Implementing Semantics-Driven Data Exchange in Brain Science: The NCANDA Case Study

B Nolan Nichols[1, 2] Jean-Baptiste Poline[3] Russell A. Poldrack[4] Arno Klein[5] Dongjin Kwon[1] Weiwei Chu[1] Kilian M Pohl[1, 2]

[1]Center for Health Sciences, SRI International, Menlo Park, [2]Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, [3]Helen Wills Neuroscience Institute, UC Berkeley, CA, [4]Psychology, Stanford University, Stanford, CA, [5]Sage Bionetworks, Seattle, WA

Acknowledgments. We acknowledge the INCF NIDASH Task Force members for their many contributions. This project is funded by NIH NIAAA and OD (NCANDA Data Analysis Component, NIH 1 U01 AA021697; BD2K Supplement, NIH 1 U01 AA021697-04S1).

Project Overview

Problem Statement

The data sharing capabilities of large studies, such as the National Consortium on Alcohol & Neurodevelopment in Adolescence (NCANDA),[1] are currently limited by:

- Data management systems with no **common API** or **data exchange standard**
- A lack of data **provenance**, from acquisition and analysis, thus hindering reproducibility and transparency

NCANDA Data Repository

- 831 Participants (ages 12-22) followed annually for 3 years across 5 sites in the US
- Multimodal imaging – approximately 20TB of Anatomical, Diffusion, and Functional MRI
- Cognitive measures – 3,000 data elements across 8 domains (e.g., memory, reward, etc.)

Project Goals

- Enhance the long-term sustainability of NCANDA by applying NIDM, the Neuroimaging Data Model, to imaging and cognitive measures
- Facilitate data sharing and exchange by developing standard terminologies, data models, and APIs **as a community**

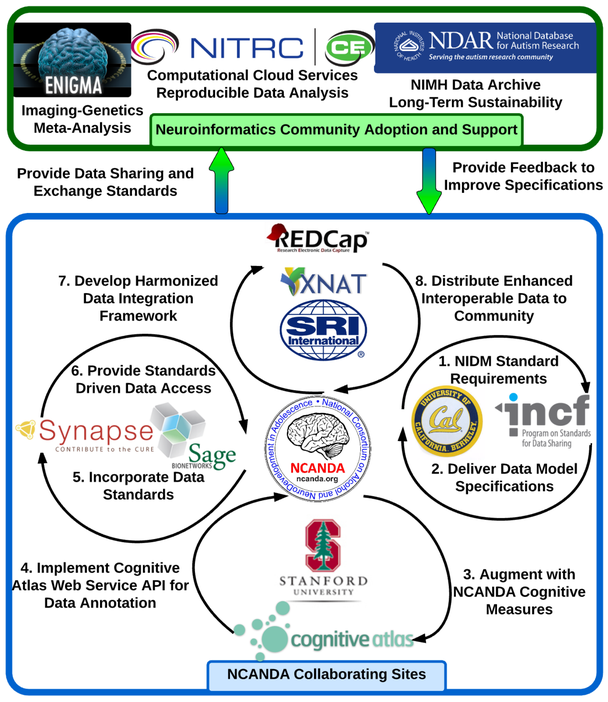

Figure 1. Deliverables and process for collaboratively enhancing the data and knowledge sharing infrastructure of NCANDA

NIDM a Linked Data Framework

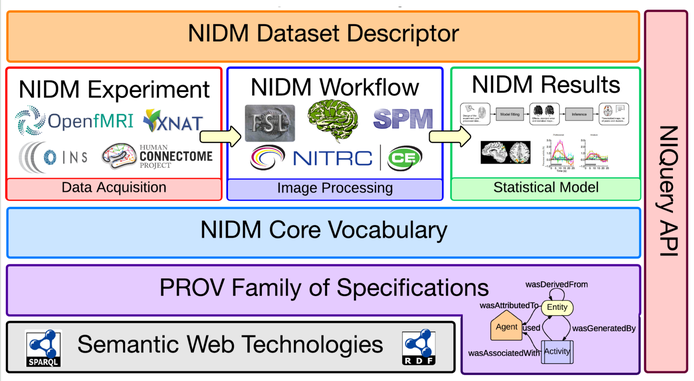

Figure 2. The technology "layer cake" used by NIDM

Neuroimaging Data Model (NIDM)[2]

- Extension of the PROV standard to represent provenance (Figure 2)
- **NIDM Experiment** - primary data capture
- **NIDM Workflow** - computational environment used for data analysis
- **NIDM Results** - minimal information sets when reporting analysis findings
- **NIDM Dataset Descriptors** - W3C HCLS Dataset Description extension
- **NIQuery API** - RESTful API wrapper around SPARQL Queries

