FAIR Sequencing Data Repository based on iRODS

Felipe O. Gutierrez

AMC - Academic Medical Center - Amsterdam, Netherlands

A.C.Camargo Cancer Center - São Paulo, Brazil

F. Oliveira Gutierrez

P.F.G. De Geest

Aldo Jongejan

Sjoerd Repping

Diogo Ferreira Patrão

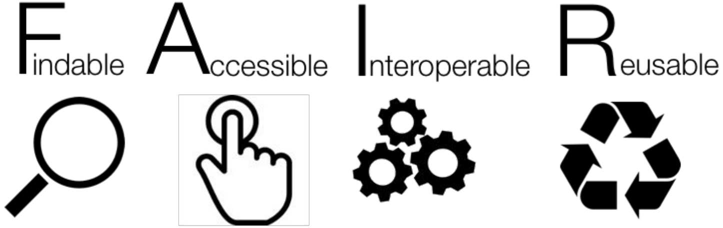
A.H.C. van Kampen

J.T. van den Berg

Silvia D. Olabarriaga

Problem

- Inadequate RDM (Research Data Management) solution for NGS data (Next Generation Sequencing):
  - Individual storage and backup
  - Dispersed datasets
  - Disconnected from metadata
  - Not FAIR



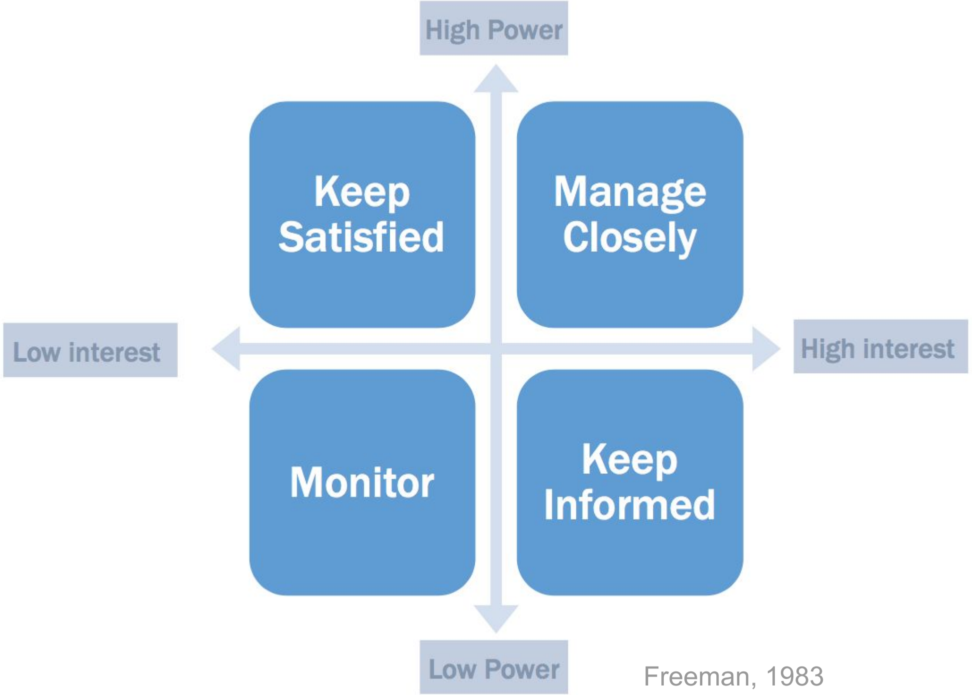
Considerations

Fit within organization

- ICT culture
- Research culture
