Comprehensive data organization made easy: NIX Versatile format for data and metadata

Adrian Stoewer[1], Christian Kellner[1], Michael Sonntag[1], Andrey Sobolev[1], Jan Benda[2], Thomas Wachtler[1], Jan Grewe[1,2]

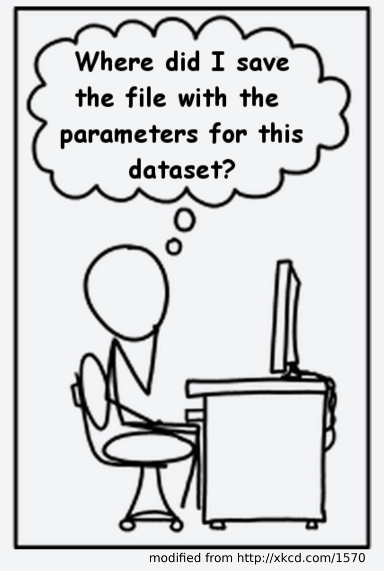
[1]German Neuroinformatics Node, Ludwig-Maximilians-Universität München, Germany; [2]Institut für Neurobiologie, Universität Tübingen, Germany

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 a fundamental part of experimental research and essential for:

- efficient data analysis
- 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

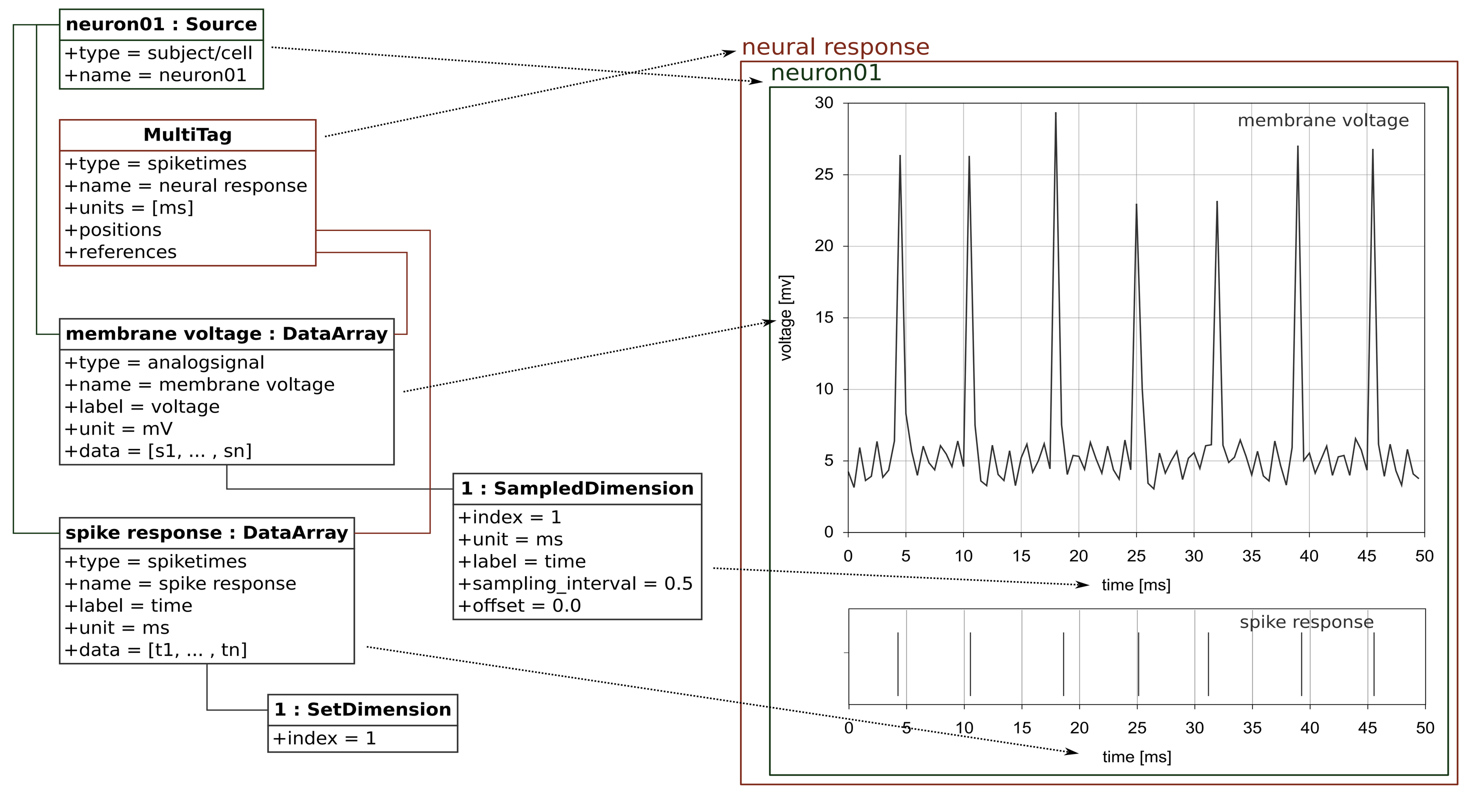
Approach

To lower the technical barriers of data management and facilitate comprehensive data organization in the laboratory, we present a versatile, open format for neuro-scientific data. Design goals:

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

Data Model

Data entities to represent (neuro)scientific data



The model provides all information to interpret the data correctly

Main Entities:

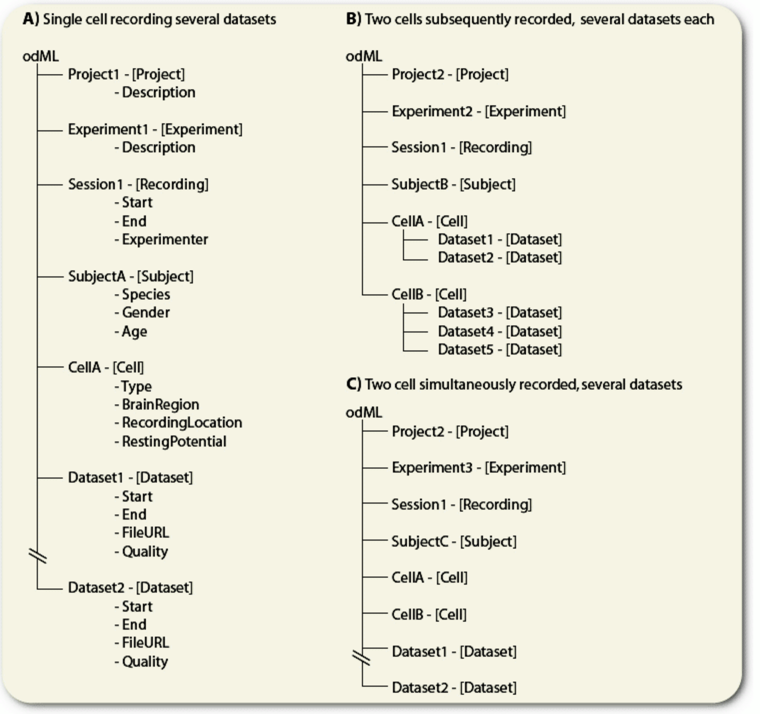
- **Array**: stores n-dimensional data with information about data type and units, defines dimensions using **Dimension** entity
- **Tag**: Defines points or regions, representing segments, spike times, events, and relationships between data

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.

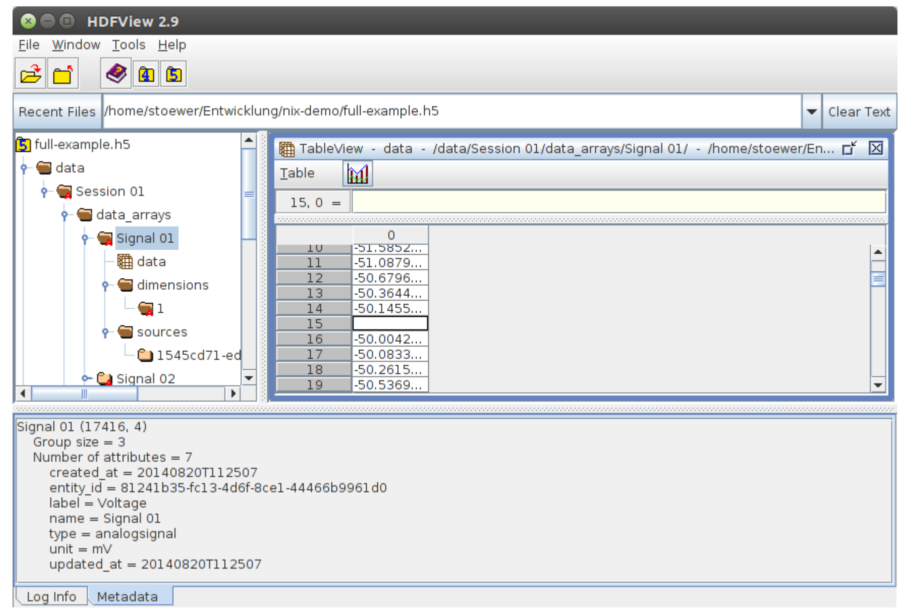
Flexible approach to store any kind of metadata

- 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.



Any kind of metadata can be stored and can be organized to reflect the structure of the experiment.

