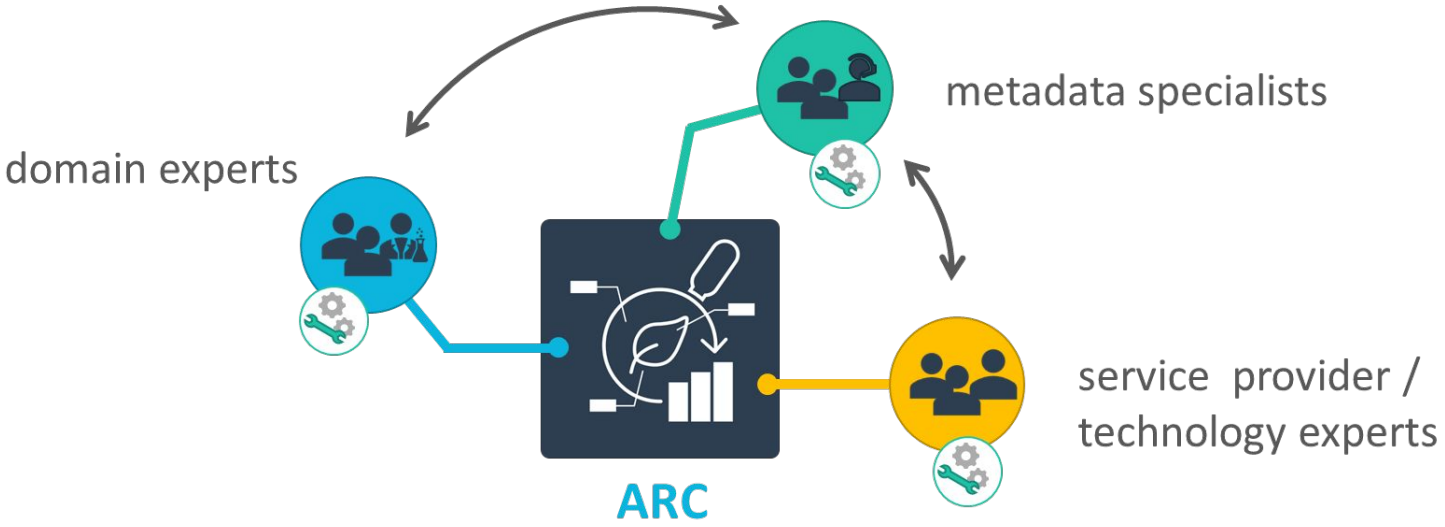
DataPLANT - Facilitating Research Data Management to combat the reproducibility crisis

