

Keeping Track of Complex Data

Benefits of Comprehensive Data Management for Efficient Data Access, Reproducibility, and Data Sharing

Thomas Wachtler

German Neuroinformatics Node

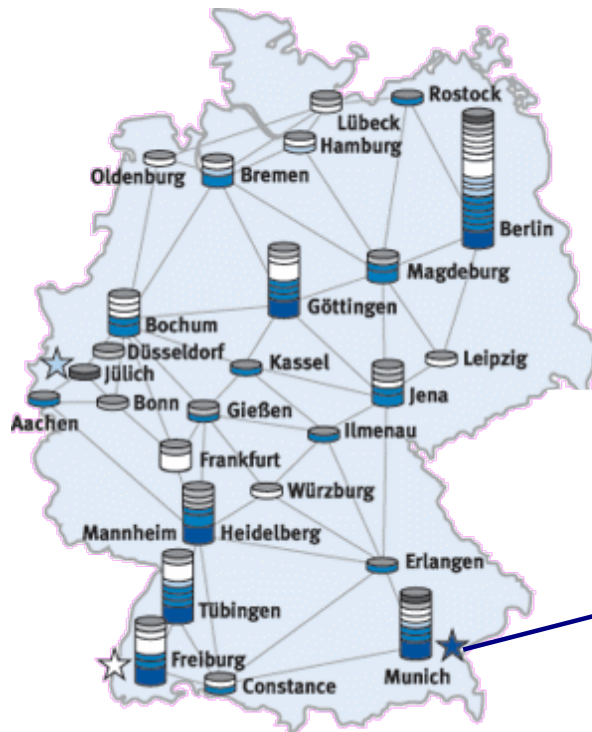
Department Biology II

Ludwig-Maximilians-Universität München

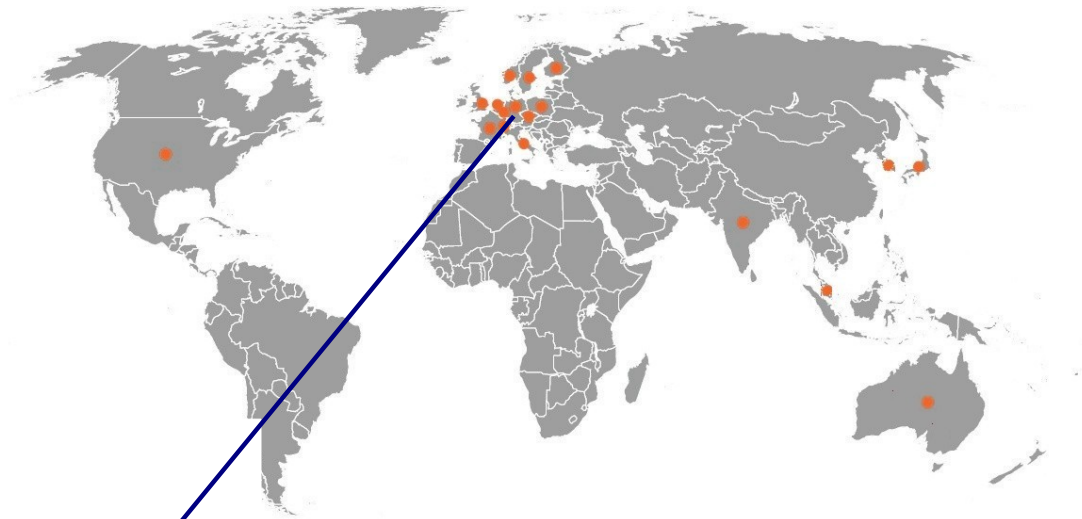


G-Node: German Neuroinformatics Node

Bernstein Network



INCF National Nodes



funded by



Federal Ministry
of Education
and Research

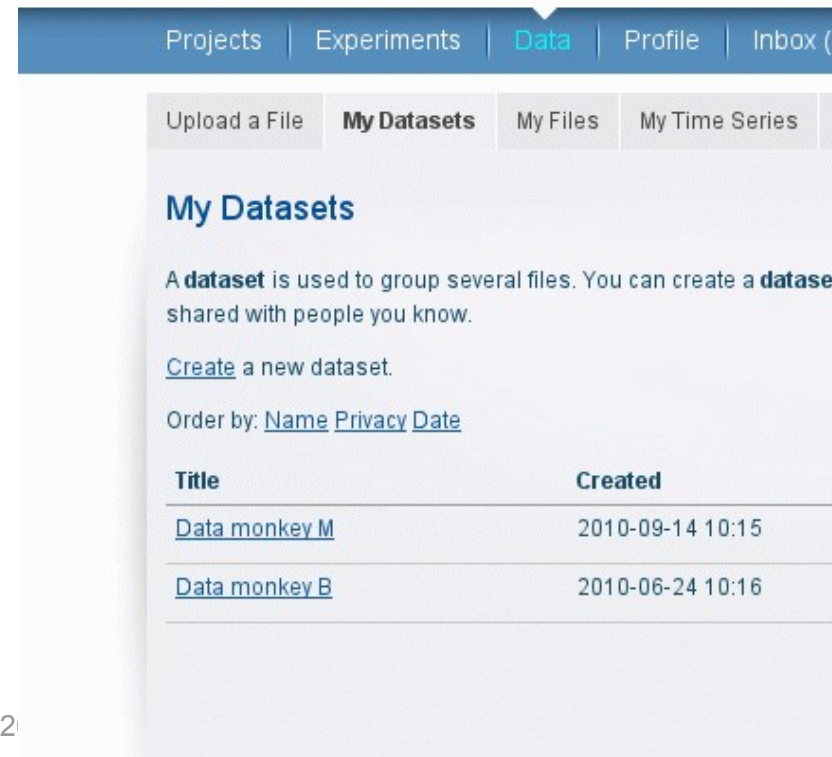
www.nncn.de

www.incf.org

German Neuroinformatics 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 data and metadata management**
- Data sharing platform
- Custom solutions for collaborative data exchange
- Hosting services
- Teaching and training



The screenshot shows a web interface for data management. At the top, there is a navigation bar with tabs for 'Projects', 'Experiments', 'Data', 'Profile', and 'Inbox'. Below this, there are tabs for 'Upload a File', 'My Datasets', 'My