Semantic Technology for Open Science:  
Creating an Ecosystem for FAIR Data

Mark A. Musen, M.D., Ph.D.  
Stanford University  
musen@Stanford.EDU

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

A Story
Purvesh Khatri, Ph.D. A self-professed “data parasite”
Khatri is using GEO to build a career from the analysis of “other people’s data”

Publicly available data offer
- Alternative experimental conditions
- Real-world heterogeneity
- No cost to perform experiments
- Human samples—with built-in IRB approval

Khatri has a pipeline for
- Searching for data sets
- Elucidating genomic signals
- Confirming those signals in “validation” data sets
- Making discoveries without ever performing a primary experiment
“Sepsis” is a difficult diagnosis

Most infections are not in blood

Clinically, sepsis looks like non-infectious inflammation

II-gene signature distinguishes sepsis patients from trauma patients

Sweeney et al. Sci Trans Med 2015
Khatri has reused public data sets to identify genomic signatures ...

• For incipient sepsis
• For active tuberculosis
• For distinguishing viral from bacterial respiratory infection
• For rejection of organ transplants

... and he has never touched a pipette!

But you can be a data parasite only if scientific data are ...

• Available in a public repository or data commons
• Findable through some sort of search facility
• Retrievable in a standard format
• Self-describing so that third parties can make sense of the data
• The product of careful planning, organization, and stewardship
• Intended to be reused, and to outlive the experiment for which they were collected
But data in public repositories are a mess!

- Investigators view their work as publishing papers, not leaving a legacy of reusable data
- Funding agencies may require data sharing, but they do not explicitly pay for it
- Creating the metadata to describe data sets is unbearably hard
- Ensuring that metadata are standardized and searchable is just about impossible
Failure to use standard terms makes datasets often impossible to search

<table>
<thead>
<tr>
<th>age</th>
<th>age [y]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>age [year]</td>
</tr>
<tr>
<td>AGE</td>
<td>age [years]</td>
</tr>
<tr>
<td>‘Age</td>
<td>age in years</td>
</tr>
<tr>
<td>age (after birth)</td>
<td>age of patient</td>
</tr>
<tr>
<td>age (in years)</td>
<td>Age of patient</td>
</tr>
<tr>
<td>age (y)</td>
<td>age of subjects</td>
</tr>
<tr>
<td>age (year)</td>
<td>age(years)</td>
</tr>
<tr>
<td>age (years)</td>
<td>Age(years)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>Age(years)</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>Age, year</td>
</tr>
<tr>
<td>age (yr)</td>
<td>age, years</td>
</tr>
<tr>
<td>age (yr-old)</td>
<td>age, yrs</td>
</tr>
<tr>
<td>age (yrs)</td>
<td>age.year</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>age_years</td>
</tr>
</tbody>
</table>

Metadata from NCBI *BioSample* Stink!

- 73% of “Boolean” metadata values are not actually *true* or *false*
  - nonsmoker, former-smoker
- 26% of “integer” metadata values cannot be parsed into integers
  - JM52, UVPgt59.4, pig
- 68% of metadata entries that are supposed to represent terms from biomedical ontologies do not actually do so.
  - presumed normal, wild_type
The problem has many causes

- A lack of statistical power in many studies
- An “art form” in conducting experiments
- An eagerness to publish early
- Outright fraud
- A system that does not make it easy or rewarding to replicate the results of other investigators
- A culture that does not yet make the preservation, description, and re-exploration of data an essential element of the enterprise
At a minimum, science needs

- Open, online access to experimental data sets
- Mechanisms to search for metadata to find relevant experimental results
- Annotation of online data sets with adequate metadata
- Use of controlled terms in metadata whenever possible
- A culture that cares about data sharing, secondary analyses, and the ability to make new discoveries from legacy data

Obstacles to Data Sharing

- It takes too much work
- It won’t get me tenure
- I might get scooped
- I might end up looking stupid
- Nobody cares anyway
**Requirement #1**: Standard ontologies to describe what exists in a dataset completely and consistently
Linnaeus created other binomial classifications

- One of **plants**—based on the arrangement of their sexual parts, rather than on broader categories
- One of **diseases**—based on observable signs and symptoms, rather than on underlying causes
- There’s a *reason* that you have never heard of these efforts!
Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class
Enter a class, e.g. Melanoma

Find an ontology
Start typing ontology name, then choose from list

BioPortal Statistics
- Ontologies: 737
- Classes: 9,665,019
- Resources indexed: 48
- Indexed Records: 39,537,360
- Direct Annotations: 95,468,433,792
- Direct RDFS Expanded Annotations: 144,769,502,012

http://bioportal.bioontology.org
Browse
Browse the library of ontologies

Submit New Ontology
Entry Type
- Ontology (996)
- Ontology File (104)

Uploaded in the Last

Category
- All Organisms (1)
- Anatomy (71)
- Animal Development (6)
- Animal Gross Anatomy (2)
- Asplenia (0)
- Biological Process (91)
- Biochemical Resources (4)

