

Informatics for Data and Resource Discovery in Addiction Resource

Summary Discussion

Panelists: Maryann Martone, Ph.D., Matthew McAuliffe, Ph.D., Elissa Chesler, Ph.D., Gwen Frishkoff, Ph.D., Sudeshna Das, Ph.D., and Jeffrey Grethe, Ph.D.

Summary

- .Identification of Common Themes
- .Identification of Common Approaches
- .Suggestions for Best Practices
- .Suggestions for their Implementation

Common Themes

- There is a need for data to be readily machine and human interpretable
- There is a need to capture information from the 'Hidden Web'
- Resources need to be searched in a community-specific manner
- There is a need to convert data, image data to descriptive data, for example
- Data integrity is the responsibility of the data producer
- Attribution/Provenance
- Reproducibility and its exceptions (HIPAA data)
- Primary Data Identifiers (*i.e.* URI, GUID, GeneID)
- There is a need to cope with unfunded mandates (BRO). Current funding almost demands the use of open source technologies and *ad hoc* or interval development.
 - Anyone can make a database
 - Uneven life cycles
 - What happens at the end of a life cycle?

Common Themes

- To what extent is there a need to integrate Data over MetaData?
- There must be a careful examination of existing metaphors in life science (caBIG, HeadIT, PhenX, CDE, NGC, CAB)
- Harmonizing new versus legacy data, when is data deemed not useful to integrate?
- How do we measure the impact of collaborative or other integration resources?
- There is an emergent need to catalog resources and naming conventions, including common software toolkits
 - Meta-NIF, rich snippets (google)
 - URI, PURL, shared naming

Common Themes

- Barriers to Data Sharing
 - Security
 - Data Structure Complexity, do you need a full-time bioinformatician to participate? Different size groups have varying barrier heights
 - Annotation Complexity and Curation
 - Static vs. Dynamic Access
 - Homology
- Data Encapsulation
 - Temporal, Spatial, Digital, textual, protocols, among others
 - Levels of Prescriptive Management
 - RDF, OWL, Ontologies, Discrete Management, Federated, Warehouse
 - Minimal Information Standards (journals, supplemental material)
 - Collaboratories (Alzforum, WormBook, StemBook, Electronic Notebooks)

Ontologies

- Community Driven
- Integration of Ontology instability and iterative development cycles
- We need to do a better job leveraging existing ontologies
- Ontological alignment/mapping
- Ontologies as Artificial Partitions
 - Do researchers put too much confidence in ontologies?

Carrots vs. Sticks

- How do you encourage investigators to participate?
 - Compulsory (NDAR)
 - Restriction (PD Online)
 - Community Standard (NIF)
 - Utility Driven (ODE)
 - Are sticks counter-productive by ultimately stunting innovation?
 - Is the answer static for all resources (ODE = analysis-based, NIF = data-based)
 - Forcing data analytics on PIs often lead to very uneven results
 - Need for community buy-in
 - Must continue to encourage researchers that data is not knowledge

Best Practices & Implementation

- Specific Recommendations
- Scope
 - Are all themes and best practices discussed here within the scope of implementation?
- Strategies
 - DB-backed Federation, semantic or RDF approach, or higher level open data standards (data registration)
- Priorities
 - For example, is it imperative to establish a means to locate community relevant data before enforcing integrity constraints on that data?
 - Where on the Data – Information – Knowledge pyramid does implementation begin?

Best Practices

Map all data to shared (public) ontologies using full URI when possible

Make all supplementary data tables machine parseable (minimal journal and author standards)

Avoid hidden semantics; think about machine access AND human access

Use primary id most closely associated to real data values wherever possible, e.g., reagents (catalog number if available), stable identifiers

Use a standardized checklist for **internal documentation** and **reporting of provenance** for each experiment dataset.

Apply **data-driven techniques** for analysis of complex, high-dimensional datasets

Classify and label patterns using **pattern definitions** and **Data mining** from the literature

Data integrity and stewardship: how long should data be kept?

