Protege, BioPortal, CEDAR: Building Complete Pipelines for Engineering Metadata
Introduction

John Graybeal
Technical Program Manager
CEDAR & BioPortal

Stanford Center for
Biomedical Informatics Research (BMIR)

- Science computing support years
  - 19: astronomy
  - 12: marine
  - 5: biomedicine
- Discipline years
  - 12: operations
  - 10: cyberinfrastructure
  - 13: metadata
  - 15: semantics
- lots: project lead
- Project Lead, Marine Metadata Interoperability Project and MMI Ontology Registry and Repository
In this talk

1. Make the case for the importance of standardized semantics in metadata.

2. Describe how to build and make available standardized semantics (e.g., vocabularies and concepts).

3. Show you how to easily build most metadata models but with standardized semantics.
Part 1:
Why are standardized semantics important?
Why?
Observatory Topology

Sensor Platforms
- Research Platforms
- Fixed Core
- Mobile Core

Physical Interface
- Marine Operations
- Shore Side Operations
- Observatory Operations
- End User Operations

Marine Net
- Marine Management
- Observatory Net
- Observatory Management

Observation Topology

Science Teams & Educators
- User Environment Management

Marine Operator
- Cyber Operator
- Administrator
from Research Data Repositories: The What, When, Why, and How by Ray Uzwyshyn
https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml
What do we want that data to look like?
We want research data to be self-describing.
“We need to develop one universal [metadata] standard to cover everyone’s use cases.”

–XKCD #927
How standards proliferate:

Situation: There are 14 competing standards.

14?! Ridiculous! We need to develop one universal standard that covers everyone's use cases.

Soon:

Situation: There are 15 competing standards.
What is ‘metadata’ again?

- It’s not just ‘data about data’—e.g., “here are my definitions of what my values mean” (though that helps)
- It’s about origins / provenance / context of the data
- This can include things like…
  - a precise description of how to parse data streams
  - the reason we collected the data streams (and who paid)
  - basically, any context related to the metadata
A story about ocean temperatures...

from Correcting Ocean Cooling, a NASA Earth Observatory blog post (https://earthobservatory.nasa.gov/features/OceanCooling)
A story about ocean temperatures...
A story about ocean temperatures…

Biased Measurements

Argo Float

XBT Instrument
A story about ocean temperatures...

[Map showing changes in heat content with color bars indicating change in W/m².]

Biased Measurements

Argo Float, XBT Instrument
A story about ocean temperatures...

1993-2003

2003-2005

'04-'06 Temperature Change at 500m (°C)

Change in Heat Content (W/m²)

-60 -30 0 30 60

-6 -3 0 3

Argo Float  XBT Instrument

Corrected

Biased Measurements
A story about ocean temperatures...

Change in Heat Content (W/m²)

1993-2003

2003-2005

Corrected

Uncorrected

Argo Float

XBT Instrument

Biased Measurements

original data

revised data

%04-'06 Temperature Change at 500m (°C)
A story about ocean temperatures...

The method gives detailed information on biases for specific vintages of XBTs, but **this information cannot be used with confidence to correct the majority of XBT profiles archived that are lacking critically important metadata**.

Depth-dependent corrections were introduced for each XBT temperature observation (the corrections were obtained by comparing XBT and CTD/Nansen bottle anomalies for overlapping 2° × 4° boxes). **Though a cruise-wise correction of the XBT is preferable, it remains beyond the scope of this study because of the absence of collocated high quality data and of the respective metadata (manufacturer, probe and acquisition system type).**

---


How much metadata do we have to keep?

Do we forever follow the “Provenance Trail” to keep any data-related context?
Metadata (redefined): “Information useful to work with data”
How do we decide what metadata to provide?