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

Main Goal

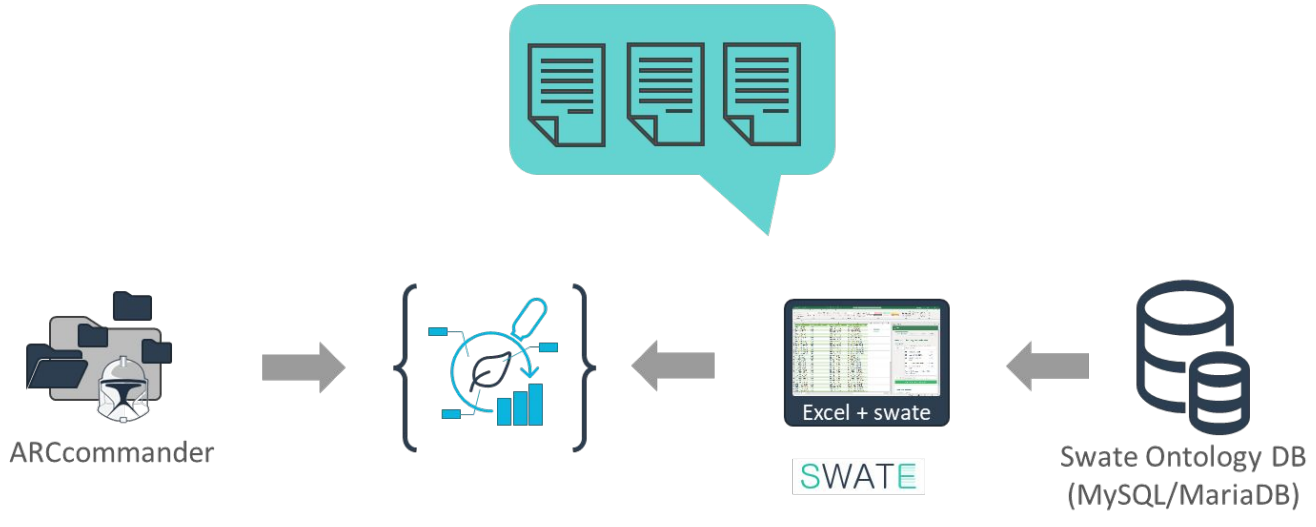


Make ARCs more accessible

- Improve Tool documentation
- Towards MIAPPE conforming ARCs

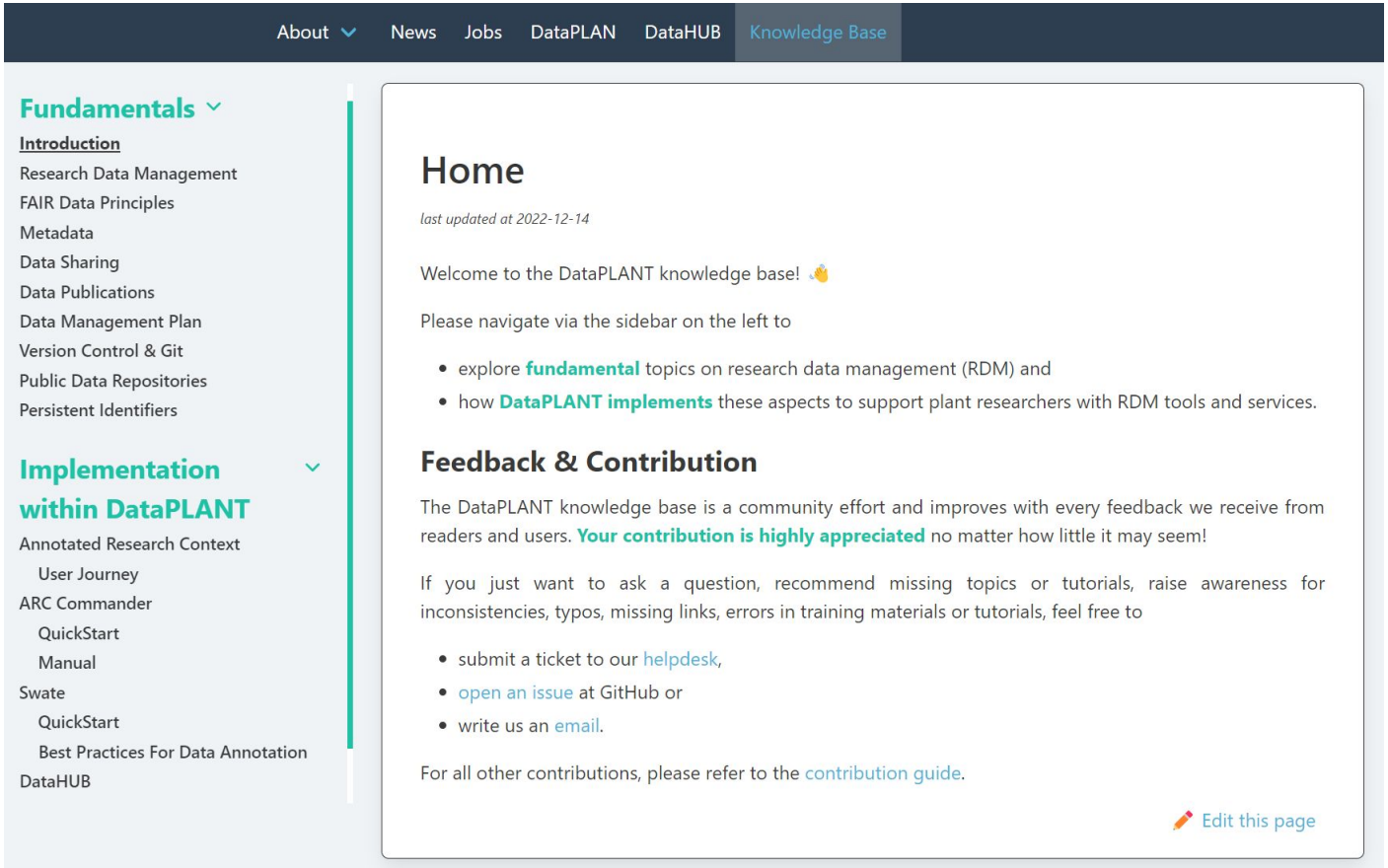
Centralization of training materials

Centralized KnowledgeBase with tutorials for consumers

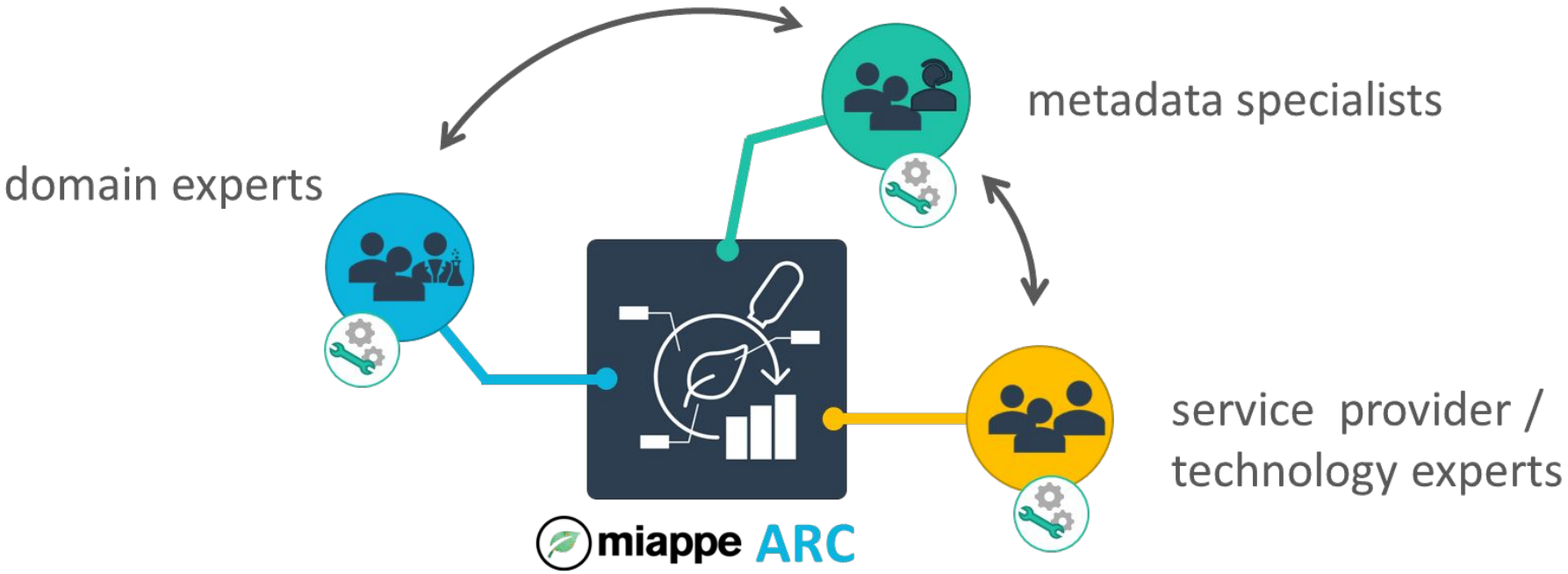


Tool specific Documentation for developers and data stewards

Centralization of training materials



Towards MIAPPE conforming ARCs



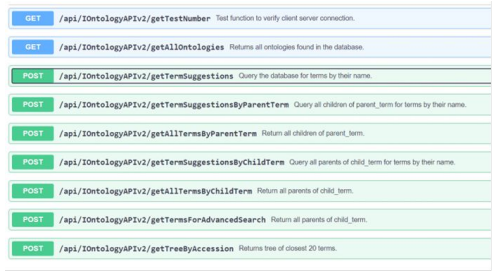
Towards MIAPPE conforming ARCs

metadata specialists

domain experts

service provider / technology experts

Towards MIAPPE conforming ARCs

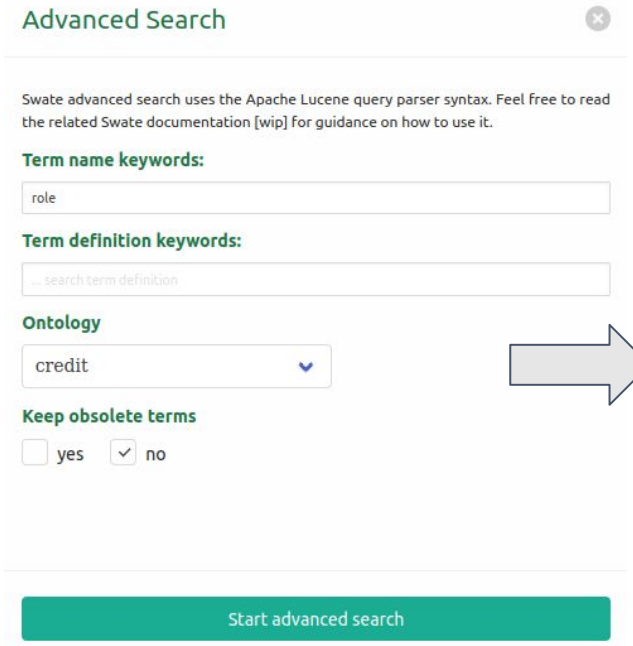


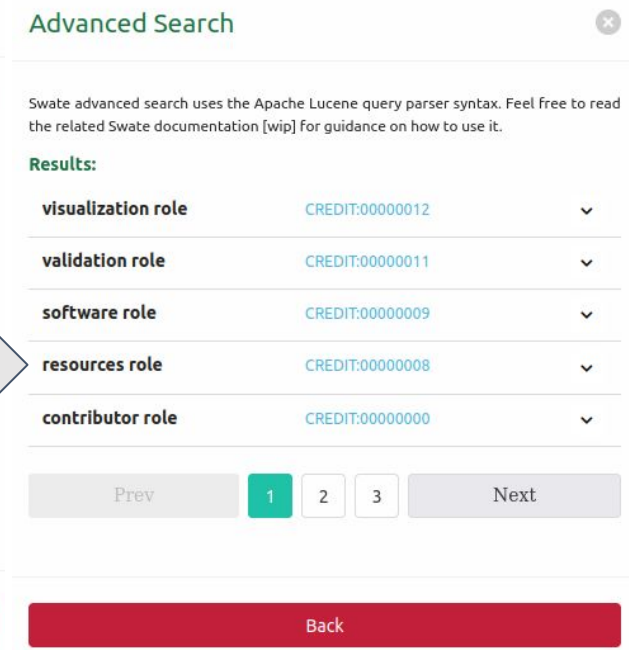
Current state:

- Ontology service consumption does work
- Addition of Ontology Terms required by MIAPPE finished

Expansion of the Ontology Service

- MIAPPE ontology was only provided in OWL format → creation of an OBO-formatted version based on the MIAPPE checklist
- OBO file of the CRediT ontology was broken → fixed the file and added it to the service





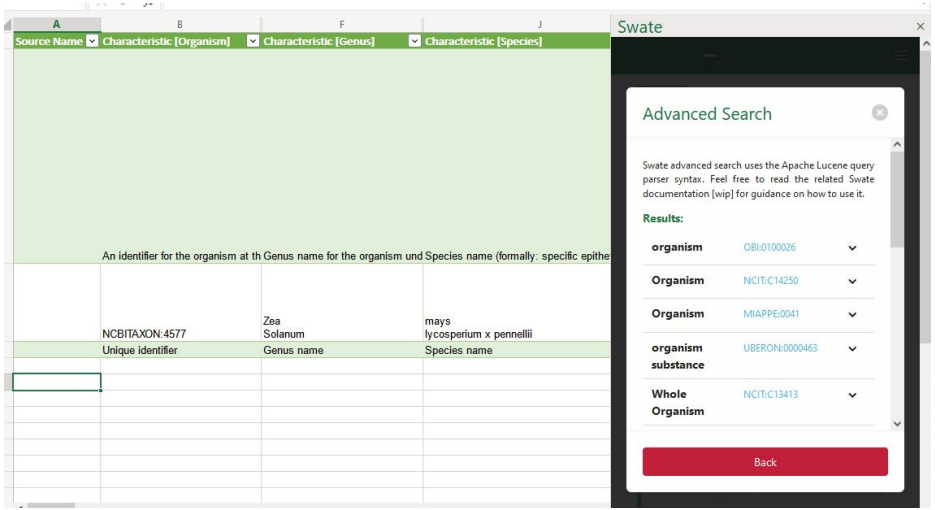
Template files generated in SWATE

Templates done - MIAPPE conform:

