The Center for Expanded Data Annotation and Retrieval

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

- Stanford University School of Medicine
- University of Oxford e-Science Centre
- Yale University School of Medicine
- Northrup Grumman Corporation
- Stanford University Libraries
At a minimum, science needs

• Open, online access to experimental data sets
• Annotation of online data sets with adequate metadata (data about the data)
• Use of controlled terms in metadata annotations
• Mechanisms to search for metadata to find relevant experimental results
Open data is about more than disclosure; it must be "fair".

- Findable
- Accessible
- Interoperable
- Reusable
Problem #1: Describe what exists in the world completely and consistently
Welcome to BioPortal, the world’s most comprehensive repository of biomedical ontologies

Ontology Visits (January 2017)

BioPortal Statistics

http://bioportal.bioontology.org
Problem #2: Describe properties of experiments completely and consistently
We need metadata to help us know ...

• What are the data about?
  – Where the data come from
  – How the data were produced
  – How the data are formatted
  – Where we can learn more about the data

• Where might we find additional datasets that we might be interested in?
To make sense of the metadata ...

• We need standards
  – For the kinds of things we want to say about the 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
The microarray community took the lead in standardizing metadata

- What was the substrate of the experiment?
- What array platform was used?
- What were the experimental conditions?
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.
Minimum Information About a Microarray Experiment

Abbreviation: MIAME

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.

Developed in: United Kingdom, France, Germany, Netherlands, Belgium, United States of America
Created in: 1999
Taxonomic range

Scope and data types
- Microarray Data
- Genome
- DNA
- DNA Microarray
- Transcriptome
- RNA
- Nucleic Acid Hybridization

Record updated: March 11, 2016, 5:33 p.m. by The BioSharing Team.

Implementing Databases (4)
- ArrayExpress
  ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MiNSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.
- Gene Expression Omnibus
  The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO.

The Immunology Database and Analysis Portal

Recommended by
- EMBO Press
- Scientific Data

In Collections
- National Child Development Study (UK)
- DNA Microarray

Related Standards

Implementing Policies
- Scientific Data’s Recommended Data Repositories
- EMBO Press Recommended Databases and Data Standards
Problem #3: Make it palatable 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
- Metadata acquisition forms
  - contribute
  - fill in
  - Scientists

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

Logos and projects:
- The Cancer Genome Atlas
- HMP: Human Microbiome Project
- IMMPORT
- GECCO: Gene Expression Computational Core
CEDAR technology is providing

- **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>BIBFRAME</td>
<td>11/5/16 1:37 PM</td>
<td>11/5/16 1:37 PM</td>
</tr>
<tr>
<td>CDEs</td>
<td>8/21/16 10:22 AM</td>
<td>8/21/16 10:22 AM</td>
</tr>
<tr>
<td>Sample templates</td>
<td>11/4/16 8:19 PM</td>
<td>11/4/16 8:19 PM</td>
</tr>
<tr>
<td>BioSample Human</td>
<td>11/17/16 3:56 PM</td>
<td>11/17/16 3:56 PM</td>
</tr>
<tr>
<td>ImmPort Study</td>
<td>5/2/16 12:38 PM</td>
<td>5/2/16 12:38 PM</td>
</tr>
<tr>
<td>Investigation</td>
<td>5/7/16 2:03 PM</td>
<td>5/7/16 2:03 PM</td>
</tr>
<tr>
<td>LINCS Antibody</td>
<td>10/12/16 3:42 PM</td>
<td>10/12/16 3:42 PM</td>
</tr>
<tr>
<td>LINCS Cell Line</td>
<td>10/12/16 3:57 PM</td>
<td>10/12/16 3:57 PM</td>
</tr>
<tr>
<td>Title</td>
<td>Created</td>
<td>Modified</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>BIBFRAME</td>
<td>11/5/16 1:37 PM</td>
<td>11/5/16 1:37 PM</td>
</tr>
<tr>
<td>CDEs</td>
<td>8/21/16 10:22 AM</td>
<td>8/21/16 10:22 AM</td>
</tr>
<tr>
<td>Sample templates</td>
<td>11/4/16 8:19 PM</td>
<td>11/4/16 8:19 PM</td>
</tr>
<tr>
<td>BioSample Human</td>
<td>11/17/16 3:56 PM</td>
<td>11/17/16 3:56 PM</td>
</tr>
<tr>
<td>ImmPort Study</td>
<td>5/2/16 12:38 PM</td>
<td>5/2/16 12:38 PM</td>
</tr>
<tr>
<td>Investigation</td>
<td>5/7/16 2:03 PM</td>
<td>5/7/16 2:03 PM</td>
</tr>
<tr>
<td>LINCS Antibody</td>
<td>10/12/16 3:42 PM</td>
<td>10/12/16 3:42 PM</td>
</tr>
<tr>
<td>LINCS Cell Line</td>
<td>10/12/16 3:57 PM</td>
<td>10/12/16 3:57 PM</td>
</tr>
</tbody>
</table>
BioSample Human

