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

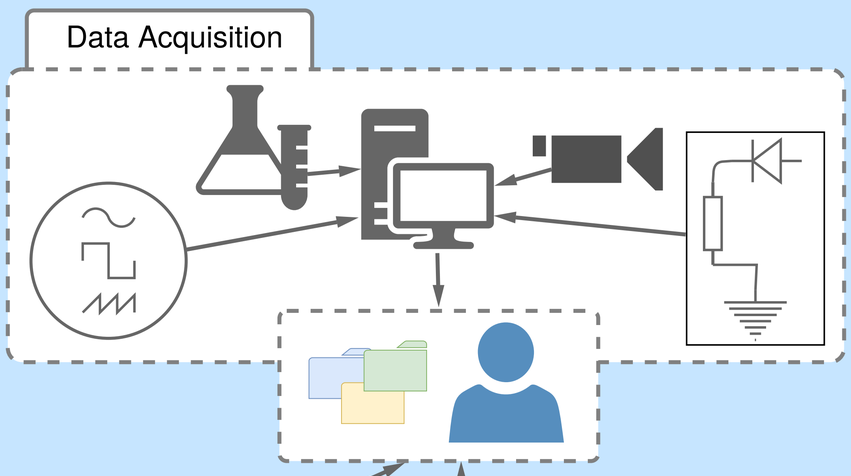
Michael Sonntag[1], Achilleas Koutsou[1], Christian Garbers[1], Jiří Vaněk[1], Mrinal Wahal[1], Thomas Wachtler[1]

[1]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
- Access data from any location
- Backup
- Built-in versioning
- Platform independent
- Secure access
- Public and private repositories



Microservice architecture

Microservices are small, self contained services that are easy to deploy and easy to maintain.

With the GIN microservices we aim to increase functionality for the Neuroscientific community while keeping the main data hosting service pure and simple.

Coordination and Collaboration

- User management
- User access levels
- On and offsite collaboration
- Online issues help coordination
- Ensure repository integrity with versioning and "Pull requests".

In-house option

Your Repository Server

Intranet

Command line client

Workflow
- Manual access via web and command line
- Script automation

Command line client

Web Browser

Collaborators

GIN or local hosting
- GIN is open source
- use the free online GIN service
- use your own in-house instance

GIN Repository Server

gin.g-node.org

Automation and validation tools

gin-valid

Data validation service

valid.gin.g-node.org

gin-proc

Data processing service

proc.gin.g-node.org

gin-dex

Data search service

gin.g-node.org/explore/data

Data publication and searchability

gin-doi

DOI service

doid.gin.g-node.org

Persistent publications
- Any public GIN repository can be published
- DOIs for paper publication data
- Research software publication
- Data set publication

Automated Data validation
- Opt in GIN service
- Automatically runs validation on selected repositories
- Supported validation formats:
  - BIDS
  - odML
  - NIX
- Easily extensible with custom formats
- Any format validation

Automated Data processing
- Opt in GIN service
- Based on snakemake and DroneCI
- Automatically runs pre-defined processing pipelines
- Triggered on repository changes
- Automatically returns specified result files safely to the users repository

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

Resources

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