Keeping track of your data with tools for comprehensive data organization

C. J. Kellner[1], A. Stoewer[1], M. Sonntag[1], A. Koutsou[1], A. Sobolev[1], J. Benda[2], *T. Wachtler[1], J. Grewe[2]

[1]Ludwig-Maximilians-Univ Munich, 82152 Planegg-Martinsried, Germany; [2]Eberhard Karls Univ., Tuebingen, Germany

Introduction

Data management challenges

Increasing complexity of experimental approaches in the neurosciences challenges efficient management of recorded data and metadata. Storing such information consistently is an essential part of experimental research and essential for:

- efficient data anlysis
- re-analysis or re-use of data
- reproducibility
- data sharing

File Formats

Consistent data organization depends crucially on available formats. However, currently existing formats have shortcomings:

- Proprietary or poorly documented
- Constraints on what can be stored
- Limited ability to store metadata
- Limited provision of software and tools to work with the data

Approach

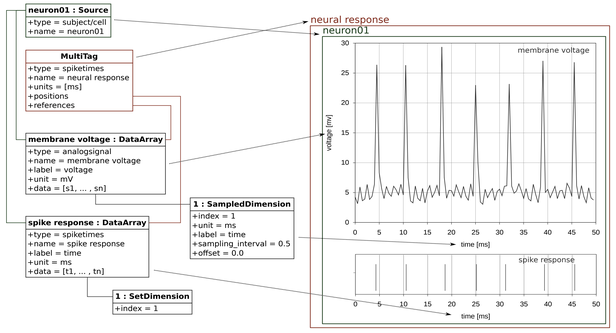
To lower the technical barriers of data management and facilitate comprehensive data organization in the laboratory, we present **NIX**, a versatile, open format for neuroscientific data.

NIX Design goals

- To store a wide variety of data types
- To store metadata with the data
- Intuitive, coherent file structure
- Easy access and integration in data analysis tools
- Support for common platforms

NIX Data Model

Data entities to represent (neuro)scientific data



The model provides all information to interpret the data correctly

Main Entities:
- **Array**: stores n-dimensional data with information about data type and units, defines dimensions using **Dimension** entity
- **Tag**: Defines points or regions, representing segments, spike times, events, and relationships between data

All entities have:
- a unique **id**: allows synchronization and identification accross files.
- a **name**: serves as a human readable identifier.
- a **type**: provides semantic context, domain-specificity.

Flexible approach to store any kind of metadata

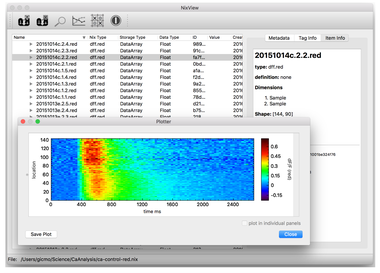
- Metadata is stored, using the odML [1] approach, as hierarchically organized structure of key-value pairs
- Any metadata can be stored, according to the specifics of the experiment or dataset.
- Metadata is linked to the data, enabling selection of data based on metadata.

Libraries and language bindings

The schema definition for HDF5 [2] represents all entities of the data model hierarchy. Easy reading and writing of the NIX file format, is provided by an **IO-library** in **C++** [3], supporting major compilers and operating systems such as **Linux**, **macOS** and **Windows**, and language bindings for **Python** [4], **Matlab** [5] and **Java** [6].
The Python package [4] additionally provides quick and easy installation without requiring the C++ library.

C++ python MATLAB Java

Access tool NixView



- NixView [7] provides convenient exploration of both data and metadata of NIX files.
- Raw data can be browsed via tabular display and easily exported to CSV.
- Facilitates plotting support for a large variety of raw data as well as the export of plots.

NIX backend for Neo

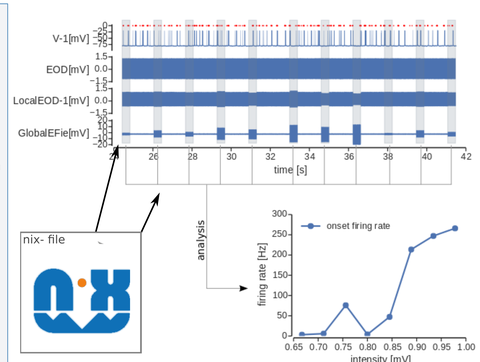
An I/O backend for Neo [8] maps the Neo data model to the NIX format:

- Data stored in any format supported by Neo can be converted to NIX.
- Enables easy storage of data analysis done with Neo compatible tools, e.g., the Elephant [9] toolkit.
- Converting data to NIX, or using NIX as a backend for Neo, enables the use of NIX as well as HDF5 tools and viewers:
  - NIX language bindings for C++, Python, Matlab, Java.
  - NixView.
  - HDFView, h5dump, h5ls, etc.

neo elephant ELECTROPHYSIOLOGY ANALYSIS TOOLKIT

Electrophysiological Data example

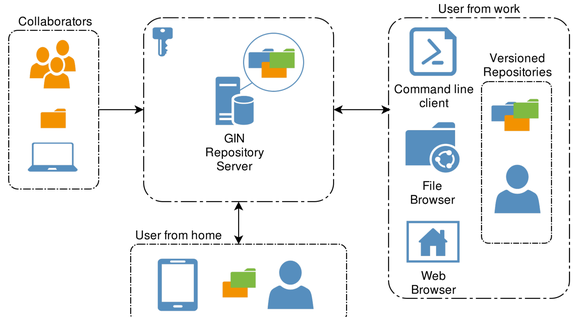
- In vivo electrophysiology in the weakly electric fish.
- 4 regularly sampled traces are recorded in parallel.
- Events in the mem-brane voltage and the EOD (top and 2nd trace, action potentials and electric organ discharges, resp.) are detected and stored in event traces.
- MultiTags are used to establish a link between recorded trace and the event. Event times point into the recorded traces.



During analysis, all data related to the same stimulus can be identified and directly retrieved via the NIX libraries.
The attached metadata is then used to create, e.g., the FI curve.

GIN Services for distributed data organization

Comprehensive data management does not stop at the file level. The **G**-Node **In**frastructure services [10] provide versioning, management and access of data repositories from multiple workplaces.



- Versioning of entire data repositories through git [11].
- Distributed data management via file and web browser or command line client.
- Secure access and shared repositories through OAuth integration.
- GIN services will provide augmented search and indexing support for NIX files.
- Free hosting of scientific data at the G-Node.

Summary

- **NIX provides a general data model for neuroscientific and other types of scientific data**
- **Enables storing all necessary information to interpret the data**
- **Relationships between data are stored explicitly.**
- **Full metadata integration (odML) enables comprehensive data organization and selection of data by metadata.**
- **HDF5 file structure reflects structure of data, easy to understand**
- **Supports other backends besides HDF5**
- **Libraries for many platforms and languages, easy to use and integrate in scientific computing environment**

Resources

Contact: nix@g-node.org

[1] Grewe et al (2011), Frontiers in Neuroinformatics 5:16
[2] https://github.com/G-Node/nix/wiki/Implementation-in-HDF5
[3] https://github.com/G-Node/nix
[4] https://github.com/G-Node/nixpy
[5] https://github.com/G-Node/nix-mx
[6] https://github.com/G-Node/nix-java
[7] http://bendalab.github.io/NixView
[8] http://neuralensemble.org/neo
[9] http://neuralensemble.org/elephant
[10] https://gin.g-node.org
[11] https://git-scm.com

Supported by BMBF grant 01GQ1302

Federal Ministry of Education and Research

Special thanks to Christian Spalthoff for the NIX logo design.