- Sustainability vision

Adhere to international community best practices

Reuse and extend existing solutions



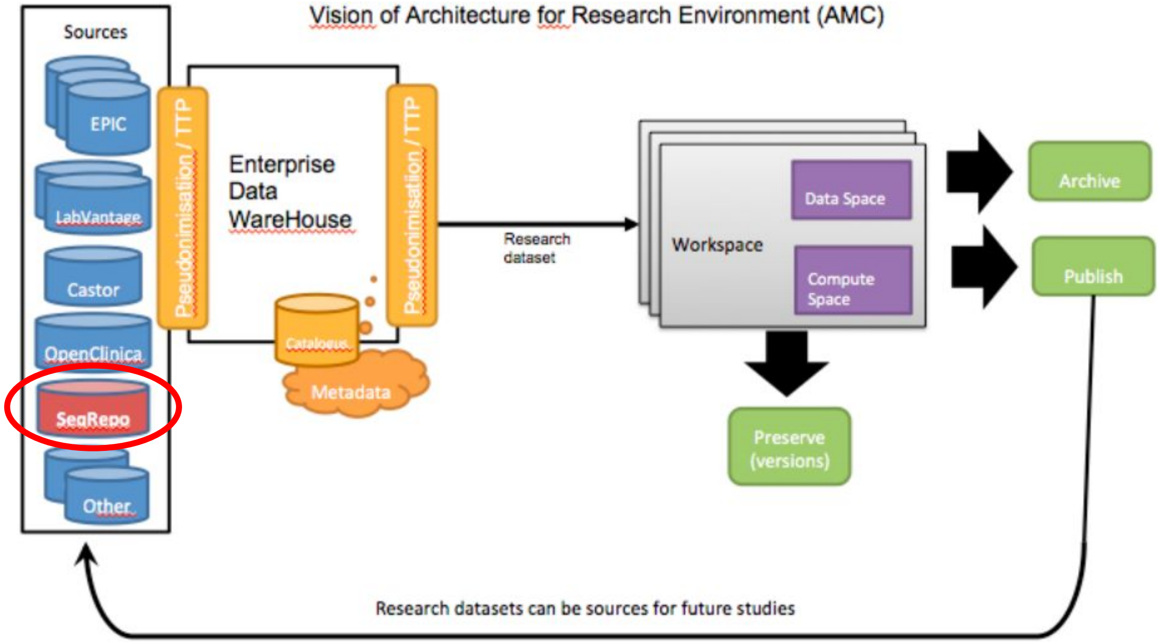
Freeman, 1983

Fit into AMC Vision for RDM

Based on NFU Data4Lifesciences WP2

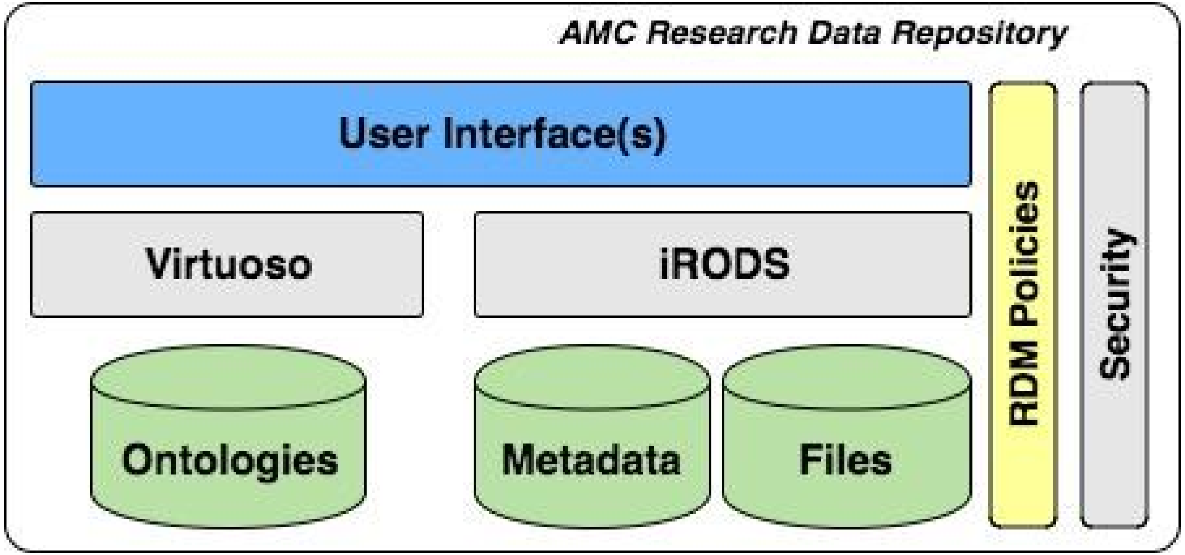
An NGS repository that is:

- Part of an ecosystem
- Controlled by AMC
- Distributed
- Scalable
- FAIR compliant
- Easy to use

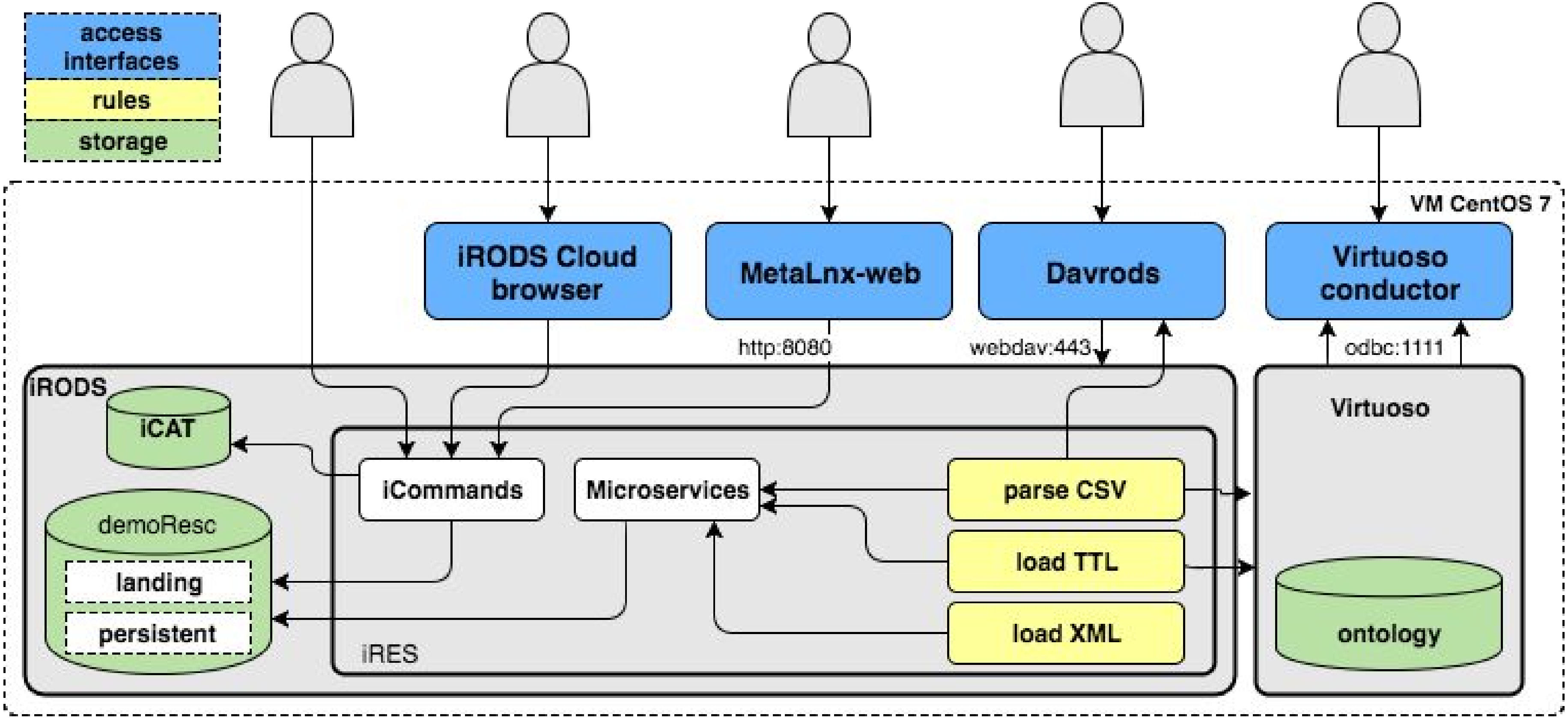


System Design

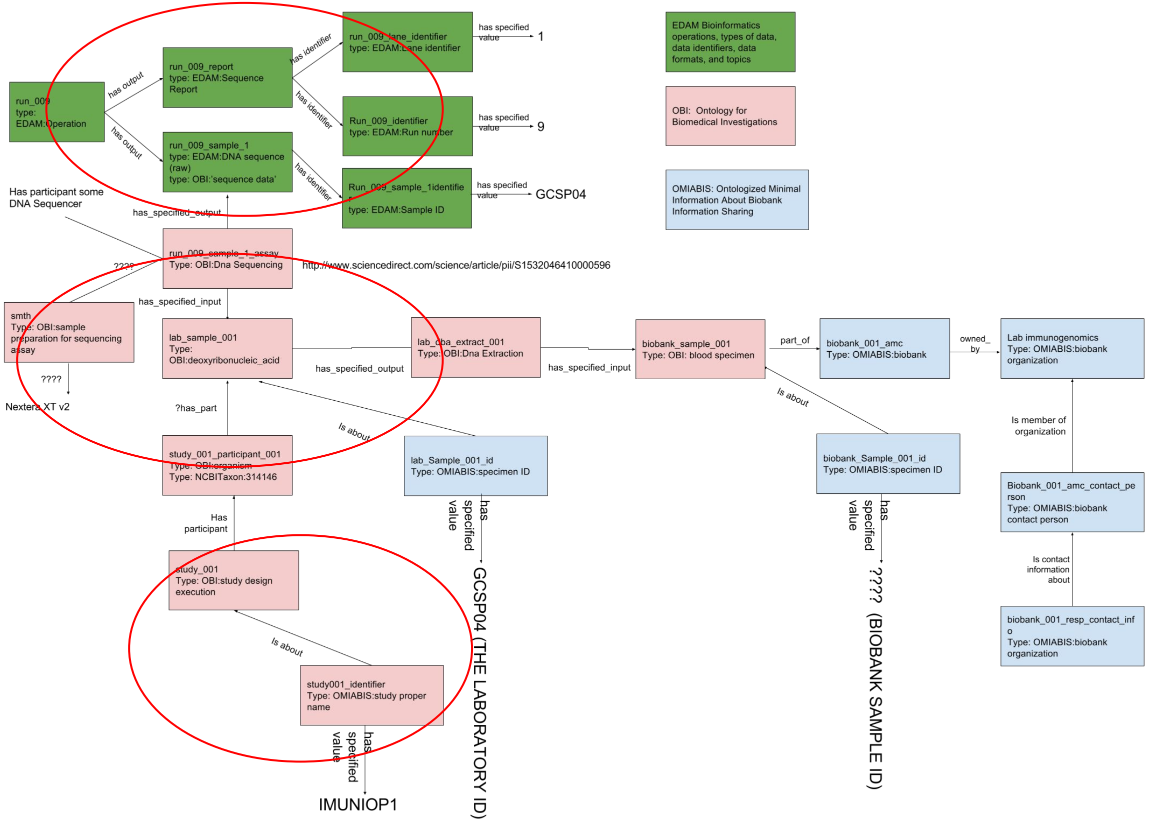
- iRODS 4.1.10
  - Middleware
  - Data virtualization
- Virtuoso 7.2
  - Triplestore
  - Supports ontologies
- User interfaces:
  - Metalnx web
  - Davrods 4.1
  - iCommands



