

An Update on SODAR: the iRODS-powered System for Omics Data Access and Retrieval

Mikko Nieminen

Senior Software Engineer, Core Unit Bioinformatics
IRODS User Group Meeting 2022, Leuven



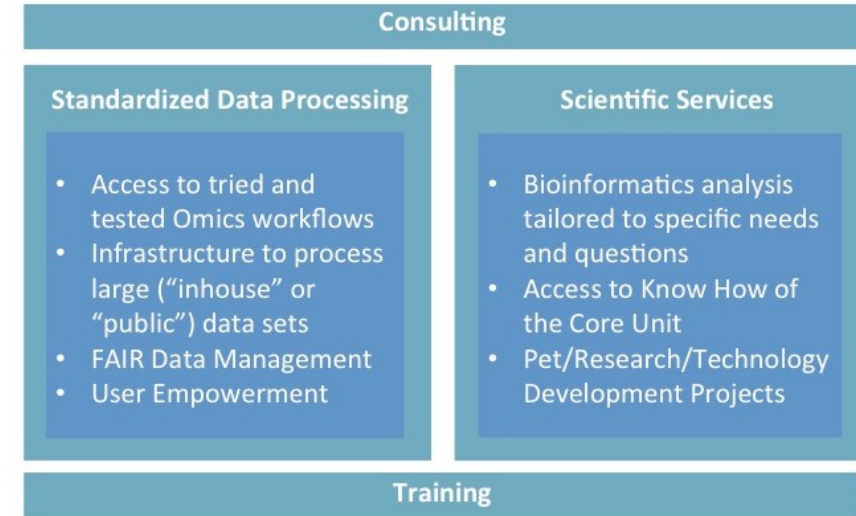
Contents

1. Background
2. The SODAR System
3. New Features
4. Live Demonstration
5. Status and Ongoing Work
6. Conclusions

Background

Core Unit Bioinformatics at BIH

- **Core Unit Bioinformatics (CUBI)**
 - We provide bioinformatics and data analysis expertise for translational research
- **Omics Data at CUBI**
 - High throughput data from various sources (sequencing, metabolomics, proteomics..)
 - Large data sizes and many measurements
- **Study Design Modeling**
 - Study metadata must be recorded in an organized fashion
 - Files relevant to studies should be easily accessible



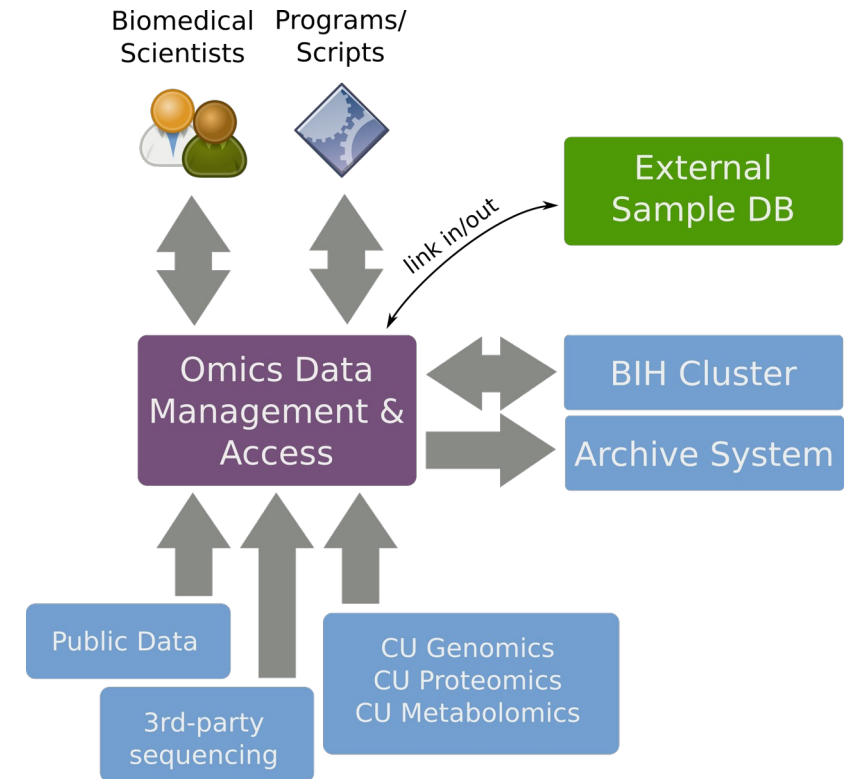
Requirements for Sustainable Data Management

