Open lightweight tools for safe and efficient data management, processing and validation

Michael Sonntag[1], Achilleas Koutsou[1], Jiří Vaněk[1], Christian Garbers[1], Christian Kellner[1], Mrinal Wahal[1], Jan Grewe[2], Thomas Wachtler[1]

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

Maintaining reproducible data workflows while keeping data in sync, backed up, and easily accessible from within and outside the lab is a key challenge in research. To minimize time and effort invested in these tasks scientists have to spend on these tasks, we provide a suite of tools designed for comprehensive, reproducible and versioned management of scientific data.

Data and Metadata Organisation

odML: Manage all Information about an Experiment

The odML Format
• Open metadata format
• Flexible hierarchical key-value storage
• Template system for reusable metadata structures

meta.g-node.org
• export odML to RDF
• access diverse metadata datasets
• all datasets are publicly available
• searchable by SPARQL via API and web

Re-usable Metadata Concepts:

https://templates.g-node.org
Re-usable building blocks to construct metadata files.
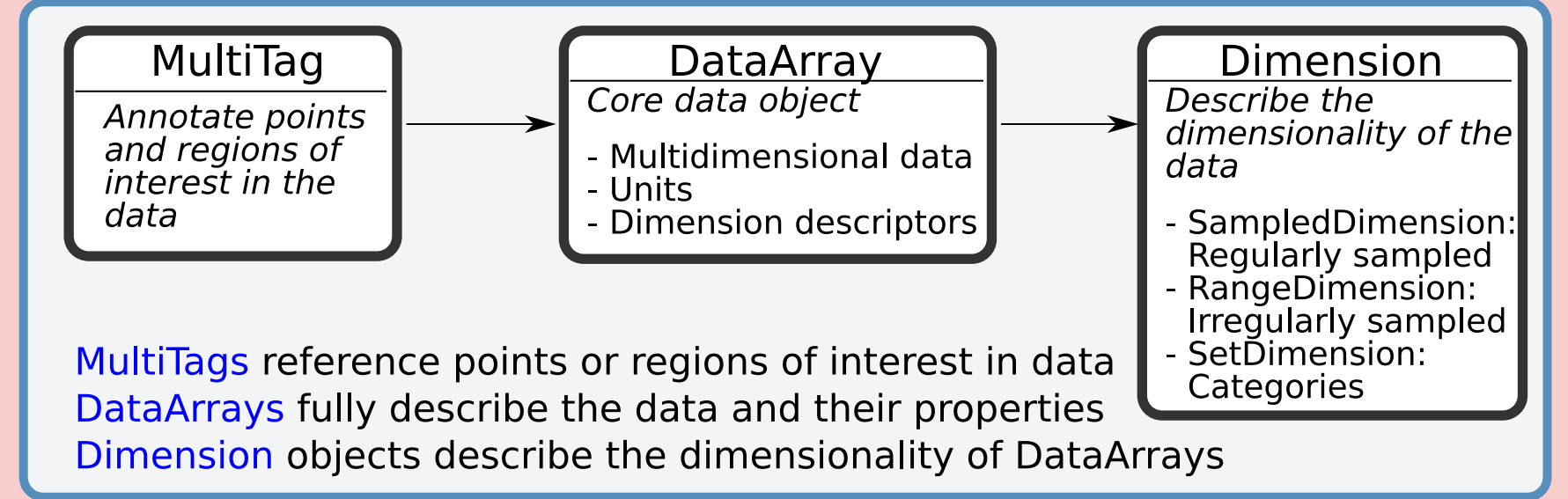
https://terminologies.g-node.org
Importable definitions to link to metadata entities.

