Steering Committee

Mark Musen,  
*Principal Investigator*

Carol Bean,  
*Project Manager*

Michel Dumontier

Olivier Gevaert

Purvesh Khatri

Steven Kleinstein

Susanna-A Sansone

Jeffrey Wiser

Kei-Hoi Cheung

Stanford Medicine

Yale University School of Medicine

University of Oxford

Northrop Grumman
Community-developed **content standards**

- To **structure**, **enrich** and **report** the description of the datasets and the experimental context under which they were produced
- To facilitate **discovery**, **sharing**, **understanding** and **reuse** of datasets

Including **conceptual model**, **conceptual schema** from which an exchange format is derived to allow data to flow from one system to another

Including **controlled vocabularies**, **taxonomies**, **thesauri**, **ontologies** etc. to use the same word and refer to the same ‘thing’

Including **minimum information reporting requirements**, or **checklists** to report the same core, essential information
In the life sciences alone… almost 600!
Researchers *hate* standards!

- Most researchers understand the value of standardized descriptions, when using third-party datasets.
- But when asked to structure their datasets, they view requests for even “minimal” information as burdensome.
Researchers *hate* standards!

- Most researchers understand the value of standardized descriptions, when using third-party datasets.
- But when asked to structure their datasets, they view requests for even “minimal” information as burdensome.

➤ There is an urgent need to lower the bar for authoring good metadata.
Our two initial use cases

Stanford Digital Repository - Online Deposit

The SDR is a service supporting long-term management of scholarly information resources at Stanford. Faculty, students, and researchers use the SDR to promote and protect the products of their work. The benefits of this service distinguish the SDR from other content storage or management options on campus: deposited scholarly content is preserved in a robust, reliable, and secure environment and is available from persistent URLs (PURLs) with optional access controls.

You  Stanford Libraries  Web Users

- deposit items  →  provides long-term preservation  →  discover your deposits via SearchWorks

access your deposits via PURLs
1. Map the landscape of content standards

bioSharing.org

**Policies Registry**
A catalogue of data preservation, management and sharing policies from international funding agencies and regulators.

**Standards Registry**
A catalogue of reporting standards and organizations that develop these. Partly compiled by linking to BioPortal, mibbi, and equator network.

**Databases Registry**
A catalogue of databases, described according to the BioDBcore guidelines, along with the standards used within them. Partly compiled with the support of OUP and re3data.org.

Logos of various organizations.
Almost 600!
2. Develop methods for creating templates

**STRUCTURE**
Authoring of Metadata Templates

Template authors define

Metadata templates

**ANNOTATE**
Annotation of Data with Metadata

Scientists contribute

 metadata acquisition forms

**EXPLORE**
Exploration and Reuse of Datasets through Metadata

Metadata repository

Scientists fill in search, reuse

Scientists contribute

Metadata acquisition forms
2. Develop methods for creating templates

**STRUCTURE**
Authoring of Metadata Templates

- Template authors define
- Metadata templates

**ANNOTATE**
Annotation of Data with Metadata

- Scientists contribute
- Metadata acquisition forms
- Metadata repository

**EXPLORE**
Exploration and Reuse of Datasets through Metadata

- Scientists fill in search, reuse
- Metadata repository

---

use ‘elements’ from content standards

create a language to represent relations among ‘elements’

use existing examples of templates

[Images and logos from bioSharing.org, W3C HCLS WGs, ImmPort, and iSAtools]
3. Develop methods to ease use of templates

- Enable researchers to help us creating templates appropriate to their needs
- Help researchers to **find** and **use** these templates to describe their experiments, and **populate** them with appropriate values (e.g. terms from ontologies)
4. Create a repository of populated templates

CEDAR repository will:
- **store** the experimental descriptions
- **facilitate** submission of datasets to our two case study repositories and progressively to other recognized online repositories
5. Exploring ways to enhance metadata

- Analyze the CEDAR repository to reveal patterns in the metadata that will enable the metadata tools to use predictive data entry to ease the task of filling out the templates.
- Augment those metadata with links to the published literature (including secondary analyses and retractions!)
- Augment those metadata with links to follow-up experiments (in online databases and in the literature)
- Allow the scientific community to comment on the experiment through structured metadata

Learn how to ease the authoring of metadata, using community standards, to enhance the richness of the experimental descriptions.