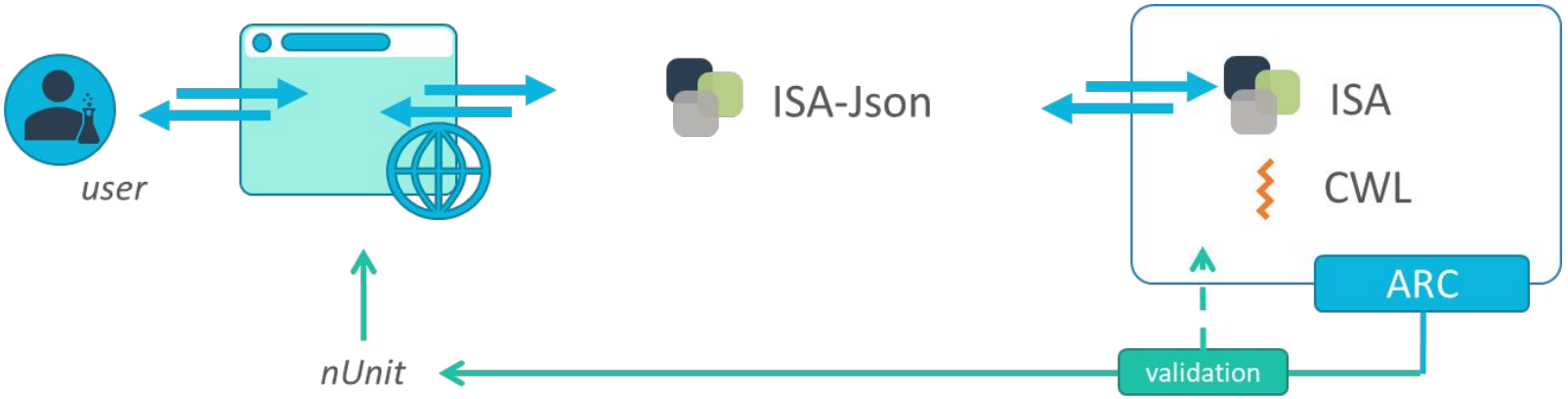
- Study metadata (study)
- Biological material + observation unit characteristics (study)
- Observation unit and sample characteristics (assay)
- Environmental and experimental factors (study factors, assay)

Templates to come - MIAPPE 2.0 ?

- Renaming of parameters and characteristics
  → more user-friendly
- Adding more parameters, characteristics and factors
  → applicable for more use-cases



Towards MIAPPE conforming ARCs



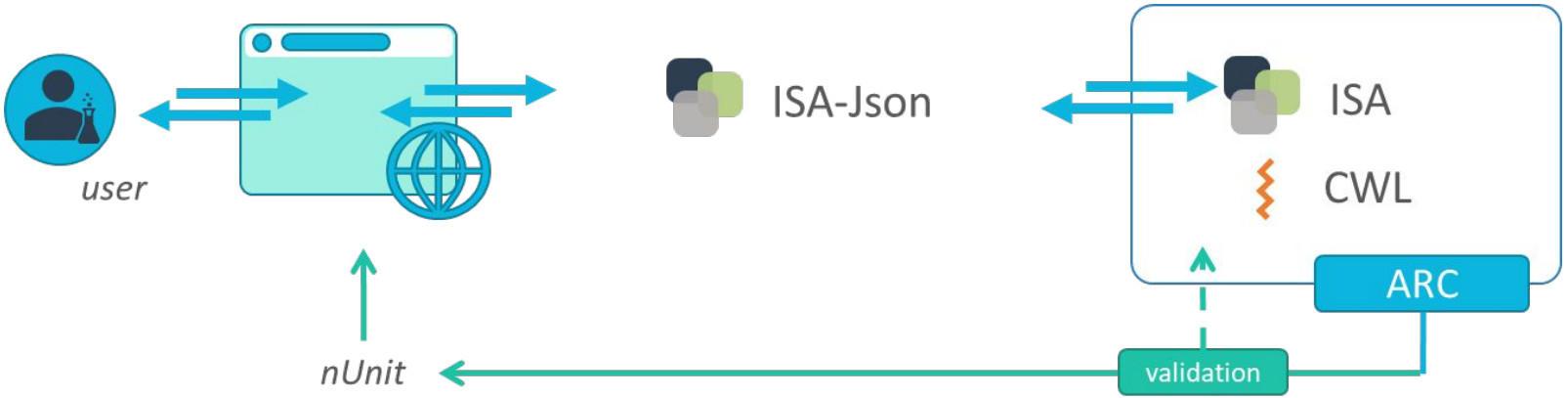
Current state:

- MIAPPE wizard does support saving state as ISA-Json
- ArcCommander able to consume ISA-Json and create an ARC from it
- Local REST connection could not be set up yet

Next steps:

- Set up REST Web-API

Towards MIAPPE conforming ARCs



Current state:

- MIAPPE wizard does support saving state as ISA-Json
- ArcCommander able to consume ISA-Json and create an ARC from it
- Local REST connection could not be set up yet

Next steps:

- Set up REST Web-API
- Profit

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

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

MIAPPE - Wizard

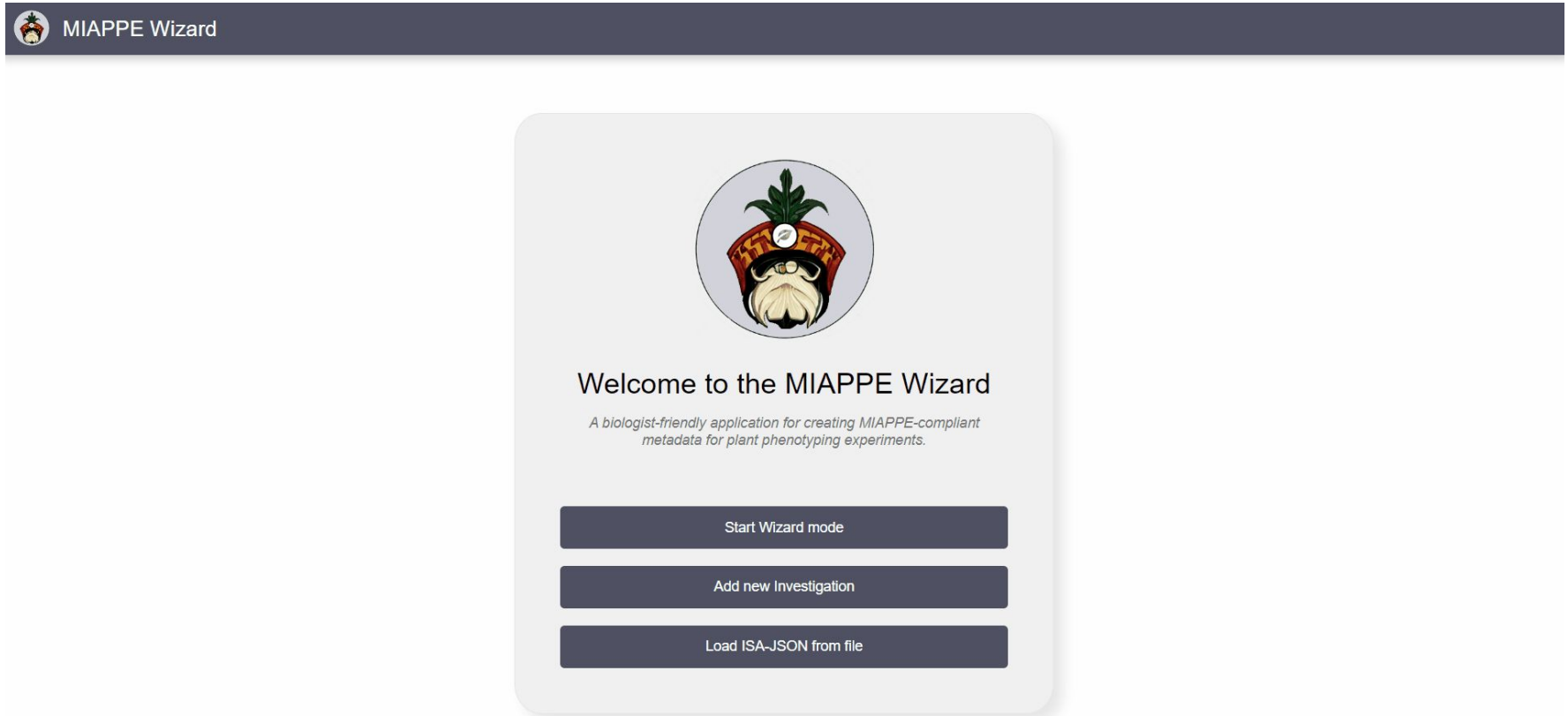
Main Goal: design & implement intuitive web app for describing phenotyping datasets