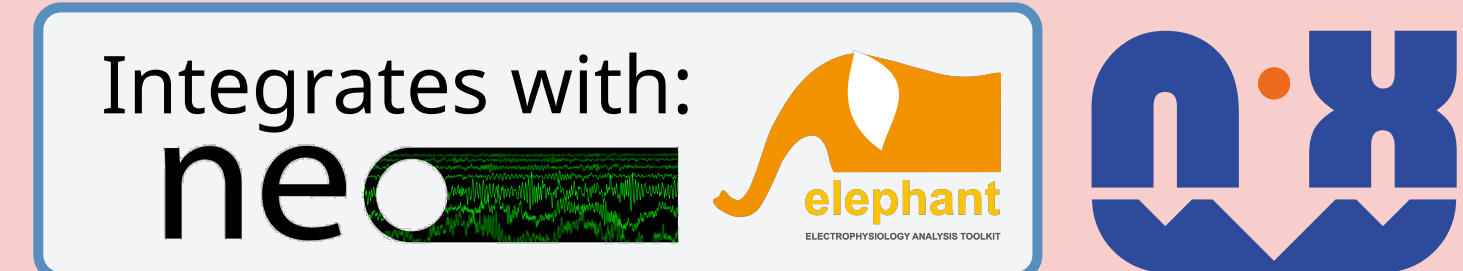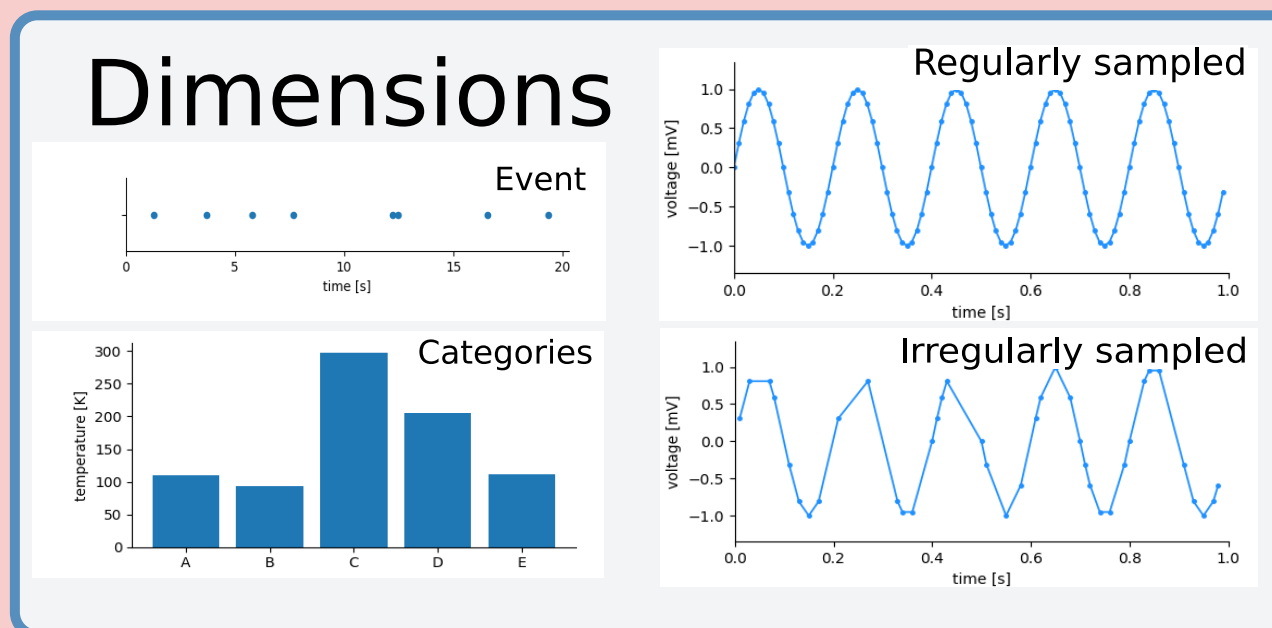