

MIAPPE Wizard: Enabling easy creation of MIAPPE-compliant ISA metadata for Plant Phenotyping Experiments

Sebastian Beier (FZ Jülich) and Daniel Arend (IPK Gatersleben)

Motivation: Build web app to generate metadata for plant phenotypic experiments

- Tight collaboration with 
- Tasks:
 - Graphical design
 - User experience
 - Data model discussions (ISA / MIAPPE)
 - Ontology connections (DataPLANT, MIAPPE, CRediT, others)
- ... and of course having fun!



Prototype GUI



miappe Wizard

[Load minimal example](#)

[Add new Investigation](#)

[Save ISA-JSON as file](#)

[Load ISA-JSON from file](#)

[Start Wizard mode](#)

Investigation

▼ People (2)

Daniel Arend
Patrick König

Investigation

submissionDate: ×

People add person

Person

Daniel

Arend

Corrensstraße 3, 06466 Seeland, Germany

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Comment: x

x

[add comment](#)

Person

ISA-JSON ([hide](#))

```
{  
  "submissionDate": "2007-04-30",  
  "people": [  
    {  
      "firstName": "Daniel",  
      "lastName": "Arend",  
      "address": "Corrensstraße",  
      "affiliation": "Leibniz I",  
      "comments": [  
        {  
          "value": "",  
          "name": "Investigatio  
        }  
      ],  
      "firstName": "Patrick",  
      "lastName": "König",  
      "address": "Corrensstraße",  
      "affiliation": "Leibniz I",  
      "comments": [  
        {  
          "value": "",  
          "name": "Investigatio  
        }  
      ]  
    }  
  ]  
}
```

Wizard questionnaire

miappe Wizard

[Load minimal example](#)

[Add new Investigation](#)

[Save ISA-JSON as file](#)

[Load ISA-JSON from file](#)

[Start Wizard mode](#)

- Investigation
- ▼ Publications (0)
- ▼ People (0)
- Studies (0)

New Investigation (Wizard-style)

What is the title of your investigation?

title:

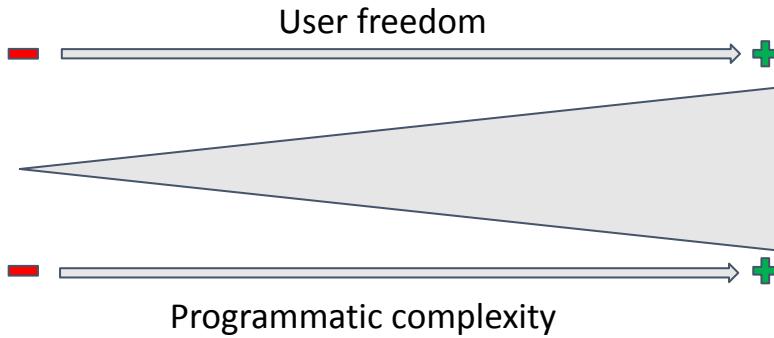
Next

[ISA-JSON \(hide\)](#)

```
{  
  "@id": "",  
  "filename": "",  
  "identifier": "",  
  "title": "",  
  "description": "",  
  "submissionDate": "",  
  "publicReleaseDate": "",  
  "ontologySourceReferences": [],  
  "publications": [],  
  "people": [],  
  "studies": [],  
  "comments": []  
}
```

Challenges

- ISA-Tab - ISA-JSON differences (hierarchy, identifier names)
 - MIAPPE (phenotyping standard) mapped to ISA-Tab but not to ISA-JSON
- Difficult to accommodate all user groups (balance between novice and experts)



- Incorporating virtual participants → limiting to certain time slots
 - Discussion Wednesday afternoon (ISA & MIAPPE)
- Connecting into other tool ecosystems (ARC Commander, pISA-Tree)

Current progress



Motivation: Build web app to generate metadata for plant phenotypic experiments

- Status:
 - Graphical design
 - Basic design (optional developer console)
 - Treeview
 - User experience
 - Form based and questionnaire dialog
 - Data model discussions (ISA / MIAPPE)
 - Generic ISA → future perspective MIAPPE
 - ISA implementation (ISA-JSON) to be discussed with community
 - Ontology connections (DataPLANT, MIAPPE, CRediT, others)
 - REST API adaptors to DataPLANT ontology
 - Extension of DataPLANT ontology to include MIAPPE & CRediT



Extending the NFDI4Microbiota Knowledge Base

Justine Vandendrope (ZB MED) and Kassian Kober
(Bielefeld University)

Bla Bla

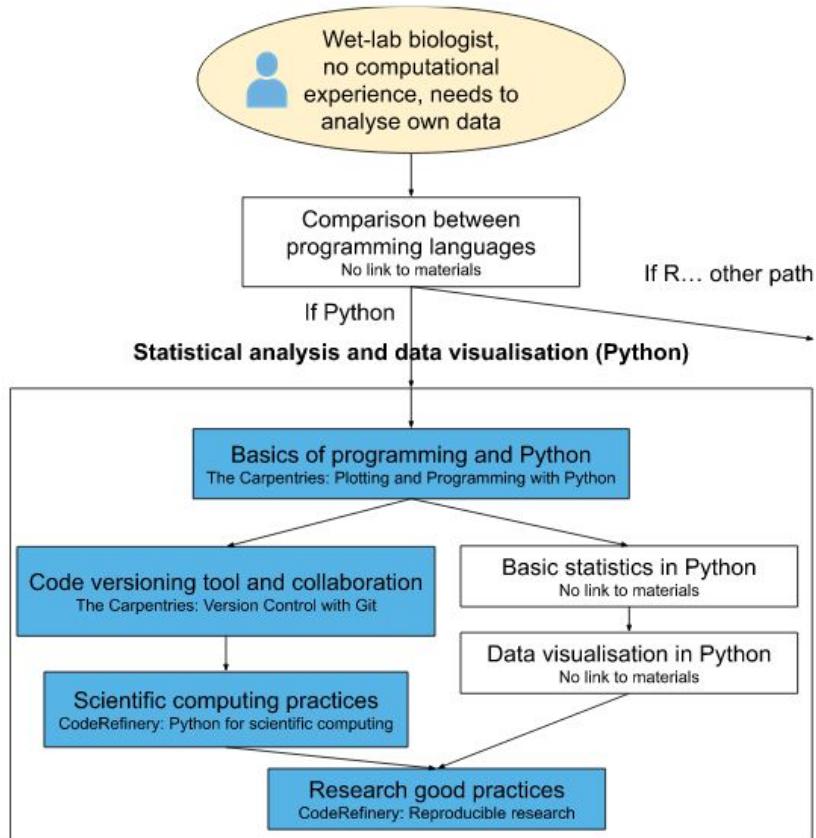


Mapping the training journey in Bioinformatics and beyond

**Lisanna Paladin (EMBL Heidelberg) and Pablo Mier
(University of Mainz)**

Quick recap

- **Aim:** mapping the “**training journey**” of different **personas** in Bioinformatics, visualising the steps needed to acquire expert knowledge in each topic, or skill.
- **For:** trainers & trainees.
- **How:** training paths



1. Gathering data: Strategy

- About
 - Personas: "Wet lab biologist that is able to visualise small dataset"
 - Skills: "Programming in R", "Data visualisation (in R)", "Programming in Python"
 - Training courses:

Data Visualization

This lesson shows how to choose the appropriate graphic to represent your data and research question, inspired by [Choosing a good chart](#), and then how to implement construction of the graphic using `ggplot2`.

Introduction to R

Data Carpentry contributors

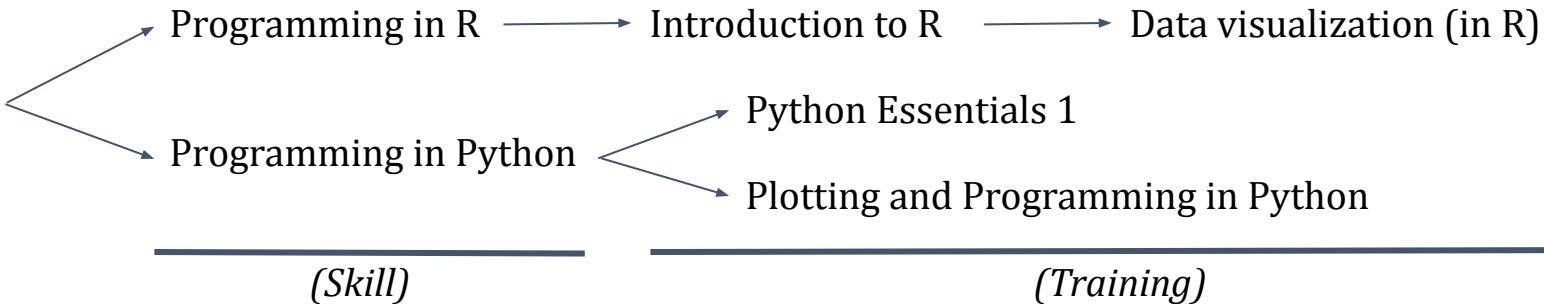
Python Essentials 1

Dive into programming, learn Python from scratch, and prepare for the PCEP – Certified Entry-Level Python Programmer certification.

Plotting and Programming in Python

This lesson is an introduction to programming in Python 3 for people with little or no previous programming experience. It uses plotting as its motivating example and is designed to be used in both [Data Carpentry](#) and [Software Carpentry](#) workshops. This lesson references [JupyterLab](#) but can be taught using alternative Python 3 interpreters as well (e.g., `repl.it`, Anaconda).

- Dependencies



1. Gathering data: Current status

18 personas (8 fields each)

Code	Persona SHORT description	Carreer stage	Skills they have already / background	Any programming skills?	Any background in life sciences?	Experience they already have	Need / motivation	Skills
P1	Wet lab biologist that is able to visualise small dataset	Early	Wet lab methods, life science background	No	Yes	Wet lab techniques	Visualise small dataset	S1 S36,S2,S3
P2	Biochemist who gets their transcriptomic dataset	Early	Biochemistry, sequencing techniques knowledge	No	Yes	-	Transcriptomic skills, Visualise small dataset	S1,S2,S3,S4,S5,S6

41 skills (5 fields each)

Code	Description	Level	Persona	Dependency	Training
S1	Programming in R	Basic	P1,P8	-	T1
S2	Data visualisation (in R)	Basic	P1,P7,P8,P12	S1	T4
S3	FAIR -> Software focused / Git	Basic	P12,P18	S15	T28,T32,T33,T36
S4	Data cleaning (with bash)	Basic	P1,P2,P5	-	-
S5	Transcriptomics theory	Basic	P1,P2	-	T54

55 training courses (14 fields each)

Code	Description	Link learners	Link trainers	Skill	Dependency	Easy to adopt from trainers?	Easy to adopt for learners?	Maintained?	Last update	Accept feedback?	Customizable?	Theoretical/Practical
T1	Introduction to R	https://datacarpentry.org/R-ecology-lesson/01-ir	https://github.com/dat	-	-	Yes	Yes	Yes	2022	Yes	Yes	Practical
T2	Introduction to Python	http://swcarpentry.github.io/python-novice-gapm/	https://github.com/swc	-	-	Yes	Yes	Yes	2022	Yes	Yes	Practical
T3	Data visualisation in Python	https://carpentries-incubator.github.io/python-int/	https://github.com/carp	T2	-	Yes	Yes	Yes	2022	Yes	Yes	Practical
T4	Data visualisation in R	https://swcarpentry.github.io/visualization-novice/	https://github.com/swc	T1	-	Yes	Yes	Yes	2022	Yes	Yes	Practical
T5	Orchestrating Single Cell Analysis	https://bioconductor.org/books/release/OSCA/	-	T1	-	-	-	Yes	2022	-	-	Practical
T6	Python essentials 1	https://pythoninstitute.org/python-essentials-1/	-	-	No	Yes	Yes	Yes	2022	No	No	Both
T7	Introduction to R Shiny	https://shiny.rstudio.com/tutorial/written-tutorial/1/	-	T1	-	-	-	Yes	-	-	-	-

2. Data visualization: our idea



Wet Lab

Persona 1
Persona 2
Persona 3

Comput Biol

Persona 4
Persona 5
Persona 6

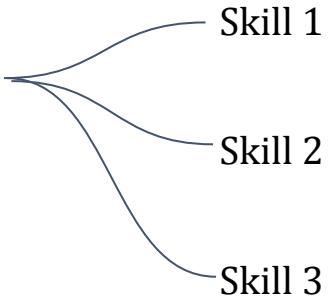
Other

Persona 7
Persona 8
Persona 9

2. Data visualization: our idea

Wet Lab

Persona 1
Persona 2
Persona 3



Comput Biol

Persona 4
Persona 5
Persona 6

Other

Persona 7
Persona 8
Persona 9

Download

metadata 1
metadata 2
metadata 3

...

(for Persona 2)

2. Data visualization: our idea

Wet Lab

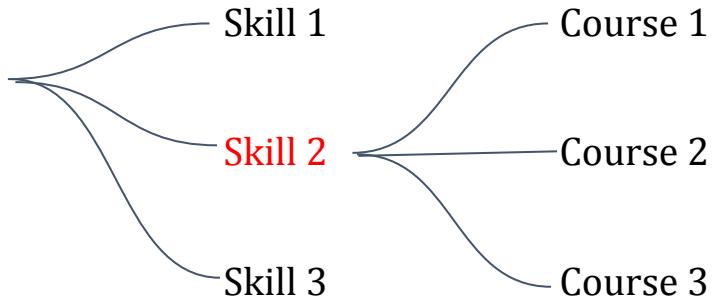
Persona 1
Persona 2
Persona 3

Comput Biol

Persona 4
Persona 5
Persona 6

Other

Persona 7
Persona 8
Persona 9



Download

metadata 1
metadata 2
metadata 3

...