Files', and 'My Time Series'. The 'My Datasets' tab is active, showing a heading 'My Datasets' and a description: 'A dataset is used to group several files. You can create a dataset shared with people you know.' Below the description, there is a link 'Create a new dataset.' and a sorting option 'Order by: Name Privacy Date'. A table lists two datasets:

Title	Created
Data monkey M	2010-09-14 10:15
Data monkey B	2010-06-24 10:16

funded by

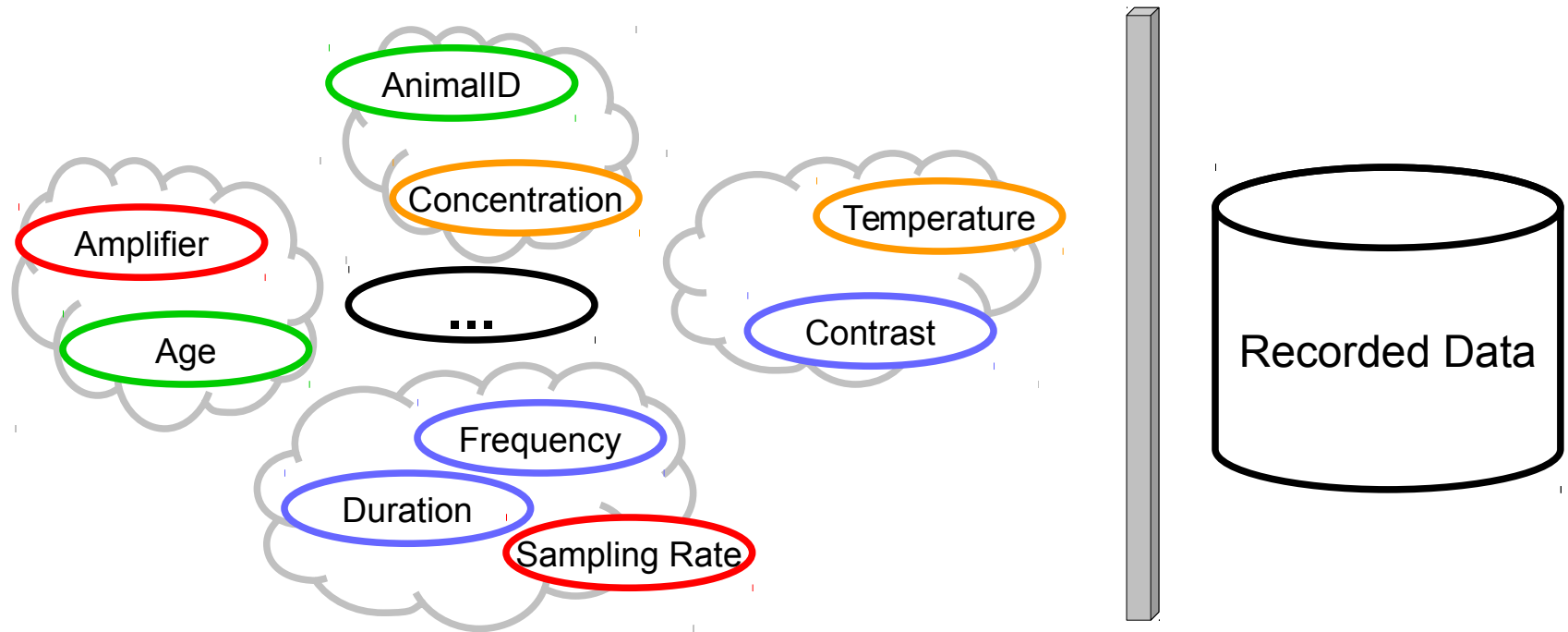
Why care about data management?

- Progress in neuroscience increasingly depends on collaborative efforts, exchange of data, re-analysis of data.
- Advances in technology and methodology dramatically increase volume and complexity of the data recorded.
- **Complexity and volume of data** pose a challenge for data organization. Collaborative work and re-use of data are hampered by the **effort** it takes to **access** and **understand** the data. Reducing this effort can enhance reproducibility and facilitate data sharing.