- successful collaboration with DataPLANT project
- monday:
  - socialising & discussing general concept
  - define concrete tasks & build sub groups
- tuesday:
  - implement first draft of wizard
  - set-up REST linkage between DataPLANT OLS & MIAPPE wizard
- wednesday:
  - hacking session together with virtual participants
  - discuss raised questions & define next steps
  - set-up ISA-Tab / ISA JSON conversion service
- thursday:
  - improve wizard GUI + creating a logo
  - integrate further ontologies into DataPLANT OLS

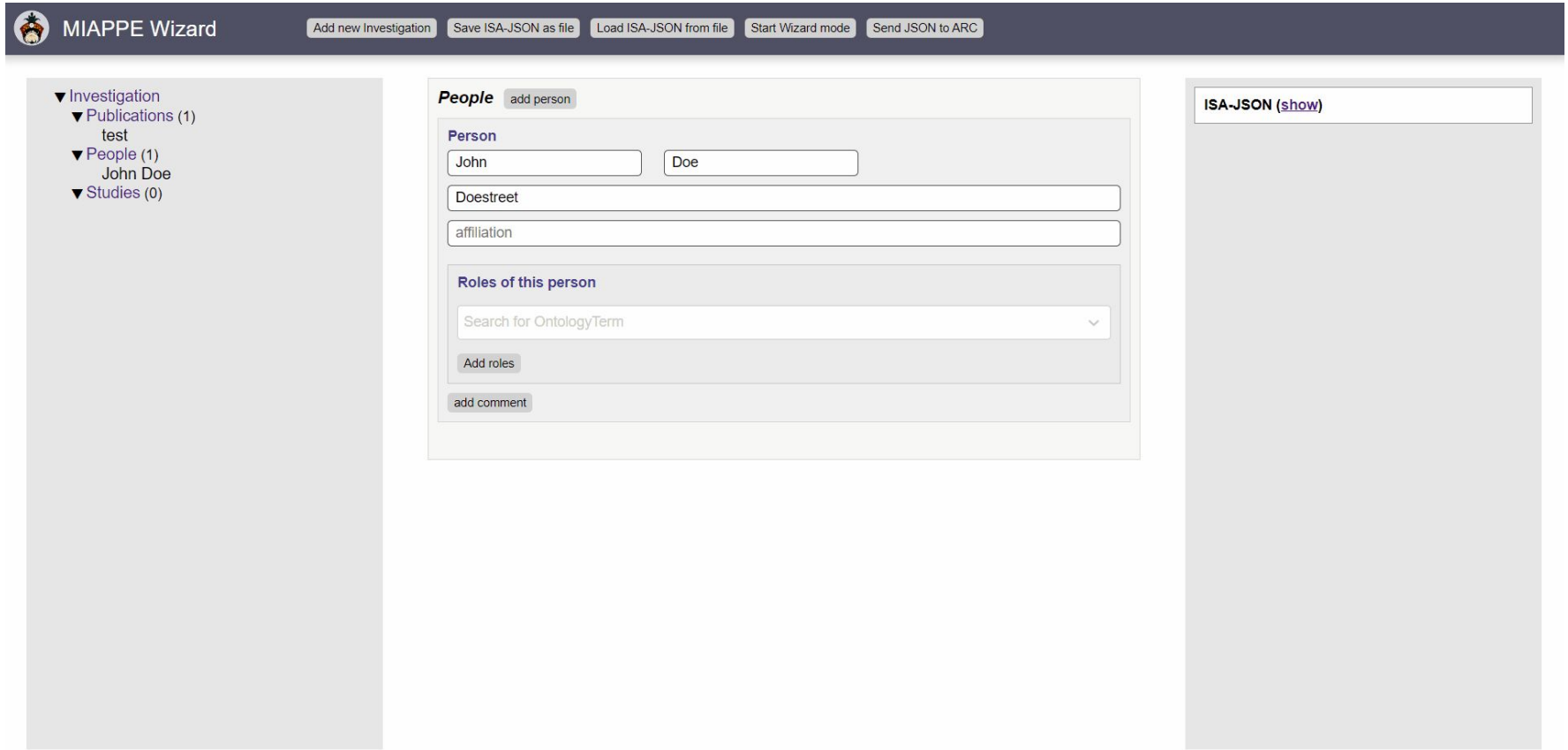
Summary & Results

- first prototype of MIAPPE Wizard deployed
  - basic form & questionnaire components implemented
  - reduced ISA/MIAPPE data model integrated
  - tree-navigation implemented
  - material upload component implemented
- REST connection to DataPLANT OLS established
  - integrated into different GUI components
- ISA JSON to ISA-Tab conversion service set-up & integrated
  - based on existing ISA framework
  - additional export function
- created project logo
- converted MIAPPE ontology (PPEO) from *owl to *obo format
- connection to DataPLANT ARC Commander prepared