- Traditional data management practices are not sufficient
 - Spreadsheets, portable hard drives..
- **Requirements**
 - Large scale centralized storage and archival of raw data
 - Maintain context between study design and stored files
 - Data protection and access control
 - Adhering to the FAIR principles (Wilkinson et.al. 2016)
 - Findability, Accessibility, Interoperability and Reuse
 - Multi-institute collaboration

The SODAR System

SODAR Design (1/2)

- **SODAR** is our solution to meet the omics data management requirements
- **Features**
 - Project based access control and data encapsulation
 - Management of study design metadata
 - Large scale data storage
 - Linking stored files to metadata
 - Tools for aiding data management in research projects
- Implemented with open source tools: Python 3, the Django web server, Vue.js, etc.



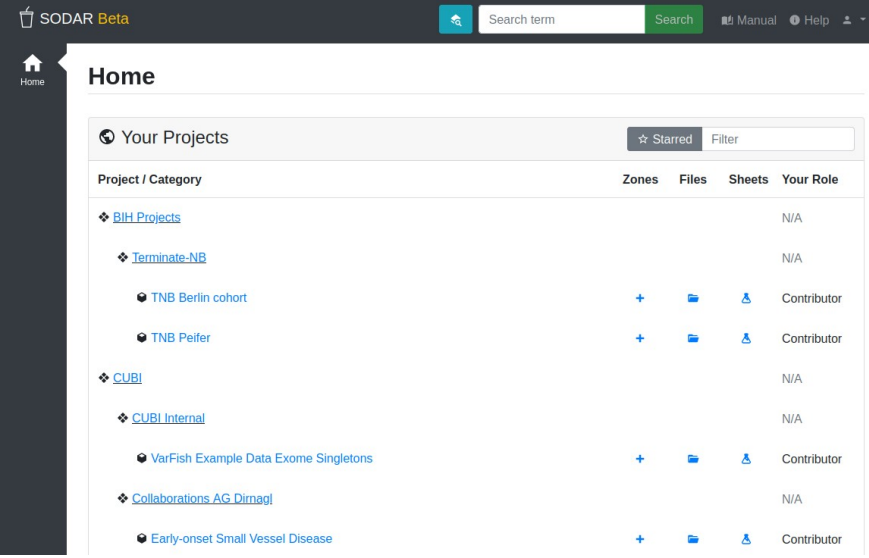
SODAR Design (2/2)

- **SODAR for the User**

- Web UI for user interaction in the browser
- REST APIs for scripts and software
- Davrods for WebDAV and random file access
- Access with existing institute credentials, support for multiple organizations

- **Projects**

- Data is organized in projects and categories
- Project-specific roles are assigned to users
- SODAR also manages iRODS user access



The screenshot shows the SODAR Beta web interface. The top navigation bar includes the SODAR Beta logo, a search bar, and links for Manual and Help. The main content area is titled 'Home' and displays 'Your Projects' with a 'Starred' filter. The projects are listed in a table with columns for Project / Category, Zones, Files, Sheets, and Your Role. The projects are grouped into categories like BIH Projects, CUBI, and Collaborations AG Dirmagl. Below the projects list, there is a section for the Berlin Institute of Health, Charité & Max Delbrück Center, which is a directory listing of the CUBI iRODS server. This section includes a URL for the index of a specific project and a table showing the parent collection details.

BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

This is a directory listing of the CUBI iRODS server.

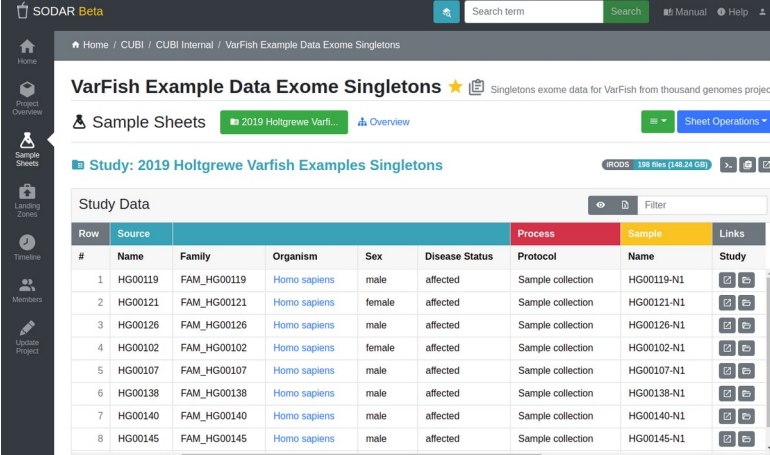
Index of /sodarZone/projects/a1/a12164ff-9753-4d32-a520-7f499f862910/sample_data/study_34dcf3de-e410-4e53-b717-38f6bf6407f0/assay_a435a0b1-9ef0-4fc7-b69f-faff954a260e/HG00119-N1-DNA1-WES1/raw_data/2019-09-18/ on sodarZone