HDF5 file schema

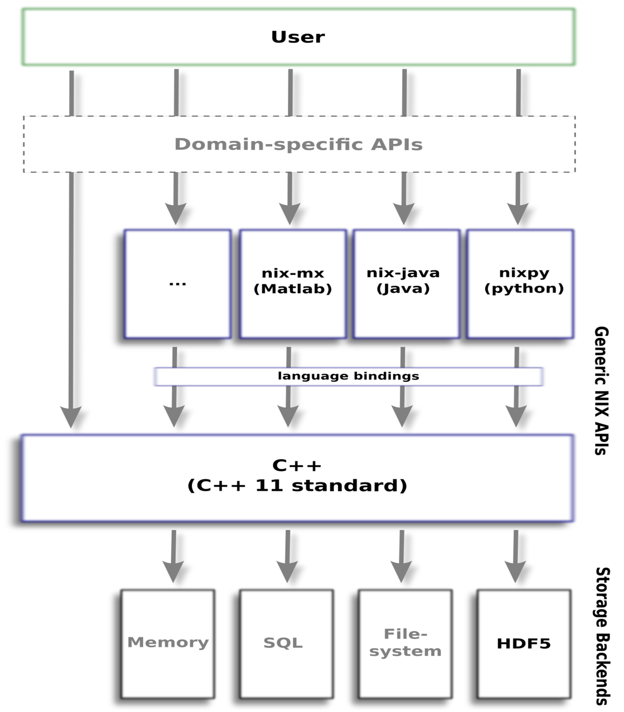


- The schema definition for **HDF5** [2] represents all entities of the data model in a flat hierarchy.
- It was designed to be **easily readable even without a special library**.

Libraries

Easy reading and writing of the NIX file format, even without deep knowledge about the exact format specification, is provided by an **IO-library** in **C++**[3], supporting major compilers and operating systems such as **Linux**, **OSX** and **Windows**, and language bindings for **Python** [4], **Matlab** [5] and **Java** [6].