(for Skill 2)

2. Data visualization: our idea

Wet Lab

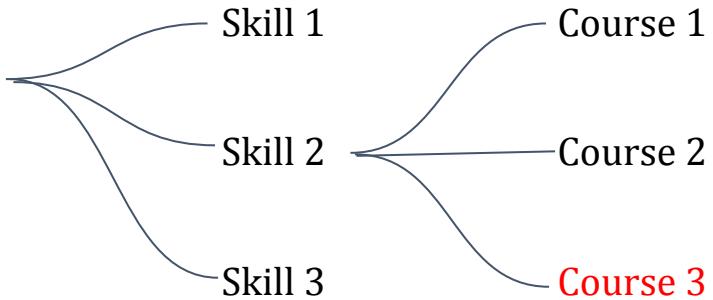
Persona 1
Persona 2
Persona 3

Comput Biol

Persona 4
Persona 5
Persona 6

Other

Persona 7
Persona 8
Persona 9



Download

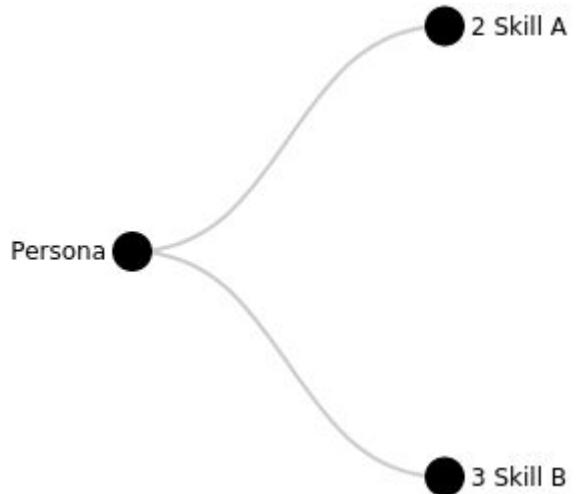
metadata 1
metadata 2
metadata 3

...

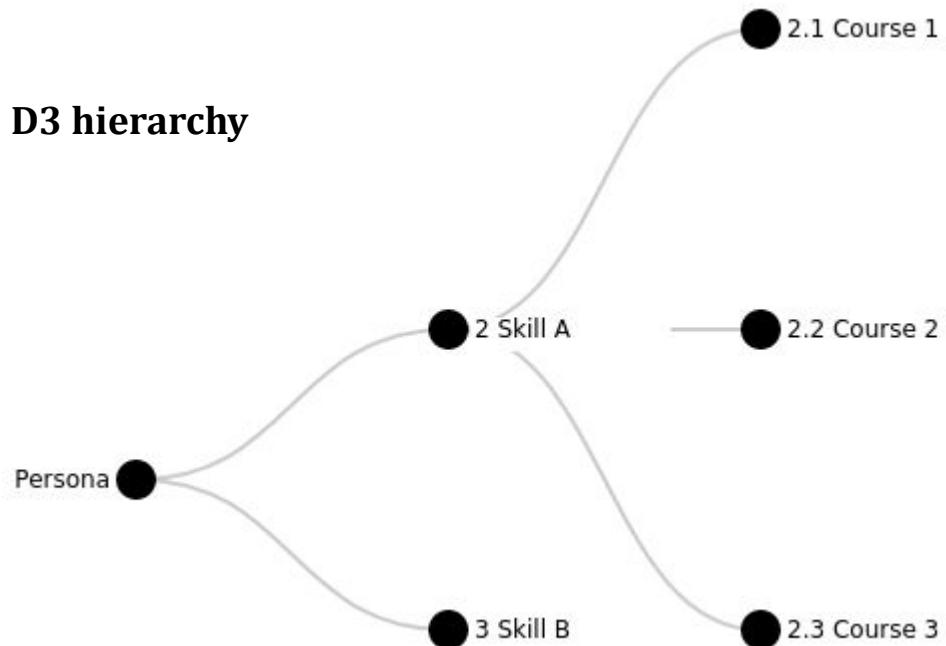
(for Course 3)

2. Data visualization: possible implementation

D3 hierarchy

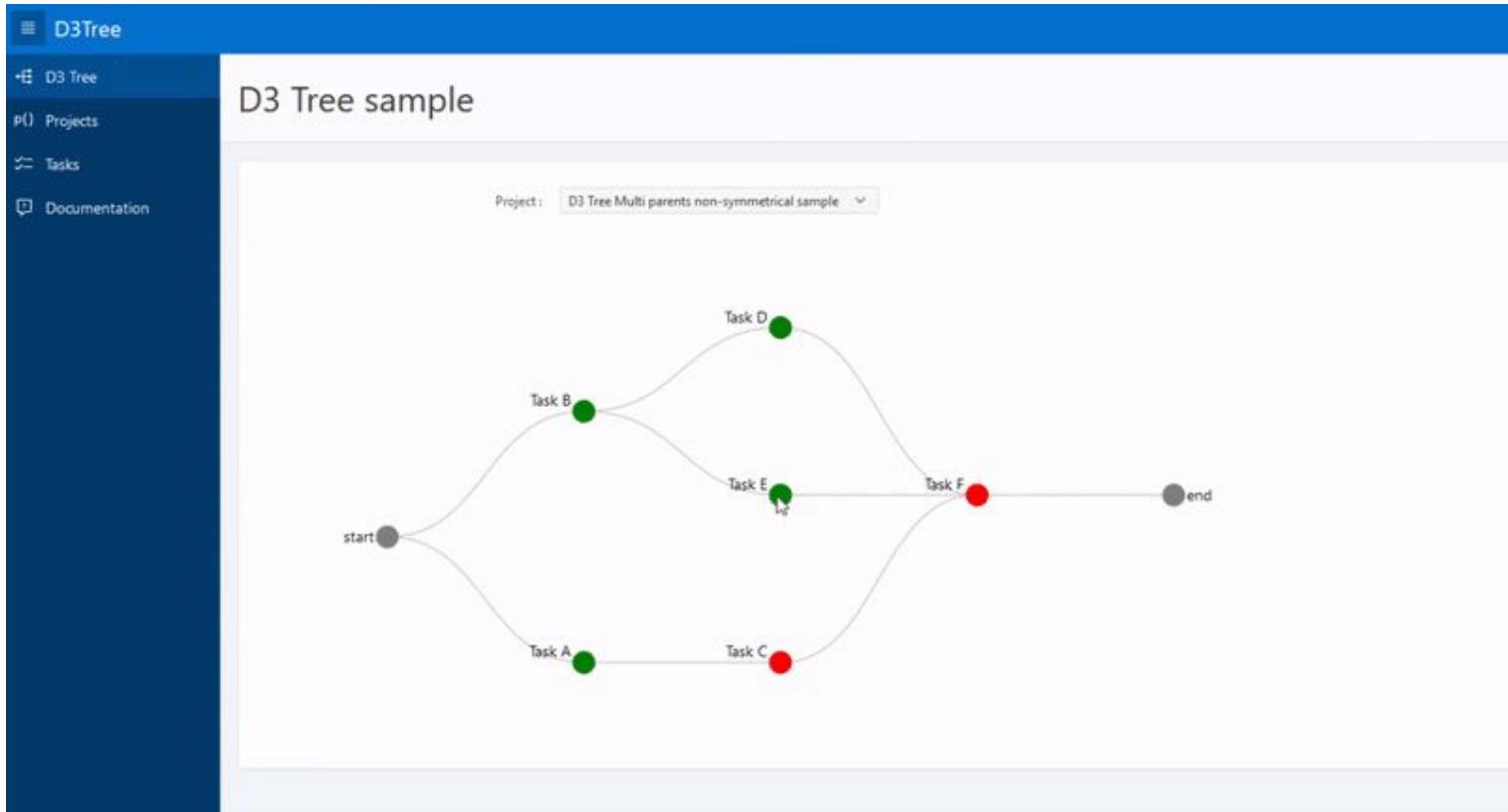


2. Data visualization: possible implementation



2. Data visualization: possible implementation

D3 hierarchy



3. Next steps



What will we do now?

Implement our data in a visual manner

What do we need?

More people to expand our persona/skill/training set (currently up to ~6 people)
D3 expert(s)

Establishing best practice guidelines for imaging-based spatially resolved transcriptomics data

Naveed Ishaque (Berlin Institute of Health at the Charité)
and Louis Kümmerle (Helmholtz Center München)

Day 1 - on boarding and motivation

- 3 introductory talks
- 2 technical methods talks
- 3 biology talks
- **Intro to TXsim**

Screenshot of a web browser showing the schedule for the de.NBI/ELIXIR-Germany Spatial Hackathon.

The page title is "spatialhackathon.github.io/schedule.html". The header includes the de.NBI/ELIXIR-Germany logo, the text "de.NBI/ELIXIR-Germany Spatial Hackathon", and navigation links for "Home", "Schedule", and "Contact".

Schedule

This workshop will take place from **Monday 12th Dec - Friday 16th Dec 2022**. The timing of this event will be **Central European Time (CET)**.

Day 1: Monday 12th December 2022 [Times in CET and PST]

13:00-15:00 [04:00-06:00]   Talks on de.NBI biohackathon and all projects

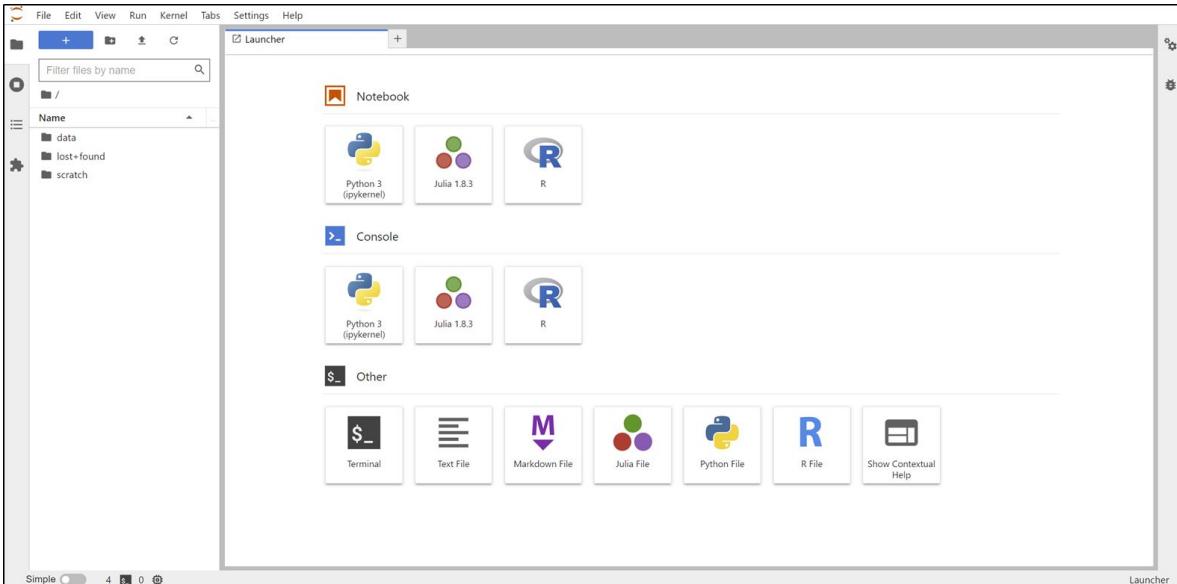
15:00-15:30 [06:00-06:30] Break

15:30-18:50 [06:30-09:50] Introductory presentations

- 15:30 [06:30]   General introduction (topic, motivation, paper, cloud, "streams"), *Naveed Ishaque, BIH*
- 15:50 [06:50]   General organisation (Slack, GitHub, tasks, cloud), *Naveed Ishaque, BIH*
- 16:10 [07:10]   Practical: SpaceHack cloud, *Naveed Ishaque, BIH*
- 17:00 [08:00]    Spatial cell type mapping in SpaceTx, *Renee (Yun) Zhang, J. Craig Venter Institute*
- 17:20 [08:20]   TXsim - a workflow to compare single cell and segmentated spatial, *Louis B Kuemmele, Helmholtz Munich*
- 17:50 [08:50]   Data + challenges: human heart and kidney, *Christoph Kuppe, RWTH Aachen*
- 18:10 [09:10]   Data + challenges: human brain, *Jennie Close, Allen Inst*
- 18:30 [09:30]  Data + challenges: mouse brain, *Michael Kunst, Allen Inst*

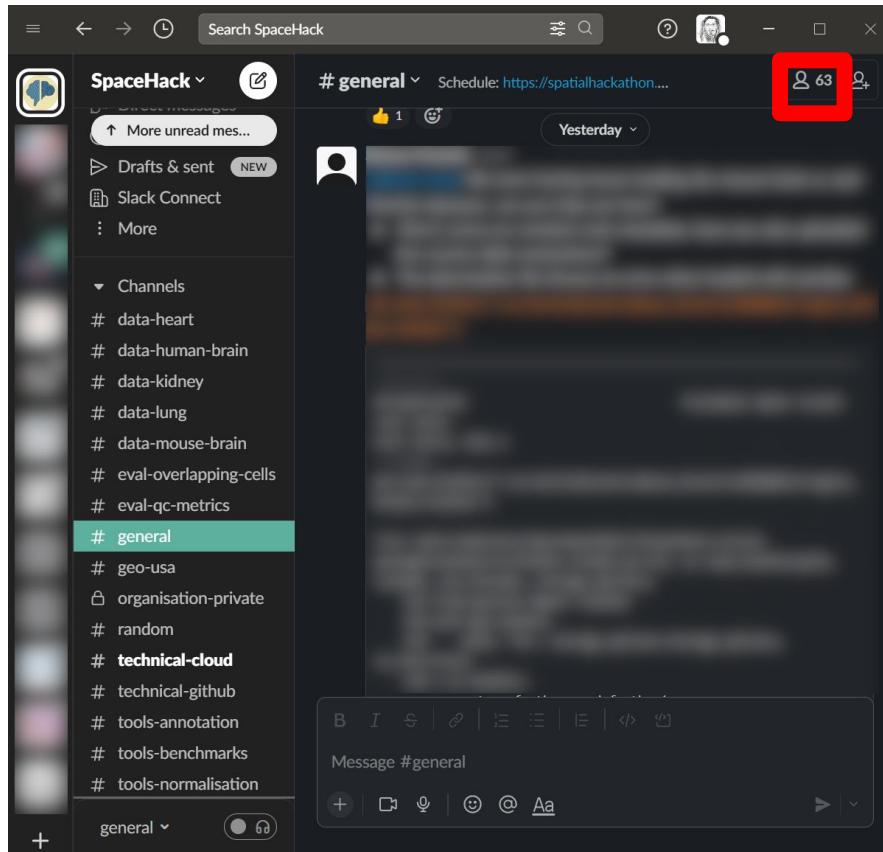
Day 2 - starting with de.NBI cloud

- ~2000 cores
- ~8 Tb RAM
- 50 Tb storage
- JupyterHub with GitHub authentication
- 2 HowTo seminars