Parent collection

Name	Size	Owner	Last modified
SRR099967_1.fastq.gz	7.4G	holtgrem@CHARITE	2019-09-23 08:27
SRR099967_1.fastq.gz.md5	55	holtgrem@CHARITE	2019-09-23 08:27
SRR099967_2.fastq.gz	7.5G	holtgrem@CHARITE	2019-09-23 08:27
SRR099967_2.fastq.gz.md5	55	holtgrem@CHARITE	2019-09-23 08:27

SODAR Data Workflow

- **Sample sheets** contain sample, process and material metadata for project studies
 - Modeled in the ISA-Tab format: isa-tools.org
 - Investigation > Study > Assay
 - Node graphs represented as spreadsheet-style tables
- **Large scale study data** is stored in iRODS
 - Sample sheets link to relevant files within assays
 - SODAR is file type agnostic, but e.g. certain collection structures are enforced
- **Landing zones** are used to upload new sample data
 - User and assay specific temporary file areas
 - Once uploaded, data is automatically validated and moved into read-only sample data repository
 - iRODS transactions with rollback on errors



SODAR Beta

Home / CUBI / CUBI Internal / VarFish Example Data Exome Singletons

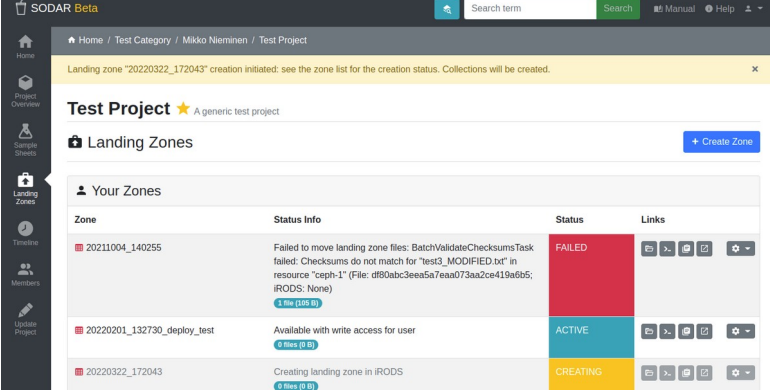
VarFish Example Data Exome Singletons ★ Singletons exome data for VarFish from thousand genomes project

Sample Sheets 2019 Holtgrewe Varf... Overview Sheet Operations

Study: 2019 Holtgrewe Varfish Examples Singletons iRODS 108 files (148.24 GB)

Study Data

Row	Source	Family	Organism	Sex	Disease Status	Process	Sample	Links
1	HG00119	FAM_HG00119	Homo sapiens	male	affected	Sample collection	HG00119-N1	
2	HG00121	FAM_HG00121	Homo sapiens	female	affected	Sample collection	HG00121-N1	
3	HG00126	FAM_HG00126	Homo sapiens	male	affected	Sample collection	HG00126-N1	
4	HG00102	FAM_HG00102	Homo sapiens	female	affected	Sample collection	HG00102-N1	
5	HG00107	FAM_HG00107	Homo sapiens	male	affected	Sample collection	HG00107-N1	
6	HG00138	FAM_HG00138	Homo sapiens	male	affected	Sample collection	HG00138-N1	
7	HG00140	FAM_HG00140	Homo sapiens	male	affected	Sample collection	HG00140-N1	
8	HG00145	FAM_HG00145	Homo sapiens	male	affected	Sample collection	HG00145-N1	



SODAR Beta

Home / Test Category / Mikko Nieminen / Test Project

Landing zone "20220322_172043" creation initiated: see the zone list for the creation status. Collections will be created.

