MeDRA code,*  *MeDRA level* | SCD |

***Notes:***

* The above table is confusing. Probably do not use table as the form. Instead, you can separate it to lists for individual rows.
* **Gilberto**: The globals is meant to support queries on common properties, rather than specific properties for each vocabulary.

Will people need to use IRI or string for work. This will depend of information on authority/annotations for other information..

Our intial use cases

NCIT has all non-private information as part of the vocabulary. RDF lacks restrictions and axioms that the OWL conversion is found, it would be moved to instances.

Distinction … development and formalism…. Conversion authority is shared service metadata ontology.

caDSR (Converted into OWL). Do we distinct RDF/OWL.

**Discussion**

List of

**GO API tools**

List of