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
SODAR Data Workflow

- **Sample sheets** contain sample, process and material metadata for project studies
  - Modeled in the ISA-Tab format: [isa-tools.org](http://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
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
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
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
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
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
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 \texttt{github.com/bihealth}

- \texttt{sodar-server}
  - The Django server for the main SODAR system, UIs and REST APIs

- \texttt{sodar-docker-compose}
  - A Docker Compose network containing all the necessary components for running SODAR
  - For evaluation, development or deploying in production

- And more...
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