National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

iRODS-based system turbocharged next-gen sequencing analysis during pandemic and beyond

RIVM: the Dutch National Institute for Public Health and the Environment

- Living in a safe, clean, healthy environment
- Preventing and controlling infectious diseases
- Good healthcare and a healthy lifestyle

Bioinformatics and Computational Services

*Bioinformatics, as related to genetics and genomics, is a scientific subdiscipline that involves using computer technology to **collect**, **store**, **analyze** and **disseminate** biological data and information, such as DNA and amino acid sequences or annotations about those sequences.*

Research environment for scientific employees/researchers and the support of the environment.

Data management system for unstructured data.

Content

- History, why did we start?
- Challenges, bioinformatics environment.
- (some) Products: Intorods, Jobengine, BioRODS, Archiving SURF.
- 2021: Amount of analyses of the previous year in half a week.
- Success stories.
- Future plans.

History, why did we start?

Bioinformatics is used in more and more places, also at the National Institute for Public Health and the Environment (RIVM).

Examples:

- the spread of pathogens
- (genetic) population research
- effects of lifestyle choices
- the origin of chronic diseases
- consequences of changes in the living environment

Challenges, bioinformatics platform

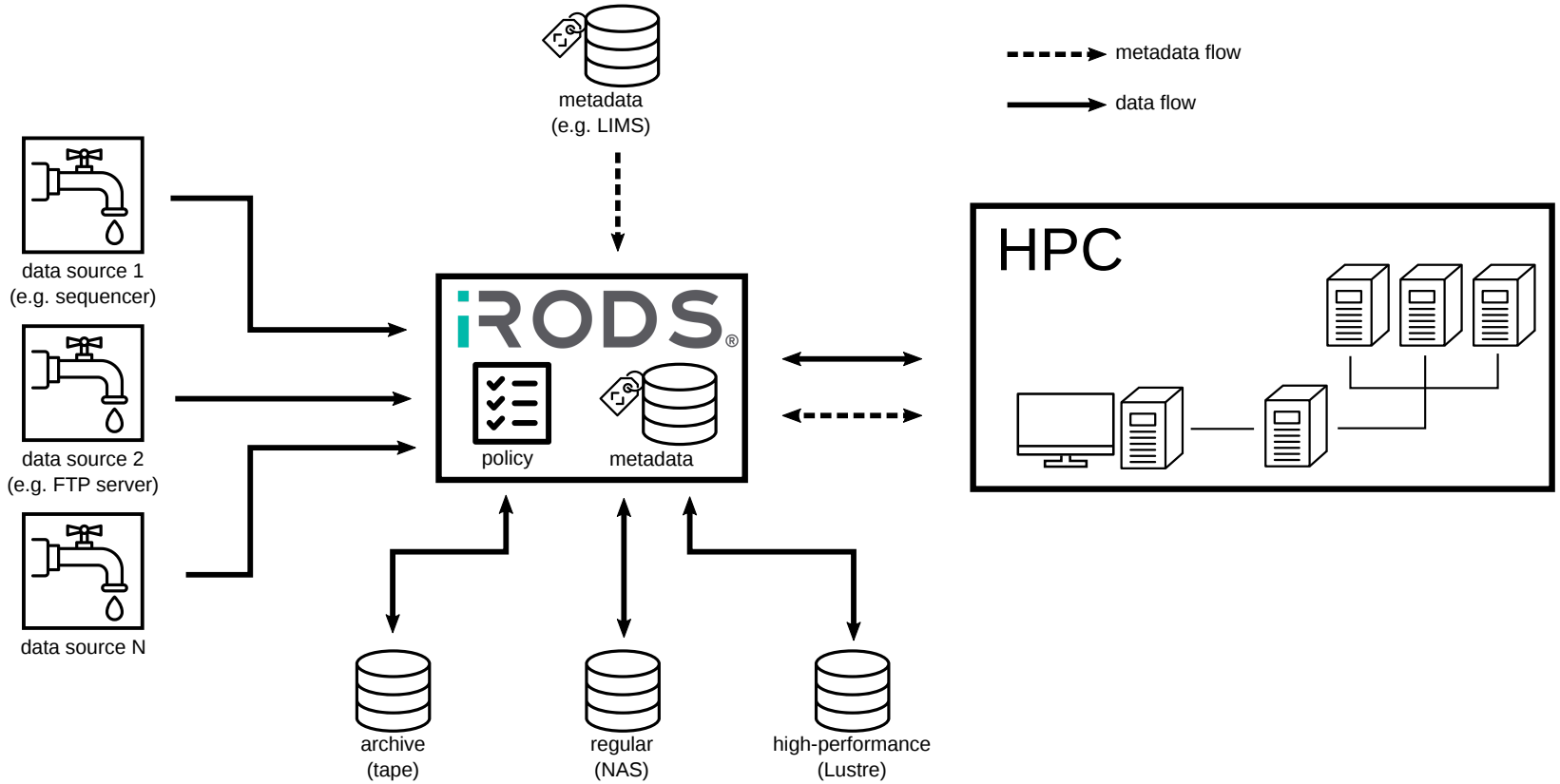
Challenges:

- Large amount of data, data should be FAIR.
- Lots of computational resources necessary to analyze the data.
- Reproducible data analyses.
- A platform where people can develop and share pipelines.

