White Paper

Introduction

Methods

Tools

The ontology is designed in Protégé (Stanford) and published at the Bioportal (NCBO). The ontology and example datasets were implemented in Sesame (OpenRDF) with OWLIM-lite (Ontotext) at sparql.cancerdata.org.

License

Conventions and best practices

URI convention

The ROO adheres to the following convention for all Unique Resource Identifiers (URI) www.cancerdata.org/ROO/[UniqueNumber]. All URIs will be resolved in future to a human readable webpage which describes the entity on www.cancerdata.org.

Minimal description of entities

Each entity shall have at least one human readable English language label specified in the rdfs:label annotation. In the case of an entity imported from an ontology the rdfs:isDefinedBy should be stated (see below). When the ROO defines the entity, the label of classes shall be in Title Case and the label of properties shall be in underscore_case.

Deprecation

When the ROO deprecates entities the following annotations shall be asserted.

- owl:deprecated shall be set to true
- rdfs:comment shall be asserted with the reason why the entity is deprecated
- [if there is an entity that replaces the deprecated entity] rdfs:seeAlso shall be asserted and point to the replacement entity.

Upon deprecation all assertions associated with the deprecated entity shall be kept but no longer maintained in future editions of the ROO. Up until version 1.0 of the ROO the deprecated classes and properties may even be removed.For viewing purposes a deprecatedClass and depecratedProperty entity are defined and all depecrated entities are a direct subClass of this class. This deprecatedClass and depecratedProperty are only created for viewing purposes.

Design principles

OWL2 QL profile

The W3C has described a number of OWL profiles (http://www.w3.org/TR/owl2-profiles/) each a subset of the full OWL 2 standard which trade-off expressive modeling power for computational efficiency when performing reasoning. The ROO adheres to the OWL 2 QL ("Query Language") profile. This profile was chosen because the ROO aims to link clinical and research data which often sits in (various) relational databases inside health care and research institutions which are best queried using the SQL query language. The OWL 2 QL profile guarantees that all ROO queries can be fully rewritten as SQL queries.

High level organization

The high level structure of the ROO is based on the UMLS Semantic Network through the use of the Semantic Types (classes) ontology (http://bioportal.bioontology.org/ontologies/STY/?p=summary) and the assertion of the Semantic Relations (properties) as specified by the UMLS (https://uts.nlm.nih.gov/). The UMLS Metathesaurus assigns a semantic type to each of the concepts it covers (e.g. the NCI Thesaurus concept). This semantic type is the preferred superclass of such a concept. However deviations as well as mutliple superclasses are allowed.

Multiple URI for the same concept

The ROO allows the specification of more than one resource for the same entity as that may help in mapping commonly used ontologies. An example is the concept of lung cancer which can be specified using ICD (C34) and NCI Thesaurus (C7377). In such a case the ROO specifies the equivalence of classes and properties using owl:equivalentClass and owl:equivalentProperty.

Re-use of other ontologies

The ROO re-uses as much as possible entities from other ontologies such as the NCI Thesaurus, ICD, Unit Ontology etc. provided that they are published at NCBO's Bioportal and are provided without any restrictions (so this excludes an ontology that requires a license such as SNOMED CT and MedDRA). The ROO uses the original URI for these imported entities. Besides the URI, the ROO states the rdfs:label annotation which is a copy of the label the source ontology specified at the time of import of the entity into the ROO. The ROO also asserts the rdfs:isDefinedBy annotation property which is the resolvable URL of the source ontology at the Bioportal (e.g. http://bioportal.bioontology.org/ontologies/NCIT for the NCI Thesaurus). The ROO does not assert any other axioms such as hierarchies and object properties of the source ontologies, may implement additional axioms and may implement axioms that are not consistent with the source ontology.

There is a preference in ontologies

- National Cancer Institute Thesaurus (NCIT)
- Units of Measurement Ontology (UO)
- Foundational Model of Anatomy (FMA)
- Semantic Types Ontology (STY)
- Semantic DICOM Ontology (SEDI)
- International Classification of Diseases, Version 10 (ICD10)

Besides these formal ontologies the ROO includes the following legacy ontologies or radiation oncology

- euroCAT: A legacy ontology used in the euroCAT project (www.eurocat.info). The euroCAT entities are all deprecated.
- Standardizing Naming Conventions in Radiation Oncology: A paper from multiple societies/groups that describes a way to standardize names for regions of interest in radiation oncology (http://dx.doi.org/10.1016/j.ijrobp.2011.09.054). This convention is specified as a subclass to Radiation Oncology Region of Interest. Each Organ at Risk is considered to be a delineation of a body part, organ or organ component.
- [in future] ICRU reports