Where we stand

- 63 participants
- 9 GitHub Repos
- 29 Slack channels
- 7 sub-projects



Where we stand

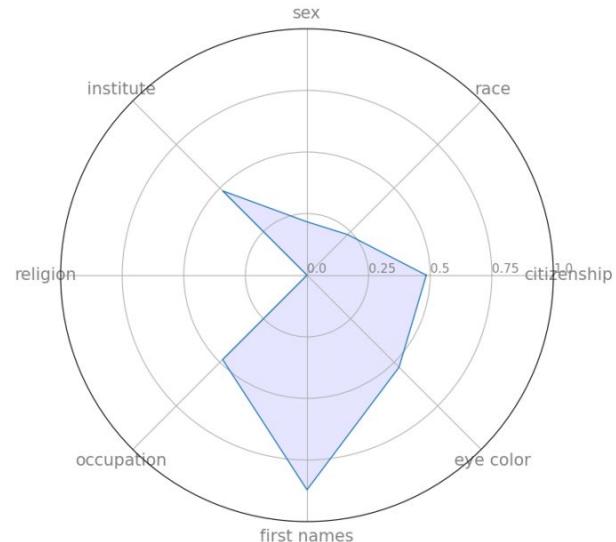
- 63 participants
- 9 GitHub Repos
- 29 Slack channels
- 7 sub-projects



A screenshot of a Slack workspace titled "SpaceHack". The left sidebar shows a list of channels, with "# general" highlighted in green. The main pane displays the "# general" channel, which has 63 members. A red box highlights the member count "63" in the top right corner of the channel header. The message list is mostly blurred, but a single message from "Yesterday" is visible. At the bottom, there is a message input field and various keyboard shortcuts.

Project 1: #xenium_omni_pipeline

- **Project motivation:** Evaluation of added value of Xenium technology and designing the analysis workflow
- **Our solution**
 - Creating necessary infrastructure using SpatialData & TXsim
 - QC of the workflow
 - Multiomics integration and analysis
- **Update points**
 - Imported Xenium & Visium into SpatialData format, napari-assisted registration
 - Napari plugin for smooth histopathological annotation
 - Xenium & scFFPE annotated
 - Visium & scRNA-seq copy number projection
 - Diversity plot of the group created



Project 2: #zarr_for_data_and_metadata

- **Project motivation**

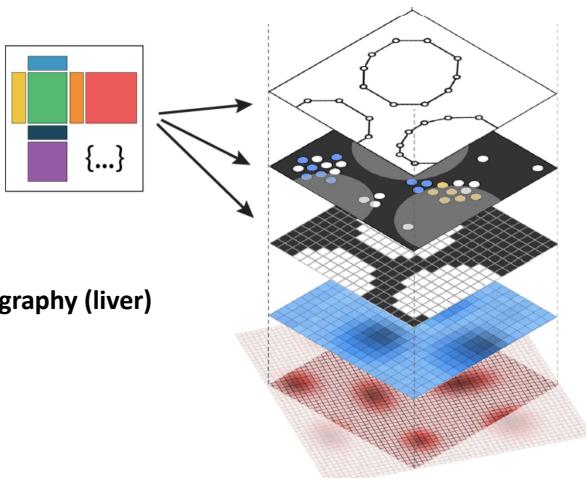
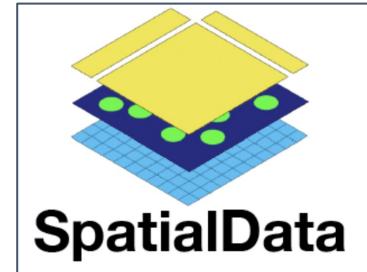
- Analyzing multiple spatial datasets with all their metadata takes a **long time and varies in quality**

- **Our solution**

- Increase **interoperability and performance** using a new standard: [SpatialData](#)
 - Add metadata for **physical distances**
 - useful for upcoming methods like Project 4: finding overlapping cells in z-dimension
 - Document and make **reproducible** using [spatial-sandbox](#)

- **Updates**

- SpatialData with physical coordinates done for **Xenium (breast)** and **Molecular Cartography (liver)**
 - Documentation of the physical distance metadata sometimes **not there or hard to find**
 - WIP: lung (Wouter-Michiel), mouse brain (Luca), heart (Benjamin)



Project 3: #segmentation-x

- Project motivation

- How do we segment the un-segmentable?**

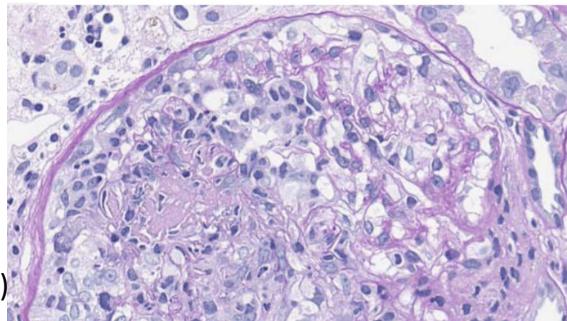
- Deep learning / expert annotation / membrane stains / segmentation-free?

- Optimize parameters and setup pipelines

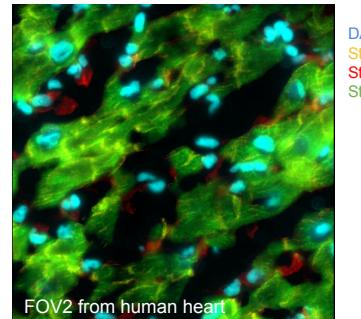
- Current tasks we are working on



SSAM + Baysor on transcripts from human heart MERFISH data (Kuppe)



Large scale data segmentation with Cellpose 2.0
(heart/in_situ_sequencing/human-in-the-loop approach)



Testing mesmer with membrane stains on heart data
+ Include Mesmer and JSTA in TXSim



Testing Cellpose2 on Kuppe data (human-in-the-loop retraining)

Project 4: #eval-overlapping-cells

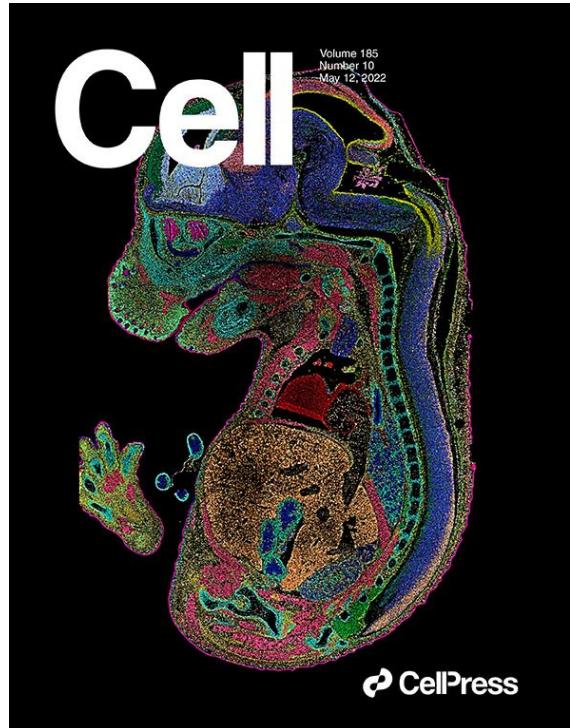
- **Project motivation:** Overlapping cells cause problems in analysing spatial gene expression.
- **Our solution:** Utilise Z-plane information to find cells that overlapping
- **Updates**
 - Adapting a script to make it work on our different data sets
 - Script close to done, 1st data set analyzed
 - Working on visualization automatization

Project 5: #brain-evaluation

- **Project motivation:** quantitative and qualitative interpretation on quality metrics for single-cell vs spatial gene expression is not trivial due to lack of general understanding of these metrics
- **Our solution:**
 - Evaluate single cells of the brain tissue of mouse and/or human
 - Create vignettes for metric measures in the TXsim pipeline.
- **Updates**
 - Run the TXsim pipeline on test dataset:
 - Resolution differences between watershed vs cellpose
 - Single cell type annotation in the spatial context
 - Modify cell sizes based on type
 - Understanding metric scripts, trying different parameters and running on test datasets.

Project 6: #stereoseq

- **Project motivation - what is the problem?**
 - Stereoseq is a new and exciting spatial technology!
 - Currently not supported in our TXsim workflow
- **Our Solution**
 - **Stere-seq:** [Xia et al. 2022] [Chen et al. 2022]
 - Enables cell type and cell-subtype characterization and profiling based on both transcriptome and spatial information.
 - Integrate different segmentation tools (eg:`ssam`, `msmer`) to work with full transcriptome data
- **Updates**
 - Data Compilation and software setup
 - Troubleshooting
 - Initial sanity checks of the input data-set



Project 7: #hyperbolic_brain



- **Project motivation - what is the problem?**
 - The cell type segmentation and annotation process can still be improved
- **Our solution**
 - Applying hyperbolic space embedding and distance metrics as an alternative to the classical euclidean-based segmentation and annotation pipeline
- **Updates**
 - Obtained Xenium Human Brain tissue dataset and applied a simple `squidpy` analysis and visualization process
 - Started the algorithm to embed the transcript levels data to hyperbolic space

Summary



- A hackathon project with 63 participants present certain challenges
 - Despite all issues, things have fallen into place
 - Integrating online participants isn't easy (sorry!)

(Bio)Schemas4NFDI, lightweight domain metadata (not only) for NFDI consortia

Steffen Neumann (IPB Halle) and Leyla Jael Castro (ZB MED)

Data Provider Department



The screenshot shows the nmrXiv interface. On the left, there's a sidebar with 'Your Space' containing 'Dashboard', 'Shared with me', 'Recent', 'Starred', and 'Trash'. The main area is titled 'IPB HALLE DASHBOARD' and shows two projects:

- Die Polei-Minze im Wandel der Zeiten** (Draft)
NMR dataset of (-)-Pulegon isolated from *Mentha pulegium*, isolated from the essential oil of *Mentha pulegium* (Pennyroyal) o...
Last updated on December 12, 2022
Created on October 24, 2022
- Linderazulen aus einer invasiven Pflanze. Delphi und sein violette Wunder.** (Draft)
NMR dataset of Linderazulene isolated from *Smyrnium perfoliatum*, collected close to the Monument to the Battle of Nations.

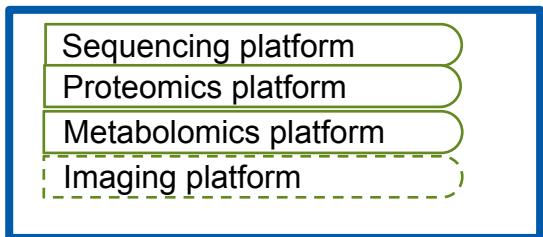
Achievements:

- Describe nmrXiv as [DataCatalog](#)
- Sample -> BioSample
- Improve MeasurementTechnique & keywords

HELP !!

Fix Bioschemas validator
[ElixirTeSS/bioschemas-validator](#)
Fix keywords as definedTerm

Data Provider Department



Achievement: 1st Bioschemas crosswalk and model draft

Lesson learned: Google Dataset Search only cares for Dataset, DataCatalog and DataDownload



Google
Dataset
Search



[OmicsDI](#)

Crosswalking Department



“For models defined in LinkML, what is the best way to export dataset metadata into Bioschemas JSON-LD for inclusion in data provider pages.”

Achievements:

- Dataset crosswalk [spreadsheet](#) between [GHGA's](#) metadata YAML & Bioschemas
- Added “exact_mapping” elements to [a hackathon copy of ghga.yaml](#)
- Developed [prototype LinkML code](#) for the translation from YAML to JSON-LD

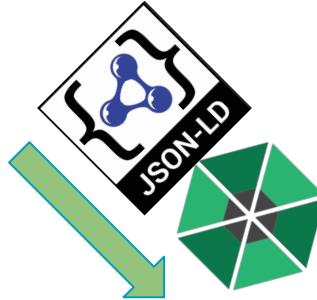
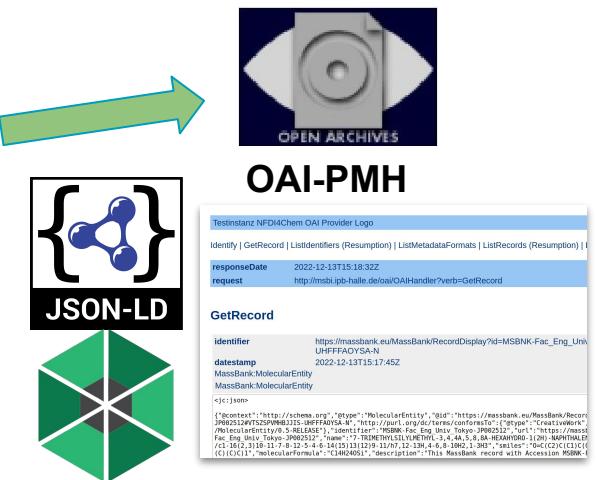
Lessons learned:

- LinkML maintainer(s) have [raised the point](#) that the original [crosswalk.yaml](#) may not be sufficient for a common solution.