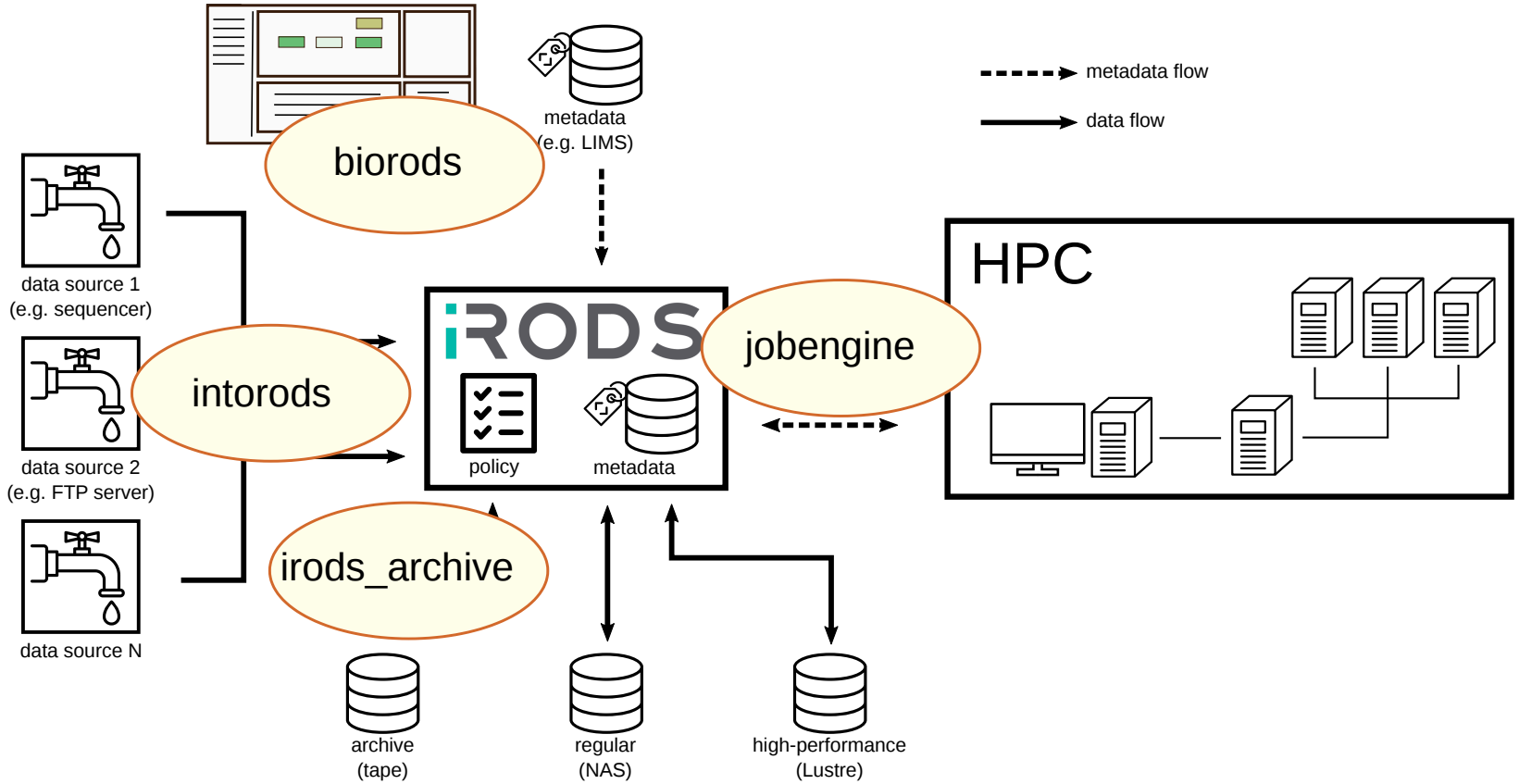
Bioinformatics platform

1. Shared Linux environment, cooperation for creating bioinformatics pipelines.
2. High performance computing (HPC) cluster, fast Lustre storage.
3. Data management system (DMS) for organizing the data flow.
4. Process automation tool to automatically start pipelines.
5. Courses (git, Linux, HPC, Snakemake) for using the environment.

RIVM iRODS architecture



In-house development



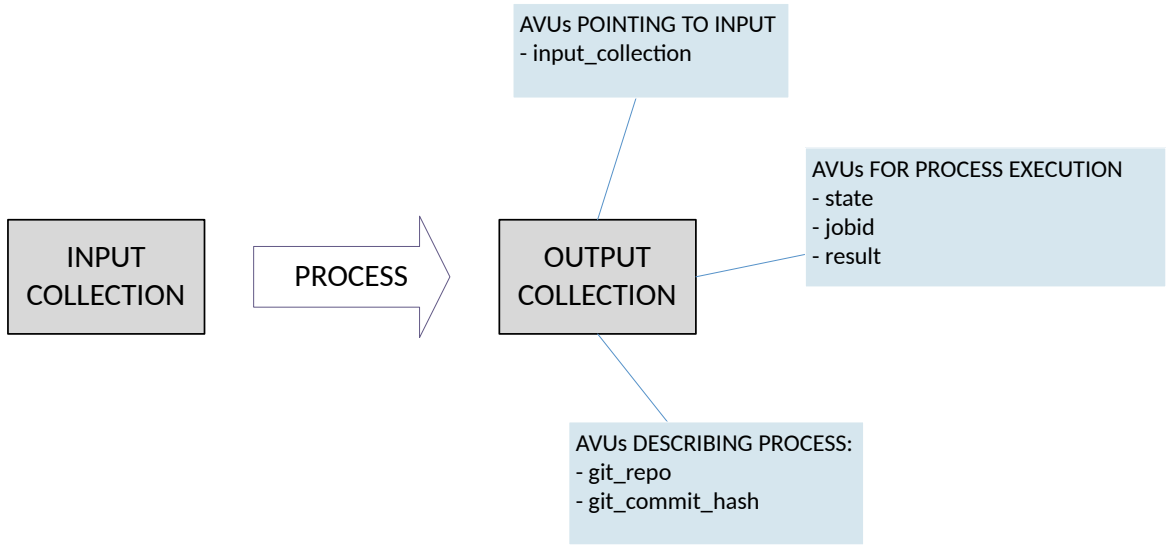
Intorods: iRODS data import tool

- multiprocess parallel data import
- reliable through use of checksum verification
- multiple import sources:
  - local file system
  - other irods instance
  - SMB
  - scp
  - (s)ftp

https://github.com/rivm-syso/intorods

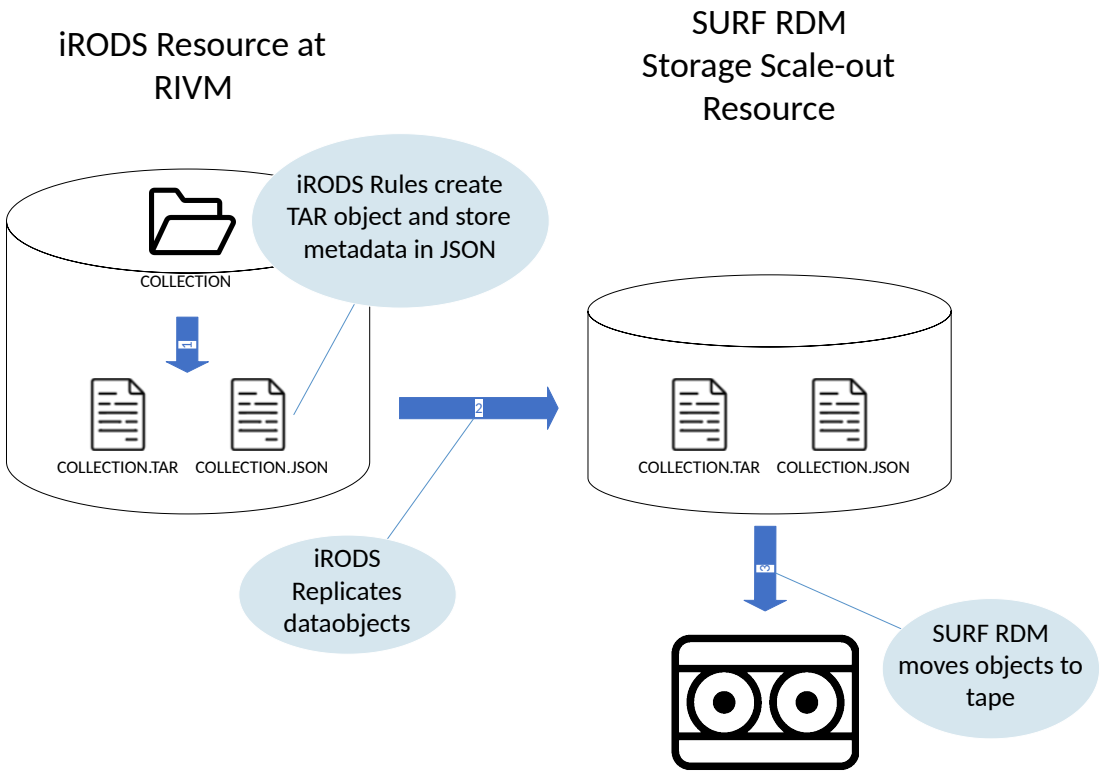
Jobengine: Automation of analysis processes

