Developing a life science infrastructure using automated workflows and data management according to FAIR by design principles

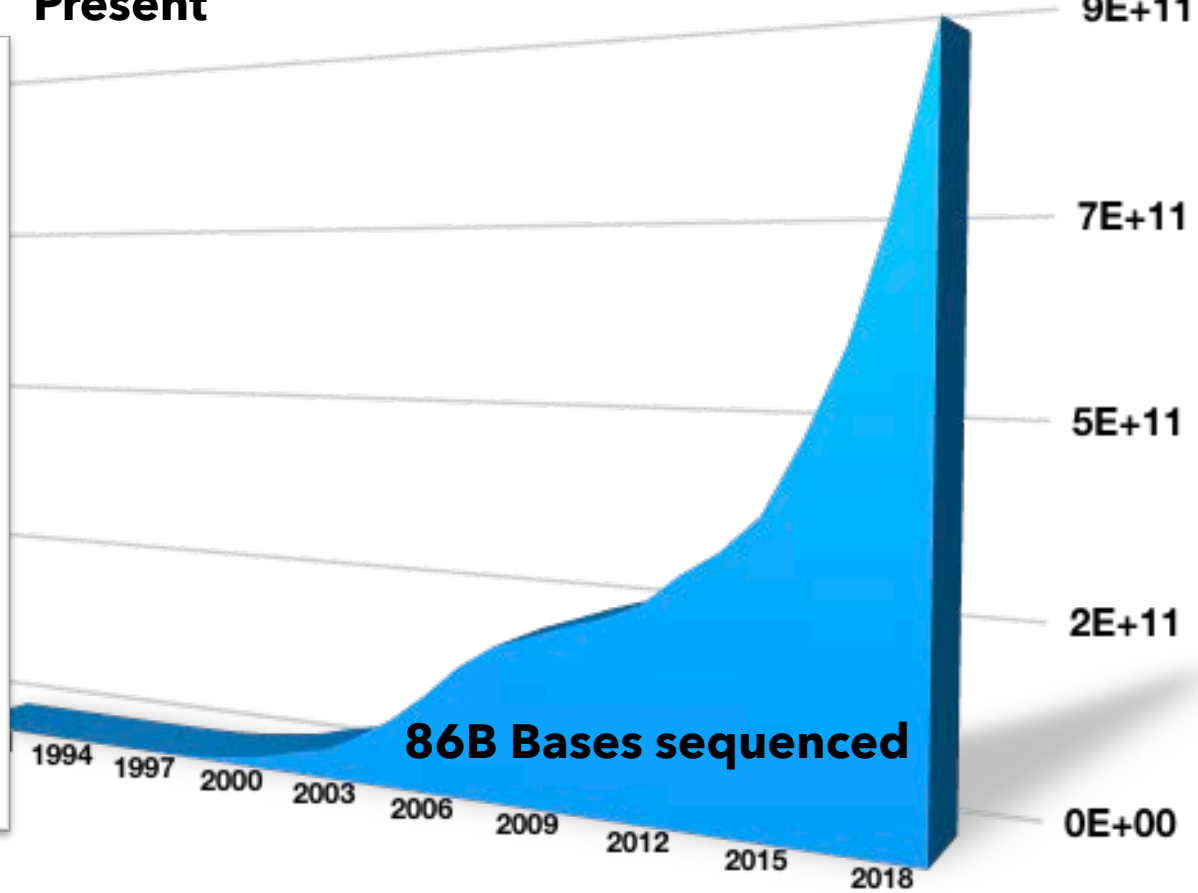
Jasper Koehorst

Data generation in Microbiology

Mostly small well documented reports

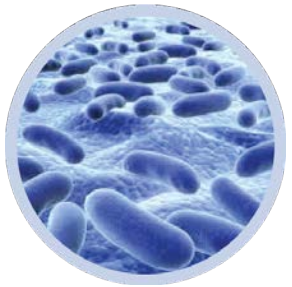
Past

Present



Data explosion

- Development of high-throughput sequencing
  - Unravelling the sequence of DNA (ATGC * 5.000.000)
  - Unravelling the composition of a sample
  - Unravelling the activity of genes (things that do the work)



That is not the only "thing" generating data… but it is the easiest and largest to capture

273+ petabyte stored at EBI (annual report 2018)

Unlocking
Microbial potential

An open infrastructure for exploring new horizons
for research on microbial communities.