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
 - interaction between owner and collaborating partner
- **Share with the world**
 - often after data have served their primary purpose
 - might be re-used for different purpose
 - should be readable and understandable without interaction with the author

Getting (the data) ready for sharing ...

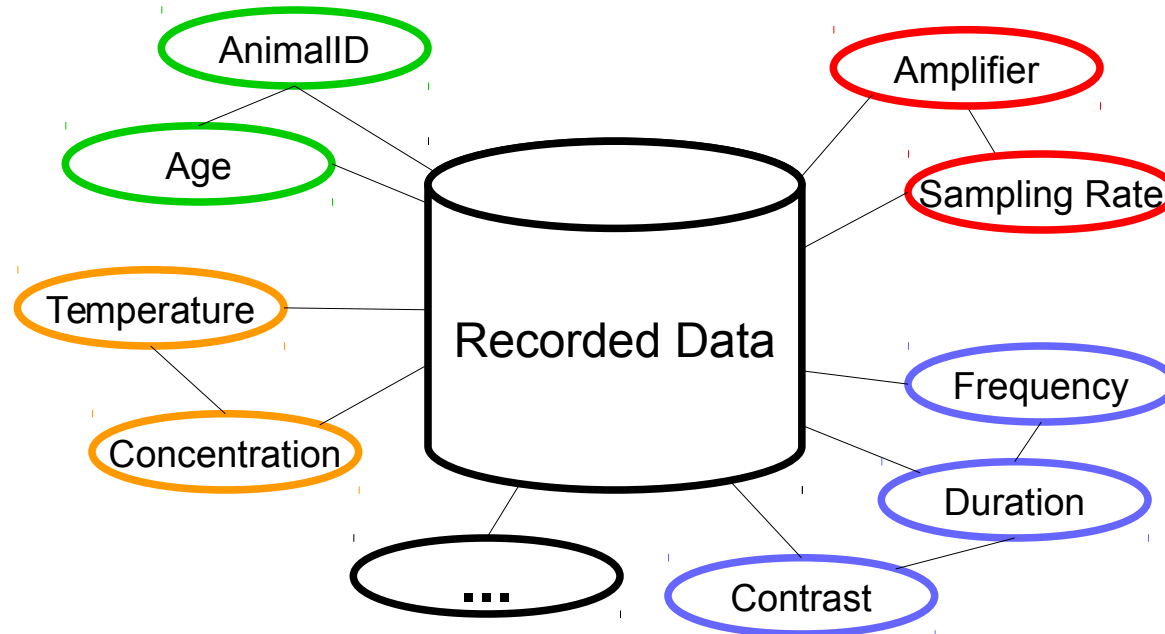


Metadata is often stored in **heterogeneous** formats, **distributed** over files, software code, file names, spreadsheets, handwritten ...

Separate organization and storage of data and metadata implies additional **effort** of identifying and selecting data

How to reduce this effort?

Getting (the data) ready for sharing ...



Integrated, standardized organization of data and metadata can reduce the overhead of searching for data and other necessary information for analysis.

This facilitates data analysis and re-analysis, reproducibility, data sharing.

Development of Tools for Efficient Data Management

Approach:

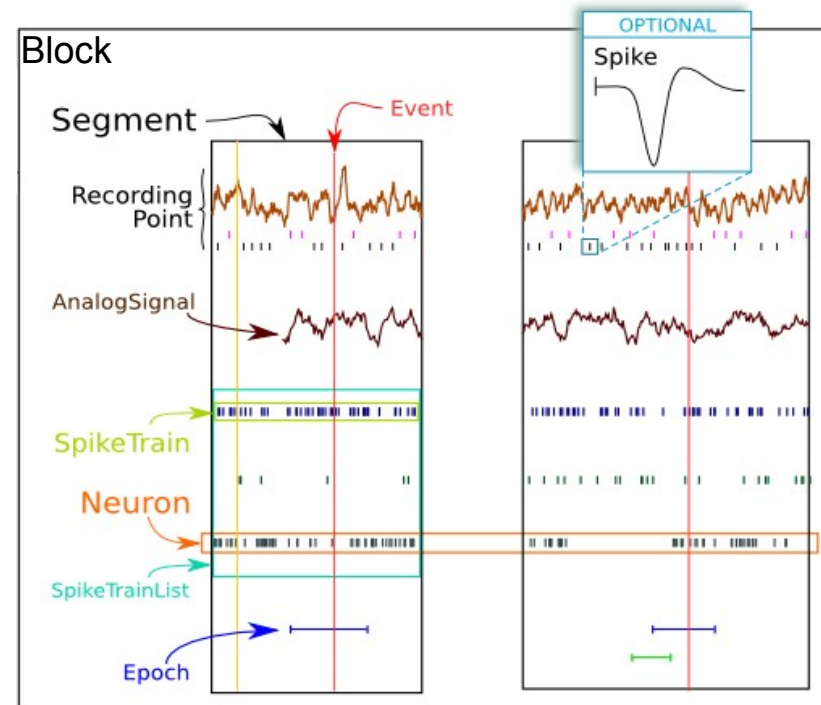
- Well-defined **data model** for neuroscience data that accounts for all types of recorded data
- Flexible methods for **data annotation** and metadata management that can be adapted to the requirements of the experiment and laboratory
- Format and tools for **integrated organization of data and metadata**, including interfaces for common tools and languages, to facilitate data access, data management, and data analysis

Neo - Data Model for Neurophysiology

<http://packages.python.org/neo>



- Common class names and concepts for electrophysiological 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, Elephant, ...)



