

SODAR – THE IRODS-POWERED SYSTEM FOR OMICS DATA ACCESS AND RETRIEVAL

Mikko Nieminen

iRODS User Group Meeting, Utrecht (2019-06-26)

CONTENT

1. Background and Goals
2. SODAR Design
3. Rare Disease Genomics Use Case
Demonstration
4. Status and Ongoing Work
5. Conclusions

Background and Goals

Core Unit Bioinformatics (CUBI) at BIH

Consulting

Standardized Data Processing

- Access to tried and tested Omics workflows
- Infrastructure to process large (“inhouse” or “public”) data sets
- FAIR Data Management
- User Empowerment

Scientific Services

- Bioinformatics analysis tailored to specific needs and questions
- Access to Know-How of the Core Unit
- Pet / Research / Technology Development Projects

Training

Omics Data at CUBI

High Throughput Data from Various Sources

- Sequencing (genomics, transcriptomics..)
- Metabolomics
- Proteomics
- High throughput equals large data sizes and many measurements
- Data is heavily processed and reduced in size
 - Many files are necessary and worth keeping

Traditional Data Management

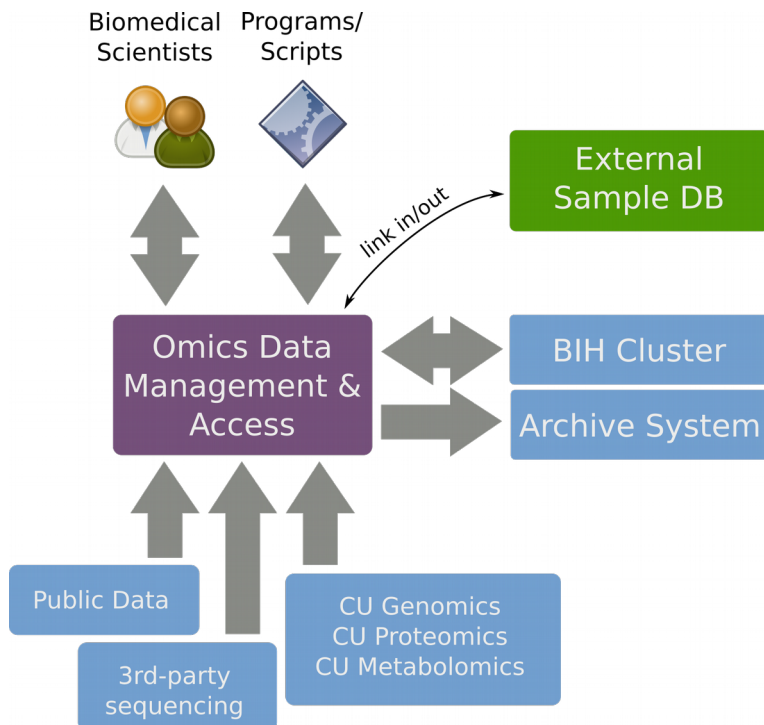
- Modeling study data in spreadsheets
- Files stored and shared using e.g. portable drives

Omics Data at CUBI

Key Requirements for Sustainable Data Management

- Large scale storage and archival of raw data
- Maintain context between study design meta-data and raw data files
- Data protection and access control
- Adhering to the FAIR principles (*Wilkinson et. al. 2016*)
 - **F**indable, **A**ccessible, **I**nteroperable, **R**euseable
- Multi-institute collaboration

Our Goals



Develop a System for Omics Data Access and Retrieval

- System to aid researchers and project owners manage and access omics data
- Support omics study design modeling
- Managed storage of large scale raw data
- Govern user access to data
- Linking data to third party systems / public data sources
- Enable collaboration between multiple organizations

Why iRODS?

Reasons for Choosing iRODS for Mass Storage

- Scalability and replication support
- Built-in meta-data functionality
- Potential in rule engine for e.g. data validation
- Flexibility: allows integration with out own infrastructure
- PAM support enables multi-organization authorization
- Nice community :)

Why not Go for Cloud?

- Data protection issues
- Cost issues
- iRODS offers better flexibility than “just” object storage
- S3 is there if needed

SODAR Design

SODAR Basics

The screenshot displays the SODAR Beta web interface. The top navigation bar includes a search bar, a 'Search' button, and links for 'Manual', 'Help', and a user profile. The left sidebar contains navigation icons for Home, Project Overview, Sample Sheets, Landing Zones, Small Files, and Timeline.

Home Page:

- Available Projects:** A table listing example projects with columns for Project, Description, and Your Role.

Project	Description	Your Role
Example Projects	Static example projects will go here	Owner
Cancer Example	Example cancer project	Owner
Germline Example	Example germline project	Owner
Proteomics Example	Proteomics model here	Owner
Transcriptomics Example	Transcriptomics example	Owner

Test Project Page:

- Test Project:** The quintessential test project.
- Project Members:** A table listing project members with columns for User, Name, Email, and Role.

