Implementing a Genomic Data Management System using iRODS at Bayer HealthCare
iRODS User Group Meeting 2015

Carsten Jahn – Bayer Business Services GmbH, R&D IT, HealthCare Research
Navya Dabbiru – Innovations Labs, Tata Consultancy Services, Hyderabad, AP India
Genomic data (e.g. DNA, RNA sequencing) is generated with less cost, massive data amounts need to be managed

- data overview - where is the data of a patient who retracted his study consent?
- directory namespace larger than one file system, e.g. archival of data
- multiple systems for data analysis: Linux commandline based and GUI applications
iRODS at Bayer

iRODS project started April 2014, productive since December 2014

Addressing requirements for

- data security and user management
- metadata import and search
- data transfer automation – sequencer to iRODS
- stable and safe operations

Implementation is supported by Tata Consultancy Services.
Data Ownership

study X
• owner(s)
• members

study owners are enabled to manage their team and grant access rights

general Bayer web application

iRODS

custom replication

iRODS group

other applications

iRODS admins are no “data admins“
Users and Metadata

Study / Project Data Owner

Bioinformatician, Data Analyst

“iRODS? Never heard about...”

“Ok, I can tell you the data has Restricted access.”

“I want to keep track of the tools (and versions) that produced this output.”
Metadata Entry:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>Unit</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>study_type</td>
<td>research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bayer_study_id</td>
<td></td>
<td></td>
<td>mandatory attribute, value is required</td>
</tr>
<tr>
<td>contact_person</td>
<td>Karl Heinz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cross_study_analysis_allowed</td>
<td>within indication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data_classification</td>
<td>Restricted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>external_reference</td>
<td>Secret</td>
<td></td>
<td></td>
</tr>
<tr>
<td>health_status</td>
<td>Internal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Metadata “Schema“ – Validation in Excel and iRODS:

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>mandatory</th>
<th>fixed dictionary</th>
<th>multiple values allowed</th>
<th>intended on study level</th>
<th>intended on project level</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>bayer_study_id</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>study ID within Bayer</td>
</tr>
<tr>
<td>contact_person</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>acting as backup for the data owner</td>
</tr>
<tr>
<td>cross_study_analysis_allowed</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>statement by data owner, if data can be used for cross study</td>
</tr>
<tr>
<td>data_classification</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>bayer data classification</td>
</tr>
<tr>
<td>data_source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&quot;creator&quot; of the data, i.e. external sequencing provider or internal</td>
</tr>
<tr>
<td>data_type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>i.e. unmapped or mapped sequence reads, variant calls, genome</td>
</tr>
<tr>
<td>entry_date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>entry date of data into Bayer</td>
</tr>
</tbody>
</table>
HPC Cluster Integration

- iRODS at Bayer HealthCare
- UGM 2015
HPC Cluster Integration

- Jobs produce metadata but files not registered yet.
- Duplicate access permissions must not overwrite data of colleague.
- Jobs produce metadata but files not registered yet.
Custom Implementations

- Study based custom ACLs
- Hierarchical Inheritance
- Audit Trails
- Bulk Metadata Operations
- Metadata Validation
- Autom. Checksum Generation
- Data Integrity and Consistency
Installing iRODS with puppet

Puppet is a configuration management system that allows to define the state of the IT infrastructure, automates every step of the software delivery process.

- Automated custom installation
- Supports multiple environments
- Easy to discard and rebuild installation on demand.
- Open topics:
  - support for version upgrades
  - security updates, patches

Hiera hosts yaml config

ICAT host

Zone name: Resource name: ....

Common

resource

Role: //functionality
Profile //functionality

iRODS

OS type
Database type
DB Auth type
Version
Auth type - Basic/ PAM - SSL
Technical Recommendations for Introducing iRODS

- involve Linux and storage admins early on
- plan effort for testing your installation
  your version of Linux and PostgreSQL, your iRODS rules, your expectations
  - document test procedures
  - deploy three similar environments: “development” / tryout system, acceptance test system, production system
- API selection
  evaluate all use case before deciding for programming platform and API
Organizational Recommendations for Introducing iRODS

- get involved with iRODS support or the user community
- save effort by incorporating external knowhow
- iRODS manual and help pages cover many aspects, but not all

- prepare a metadata and data access concept

- Metadata collection is a challenging task...

- user training – iRODS is not a distributed file system
  (e.g. replication – same file name, different content is possible)
Questions & Answers
Contact

**Carsten Jahn**  
Bayer Business Services  
**Phone:** +49 – 30 468 12837  
**E-mail:** carsten.jahn@bayer.com

**Navya Dabbiru**  
Tata Consultancy Services  
**Phone:** +49 – 17680892216  
**E-mail:** navyaanantayashasri.dabbiru@bayer.com
Forward-Looking Statements

This presentation may contain forward-looking statements based on current assumptions and forecasts made by Bayer Group or subgroup management.

Various known and unknown risks, uncertainties and other factors could lead to material differences between the actual future results, financial situation, development or performance of the company and the estimates given here. These factors include those discussed in Bayer’s public reports which are available on the Bayer website at www.bayer.com.

The company assumes no liability whatsoever to update these forward-looking statements or to conform them to future events or developments.
Image Credits

Slide “Improving Data Management”
Kinghorn Centre for Clinical Genomics, HSXcutout
https://creativecommons.org/licenses/by-nd/2.0/
Uwe Hermann, Organized
https://creativecommons.org/licenses/by-sa/2.0/
Greg Emmerich, Vector DNA
https://creativecommons.org/licenses/by-sa/2.0/
Thank you!