UNIVERSITY OF MIAMI
CENTER for
COMPUTATIONAL
SCIENCE

Data Standards for Data Integration

Stephan Schürer
University of Miami

SPARC Workshop, NIH, Feb 25-26 2015

sschurer@med.miami.edu

UNIVERSITY
OF MIAMI

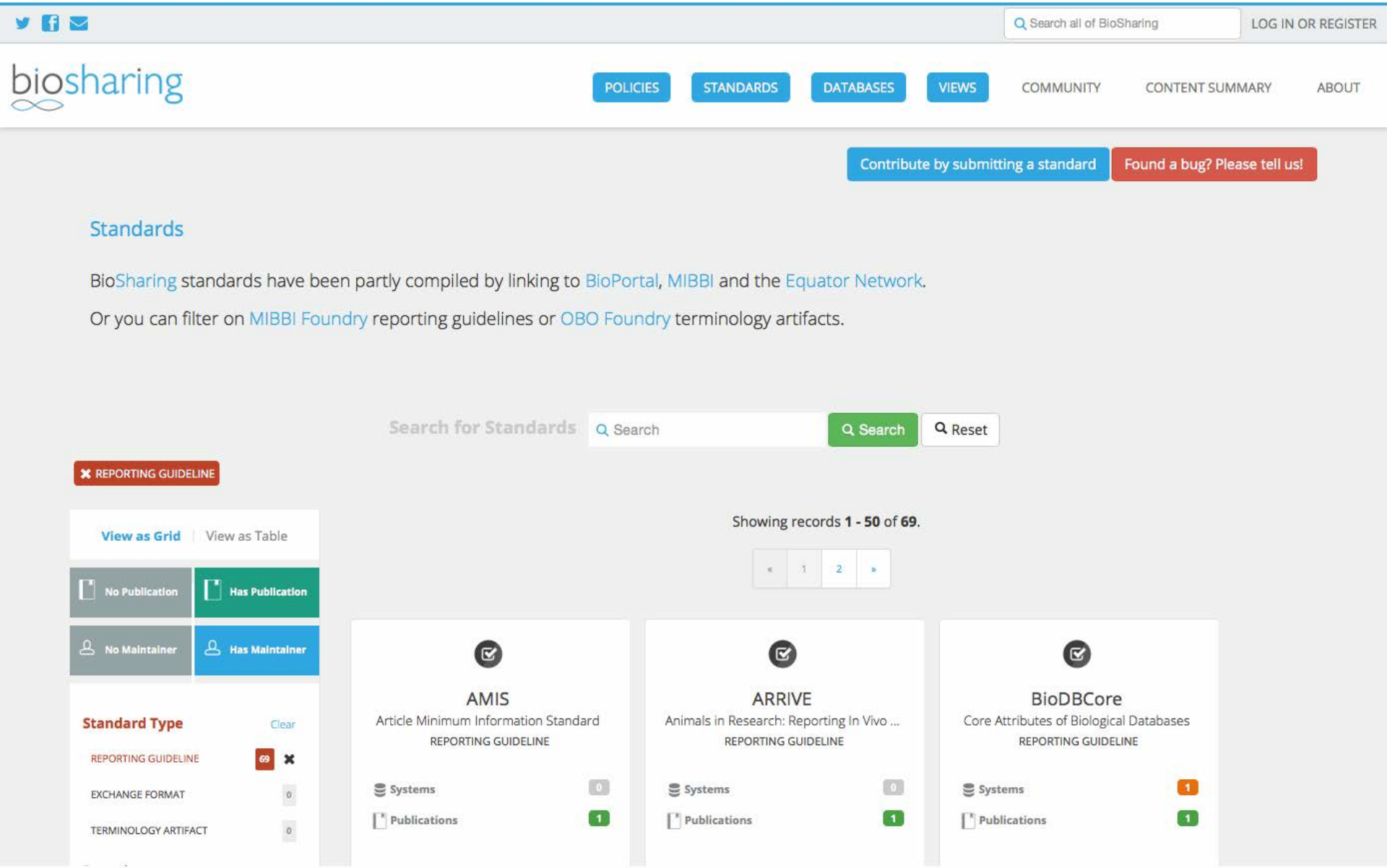
Types of data standards

- **Reporting guideline (checklist)** specifies what information need to be captured about an experiment for a particular purpose
- **(Controlled) vocabulary** terminological resource that provides the identification and definition of entities