Garcia et al (2014) Front. Neuroinf. 8:10

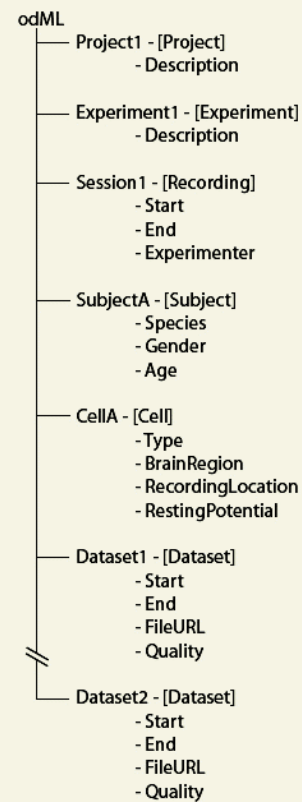
odML - flexible, extensible Metadata format

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

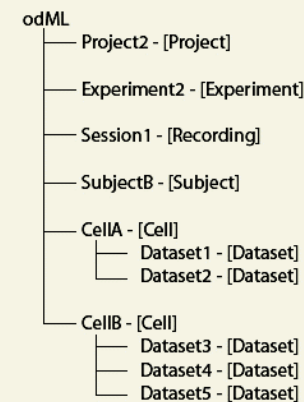


- separation of format and content
- format: hierarchical structure of key-value pairs: simple, flexible, inherently extensible → can be adapted to the specifics of the lab or experiment
- can carry **any metadata** → no information is lost
- machine readable, **facilitates automated collection** of metadata in the laboratory
- community-driven standardization through shared **terminologies**
- tools available (libraries, editor, apps)

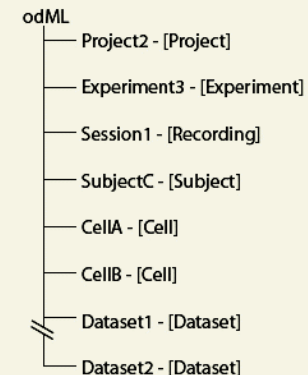
A) Single cell recording several datasets