Assertions that something is false

The ROO adheres to the open world assumption of the Semantic Web which states that the absence of statement does not mean the statement is

not true. A typical example being the patient did not receive chemotherapy. Using the ROO, the absence of a group of statements of the form

```
ex:Patient1 ex:hasProcedure ex:Procedure1 .

ex:Procedure1 rdf:type ex:Chemotherapy .
```

is not sufficient to conclude that a patient did not have chemotherapy. Rather it has to be explicitly defined that a patient belongs to the class of things that did not receive chemotherapy using owl:Restriction and owl: complementOf like:

```
ex:ExampleNoChemo rdf:type owl:NamedIndividual ,
                  [ rdf:type owl:Class ;
                    owl:complementOf [ rdf:type owl:Restriction ;
                                       owl:onProperty ex:hasProcedure ;
                                       owl:someValuesFrom ex:Chemotherapy
                                     ]
                  ] .
```

Representing specified values, attributes and units

One needs to be able to specify attributes such as units, or modifiers such as severity etc. Suppose one would like to state that a measurement of dose has the unit Gray. There are two basic ways to represent this in OWL (see also http://www.w3.org/TR/swbp-specified-values/). Either as sets of individuals (e.g. the unit “Gray” is a owl:NamedIndivual) or as subclasses (e.g. the unit “Gray” is a owl:Class). In the ROO we choose the latter. E.g. to state that Measurement1 was 50.0 Gray it should be asserted that Measurement1 is restricted to the things that have Gray as its unit.

```
ex:Measurement1

  roo:has_value “50.0”^^xsd:double;

  rdf:type [owl:Restriction;

            owl:onProperty roo:has_unit;

            owl:someValuesFrom uo:gray]
```

Representing date and time

To represent data and time the following conventions is

```
ex:Measurement1

  roo:at_date_time "2002-05-30T09:00:00 "^^xsd:dateTime
```

Results

Discussion

Implementers of this ontology should choose tools which have it is important

Colophon

Title

The Radiation Oncology Ontology (ROO): an ontology for publishing Linked Data for radiotherapy

Authors

Andre Dekker [1], Johan van Soest [1], M. Scott Marshall [1], Wolfgang Wiessler [2], Ying Xiao [3], Jiazhou Wang [4], Zhen Zhang [4], Lois Holloway [5], David Thwaites [6], Andrea Damiani [7], Vincenzo Valentini [7], Geetha Mahadevaiah [8]

Affiliations

[1] MAASTRO / [2] Varian / [3] RTOG / [4] Fudan / [5] Liverpool / [6] Sydney / [7] Gemelli Rome / [8] Philips

Keywords

Semantic Web, Ontology, Linked Data, Radiotherapy, Cancer

Action Items

- OWL QL?
- •Ok for now, but we need to get a better understanding of how severe the limit is
- •If QL is too limited, the preference is to leave QL. But the problem may be that the queries take too long. IN that case we will and not use query rewriting
- •Alternative is online reasoning or pre-assertions
- Access to git for Wolfgang?
- Assumption in terms of infra
  1.We don't allow patient data leak to the outside world
  2.High level hardware at each site is not going to work
  1.Only basic reasoning is doable RDFS with NCIT
  3.Automated learning application only needs to query the patient data endpoint
  1.Pre-assertion of inferred triples possible (just RDFS reasoning), distributed sparql does not need inferencing
  2.Expanded SPARQL both in the WHERE (easy) and in the SELECT (poss. with CONSTRUCT)

  subClassOf vs. descendents

  Disease1 a ncit:StageII_NSCLC

  Patient 1 :has Disease 1

  :hasNSCLC rdfs:range ncit:NSCLC

  ncit:StageII_NSCLC rdfs:subClassOf ncit:NSCLC

  Infer: Patient 1 :hasNSCLC Disease 1

  -> WHERE ?pat :has ?dis

  ?dis rdfs:subClassOf* ncit:NSCLC

  Disease1 a ncit:StageII_NSCLC

  :Patient 1 :hasNSCLC :Disease 1

  :hasNSCLC rdfs:range ncit:NSCLC

  ncit:StageII_NSCLC rdfs:subClassOf ncit:NSCLC

  Infer: Disease1 a ncit:NSCLC

  -> WHERE ?pat roo:has* ?dis
- No argument against more specific properties.
- Diagnosis1 :atDate "1-1-2014"

  Diagnosis1 :found Disease1

  Like to infer:

  Disease1 :atDate "1-1-2014"

  >>Decision Diagnosis1 :atDate Date1

  Date1 :hasStartDate/hasEndDate/hasExactDate"1-1-2014"

  Diagnosis1 :found Disease1
- Central service

  Learning connector ok solution for many sites

Timeline->Wolfgang

License, duCAT?

