Brain imaging datasharing

The development of NIDM

The missing principles and tools

Jean-Baptiste Poline

UC Berkeley Brain Imaging Center

INCF Nidash TF
Outline

• Background / Introduction: Software development and sociological aspects
• Why NIDM? How is it developed?
• Current use cases
  • Experimental data
  • Freesurfer
  • FMRI statistical results (SPM, FSL)
• Future of NIDM
  • Link with other projects
  • The big picture: standardizing the meta data
Software development in neuroimaging: background

- Groups want an easy media for dissemination of their methods (SPM, FSL, Afni, etc)
- Code is often open, but often a “lab” enterprise
- Generally, software is still poorly considered in research (e.g. numpy author)
- Poor interoperability between tools: No W3C for neuroimaging. Nifti experience.
- Publications (the currency) are not adapted

No provenance of results and data, little testing
Sociological consequences

- Groups have to “promote” their software
  - A medium for competitive advantage
  - Software suite silos

- Code to reproduce paper is not available and is not reviewed – grants panels do not often care
- Poor code development standard & training of scientists

- Out of NIH 400M USD for fMRI – how much for software development?
- How much of this across institutions/countries?
Consequences for the scientific community

- Lack of harmonization and standard development leads to **inefficient** research (eg Nifti standard)
- Lack of **re-usability**: A big **waste of resources**
- **Lack of detailed provenance** information
- Lack of good testing and development standard: **reproducibility issue** (Donoho/Claerbout)

We need to respond to the reproducibility crisis
Software: The missing principles

- Publicly funded research should lead to public research products. Not only help me get the next grant.
- Research is about advancing knowledge: Software should help to:
  - Reproduce and validate - Tested
  - Provide with provenance
  - Help us be efficient (limited resources)
- Adopting a culture of software and data sharing: long term versus short term
  - Re-use rather than re-develop
  - Collaborate when ever it is possible
NIDASH-Neuroimaging Data Model

Goals

- Comprehensive data sharing
- Enhance reproducibility
- Enhance re-usability
- Increase interoperability and efficiency
  - Discover data
  - Access and use data
- Enable new research and idea

Challenges

- No easy tools
- Missing metadata
- Undiscoverable
- Multiple software
- Limited provenance
- Very limited funding
- No common standards

We need a common language, and a method to construct it
NIDM: What are we trying to do?
How do we avoid this?

**How Standards Proliferate:**

(See: A/C chargers, character encodings, instant messaging, etc.)

**Situation:**
There are 14 competing standards.

**14?! Ridiculous! We need to develop one universal standard that covers everyone's use cases.**

**Soon:**

**Situation:**
There are 15 competing standards.
NIDM: solving for both technical issues and social engineering

- NIDM: A data model shared and co-developed
- NIDM development: Nidash methodology
  - Weekly call and crowd sourced minutes (google documents)
  - Git / Github for the development, BSD license
  - Pull requests / issues for discussion on terms
  - Hackathon
- Tools developers on board
- Technical: semantic web solution
Community development – get the right people on board
So, what is NIDM based on?

NIDM Component Layer Cake

Level 4
NIDM Dataset Descriptor

Level 3
NIDM Experiment
NIDM Workflow
NIDM Results

Data Acquisition
Image Processing
Statistical Model

Level 2
NIDM Core Vocabulary

Level 1
PROV Family of Specifications

Level 0
Semantic Web Technologies

PROV Model

http://www.w3.org/TR/prov-primer/
## NIDM-Experiment

### NIDM Experiment Elements

<table>
<thead>
<tr>
<th>Project</th>
<th>Study</th>
<th>Acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Object</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Project Agent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Project Process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Study Object</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Study Agent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Study Process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Acquisition Object</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Acquisition Agent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Acquisition Process</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PROV Type Key**
- Agent
- Activity
- Entity
- Attributes

**Credits:** Nolan Nichols et al.

How do we involve larger community

**NEUROIMAGING DATA MODEL**

**NIDM SPECIFICATIONS**

- NIDM-Overview: a listing and description of elements.
- NIDM-Primer: a description of the overall framework.
- NIDM-Experiment: a specification for how to represent neuroimaging studies and acquisitions.
- NIDM-Results: a specification for how to represent neuroimaging analysis.
Use Case 1: Conte DB