Collaborative Data Modeling

- Use-cases are identified by working with scientists and other stakeholders (Figure 3)
- Open approach to form consensus around terms and relationships
- Community participation through reporting issues or during weekly web-based meetings
- Formal specifications define standards, for example, see: http://nidm.nidash.org/specs

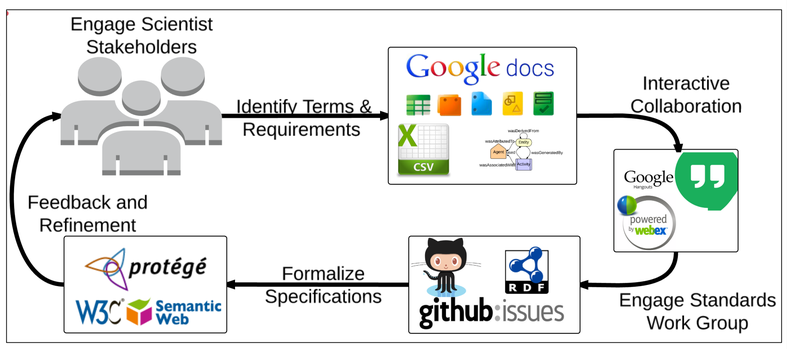

Figure 3. Collaborative process used to develop NIDM object models in a distributed and grass roots fashion[3]

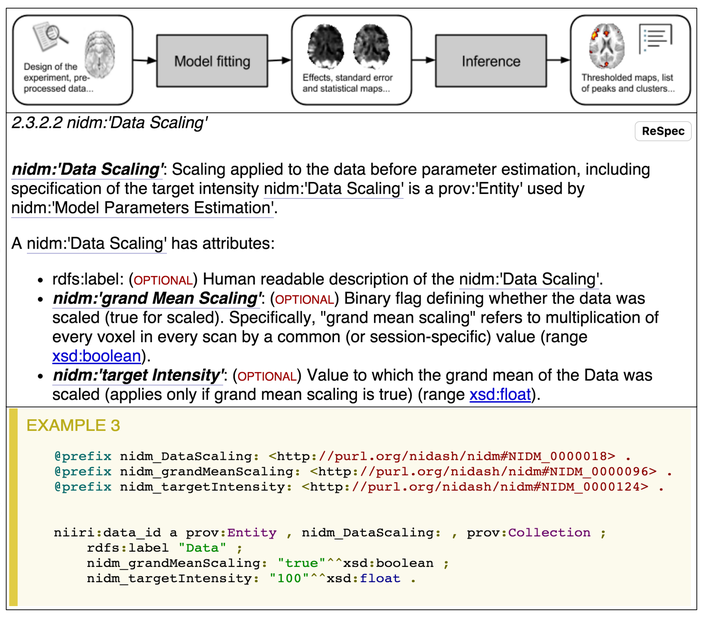

Figure 4. Example of the NIDM Results Specification

Formal Specifications

- Object models are organized around the identified use-cases
- A conceptual diagram of an object model is used to identify sub-components (Figure 4A)
- Figure 4B is a screenshot of a term from an automatically generated specification from an OWL file
- All terms are given a definition and properties in an OWL Ontology document developed using Protégé
- Examples for each term are provided using the human-readable Turtle syntax for RDF
- Prefixes are provided for all predicates for enhanced readability, for example, in SPARQL queries (Figure 4C)

Use Case: NCANDA Data Sharing and Apps

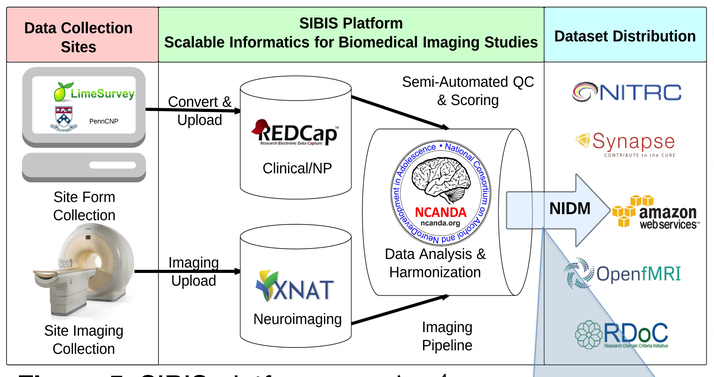

Figure 5. SIBIS platform overview[4]

Curating Data using NIDM

- A NIDM Experiment object model will be specified to capture both data and data dictionary content
- This mapping will "lift" cognitive and brain measures into RDF and enable semantic annotation and query
- Terms for Common Data Elements will be curated with URIs that dereference to online documentation
- A Web API will enable SPARQL-based views of NIDM for Apps