- **Data exchange format** is a specification how data are encoded to be computer-readable / -processable

- **Data structure** refers to organization of data, data schema, entity relations

BioSharing Standards (http://www.biosharing.org)



Checklists – Minimum Information Guidelines

mibbi

Bioscience reporting guidelines and tools

- Portal
- Foundry
- About

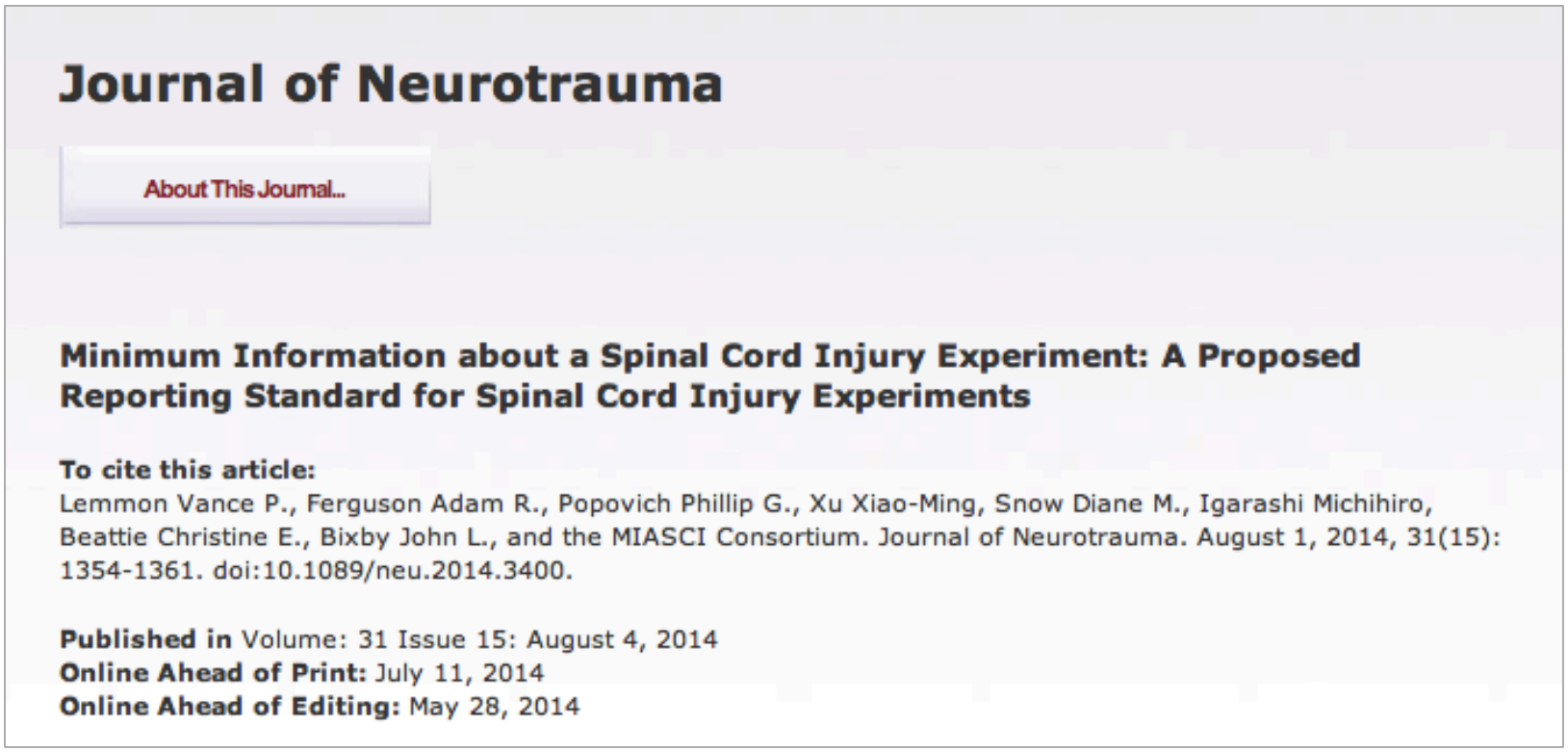
Minimum Information guidelines from diverse bioscience communities