Heart rate description

UCI Conte Center on Brain Programming in Adolescent Vulnerabilities (contecenter.uci.edu)
Dave Keator

Comprehensive and long lasting markup of a complex dataset
FreeSurfer Domain Object

David Keator et al., Neuroimage, 2013
Use case 2: FreeSurfer

Example of converting single files

```python
In [5]:
g = prov.ProvBundle()
e1 = create_entity(g, "bert", os.path.join(sdir, "mri/T1.mgz"))
e2 = create_entity(g, "bert", os.path.join(sdir, "label/lh.BA6.label"))
fsdir_collection = g.collection(niiri['foo'])
g.hadMember(fsdir_collection, e1)
print g.rdf().serialize(format='turtle')
```

- Link to FMA
- Link to other tools
- Link to other data
- Github incf-NIDASH

_Credit: Satra Ghosh_
Use Case 3: fMRI statistical results
Usual (x,y,z) meta analyses

neurosynth.org

Neurosynth is a platform for large-scale, automated synthesis of functional magnetic resonance imaging (fMRI) data.

It takes thousands of published articles reporting the results of fMRI studies, chews on them for a bit, and then spits out images that look like this:

An automated meta-analysis of 420 studies of pain

Database Status
413429 activations reported in 11406 studies
Interactive, downloadable meta-analyses of 3107 terms

Tal Yarkoni
Meta analyses - SPM+FSL(+Afni)

Smoothness, Error model, Contrast direction, ... marked up
Tags:
* Cog Atlas
* CogPo
* Demogr.
* ROI

368 articles corresponding to the search "pain"
Storing contrast maps and meta data

NeuroVault
A public repository of unthresholded brain activation maps

What is it?
A place where researchers can publicly store and share unthresholded statistical maps produced by MRI and PET studies.

Why use it?
- Interactive visualization
- A permanent URL
- Publicly shareable
- Improves meta-analyses

Supported by

Chris Gorgolewski
Queries: where power will be demonstrated

- Standard SPARQL language
- Set of resources: local and/or distributed
- Create a common graph across resources

- Size of graphs?
- Efficiency?
- Tools: Virtuoso – 4store - Dydra - Talis:
  - Manage increasing large number of triples
- Training + Tools for neuroscientists
Queries

In [12]: qres = gmap.query("""SELECT DISTINCT ?subclasses
    WHERE {
        ?subclasses rdfs:subClassOf nidm:FSIQ.
    }""")
print qres.serialize(format='csv')

subclasses
http://fcon_1000.projects.nitrc.org/indi/abide/ABIDE_FIQ
http://www.birncommunity.org/collaborators/function-birn/FSIQ

Add the Mappings to the Joined Datasets

In [15]: mapped_graph = union_graph + gmap

Now use the mappings to filter the results!

In [17]: qres = mapped_graph.query("""SELECT DISTINCT ?s ?subclasses ?o
    WHERE {
        ?subclasses rdfs:subClassOf nidm:FSIQ.
        ?s ?subclasses ?o
    }""")
print qres.serialize(format='csv')

s, subclasses, o
SELECT DISTINCT *
WHERE
{ ?peak a nidm:Peak .
  ?cluster a nidm:Cluster .
  ?peak prov:wasDerivedFrom ?cluster .
  ?peak prov:atLocation ?coordinate .
  ?coordinate nidm:coordinate1 ?x .
  ?coordinate nidm:coordinate3 ?z .
  OPTIONAL { ?peak prov:value ?value }.
  OPTIONAL { ?peak nidm:pValueFWER ?pvalcor }.
  ?cluster
  prov:wasDerivedFrom/prov:wasGeneratedBy/prov:used ?statmap .
  ?statmap a nidm:StatisticMap .
  ?statmap nidm:statisticType ?stat .
}
ORDER BY ?cluster ?peak

