



Data Standards for Data Integration

Stephan Schürer
University of Miami

SPARC Workshop, NIH, Feb 25-26 2015

sschurer@med.miami.edu



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>)



Search all of BioSharing

LOG IN OR REGISTER



POLICIES

STANDARDS

DATABASES

VIEWS

COMMUNITY

CONTENT SUMMARY

ABOUT

Contribute by submitting a standard

Found a bug? Please tell us!

Standards

BioSharing standards have been partly compiled by linking to [BioPortal](#), [MIBBI](#) and the [Equator Network](#).

Or you can filter on [MIBBI Foundry](#) reporting guidelines or [OBO Foundry](#) terminology artifacts.

Search for Standards

Search

Search

Reset

REPORTING GUIDELINE

View as Grid | View as Table

No Publication

Has Publication

No Maintainer

Has Maintainer

Standard Type

Clear

REPORTING GUIDELINE

69 X

EXCHANGE FORMAT

0

TERMINOLOGY ARTIFACT

0

Showing records 1 - 50 of 69.

« 1 2 »



AMIS

Article Minimum Information Standard
REPORTING GUIDELINE

Systems

0

Publications

1



ARRIVE

Animals in Research: Reporting In Vivo ...
REPORTING GUIDELINE

Systems

0

Publications

1



BioDBCore

Core Attributes of Biological Databases
REPORTING GUIDELINE

Systems

1

Publications

1

Checklists – Minimum Information Guidelines



- 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	I	L	X
GIATE	Guidelines for Information About Therapy Experiments			
MIABE	Minimal Information About a Bioactive Entity	I	L	X
MIABIE	Minimum Information About a Biofilm Experiment			
MIACA	Minimal Information About a Cellular Assay			
MIAME	Minimum Information About a Microarray Experiment	I	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	I	L	X
MIASE	Minimum Information About a Simulation Experiment	I	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	I	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

Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies.

For help using BioPortal, click on this icon: ?

Search all ontologies: Enter concept, e.g. Melanoma. Search

Find an ontology: bio. Explore

Search resources: Enter a concept, e.g. Melanoma. Search

Ontology Visits (January 2015)

RxNorm (RXNORM)	12178
Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT)	9423
Current Procedural Terminology (CPT)	7000
Medical Dictionary for Regulatory Activities (MEDDRA)	6909
National Drug File - Reference Terminology (NDFRT)	3584
More	

Statistics

Ontologies	421
Classes	5,848,633
Resources Indexed	48
Indexed Records	39,464,136
Direct Annotations	95,466,433,792
Direct Plus Expanded Annotations	144,789,582,932

EBI Ontology Lookup Service OLS

Ontology Lookup Service

Enter Ontology Term: pyridine

Search Ontology: Chemical Entities of Biological Interest (CHEBI) Browse

Term Name: (Include obsolete terms)

Term ID:

pyridine

Pyridine

pyridine ring

pyridine oxide

pyridine rings

Benzo[c]pyridine

pyridine-1-oxide

pyridine-N-oxide

pyridine N-oxides

pyridine alkaloid

pyridine nucleoside

pyridine nucleotide

dibenz[b,e]pyridine

pyridine nucleosides

pyridine nucleotides

pyridine [CHEBI:16227]

Search button to quickly obtain all pertinent information for this term. Searches are case-sensitive, so ensure that the proper ontology prefix is used (GO, rather than go; or Go).

Introduction

The Ontology Lookup Service is a spin-off of the PRIDE project, which required a centralized query interface for ontology and controlled vocabulary lookup.

The OLS provides a web service interface to query multiple ontologies from a single location with a unified output format. The OLS can integrate any ontology available in the Open Biomedical Ontology (OBO) format.

OLS Statistics

Version 1.21

Ontologies 83

Terms 2577564

Last Thu Feb 19 17:08:15 GMT updated 2015

See the full breakdown of loaded ontologies here and load statistics here.

OBO Foundry

The Open Biological and Biomedical Ontologies

Home | Contact

Ontologies Resources Participate About

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other relevant efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the OBO Foundry principles, discussion fora, technical infrastructure, and other services to facilitate ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The link to the term request trackers for the listed ontologies.