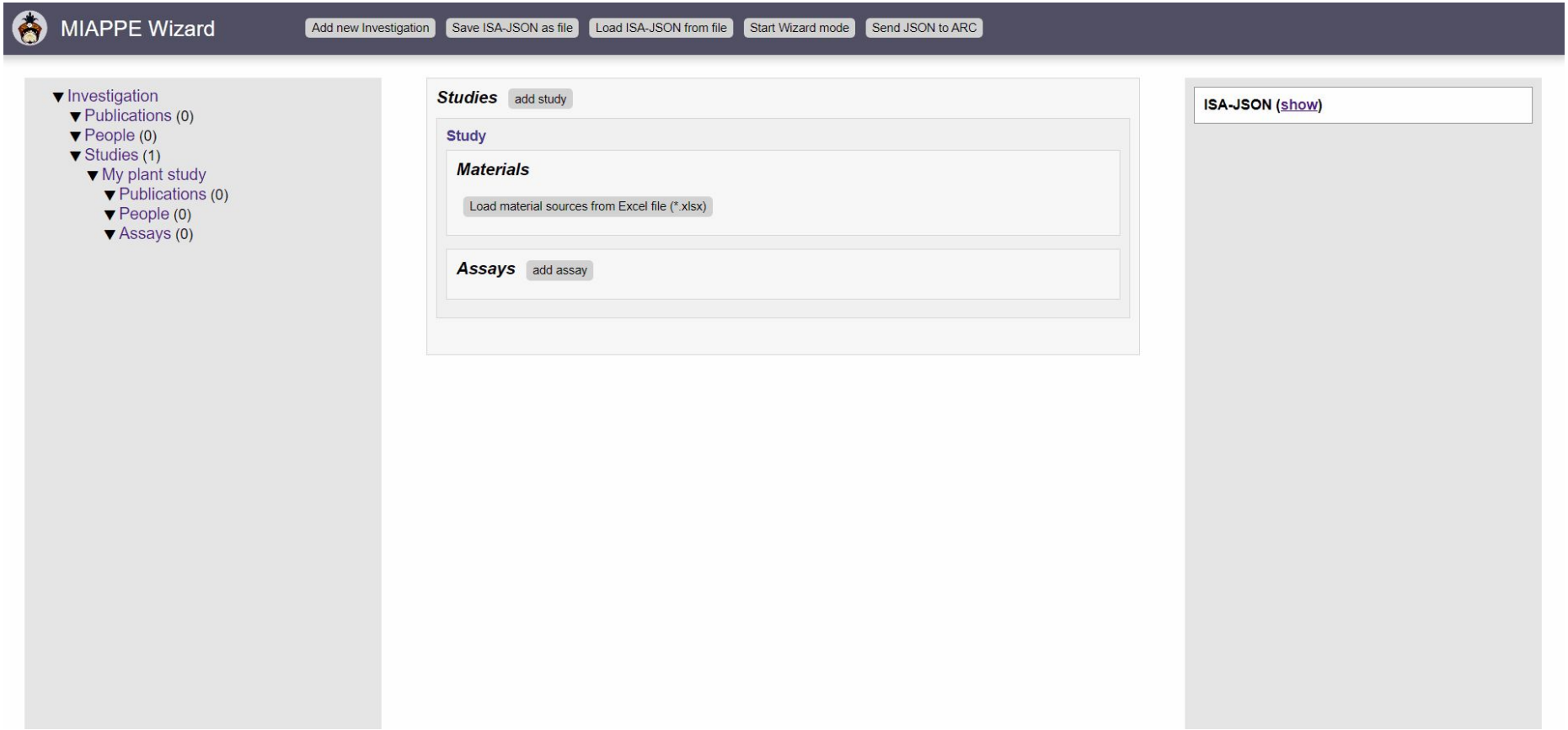
Preview - Graphical UI



Preview - Ontology LookUp



Preview - Material Loading



Future Tasks

- ISA representations (TAB <-> JSON) not fully consistent
- MIAPPE mapping to ISA-Tab does not fit 100%
  - suggest already certain improvements & updates
  - continue discussions in the future

- MIAPPE to ISA JSON mapping not available
  - first draft created & shared with MIAPPE community for discussion

- extend Wizard implementation
  - integrate full MIAPPE data model
  - add components to load result files

- test ARC generation within MIAPPE Wizard

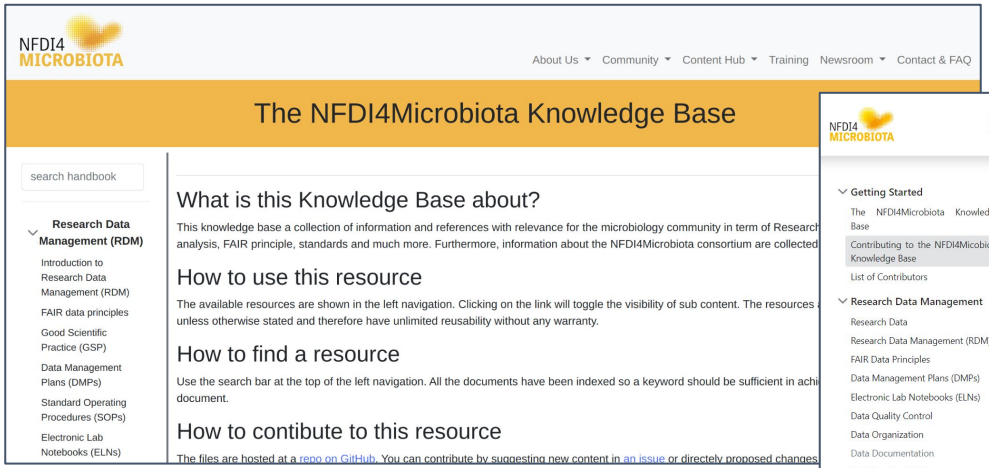
Extending the NFDI4Microbiota Knowledge Base

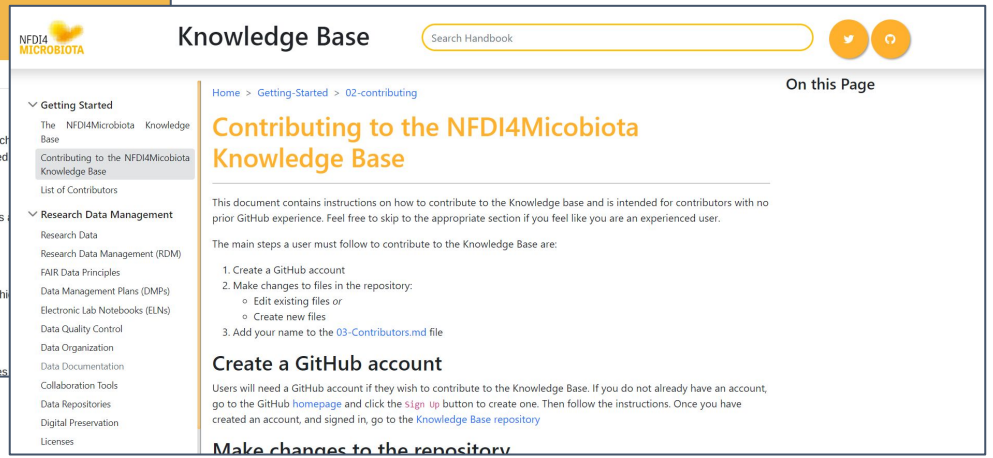
Justine Vandendorpe (ZB MED) and Kassian Kobert (Bielefeld University)

UI Overhaul

- Bugfixes
- UI/UX improvements
- UI Overhaul

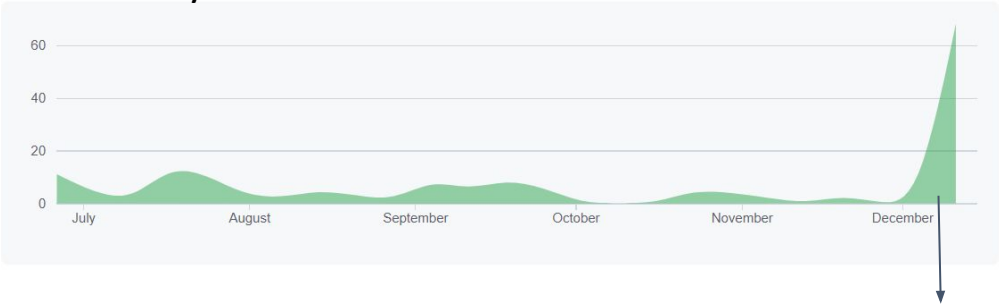
https://nfdi4microbiota.github.io/nfdi4microbiota-knowledge-base/





Activity This Week

Git Activity:



⅓ of all PRs

Content Creation:

4843 Words -> 11845 Words Content (+144%)

Activity This Week

Next Steps:

- Content Enrichment
- Advertisement

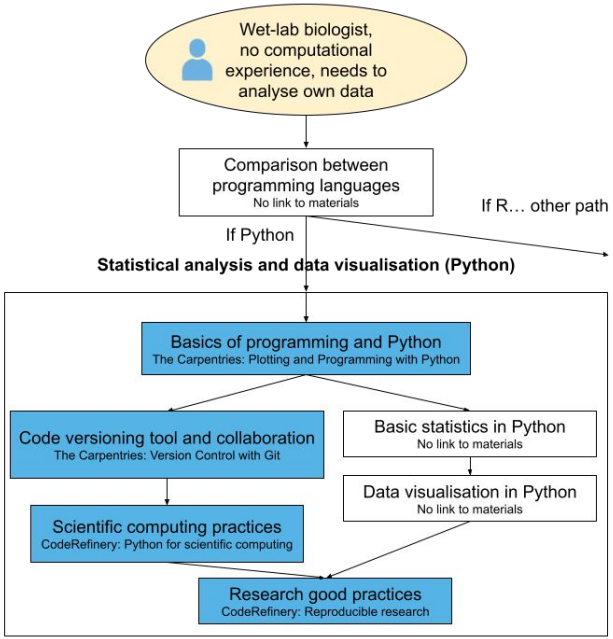
Mapping the training journey in Bioinformatics and beyond

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

The aim

In this project, we aim at mapping the **training journey** of different personas in Bioinformatics, visualising the steps needed to acquire expert knowledge in each topic, or skill. The focus of the project is to **link the steps with available training materials** of **short training courses** (compatible with post-graduation learning), with comments on their quality, accessibility and potential integration. This allows (i) **map** the relationship/**dependencies** between open training content and (ii) **identify areas** where training materials **should be developed**/enhanced by the community.