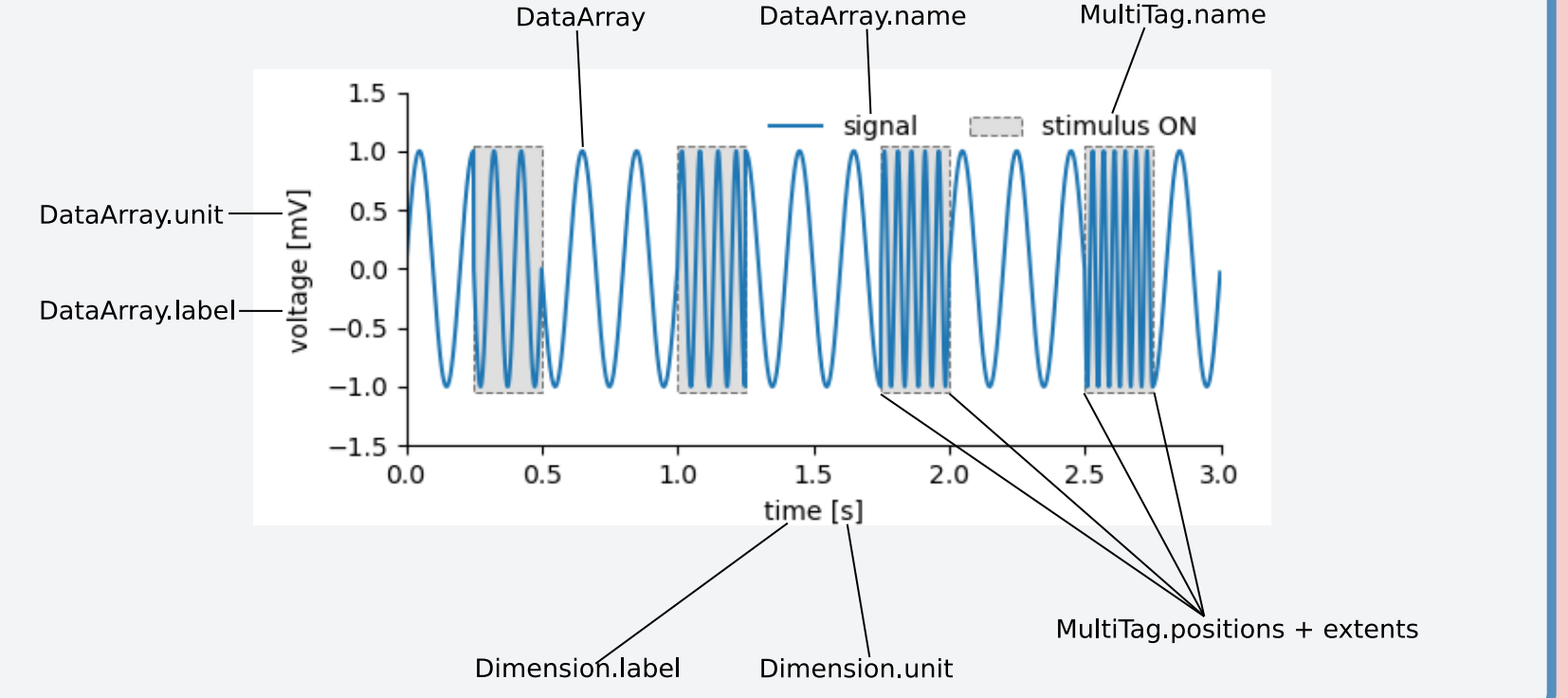Simplified data model of odML and NIX