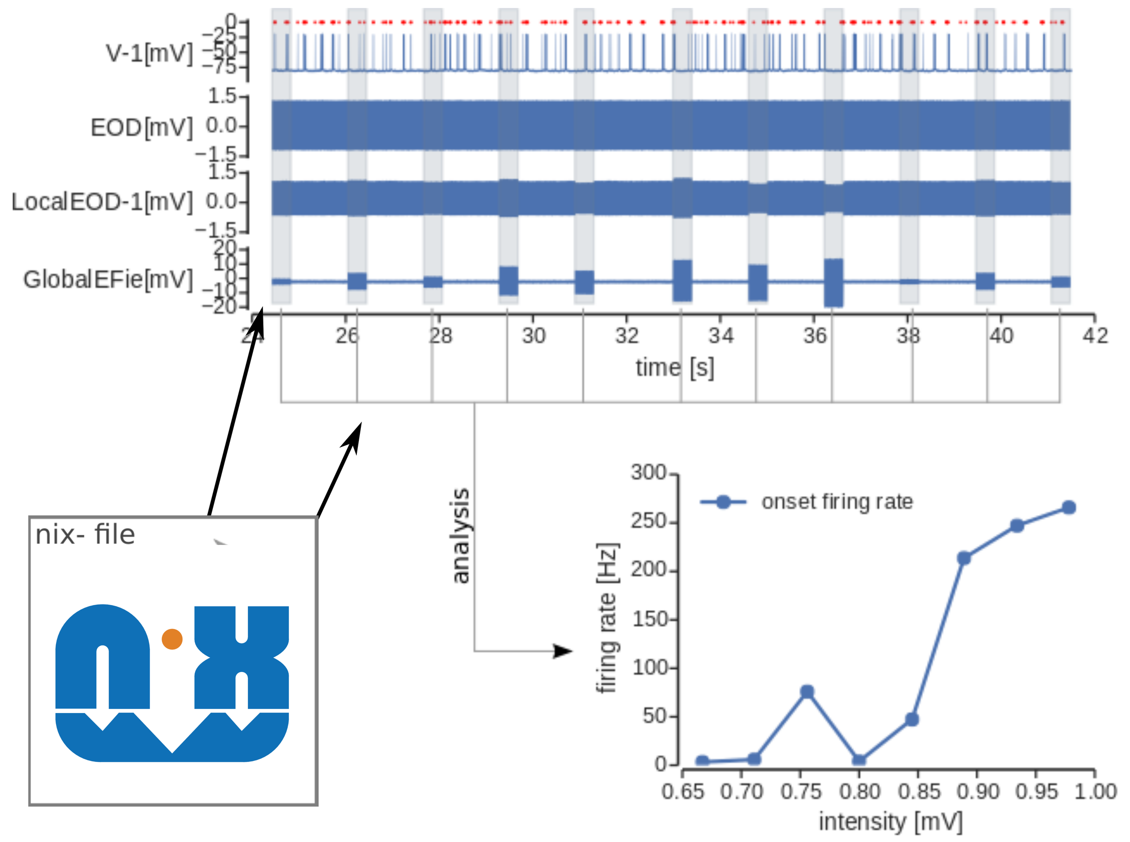
Modular architecture



Examples

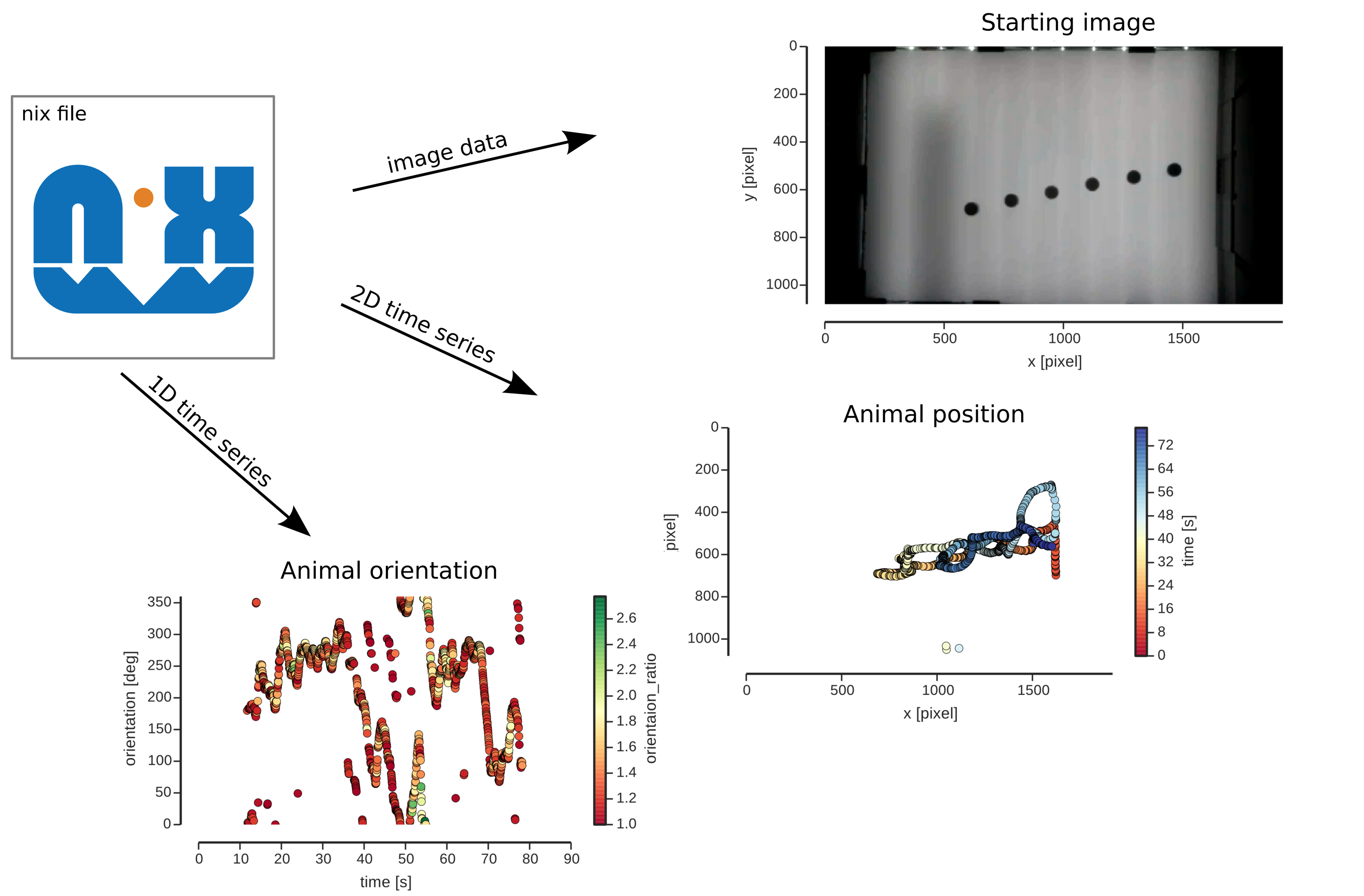
Electrophysiological Data

- 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 stame stimulus can be identified and directly retrieved via the NIX libraries.
The attached metadata is then used to create, e.g., the FI curve.

Behavioral Data



- In behavioral experiments the animal movement is video-tracked.
- Image data, animal positions and its orientation are stored in the NIX file.
- Information about the recording session, hardware and software settings are also stored and linked to the data

Summary

- **NIX provides a general format 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

This work was performed in connection with the activities of the HDF5 working group of the INCF Electrophysiology Data Sharing Task Force.
Special thanks to Christian Spalthoff for the NIX logo design.

Supported by BMBF grant 01GQ1302

Federal Ministry of Education and Research

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