System Architecture



Stewardship: Ontologies



run_009
type:
EDAM:Operation

has output

run_009_report
type: EDAM:Sequence
Report

has output

run_009_sample_1
type: EDAM:DNA sequence
(raw)
type: OBI:'sequence data'

has identifier

run_009_lane_identifier
type: EDAM:Lane identifier

has specified
value

1

has identifier

Run_009_identifier
type: EDAM:Run number

has specified
value

9

has identifier

Run_009_sample_1identifie
r
type: EDAM:Sample ID

has specified
value

GCSP04

EDAM Bioinformatics
operations, types of data,
data identifiers, data
formats, and topics

OBI: Ontology for
Biomedical Investigations

OMIABIS: Ontologized Minimal
Information About Biobank
Information Sharing

Has participant some
DNA Sequencer

has_specified_output

run_009_sample_1_assay
Type: OBI:Dna Sequencing

????

http://www.sciencedirect.com/science/article/pii/S1532046410000596

has_specified_input

smth
Type: OBI:sample
preparation for sequencing
assay

????

Nextera XT v2

lab_sample_001
Type:
OBI:deoxyribonucleic_acid

has_specified_output

lab_dna_extract_001
Type: OBI:Dna Extraction

has_specified_input

biobank_sample_001
Type: OBI: blood specimen

part_of

biobank_001_amc
Type: OMIABIS:biobank

owned_
by

Lab immunogenomics
Type: OMIABIS:biobank
organization

?has_part

Is about

study_001_participant_001
Type: OBI:organism
Type: NCBITaxon:314146

lab_Sample_001_id
Type: OMIABIS:specimen ID

has
specified
value

GCSP04 (THE LABORATORY ID)

Is about

biobank_Sample_001_id
Type: OMIABIS:specimen ID

has
specified
value

???? (BIOBANK SAMPLE ID)

Is member of
organization

Biobank_001_amc_contact_pe
rson
Type: OMIABIS:biobank
contact person

Is contact
information
about

biobank_001_resp_contact_inf
o
Type: OMIABIS:biobank
organization

Has
participant

study_001
Type: OBI:study design
execution

Is about

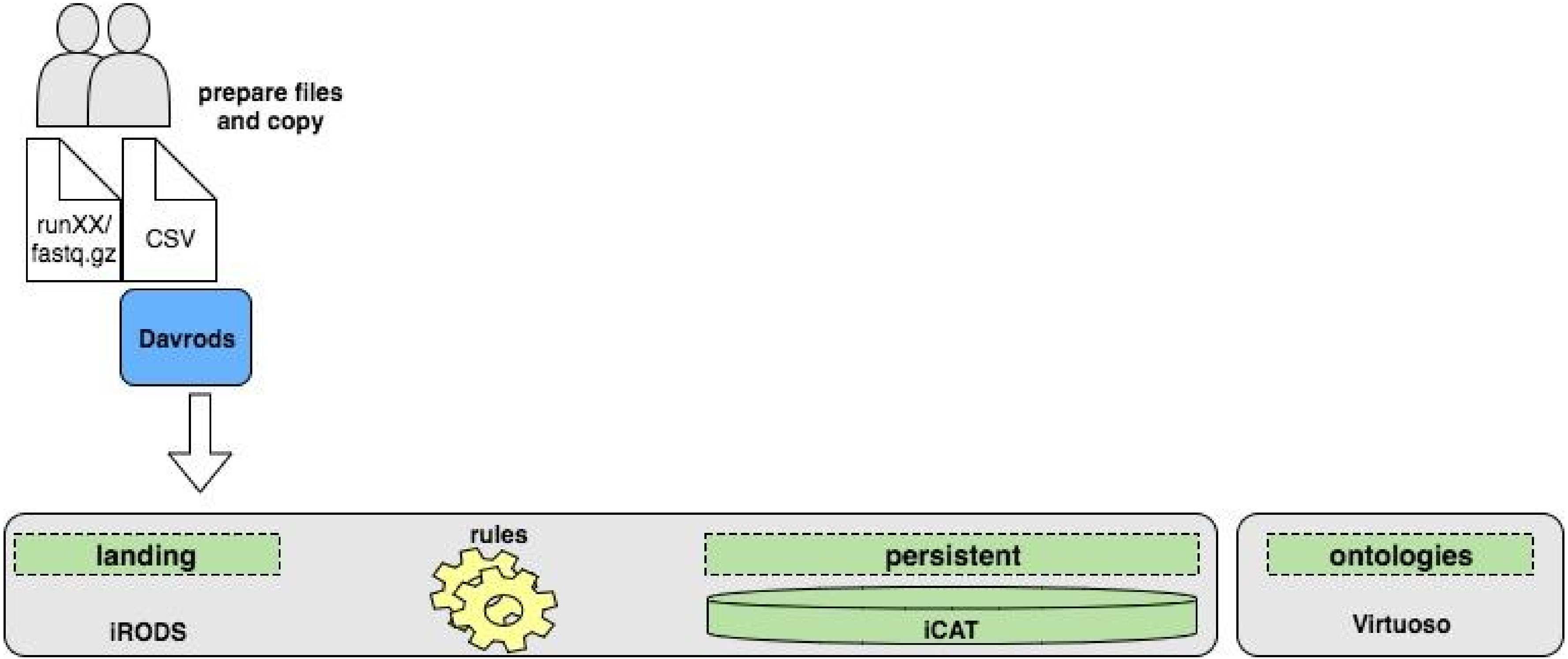
study001_identifier
Type: OMIABIS:study proper
name

