Unleashing the value of Common Data Elements through the CEDAR Workbench

Translating Text and Ontologies

S70

Mark A. Musen

Author list: Martin J. O’Connor, Denise B. Warzel, Marcos Martínez-Romero, Josef Hardi, Debra Willrett, Attila L. Egyedi, Aras Eftekhari, John Graybeal, Mark A. Musen

1Stanford University; 2National Cancer Institute; 3Attain LLC
Disclosure

I and my spouse/partner have no relevant relationships with commercial interests to disclose.
What is this colored picture about?
**Sample GSM1253028**

**Status**
Public on Jun 27, 2014

**Title**
Dengue Fever Patient 3

**Sample type**
RNA

**Source name**
Whole Blood of Dengue infected patient at acute infection time point

**Organism**
Homo sapiens

**Characteristics**
- subject id: 3
- infection: DENV
- status: DF
- tissue: whole blood

**Treatment protocol**
Blood was collected on CPT tubes (Vacutainer® with Sodium Citrate; BD). Whole blood specimens were preserved on RNALater buffer (Ambion) and stored in -80°C

**Extracted molecule**
total RNA

**Extraction protocol**
RNA was isolated using the RiboPure-Blood kit (Ambion) and alpha and beta globin mRNA was depleted by GLOBINclear™ Kit (Ambion) according to the manufacturer's protocol. All RNA samples were checked for purity using a ND-1000 spectrophotometer (NanoDrop Technologies) and for integrity by electrophoresis on a 2100 BioAnalyzer (Agilent Technologies).

**Label**
biotin
Metadata authoring is hard

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td># Use this template for 3' or whole Gene expression studies when summarization probe set data will be provided as CHP files.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Do NOT submit CHP files unless they are relevant to your analysis (instead, use the Matrix table option to submit the relevant data, e.g. Bioconductor).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Incomplete submissions will be returned. Click the Metadata Example tab below to view a completed worksheet.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># A complete submission will consist of: (1) a completed metadata worksheet, (2) the CHP files, and (3) the original CEL files.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Field names (in blue on this page) should not be edited. Hover over cells containing field names to view field content guidelines or,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># CLICK HERE for Field Content Guidelines Web page.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SERIES**

1. **title**
2. **summary**
3. **summary**
4. **overall design**
5. **contributor**
6. **contributor**

**SAMPLER**

1. **Sample name**
2. **title**
3. **CHP file**
4. **source name**
5. **organism**
6. **characteristics: tag**

- **Unique title** (less than 120 characters) that describes the overall study.

"Firstname,Initial,Lastname".

Example: "John,H,Smith" or "Jane,Doe".

- **Unique title** that describes the Sample. We suggest that you use the convention:
  [biomaterial]-[condition(s)]-[replicate number], e.g., Muscle_exercised_60min_rep2.

- Replace 'tag' with a biosource characteristic (e.g. "gender", "strain", "tissue", "developmental stage", "tumor stage", etc), and then enter the value for each sample beneath (e.g. "female", "129SV", "brain", "embryo", etc). You may add additional characteristics columns to this template (see 'Metadata Example' spreadsheet).
Poor metadata

An analysis of metadata from NCBI’s BioSample

- 73% of “Boolean” values
  - nonsmoker, former-smoker
- 26% of “integer” values
  - JM52, UVPgt59.4, pig
- 68% of ontology terms
  - presumed normal, wild_type

Gonçalves, R. S., & Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. Scientific Data, 6, 190021. https://doi.org/10.1038/sdata.2019.21
FAIR Principles to enhance the value of digital resources and their metadata

https://www.nature.com/articles/sdata201618

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier […] Barend Mons

Scientific Data 3, Article number: 160018 (2016) | Download Citation

An Addendum to this article was published on 19 March 2019

Findable, Accessible, Interoperable, Reusable
The CEDAR approach to metadata

Authoring of Metadata Templates
- Template authors (e.g., standards committees)
  - define
  - Metadata templates

Annotation of Data with Metadata
- Metadata acquisition forms
  - contribute
  - fill in

Exploration and Reuse of Datasets through Metadata
- Metadata repository
  - search, reuse

Scientists

CEDAR
The CEDAR approach to standards

• Operating on Big Data requires all kinds of standards
• We don’t want to be in the standards business ourselves
• We want to be able to accommodate the standards that come from the biomedical community
• We need an adaptable infrastructure where standard specifications are themselves editable
Minimum Information About a Microarray Experiment

General Information
MIAME is intended to specify all the information necessary for an unambiguous interpretation of a microarray experiment, and potentially to reproduce it. MIAME defines the content but not the format for this information.

Homepage: http://www.fged.org/projects/miame/
Countries that developed this resource: Belgium, France, Germany, Netherlands, United Kingdom, United States
Created in 1999
Taxonomic range

Knowledge Domains
- DNA Microarray
- Messenger Ribonucleic Acid
- Gene Expression Data
- Genome
- Nucleic Acid Hybridization
- Ribonucleic Acid

Subjects
- Life Science
- Transcriptomics

In the following recommendations:

How to cite this record: FAIRsharing.org: MIAME; Minimum Information About a Microarray Experiment; DOI: https://doi.org/10.25504/FAIRsharing.32b10v; Last edited: Feb. 14, 2019, 5:16 p.m.; Last accessed: Oct 23 2019 10:05 p.m.
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.
CEDAR takes advantage of ontology standards

- Standard templates derived from community-based minimal information models
- Templates rendered as frames that can be instantiated with standard values
- Template slots filled using standard ontologies and value sets
But there are different kinds of specifications!

- Templates describing classes of experiments
- Ontologies describing potential values
- Metadata (CDEs) describing reusable question specifications for collecting and reporting data
NCI uses “common data elements” as metadata for fields in CRFs
ISO/IEC 11179

- International standard for representing metadata
- Provides:
  - Guidelines for the naming and definition of data elements
  - Information about the metadata that must be captured for data elements
  - Rules for the way data elements are created and registered
NCI’s Cancer Data Standards Repository (caDSR)

- One of the largest CDE registries (over 60,000 CDEs)
- Based on ISO/IEC 11179
Bringing CDEs to the masses

• Practical challenges to reusing CDEs limit their adoption by the biomedical community
  • ISO/IEC 11179 does not specify implementation-level details
  • Complex ISO conformance requirements

• We extended the CEDAR Workbench to support CDEs
• We incorporated over 49K caDSR CDEs into CEDAR
• These CDEs are publicly available on CEDAR
We developed a CDE ingestion workflow
What CDEs Bring CEDAR

Template slots for experimental metadata …

- Can still refer to **ontologies and value** sets as the source of values for data items
- Are able to refer to **CDEs** as the source of values or datatype restrictions for data items

CEDAR is able to interoperate with both kinds of standards
Representing CDEs in CEDAR will allow authoring of CEDAR templates that can provide the basis for eCRFs
Summary

• We extended CEDAR to natively support CDEs
• We ingested a library of over 49K CDEs and associated value sets
• These CDEs are publicly available for general use on CEDAR
  • Users can create templates using CDE-based fields
  • Users can create FAIR metadata → Enhanced data Findability
Thank you!

Email me at: musen@Stanford.EDU