- If you want to register your checklist to MIBBI, please contact the BioSharing team
- Excel spreadsheet and XML document (schema) describing all registered projects

Bioscience projects registered with MIBBI

CIMR	Core Information for Metabolomics Reporting	T	L	X
GIATE	Guidelines for Information About Therapy Experiments			
MIABE	Minimal Information About a Bioactive Entity	T	L	X
MIABiE	Minimum Information About a Biofilm Experiment			
MIACA	Minimal Information About a Cellular Assay			
MIAME	Minimum Information About a Microarray Experiment	T	L	X
MIAPA	Minimum Information About a Phylogenetic Analysis			
MIAPAR	Minimum Information About a Protein Affinity Reagent			
MIAPE	Minimum Information About a Proteomics Experiment			
MIAPepAE	Minimum Information About a Peptide Array Experiment			
MIARE	Minimum Information About a RNAi Experiment	T	L	X
MIASE	Minimum Information About a Simulation Experiment	T	L	X
MIASPPE	Minimum Information About Sample Preparation for a Phosphoproteomics Experiment			
MIATA	Minimum Information About T Cell Assays			
MICEE	Minimum Information about a Cardiac Electrophysiology Experiment			
MIDE	Minimum Information required for a DMET Experiment			
MIFlowCyt	Minimum Information for a Flow Cytometry Experiment	T	L	X

Minimum Information Standard may not exist



Journal of Neurotrauma

About This Journal...

Minimum Information about a Spinal Cord Injury Experiment: A Proposed Reporting Standard for Spinal Cord Injury Experiments

To cite this article:
Lemmon Vance P., Ferguson Adam R., Popovich Phillip G., Xu Xiao-Ming, Snow Diane M., Igarashi Michihiro, Beattie Christine E., Bixby John L., and the MIASCI Consortium. Journal of Neurotrauma. August 1, 2014, 31(15): 1354-1361. doi:10.1089/neu.2014.3400.

Published in Volume: 31 Issue 15: August 4, 2014
Online Ahead of Print: July 11, 2014
Online Ahead of Editing: May 28, 2014

Regenbase: Integration of diverse data related to nerve regeneration in the context of spinal cord injury

http://regenbase.org

Vocabulary vs. ontology

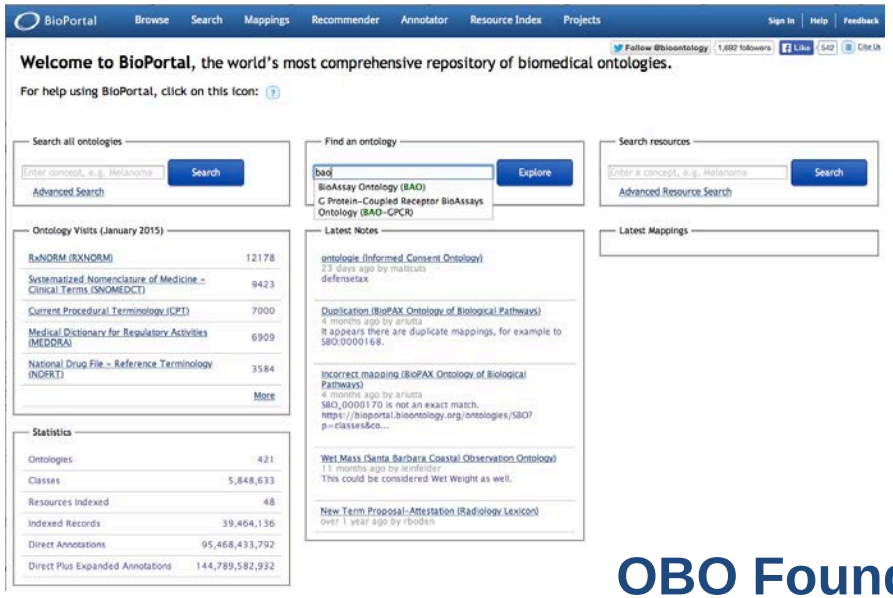
- Controlled vocabularies / thesauri
  - describe what things mean (link terms to human description)
  - Entities with identity criteria
  - Share knowledge in a common language
  - Natural language synonyms for search and text mining

