Making it Easier, Possibly Even Pleasant, to Author Rich Experimental Metadata
High Quality Metadata are Essential for Large-Scale Reuse and Biomedical Discovery
MIAAVE (Minimum Information About a Microarray Experiment) is a standard for describing microarray experiments. MIAAVE is based on the use of MAGE-TAB and MAGE-ML, which are designed to provide a framework for describing microarray experiments in a structured and machine-readable format.

MIAAVE does not specify any particular tool or method for conducting microarray experiments. However, it recommends the use of MAGE-TAB and MAGE-ML to ensure that the data is consistent and can be easily shared and analyzed.

The six most critical elements contributing towards MIAAVE are:

1. The raw data for each hybridization (e.g., CEL or MIAME)
2. The final processed (navigated) data for the set of hybridizations 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 set subject is part of which sample, with which hybridizations are technical, which are biological replicates)
5. The essential annotation of the array (e.g., Gene identifiers, genomic coordinates, probe oligonucleotide sequences for reference commercial array calibration)
6. The essential laboratory and data processing protocols (e.g., what method has been used to obtain the final processed data)

For more details, see MIAAVE 2.0.
<table>
<thead>
<tr>
<th>Minimal Information About a Cardiac Electrophysiology Experiment</th>
<th>MICEE</th>
<th>MTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Information About a Cell Assays</td>
<td>MIASPE</td>
<td>MIASE</td>
</tr>
<tr>
<td>Minimum Information About a Sample Preparation for a Phosphoproteomics Experiment</td>
<td>MIARE</td>
<td>MIRAPE</td>
</tr>
<tr>
<td>Minimum Information About a Simulation Experiment</td>
<td>MIARE</td>
<td>MIRAPE</td>
</tr>
<tr>
<td>Minimum Information About a RNA Experiment</td>
<td>MIARE</td>
<td>MIRAPE</td>
</tr>
<tr>
<td>Minimum Information About a Peptide Array Experiment</td>
<td>MIARE</td>
<td>MIRAPE</td>
</tr>
<tr>
<td>Minimum Information About a Protein Affinity Regent</td>
<td>MIAPAR</td>
<td>MIAPA</td>
</tr>
<tr>
<td>Minimum Information About a Physiological Analysis</td>
<td>MIAPAR</td>
<td>MIAPA</td>
</tr>
<tr>
<td>Minimum Information About a Microarray Experiment</td>
<td>MIAPAR</td>
<td>MIAPA</td>
</tr>
<tr>
<td>Minimal Information About a Cellular Assay</td>
<td>MIACA</td>
<td>MIABE</td>
</tr>
<tr>
<td>Minimum Information About a Biology Experiment</td>
<td>MIABE</td>
<td>MIABE</td>
</tr>
<tr>
<td>Minimum Information About a Bioactive Entity</td>
<td>MIABE</td>
<td>MIABE</td>
</tr>
<tr>
<td>Guidelines for Information About Therapy Experiments</td>
<td>MIABE</td>
<td>MIABE</td>
</tr>
<tr>
<td>Core Information for Metabolomics Reporting</td>
<td>MIR</td>
<td>MIR</td>
</tr>
</tbody>
</table>

### Bioscience Projects Registered with MiBi

- XML document containing all registered projects (from this schema, same information as the Excel spreadsheet)
- Summary spreadsheet of all registered projects
- Registration form for the MiBi Portal (please return to ChrisPayton[@]gmail.com)
http://www.w3.org/TR/hcls-dataset/
To be Findable:

1. (meta)data are assigned a globally unique and eternally persistent identifier
2. (meta)data are described with rich metadata
3. (meta)data are registered or indexed in a searchable resource
4. metadata specify the data identifier

To be Accessible:

1. (meta)data are retrievable by their identifier using a standardized communications protocol
1.1 the protocol is open, free, and universally implementable
2. metadata are eternally accessible, even when the data are no longer available
3. metadata are registered or indexed in a searchable resource

To be Interoperable:

1. (meta)data use vocabularies that follow FAIR principles
2. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
3. (meta)data include qualified references to other (meta)data

To be Reusable:

1. (meta)data include qualified references to other (meta)data

Data Sharing Plans → FAIR – Findable, Accessible, Interoperable, Reusable
The Good News:
Minimal information checklists, such as MIAME, are being advanced from all sectors of the biomedical community.

The Bad News:
Investigators view requests for even "minimal" information as burdensome.
Unguided Data Entry is Problematic
Research Questions

• What are the most effective approaches to maximize reuse of pre-defined metadata elements?

• How can we get metadata authors to generate more metadata in shorter time periods?

• To what extent can we successfully predict metadata values from existing metadata?
  – Can we use manuscripts, published work, user history, social media to improve prediction?
  – To what extent can we successfully predict metadata values from existing metadata?

• To what extent can experts and crowds find, verify, and fix metadata?

• To what extent does ontology-based metadata improve the precision and recall of dataset search?