Group
- BIBIO (x)
- BIB (x)
- CSAR (x)
- CTSA (x)
- OBO Foundry (x)

Search...
Showing 696 of 1556 Best

Current Procedural Terminology
Updated: 10/17

Medical Dictionary for Regulatory Activities (MEDDRA)
Medical Dictionary for Regulatory Activities Terminology (MedDRA)
Updated: 10/17

RxNorm (RxNorm)
RxNorm Vocabulary
Updated: 10/17

SNOMED CT (SNOMEDCT)
SNOMED Clinical Terms
Updated: 10/17

National Drug Data File (NDDF)
National Drug Data File Plus Source Vocabulary
Updated: 10/17

BioPortal User Activity—Interactive
UI Page Views (thousands per month)
BioPortal User Activity—Programmatic

REST API END USER Requests (millions per month)

Who is using NCBO technology?
Requirement #2: Describe properties of experiments completely and consistently

When we look at experimental data ...

• How can we tell what the data are about?
• How can we search for additional data that we might be interested in?
• How can we group data into similar categories?
We need metadata to describe

- The digital context (properties of the file)
- The investigators and stakeholders
- The scientific context
  - The motivation for the experiment
  - The data that were collected
  - The methods of the experiment
  - The instruments that were used
  - When and where the data were collected
- The parameters of the data

---

The microarray community took the lead in standardizing biological metadata

- What was the substrate of the experiment?
- What array platform was used?
- What were the experimental conditions?

---

DNA Microarray
But it didn’t stop with MIAME!

- Minimal Information About T Cell Assays (MIATA)
- Minimal Information Required in the Annotation of biochemical Models (MIRIAM)
- MINImal MEtagemome Sequence analysis Standard (MINIMESS)
- Minimal Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)
FAIRsharing.org lists more than 100 guidelines for reporting experimental data

Minimal Information Guidelines are not Models

- MIAME and its kin specify only the “kinds of things” that investigators should include in their metadata
- They do not provide a detailed list of standard metadata elements
- They do not provide datatypes for valid metadata entries
- It takes work to convert a prose checklist into a computable model
To make sense of metadata ...

- We need computer-based standards
  - For what we want to say about experimental data
  - For the language that we use to say those things
  - For how we structure the metadata
- We need a scientific culture that values making experimental data accessible to others

**Requirement #3**: Make it palatable to describe experiments completely and consistently
A Metadata Ecosystem

- **Scientists** perform experiments and generate datasets
- **Community-based standards groups** create minimal information models (like MIAME) to define requirements for annotating particular kinds of experimental data in a uniform manner
- **CEDAR users** create formal *metadata templates* based on community standards
- **CEDAR users** fill in metadata templates to create instances of metadata
- **CEDAR** stores the datasets (and the metadata instances) in public repositories
The CEDAR Approach to Metadata

The CEDAR Workbench provides

• Mechanisms
  – To author metadata templates that reflect community standards
  – To fill out templates to encode experimental metadata

• A repository of metadata from which we can
  – Learn metadata patterns
  – Guide predictive entry of new metadata

• Links to BioPortal to ensure that metadata are encoded using appropriate ontology terms
Some key features of CEDAR

• All semantic components—template elements, templates, ontologies, and value sets—are managed as first-class entities
• User interfaces and drop-down menus are not hardcoded, but are generated on the fly from CEDAR’s semantic content
• All software components have well defined APIs, facilitating reuse of software by a variety of clients
• CEDAR generates all metadata in JSON-LD, a widely adopted Web standard that can be translated into other representations

The CEDAR Approach to Metadata
The CEDAR Workbench
The CEDAR Workbench
Brain
Parkinson's disease (DOID) (39%)
Central nervous system lymphoma (DOID) (27%)
Autistic disorder (DOID) (22%)
Melanoma (DOID) (5%)
Schizophrenia (DOID) (1%)
Edwards syndrome (DOID) (2%)
AIRR is providing our first experience uploading CEDAR-authored metadata directly to NCBI.
LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbating agents.
Where is this ecosystem leading?

• Formalization of scientific knowledge in terms of ontologies
• Use of ontologies and templates to standardize scientific communication
• In the long term …
  – Dissemination of the results of scientific investigations purely as machine-interpretable knowledge
  – Intelligent agents that will allow help us to synthesize scientific knowledge from online representations
Imagine a Siri-like intelligent agent on the Web ...

- That tells you when new studies have been done in disciplines that you care about
- That can compare and contrast study results across dimensions that matter to you
- That can identify those studies whose subjects best match the characteristics of particular patients
- That can tell investigators when the methods of their proposed studies are nearly identical to those of earlier studies

Once scientific knowledge is disseminated in machine-interpretable form ...

- Technology such as CEDAR will assist in the automated “publication” of scientific results online
- Intelligent agents will
  - Search and “read” the “literature”
  - Integrate information
  - Track scientific advances
  - Re-explore existing scientific datasets
  - Suggest the next set of experiments to perform
  - And maybe even do them!