Vocabulary vs. ontology

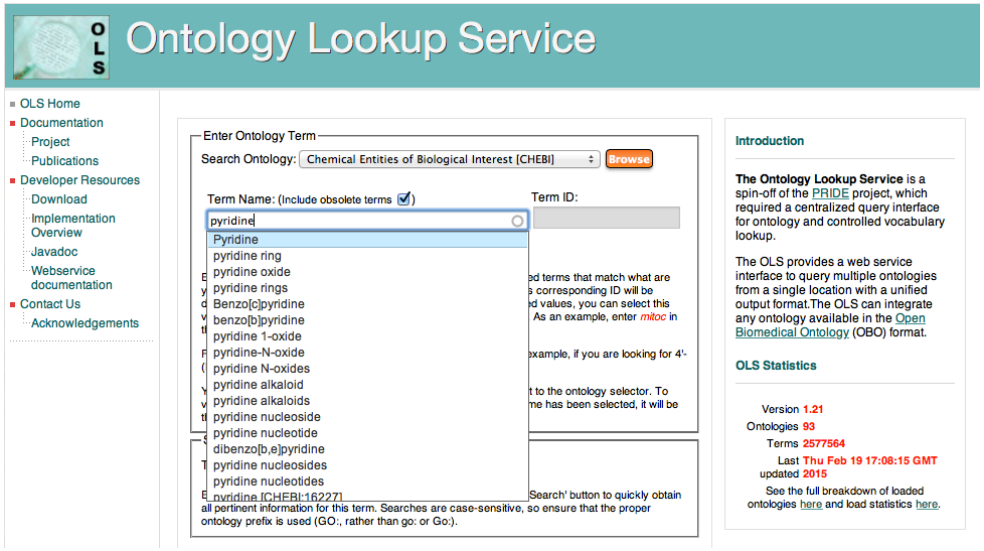
- Ontologies
  - Contains entities (classes) and their relationships (object properties)
  - Capture / abstract knowledge using logical axioms
  - Explicit specification (OWL-DL)
    - Building formal (computable) models
    - Computing with knowledge (reasoning engines)
    - Foundation of Semantic Web information systems

Ontology resources

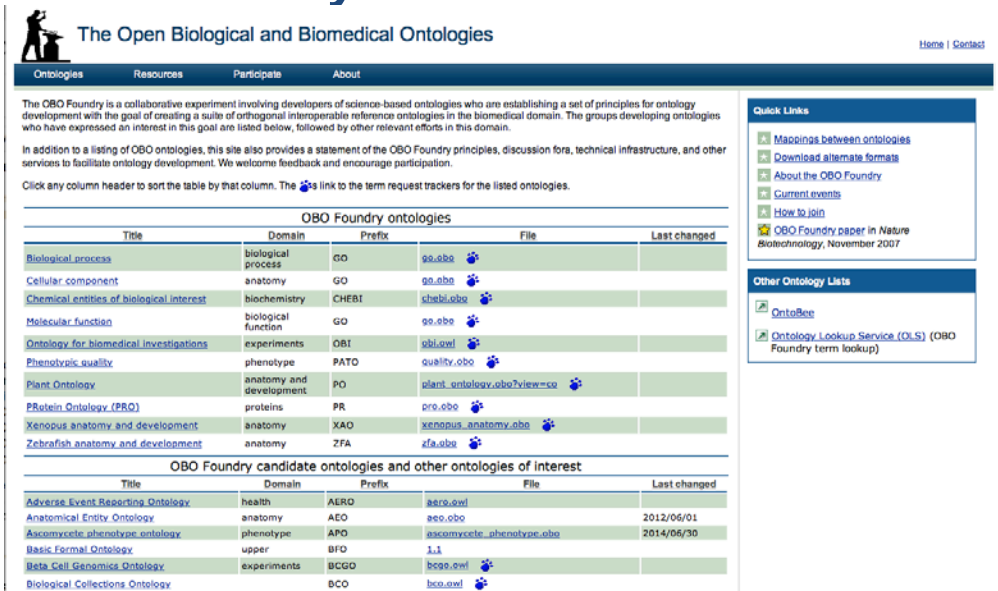
NCBO Bioportal



EBI Ontology Lookup Service OLS



OBO Foundry



Metadata specifications

Metadata: Data not directly measured in an experiment (or obtained in a study)

Why metadata:

- Facilitate data replicability, reproducibility, reuse
- Interpret results, perform data analysis, hypotheses
- Repurpose data for other projects
- Information systems (search, query, data integration and exchange)

What metadata to capture in a standardized format with controlled vocabularies (and formal descriptions)?