Querying / Harvesting Department



MassBank
High Quality Mass Spectral Database



HELP !!
~~python xtree~~
~~xpath and namespaces~~



Achievements:

- MassBank in OAI-PMH
- Harvesting bioschemas through OAI-PMH

Google
Dataset
Search

Lessons learned:

- Only indexing **DataSet**
- Re-indexing needs **Last-changed-date**

Towards FAIR Computational Metabolomics Workflows - Improving Provenance Collection

Mahnoor Zulfiqar (University of Jena)

Kristian Peters (IPB Halle)

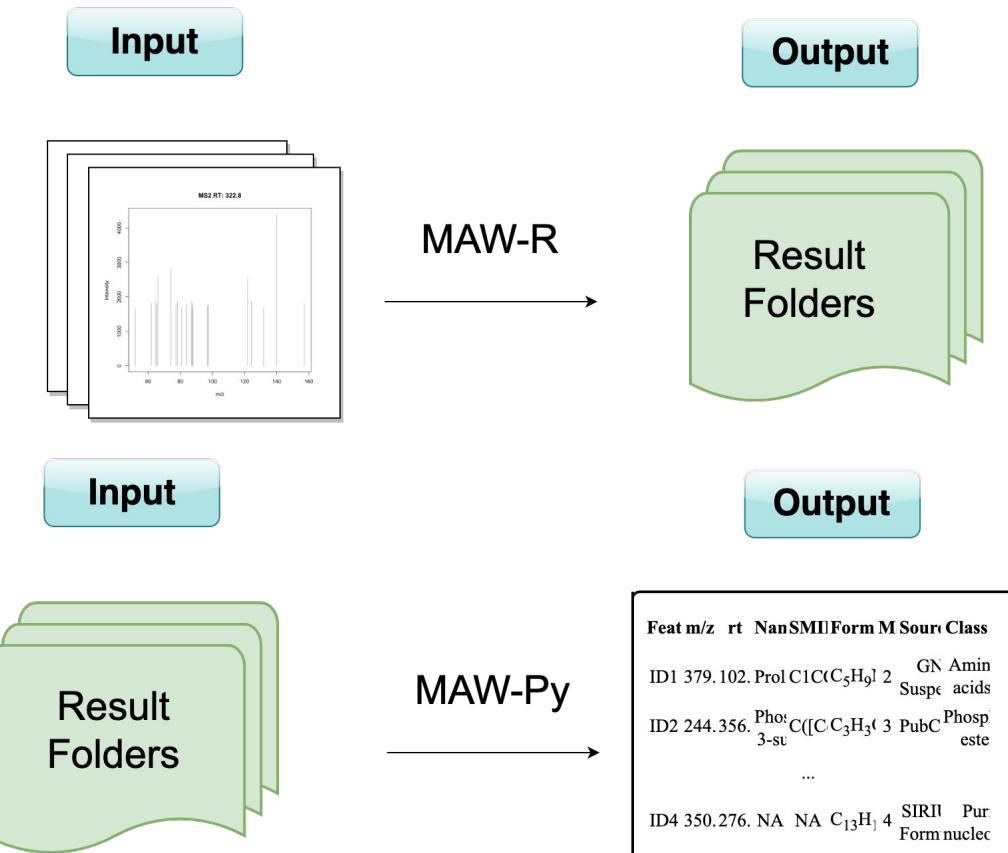
Michael R. Crusoe (ELIXIR-DE)

Example Workflow



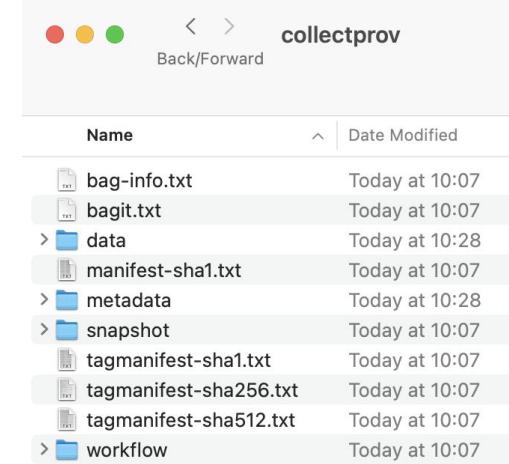
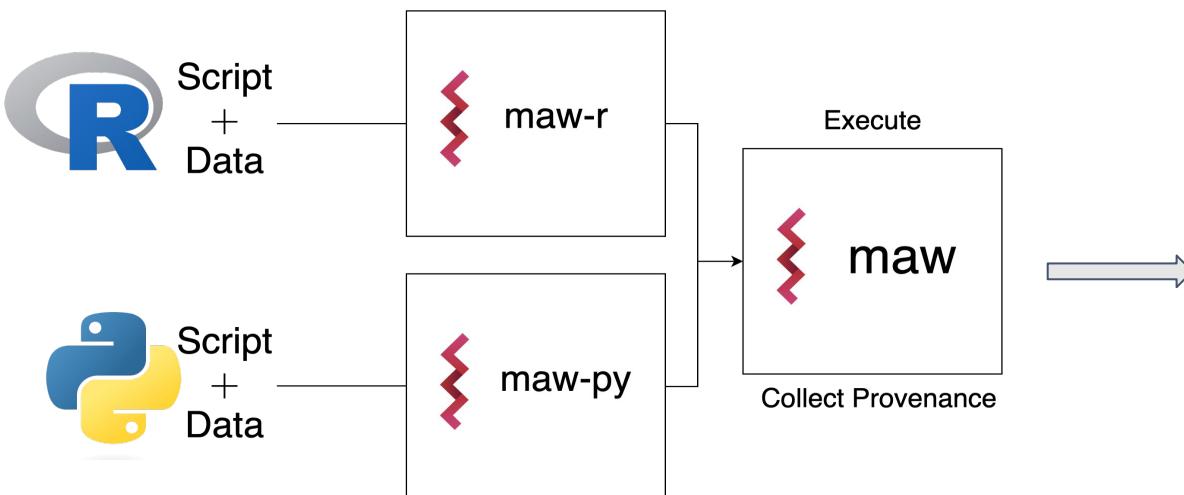
Metabolome Annotation Workflow

<https://github.com/zmahnoor14/MAW>



Progress

- Goals achieved:
 - Wrote CWL CommandLineTool descriptions for the [R step](#) and the [Python step](#)
 - Combined those steps into a CWL workflow [MAW](#)
 - Upgraded the [Docker container for the R step](#) to not need to be run interactively; rebuilt for linux/amd64
 - Collected workflow execution provenance using [CWLprov](#)



Goals



- Further goals
 - Convert the CWLProv RO-Bundle to a [Workflow Run RO-Crate](#) using [runcrate](#)
 - Add annotations using the EDAM ontology
 - Generate provenance of the separate analysis steps with the R step and the Python step; explore how to pass that information to CWLProv/RO-Crate
 - Run the workflow using toil-cwl-runner on the ARA Slurm cluster

Find us here:

SR 8 (2nd floor)

Slack channel: #fair-workflow

Interactive data analysis and visualization in the web browser

Asis Hallab (FZ Jülich) and Ata Ul Haleem (FZ Jülich)

Project turns out to be an ugly wallflower



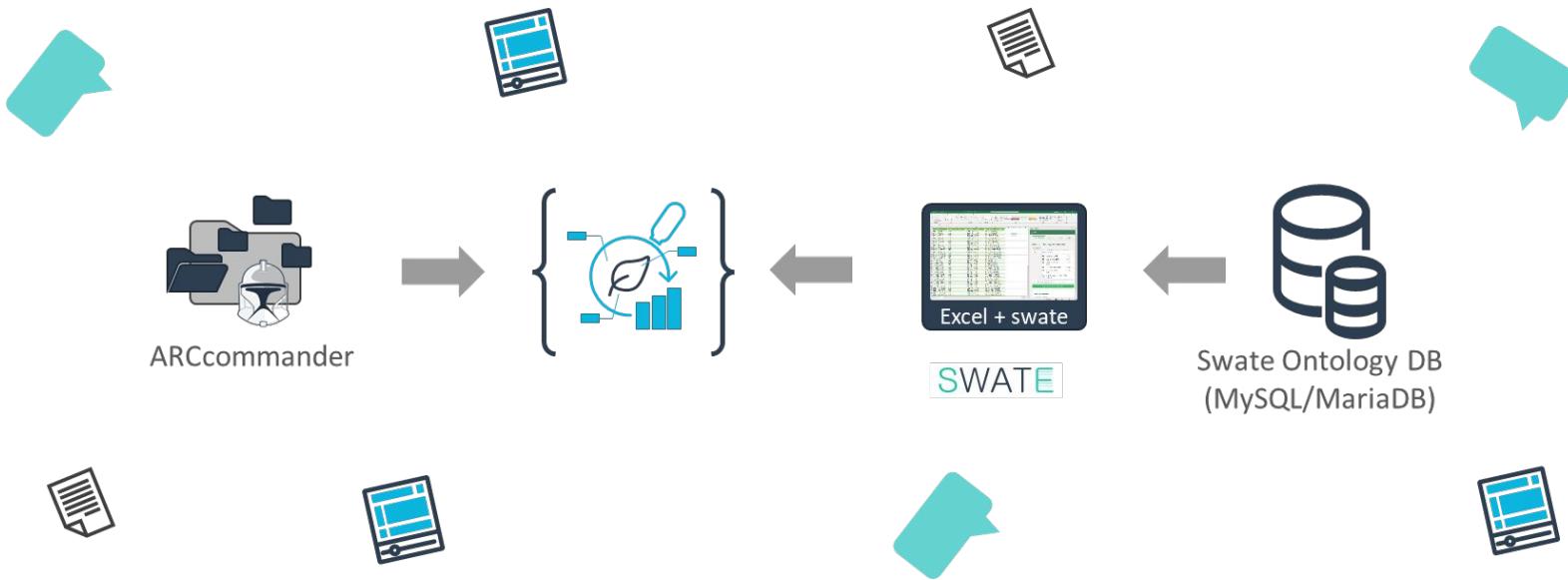
No advances made;
we are sorry.

But hey,
my luggage arrived and
my co-lead's son is getting better..

DataPLANT - Facilitating Research Data Management to combat the reproducibility crisis

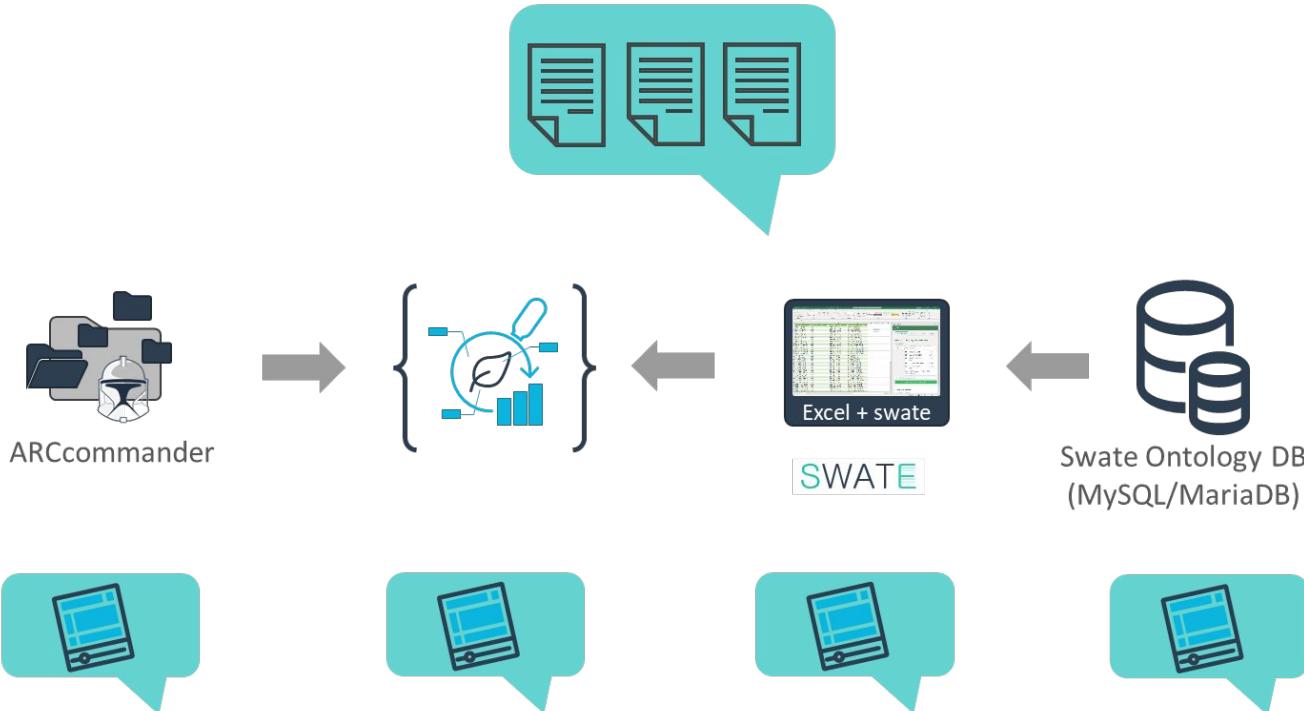
Elisa Senger (FZ Jülich) and Lukas Weil (TU Kaiserslautern)

Centralization of training materials



Centralization of training materials

Centralized KnowledgeBase with tutorials for consumers



Tool specific Documentation for developers and data stewards

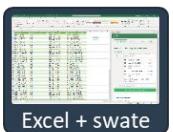
Centralization of training materials

Centralized KnowledgeBase with tutorials for consumers

The screenshot shows the deNBI KnowledgeBase homepage. On the left, there is a sidebar with navigation links: 'Fundamentals' (Introduction, Research Data Management, Data Management Principles, Measures, Data Models, Data Protection, Data Management Plan, Version Control, etc.), 'Implementation within DataPLANT' (DataPLANT Overview, DataPLANT Components, Source + workflow annotation tool for DataPLANT, DataPLANT, our Data Management Plan Generator), and 'Training & Tutorials' (Quickstart on ARCs, ARCCommander Quickstart, ARCCommander RBS). The main content area is titled 'Home' and includes sections for 'Welcome to the deNBI/KNIT Knowledge Base!', 'Feedback & Contribution', and 'For all other contributions, please refer to the Contribution guide.'