- Used to run *pipeline code* on datasets stored in iRODS
- Controlled by *metadata* on the output collection of the pipeline process
- Starts and monitors tasks on the compute cluster
- Can handle a single pipeline run, or a group of pipelines called a *process group*

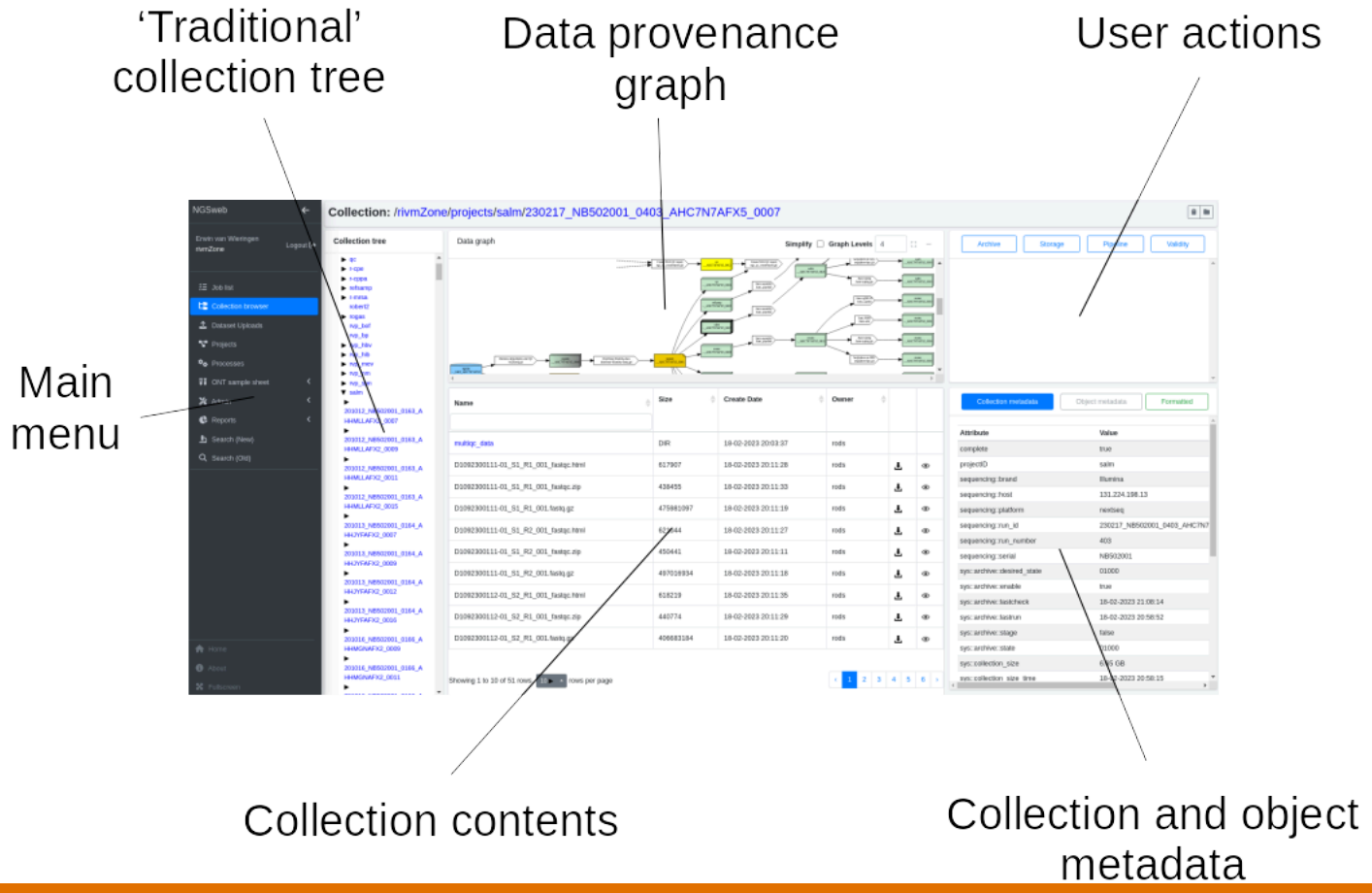