- Given our metadata definition: “Information useful to work with data”
  - The question becomes: useful to whom?
- To compensate for human beings’ lack of imagination, let’s ask What if…
  - Someone else has to understand this data?
  - And they aren’t working in this domain?
  - And don’t speak our language fluently (or at all)?
  - And it’s 10 years in the future?
  - And the meaning of some of these terms have changed?
The choice is yours, because the end goal is yours

- The premise of FAIR data: We want this data to be findable, understandable, interoperable, and reusable across all factors that may interfere
- Factors we want to consider include individuals, languages, domains, time frames, evolving scientific understanding…
- Well, as many as is reasonable, anyway
- Use engineering (and financial) judgment to decide how far to go to satisfy the above
And why are semantics important?

- The premise of FAIR data: We want this data to be findable, understandable, interoperable, and reusable across all factors that may interfere.
- Factors include individuals, languages, domains, time frames, evolving scientific understanding…

- For these goals, and these factors that interfere with them, meaning is critical.
- So you want an unambiguous ‘Rosetta Stone’ that helps interpret the data.
Part 2: How do we create (and make available) standardized semantics?
Semantic Introduction
(Good news!)
In semantics, there IS one standard to rule them all!
Primer: Resource Description Framework (RDF)

An Astonishingly Simple Guide
## RDF: Making Statements with Triples

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a class)</td>
<td>(aka Property)</td>
<td>data or class</td>
</tr>
<tr>
<td>identifier (IRI)</td>
<td>identifier (IRI)</td>
<td>“string” or identifier (IRI)</td>
</tr>
<tr>
<td><a href="https://ex.org/dataset">https://ex.org/dataset</a></td>
<td><a href="https://ex.org/createdOn">https://ex.org/createdOn</a></td>
<td>&quot;20190226T145903Z&quot;</td>
</tr>
<tr>
<td>ex:dataset</td>
<td>ex:createdOn</td>
<td>&quot;20190226T145903Z&quot;</td>
</tr>
<tr>
<td>dataset</td>
<td>createdOn</td>
<td>&quot;20190226T145903Z&quot;</td>
</tr>
</tbody>
</table>

The power of RDF comes from (a) declaring logical attributes of properties, like transitivity; and (b) creating shared rules about certain classes and properties. RDFS, SKOS, and OWL add more of these shared classes and properties (and describe rigorously how they work).
Metadata as RDF Graph for a study

tinstances:55417
  rdf:type
  schema:name Immune biomarkers study
  schema:description Immune biomarkers ...

myschema:hasPI
  telements:557
    rdf:type
    schema:name Dr. PI.

myschema:hasInstitution
  telements:37
    rdf:type
    schema:name
    schema:postalCode
    Stanford
    94305
Here’s how simple RDF is
(* = nice on specific topic)

*(I was putting together this list and my browser died…)*

- http://dev.iptc.org/Introduction-To-RDF
- https://www.cambridgesemantics.com/blog/semantic-university/learn-rdf/
- RDF and RDFS (slides)
- Understanding Linked Data Formats *
- Data Modeling with RDF(S) *
- https://www.w3.org/TR/rdf11-concepts/ (technical but precise)
- What is Linked Data (12-minute video, totally basic for your non-tech friends)
2A: Create
How do I create a file with RDF statements?

- A lot of options, from hand-editing the text (I still do this) to free to expensive
- We will briefly show one free option (pair), the one we created 30+ years ago
Protégé Community

- Over 365,000 registered users
- Over 21,000 mailing list subscribers
- Academia, industry, government
- Over 76,000 downloads of Protégé v5.5.0
- Over 141,000 downloads of 5.5.0 beta series in 2019
- 55,000+ projects; 70,000+ users on https://webprotege.stanford.edu
Who uses Protégé?