Clearly indicate date of accession of all data sources, and establish sufficient archiving or 'time-machine' for electronic research reproducibility within a grant revision or publication cycle. Approximately 2-3 years is a recommended minimum archive.

Enable traceable workflows and query paths to facilitate research reproducibility.

Establish local or central user-project storage and data sharing capabilities to facilitate collaboration and research reproducibility.

Best Practices

Best Practices

- Map all data to shared (public) ontologies using full URI when possible
- Identify resources through stable URI
- Make all supplementary data tables machine parseable
- Avoid hidden semantics; think about machine access AND human access
- Use identifiers in text for key items, e.g., reagents (catalog number if available)

Best Practices

BEST PRACTICE #1 (DATA PROVENANCE). Use a standardized checklist for *internal documentation* and *reporting of provenance* for each experiment dataset.

BEST PRACTICE #2 (DATA ANALYSIS— WITHIN DATASETS). Apply *data-driven techniques* for analysis of complex, high-dimensional datasets

BEST PRACTICE #3 (DATA METRICS). Describe patterns in complex, high-dimensional data using multiple metrics that fully characterize key dimensions of the data (e.g., temporal/spectral, spatial/topographic, individual conditions as well as contrasts where possible). Report full set of *data metrics*, along with *data provenance* (e.g., as Supplement). Note this "best practice" complements BP#1-2.

BEST PRACTICE #4 (DATA CLASSIFICATION & LABELING). Classify and label patterns using *pattern definitions* from the literature ("knowledge-driven" approach, to promote data sharing & integration)

BEST PRACTICE #5 (DATA ANALYSIS — ACROSS DATASETS). Classify and label patterns using *data mining* from the literature ("knowledge-driven" approach)

Best Practices

- **Data Integrity Principle:** Ensuring the integrity of research data is essential for advancing scientific, engineering, and medical knowledge and for maintaining public trust in the research enterprise. Although other stakeholders in the research enterprise have important roles to play, researchers themselves are ultimately responsible for ensuring the integrity of research data.
- **Data Access and Sharing Principle:** Research data, methods, and other information integral to publicly reported results should be publicly accessible.
- **Data Stewardship Principle:** Research data should be retained to serve future uses. Data that may have long-term value should be documented, referenced, and indexed so that others can find and use them accurately and appropriately

Best Practices

- Use primary id most closely associated to real data values wherever possible, e.g. when a microarray is in use, use the microarray probe set id, not the gene symbol presently associated with that id. Similarly, report genetic associations with respect to genomic features, markers or genotypes, not to build-specific map coordinates.
- Build structured vocabulary into all levels of data annotation
- Allow unstructured annotation (free text) to accompany structured annotation to capture information not presently well integrated into existing ontology
- Facilitate data entry and annotation such that users comply with standards for submission.
- Establish and enforce standards documents for data submission
- Seek robust and generalized mediators, rather than individual connections among research web sites.
- Create a platform for sharing of published and “unpublished” data, with a ‘caveat emptor’ indication of data source and curation status.
- Clearly indicate date of accession of all data sources, and establish sufficient archiving or ‘time-machine’ for electronic research reproducibility within a grant revision or publication cycle. Approximately 2-3 years is a recommended minimum archive.
- Enable traceable workflows and query paths to facilitate research reproducibility.
- Establish local or central user-project storage and data sharing capabilities to facilitate collaboration and research reproducibility.

Best Practices

- The Web - Use it. Contribute to it.
- Open Linked Data on the Web. Publish it. Support it.
- Open Access Journals. Publish in them wherever possible.
- Website development Use an open source CMS so that code is contributed to the public good.
- Open content. Use the Creative commons license wherever possible.
- Wheels. Don't re-invent them.
- Mash-ups. Do them.

Best Practices

- Stable identifiers: Every time the database updates, the identifiers should not change and cause all pre-indexed links to those data records break;
- Access: For increased utilization of the data, stable access needs to be provided either through a public connection to the database, a periodic dump of the database contents or through structured web services;
- Sessions: For general information results and data should be accessible using a static (i.e. non session based or stateless) URL;
- Vocabulary: Use a standard terminology and avoid symbolic notations where possible.