NDAR: A Model for Results Reporting to the National Data Service

The National Data Service seeks to develop a unified framework for storing, sharing, publishing, locating, and verifying data that are heterogeneous and come from multiple scientific disciplines. The National Database for Autism Research (NDAR) represents an existing platform, controlled by the NIH, which currently provides such services for the Autism Research community. With data shared on 77,500 de-identified research subjects harmonized to a community maintained data definition, NDAR is being held up as the model for data sharing across the broader Mental Health research community[1]. NDAR's capabilities, which may be relevant to others, include:

- A universal Global Unique Identifier (GUID), The NDAR GUID[i] is a universal subject ID that allows researchers to share data specific to a study participant without exposing personally identifiable information (PII) and makes it possible to match participants across labs and research data repositories.
- A community defined Data Dictionary[ii], the NDAR dictionary comprises over 400 data structures consisting of 60,000 discrete data elements.
- A Validation Tool[iii] ensuring data quality by confirming that all data submitted are first harmonized to the community data definition and data are validation to that data standard.
- A data-sharing regimen[iv] with the expectation that captured data is shared as it is acquired, and analyzed data is shared to a precise standard at the time of publication. NDAR will soon provide a DOI for each project/finding.
- Data federation and integration into public and private data repositories. NDAR does this for those that fund and control relevant research data wherever it exists, nationally and internationally.
- Computational approaches[v] designed to move computation to the data and provide mechanisms to receive computational results and make them available for software reuse, replication, and query[vi] in the same way as those results supporting scientific findings[vii].

The current practice of investigator defined data sharing is often insufficient in supporting result replication in the life sciences. More than just captured data is needed to replicate findings. The methods, software, tools, and results are also necessary. Furthermore, in the life sciences, research subjects by cohort definition, outcome measures defined to a common definition, and detailed analyses supporting research results are also important. We have defined such a model for autism that includes the following attributes (see http://ndar.nih.gov/data_from_papers.html):

Category	**Examples**
Record Info	Repository-specific ID, DOI, *etc.* Name, URL, contact info for data repository. Version number, dates, etc.
Bibliographic Descriptors	Dataset description, authors, investigators, curators, dataset URL, dates, versions, etc.
Methodological Descriptors	
• Cohorts	Grouping of data (*e.g.* research subjects or computational techniques) for analysis (*i.e.*, case/control, parent/child/sibling, pipeline).
• Measures	List of measures and data definition for the measures used.
• Study Design	Study description, controlled, observational, arms/comparisons, interventions, *etc.*
• Analysis	Keywords describing data analysis, *i.e.*, data transformation, statistical methods, equipment, software, algorithms, etc.
Data Descriptors	IDs for individual data items from collection of data; one data item could be used in multiple studies. Data Types (Raw, Analyzed, Sequencing, Microarray, *etc.*) URLs for accessing data (http, ftp, cloud resource *etc.*) Data Access/Use Restrictions (if any).

The National Data Service should consider the definition and promotion of such a standard for result replication, one that is extended to the life sciences, paving the way for independent validation of published results. The development and promotion of such a standard is practical and should be made essential for all repositories holding life sciences data. Then, publishers can raise the data sharing expectation, ensuring that the data are shared at time of publication and are easily accessible. NDAR would be happy to contribute to the definition and adoption of such a standard.

---

[1] Insel TR. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *Am J Psychiatry*, 2014 Apr 1; 171(4):395-7.

[i] NDAR GUID - http://ndar.nih.gov/standards.html#GUID
[ii] Data Dictionary - https://ndar.nih.gov/ndar_data_dictionary.html?type=All&source=NDAR&category=All
[iii] Validation Tool - https://ndar.nih.gov/ndarpublicweb/tools.html#validation
[iv] Data Sharing Regimen - http://ndar.nih.gov/contribute_data_sharing_regimen.html
[v] Computational Approaches - http://ndar.nih.gov/cloud_overview.html
[vi] Query - http://ndar.nih.gov/query_concept.html
[vii] Scientific Findings - http://ndar.nih.gov/data_from_papers.html