Modular reactor systems

WU-ETE

Parallel cultivation

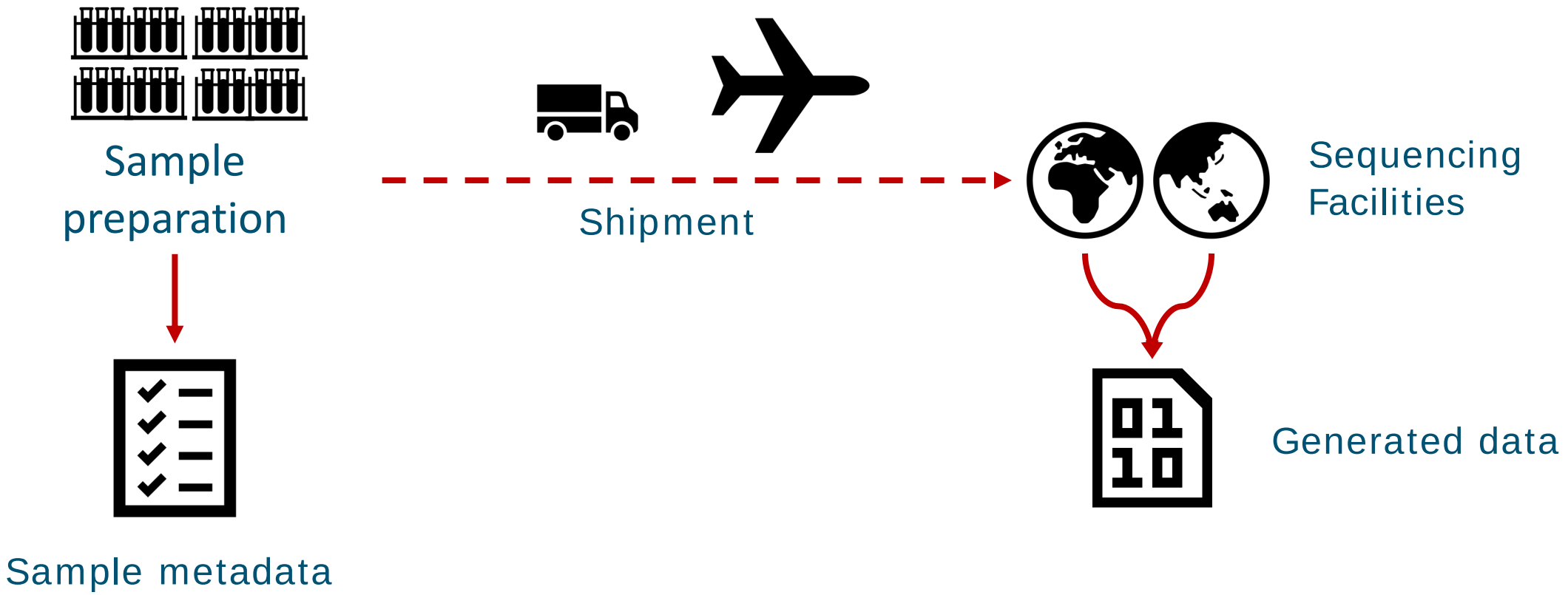
TUD-BT

Biodiscovery suite

WU-MIB

Dry laboratory WU-SSB

Data ingestion, sequence data (omics workflow)



Metadata

- All project data is stored according to the:
  - **P**roject (information)
  - **I**nvestigation (experimental design)
  - **S**tudy (Patient, Animal, reactor ID)
  - **A**ssay (Datasets)