has
specified
value

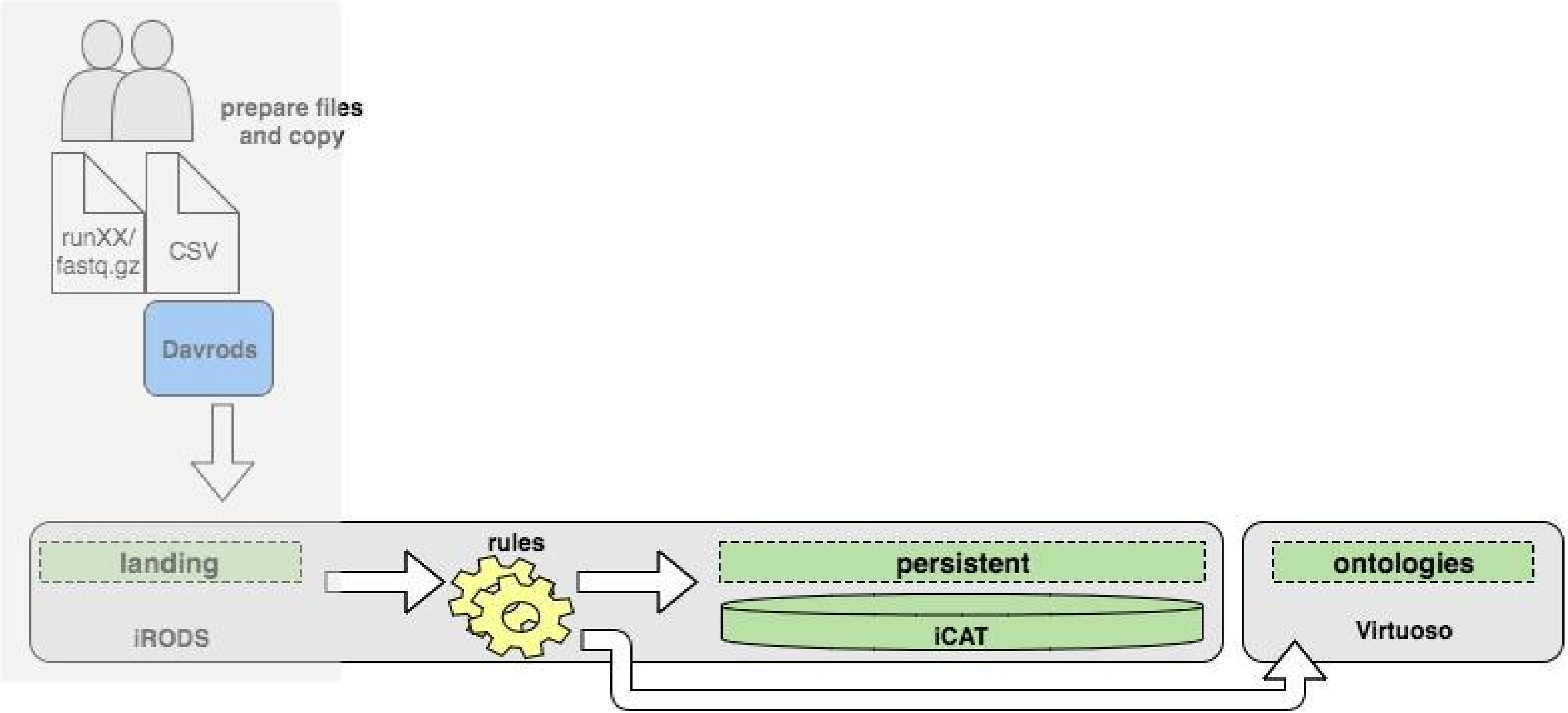
IMUNIOP1

- **EDAM**
  Ontology for bioinformatics operations, types of data, data identifiers, data formats, and topics
- **OMIABIS**
  Ontologized Minimum Information About Biobank data Sharing (MIABIS)
- **OBI**
  Ontology for Biomedical Investigations
- **EFO**
  Experimental Factor Ontology