A useful distinction of metadata

Model metadata:

- Required to understand, interpret, and meaningfully integrate experimental results
- Typically queryable in software systems
- Important parameters to describe conclusions (data visualizations)

Confounder metadata:

- Non model metadata required to replicate and reproduce experimental results
- Needed for “data forensics” (e.g. batches of reagents, maintenance of experimental equipment, etc.)

Standardized metadata

Capture all (detailed descriptions, SOP)

Make model metadata explicit (controlled vocabulary, standard format)

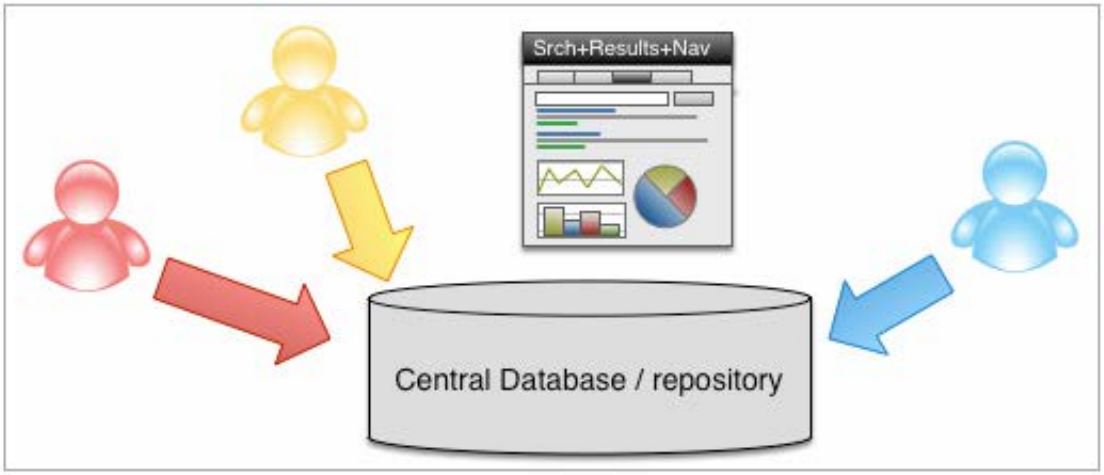
But what's really model metadata?