User	Name	Email	Role
mikkopen@CHARITE	Mikko Nieminen	mikko.nieminen@bihealth.de	project owner ★
admin	Admin User	admin@example.com	project delegate ☆
alice	Alice Example	alice@example.com	project contributor

SODAR for the User

- Web site for user interaction
- REST APIs for programmatic access
- Access with existing institute credentials, supports multiple organizations

Projects and Roles

- Data is organized in projects and categories
- Project-specific roles are assigned to users
- Project meta-data and application data maintained in the SODAR database, certain meta-data also mirrored in iRODS
- Audit trails generated by the system with the ability to log project activity
- ID management: UUIDs generated for each project object, access via UUID

Study Design via Sample Sheets

Study Data

Row	Source	Name	Sex	Family	Father	Mother	Hpo Terms	Disease Status	External Links	Study
1	daughter	female	FAM_Demo	father	mother	-	unaffected	ID 19-0002		
2	father	male	FAM_Demo	0	0	-	unaffected	ID 19-0004		
3	mother	female	FAM_Demo	0	0	-	unaffected	ID 19-0003		
4	son	male	FAM_Demo	father	mother	Hand polydactyly	affected	ID 19-0001		

Assay: Cubi Utrecht Demo Genome Sequencing Nucleotide Sequencing

Assay Data

Row	Source	Sample	Process	Extract Name	Process	iRODS
1	daughter	daughter-N1	Nucleic acid extraction WGS	daughter-N1-DNA1	Library construction WGS	
2	father	father-N1	Nucleic acid extraction WGS	father-N1-DNA1	Library construction WGS	

Sample Sheets for Study Design

- Sample sheets contain sample and process meta-data for project studies
- Modeled in the ISA-Tools standard: <https://isa-tools.org/>
- Investigation > Study > Assay
- Graph models commonly represented as tables
- SODAR features a built-in browser to view and search the sample sheets
- Links out to raw data and external tools from e.g. specific samples
- CUBI altamISA parser used to read and write ISA model files (GitHub: [bihealth/altamisa](https://github.com/bihealth/altamisa))

Data File Management in iRODS

The screenshot shows the SODAR Beta web interface. The top navigation bar includes a search bar, 'Manual', and 'Help'. The left sidebar contains icons for Home, Project Overview, Sample Sheets, Landing Zones, Small Files, Timeline, Members, and Update Project. The main content area displays the 'CUBI Utrecht Demo' project, which is an iRODS UGM 2019 Demo. Below the project name is a 'Landing Zones' section with a 'Zone Operations' button. A table titled 'Your Zones' lists three zones:

Zone	Assay	Status Info	Status	Links
20190625_202553	Cubi Utrecht Demo Genome Sequencing Nucleotide Sequencing	Validating 51 files, write access disabled 51 files (28.2 GB)	VALIDATING	[Icons for file operations]
20190625_202619_keep_this	Cubi Utrecht Demo Genome Sequencing Nucleotide Sequencing	Available with write access for user 0 files (0 B)	ACTIVE	[Icons for file operations]
20190625_202628_to_be_deleted	Cubi Utrecht Demo Genome Sequencing Nucleotide Sequencing	Landing zone deleted	DELETED	[Icons for file operations]

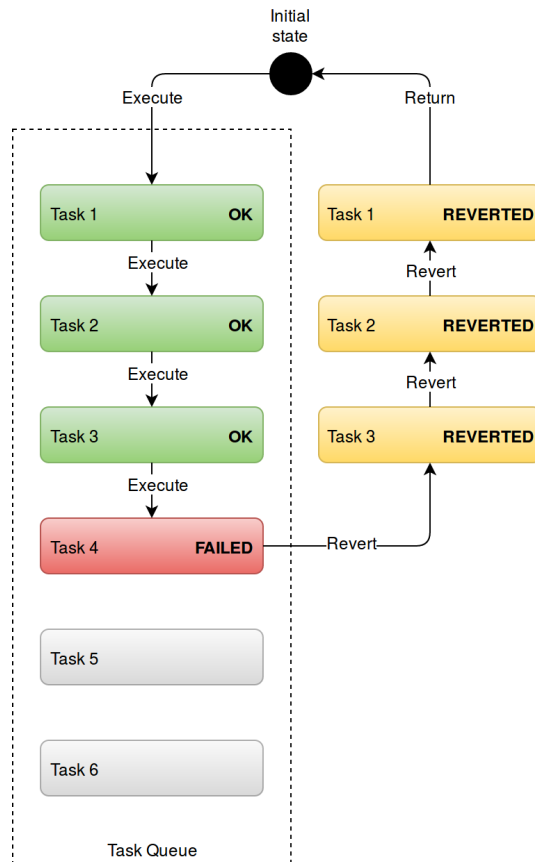
Data Files in iRODS

- Files organized in collections by project
- User access managed by SODAR
- Access via the same pre-existing institute credentials
- Links to iRODS resources provided in the web UI

