Requirements for Efficient Data Management in Neurophysiology

Thomas Wachtler

German Neuroinformatics Node
Department Biology II
Ludwig-Maximilians-Universität München
INCF Program on Standards for Datasharing

**Mission:** to develop generic standards and tools to facilitate the recording, sharing, and reporting of metadata; and thereby facilitate the archiving and sharing of neuroscience data.

**Focus areas:** electrophysiology and brain imaging data

Washington Irving, in the public domain (Wikipedia)

www.incf.org/about/programs/datasharing
G-Node: German National Node of the INCF

Bernstein Network

INCF National Nodes

www.nncn.de

www.incf.org
INCF German Node (G-Node):
Focus on Neuroinformatics Solutions for Electrophysiology

Development of tools and services for cellular and systems electrophysiology, facilitating data access, data analysis and data sharing

- Data conversion tools
- Methods for metadata management
- Data sharing platform
- Custom solutions for collaborative data exchange
- Teaching and training

www.g-node.org
Levels of Data Sharing

- **Share with yourself (and your colleagues/students/supervisors)**
  - data management within a lab
  - all data that is recorded
  - keep all information, document 'hidden' knowledge to enable future access for re-analysis

- **Share with collaborator**
  - specific datasets
  - specific purpose, specific set of metadata required
  - interaction between owner and collaborating partner

- **Share with the world**
  - typically after data have served their primary purpose
  - data underlying a publication
  - landmark datasets, benchmark datasets
  - should be readable and understandable without interaction with the author
Metadata

Information about the conditions under which data have been acquired
Essential for meaningful analysis and sharing of the data

Levels of (Meta)Data

<table>
<thead>
<tr>
<th>Proper Metadata</th>
<th>&quot;Soft&quot; Metadata</th>
<th>&quot;Hard&quot; Metadata</th>
<th>Data Complement (units, ...)</th>
<th>Format Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(creator, dataset identifier, ...)</td>
<td>(purpose of study, scientific context, ...)</td>
<td>(experimental parameters, stimulus description, ...)</td>
<td>Recorded Data</td>
<td></td>
</tr>
</tbody>
</table>

Sobolev et al 2014
Challenges in Neurophysiology

• Heterogeneity of data due to variety of
  - Electrode configuration (single electrode, multiple single electrodes, tetrodes, electrode arrays, ...)
  - Types of Signals (intracellular, field potentials, single unit spikes, multi unit activity ...
  - Species (invertebrate, non-mammalian, rodent, non-human primate, humans, ...)
  - Preparation (culture, slice, in vivo, ...)
  - Stimuli / Experimental Paradigm (current injection, pharmacology, presynaptic stimulation, sensory stimulation all modalities, behavioral task, ...)
  - Data file formats (large number of proprietary formats, in-lab custom formats)
  - ...

• Currently no standards exist for data access, data annotation, stimulus description, etc.
Finding the balance between flexibility and standardization

One possible approach:

• Defined **data model** for electrophysiological data that accounts for all types of recorded data
• Flexible methods for **data annotation** and metadata management that can be adapted to the requirements of the experiment and laboratory
NEO - Data Model for Neurophysiology

http://packages.python.org/neo

- Common class names and concepts for electrophysiological (recorded or simulated) data
- Consistent data organization
- Easy to adopt
  I/O modules for various file formats are provided
  Used by several software packages (OpenElectrophy, G-Node tools, NeuroTools, SpykeViewer, ...)

odML - flexible, extensible Metadata format

http://www.g-node.org/odml

- separation of format and content
- format: hierarchical structure of key-value pairs: simple, flexible, inherently extensible → can be adapted to the specifics of the lab or experiment
- can carry any metadata → no information is lost
- facilitates automated collection of metadata in the laboratory
- community-driven standardization through shared terminologies (odML terminologies, ontology of experimental neurophysiology (OEN))

Requirements for efficient data management, analysis, and sharing

- Representation of recorded data by standardized data model
- Representation of metadata in flexible, extensible way
- Store any metadata necessary together with the data
- Integrated organization of data and metadata that reflects the structure of the experiment
- Fine grained annotation of data items
- Ability to select data by metadata