• How can we get metadata authors to generate more reuse of pre-defined metadata elements?
CEDAR technology will give us

- To explore, verify, and collaboratively augment experimental metadata even when the data are located elsewhere
- To index, query, and reason about ontology-based metadata
- To efficiently guide predictive entry of new metadata and structured metadata
- To efficiently search for unstructured, semi-structured, and unstructured metadata
- To discover datasets that meet a plurality of dataset and experimental requirements
- To facilitate the capture of experimental metadata that conform to one or more community standards
- To facilitate the capture of experimental metadata that standard while reusing previously defined elements
- To author metadata templates that define a community
- To learn metadata patterns from unstructured, semi-structured, and structured metadata

Methods

• Experimental requirements

Tools
CEDAR will empower users to meet and exceed the minimal metadata standards—

- We need a more expressive—and computable—framework for describing metadata
- Emphasis traditionally has been on development of simple checklists of metadata elements
- Little practical consideration for richness and quality
- Metadata validation against community standards
- Ecosystems of tools and data
- Common knowledge representations (interoperable)
- Using shared value sets (search, browse, query)

- Emphasis traditionally has been on development of simple checklists of metadata elements
- Using shared value sets (search, browse, query)
- Metadata validation against community standards

• We need a more expressive—and computable—framework for describing metadata
• Emphasis traditionally has been on development of simple checklists of metadata elements
• Little practical consideration for richness and quality
• Metadata validation against community standards
• Ecosystems of tools and data
• Common knowledge representations (interoperable)
• Using shared value sets (search, browse, query)
CEDAR Team

Mark A. Musen (PI) • Michel Dumontier (Co-I) • Purvesh Khatri (Co-I) • Olivier Gevairt (Co-I) • Rafael Shanekar • Attila Egyedi • Mary Panahiazar • Marcos Martinez Romero • Martin O’Connor • John Greybear • Jeff Wisser • Northrop Grumman (HIPC/Import) • Kei Chang • Steven H. Kleinstein • Yale University • Alejandra Gonzalez-Beltran • Philippe Rocca-Serra • Susanna Sansone • Oxford University • Kim Durant • Stanford Library • Jeff Wisser • Northrop Grumman (HIPC/Import) • Kei Chang • Steven H. Kleinstein • Yale University • Alejandra Gonzalez-Beltran • Philippe Rocca-Serra • Susanna Sansone • Oxford University • Kim Durant • Stanford Library
We also want to help users compose templates and auto-magically suggest metadata values.
Can we use free text to predict structured and semi-structured values?

<table>
<thead>
<tr>
<th>Sample GSM1230698</th>
<th>Query DataSets for GSM1230698</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Status</strong></td>
<td>Public on Oct 02, 2014</td>
</tr>
<tr>
<td><strong>Title</strong></td>
<td>SNG-M_PTX_1</td>
</tr>
<tr>
<td><strong>Sample type</strong></td>
<td>RNA</td>
</tr>
<tr>
<td><strong>Source name</strong></td>
<td>SNG-M Paclitaxel 24h</td>
</tr>
<tr>
<td><strong>Organism</strong></td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>
| **Characteristics** | cell line: SNG-M  
          | cell-type: endometrial cancer |
| **Treatment protocol** | Treated with eribulin and paclitaxel at 10xIC50 conc. for 24 hours. IC50 determined in growth inhibition assay for cell line separately. |
| **Growth protocol** | Cell lines were growing in growth media recommended by ATCC. |
| **Extracted molecule** | total RNA                       |
| **Extraction protocol** | Total RNA was extracted using RNeasy Mini kit (Qiagen). |
| **Label**         | biotin                         |
| **Label protocol** | Biotinylated fragmented cRNA was used. |
| **Hybridization protocol** | We used manufacture recommended protocol (Affymetrix). Arrays were washed and stained using Affymetrix Fluidics Station 450 |
| **Scan protocol** | Arrays were scanned using Affymetrix GeneChip Scanner 3000 |
| **Description**   | drug                            |
| **Data processing** | Gene chips were analyzed using Affymetrix Microarray Analysis Suite (MAS) version 5. RMA normalization was performed using Affymetrix Power Tools version 1.12.0 |
Using multi-label tree

Predicting semi-structured metadata

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Label</th>
<th>Molecule</th>
<th>Organism</th>
<th>Type</th>
<th>GPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA =&gt; SVM</td>
<td>88.90%</td>
<td>87.64%</td>
<td>87.43%</td>
<td>94.87%</td>
<td>32.69%</td>
</tr>
<tr>
<td>LDA =&gt; Decision Tree</td>
<td>93.30%</td>
<td>86.80%</td>
<td>95.45%</td>
<td>95.01%</td>
<td>73.00%</td>
</tr>
</tbody>
</table>

Accuracy = % correctly classified samples

Average recall 82% (for all keys)
Average precision 79% (for all keys)
Accuracy of 72% to predict 39 most-occurring keys

Predicting structured metadata (accuracy)
We welcome new partners to:

- Learn about your metadata authoring workflow
- Evaluate CEDAR technology
- Incorporate this technology into your curation workflow

Metadata: The Next Frontier