Data Uploads via Landing Zones

- Files in project repositories are read-only
- Upload through user-specific landing zones
- Data validation → Rules for accepting data into repository

Managing iRODS Transactions



SODAR Taskflow: an In-House Transaction Engine

- Handles automated validation and moving of landing zone data into project repository within iRODS
- Reverts the transaction if failures are encountered → user can go back to alter their data in the landing zone
- Locks each project during transactions, to prevent data corruption
- REST API based Python service, uses Openstack Taskflow
- Updates transaction status in the SODAR web interface via its API
- Also makes use of iRODS rules (to be expanded in the future)

Accessing iRODS Data



This is a directory listing of the CUBI iRODS server.

Index of [/omicsZone/projects/fb/fb986607-b4fe-4cbf-870a-d441078bf80b/sample_data/study_8566e7c1-b4df-427e-8bb2-ffcd20430643/assay_50077a31-f4b8-4ab6-8547-4719dc960f5e/](#) on omicsZone

Parent collection

Name	Size	Owner	Last modified
CNMC_CNHS00037_00_032-N1-DNA1-WES1/		rods	2018-06-26 11:04
CNMC_CNHS00037_01_032-N1-DNA1-WES1/		rods	2018-06-26 11:04
CNMC_CNHS00047_00_041-N1-DNA1-WES1/		rods	2018-06-26 11:04
CNMC_CNHS00047_01_041-N1-DNA1-WES1/		rods	2018-06-26 11:04
CNMC_CNHS00047_02_041-N1-DNA1-WES1/		rods	2018-06-26 11:04
GRCh37/		rods	2018-06-27 14:54

Directory listing generated by Davrods.
Adwaita icons are cc-by-sa, by the GNOME Project.



Davrods

- DAV mounting
- Web-based file browsing
- Random access to large files

Integrative Genomics Viewer (IGV)

- Automated session file generation and serving
- Generated from sample sheets by SODAR, linking to iRODS files via Davrods

iCommands

- Working in landing zones also possible for command line and scripts

SODAR Core

Core Features as a Separate Project

- Project management & UI framework
- Reusable project apps
- Ability to create and install new apps in a plugin fashion
- Can be used to build new sites with their own configuration, applications and functionality
- Allows sharing project access between multiple sites
- Python package containing installable Django apps and an example site

Availability

- Publicly available In GitHub: [bihealth/sodar_core](https://github.com/bihealth/sodar_core)
- Latest release: v0.6.2 (2019-06-21)

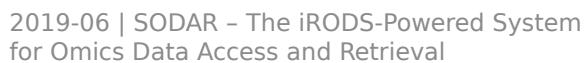
SODAR Technology

Web UIs and Applications

- Python 3
- Django
- Bootstrap
- Font Awesome
- JQuery
- Vue.js
- Ag-Grid
- Node/Webpack

Back-End and iRODS

- Davrods
- Python-Irodsclient
- AltamISA (ISA-Tools parser developed in CUBI)
- OpenStack Taskflow & Tooz
- Celery
- PostgreSQL
- Redis



Rare Disease Genomics Use Case Demonstration

Status and Ongoing Work

Status and Ongoing Work

SODAR Usage

- Deployed at CUBI in beta
- Second instance in use at Uni. Bonn
- Actively used in dozens of projects with collaborators
- Talks with other organizations interested in adopting SODAR

SODAR Development

- Source code will be published, as well as submitting scientific publications
- SODAR Core already made public on GitHub
- SODAR Core in use as the platform for several other CUBI software projects (Varfish, Digestiflow..)
- Development is ongoing

Ongoing and Future Work

- Integrated editor for sample sheets
- More advanced validation of data in iRODS
- A more comprehensive REST API
- Etc., etc.

Conclusions

Conclusions

SODAR

- Has proven to be a valuable aid to researchers in CUBI omics projects
- Interest from several organizations
- Core parts also in active use by several other systems
- SODAR and its parts are expected to evolve further

iRODS in SODAR

- iRODS was our choice when starting to build initial prototypes
- Remains as the mass storage platform of choice
- Utilized comprehensively from iCommands to Python APIs and Davrods
- We envision more use for e.g. the rule engine in the future..
- Deployment to be scaled up in the future as well

Acknowledgements

Collaboration

- Special thanks to Chris Smeele for his work with Davrods
- Numerous BIH researchers and collaborators using the system, reporting bugs etc.

CUBI

- Dieter Beule and Manuel Holtgrewe for requirements, support and feedback
- Mathias Kuhring for work with the altamISA parser
- Franziska Schumann for code contributions



THANK YOU!



CONTACT

Mikko Nieminen

Senior Software Engineer

**Berlin Institute of
Health (BIH)**

mikko.nieminen@bihealth.de

www.bihealth.org