Archiving: SURF RDM Storage Scale-out

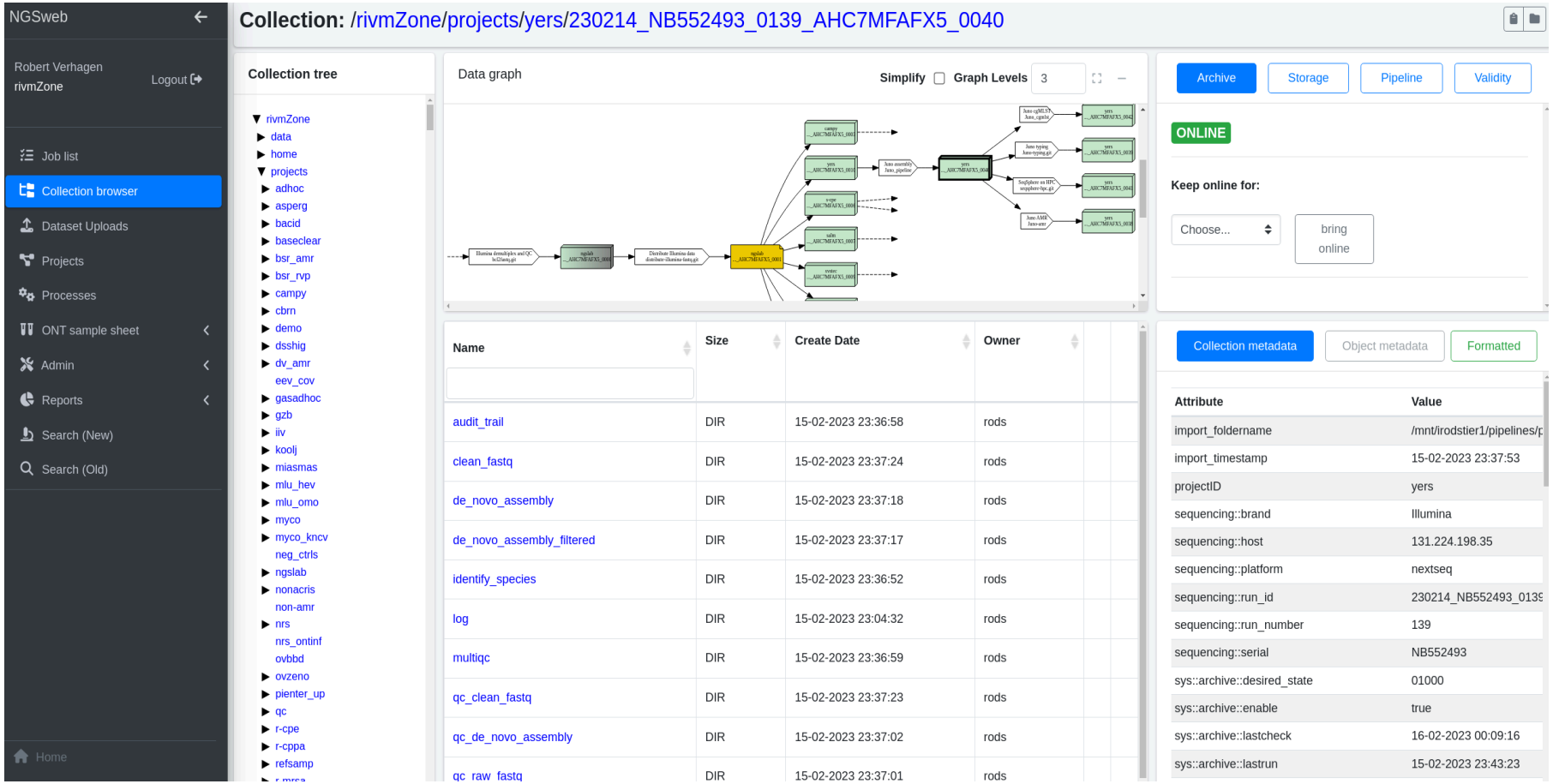
- **Research Data Management Storage Scale-out** using SURF iRODS Consumer.
- iRODS rules to archive collections



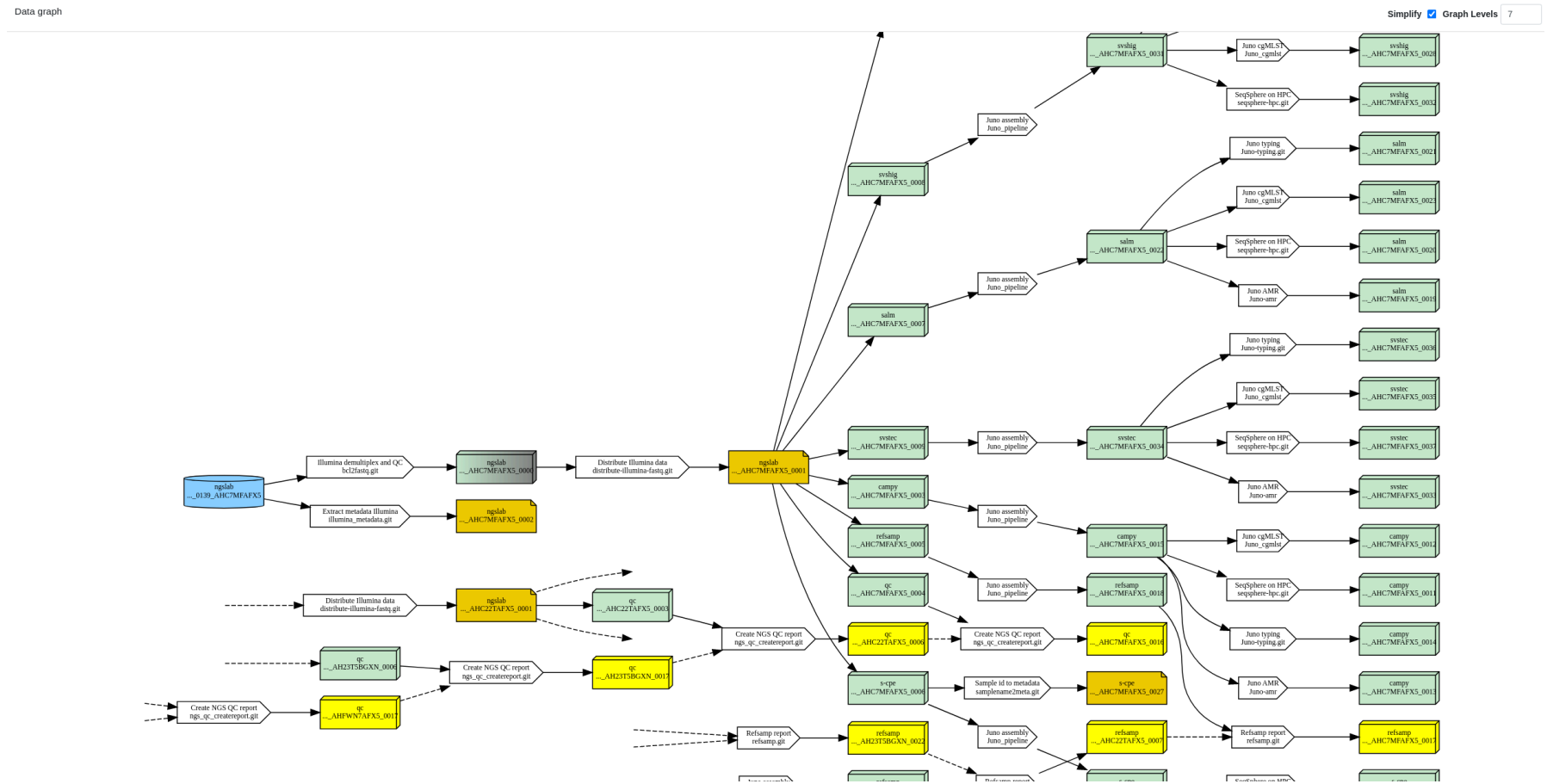
BioRODS: Web interface for users and admins