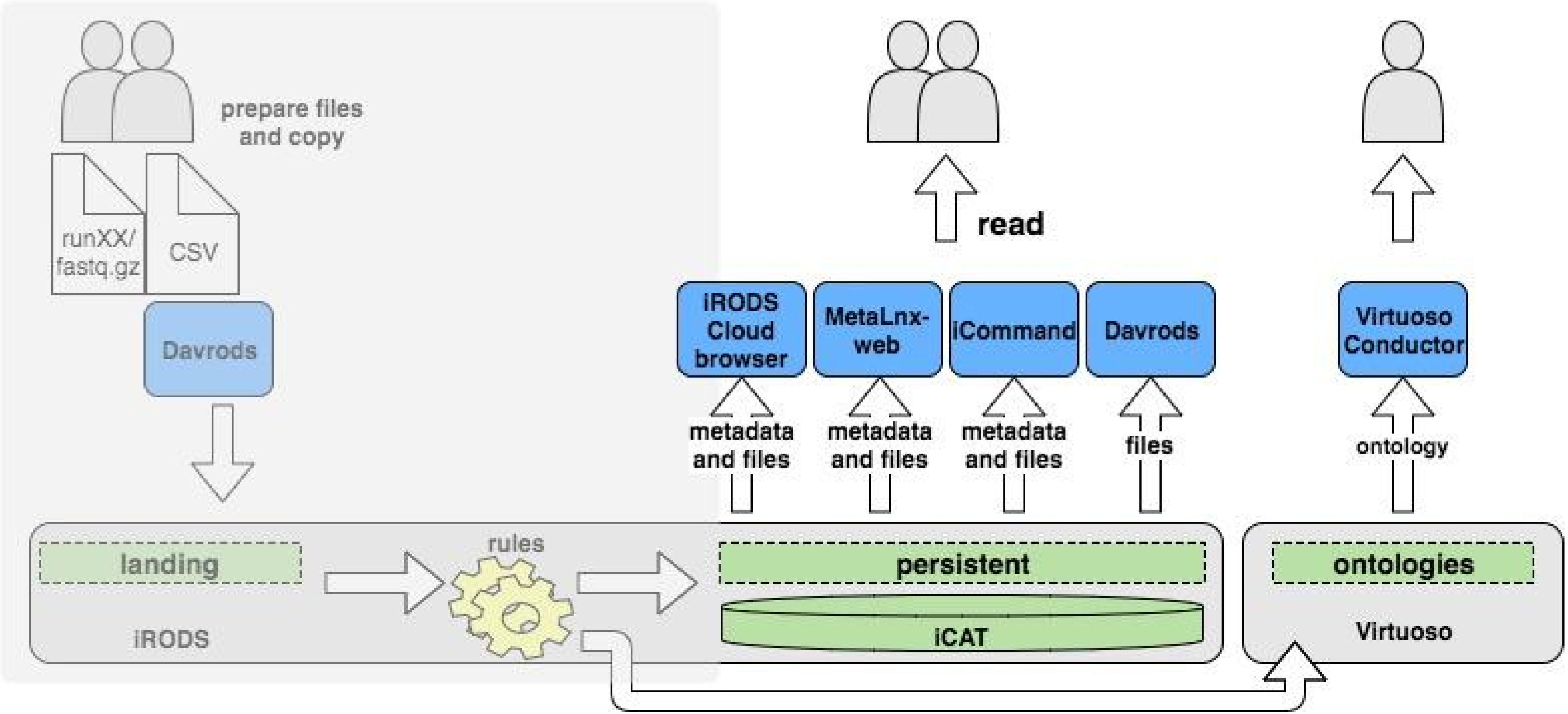
Workflow: Data Ingestion



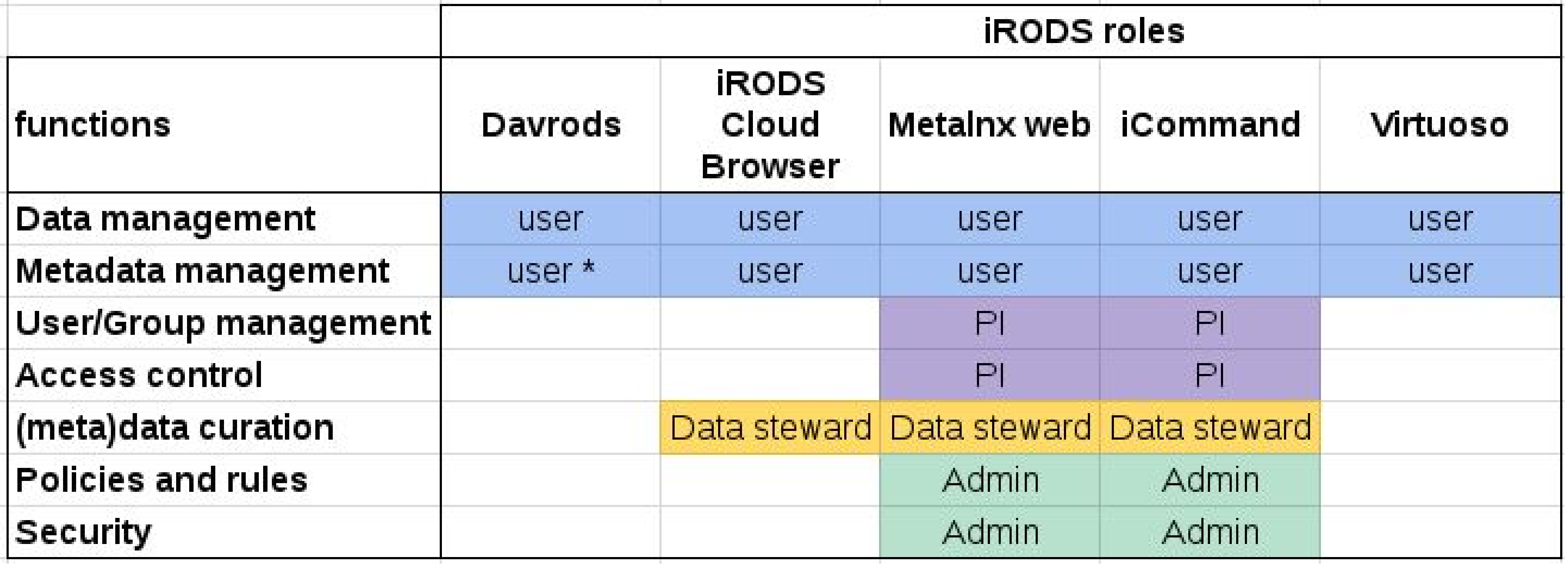
Workflow: (meta)data Registration



Workflow: (meta)data Retrieval

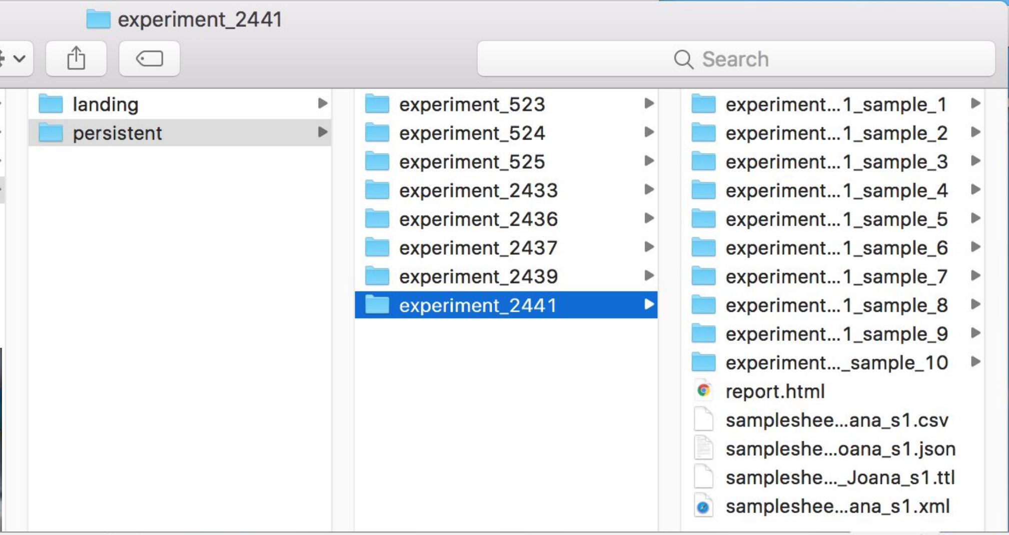


Access and Security



	iRODS roles				
functions	**Davrods**	**iRODS Cloud Browser**	**Metalnx web**	**iCommand**	**Virtuoso**
Data management	user	user	user	user	user
Metadata management	user *	user	user	user	user
User/Group management			PI	PI	
Access control			PI	PI	
(meta)data curation		Data steward	Data steward	Data steward	
Policies and rules			Admin	Admin	
Security			Admin	Admin	

Status



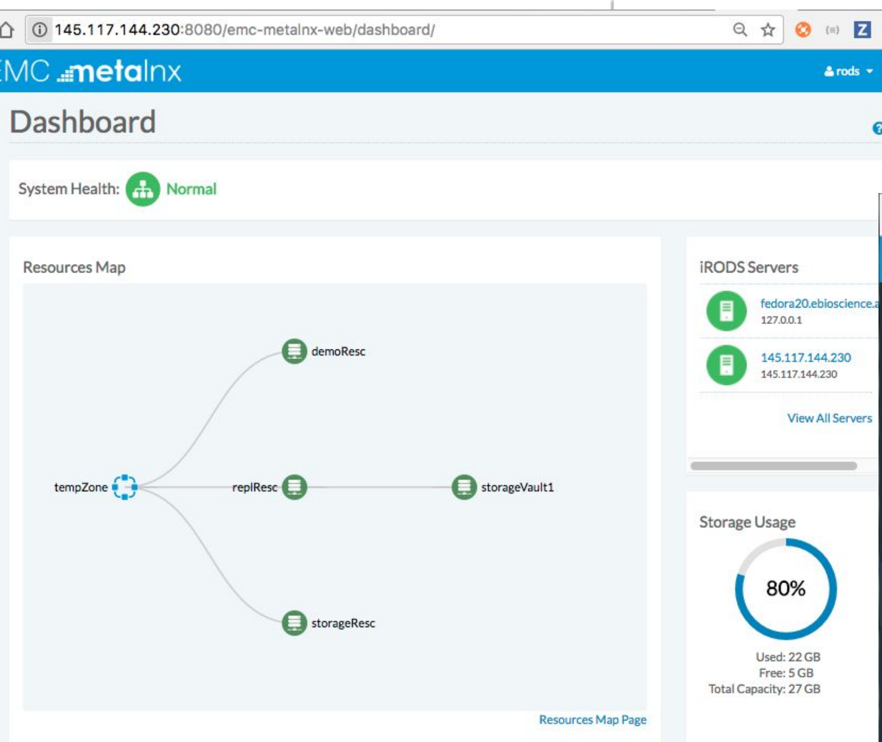
experiment_2441

Macintosh HD
Network
Remote Disc
ugp-repm...ce.amc.nl

landing
persistent

experiment_523
experiment_524
experiment_525
experiment_2433
experiment_2436
experiment_2437
experiment_2439
experiment_2441

experiment...1_sample_1
experiment...1_sample_2
experiment...1_sample_3
experiment...1_sample_4
experiment...1_sample_5
experiment...1_sample_6
experiment...1_sample_7
experiment...1_sample_8
experiment...1_sample_9
experiment..._sample_10
report.html
samplesshee...ana_s1.csv
sampleshe...oana_s1.json
sampleshe..._Joana_s1.ttl
sampleshee...ana_s1.xml