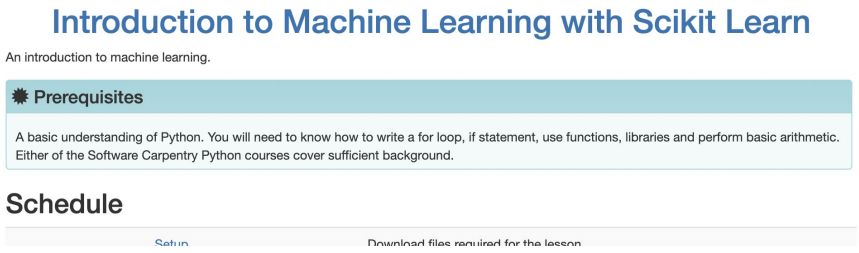
Persona	Skill	Training course
Use-case/example, defined by their **background** and **need**	**Technical ability or topical knowledge** that can be acquired through one or more courses	Materials **openly available** and easy to adopt **for learners** (+ info about how to adopt from trainers)



Data collection

Persona	Skill	Training course

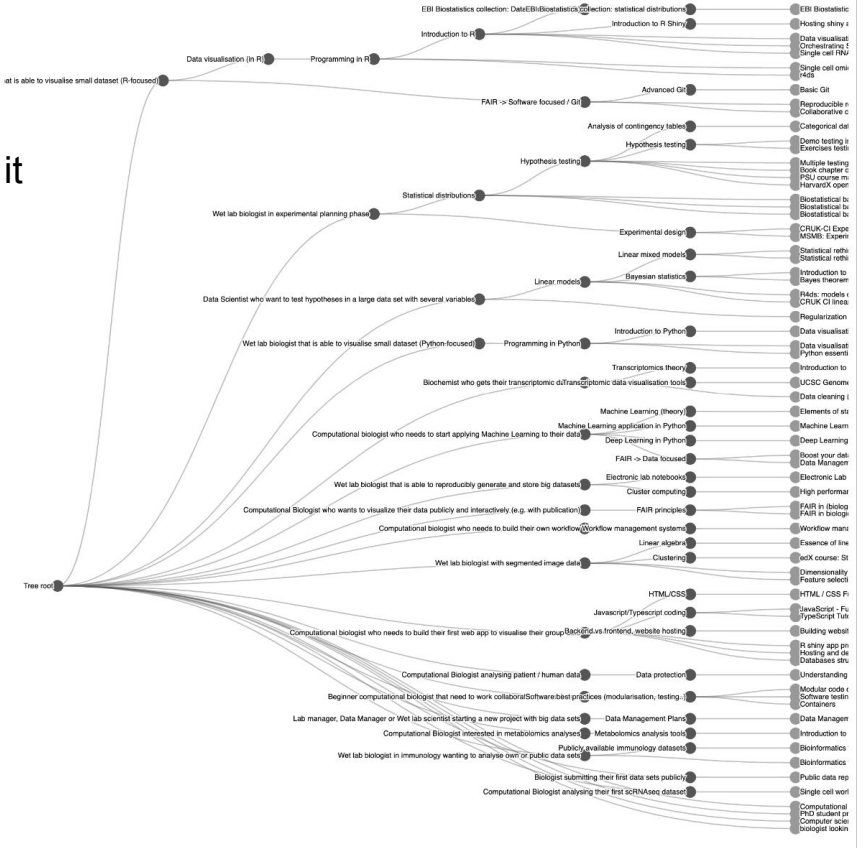
• Career stage
• Skills they have already / background
  ○ Any programming skills?
  ○ Any background in life sciences?
• Experience they already have
• Need / motivation (this is central, as the prerequisite skills could be captured by another path)
	• Meaningful name of the skill	• Is the content right for the step? Too much / not enough? Too theoretical / too much hands-on?
		• Easy to adopt for learners?
		  ○ Is the format compatible with self-learning?
		  ○ Includes tests / ways to assess understanding?
		• Easy to adopt from trainers?
		  ○ Copyright issues? Does the course include all the raw materials to be used in the course?
		• Are the materials actively maintained by a community? Is there a structured way to provide feedback about materials? Is it possible to customise them?

Computational biologist (background) → start applying Machine Learning to their data (need)	Machine Learning application in Python	

Challenges

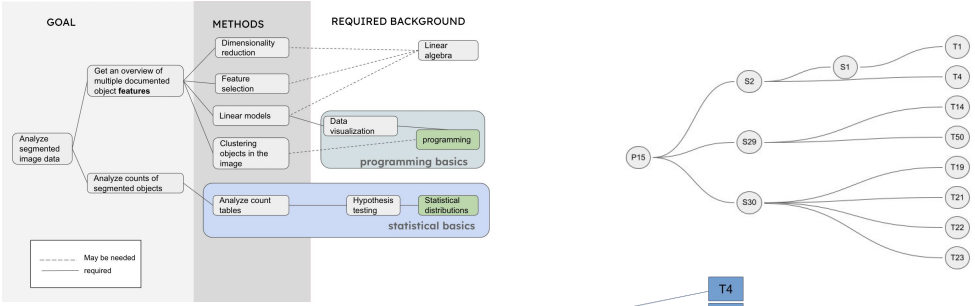
-
- Complexity of the data
- Subjectivity: of the annotations, of the way to visualise it

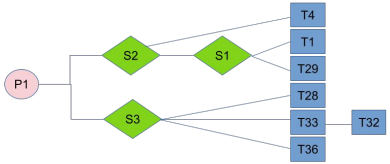
We concluded that (1) we shouldn't "force the data to the tool" and (2) the added value is really in each expert's freedom to annotate the information the way they prefer.



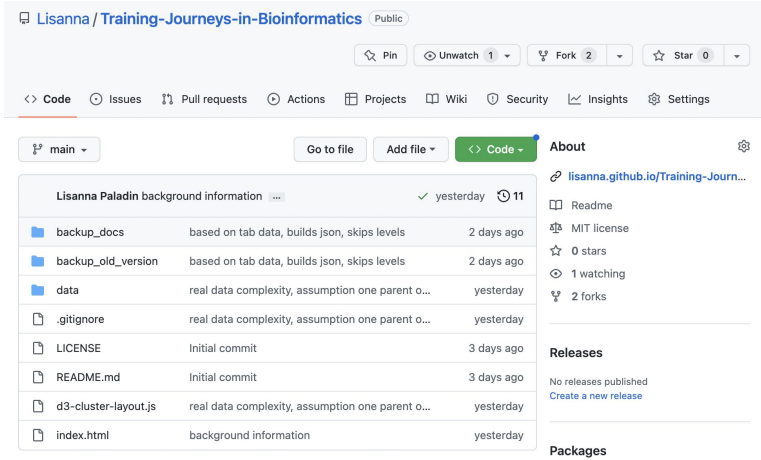
Results

- A repository to store the data with preliminary visualisation
- **22 personas** (examples) annotated by experts with the skills they need to meet their need, and the paths through the openly available training materials to do so.

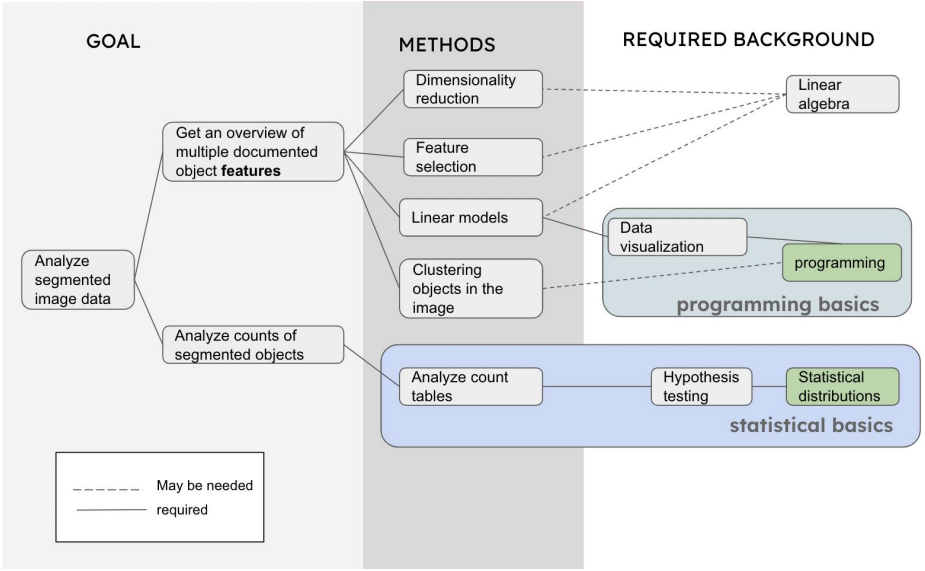




P1	Wet lab, visualize small dataset (R)	T28	Reproducible research software
S1	Programming (R)	T29	R4ds: R for Data Science
S2	Data visualization (R)	T32	Basic Git
S3	FAIR → Software focused / Git	T33	Advanced Git
T1	Introduction to R	T36	Collaborative coding
T4	Data visualization (R)		