Title	Domain	Prefix	File	Last changed
Biological process	biological process	GO	go.obo	
Cellular component	anatomy	GO	go.obo	
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo	
Molecular function	biological function	GO	go.obo	
Ontology for biomedical investigations	experiments	OBIT	obit.owl	
Phenotypic quality	phenotype	PATO	quality.obo	
Plant Ontology	anatomy and development	PO	plant_ontology.obo?view=so	
Protein Ontology (PRO)	proteins	PR	pro.obo	
Xenopus anatomy and development	anatomy	XAO	xenopus_anatomy.obo	
Zebrafish anatomy and development	anatomy	ZFA	zfa.obo	

Title	Domain	Prefix	File	Last changed
Adverse Event Reporting Ontology	health	AERO	aero.owl	
Anatomical Entity Ontology	anatomy	AEO	aao.obo	2012/06/01
Ascomycete phenotypic ontology	phenotype	APD	ascomycete_phenotypic.obo	2014/06/30
Basic Formal Ontology	upper	BFO	1.1	
Beta Cell Genomics Ontology	experiments	BCCO	bcgo.owl	
Biological Collections Ontology		BCO	bco.owl	

Quick Links

- Mappings between ontologies
- Download alternate formats
- About the OBO Foundry
- Current events
- How to join
- OBO Foundry paper in Nature Biotechnology, November 2007

Other Ontology Lists

- OntoBe
- Ontology Lookup Service (OLS) (OBO Foundry terms lookup)

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)

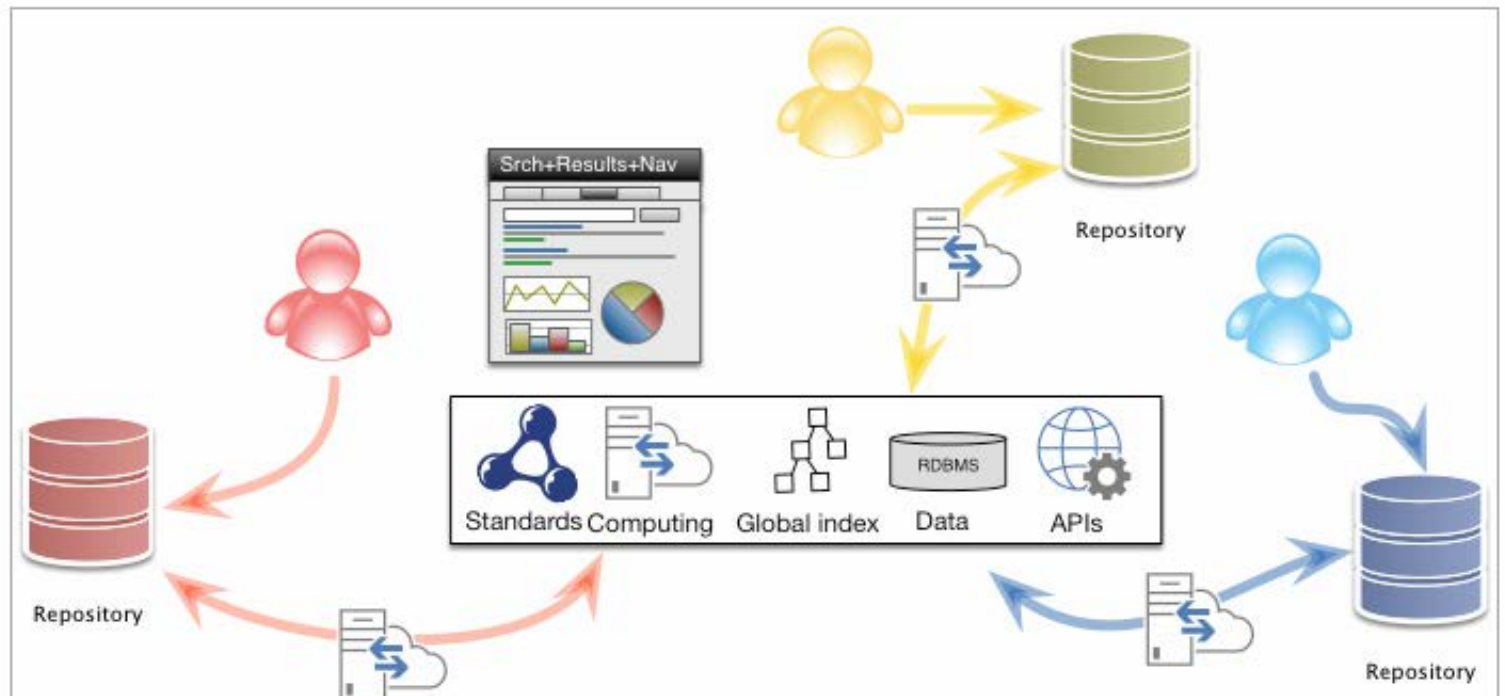
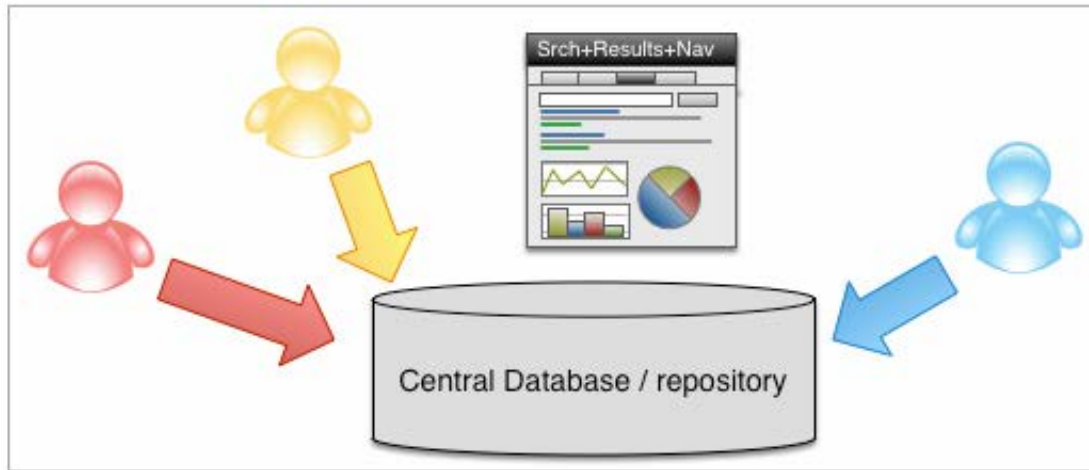
Make model metadata explicit (controlled vocabulary, standard format)

