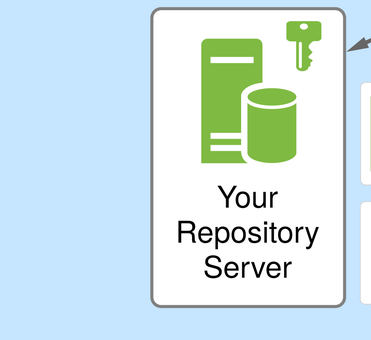
Microservice infrastructure for continuous validation, processing and indexing of research data on an open platform

Michael Sonntag, Achilleas Koutsou, Christian Garbers, Jiří Vaněk, Mrinal Wahal, Thomas Wachtler

German Neuroinformatics Node, Department Biologie II, Ludwig-Maximilians-Universität München, Germany

Research requires data to continuously be processed, analyzed and visualized; data needs to be quality checked, verified and backed up. Data and metadata need to be made publicly available in a manner that is easy to find and use. Many of these tasks can be automated, which usually leads to fewer errors and a higher results quality. To facilitate these tasks, we introduce a suite of microservices for the G-Node data infrastructure (GIN), an open platform for collaboration and sharing of research data and code.

GIN Microservices for Data Storage, Processing, Validation and Data Publication



GIN Core Features

- Secure remote access
- Versioning of datasets
- Access control
  - private
  - shared
  - public
- Data publication

Self Hosted Option

GIN or Local Hosting

- GIN is open source
- use the free online GIN service
- use your own self hosted instance

Microservice Architecture

GIN features a lightweight service framework that is easily extensible and maintainable.
Services can be added, modified or removed without interfering with core functionality or other loosely coupled services.

With the GIN microservices we aim to help improve data quality in the Neuroscientific community while keeping the main data hosting service pure and simple.

Coordination and Collaboration

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

Automation and Validation Tools

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

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

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

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

Poster presented at
INCF Neuroinformatics 2019
Warsaw, Poland

Contact:
dev@g-node.org

GIN (RRID:SCR_015864): https://gin.g-node.org
BIDS (RRID:SCR_016124): http://bids.neuroimaging.io
NIX (RRID:SCR_016196): http://www.g-node.org/nix
odML (RRID:SCR_001376): http://www.g-node.org/odml
SnakeMake (RRID:SCR_003475): https://doi.org/10.1093/bioinformatics/bts480
DroneCI: https://drone.io/ https://drone.io/

Supported by BMBF grants 01GQ1302, 01GQ1509

Federal Ministry of Education and Research