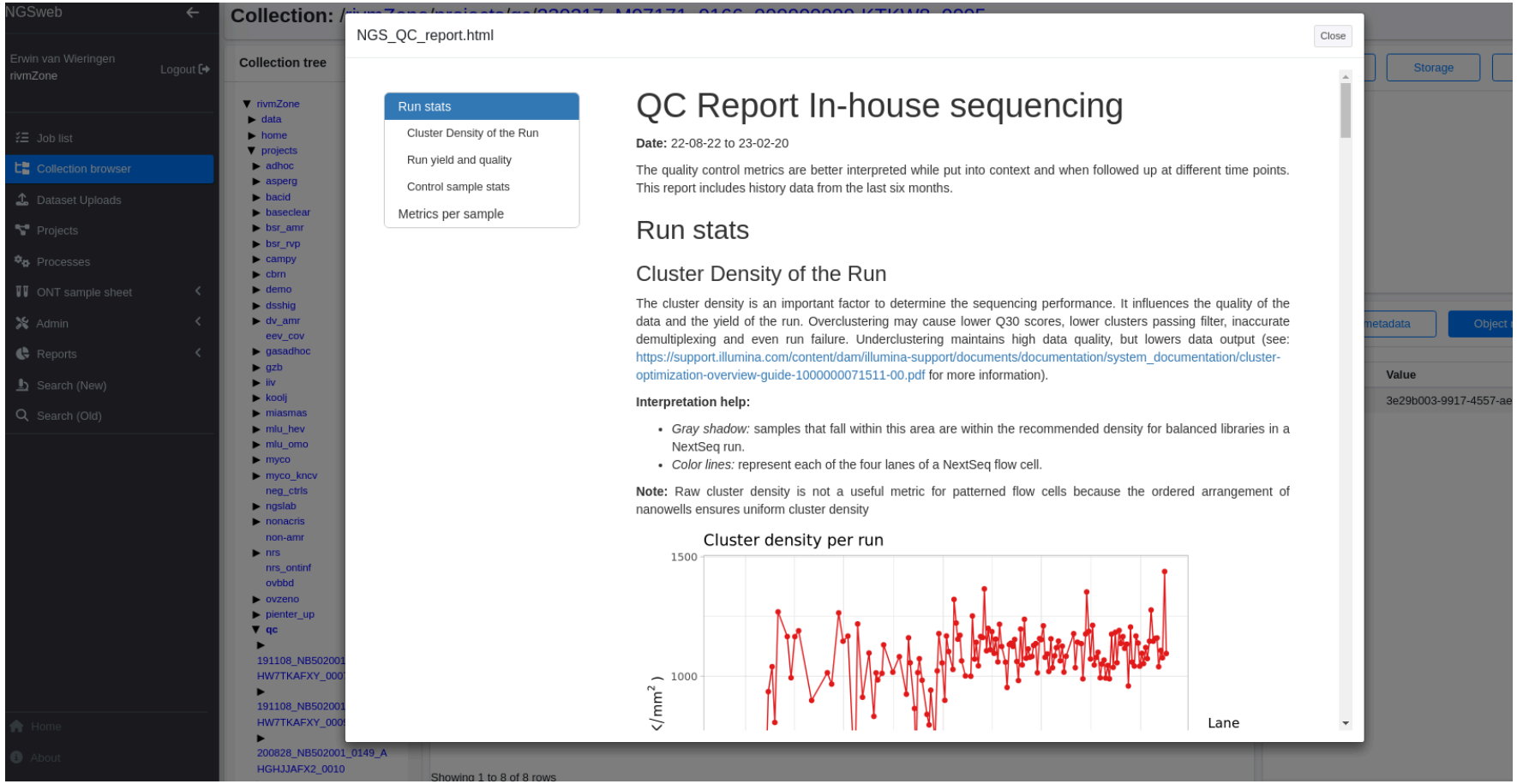
Collection browser (1/3)



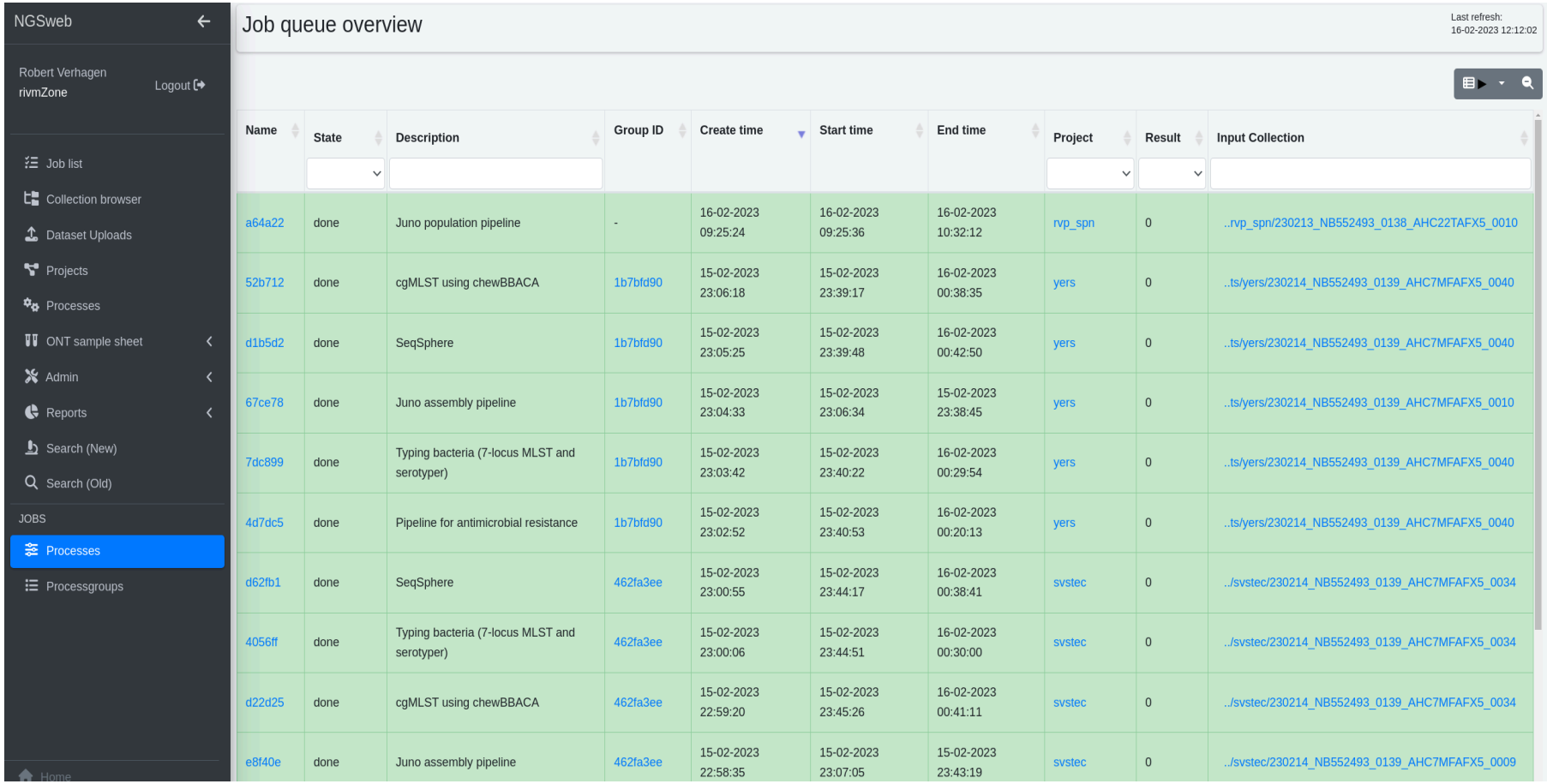
Collection browser data provenance graph (2/3)