B) Two cells subsequently recorded, several datasets each

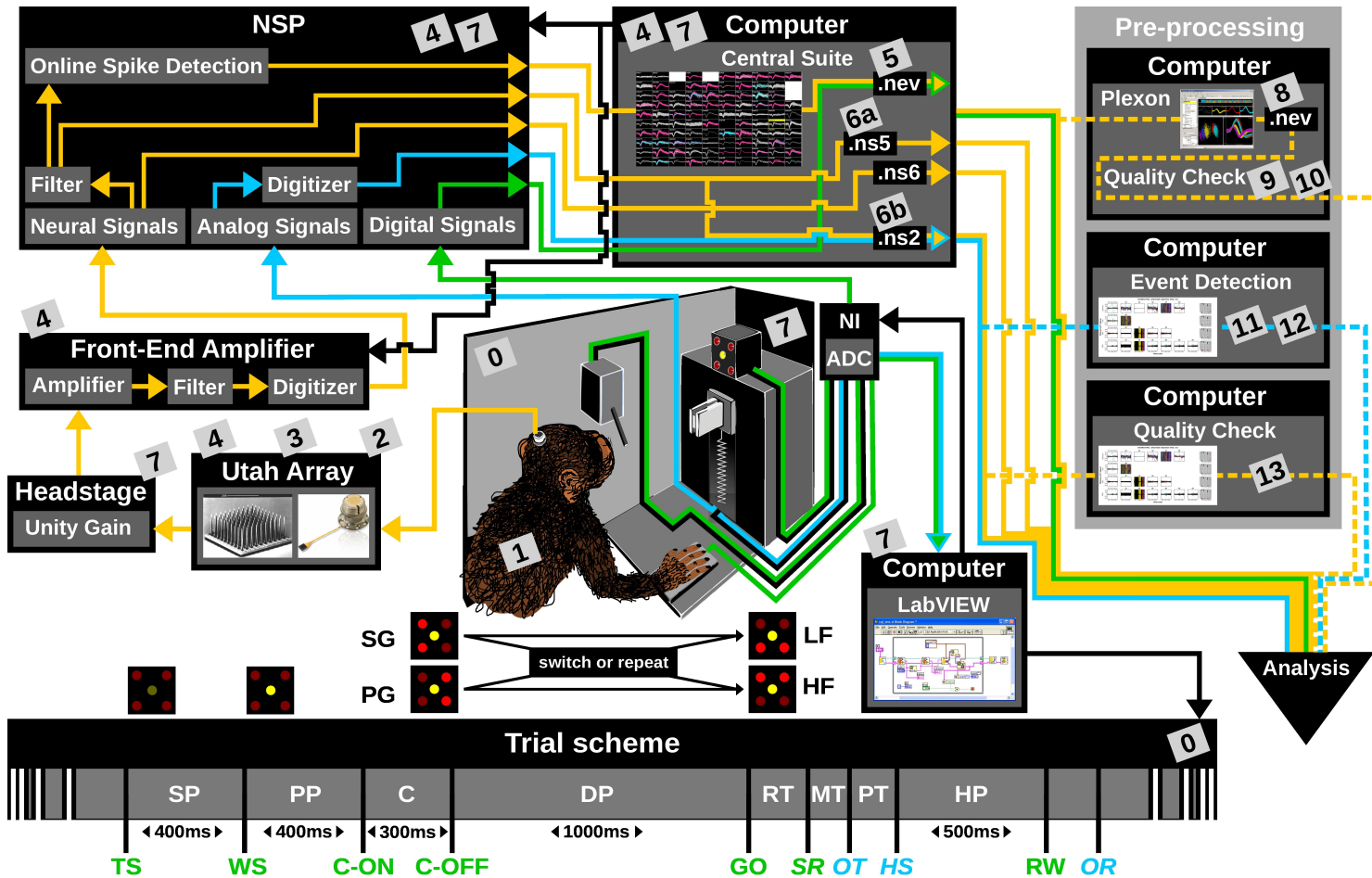


C) Two cell simultaneously recorded, several datasets





Example: collecting metadata from different sources

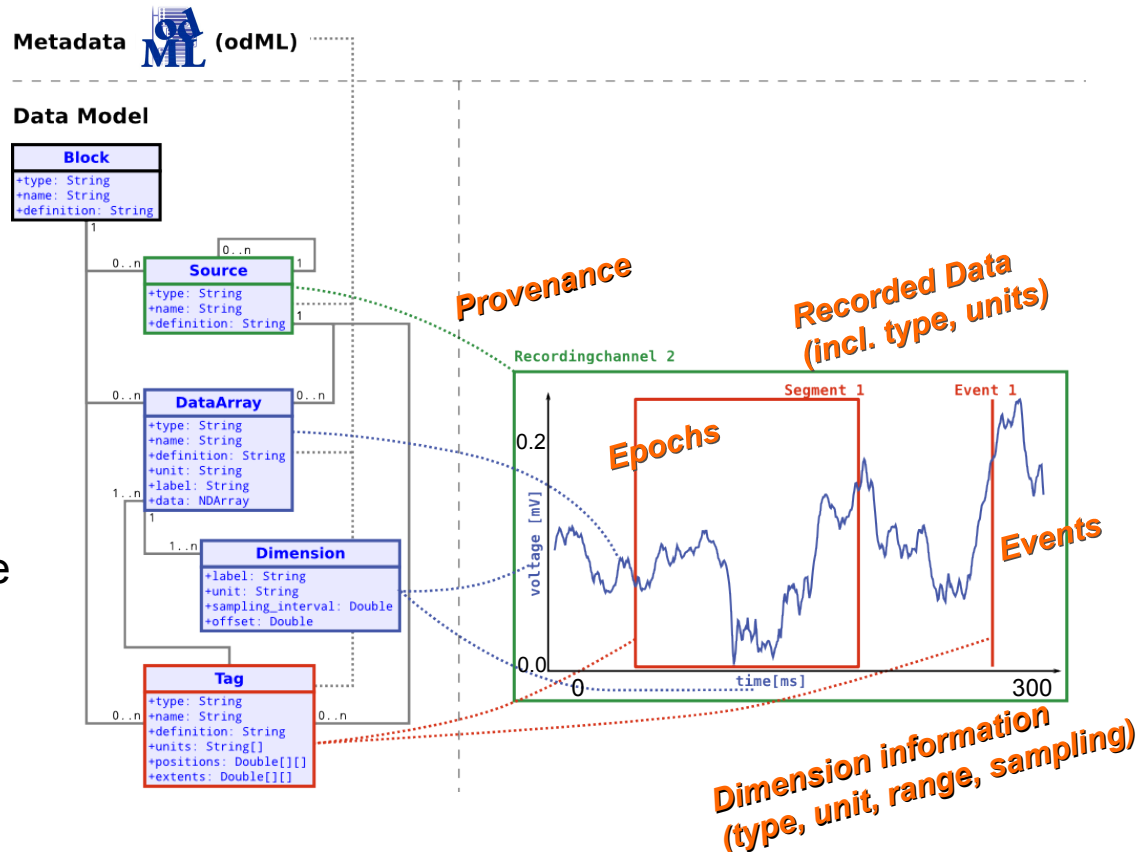


NIX – integration of data and metadata



<http://www.g-node.org/nix>

- general data model (derived from **Neo**) to represent recorded data, derived data, relations of data
- flexible data model for metadata (**odML**) for comprehensive annotation of data
- file backend: HDF5 file format
 - structure reflects data model, easy to understand
 - other storage backends possible
- libraries for different languages (C++, Python, Matlab, Java)
 - integration in data acquisition and analysis tools