ARCCommander



SWATE



Swate Ontology DB
(MySQL/MariaDB)

The screenshot shows the ARCCommander Dev Docs index page. It includes a table of contents for 'ARC Commander Dev Docs' and a section for 'Index' with a note about last update.

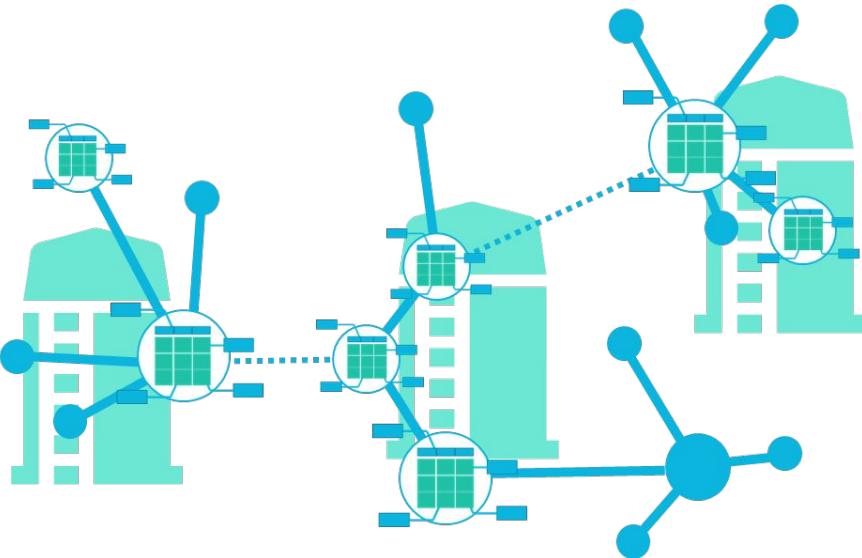


The screenshot shows the Feature Documentation page for the Swate Ontology DB. It includes sections for 'Feature Documentation' (Data, Building Sets, Creating Annotations, Adding Objects, Adding Cells, Adding Cells with ontology terms, Adding Cells with term search, Templates, T-ISA-JSON), 'Tutorials' (Data, Project Description, Development Setup, Project Revisions), and 'DataPLANT Support'.

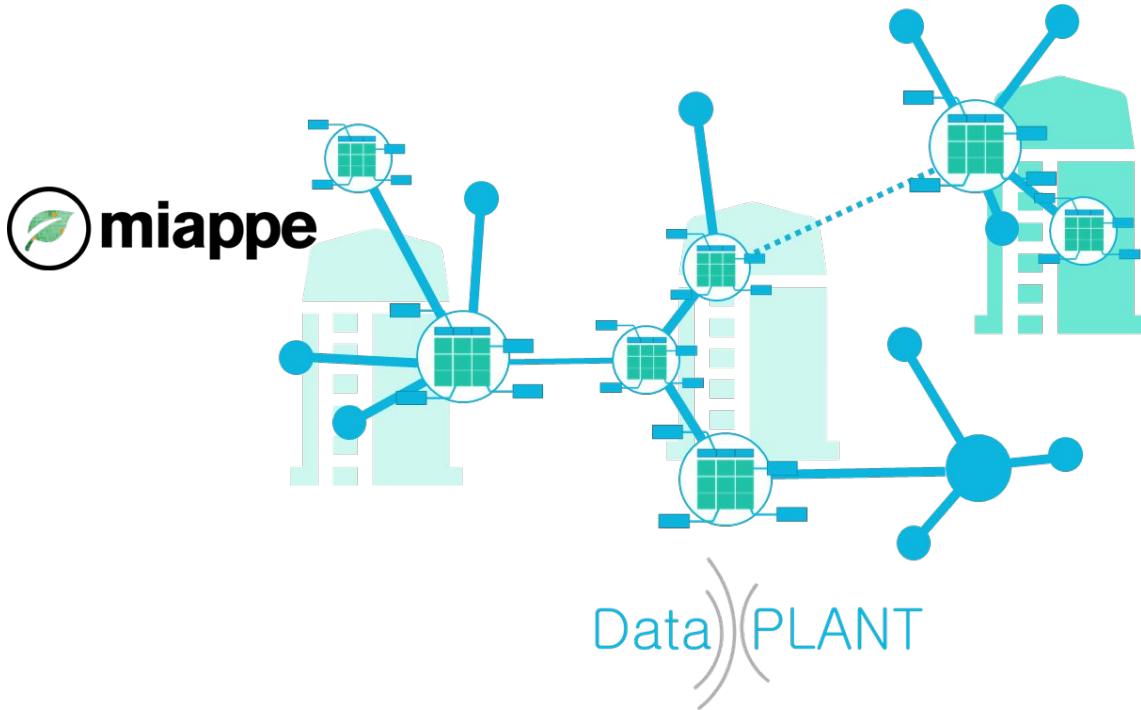


Tool specific Documentation for developers and data stewards

Creating a MIAPPE template



Creating a MIAPPE template



Creating a MIAPPE template

The screenshot illustrates the process of creating a MIAPPE template, specifically focusing on the 'datafile / sample' section. The data is organized into a table with columns: Source Name, Characteristics [sample label], Factor [temperature unit], and Data File Name.

Annotations:

- source**: A box labeled 'source' is placed over the first column of the table.
- characteristic**: A box labeled 'characteristic' is placed over the second column.
- factor**: A box labeled 'factor' is placed over the third column.
- datafile / sample**: A box labeled 'datafile / sample' is placed over the fourth column.

Annotation building block selection dialog (Swate):

The 'Swate' window shows the 'new parameter' annotation building block selection. It lists several annotation building blocks:

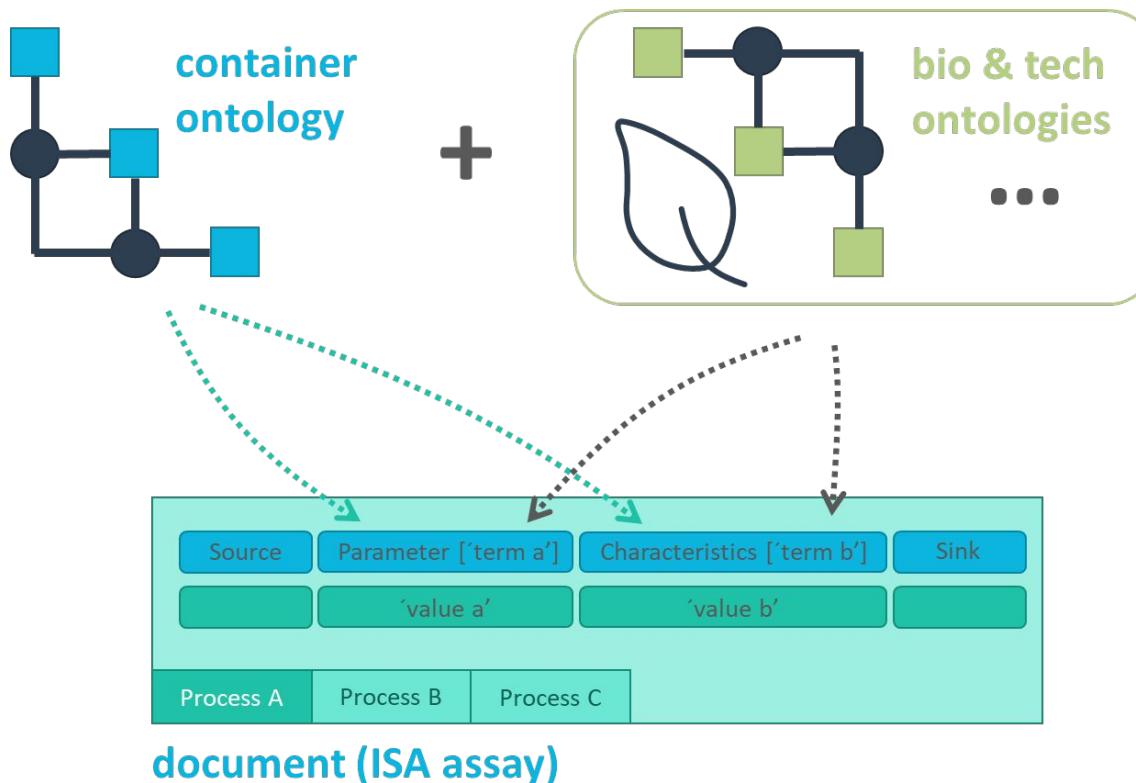
- instrument
- instrument model
- instrument vendor
- medical instrument
- MascotInstrument

Below the list, there is a search bar and a note: "Cant find the Term you are looking for? Use Advanced Search".

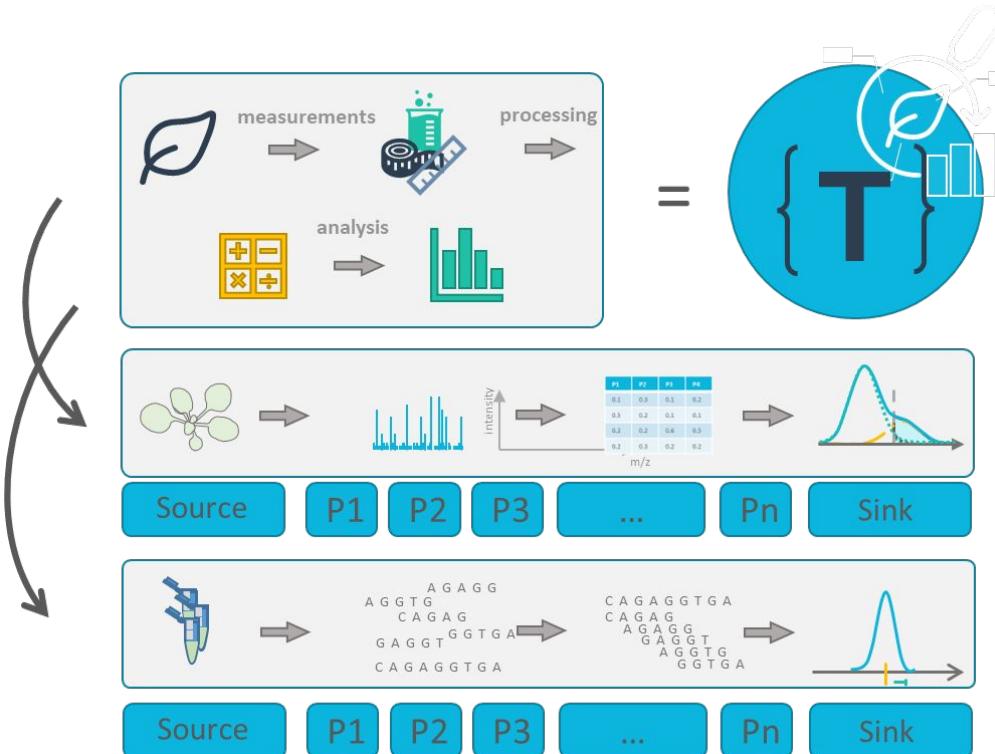
Excel Table Data:

Source Name	Characteristics [sample label]	Factor [temperature unit]	Data File Name
Heat_15A_OD_R1	15N	32.00 degree Celsius	Heat_15A_OD_R1.wiff
Heat_15A_OD_R2	15N	32.00 degree Celsius	Heat_15A_OD_R2.wiff
Heat_180A_OD_R1	15N	32.00 degree Celsius	Heat_180A_OD_R1.wiff
Heat_180A_OD_R2	15N	32.00 degree Celsius	Heat_180A_OD_R2.wiff
Heat_2880A_OD_R1	15N	32.00 degree Celsius	Heat_2880A_OD_R1.wiff
Heat_2880A_OD_R2	15N	32.00 degree Celsius	Heat_2880A_OD_R2.wiff
Heat_5760A_OD_R1	15N	32.00 degree Celsius	Heat_5760A_OD_R1.wiff
Heat_5760A_OD_R2	15N	32.00 degree Celsius	Heat_5760A_OD_R2.wiff
Heat_5760A_15D_R1	15N	32.00 degree Celsius	Heat_5760A_15D_R1.wiff
Heat_5760A_15D_R2	15N	32.00 degree Celsius	Heat_5760A_15D_R2.wiff
Heat_5760A_180D_R1	15N	32.00 degree Celsius	Heat_5760A_180D_R1.wiff
Heat_5760A_180D_R2	15N	32.00 degree Celsius	Heat_5760A_180D_R2.wiff
Heat_5760A_2880D_R1	15N	32.00 degree Celsius	Heat_5760A_2880D_R1.wiff
Heat_5760A_2880D_R2	15N	32.00 degree Celsius	Heat_5760A_2880D_R2.wiff
Cold_15A_OD_R1	15N	4.00 degree Celsius	Cold_15A_OD_R1.wiff
Cold_15A_OD_R2	15N	4.00 degree Celsius	Cold_15A_OD_R2.wiff
Cold_180A_OD_R1	15N	4.00 degree Celsius	Cold_180A_OD_R1.wiff
Cold_180A_OD_R2	15N	4.00 degree Celsius	Cold_180A_OD_R2.wiff
Cold_2880A_OD_R1	15N	4.00 degree Celsius	Cold_2880A_OD_R1.wiff
Cold_2880A_OD_R2	15N	4.00 degree Celsius	Cold_2880A_OD_R2.wiff
Cold_5760A_OD_R1	15N	4.00 degree Celsius	Cold_5760A_OD_R1.wiff
Cold_5760A_OD_R2	15N	4.00 degree Celsius	Cold_5760A_OD_R2.wiff
Cold_5760A_15D_R1	15N	4.00 degree Celsius	Cold_5760A_15D_R1.wiff
Cold_5760A_15D_R2	15N	4.00 degree Celsius	Cold_5760A_15D_R2.wiff
Cold_5760A_180D_R1	15N	4.00 degree Celsius	Cold_5760A_180D_R1.wiff
Cold_5760A_180D_R2	15N	4.00 degree Celsius	Cold_5760A_180D_R2.wiff
Cold_5760A_2880D_R1	15N	4.00 degree Celsius	Cold_5760A_2880D_R1.wiff
Cold_5760A_2880D_R2	15N	4.00 degree Celsius	Cold_5760A_2880D_R2.wiff
Cold_5760A_5760D_R1	15N	4.00 degree Celsius	Cold_5760A_5760D_R1.wiff
Cold_5760A_5760D_R2	15N	4.00 degree Celsius	Cold_5760A_5760D_R2.wiff
Highlight_15A_OD_R1	15N	22.00 degree Celsius	Highlight_15A_OD_R1.wiff
Highlight_15A_OD_R2	15N	22.00 degree Celsius	Highlight_15A_OD_R2.wiff
Highlight_180A_OD_R1	15N	22.00 degree Celsius	Highlight_180A_OD_R1.wiff
Highlight_180A_OD_R2	15N	22.00 degree Celsius	Highlight_180A_OD_R2.wiff

Creating a MIAPPE template



Creating a MIAPPE template



Creating a MIAPPE template



Swate

Building Blocks

Add annotation building blocks (columns) to the annotation table.