Coll. browser contents and integrated document viewer (3/3)

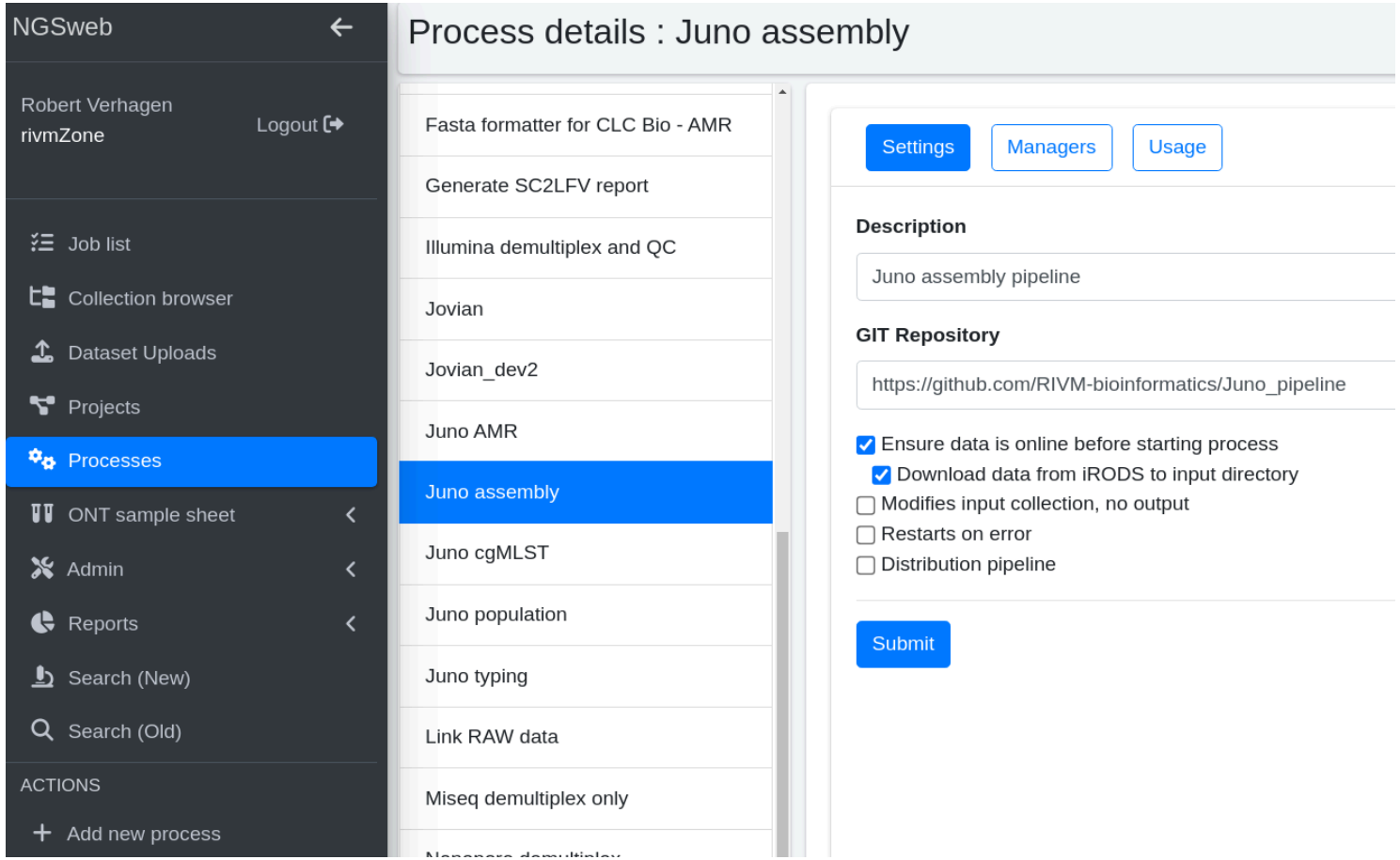


Job list

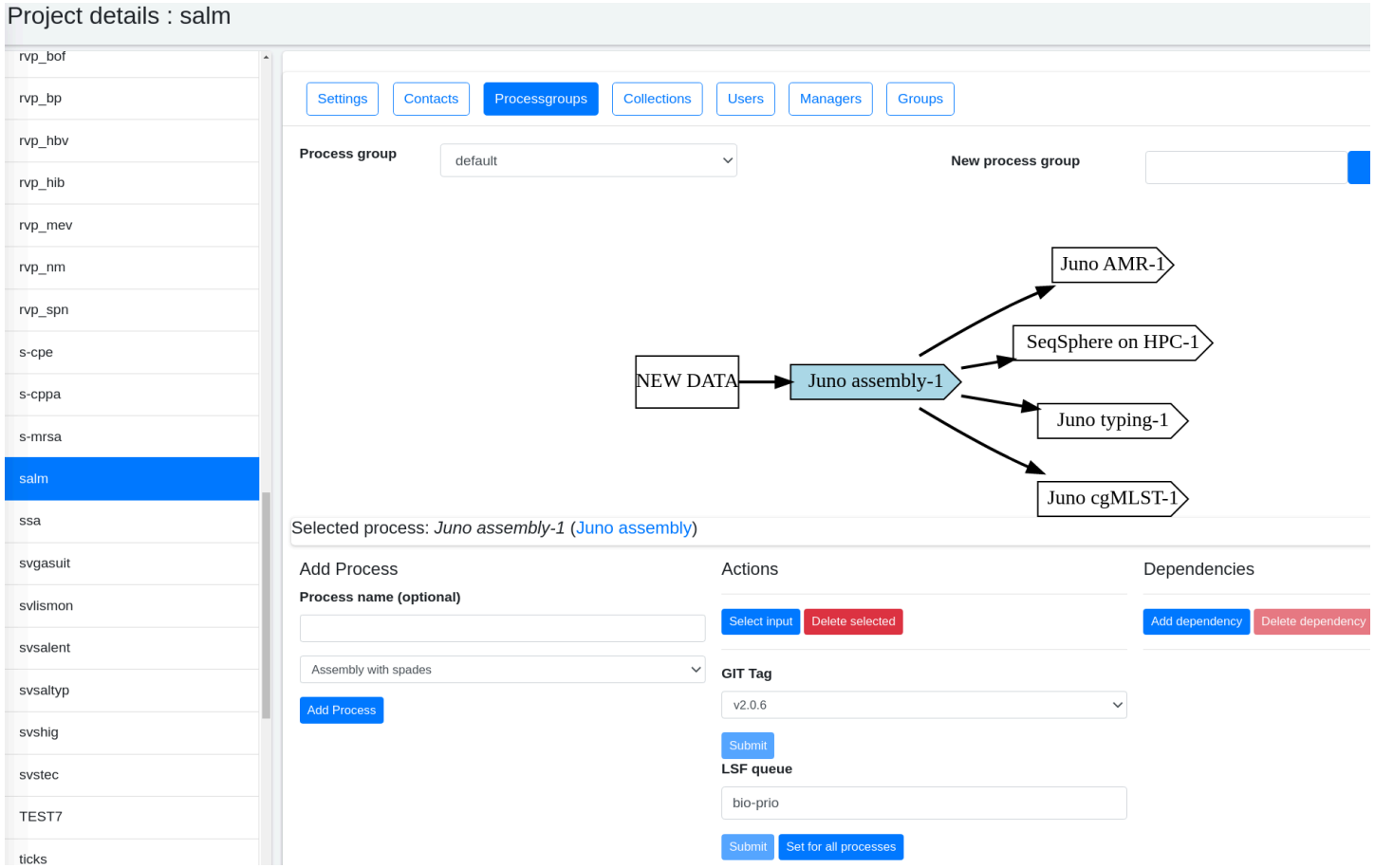


Name	State	Description	Group ID	Create time	Start time	End time	Project	Result	Input Collection
a64a22	done	Juno population pipeline	-	16-02-2023 09:25:24	16-02-2023 09:25:36	16-02-2023 10:32:12	rvp_spn	0	..rvp_spn/230213_NB552493_0138_AHC22TAFX5_0010
52b712	done	cgMLST using chewBBACA	1b7bfd90	15-02-2023 23:06:18	15-02-2023 23:39:17	16-02-2023 00:38:35	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
d1b5d2	done	SeqSphere	1b7bfd90	15-02-2023 23:05:25	15-02-2023 23:39:48	16-02-2023 00:42:50	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
67ce78	done	Juno assembly pipeline	1b7bfd90	15-02-2023 23:04:33	15-02-2023 23:06:34	15-02-2023 23:38:45	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0010