Format…

- ISA-tab is standard
- minimal information system

isatools

X

RDF

Metadata

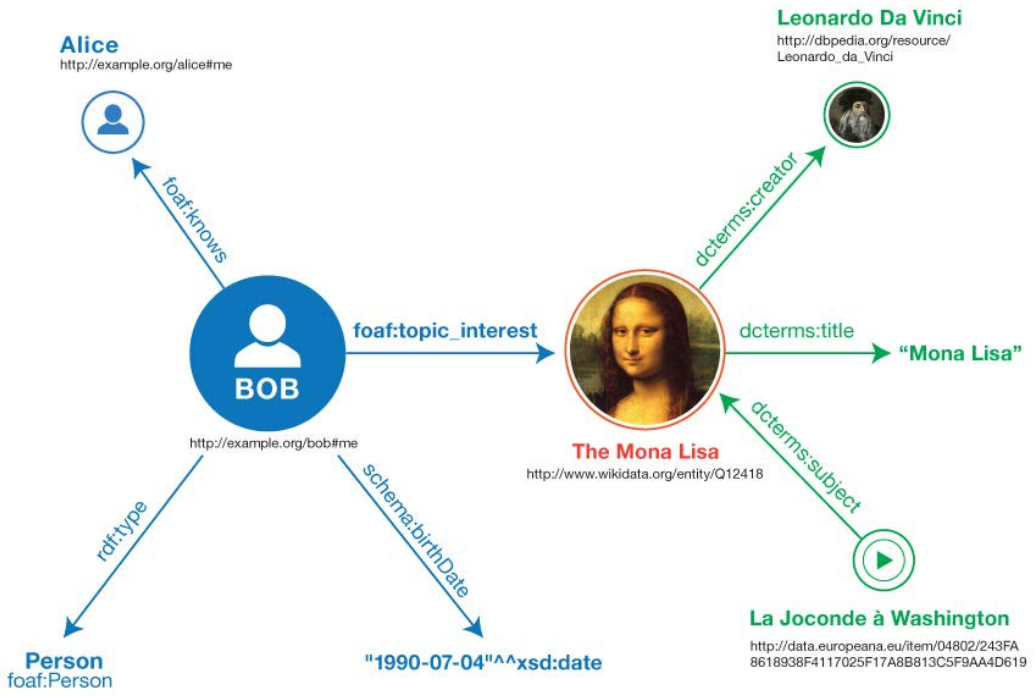
Generic

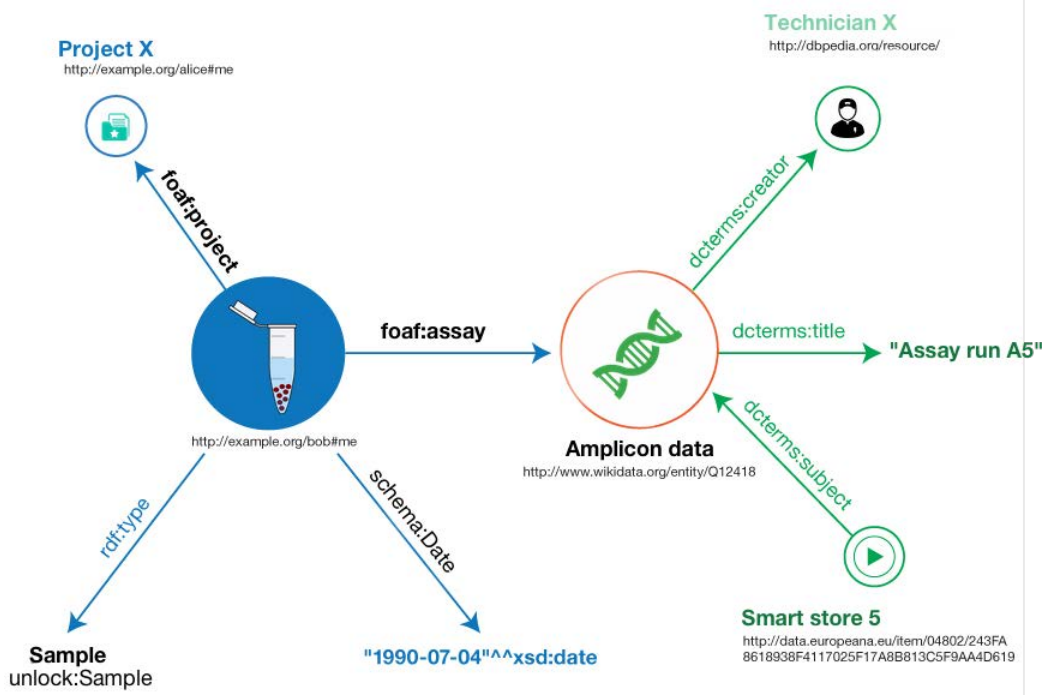
- To what project does it belong to?
- Who is responsible?
- Contact details

Specific

- Environmental conditions
- Collection time / date
- Geographical locations
- Host

Semantic Translation (RDF)





