Gaps in actionable scholarly publishing metadata

Megan O’Donnell
Head of Research Data Services
Iowa State University, University Library
Disclaimer: I’m not a metadata librarian

Head of Research Data Services
part time subject librarian

- I manage an institutional data repository.
- I curate and publish data sets.
- I also advise, consult, and teach.
- Publication metadata is data.
Libraries are:

- Mechanisms of discovery and knowledge making.
- Heavy consumers AND producers of (meta)data.
- Information and data management experts.
- Maintainers of official records.
- And publishers.

But not most scholarly publishing records.
Typical metadata

Institutional Repositories

• Original works (ETD, tech reports, etc.)
  • Without assigned DOIs they’re not part of the scholarly linked data environment.

Green OA

• Can link back to the official publication but this generates multiple sources of valuable metrics.
Typical metadata

People IDs (ORCID):

- Typically, opt-in but with uneven adoption.
- Still not a part of university systems.
  - Not well integrated with **HR systems**.
  - Not “baked in” to most **library systems**.
- Highly reliant on authors to create, maintain, and **link**.
  - *Active and retroactive record updates from authors and libraries are possible.*
Emerging metadata

Resources and Centers (HPC, imaging, etc.):

• Track use and justify funding for expensive resources and expertise.
  • Who did we help? What research did we enable?
• Standards are needed.
  • Attribution as authors? By citation?
  • Libraries can manage and update records for local resources.
Emerging metadata

Billing Verification for R&P Agreements

- Wrong institutions listed
  - College / Lab / Center ≠ University
- Email address errors.
  - Library is never contacted.
  - Email addresses belong to child-orgs.
- Query ROR for related organizations?
  - Not granular enough?
  - Doesn’t track web domains.
Problematic metadata

Funding and awards

• **Standards are needed.**
  • Award ID numbers
  • Discovery and verification
  • Attribution guidelines and metadata.
• Support for different award types:
  • Cooperative agreements, contracts, etc.

Ames Laboratory operates under a DOE contract and has project numbers.

CSAFE is funded through a cooperative agreement with NIST and four institutions. We have been given three different numbers over four-years.
Incomplete metadata

Data, data sets, datasets, and databases

• Lack of metadata coordination between publishers and repositories.
  • ORCIDs*, funding,* article publication date,* ...
  • Publications are not updated with links to data
  • data is updated with links to publications.
• You cannot query which publications:
  • Have data or claim to have data.*

Poor metadata & discovery chaos

- **Item type** is underdeveloped for non-traditional items.

**Data set =**
- Tabular data (yes)
- File sets (often)
- Tables (no)
- Graphs (no)
- PDFs (no)

- This is not useful metadata.

---

**TABLE 1.** Crude and age-adjusted prevalence of self-reported physician-diagnosed diabetes among Asians and Native Hawaiians or other Pacific Islanders (NHPI) aged ≥18 years, in 32 states, the District of Columbia, and Guam — Behavioral Risk Factor Surveillance System, United States, 2011–2014

