Of Course Data Stewards Benefit from Training. But What They Really Need is Better *Technology*

Mark A. Musen, M.D., Ph.D
and the CEDAR Team
musen@Stanford.EDU
A Story
Purvesh Khatri, Ph.D. A self-professed “data parasite”
Gene Expression Omnibus (GEO)
Khatri has reused public datasets in GEO 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 the online datasets that Khatri studies 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.
# Use this template for 3' or whole Gene expression studies when summarization probe set data will be provided as CHP files.
# 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. Bioconductors).
# Incomplete submissions will be returned. Click the Metadata Example tab below to view a completed worksheet.
# A complete submission will consist of: (1) a completed metadata worksheet, (2) the CHP files, and (3) the original CEL files.
# Field names (in blue on this page) should not be edited. Hover over cells containing field names to view field content guidelines or, # CLICK HERE for Field Content Guidelines Web page.

**SERIES**

# This section describes the overall study.

<table>
<thead>
<tr>
<th>title</th>
<th>summary</th>
<th>overall design</th>
<th>contributor</th>
</tr>
</thead>
</table>

**SAMPLES**

# The Sample names in the first column are arbitrary but they must match the column headers of the Matrix table (see next worksheet).

<table>
<thead>
<tr>
<th>Sample name</th>
<th>title</th>
<th>CHP file</th>
<th>source name</th>
<th>organism</th>
<th>characteristics: tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMPLE 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PROTOCOLS**

# This section includes protocols and fields which are common to all Samples.

# Protocols which are applicable to specific Samples or specific channels should be included in additional columns of the SAMPLES section instead.

<table>
<thead>
<tr>
<th>growth protocol</th>
<th>treatment protocol</th>
<th>extract protocol</th>
<th>label protocol</th>
<th>hyb protocol</th>
</tr>
</thead>
</table>

**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).**

**[Optional] Describe the conditions that were used to grow or maintain organisms or cells prior to extract preparation.**
Failure to use standard terms makes datasets often impossible to search

age
Age
AGE
`Age
age (after birth)
age (in years)
age (y)
age (year)
age (years)
Age (years)
Age (Years)
age (yr)
age (yr-old)
age (yrs)
Age (yrs)

age [y]
age [year]
age [years]
age in years
age of patient
Age of patient
age of subjects
age(years)
Age(years)
Age(ys.)
Age, year
age, years
age, yrs
age.year
age_years
An Analysis of Metadata from *BioSample*

- 85% of submissions avoid using a predefined “package” for regularizing metadata
- 73% of “Boolean” metadata values are not actually *true* or *false*
- 26% of “integer” metadata values cannot be parsed into integers
- 68% of metadata entries that are supposed to represent terms from biomedical ontologies do not actually do so.
At a minimum, scientists need

- Open, online access to experimental data sets
- Annotation of online data sets with adequate metadata
- Use of controlled terms in metadata whenever possible
- Technology that can help them curate their data—not training to instill specific skills
Open data is about more than disclosure, it must be "Fair".

- Findable
- Accessible
- Interoperable
- Reusable
Requirement #1: Have standard terms to describe what exists in a dataset completely and consistently
Welcome to BioPortal, the world’s most comprehensive repository of biomedical ontologies

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

Find an ontology
Start entering ontology name, e.g. Cancer, then choose from list
Browse Ontologies

Ontology Visits (February 2018)

BioPortal Statistics
Ontologies 692
Classes 8,848,090
Resources Indexed 48
Indexed Records 39,537,360
Direct Annotations 95,468,433,792
Direct Plus Expanded Annotations 144,789,582,932

http://bioportal.bioontology.org
<table>
<thead>
<tr>
<th>Entry Type</th>
<th>Category</th>
<th>Group</th>
<th>Ontology ID</th>
<th>Uploaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Dictionary for Regulatory Activities (MEDDRA)</td>
<td>Medical Dictionary for Regulatory Activities Terminology (MedDRA)</td>
<td>BIBLO</td>
<td>10</td>
<td>2/6/17</td>
</tr>
<tr>
<td>RxNORM (RXNORM)</td>
<td>RxNorm Vocabulary</td>
<td>Biomedical Resources</td>
<td>7</td>
<td>2/6/17</td>
</tr>
<tr>
<td>SNOMED CT (SNOMEDCT)</td>
<td>SNOMED Clinical Terms</td>
<td>Biomedical Resources</td>
<td>22</td>
<td>2/6/17</td>
</tr>
<tr>
<td>National Drug Data File (NDDF)</td>
<td>National Drug Data File Plus Source Vocabulary</td>
<td>Biomedical Resources</td>
<td>1</td>
<td>2/6/17</td>
</tr>
</tbody>
</table>
Who is using BioPortal technology?
Requirement #2: Describe properties of experiments completely and consistently
Minimum Information About a Microarray Experiment - MIAME

MIAME describes the Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. [Brazma et al., Nature Genetics]

The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)

For more details, see MIAME 2.0.
But it didn’t stop with MIAME!