145.117.144.230:8080/emc-metalnx-web/dashboard/

Dashboard

System Health: Normal

Resources Map

iRODS Servers

fedora20.ebioscience.a
127.0.0.1

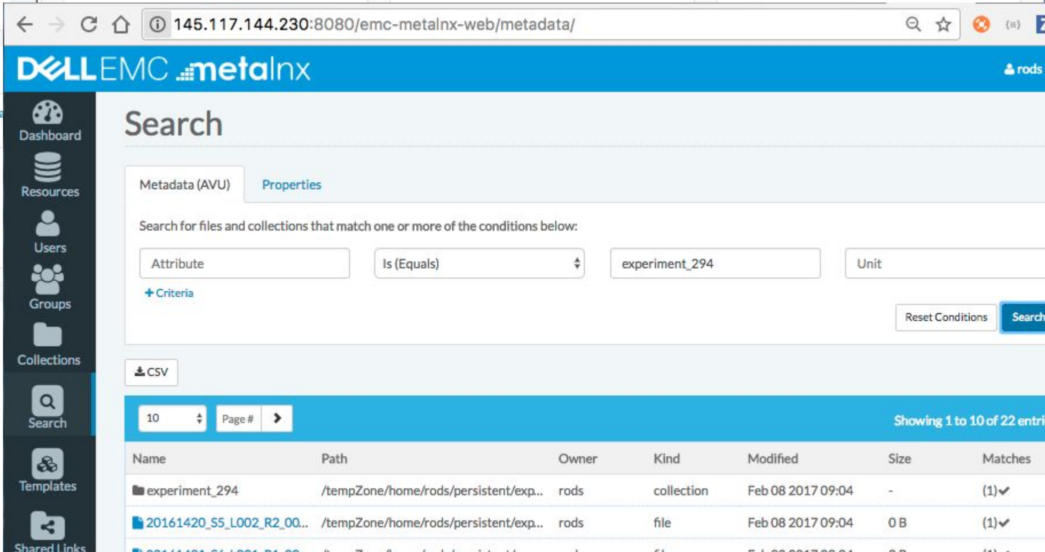
145.117.144.230
145.117.144.230

View All Servers

Storage Usage

80%

Used: 22 GB
Free: 5 GB
Total Capacity: 27 GB



145.117.144.230:8080/emc-metalnx-web/metadata/

Search

Metadata (AVU) | Properties

Search for files and collections that match one or more of the conditions below:

Attribute | Is (Equals) | experiment_294 | Unit

+ Criteria

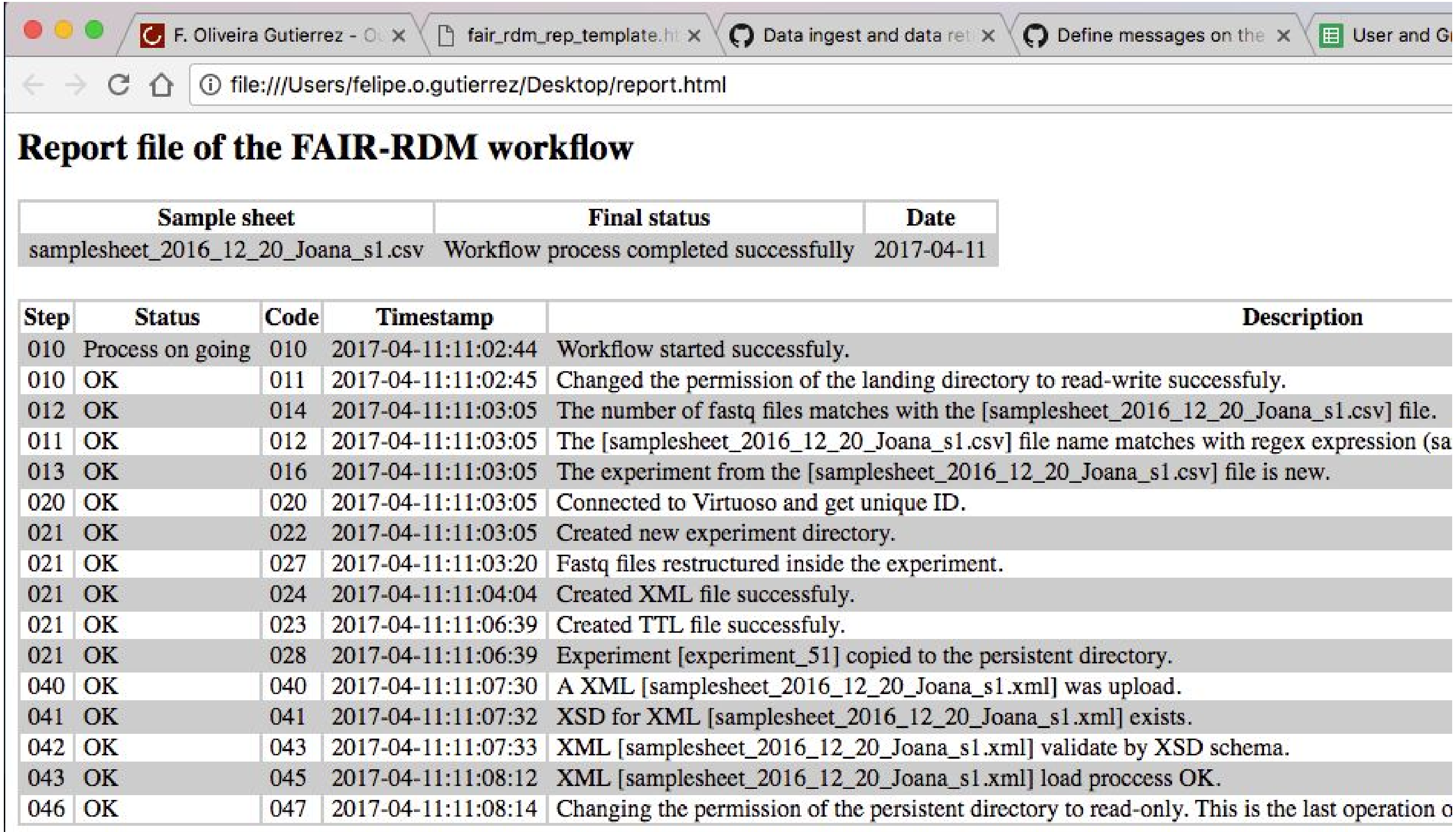
Reset Conditions | Search

CSV

10 | Page # | Showing 1 to 10 of 22 entri

Name	Path	Owner	Kind	Modified	Size	Matches
experiment_294	/tempZone/home/rods/persistent/exp...	rods	collection	Feb 08 2017 09:04	-	(1)✓
20161420_S5_L002_R2_00...	/tempZone/home/rods/persistent/exp...	rods	file	Feb 08 2017 09:04	0 B	(1)✓