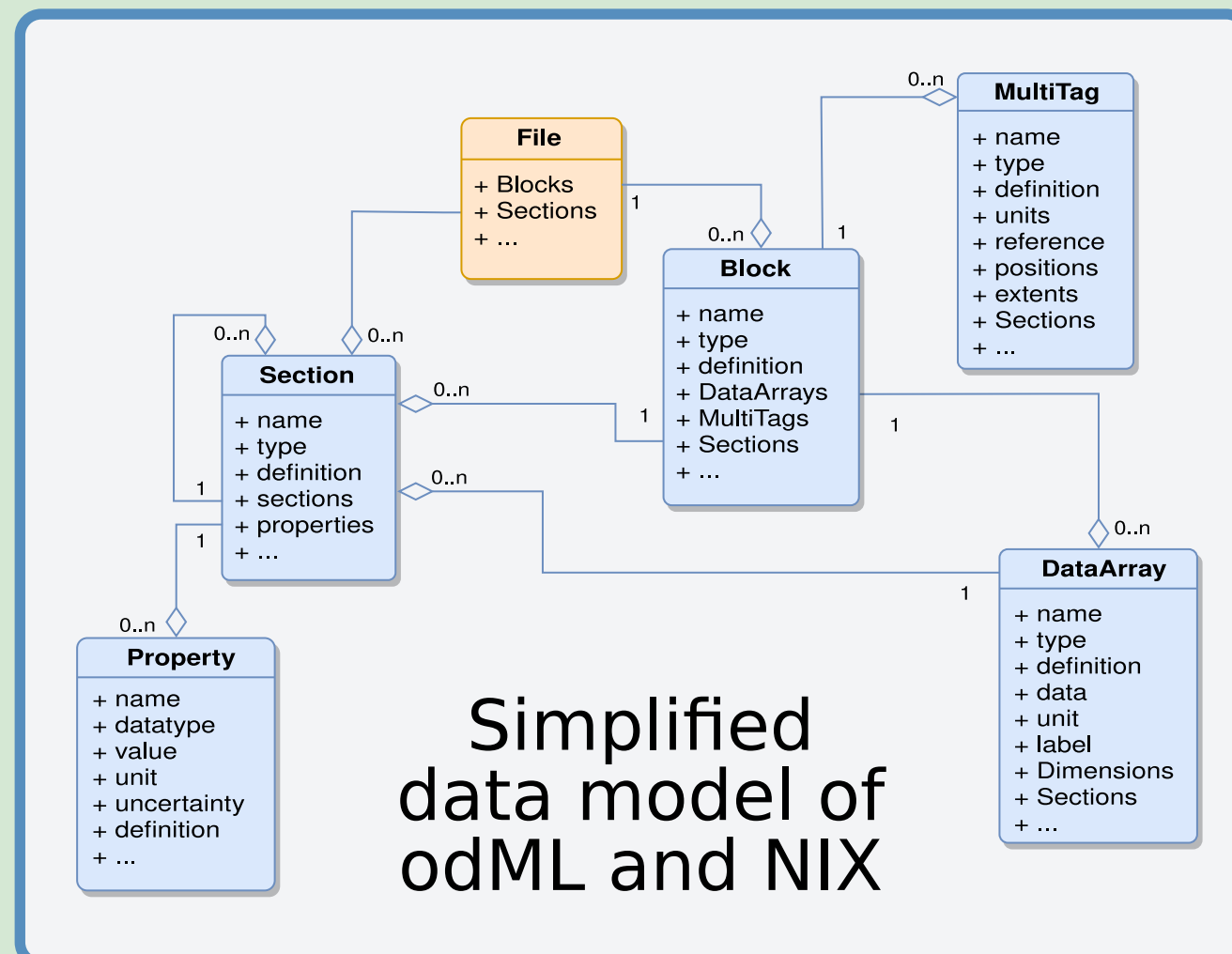
NIX: Manage Data and Metadata in one Versatile Format