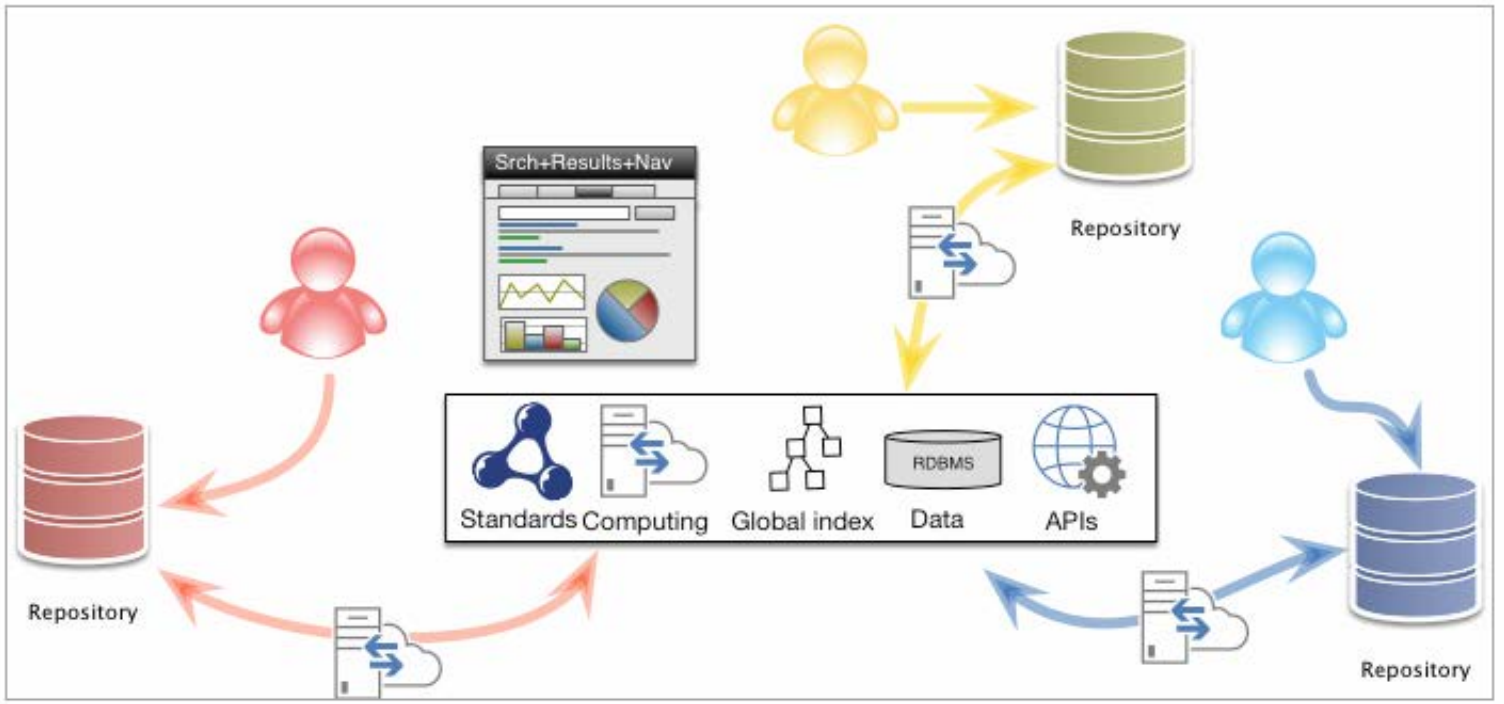
Data and informatics use cases

- Types of queries and analyses
- Integration with other data sources
- Information systems / UI components
- Consider re-use of data for other projects

LINCS Metadata Standards: Vempati et al *J Biomol Screen* 2014

Data Coordination





Data set IDs and provenance

Permanent ID via (authoritaitve) repository or data publication

- DOI
- PURL

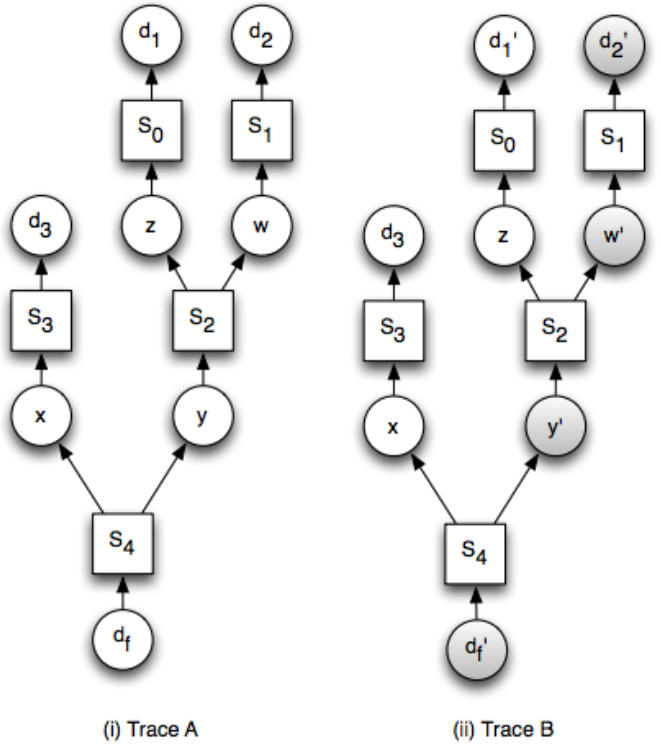
Capture data provenance

- PROV-O: The PROV Ontology (W3C) http://www.w3.org/TR/prov-o/
- PAV (Provenance, Authoring Versioning) Ontology http://purl.org/pav/

Provenance

The link between source data, computation / processing and derived data / results

- static verifiable record
- track changes
- compare / discrepancies
- repeat / reproduce
- Citation
- version
- data release