Report file

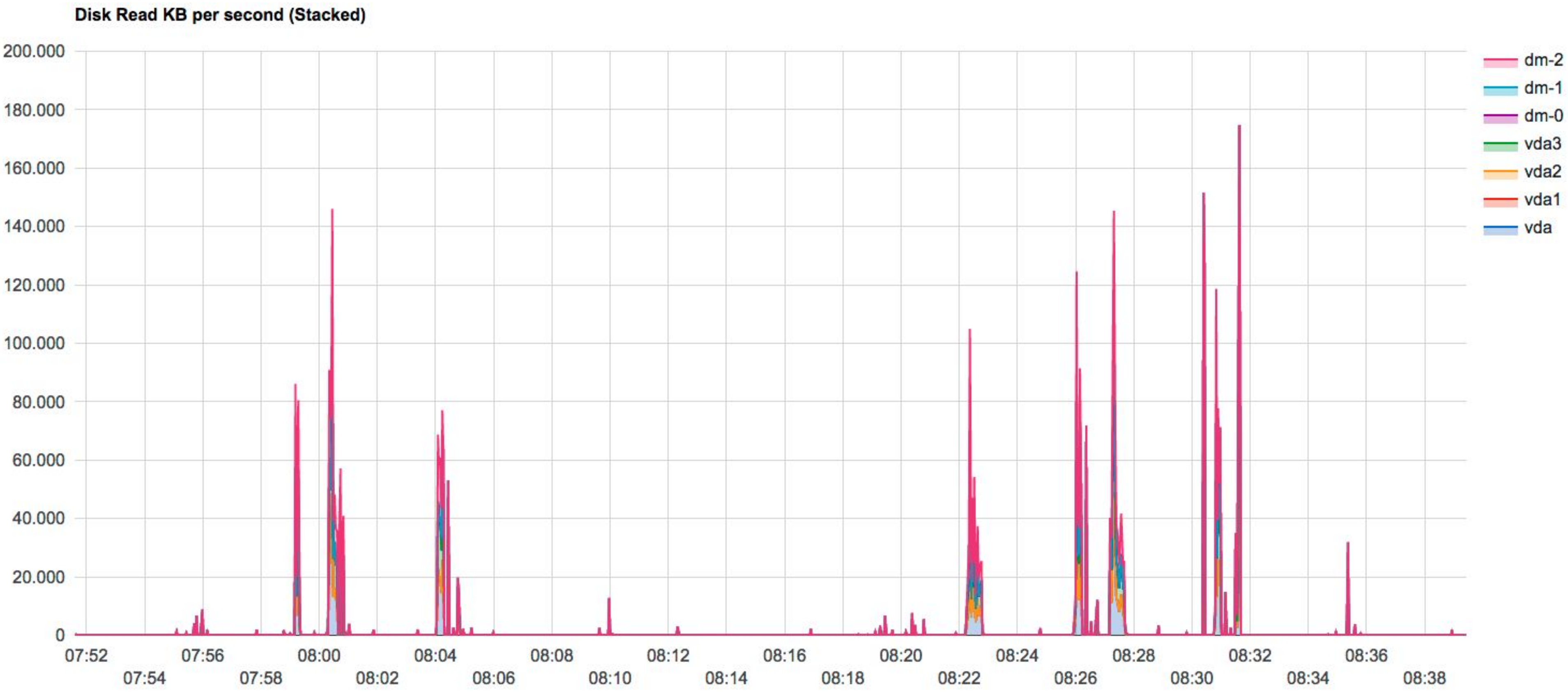


Report file of the FAIR-RDM workflow

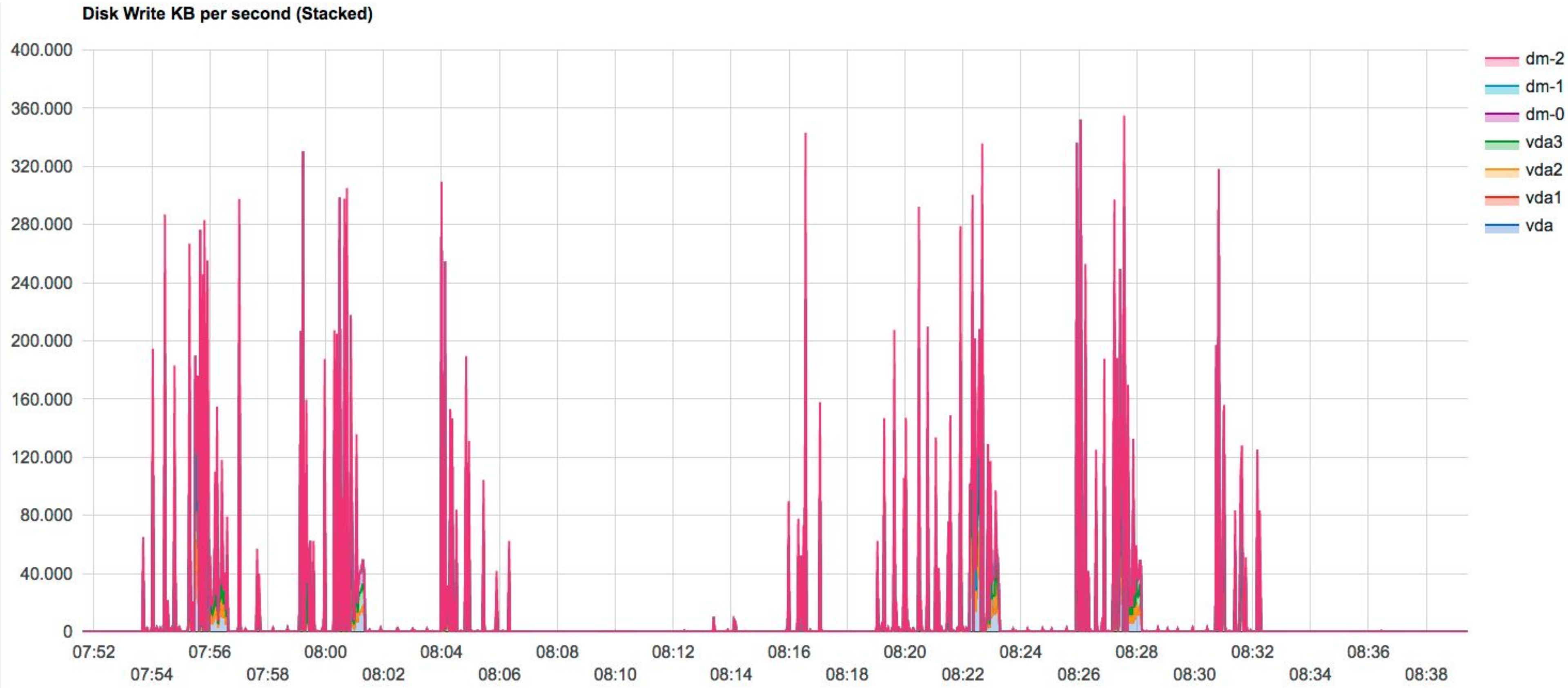
Sample sheet	Final status	Date
samplesheet_2016_12_20_Joana_s1.csv	Workflow process completed successfully	2017-04-11