The NIX Format
• Open data format
• Raw data, analysis results, and metadata in the same file
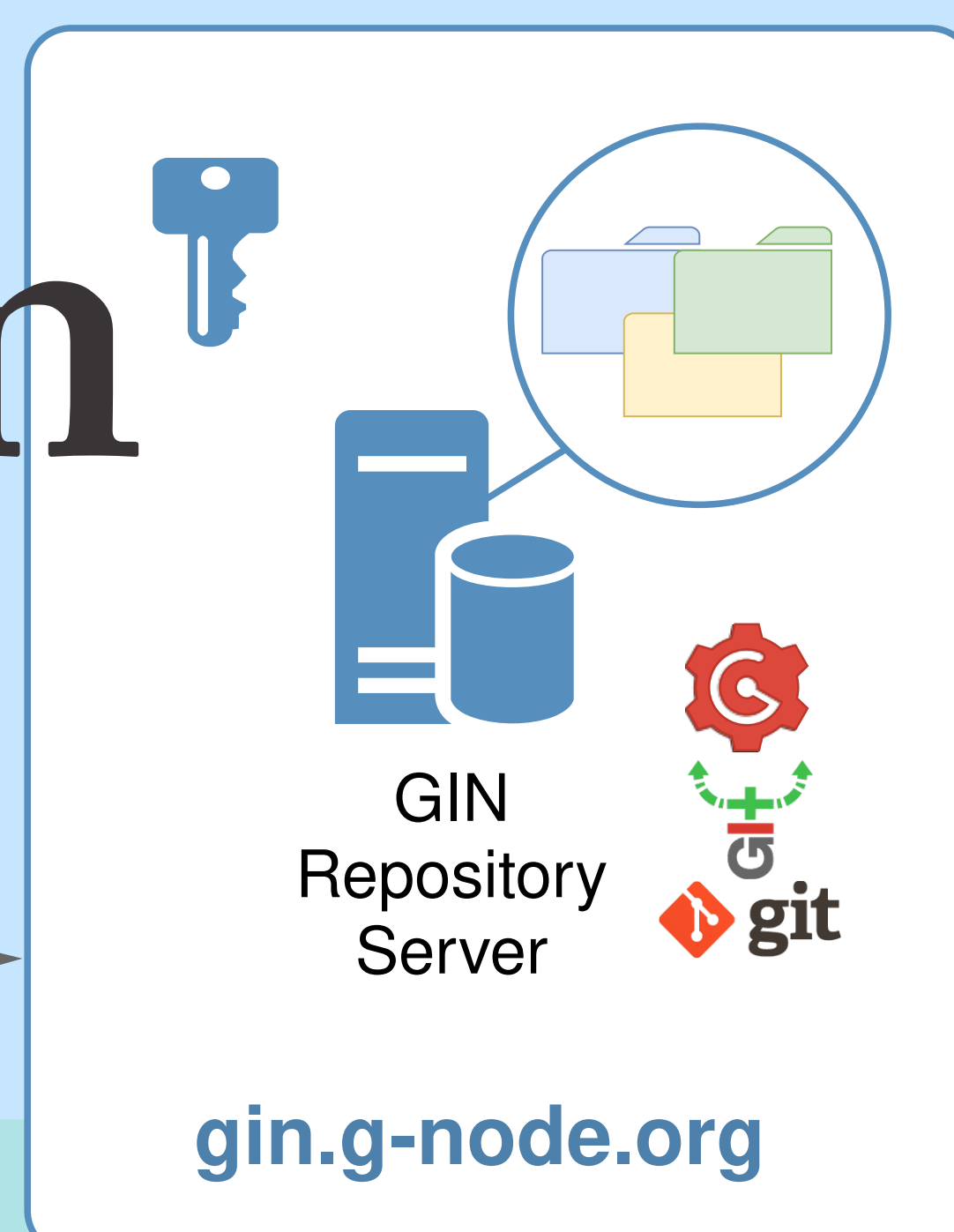