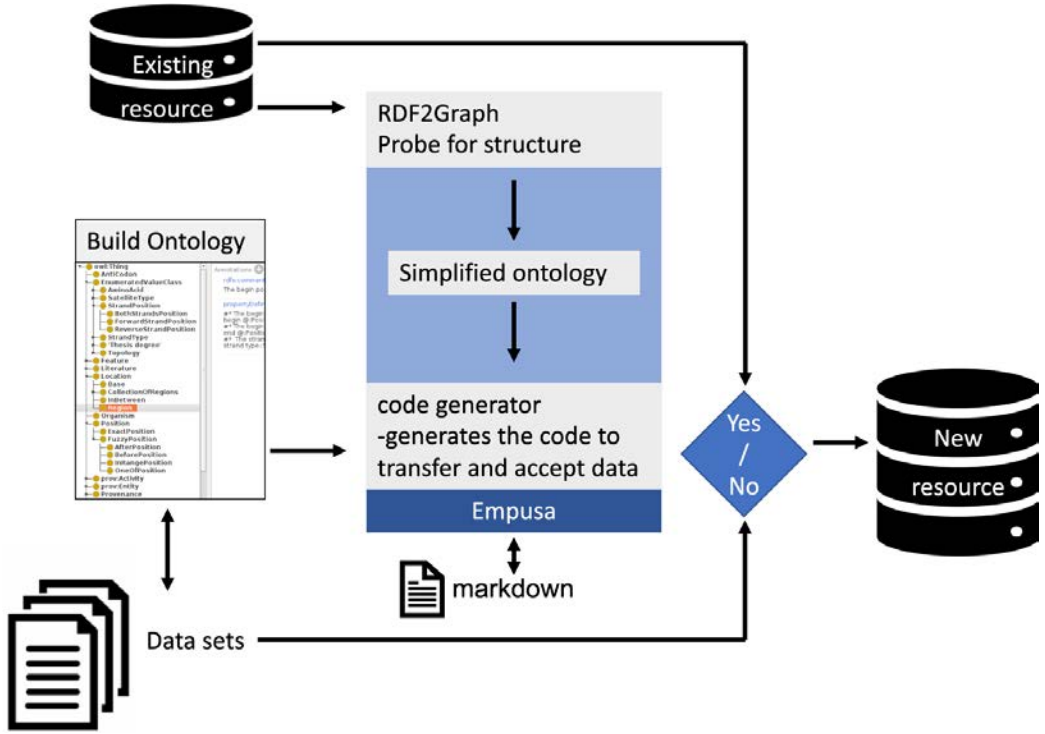
Empusa validator

Article | Open Access | Published: 04 November 2019

The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation

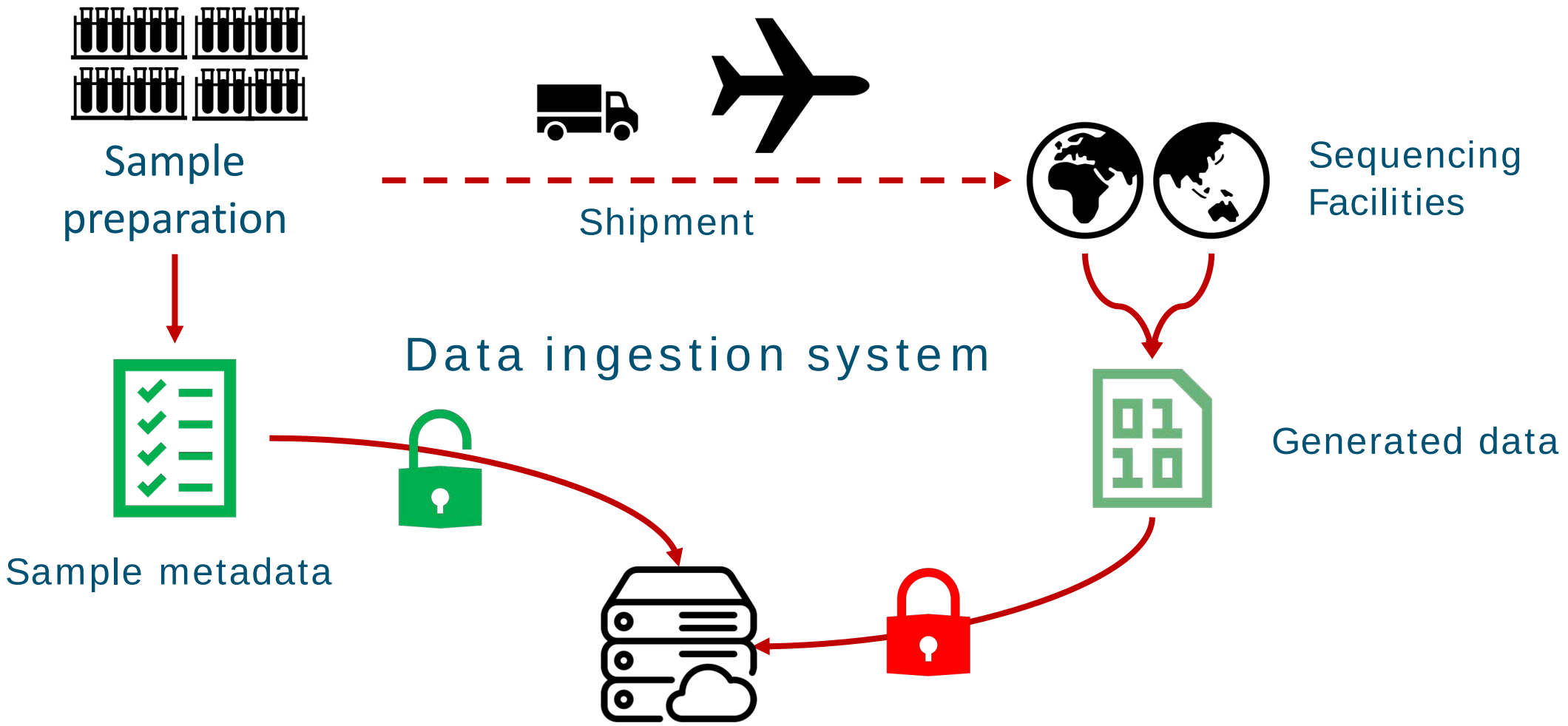
Jesse C. J. van Dam, Jasper J. Koehorst, Jon Olav Vik, Vitor A. P. Martins dos Santos, Peter J. Schaap & Maria Suarez-Diez

Scientific Data **6**, Article number: 254 (2019) | Cite this article



- Metadata ontology
- OWL/ShEx format
- Varying requirements
- Continuous validation through XSLX > RDF