['http://iri.nidash.org/cluster_0001', '4.126074e-10',
 'http://iri.nidash.org/peak_0001',
 'http://iri.nidash.org/coordinate_0001', '-48.1', '-9.24', '-73.7',
 'http://www.incf.org/ns/nidash/nidm#ZStatistic',
 'http://iri.nidash.org/z_statistic_map_id', 'None',
 '6.14', 'None']

['http://iri.nidash.org/cluster_0001', '7.705712e-10',
 'http://iri.nidash.org/peak_0002',
 'http://iri.nidash.org/coordinate_0002', '-38.1', '-18.0', '-53.4',
 'http://www.incf.org/ns/nidash/nidm#ZStatistic',
 'http://iri.nidash.org/z_statistic_map_id', 'None',
 '6.14', 'None']

['http://iri.nidash.org/cluster_0001', '4.462172e-09',
 'http://iri.nidash.org/peak_0003',
 'http://iri.nidash.org/coordinate_0003', '-29.6', '-16.9', '-73.8',
 'http://www.incf.org/ns/nidash/nidm#ZStatistic',
 'http://iri.nidash.org/z_statistic_map_id', 'None',
 '5.75', 'None']
NIDM-results viewer
NIDM-Results\(^2\)

- FS converter
- FSL converter
- Native SPM exporter
- AFNI engaged in the project

- OpenFMRI data nidm results for both SPM and FSL
- Neurovault ingests nidm results
- BIDS to NIDM soon

- Nidm used in Conte, NCANDA, NKI soon & other projects

Four grants have been submitted to support some aspects of NIDM
How do we understand each others?

- We need a common goal
- We need a common grammar (NIDM)
- We need a common vocabulary
- We need a way to re-use “terms” and augment the vocabulary

→ **NIDM: A workflow to curate terms openly**
Finding terms: where should I look

Cognitive Atlas

DCT OBO RDFS HCSI NCIT STATO NIF NIDM

The OBO Foundry

Welcome to the Foundational Model of Anatomy

Project of the Structural Informatics Group at the University of Washington Department of Biological Structure and Biomedical and Health Informatics, Department of Medical Education and Biomedical Informatics.

Checking the GitHub project for more info.

<table>
<thead>
<tr>
<th>CHEBI</th>
<th>Chemical entities of biological interest</th>
<th>A structured classification of biological relevance. <a href="chebi">Detail</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>zfa</td>
<td>Zebras and development</td>
<td>A structured controlled vocabulary for the development of the Zebrafish. <a href="zfa">Detail</a></td>
</tr>
<tr>
<td>xao</td>
<td>Xenopus anatomy and development</td>
<td>Anatomy and development of Xenopus (Xenopus laevis). <a href="xao">Detail</a></td>
</tr>
<tr>
<td>pr</td>
<td>Protein Ontology</td>
<td>an ontological representation. <a href="pr">Detail</a></td>
</tr>
</tbody>
</table>

Why this is –may be– working?

• Spirit based on open source development movement has proven to be efficient – get inspired by Open Science Foundation type of initiative
• Based on a strong neuroimaging and python open source ecosystem:
  • Nipy, Nipype, NiTime, Dipy, Nilearn, MNE, etc
• Because we are trying to solve a problem in a domain we know well (500 years of experience – and yet some fresh brains)
• Funding: opportunity with NIDASH / INCF

Because of the individuals in the group
What are the difficulties?

- Coordination: short vs long terms goals
- Software development requires sustained resources
- Small developer community
- Software development: PhDs and grants not common – publications focus - but see Giga science
- Most researchers have domain knowledge, but not the technical skills
- Steep learning curve to understand all technologies involved: need for training
Current/Future work

- NIDM experiment:
  - Many more datasets NKI, NIDM-BIDS
  - Link with NDAR
- LORIS, XNAT, COINS, HID, etc: engage to read and export NIDM-experimental – Common API
- NIDM-Workflow:
  - Integrate Nipype, C-Pack, C-BRAIN, AA, etc, with a common provenance model
  - A common language for processing
- NIDM-Results:
  - Other software, methods (resting state networks....)
Acknowledgements

TO MATHEW, ROSA, EVA-CHRISTINE, LINDA
And all at the INCF Secretariat