Test Project ★ A generic test project

Landing Zones Create Zone

Your Zones

Zone	Status Info	Status	Links
20211004_140255	Failed to move landing zone files: BatchValidateChecksumsTask failed: Checksums do not match for "test3_MODIFIED.txt" in resource "ceph-1" (File: df80abc3eea5a7eaa073aa2ce419a6b5; iRODS: None) View (105 B)	FAILED	
20220201_132730_deploy_test	Available with write access for user View (5 B)	ACTIVE	
20220322_172043	Creating landing zone in iRODS View (5 B)	CREATING	

Status at Last Presentation (UGM 2019)

- SODAR in development, in use at CUBI
 - Used in dozens of projects
 - Parts of source code made public
- **Features**
 - Import, viewing and searching of ISA-Tab sample sheets
 - File uploads to iRODS via landing zones
 - Linking to iRODS files from sample sheets
 - IGV genome browser integration from sample sheets
 - Limited REST API for specific functionalities

New Features

New Features: Sample Sheets (1/3)

- **Sample Sheet Creation from Templates**
 - Create ISA-Tab compatible sample sheets in the SODAR UI
 - Multiple templates are available for different types of research projects
 - Templates are created with Cookiecutter
 - In the future, we intend to make it easy to introduce new templates

Create from "Generic RNA sequencing ISA-tab template"

Directory Name*

Template_Test

Investigation directory name and assay prefix

investigation_title*

Investigation Title

sample_names*

alpha,beta,gamma

a_measurement_type*

transcription profiling

a_measurement_types*

{"transcription profiling": {"accession": "http://purl.obolibrary.org/obo/OBI_0000424", "source": "OBI"}}

a_technology_type*

nucleotide sequencing

a_technology_types*

{"nucleotide sequencing": {"accession": "http://purl.obolibrary.org/obo/OBI_0000626", "source": "OBI"}}

New Features: Sample Sheets (2/3)

- **Sample Sheet Editing**

- Sample sheet ISA-Tabs can be edited in the SODAR UI
- Editing cell values
- Restricting columns to a specific format
- Inserting and deleting rows
- Ontology term lookup
- Sheet version management with comparison, restoring and exporting
- Maintaining full ISA-Tab TSV compatibility at all states of editing
- Not a 100% feature complete ISA-Tab editor (yet), but usable

Study Data					+ Insert Row		
Row	Source		Process				
#	Name	Age	Protocol	Instrument			
1	0814	91 day	sample collection	scalpel			
2	0815	91 day	library preparation	scalpel			
3	0815	91 day	nucleic acid sequencing	scalpel type A; scalpel			
4	0816	-	sample collection	scalpel			
5	0817	149 day	sample collection	scalpel			

Disease Status

Paste

Editable

☒

Format

select

Options

affected
unaffected

Default Value

Default Fill

☐

New Features: Sample Sheets (3/3)

- **Ontology Term Lookup**

- Import common ontologies into SODAR
- Query via local API in UI
- Examples of supported ontologies for import: HP, NCBITAXON, OMIM, ORDO, UBERON...
- Manual term editing also supported
- Support for multiple ontologies and terms per cell

A001: Hpo Terms Paste

limb muscle HP ☐ Sort by ontology

[HP:0007156] Asymmetric limb muscle stiffness

[HP:0009053] Distal lower limb muscle weakness

[HP:0040267] Distal upper limb muscle hypertrophy

[HP:0008959] Distal upper limb muscle weakness

[HP:0030198] Fatigable weakness of distal limb muscles

[HP:0030200] Fatigable weakness of proximal limb muscles

[HP:0009055] Generalized limb muscle atrophy

[HP:0009028] Generalized weakness of limb muscles

Name	Ontology	Accession	
Abnormal location of ears	HP	http://purl.obolibrary.org/obo/HP_...	↑ ↓ ✎ ✖
Mild expressive language delay	HP	http://purl.obolibrary.org/obo/HP_...	↑ ↓ ✎ ✖
obsolete Anaphylactoid purpura ⚠	HP	http://purl.obolibrary.org/obo/HP_...	↑ ↓ ✎ ✖

+

Ontology Access

Import Ontology

OBO Format Ontologies

Name	Title	ID	Terms	Imported	
CL	Cell Ontology	cl	5605	2021-05-04 14:28	⚙
DUO	Data Use Ontology	duo	41	2021-05-04 13:37	⚙
HP	Human Phenotype Ontology	hp.obo	16173	2021-05-04 13:38	⚙
MS	Mass Spectrometry Ontology	ms	3016	2021-05-04 13:41	✎ Update Ontology ✖ Delete Ontology
NCBITAXON	NCBI Taxonomy	ncbitaxon/subsets/taxslim	14168	2021-05-04 13:41	⚙
OMIM	Online Mendelian Inheritance in Man	OMIM.csv	54658	2021-05-04 13:43	⚙