Step	Status	Code	Timestamp	Description
010	Process on going	010	2017-04-11:11:02:44	Workflow started successfuly.
010	OK	011	2017-04-11:11:02:45	Changed the permission of the landing directory to read-write successfuly.
012	OK	014	2017-04-11:11:03:05	The number of fastq files matches with the [samplesheet_2016_12_20_Joana_s1.csv] file.
011	OK	012	2017-04-11:11:03:05	The [samplesheet_2016_12_20_Joana_s1.csv] file name matches with regex expression (sa
013	OK	016	2017-04-11:11:03:05	The experiment from the [samplesheet_2016_12_20_Joana_s1.csv] file is new.
020	OK	020	2017-04-11:11:03:05	Connected to Virtuoso and get unique ID.
021	OK	022	2017-04-11:11:03:05	Created new experiment directory.
021	OK	027	2017-04-11:11:03:20	Fastq files restructured inside the experiment.
021	OK	024	2017-04-11:11:04:04	Created XML file successfuly.
021	OK	023	2017-04-11:11:06:39	Created TTL file successfuly.
021	OK	028	2017-04-11:11:06:39	Experiment [experiment_51] copied to the persistent directory.
040	OK	040	2017-04-11:11:07:30	A XML [samplesheet_2016_12_20_Joana_s1.xml] was upload.
041	OK	041	2017-04-11:11:07:32	XSD for XML [samplesheet_2016_12_20_Joana_s1.xml] exists.
042	OK	043	2017-04-11:11:07:33	XML [samplesheet_2016_12_20_Joana_s1.xml] validate by XSD schema.
043	OK	045	2017-04-11:11:08:12	XML [samplesheet_2016_12_20_Joana_s1.xml] load proccess OK.
046	OK	047	2017-04-11:11:08:14	Changing the permission of the persistent directory to read-only. This is the last operation o

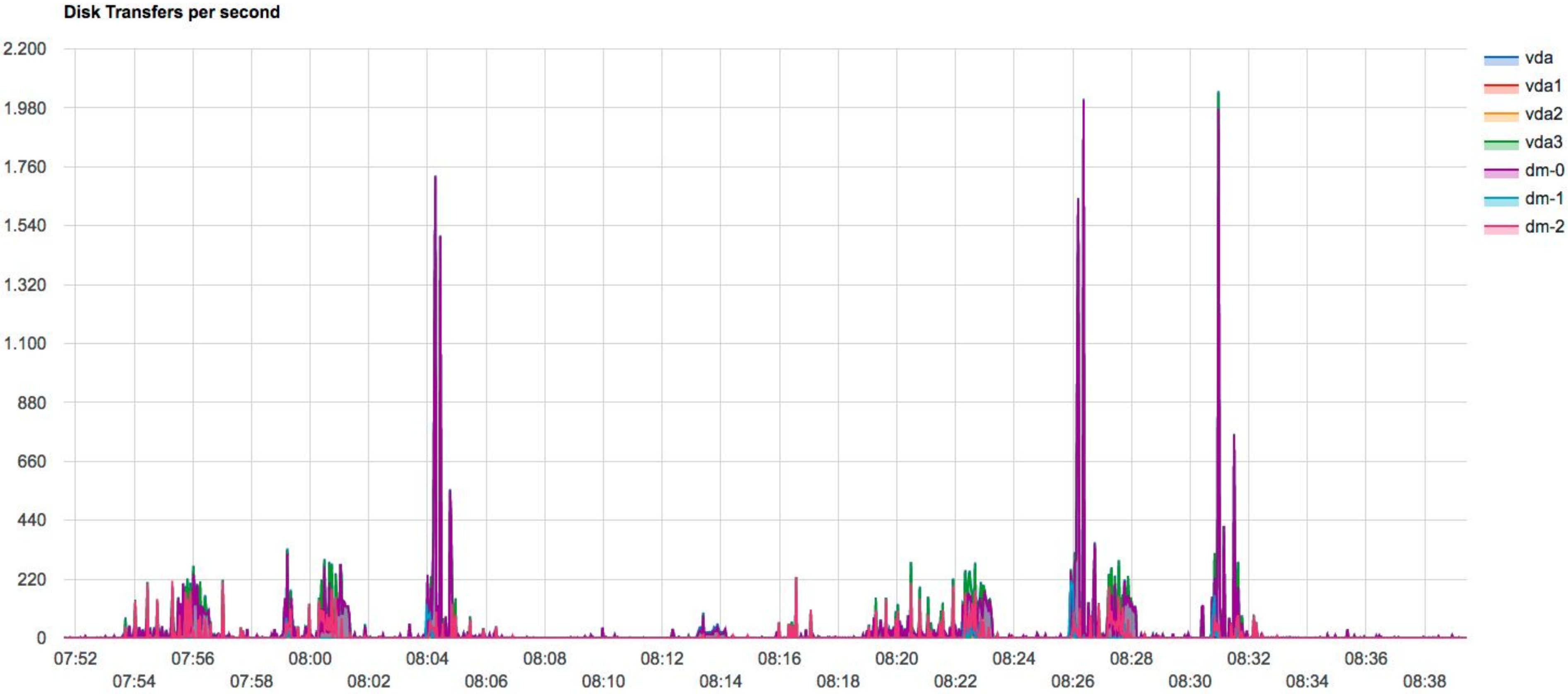
nmon read KB/s



nmon write KB/s



nmon IOPs



Disk Transfers per second
2.200
1.980
1.760
1.540
1.320
1.100
880
660
440
220
0
07:52
07:54
07:56
07:58
08:00
08:02
08:04
08:06
08:08
08:10
08:12
08:14
08:16
08:18
08:20
08:22
08:24
08:26
08:28
08:30
08:32
08:34
08:36
08:38
vda
vda1
vda2
vda3
dm-0
dm-1
dm-2

Qualitative & Quantitative questions

- (meta)data preparation? Clear, doable, easy, ...
- (meta)data upload? Type, size, quantity, integrity, ...
- Rule processing? Report file clear and easy, system delay feedback, ...
- (meta)data retrieval? Findable, Accessible, Organized, Interoperable, Reusable, ..
- Concurrent users, variation on the number and size of files.

Acknowledgements

KEBB:

- Barbera van Schaik
- Allard van Altena

ADICT: Hans van den Berg
UvA ICTS: Joyce Nijkamp

Medical Library: Lieuwe Kool
Clinical Research Unit: Rudy Scholte

Reproductive medicine: Sjoerd Repping
Genetic Metabolic Diseases: Frédéric Vaz
Immunogenomics: Niek de Vries