<table>
<thead>
<tr>
<th>State/Territory</th>
<th>Race</th>
<th>No. of respondents</th>
<th>Crude prevalence</th>
<th>Age-adjusted prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>Asian</td>
<td>313</td>
<td>7.1 (4.5–11.1)</td>
<td>9.0 (5.6–14.3)</td>
</tr>
<tr>
<td>Arizona</td>
<td>Asian</td>
<td>108</td>
<td>3.4 (2.0–5.8)</td>
<td>4.9 (2.9–8.0)</td>
</tr>
<tr>
<td>California</td>
<td>Asian</td>
<td>3,102</td>
<td>8.9 (7.7–10.3)</td>
<td>9.7 (8.4–11.1)</td>
</tr>
<tr>
<td>Colorado</td>
<td>Asian</td>
<td>589</td>
<td>5.2 (3.5–7.8)</td>
<td>7.1 (4.9–10.1)</td>
</tr>
<tr>
<td>Connecticut</td>
<td>Asian</td>
<td>641</td>
<td>4.8 (3.3–7.0)</td>
<td>6.0 (4.6–11.3)</td>
</tr>
<tr>
<td>District of Columbia</td>
<td></td>
<td>351</td>
<td>2.8 (1.6–4.9)</td>
<td>3.1 (3.0–8.5)</td>
</tr>
<tr>
<td>Florida</td>
<td>Asian</td>
<td>661</td>
<td>7.6 (5.5–10.5)</td>
<td>10.2 (7.5–13.7)</td>
</tr>
<tr>
<td>Georgia</td>
<td>Asian</td>
<td>371</td>
<td>4.7 (2.8–7.6)</td>
<td>6.9 (5.0–14.5)</td>
</tr>
<tr>
<td>Hawaii</td>
<td>NHPI</td>
<td>930</td>
<td>10.7 (8.4–13.6)</td>
<td>13.9 (11.1–17.3)</td>
</tr>
<tr>
<td>Asian</td>
<td>930</td>
<td></td>
<td></td>
<td>8.8 (6.0–9.6)</td>
</tr>
<tr>
<td>Illinois</td>
<td>Asian</td>
<td>520</td>
<td>5.7 (3.8–8.4)</td>
<td>6.9 (4.2–9.2)</td>
</tr>
<tr>
<td>Indiana</td>
<td>Asian</td>
<td>318</td>
<td>6.3 (2.9–10.2)</td>
<td>8.5 (5.6–15.6)</td>
</tr>
<tr>
<td>Iowa</td>
<td>Asian</td>
<td>230</td>
<td>5.2 (3.2–8.5)</td>
<td>6.6 (4.4–10.2)</td>
</tr>
<tr>
<td>Kansas</td>
<td>Asian</td>
<td>689</td>
<td>4.2 (2.0–6.0)</td>
<td>5.7 (4.0–8.9)</td>
</tr>
<tr>
<td>Maryland</td>
<td>Asian</td>
<td>965</td>
<td>7.1 (5.4–9.2)</td>
<td>7.1 (5.3–9.2)</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>Asian</td>
<td>628</td>
<td>5.5 (4.1–7.3)</td>
<td>6.7 (4.9–9.1)</td>
</tr>
<tr>
<td>Michigan</td>
<td>Asian</td>
<td>470</td>
<td>4.9 (3.1–7.4)</td>
<td>5.6 (4.3–10.2)</td>
</tr>
<tr>
<td>Minnesota</td>
<td>Asian</td>
<td>892</td>
<td>3.8 (2.6–5.7)</td>
<td>4.2 (2.7–5.9)</td>
</tr>
<tr>
<td>Nebraska</td>
<td>Asian</td>
<td>455</td>
<td>4.3 (2.6–6.9)</td>
<td>4.6 (2.9–7.0)</td>
</tr>
<tr>
<td>Nevada</td>
<td>Asian</td>
<td>485</td>
<td>12.2 (8.8–16.8)</td>
<td>14.0 (10.0–17.7)</td>
</tr>
<tr>
<td>New Jersey</td>
<td>Asian</td>
<td>251</td>
<td>7.8 (5.6–10.5)</td>
<td>9.1 (6.9–13.7)</td>
</tr>
<tr>
<td>New Mexico</td>
<td>Asian</td>
<td>250</td>
<td>7.9 (5.0–12.3)</td>
<td>9.5 (6.5–14.7)</td>
</tr>
</tbody>
</table>

---

An image of a data table accompanied by the French phrase: "this is not a data set.” (this is play off Rene Magritte’s “The Treachery of Images,” 1929).
Summary of gaps and next steps:

• More standards and standardization.
• Metadata verification (who and how).
• Post-publication updates and enrichment.
• Communication between data and text publishers.
• Goals and ethics for scholarly publication metadata (it is data after all...).