7dc899	done	Typing bacteria (7-locus MLST and serotyper)	1b7bfd90	15-02-2023 23:03:42	15-02-2023 23:40:22	16-02-2023 00:29:54	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
4d7dc5	done	Pipeline for antimicrobial resistance	1b7bfd90	15-02-2023 23:02:52	15-02-2023 23:40:53	16-02-2023 00:20:13	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
d62fb1	done	SeqSphere	462fa3ee	15-02-2023 23:00:55	15-02-2023 23:44:17	16-02-2023 00:38:41	svstec	0	../svstec/230214_NB552493_0139_AHC7MFAFX5_0034
4056ff	done	Typing bacteria (7-locus MLST and serotyper)	462fa3ee	15-02-2023 23:00:06	15-02-2023 23:44:51	16-02-2023 00:30:00	svstec	0	../svstec/230214_NB552493_0139_AHC7MFAFX5_0034
d22d25	done	cgMLST using chewBBACA	462fa3ee	15-02-2023 22:59:20	15-02-2023 23:45:26	16-02-2023 00:41:11	svstec	0	../svstec/230214_NB552493_0139_AHC7MFAFX5_0034
e8f40e	done	Juno assembly pipeline	462fa3ee	15-02-2023 22:58:35	15-02-2023 23:07:05	15-02-2023 23:43:19	svstec	0	../svstec/230214_NB552493_0139_AHC7MFAFX5_0009