Data ingestion, sequence data

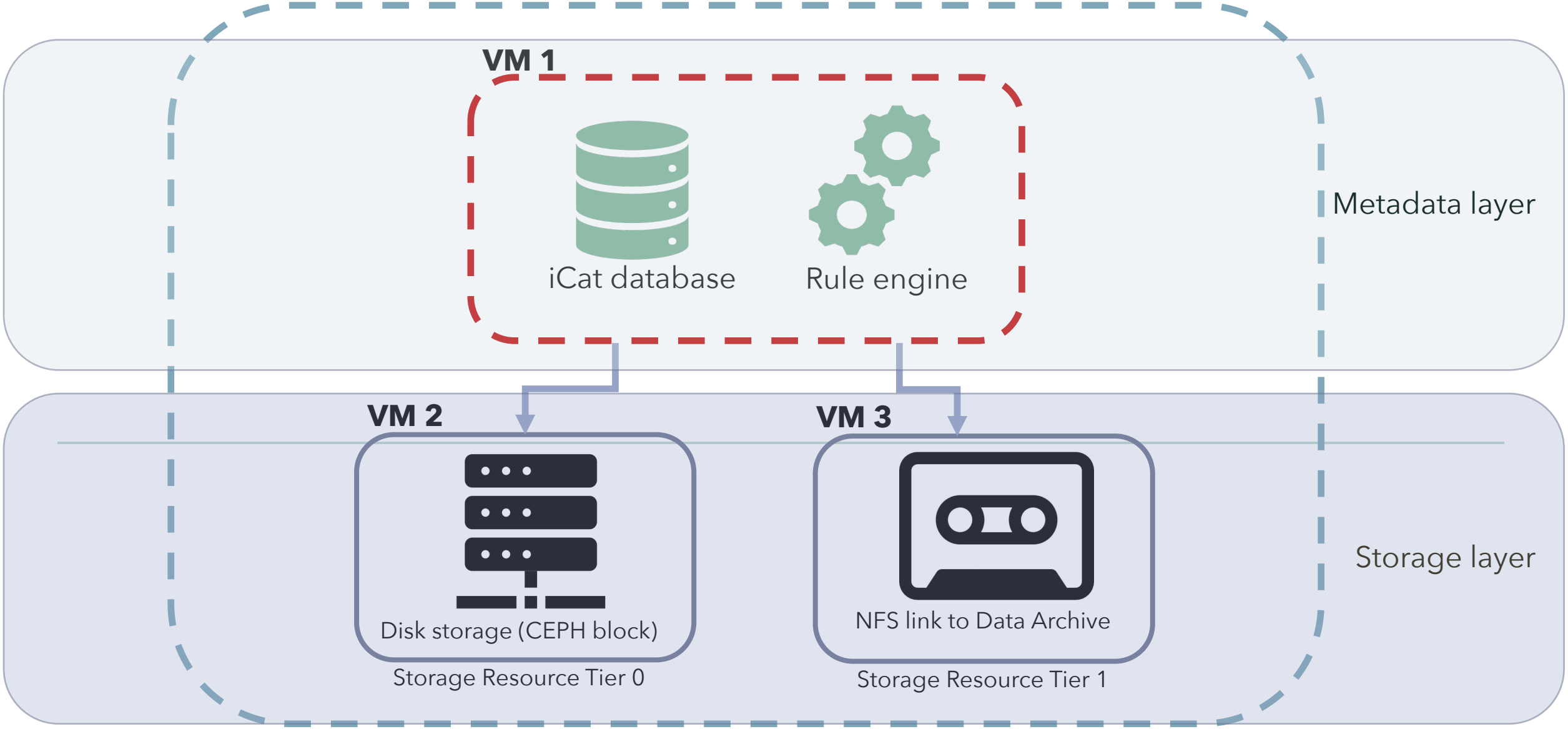


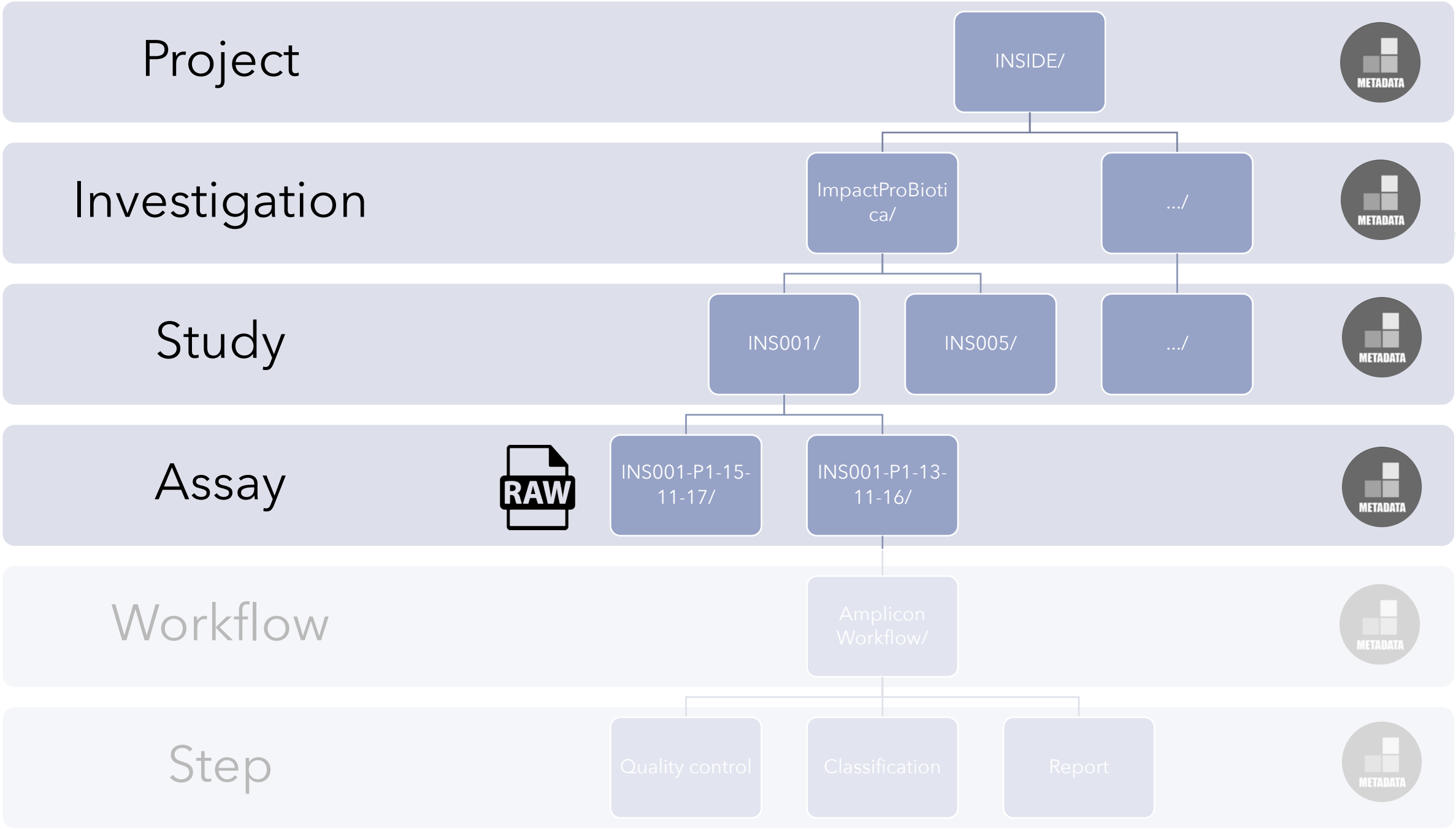
Storage, storage, storage

Reasons why

- Data storage
- Metadata storage (AVU triples)
- Mount (e.g. webdav, for lab equipment / users)
- API (Python / Java)
- Query interface (iquest)

iRODS zone UNLOCK hosted at SURF




Project
INSIDE/
METADATA
Investigation
ImpactProBiotica/
.../
METADATA
Study
INS001/
INS005/
.../
METADATA
Assay
RAW
INS001-P1-15-11-17/
INS001-P1-13-11-16/
METADATA
Workflow
Amplicon Workflow/
METADATA
Step
Quality control
Classification
Report
METADATA

Example

*Every individual pig, human, bioreactor is a study

Project Investigation Study* AssayType Data Assay

/INSIDE/ I_ImpactOfProbiotica/S_INS001/Amplicon/ INS001-P1-13-11-16/Unprocessed/1.fastq.gz

/INSIDE/ I_ImpactOfProbiotica/S_INS001/Amplicon/ INS001-P1-13-11-16/Processed/

AmpliconWorkflow/Step_2_NGTaX_2.0_#1/x.biom

/x.rdf

Workflows

Compute

Majority of the data generated can be automatically (pre-) processed

- Quality control
- Characterisation
- Conversion

- Local compute cluster 
- HPC Wageningen 
- HPC SURF Sara 

- Easy to develop modules
- Easy to deploy
- Don’t care where it runs
  - Local
  - Remote
  - Cloud