• Minimal Information About T Cell Assays (MIATA)
• Minimal Information Required in the Annotation of biochemical Models (MIRIAM)
• MINIimal MEtagenome Sequence analysis Standard (MINIMESS)
• Minimal Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)
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
Requirement #3: Make it easy to describe experiments completely and consistently.
The CEDAR Approach to Metadata

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

- **Annotation of Data with Metadata**
  - Contribute
  - Fill in
  - Metadata acquisition forms

- **Exploration and Reuse of Datasets through Metadata**
  - Search, reuse
  - Metadata repository

[Diagram showing the process flow with icons and labels]
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
<table>
<thead>
<tr>
<th>Title</th>
<th>Created</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEO</td>
<td>9/5/17 9:48 AM</td>
<td>9/5/17 10:24 AM</td>
</tr>
<tr>
<td>BioCADDIE</td>
<td>9/5/17 9:48 AM</td>
<td>9/5/17 10:24 AM</td>
</tr>
<tr>
<td>Optional Attribute</td>
<td>9/5/17 10:38 AM</td>
<td>9/5/17 10:38 AM</td>
</tr>
<tr>
<td>ImmPort Investigation</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 10:21 AM</td>
</tr>
<tr>
<td>LINCS Cell Line</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 9:49 AM</td>
</tr>
<tr>
<td>LINCS Antibody</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 9:49 AM</td>
</tr>
<tr>
<td>ImmPort Study</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 9:49 AM</td>
</tr>
<tr>
<td>Title</td>
<td>Created</td>
<td>Modified</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>GEO</td>
<td>9/5/17 9:48 AM</td>
<td>9/5/17 10:24 AM</td>
</tr>
<tr>
<td>BioCADDIE</td>
<td>9/5/17 9:48 AM</td>
<td>9/5/17 10:24 AM</td>
</tr>
<tr>
<td>Optional Attribute</td>
<td>9/5/17 10:38 AM</td>
<td>9/5/17 10:38 AM</td>
</tr>
<tr>
<td>ImmPort Investigation</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 10:21 AM</td>
</tr>
<tr>
<td>LINCS Cell Line</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 9:49 AM</td>
</tr>
<tr>
<td>LINCS Antibody</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 9:49 AM</td>
</tr>
<tr>
<td>ImmPort Study</td>
<td>9/5/17 9:49 AM</td>
<td>9/5/17 9:49 AM</td>
</tr>
</tbody>
</table>
**BioSample Human**

- **Sample Name**: 056
- **Organism**: Homo sapiens
- **Tissue**: skin of body
- **Sex**: Male
- **Isolate**: N/A
- **Age**: 74
- **Biomaterial Provider**: Life Technologies

**Attribute (1)**
- **Name**: disease
- **Value**: dermatitis

**Attribute (2)**
- **Name**: description
- **Value**: Cell line was cultured until the 5th passage

**Attribute (3)**
- **Name**: treatment
- **Value**: 350mg brodalumab
The CEDAR Workbench
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Source</th>
<th>Identifier</th>
<th>No. Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Find terms in BioPortal or **Create New Terms** to constrain the values of the 'Tissue' field

<table>
<thead>
<tr>
<th>TERM</th>
<th>DEFINITION</th>
<th>TYPE</th>
<th>SOURCE</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>tissue</td>
<td>Multicellular anatomical structure that consists of many cells of one or a few types, arranged in an extracellular...</td>
<td>Class</td>
<td>UBERON</td>
<td>UBERON_0000479</td>
</tr>
<tr>
<td>tissue</td>
<td>-</td>
<td>Class</td>
<td>MA</td>
<td>MA_0003002</td>
</tr>
<tr>
<td>Tissue</td>
<td>-</td>
<td>Class</td>
<td>NIFSTD</td>
<td>birnlex_19</td>
</tr>
<tr>
<td>tissue</td>
<td>Anatomical structure, that consists of similar cells and intercellular matrix, aggregated according to genetically...</td>
<td>Class</td>
<td>TAO</td>
<td>CARO_0000043</td>
</tr>
</tbody>
</table>
Ontology: UBERON

Term Details:
- Name: tissue
- Id: http://purl.obolibrary.org/obo/UBERON_0000479
- Definition: Multicellular anatomical structure that consists of many cells of one or a few types, arranged in an extracellular matrix such that their long-range organisation is at least partly a repetition of their short-range organisation.

Click to add all the descendants of the selected term

Add
The CEDAR Workbench
BioSample Human

- **Sample Name**: 056
- **Organism**: Homo sapiens
- **Tissue**: lung
- **Sex**: Male
- **Age**: 74
- **Biomaterial Provider**: Life Technologies

### Disease

- **Name**: lung cancer (DOID)
- **Value**: 61%
- **Name**: chronic obstructive pulmonary disease (DOID)
- **Value**: 31%
- **Name**: lung squamous cell carcinoma (DOID)
- **Value**: 5%
- **Name**: idiopathic pulmonary fibrosis (DOID)
- **Value**: 4%
- **Name**: lung adenocarcinoma (DOID)
- **Value**: 4%
- **Name**: adenocarcinoma (DOID)
- **Value**: 3%
- **Name**: carcinoma (DOID)
- **Value**: 2%
Parkinson's disease (DOID) (39%)
central nervous system lymphoma (DOID) (27%)
autistic disorder (DOID) (22%)
melanoma (DOID) (5%)
Edwards syndrome (DOID) (2%)
schizophrenia (DOID) (1%)
The CEDAR Workbench

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

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

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

Graphical elements and logos (e.g., CEDAR, HMP, ImmPort, The Cancer Genome Atlas)
Biomedical Researchers Clearly are Ahead of the Pack

• They have been creating standard ontologies for years
• They are proposing increasing numbers of “minimal information models” that are ripe for conversion to formal metadata templates
• They are beginning to turn to technology such as CEDAR to enhance their online datasets
The CEDAR Approach is Generalizable to Other Areas of Science

• The building blocks needed for developing high-quality metadata are clear:
  – Standard ontologies
  – Stanford templates

• Nothing in CEDAR is hardwired to the life-sciences domain

• Most important: Operators are standing by ....