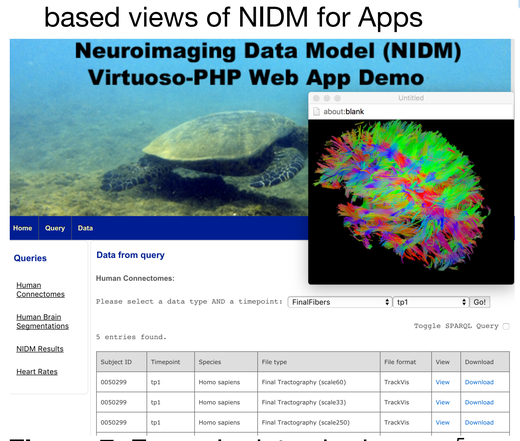

Figure 7. Example data sharing app[5]

Data Exploration Web Apps

- By linking brain models and statistical data, apps can be developed for web based exploration
- Figure 8 displays a demo app for visualizing shape distributions when a given brain structure is clicked on

Data Integration Platform

- Each NCANDA site transfers clinical and imaging data to a central repository by accessing a REDCap and XNAT server.
- We developed SIBIS, Scalable Informatics for Biomedical Imaging Studies, to integrate data streams and performs semi-automatic quality assessments with issue tracking (Fig 5)
- Data will be curated using NIDM for distribution to data sharing repositories, such as Sage Bionetwork's Synapse

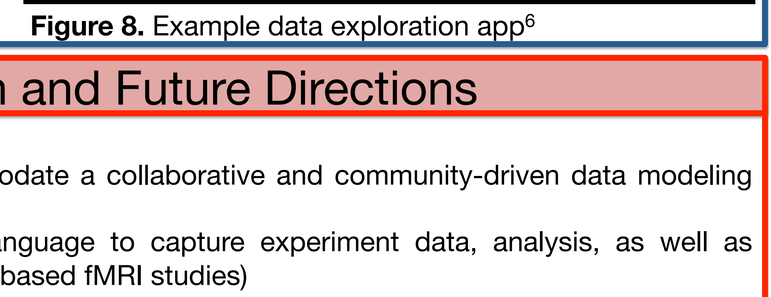

Figure 6. Lifting NCANDA into RDF with annotations

Data Sharing Web Apps

- Using NIDM, data will be "self describing" and will enable the development of Linked Data web applications to facilitate sharing
- Curated data will be annotated data to enable browsing by concepts rather than unrelated variables in tables
- Figure 7 displays a demo app built on top of example NIDM object models for brain data with browser-based 3D visualization and semantic tag-based filtering

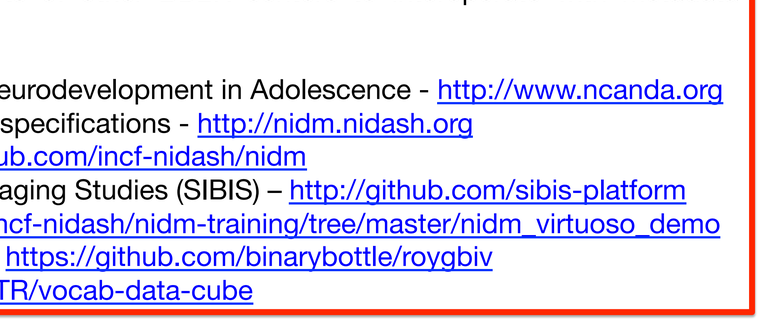

Figure 8. Example data exploration app[6]

Discussion and Future Directions

Current Thinking

- NIDM is flexible enough to accommodate a collaborative and community-driven data modeling approach
- Provenance is a general enough language to capture experiment data, analysis, as well as neuroimaging results (e.g., from task-based fMRI studies)

Technical Challenges

- Creating models is time intensive and challenging to scale across measures and datatypes
- Many resources may contain PHI (e.g., 3D anatomical images), leading to challenges for authentication across distributed sources

Social Challenges

- Other data sharing standards and ontologies exist without a neuroimaging focus, necessitating cross communication and feedback early in the process to avoid "reinventing the wheel"

Open Questions

- Can RDF Data Cubes[7] be used as an extension of NIDM and PROV for modeling tabular (i.e., csv) datasets, yet maintaining a conceptual/dimension-based index of experimental data?
- How can we best leverage the efforts of other BD2K centers to interoperate with metadata standards and web APIs?

References

1. National Consortium on Alcohol and Neurodevelopment in Adolescence - http://www.ncanda.org
2. Neuroimaging Data Model (NIDM) and specifications - http://nidm.nidash.org
3. NIDM GitHub Repository – https://github.com/incf-nidash/nidm
4. Scalable Informatics for Biomedical Imaging Studies (SIBIS) – http://github.com/sibis-platform
5. NIDM Demo VM - https://github.com/incf-nidash/nidm-training/tree/master/nidm_virtuoso_demo
6. ROYGBIV Shape Visualization Toolkit - https://github.com/binarybottle/roygbiv
7. RDF Data Cubes - http://www.w3.org/TR/vocab-data-cube