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

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

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

Optional Attribute (3)
- Name: treatment
- Value: 350mg brodalumab
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

Scientists

The Cancer Genome Atlas
**BioSample Human**

**Template Name**

BioSample Human

**Template Description**

BioSample Human

- **Sample Name**

- **Organism**

- **Tissue**

  - Type of tissue the sample was taken from
<table>
<thead>
<tr>
<th>FIELD TYPE</th>
<th>VALUES</th>
<th>MULTIPLE</th>
<th>REQUIRED</th>
<th>SUGGESTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Type</td>
<td>Source</td>
<td>Identifier</td>
<td>No. Values</td>
</tr>
</tbody>
</table>

**Sample Name**

**Organism**

Type of tissue the sample was taken from

**Enter Field Title**

Tissue
Find terms in BioPortal or Create New Terms to constrain the values of the 'Tissue' field

500 results for the query 'Tissue'. Click on a term below to select it

<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:

<table>
<thead>
<tr>
<th>Name</th>
<th>tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td><a href="http://purl.obolibrary.org/obo/UBERON_0000479">http://purl.obolibrary.org/obo/UBERON_0000479</a></td>
</tr>
<tr>
<td>Definition</td>
<td>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.</td>
</tr>
</tbody>
</table>

Click to add all the descendants of the selected term

Add
The CEDAR Approach to Metadata
BioSample Human

- Sample Name: 056
- Organism: Homo sapiens
- Tissue: brain
- Sex: male
- Isolate: N/A
- Age: 74 years
- Biomaterial Provider: Life Technologies

Disease:
- 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 Approach to Metadata
ImmPort’s Mission

ImmPort is a long-term, sustainable data warehouse for the purpose of promoting re-use of immunological data generated by NIAID DAIT and DMID funded investigators. ImmPort supports analysis of flow cytometry results and HLA genetic associations.

What is ImmPort

Flow Cytometry Analysis (FLOCK)

Flow cytometry analysis component includes:
- Automated cell population identification
- Result visualization in 2D and 3D
- Statistical analysis of population characteristics
- Automated mapping of populations across multiple samples

Open ImmPort

- Browse and search for shared study data
- Cytokine and cell interaction literature mining: ImmuneXpresso
- Example R and Python analysis code
- Cytokine registry
- Cell Ontology Visualizer

Data Release

March 2016 - Data Deluge - 48 Studies released! Human Immunology Project Consortium (HIPC) teams from Stanford, Baylor, Dana Farber, Yale, Emory and Seattle have contributed extensively. ImmPort is also sharing March of Dimes microbiome study data SDY465 and data from 5 RELIVE clinical trials (SDY289 to SDY292, SDY294). Combined with human T cell distribution data from the Donna Farber team of Columbia Center for Translational Immunology SDY702 and you have a MAJOR data deluge! What data can you expect to find? Multi-year flu vaccine studies, HAI, ELISpot, CyToF, flow cytometry, clinical assessments and much more. Get the scoop on all the new data in ImmPort Data Release 18 notes.
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 perturbing agents.
About the AIRR Community

The Adaptive Immune Receptor Repertoire (AIRR) Community is a community-driven organization that is organizing and coordinating stakeholders in the use of NGS technologies to study antibody (Ab)/B-cell and T-cell receptor (TcR) repertoires. Recent advances in sequencing technology have made it possible to sample the immune repertoire in exquisite detail. AIRR sequencing has enormous promise for understanding the dynamics of the immune repertoire in vaccinology, infectious disease, autoimmunity, and cancer biology, but also poses substantial challenges. To meet these challenges, we have established the AIRR Community.
Open data is about more than disclosure, it must be "Fair".

• Findable
• Accessible
• Interoperable
• Reusable
Authoring more complete, more standardized metadata ...

- Will aid dataset
  - Discovery
  - Exploration
  - Integration
  - Secondary use
- Will aid communication of scientific results
- Will aid knowledge management within research organizations
- Will make data more FAIR