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

Data Coordination



Data set IDs and provenance

Permanent ID via (authoritative) repository or data publication

- DOI
- PURL

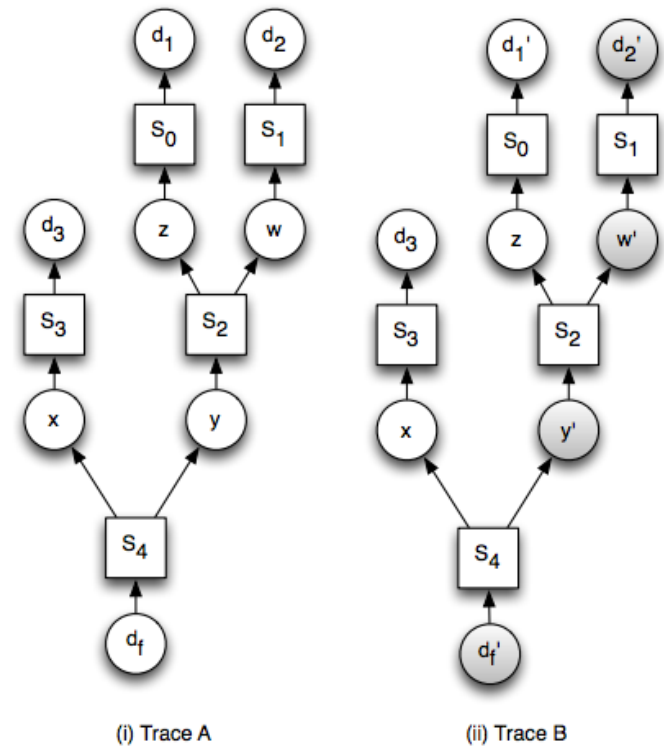
Capture data provenance

- 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