New Features: APIs

- **REST API**

- REST APIs now implemented for most SODAR features
- Project creation and access control
- Sample sheet import/export
- Landing zone management

- **Access Tokens**

- API access tokens can be generated and managed in the UI
- Can be set to expire

API Views

```
class samplesheets.views_api.InvestigationRetrieveAPIView(**kwargs) \[source\]
```

Retrieve metadata of an investigation with its studies and assays.

This view can be used to e.g. retrieve assay UUIDs for landing zone operations.

URL: `/samplesheets/api/investigation/retrieve/{Project.sodar_uuid}`







Methods: `GET`

Returns:

- `archive_name`: Original archive name if imported from a zip (string)
- `comments`: Investigation comments (JSON)
- `description`: Investigation description (string)
- `file_name`: Investigation file name (string)
- `identifier`: Locally unique investigation identifier (string)
- `iroids_status`: Whether iRODS collections for the investigation have been created (boolean)
- `parser_version`: Version of altamISA used in importing (string)
- `project`: Project UUID (string)
- `sodar_uuid`: Investigation UUID (string)
- `studies`: Study and assay information (JSON, using study UUID as key)
- `title`: Investigation title (string)

API Tokens

[+ Create Token](#)

#	Created	Expires	Key	
1	2022-07-04 15:25	2022-07-05 21:25	c3829275	 
2	2022-07-04 15:25	Never	cab6db39	 
3	2022-07-04 15:25	2022-07-05 05:25	8b106722	 

New Features: iRODS (1/2)

- **Ticket-based Access Control**

- Enable ticket-based access for specific iRODS collections in the project sample data repository
- Allows access from external software
- Used for integrating with the UCSC Genome Viewer
- This will be expanded for more generic use cases

- **File Deletion Requests**

- Users can request for deletion in case of e.g. mistakes
- Project owner or delegate must accept requests
- Requests for moving/renaming to be added in the future

Create iRODS Access Ticket

Path*

Demo Assay / DemoHub1

Path to iRODS collection

Label

Ticket label (optional)

Expiry date










mm/dd/yyyy

DateTime of ticket expiration (leave unset to never expire; click x on righthand-side of field to unset)

Cancel Create

iRODS Delete Requests

Project Sheets + Create Request

Path	User	Created	Status	
13feadf/test2.txt ⓘ	 carol	2021-11-09 14:59	REJECTED	
75tgerb53/test1.txt ⓘ	 carol	2021-11-09 14:59	ACCEPTED	
2fewfd/test2.txt ⓘ	 carol	2021-11-09 14:30	ACTIVE	
3140-N1-DNA1-WGS1/test1.txt ⓘ	 carol	2021-11-09 14:30	 Update Request  Delete Request	

New Features: iRODS (2/2)

- **Authentication with SODAR**
 - PAM auth via SODAR if not using external LDAP
- **Admin Tools**
 - Tools for locating orphaned or misplaced files (not corresponding to project study design)
- **Command Line Tooling**
 - Command line tools have been developed for SODAR and iRODS operations
 - Using the SODAR REST API, iRODS Python client and iCommands
 - For e.g. standardized ingestion of specific files