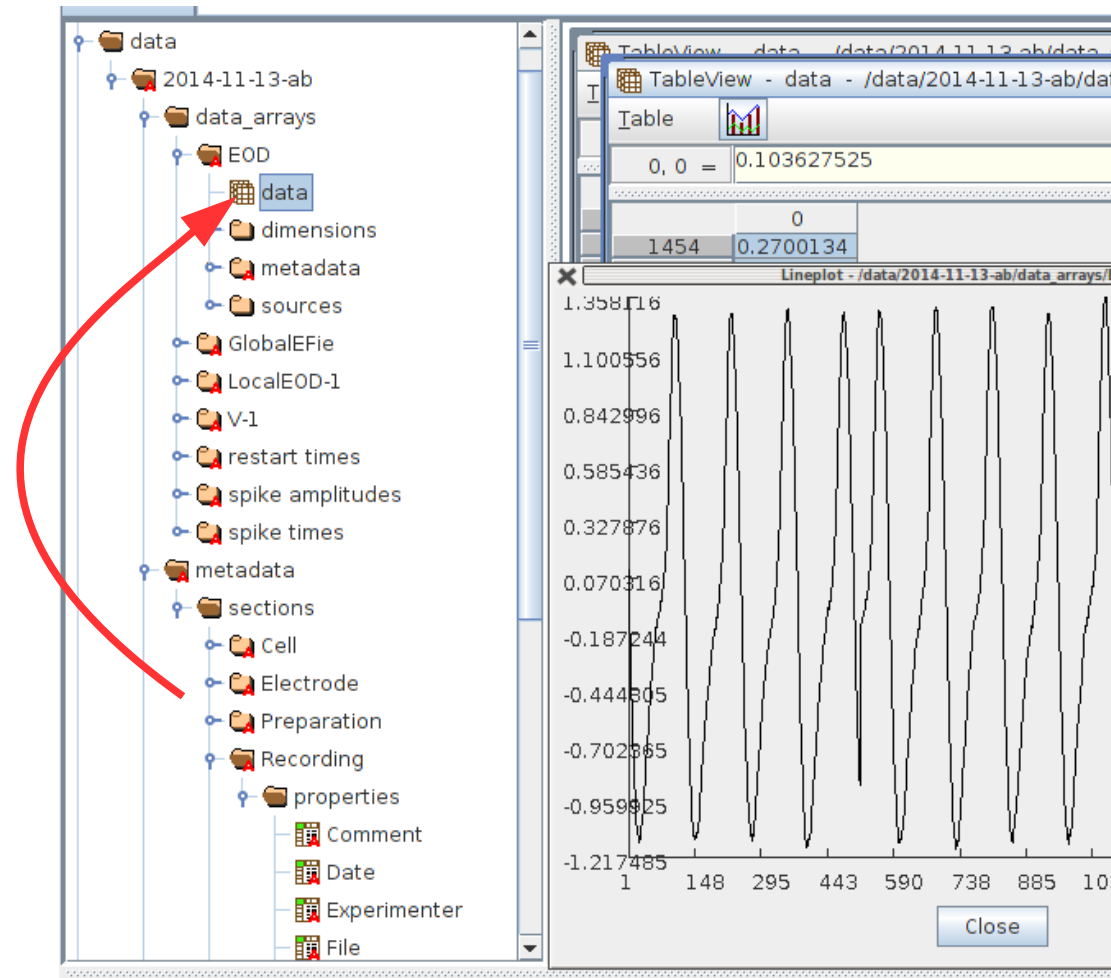
<https://github.com/G-Node/nix>



Benefits of integrated data management

Efficient data access:

- Querying data by metadata
- *“Give me all spike trains of single unit #4 from trials where the stimulus had a contrast of 0.5”*
- Facilitates automated analysis
- Seamless integration of data access into the lab data processing workflow