Example



Linear models

Back to overview

Title	format	Link learner	Link trainer	comments
Models chapter in R4ds	Book chapter	https://r4ds.had.co.nz/index.html		Book chapter with exercises, using R/tidyverse
CRUK CI linear models	slides	https://bioinformatics-core-shared-training.github.io/linear-models-r/	https://github.com/bioinformatics-core-shared-training/linear-models-r	
HarvardX open online training	Mixed (text, videos)	http://rafalab.dfci.harvard.edu/pages/harvardx.html		comprehensive online course, covers also more basic and advanced topics

Hypothesis testing

Back to overview

Title	format	Link learner	Link trainer	comments
MSMB book chapter on hypothesis testing	Book chapter	https://www.huber.embl.de/msmb-quarto/06-chap.html		
Lecture hypothesis testing	Slides, demo in R, exercises	https://www.huber.embl.de/users/kaspar/biostat_2021/		
PSU course materials on hypothesis testing	text	https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/hypothesis-testing		
EBI Biostatistics collection: Hypothesis testing	Interactive tutorial	https://www.ebi.ac.uk/training/online/courses/biostatistics-introduction/getting-data-from-resource-name/		Recommended starting point
HarvardX open online training	Mixed (text, videos)	http://rafalab.dfci.harvard.edu/pages/harvardx.html		comprehensive online course, covers also more basic and advanced topics

Acknowledgments and next steps

Sarah Kaspar

Pablo Mier Muñoz

Lisanna Paladin

Karega Pauline

Julia Philipp

Daniel Wibberg

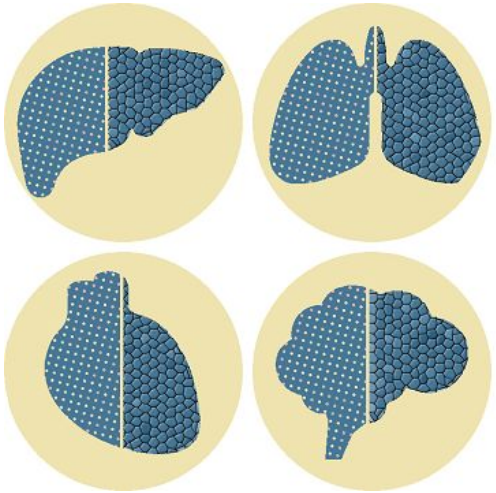
and more!

- Update GitHub repo with a report on the **challenges encountered and solution adopted**
- Add the information about the **training journeys** as annotated by the experts
- Establish **contribution guidelines**
- Establish common **format** for the **annotation** and visualisation (another hackathon? is it possible at all?)

Thank You

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)

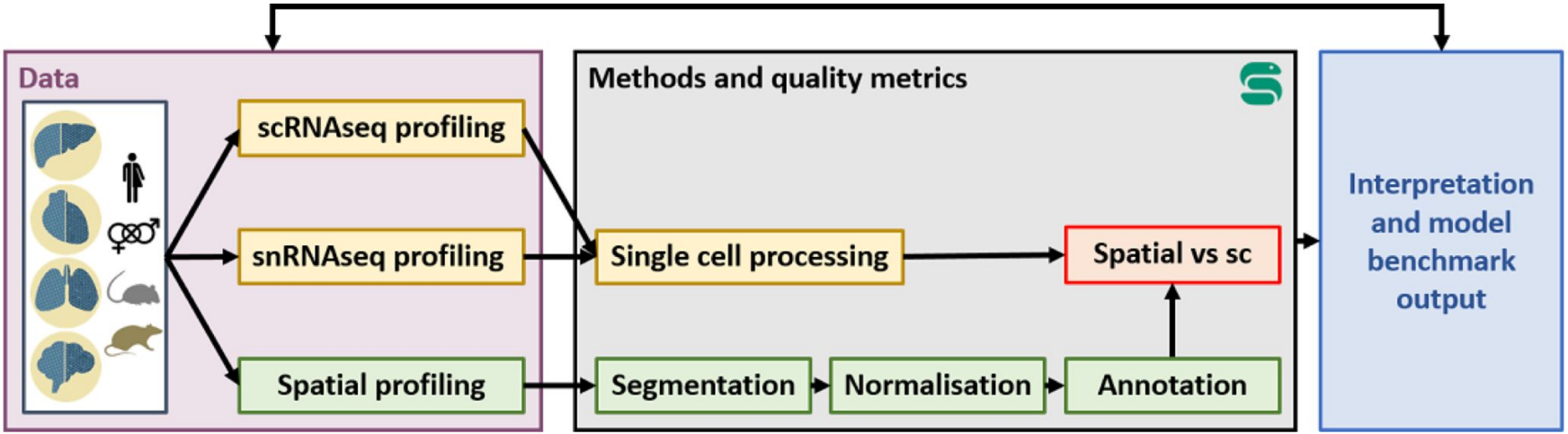


SpaceHack

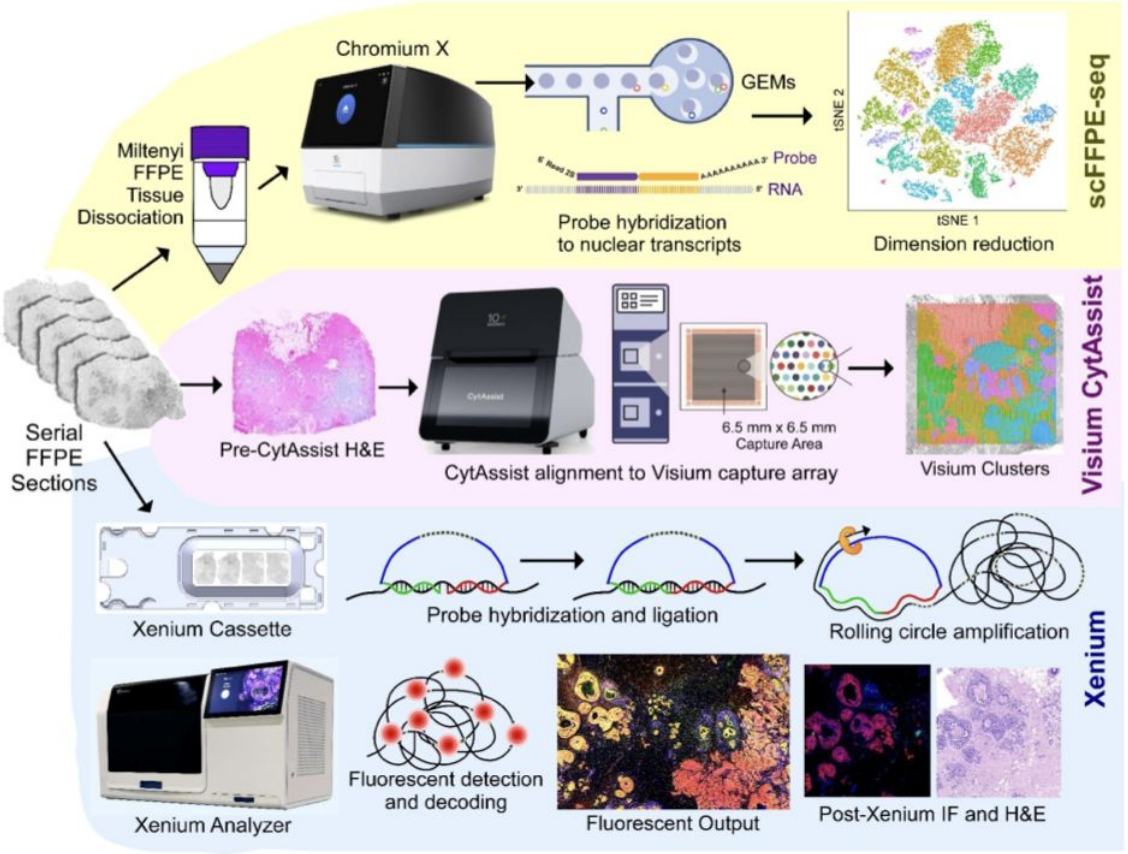
Establishing guidelines for cell segmentation and annotation in spatial transcriptomics data

Ambitious goal… how far did we get?