Parameter	experiment start
Experiment Start Date	NCIT:C90487
Definition: "The date on which an experiment begins." []	
Experiment	NCIT:C42790
Ad experiment name	
Experiment Name	NCIT:C181729
experiment performer	EFO:0000647
Can't find the Term you are looking for? Try Advanced Search!	
Still can't find what you need? Get in contact with us!	

[More about Parameter:](#)

Please select an ontology (optional)
Keywords: experiment start date

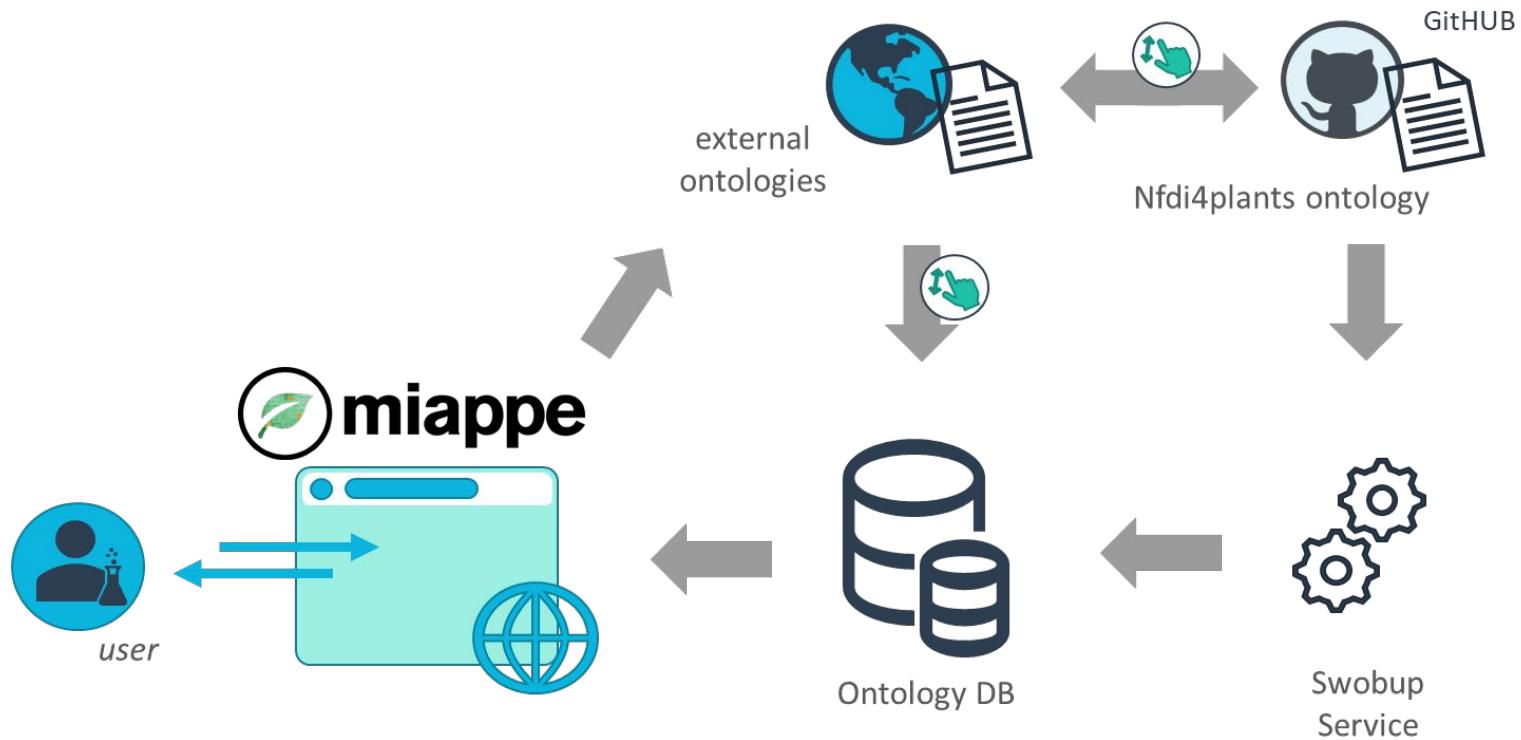
Terms with 'experiment start date' included in their label:

1. http://purl.obolibrary.org/obo/NCIT_C90487 (NCIT):
 - Experiment Start Date in Ontobee: NCIT
2. http://purl.obolibrary.org/obo/AGRO_00000684 (AGRO):
 - experiment start date in Ontobee: AGRO

B	E	H
Parameter [MIAPPE version]	Parameter [Experiment Start Date]	Parameter [Experiment End Date]
The version of MIAPPE used.	Date and, if relevant, time when the exp	Date and, if relevant, time when the exp
1.1	2002-04-04 2006-09-27T10:23:21+00:00	27.11.2002

A	D	G
Characteristic [Rooting medium]	Characteristic [Container type]	Characteristic [Container volume]
An abiotic plant treatment (EO:000719)	Type of container used to grow/treat	Volume that is available to the roots. XE
hydroponic plant culture media; in vitro	pot; Petri dish; well; tray	[L]
Plant Environment Ontology:'EO_0007	Text	Numeric

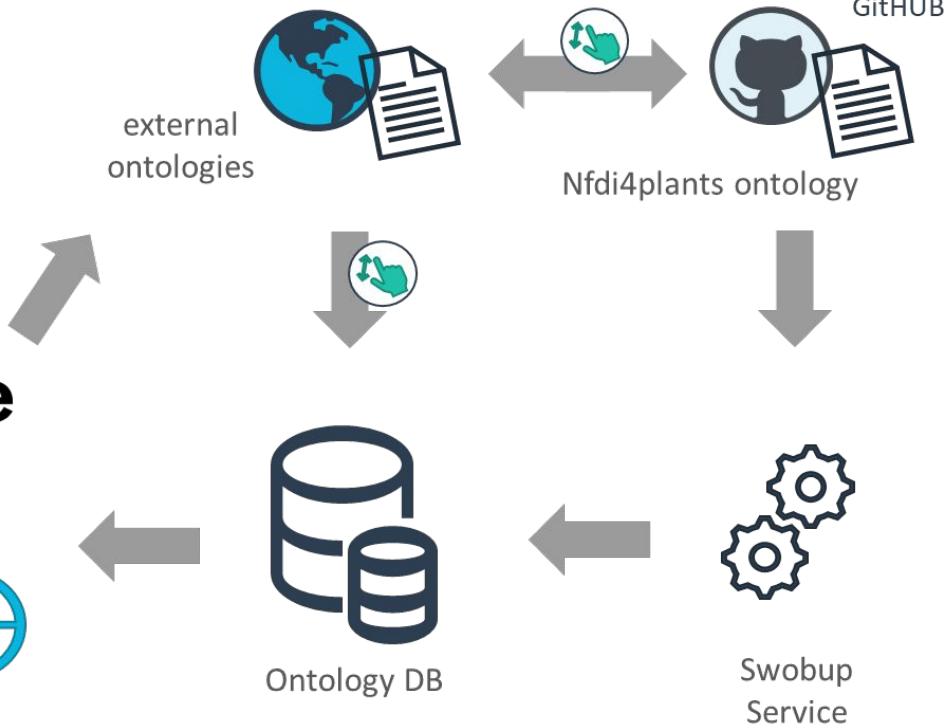
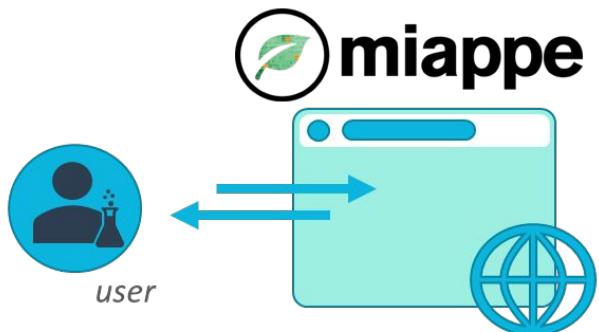
DataPlant Ontology Service



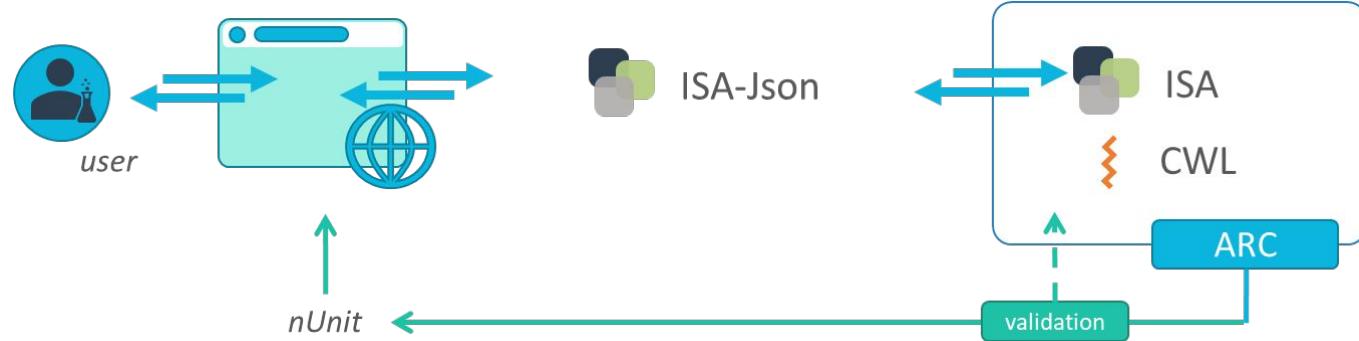
DataPlant Ontology Service



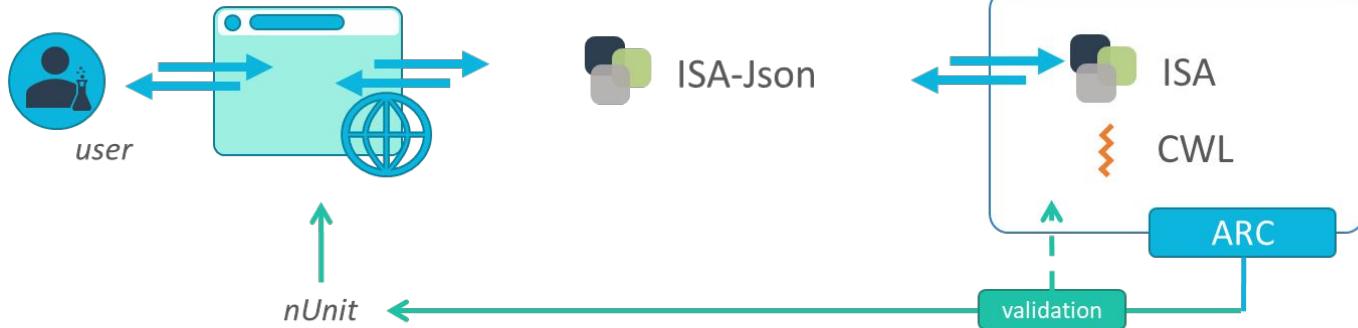
GET	/api/IOntologyAPIv2/getTestNumber	Test function to verify client server connection.
GET	/api/IOntologyAPIv2/getAllOntologies	Returns all ontologies found in the database.
POST	/api/IOntologyAPIv2/getTermSuggestions	Query the database for terms by their name.
POST	/api/IOntologyAPIv2/getTermSuggestionsByParentTerm	Query all children of parent_term for terms by their name.
POST	/api/IOntologyAPIv2/getAllTermsByParentTerm	Return all children of parent_term.
POST	/api/IOntologyAPIv2/getTermSuggestionsByChildTerm	Query all parents of child_term for terms by their name.
POST	/api/IOntologyAPIv2/getAllTermsByChildTerm	Return all parents of child_term.
POST	/api/IOntologyAPIv2/getTermsForAdvancedSearch	Return all parents of child_term.
POST	/api/IOntologyAPIv2/getTreeByAccession	Returns tree of closest 20 terms.