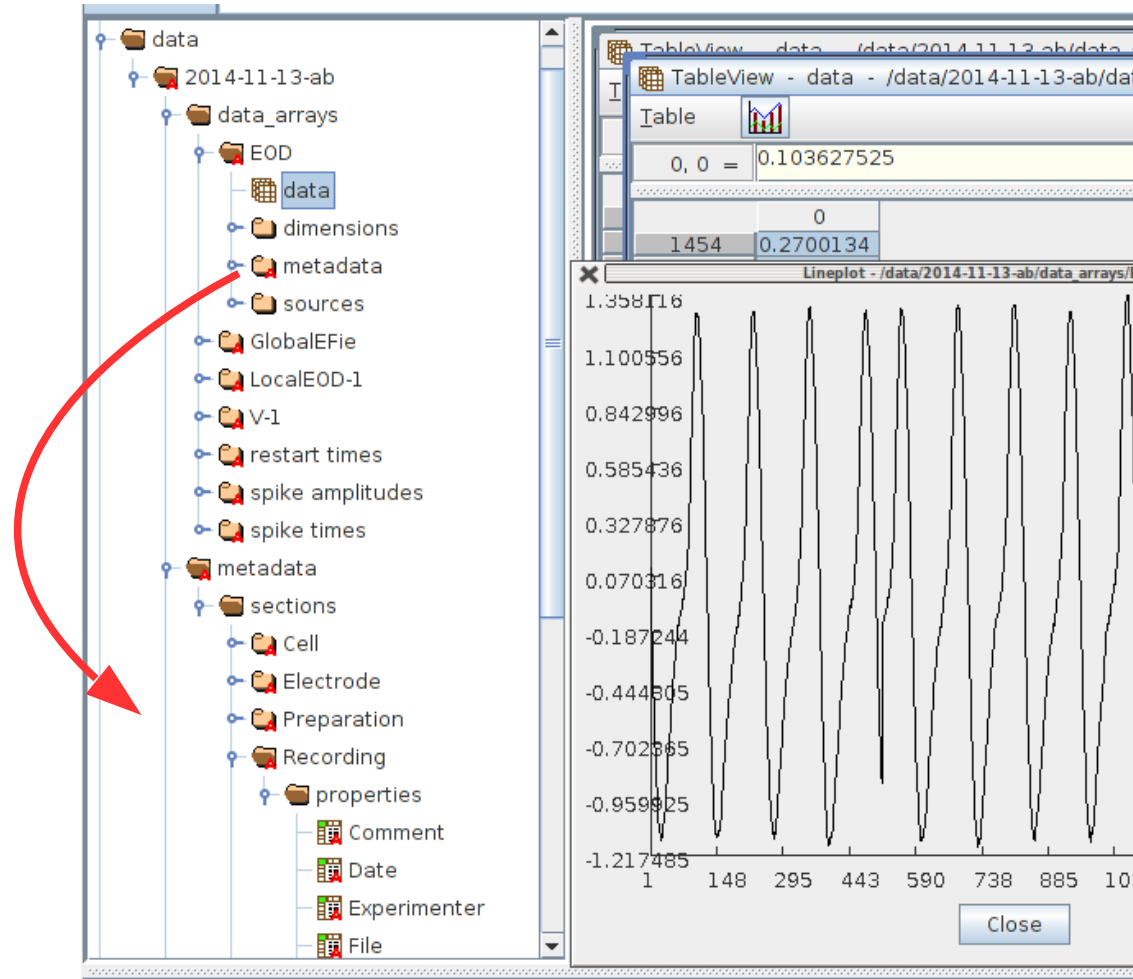




Benefits of integrated data management

Reproducibility:

- Identifying experimental conditions of recorded data
- *“What was the frequency of the stimulus that elicited this recorded response?”*
- Analysis results with provenance information can be stored consistently using the same format



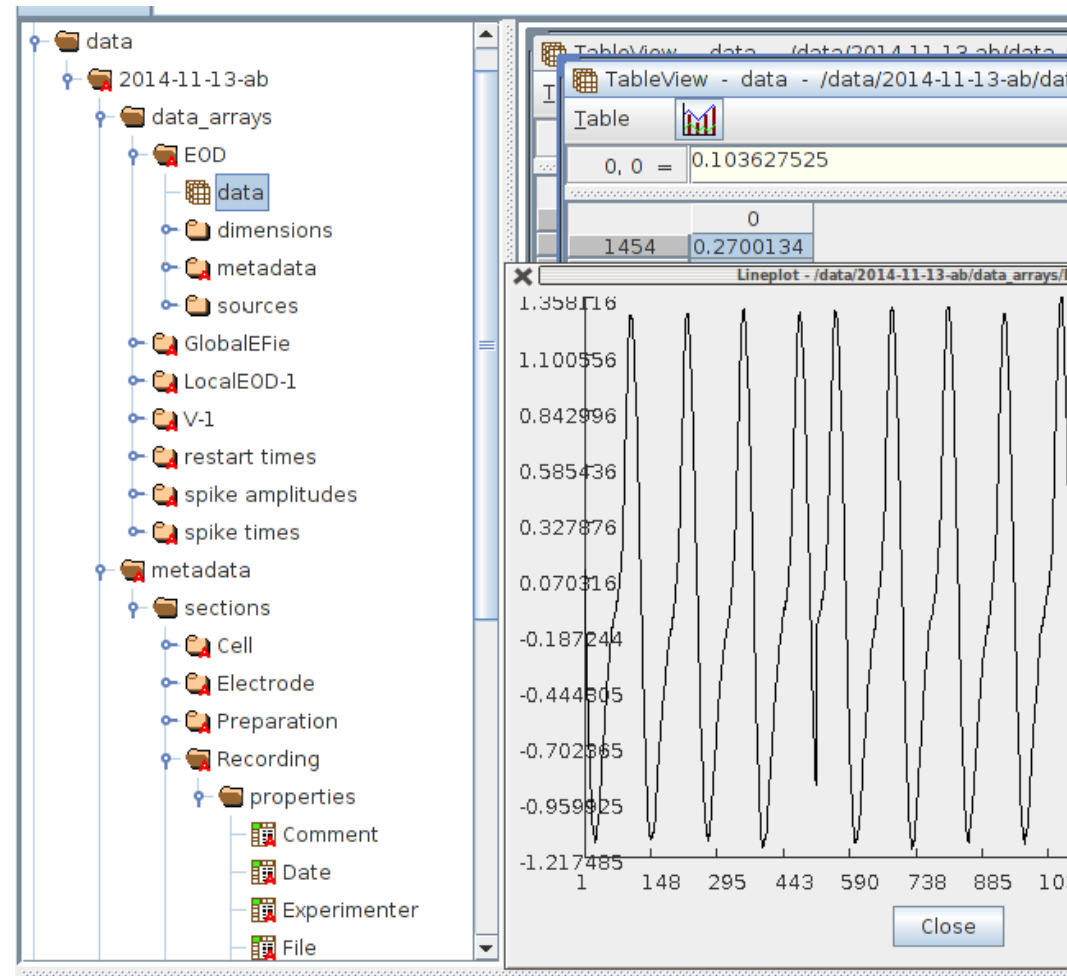


Benefits of integrated data management

Easy Data Sharing:

- For data provider: Minimizes need for preparation of data for sharing
- For data consumer: Minimizes need for communication
- Enables correct interpretation of stored data items
- Enables exploring data and automated data selection

Easy to understand for humans, but also machine-readable

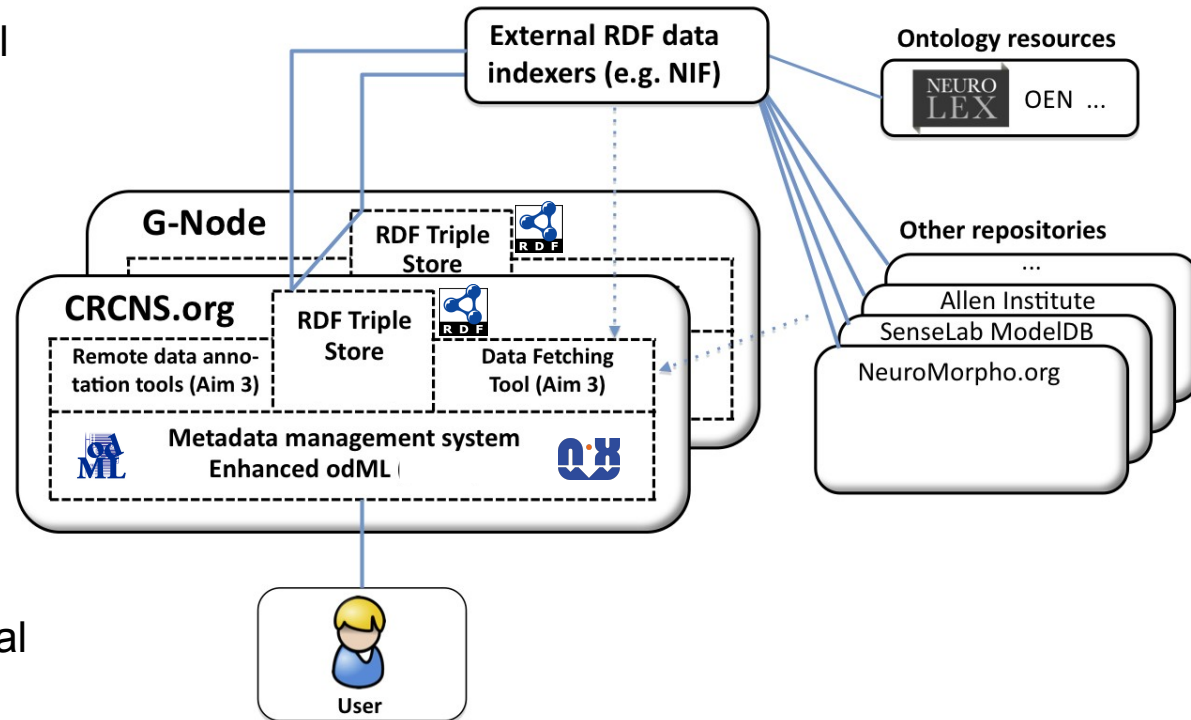


CRCNS US-German Data Sharing: Integrating distributed data sources

Collaboration with Fritz Sommer, UC Berkeley



- consistent annotation of neurophysiological data on CRCNS.org and G-Node portal using common format and terminologies (odML)
- metadata available via semantic web technologies, integration with NIF, NeuroMorpho.org, etc.
- data integration use cases: collecting datasets from distributed sources (e.g., morphological and physiological data)



Acknowledgments

G-Node Team

Christian Garbers, Christian Kellner, Achilleas Koutsou,
Andrey Sobolev, Michael Sonntag, Adrian Stoewer,
Jan Grewe, Andreas Herz, Willi Schiegel, Tiziano Zito

Collaborators, Contributors and Supporters

Hiroyuki Ai, Francesc Alted, Rembrandt Bakker, Jan Benda, Anubhav Chaturvedi,
Andrew Davison, Michael Denker, Markus Diesmann, Gaute Einevoll, Felix Franke,
Hagen Fritsch, Samuel Garcia, Daniel Gonzalez, Sonja Grün, Michael Hanke,
Hidetoshi Ikeno, Petr Jezek, Arvind Kumar, Ajayrama Kumaraswamy, Yann Le Franc,
Aljoscha Leonhardt, Philipp Meier, Balint Morvai, Roman Moucek, Dipanjan Mukherjee,
Matthias Munk, Martin Nawrot, Cristina Precup, Robert Pröpper, Raphael Ritz,
Jürgen Rybak, Michael Schmuker, Christine Seitz, Fritz Sommer, Zbyszek Szmek,
Christian Tatarau, Alvaro Tejero Cantero, Kay Thurley, Lyuba Zehl