- World Health Organization (WHO) – ICD-11
- National Cancer Institute (NCI) – NCI Thesaurus
- Gene Ontology Consortium – Gene Ontology
- Object Mgmt. Group & Enterprise Data Mgmt. Council - FIBO
- The Open Biological and Biomedical Ontologies (OBO Foundry)
- Foundational Model of Anatomy (FMA)
Protégé Web Site and Wiki

URL: [http://protege.stanford.edu](http://protege.stanford.edu)
- Download new versions of Protégé
- Subscribe to mailing lists
- Learn about support options
- Links to Protégé wiki, Facebook, and Twitter

URL: [http://protegewiki.stanford.edu](http://protegewiki.stanford.edu)
- WebProtégé user documentation
- Protégé Desktop user documentation
- Developer documentation
- Plug-in and ontology libraries
Mailing Lists

Get free support for all of your Protégé questions via our mailing lists. The lists are actively monitored by Stanford's Protégé team, as well as many experienced Protégé users from the community at large.

Click the Subscribe button next to the list you wish to join. Please note that you must be subscribed to a list before you can post messages.

Protégé User Support
User support for all versions of WebProtégé and Protégé Desktop.

Protégé Developer Support
Developer support for all versions of WebProtégé and Protégé Desktop.

Protégé Announcements
Low traffic, announcement-only list for new releases, availability of short courses, and information regarding Protégé Conferences.
Protégé on GitHub

All software (Protégé, WebProtégé, plugins, etc.) is on GitHub

https://github.com/protegeproject

Protégé: https://github.com/protegeproject/protege

WebProtégé: https://github.com/protegeproject/webprotege

Make feature requests & bug reports in our GitHub issue trackers
The next Protégé Short Course will be held at Stanford University on October 21-23, 2019. Registration is now open! Register early to get reduced rates. See details and register at protege.stanford.edu/shortcourse

We are super excited to announce the release of Protégé 5.5 today! This major release comes with several new features and UI enhancements, and fixes multiple bugs.

Download it from here: http://goo.gl/RHyJeZ
Check the release notes here: https://goo.gl/gqP6LJ
2B: Make Available
Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies
Storing Semantic Content

- NCBO BioPortal (900+ community ontologies)
  - Started in 2004/2005 with National Institutes of Health grant
  - Goal: represent biomedical knowledge (concepts) in one place
  - Emphasized practicality over semantic fidelity
- Significant funding over first 12 years, largely by NIH
  - Developed large number of features, deployable ‘appliance’
  - Started as relational system, migrated to triple store backend…
  - While originally focused on biomedicine, now branching out

Biomedical and more!
Where are we now?

- BioPortal (900+ community ontologies)
  - Extremely large user base (>12K)
  - Extremely large usage (408K hits/mo, 1.4M unique users/yr)
- Large number of deployments => OntoPortal Alliance
  - Community center of support at https://ontoportal.org
  - Opens up new avenues for pursuing support
- Significant ongoing support (>99.5% uptime, new triple store)
OntoPortal Deployment Software

- OntoPortal Virtual Appliance
  - Download as VMWare Virtual Appliance OVF (44 requests, 11 active))
  - Runnable as Amazon Web Service AMI (30)
  - Installable (and even maintainable) by us for you (for a fee)

- Significant upgrades in recent 3.0 release
  - Improved ability to deploy systems and ‘call home’ re upgrades
  - Latest UIs, APIs, and bug fixes
  - Allegrograph triple store (actually quad store) support
  - Self-installable in full preview mode; free license for most users

https://ontoportal.github.io/administration
BioPortal Features

- **Submission/Ingest**
  - Format Conversions
  - Versioning
  - Differencing
  - Metrics re Vocab.
  - Interpretation/Property Mapping
- **Search**
  - Ontology Name
  - Terms
  - Best Terms
  - Attribute/Value
  - Metadata
  - Triples
- **Subselection**
  - Views (partial voc.)
  - Slices (partial repo)
  - Preferred Vocabs
- **Terms**
  - Browsing
  - Create ID
  - Resolve ID
  - Term Pages
  - Mappings
    - Auto-syntax
    - Auto-identifiers
    - Manual-user
- **Vocab. Recommend**
  - Keyword-based
  - Text-based
  - Multi-vocabulary
  - Controllable criteria
  - Selectable vocabs
- **Annotation**
  - Keyword/text based
  - Selectable vocabs