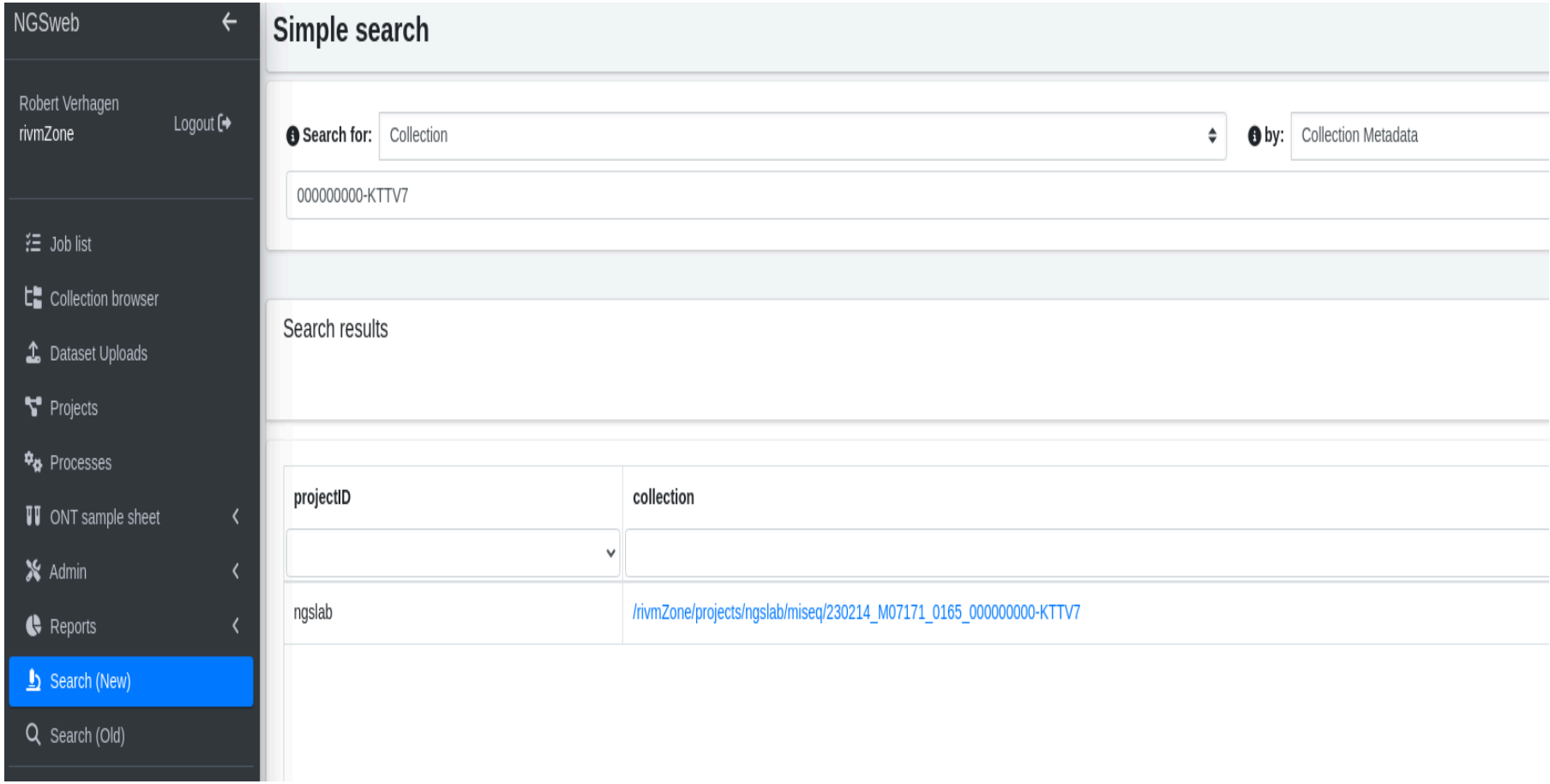
Processes


NGSweb
Robert Verhagen
rivmZone
Logout
Job list
Collection browser
Dataset Uploads
Projects
Processes
ONT sample sheet
Admin
Reports
Search (New)
Search (Old)
ACTIONS
Add new process
Process details : Juno assembly
Fasta formatter for CLC Bio - AMR
Generate SC2LFV report
Illumina demultiplex and QC
Jovian
Jovian_dev2
Juno AMR
Juno assembly
Juno cgMLST
Juno population
Juno typing
Link RAW data
Miseq demultiplex only
Settings
Managers
Usage
Description
Juno assembly pipeline
GIT Repository
https://github.com/RIVM-bioinformatics/Juno_pipeline
Ensure data is online before starting process
Download data from iRODS to input directory
Modifies input collection, no output
Restarts on error
Distribution pipeline
Submit

Project processgroups

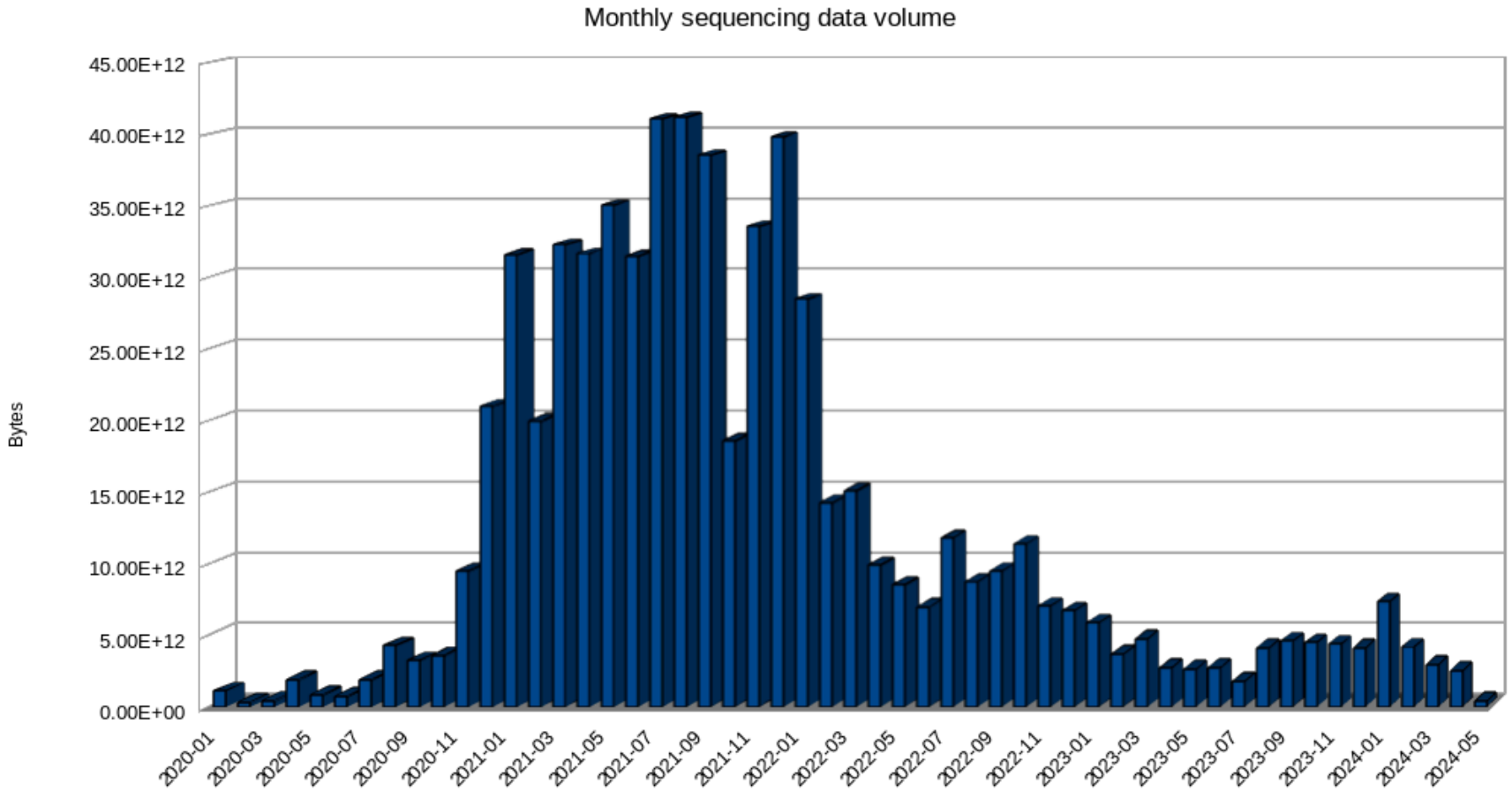


Search


NGSweb
Robert Verhagen
rivmZone
Logout
Job list
Collection browser
Dataset Uploads
Projects
Processes
ONT sample sheet
Admin
Reports
Search (New)
Search (Old)
Simple search
Search for:
Collection
by:
Collection Metadata
000000000-KTTV7
Search results
projectID
collection
ngslab
/rivmZone/projects/ngslab/miseq/230214_M07171_0165_000000000-KTTV7

The pandemic: scaling up

COVID-19 pandemic resulted in sequencing data volumes increasing to 100-fold



Succes stories

- Scale up in a short time.
- Create workplace where people can do research and run production.
- Share pipelines, also externally e.g. https://github.com/RIVM-bioinformatics/juno-assembly
- Automating import of sequence data and the starting of pipelines saved a lot of time.
- One place for datasets and archiving to Surf gives a big cost reduction in storage usage.

Future plans

- All (semi) unstructured datasets in iRODS.
- Use SURF and cloud compute for upscaling the HPC.
- Data sharing externally.
- Make our code Open Source.