Open Issues

- •Re-use of other ontologies NCI Thesaurus, SNOMED CT, ICD, Unit Ontology etc. (cont.)
- •Is it feasible?
- •Import or central service
- •Federation across services
- •URI convention
- •http://www.cancerdata.org/roo/[unique number]
- •http://www.cancerdata.org/cohort/0001/Patient123
- •URIs resolve to a human readable webpage which describes the entity on www.cancerdata.org
- •# vs. /
- •Extend it
- •Umbrella protocol
- •DVH
- •Naming conventions
- •Data provenance
- •1) :hasDataSource (tricky for literals)
- •>> 2) Separate graphs
- •Inferred vs asserted

Misc notes (will be deleted soon)

Use: Share and learn from distributed datasets coming from EHR and related data sources in radiotherapy centers

Reasoning

We will expect all implementers of the ROO to switch on reasoning with the following OWL constructs supported

What type of questions:

- Is the outcome of lung cancer patients with stage IV disease different from patients with Stage I disease?
- Is the patient population of center x different from center y
- Are outcomes in center x similar to the outcomes in center y.

Who will use:

- Researchers in radiation oncology
- Registries during data transfer

Who will maintain: ESTRO – ASTRO

Simple and Complex properties

Necessary and Necessary and Sufficient conditions

Property constraints

If a class has one child, modelling error.

If a class has 12+ children, modelling error

Decide which OWL 2 EL, QL, RL?

Version use in protégé important

Example query for not-punning

PREFIX roo:<http://www.cancerdata.org/roo#>

PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

PREFIX ncit:<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>

PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

PREFIX owl:<http://www.w3.org/2002/07/owl#>

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

```
SELECT ?person ?ageValue ?ageUnit
WHERE {
?person ?hasAge ?age.
?hasAge rdfs:label "hasAge"@en.
?age ?hasValue ?ageValue.
?hasValue rdfs:label "hasValue"@en.
?age a ?x.
?x owl:onProperty roo:ROO100008.
?x owl:someValuesFrom ?unit.
?unit rdfs:label ?ageUnit.
}
```

Class Matrix tab for a number of labels.

Je kun tin v3 een deel van de ontologie inladen

Don't use SWRL for things you can do in OWL.

SWRL moet een reasoned hebben die werkt, is dat in Sesame?

Man hasGender some Male?

Stating that something is not the case

Suppose you would want to say that a patient did not (ever) receive chemotherapy. They way to do this is to use a complement of things that did receive a chemotherapy procedure.

```
:ROO100024 rdf:type <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C16960> ,
                owl:NamedIndividual ,
                [ rdf:type owl:Class ;
                  owl:complementOf [ rdf:type owl:Restriction ;
                                     owl:onProperty :ROO100005 ;
                                     owl:someValuesFrom <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C15632>
                                   ]
                ] ;

        rdfs:label "ExampleNotChemo"@en .
```

Versioning needs to be there, deprecation, not renaming, backward compatibility

List of supported ontologies resolved implementers should resolve those

Data provenance euroCAT in euroCAT has data sources, trust

Stating something about statement hasDate?

Timing down in OWL/SPARQL SWRL ? Before/after chemotherapy

ex:hasValue “50.0”^^xsd:double;

CONSTRUCT

?Person roo:hasEvent Birth1.

Birth1 roo:atDateTime ?BirthDate

WHERE ?Person roo:hasBirthDate ? BirthDate

ex:Patient1 rdf:type

owl:NamedIndividual,

rdf:type [owl:Restriction;

owl:onProperty ex:hasEvent;

owl:someValuesFrom

rdf:type ex:Birth, rdf:type [owl:Restriction; owl:onProperty ex:atDateTime “2008-12-12”]

ROO: In a conformance statement it should be specified that says which ontologies

ROO: Have to have query wring SW objects so the following namespaces are supported, federation across services

?? investigate SWobjects license ?

Patients

For testing purposes a patient datasets were chosen to showcase the main features of the ROO. The selected datasets were a) 377 consecutive stage I-IIIB inoperable NSCLC patients treated at MAASTRO Clinic between May 2002 and 2006 which were used as the training set for a prognostic model for 2-year survival, b) 407 patients treated XXX which were used to build a model for radiation-induced dyspnea and c) 469 patients treated XXX which were used to build a model for radiation induced dysphagia. As there are patients which occur more than once in a dataset, a total of XXX unique patients were available. The data can be found at www.cancerdata.org and sparq.cancerdata.org and can be used for research purposes. The sharing of data was approved by our Internal Review Board.