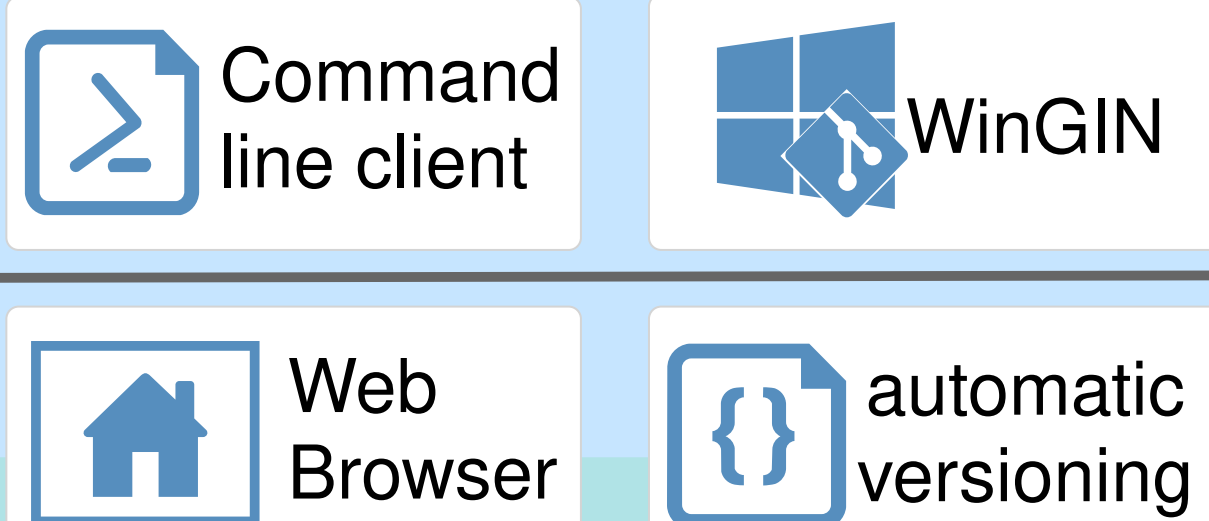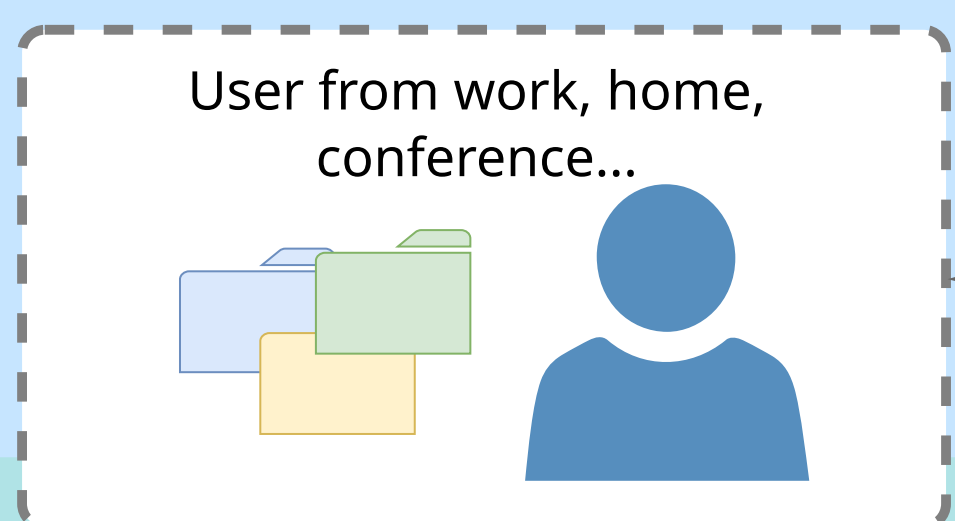• Self-contained and descriptive associations between data, analysis results, and metadata