iRODS metadata & rule system

Combined with the **C**ommon **W**orkflow **L**anguage

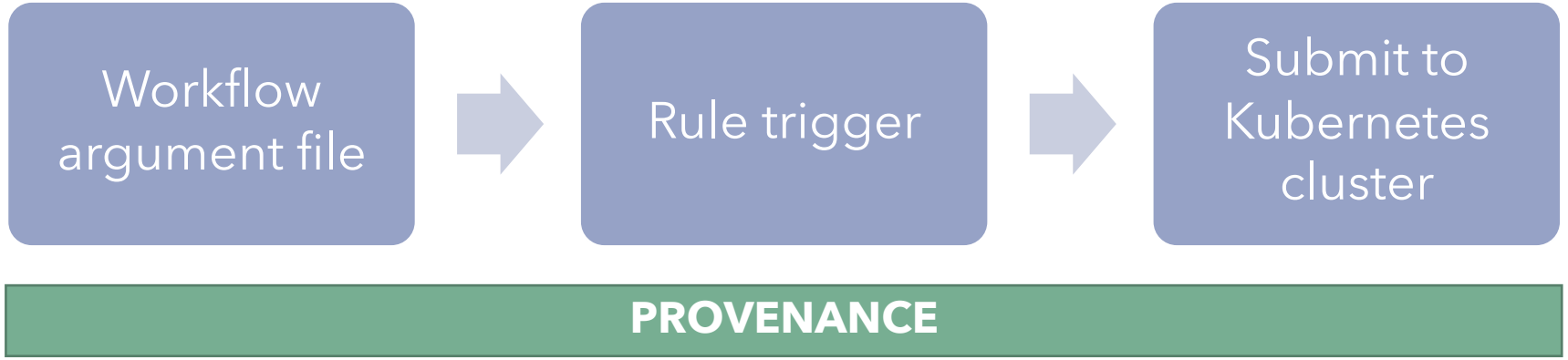
COMMON WORKFLOW LANGUAGE

File: amplicon_run.yml **(contains parameters)**

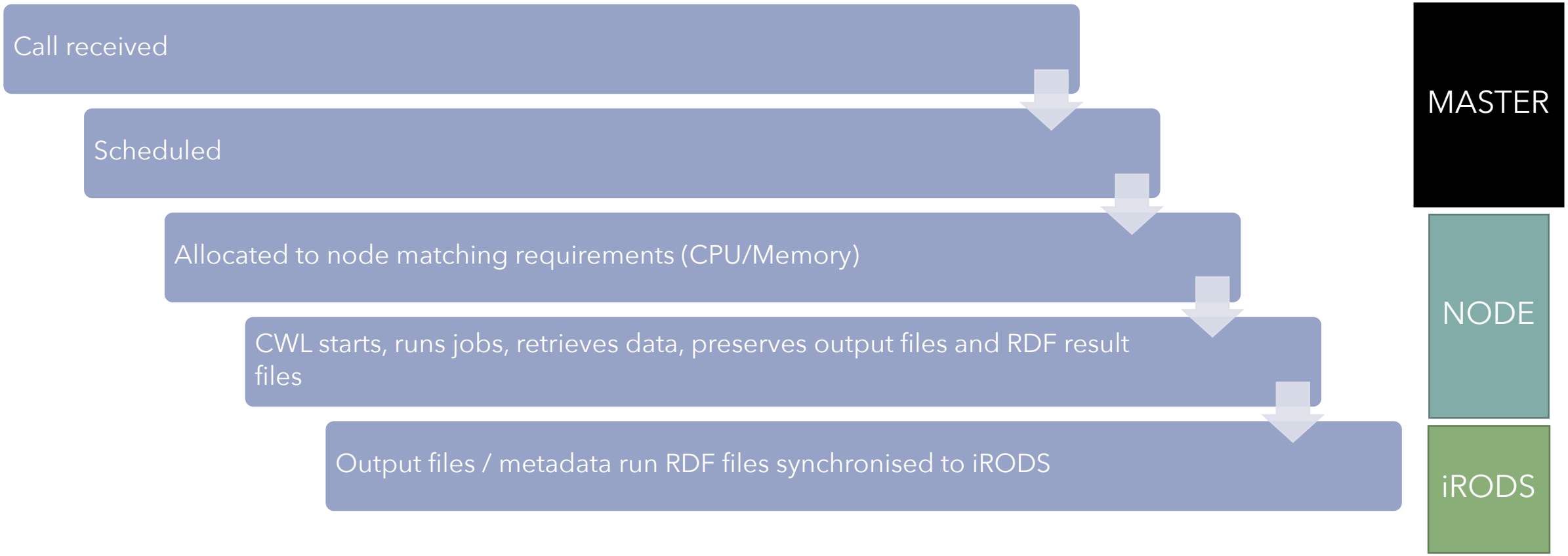
Attribute – workflow

Value - workflow_amplicon.cwl **(CWL to execute)**

Unit – Running (**Status)**



Kubernetes workflow



A (public) use case: **DIABIMMUNE**

Development of the gut microbiome of babies (followed over 600 days)

PROJECT	INVESTIGATIONS	STUDIES	VARIABLES	DNA SAMPLES
1	3	289	60	3204

Raw data storage		1 hour
Metadata registration		days (±1 hour)
Organise data		1 hour
Deploy workflow files (yml)		<1 hour
Kubernetes processing	(embarrassingly parallel)	12 hours
Result storage		

Jupyter notebook integration

- Metadata as RDF
- SPARQL queries
  - Sample information / parameters
  - Provenance of workflows or datasets

Microbial composition of babies born through c-section and when they did not eat fruit yet

jupyter

SPARQL

API

iRODS

Arthur Newton

David Salek

SURF SARA

Bart Nijsse

Peter Schaap

Willem Jan Knibbe

WAGENINGEN UNIVERSITY & RESEARCH

WUR is serious about Data

WDCC