A sea of standards for omics data: sink or swim?

Jessica D Tenenbaum,¹ Susanna-Assunta Sansone,² Melissa Haendel³

Tenenbaum JD, et al. J Am Med Inform Assoc 2013;0:1–4. doi:10.1136/amiajnl-2013-002066

Open Access Journal Club

10/3/13
Trend toward data sharing

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

A Notice by the Science and Technology Policy Office on 11/04/2011

Open Access for Research Publications

December 10, 2009

Byline: a GenomeWeb staff reporter
Newsletter: GenomeWeb Daily News

Upcoming Changes to Public Access Policy Reporting Requirements and Related NIH Efforts to Enhance Compliance

Notice Number: NOT-OD-12-160

Key Dates
Release Date: November 16, 2012
Report: Recommendations to NIH Director

• Data and Informatics Working Group Report to Advisory Committee to the Director of NIH (6/12)
  – expert advice on management, integration, and analysis of large biomedical research datasets
  – Goals included “advance basic and translational science by facilitating and enhancing the sharing of research-generated data”
  – Recommendation 1a. Establish a Minimal Metadata Framework for Data Sharing
IOM: Toward Precision Medicine

• Revised disease taxonomy based on molecular mechanisms
• Build information commons- data on large populations of patients become broadly available for research use
• Build knowledge network- add value by converting data to knowledge in context of biology and clinical care
2009: Let’s build a standards-compliant omics data repository

- What does it mean to build an omics data repository that is standards compliant?
- What standards exist?
- What is a data standard?
- What’s the best one for our purposes?
But how much do we know about these standards?

- Which ones are mature enough for me to use or recommend?
- What are the criteria to evaluate status and value?
- How can I get involved to propose extensions or modifications?
- Which tools and databases implement which standards?
- I work on plants, are these just for biomedical applications?

Courtesy of SA Sansone
The Punchline[s]

1. Many different definitions for what constitutes a ‘data standard’.
2. No one standard is the ‘right’ standard—depends on particular needs.
3. Resource are needed to help researchers navigate the standards landscape.
Exercise: Identify standards in genomics

<table>
<thead>
<tr>
<th>Standard</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIAME</td>
<td>Reporting guideline</td>
</tr>
<tr>
<td>ISA-TAB</td>
<td>Exchange format</td>
</tr>
<tr>
<td>MAGE-TAB</td>
<td>Exchange format</td>
</tr>
<tr>
<td>MAGE-ML</td>
<td>Exchange format</td>
</tr>
<tr>
<td>SOFT</td>
<td>Exchange format</td>
</tr>
<tr>
<td>MIMiML</td>
<td>Exchange format</td>
</tr>
<tr>
<td>GO</td>
<td>Terminology artifact</td>
</tr>
<tr>
<td>EFO</td>
<td>Terminology artifact</td>
</tr>
<tr>
<td>OBI</td>
<td>Terminology artifact</td>
</tr>
<tr>
<td>MGED Ontology</td>
<td>Terminology artifact</td>
</tr>
<tr>
<td>MAGE-OM</td>
<td>Object model</td>
</tr>
<tr>
<td>FuGE</td>
<td>Object model</td>
</tr>
<tr>
<td>SEND</td>
<td>Exchange format</td>
</tr>
<tr>
<td>GEML</td>
<td>Exchange format</td>
</tr>
<tr>
<td>FUGO</td>
<td>Terminology artifact</td>
</tr>
<tr>
<td>MAML</td>
<td>Exchange format</td>
</tr>
</tbody>
</table>
## Use cases

<table>
<thead>
<tr>
<th>Level of Rigor</th>
<th>Use case example</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Inter-lab collaboration</td>
<td>Data should meet minimal standards for structure and documentation to enable comprehension, but answers to questions are just an email/phone call/hallway away. At least until that person leaves the lab.</td>
</tr>
<tr>
<td>Medium</td>
<td>Publishing</td>
<td>Data should use standardized formats and annotation sufficient to enable both comprehension and reproducibility, with little or no interaction with the data owner.</td>
</tr>
<tr>
<td>High</td>
<td>Make available through public data repository</td>
<td>In addition to being comprehensible and reproducible, annotation should be structured in a way that enables querying for datasets that match specific criteria.</td>
</tr>
</tbody>
</table>
Standards Criteria

- The standard itself
- Adoption and user community
- Additional factors
The Standard Itself

• Specification documentation
• Ease of implementation (e.g. need for programmer support)
• Human and machine readability
• Formal structure
• Expressivity—breadth of information that can be represented
• Ease of use, e.g., minimal required fields, text-based interface familiarity to biologists.
Adoption and User Community

- Broad adoption and implementation, outside initial group
- Support supplied by the user community
- Use by community databases
- Software development that supports the standard (e.g., for curating, submitting to databases)
- Responsiveness to community requests
- Availability of examples of use
- Requirements of relevant authoritative bodies, e.g., funders, publishers, etc.
Additional Factors

• Integration/compatibility with other standards
• Extensibility and flexibility to cover new domains
• Conversion and mapping, when applicable
• Cost (e.g., open vs. licensing fee)
Potential resources to assist in standards selection and adoption

- Lay person’s primer to standards
- Consumer reviews
- Standards selection wizard
- Standards adoption helpdesk
- Quality assurance tools
Acknowledgments

• Contributors to the BioSharing catalog
• CTSA Omics data standards working group
  • Simon Lin
  • Bill Barry
  • David Beck
  • Colette Blach
  • Jim Cimino
  • Todd Ferris
  • Carol Haynes
  • Curtis Hendrickson
  • Carol Hill
  • Ken Kawamoto
  • Tahsin Kurc
  • John Osborne
  • Jeff Pennington
  • Sarah Wheelan

• Helpful Resources
  • Mark Musen
  • Richard Scheuermann
  • Kristi Eckeron
  • Russ Altman

Funding
  • NIH UL1RR024128
  • David H. Murdock
  • NIH R24OD011883
  • CTSA 10-001:100928SB23
  • Oxford e-Research Centre
  • UK Biotechnology and Biological Sciences Research Council (BBSRC) BB/I000771/1 and BB/I025840/1