PDIFF, Woodman 2011

RDBMS vs Semantic Web technologies

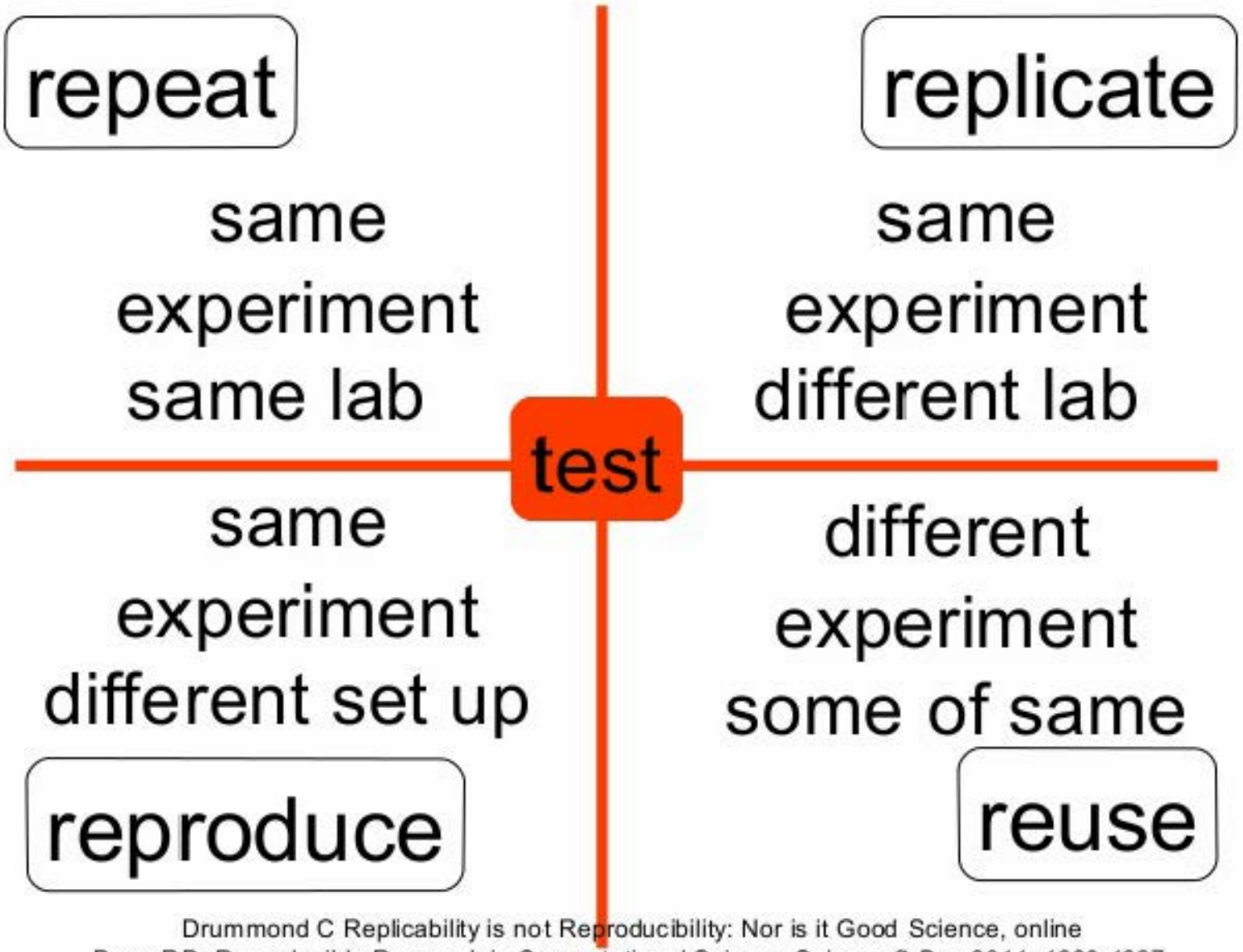
RDBMS

- Closed world assumption
- No reasoning support
- Need to know schema for highly specialized queries
- Data sharing not easy, no semantics
- Efficient RDBMS access
- Established technology
- Industry standard

Ontologies

- Open world assumption
- Reasoning support
- Provide restriction-free framework (formal semantics)
- Easy data/knowledge sharing
- Triple store access
- Relatively early stage
- Standards emerging

Replicability vs Reproducibility



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*