Centralized interfaces



Centralized interfaces



```
<?xml version="1.0" encoding="utf-8"?>
<test-results date="2022-12-14" name="fsi" total="6" errors="0" failures="1" ignored="0" not-run="0" info="0" test-time="0.080" test-time-unit="second">
  <environment executor-version="9.0.4" clr-version="7.0.0" os-version="Microsoft Windows NT 10.0.19044.0" culture-info="FSI_0000.SummaryfileWriter+XAttr@008-1[System.Object]de-DEFSI_0000.SummaryfileWriter+XAttr@008-1[System.Object]" />
  <test-suite type="Assembly" name="fsi" executed="True" result="Failure" success="False" time="0.080" results="1">
    <test-case name=" [ ISA; Semantic; Assay; Term ] " executed="True" result="Failure" success="False" failure="true">
      <failure>
        <message><![CDATA[Actual entity is not valid: Path: C:\Users\Admin\testARC\assays\assay1\assay1</message>
      </failure>
    </test-case>
    <test-case name=" [ ISA; Schema; Study; SourceNameColumn ] " executed="True" result="Success" success="True" />
    <test-case name=" [ ISA; Schema; Study; SampleNameColumn ] " executed="True" result="Success" success="True" />
    <test-case name=" [ ISA; Schema; Assay ] " executed="True" result="Success" success="True" time="0.001" />
    <test-case name=" [ ISA; Semantic; Assay; Term ] " executed="True" result="Success" success="True" />
    <test-case name=" [ ISA; Plausibility; Study; Factor ] " executed="True" result="Success" success="True" />
  </results>
</test-suite>
</test-results>
```

	Name	Date modified
▼ Today		
	isa.investigation.xlsx	14.12.2022 09:44
	.arc	14.12.2022 09:44
	assays	14.12.2022 09:44
	runs	14.12.2022 09:44
	studies	14.12.2022 09:44
	workflows	14.12.2022 09:44

A	B	C	D	E	F
Source Name	Parameter	Term Source	Term Accession	Parameter	Unit
run_35_A	control	EFO	http://purl.org/0.25	degree	Celsius
run_35_B	control	EFO	http://purl.org/0.25	degree	Celsius
run_35_C	control	EFO	http://purl.org/0.25	degree	Celsius
run_40_A	control	EFO	http://purl.org/0.25	degree	Celsius
run_40_B	control	EFO	http://purl.org/0.25	degree	Celsius
run_40_C	control	EFO	http://purl.org/0.25	degree	Celsius
run_35_A	control	EFO	http://purl.org/0.25	degree	Celsius
run_35_B	control	EFO	http://purl.org/0.25	degree	Celsius
run_35_C	control	EFO	http://purl.org/0.25	degree	Celsius
run_40_A	control	EFO	http://purl.org/0.25	degree	Celsius
run_40_B	treatment	OGMS	http://purl.org/0.35	degree	Celsius
run_35_C	treatment	OGMS	http://purl.org/0.35	degree	Celsius
run_35_A	treatment	OGMS	http://purl.org/0.35	degree	Celsius
run_35_B	treatment	OGMS	http://purl.org/0.35	degree	Celsius
run_35_C	treatment	OGMS	http://purl.org/0.35	degree	Celsius
run_35_A	treatment	OGMS	http://purl.org/0.35	degree	Celsius
run_35_B	treatment	OGMS	http://purl.org/0.35	degree	Celsius

The ELIXIR::GA4GH Cloud

Mohsen Pourjam (TUM)
Alexander Kanitz (ELIXIR-CH)

- Pipeline to **process** 16S rRNA gene sequences
- Deciphering **microbial composition** from 16S rRNA

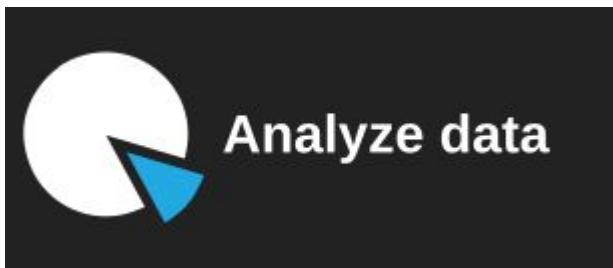
What if we could enrich our hypothesis with datasets acquired with the same hypotheses?

- How can I **find** datasets produced to address the similar hypothesis?
- How can I **process** my data as well as other projects data?
- Would my results be **comparable** to other projects results?
- How can I **store** it with relevant metadata?
- How can I **share** or use shared data?

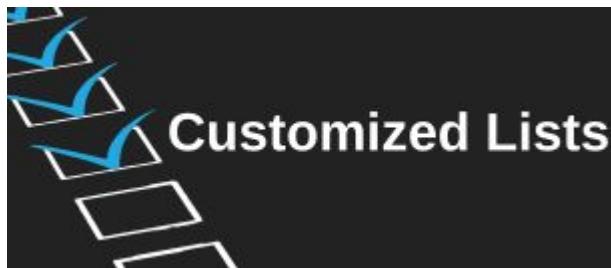
What does IMNGS2 provide?



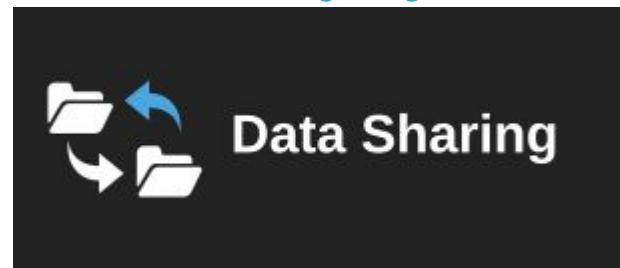
www.imngs2.org



Analyze data



Customized Lists



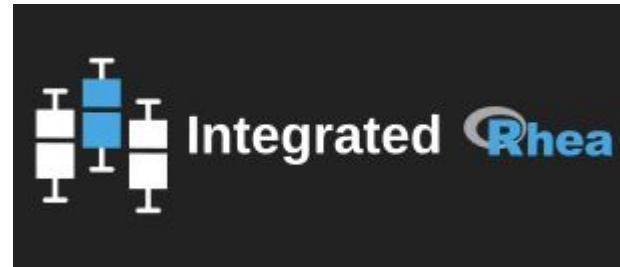
Data Sharing



Massive Integration

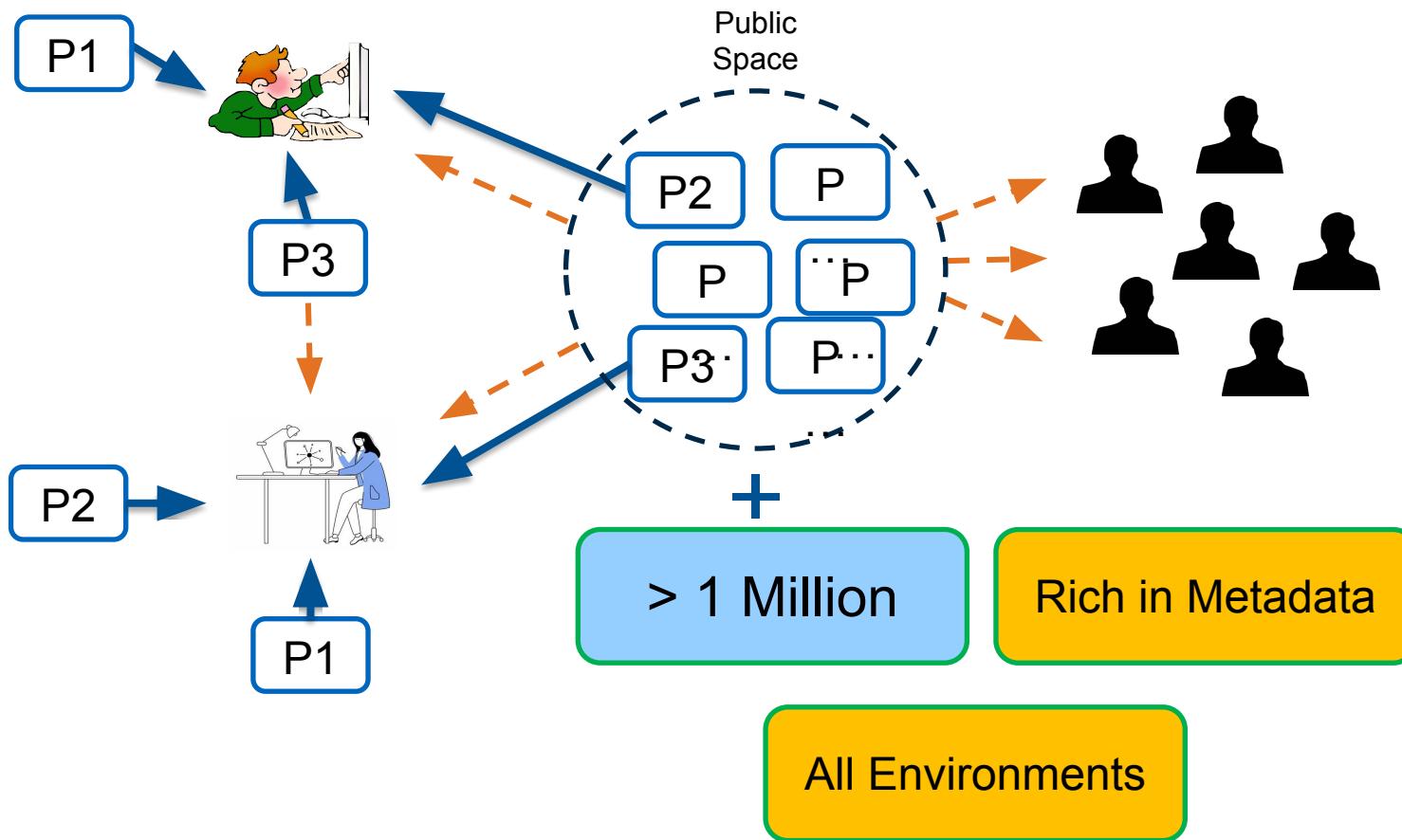


TiC



Integrated Rhea

IMNGS2:



Where is the problem?



- 3 min * 1,050,000 ≈ 6 years
- The solution is **parallelization**
- Still the problem is there → lack of infrastructure
- An unsuccessful try with ELIXIR Greece

Migrate to workflow language

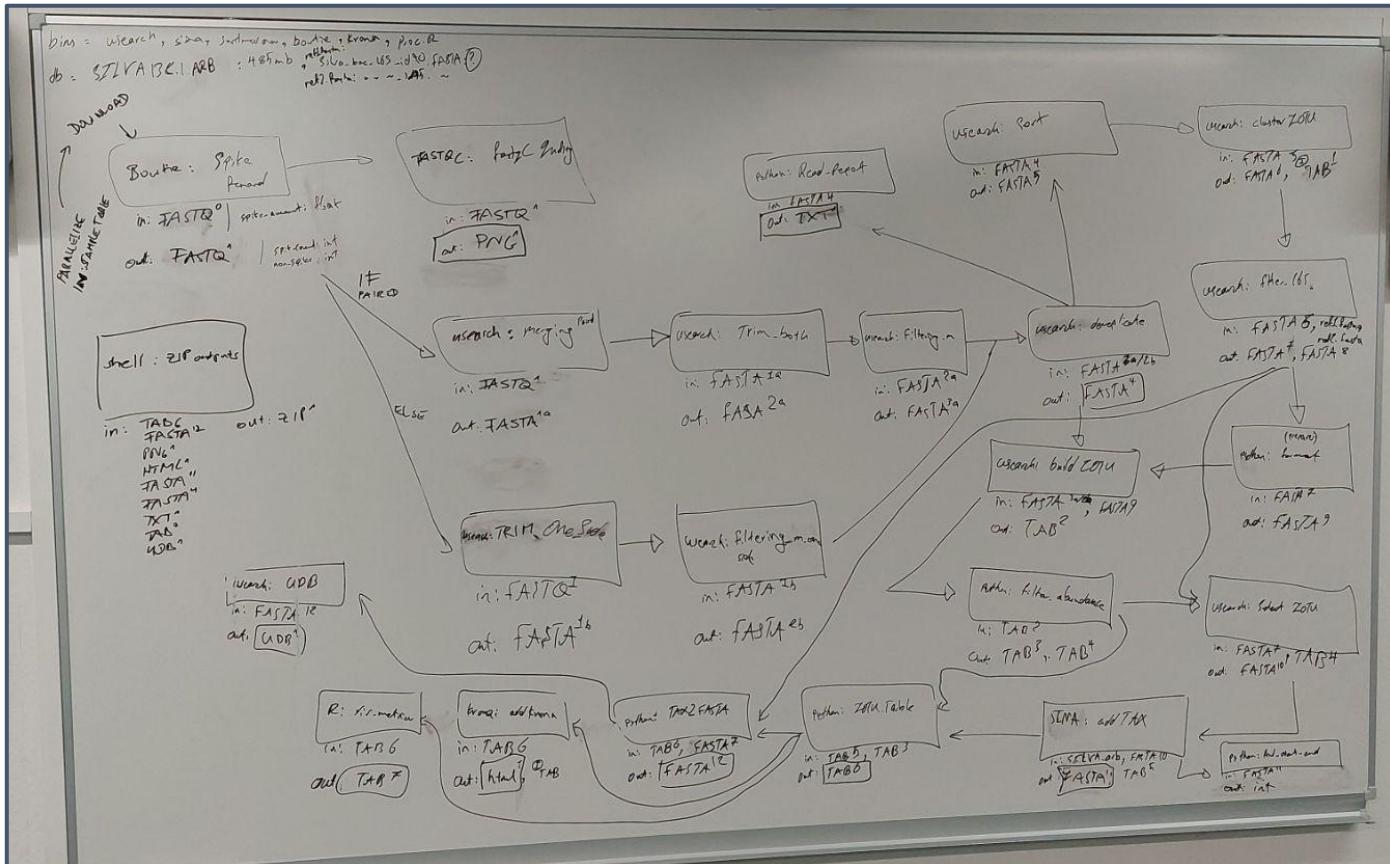
Why?

- Better for parallelization / horizontal scaling
- Resource provisioning per step allows optimizing resource use (highly relevant for >1 million jobs)
- Many execution options (HPC, cloud etc.)
- Allow partial (& automatic) reruns
- More portable (containerization; easier to deploy for others)
- Easier to maintain

Why Snakemake?

- Supported by ELIXIR::GA4GH Cloud
- Has GA4GH TES backend for task-level federation
- Supported by Amazon Genomics CLI as alternative
- Easy to migrate from Python-based pipeline; custom functions can be executed +/- as is via `run` directive, no need to “scriptify”

Step 1: Manual DAG :)



Step 2: Identify container images



Containerized workflows are most portable!