**Widgets**
- Term completion
- Term browsing
- Visualization

**Social**
- Reviews
- Notes/Comments
  - Ontology
  - Term

**Usage Tracking**
- # Requests
- # Downloads
- # Page hits
- # API requests

**API Features**
- Resource Index
- Visualization
The Plan So Far

- Define your concepts
  - Find concepts and definitions in existing terminologies/ontologies, OR
  - Transform your existing dictionaries into standard such as SKOS, OR
  - Create a controlled terminology or ontology based on your needs (using Protégé, WebProtégé, or other tools), OR
  - Mix and match the above

- Put your concepts into BioPortal so they can be seen and used
  - (You can keep them private if you must, but CEDAR users will see them)
  - BioPortal can read any RDF, SKOS, or OWL file (good practices help!)
  - BioPortal will make your ontology and its concepts easy to find
Part 3:
How do we make metadata using standardized semantics?
The CEDAR Vision

A simple ‘life of metadata’ for users

- Compatibility with best known practices and standards
- Use of existing semantic and data structure resources
- Interoperability with existing workflows/systems
- End-to-end improvement in user-supplied metadata
CEDAR Workspace elements

- Metadata templates
- Template elements: (groups of fields and/or elements)
- Template fields
- Instantiated metadata: (metadata instances)
- Folders
Model Specifications

CEDAR Template Model V1.5.0

One of the main goals of the CEDAR project is to build an infrastructure for the creation and storage of machine-readable metadata templates. Metadata templates provide detailed definitions of the metadata that describes a particular data resource. For example, a single study instance that produces a data file (the data resource) may be described with metadata in a comma-separated values (CSV) file. The metadata template describes what is in that CSV file, and applies to other CSV files following the same specification.

This report describes the format of the templates, elements, and instances that make up the CEDAR metadata framework. The specification allows CEDAR metadata to be used by CEDAR, and exchanged with other metadata systems.

We also have written a paper describing the model, which provides a more formal written description but has less technical detail.

Release Date: October 15, 2018
Attachment: CEDAR Template Model v1.5.0.pdf
Author List: Martin O’Connor, Marcos Martínez-Romero, and the CEDAR team
Artifact Type: Report

An Open Repository Model for Acquiring Knowledge about Scientific Experiments

Martin J. O’Connor, Marcos Martínez-Romero, Attila L. Egyedi, Debra Willrett, John Graybeal, and Mark A. Musen

Stanford Center for Biomedical Informatics Research
Stanford, CA 94305, U.S.A.
martin.connor@stanford.edu

Abstract. The availability of high-quality metadata is key to facilitating discovery in the large variety of scientific datasets that are increasingly becoming publicly available. However, despite the recent focus on metadata, the diversity of metadata representation formats and the poor support for semantic markup typically result in metadata that are of poor quality. There is a pressing need for a metadata representation format that provides strong interoperability capabilities together with robust semantic underpinnings. In this paper, we describe such a format, together with open-source Web-based tools that support the acquisition, search, and management of metadata. We outline an initial evaluation using metadata from a variety of biomedical repositories.

https://metadatacentral.org/cedar-template-model
Demonstration
To see more demonstrations visit YouTube

- Playlists: https://www.youtube.com/c/MetadataCenterOrg/playlists
- Short videos of CEDAR features: GO FAIR M4M DeiC Workshop
Using Principled Approaches

- All knowledge artifacts managed as first-class identifiable entities
- All software components have well-defined APIs
- All content represented in JSON and JSON-LD
- Domain-independent metadata specification
- On-the-fly validation of metadata templates and content
Deployable for Any Occasion

LINCS uses it to get metadata for their data repository.
The AIRR community uses it to submit metadata to the National Center for Biotechnology Information (NCBI) repositories.