MultiTag — Annotate points and regions of interest in the data
DataArray — Core data object: Multidimensional data, Units, Dimension descriptors
Dimension — Describe the dimensionality of the data: SampledDimension: Regularly sampled, RangeDimension: Irregularly sampled, SetDimension: Categories

MultiTags reference points or regions of interest in data
DataArrays fully describe the data and their properties
Dimension objects describe the dimensionality of DataArrays

Fully annotated file contains all information required to understand the data and their properties

Dimensions
Event
Categories
Regularly sampled
Irregularly sampled

Integrates with: neo, elephant, NIX

Data Storage, Collaboration, and Data Publication

GIN Core Features
• Secure remote access
• Versioning of datasets
• Access control: private, shared, public
• Data publication

User from work, home, conference…
Command line client
WinGIN
Web Browser
automatic versioning

gin
GIN Repository Server
gin.g-node.org

automatic versioning
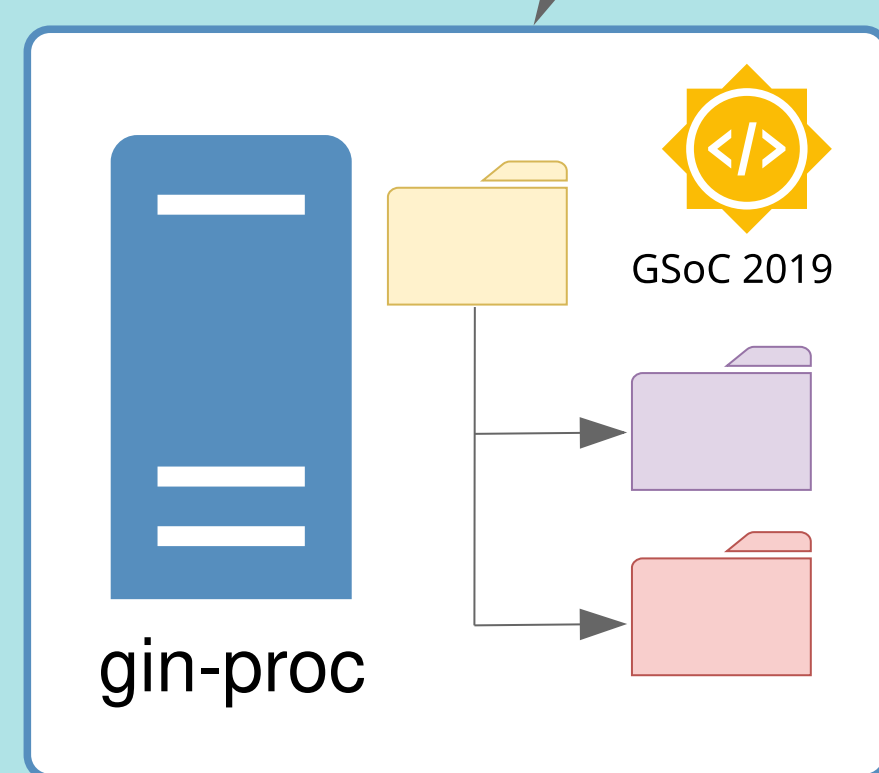Collaborators

Coordination and Collaboration

• User management
• User access levels
• Features for collaborative research
- Remote access
- Version control
- Issues
- Pull requests

Find us at the G-Node booth for demonstrations.

Automation and Validation Tools

gin-valid
Data validation service
valid.gin.g-node.org

Automated Data Validation
• Automatically runs validation on selected repositories
• Supported validation formats:
- BIDS - odML - NIX
• Easily extensible to more formats
• Format validation contributions are welcome

gin-proc
GSoC 2019
Data processing service
proc.gin.g-node.org

Automated Data Processing
• Automatically runs pre-defined processing pipelines
• Triggered on repository changes
• Automatically returns specified results
• Based on SnakeMake and DroneCI

Data Publication and Searchability

gin-dex
Data search service
gin.g-node.org/explore/data

Findable Data via GIN
• GIN provides automatic indexing of all text based files
• Online search for repository content
• Interactive rendering of
• Markdown • YAML
• JSON • XML

gin-doi
DOI service
doid.gin.g-node.org

Persistent Identifiers
• Any public GIN repository can be registered
• Make your code and data citable
• DOIs for:
- Data related to publications
- Research software
- Whole data sets

Resources and References