Live Demonstration

Status and Ongoing Work

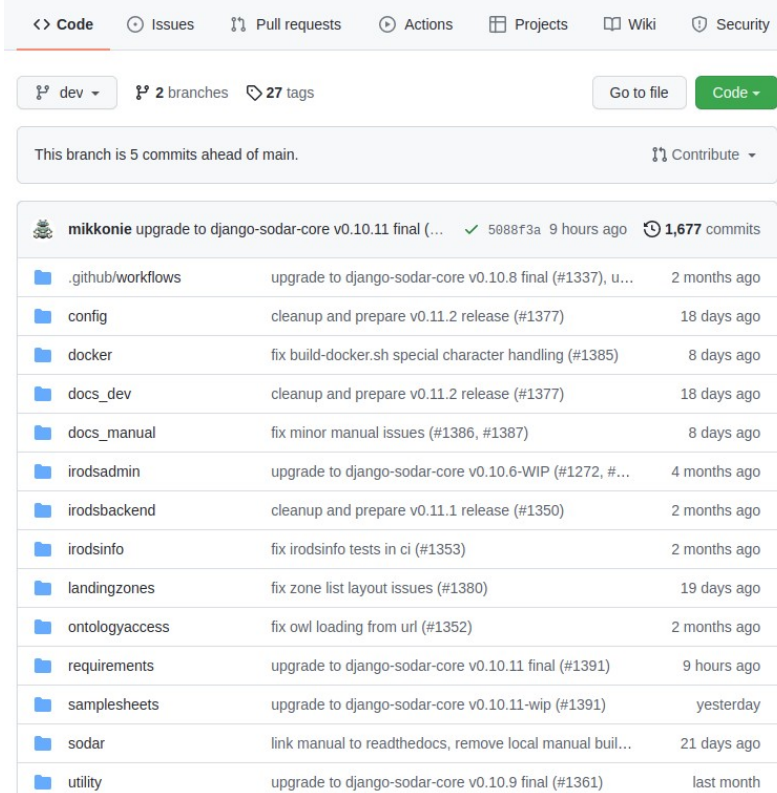
Status and Ongoing Work (1/3)

- **Development and Deployment Status**

- SODAR is in beta phase, development is ongoing
- The main CUBI SODAR instance is hosted in our private network
- In use for several years in a large number of projects at BIH with collaborators
 - 350TB+ of data stored in iRODS
 - 300+ projects
 - 300+ users

Status and Ongoing Work (2/3)

- SODAR source code and related resources are available under the MIT license at [**github.com/bihealth**](https://github.com/bihealth)
- **sodar-server**
 - The Django server for the main SODAR system, UIs and REST APIs
- **sodar-docker-compose**
 - A Docker Compose network containing all the necessary components for running SODAR
 - For evaluation, development or deploying in production
- And more...



<> Code	Issues	Pull requests	Actions	Projects	Wiki	Security
dev	2 branches	27 tags	Go to file	Code		
This branch is 5 commits ahead of main. Contribute						
mikkonie upgrade to django-sodar-core v0.10.11 final (588f3a 9 hours ago) 1,677 commits						
.github/workflows	upgrade to django-sodar-core v0.10.8 final (#1337), u...					2 months ago
config	cleanup and prepare v0.11.2 release (#1377)					18 days ago
docker	fix build-docker.sh special character handling (#1385)					8 days ago
docs_dev	cleanup and prepare v0.11.2 release (#1377)					18 days ago
docs_manual	fix minor manual issues (#1386, #1387)					8 days ago
irodsadmin	upgrade to django-sodar-core v0.10.6-WIP (#1272, #...					4 months ago
irodsbackend	cleanup and prepare v0.11.1 release (#1350)					2 months ago
irodsinfo	fix irodsinfo tests in ci (#1353)					2 months ago
landingzones	fix zone list layout issues (#1380)					19 days ago
ontologyaccess	fix owl loading from url (#1352)					2 months ago
requirements	upgrade to django-sodar-core v0.10.11 final (#1391)					9 hours ago
samplesheets	upgrade to django-sodar-core v0.10.11-wip (#1391)					yesterday
sodar	link manual to readthedocs, remove local manual buil...					21 days ago
utility	upgrade to django-sodar-core v0.10.9 final (#1361)					last month

Status and Ongoing Work (3/3)

- **Ongoing Work**

- SODAR publication to be submitted
- Publicly available demo server will be launched
- Improved iRODS ticket access support for external software
- Support for study level sample data in iRODS
- Enable easy providing of custom sample sheet templates
- Building towards a feature-complete sample sheet editor
- More command line tooling making use of the APIs
- Upgrade to iRODS 4.3 :)

Conclusions

Conclusions

- **SODAR**

- SODAR is an integral part of CUBI data management
- Major improvements in metadata management and mass storage
- External tooling makes extensive use of the REST APIs in SODAR
- The project has been made publically available
- Development is ongoing

- **Experiences with iRODS**

- IRODS has been used for file storage in SODAR since the beginning
- Used through the Python client, Davrods and iCommands
- Support from iRODS has been very helpful
- We have become a consortium member since the previous presentation

Acknowledgements

- **Collaboration**

- Developers of iRODS, Davrods and the iRODS Python Client
- iRODS support for valuable help
- BIH researchers and collaborators using SODAR for feedback, suggestions, bug reports, etc.

- **CUBI**

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

Thank You

www.cubi.bihealth.org