- Bowtie 2: `quay.io/biocontainers/bowtie2:2.4.4--py37hafa4d4c_1` /
- Krona: `quay.io/biocontainers/krona:2.7.1--pl526_5`
- SINA: `quay.io/biocontainers/sina:1.6.0--hc7f9b0f_1`
- SortMeRNA: `quay.io/biocontainers/sortmerna:4.2.0--h9ee0642_1`
- R: `r-base:3.5.2` or `rocker/r-ver:3.5.2` (would prefer the latter, see [here](#))
- Python: `python:3.7.16-slim-bullseye`

✓ **BioContainers or official images** available for tools* 😊
⇒ ***Great for reproducibility!***

(Custom images may or may not be there tomorrow, next month, in 5 years)

* Well, almost all - more on that later

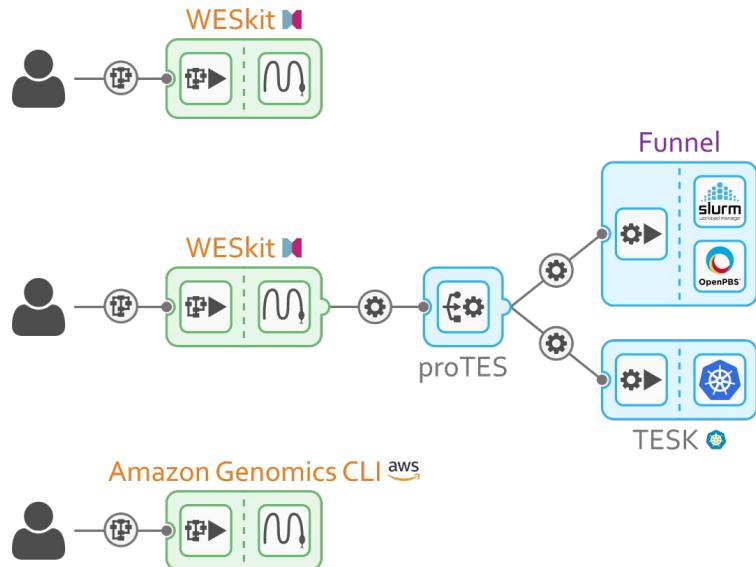
Step 3: Prepare set of test files



- ✓ *Set of input files prepared*

Next steps

- Generate Snakemake skeleton with all steps from manual DAG
- Divide up steps into individual issues
- Test execution on cloud infrastructure
 - **WESkit only**
@ ELIXIR DE
 - **WESkit + TES backend**
@ ELIXIR CZ/FI/GR
 - **Amazon Genomics CLI**
@ us-east-1
(because SRA s3 bucket is there)



Problems (1/2)



- **Current pipeline uses USEARCH, which is proprietary**
 - No FOSS!
 - Have license (>\$1k), but *can only run 1 process at a time*, so no parallelization; so either run 6 years OR pay up to \$1,000,000,000 for one million licenses :)
 - Alternative: use open source VSEARCH instead (but need refactoring)
- **Caching containers**
 - Some hundreds of Mb would need to be downloaded for each step if containers cannot be cached
 - Implement custom solution

Problems (1/2)



- **Staging persistent input files**
 - ~500 MB that would otherwise need to be downloaded for each step
 - Currently not possible based on specs ([issue exists](#))
 - Need to come up with custom solution in ELIXIR::GA4GH Cloud until interoperable solution is agreed upon for TES
- **Staging inputs & outputs, managing intermediates**
 - Sample files (FASTQ) are not big (~100 MB), but >1,000,000 of them still means moving around and storing >100 TB!
 - Find out what's the most efficient way of transferring data (download in each process vs batch download vs download all of them)
 - Delete intermediates upon completion; possibly move out results (but “only” 4 TB for all runs)



in the room



Who pays the bill? :)

ELIXIR may cover, but not yet sure, working on it

Could use help of...



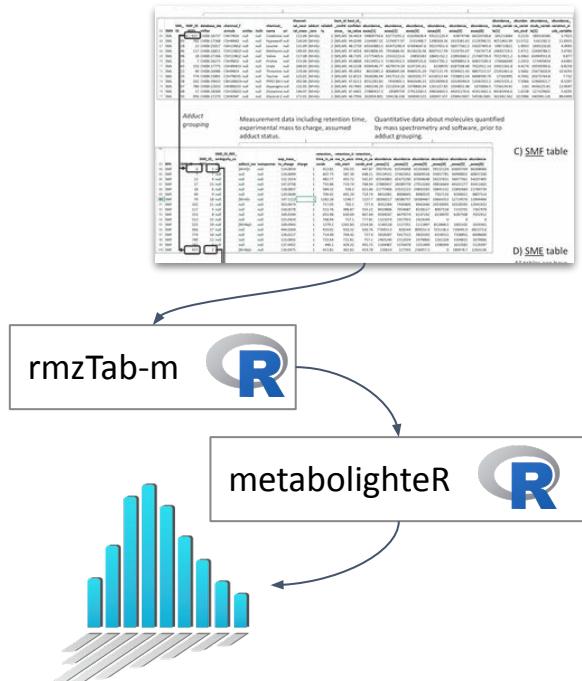
- Snakemake devs
- Anyone wishing to play with the ELIXIR::GA4GH Cloud :)

Improving interoperability support for authoring, editing and conversion of mzTab-M for Lipidomics Tools

**Eduardo Jacobo Miranda Ackerman (MPI-CBG Dresden),
Daniel Krause (FZ Borstel), Nils Hoffmann (FZ Jülich),
Olena Mokshyna (IOCB Prague), w/ support by Steffen
Neumann (IPB Halle)**

mzTab-M Support - progress

- LipidXplorer writes mzTab-M (Jacobo)
- development of lxFPostman support still ongoing (Daniel)
- R implementations
 - Google Colab Notebook for testing of conversion of mzTab-M to ISAtab for MetaboLights submission (using MTBL263 as an example) + mzMLs (Nils)
 - Writing of MAF file from mzTab-M (Steffen)
- MZmine 3 support (Olena)
 - testing of import function, documentation of current limits / issues
 - implementation and testing of export of mzTab-M
 - MZmine compound and spectral database annotations
- Help needed:
 - ISAtab creation (Study and assays from mzTab-M in R)
 - Configurable WebComponents for use in RShiny and Java Webapp for autocomplete of suitable CV terms (against existing REST API)
 - Anyone willing to participate in MZmine3 development (Java) is welcome

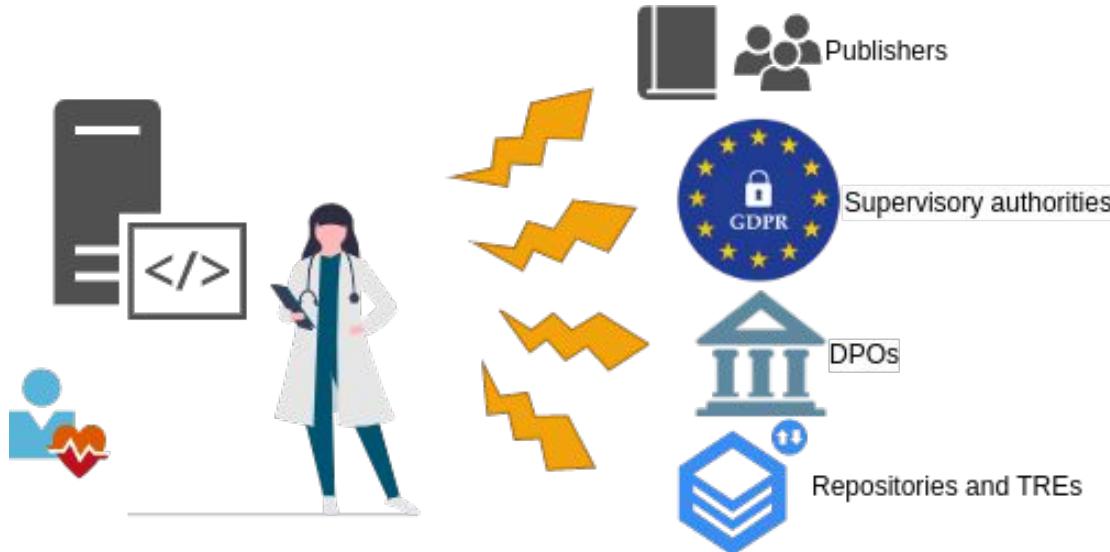


Towards a minimum information checklist for biomedical research projects with sensitive human data (only remote)

Pinar Alper, Vilém Děd, Christoph Kämpf, Valérie Barbié ,
Marina Popleteeva, Nene Djenaba Barry, Frédéric Erard

Recap

- The GDPR requires a written form for the Record of Processing Activities (Art. 30).
- The record **format can be chosen freely**, and it can be created on paper or numerically.
- Public organizations have to communicate the record to any person who demand it.

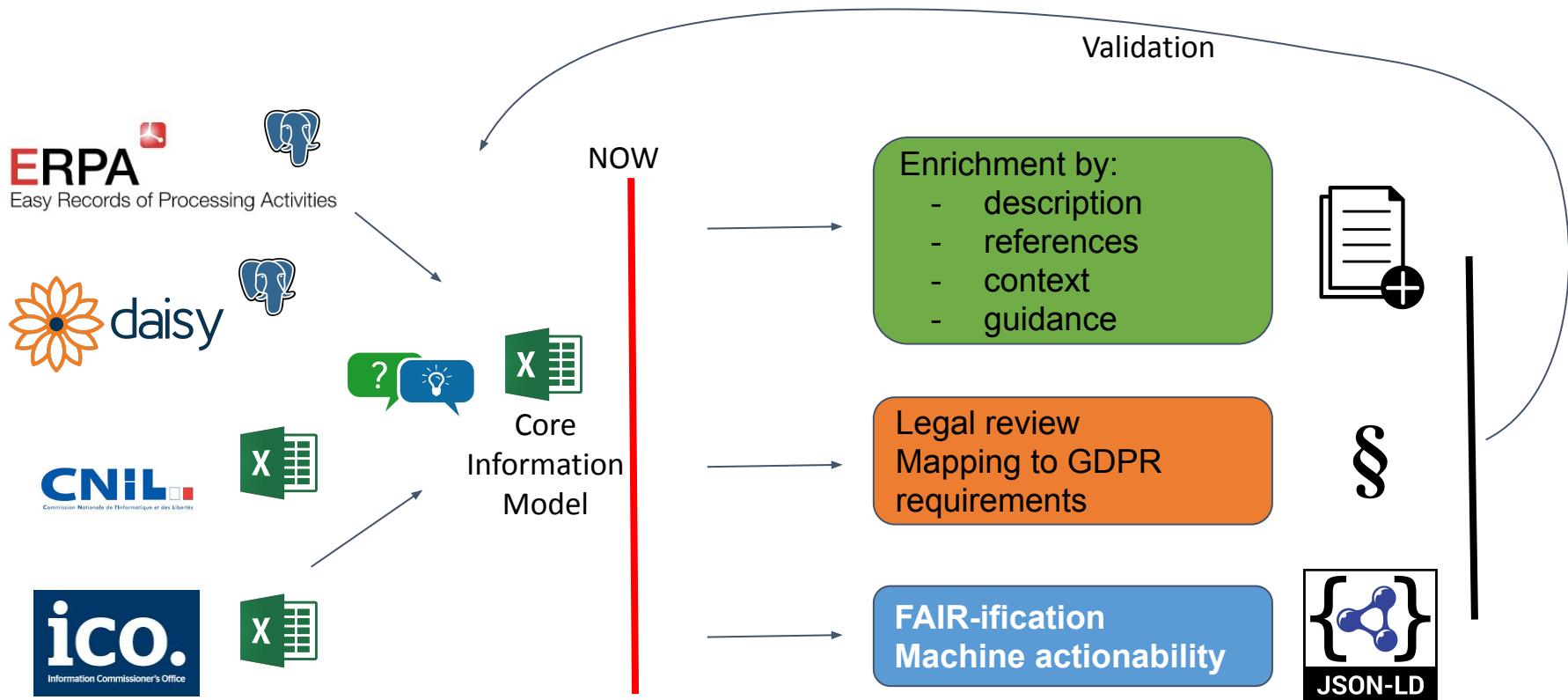


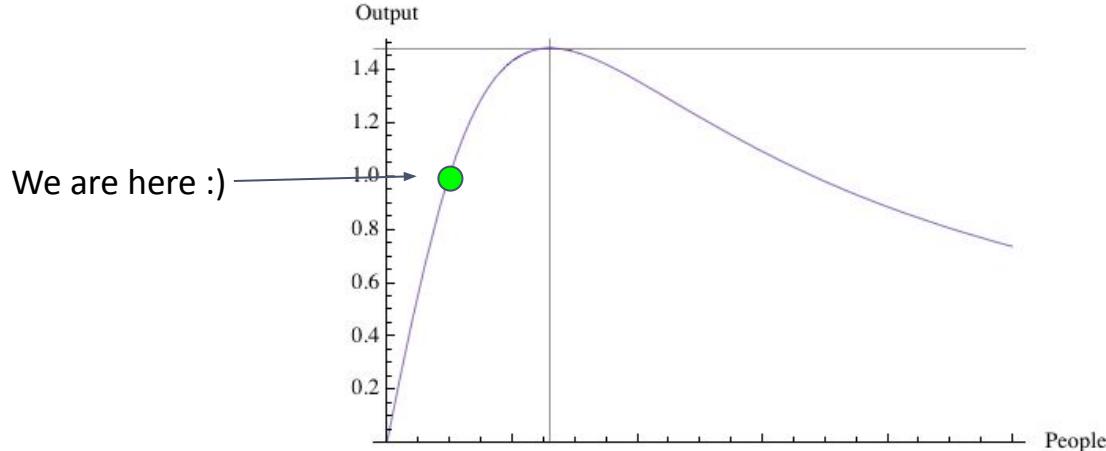
Reporting of sensitive data is big unknown for many stakeholders.

Lack of standards on format, content, scope, ...

Minimal Information for Record
Of Processing Activities
(MIROPA)

Current progress / Next steps





Slack:
#checklist_for_human_data



Feel free to join if you have:

- Knowledge of minimal information checklists
- Machine actionable representations for checklists (JSON, JSON-LD)
- Interest in GDPR and legal requirements for processing of sensitive data
- Experience in sensitive data publishing
- Experience in reporting data processing to supervisory authorities