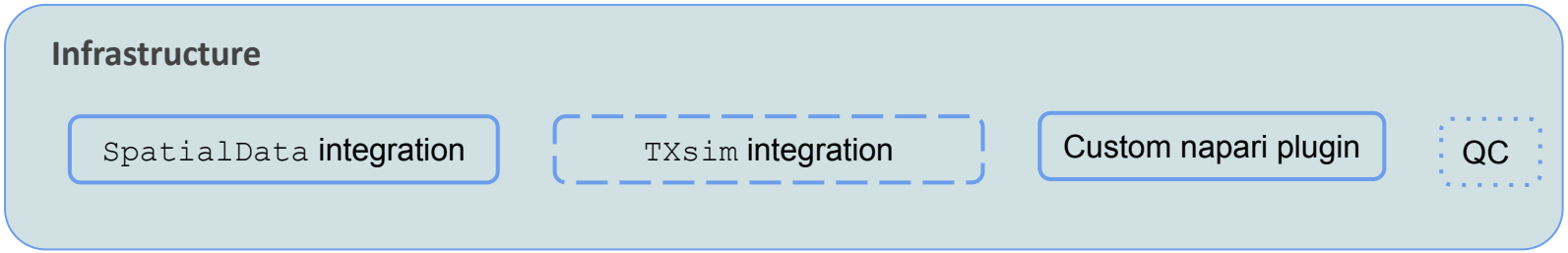
TXsim - a workflow for benchmarking imaging based spatial transcriptomics



Project 1: #xenium_omni_pipeline



Project 1: #xenium_omni_pipeline

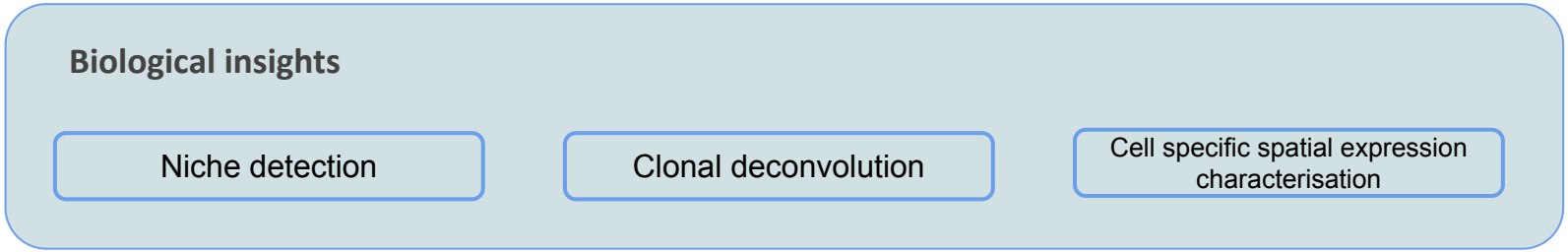


Xenium 10x

Visium 10x

Chromium 10x

Cell atlas



legend:

success

partial success

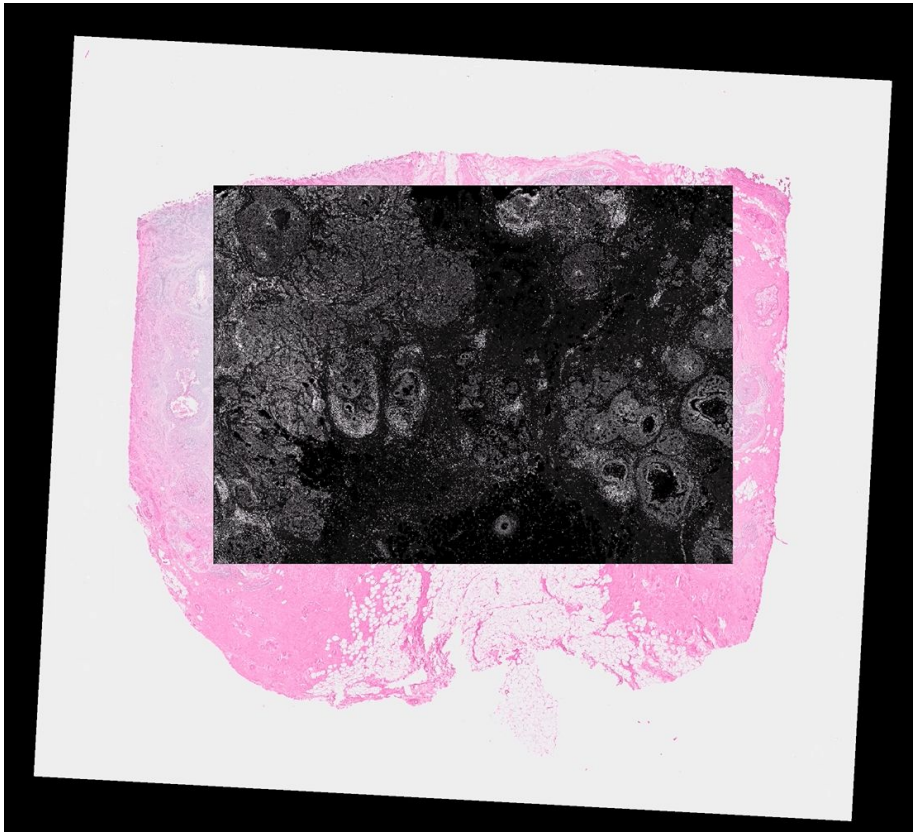
needs more time

Project 1: #xenium_omni_pipeline

- Integrated Xenium-visium data with napari Interactive layers alignment)
- Implemented custom Napari smooth polygon drawing
- Configured txsim pipeline (but didn't get results for our data)

Code for alignment

```
import spatialdata as sd

xenium_sdata = sd.SpatialData.read(xenium_path)
visium_sdata = sd.SpatialData.read(visium_path)

merged = sd.SpatialData(
    images={
        "xenium": xenium_sdata.images["morphology_mip"],
        "visium": visium_sdata.images["CytAssist_FFPE_Human_Breast_Cancer"],
    }
)

Interactive = Interactive(sdata=merged)
merged.write("merged.zarr")

affine = sd.compute_transformations(reference='xenium_landmarks', moving='visium_landmarks')
sd.set_transform(sdata['visium'], affine)
```



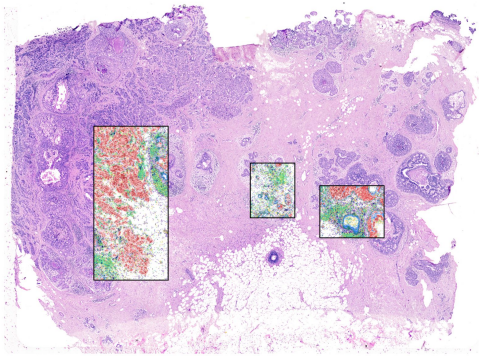
Project 1: #xenium_omni_pipeline

- Integrated Xenium-visium data with napari Interactive layers alignment)
- Implemented custom Napari smooth polygon drawing
- Configured txsim pipeline (but didn't get results for our data)

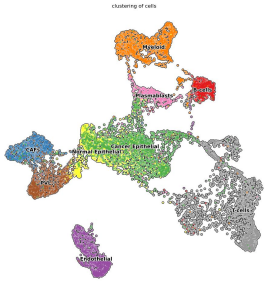

```python
import spatialdata as sd

xenium_sdata = sd.SpatialData.read(xenium_path)
visium_sdata = sd.SpatialData.read(visium_path)

merged = sd.SpatialData(
    images={
        "xenium": xenium_sdata.images["morphology_mip"],
        "visium": visium_sdata.images["CytAssist_FFPE_Human_Breast_Cancer"],
    }
)

Interactive = Interactive(sdata=merged)
merged.write("merged.zarr")

affine = sd.compute_transformations(reference='xenium_landmarks', moving='visium_landmarks')
sd.set_transform(sdata['visium'], affine)
```

Project 1: #xenium_omni_pipeline

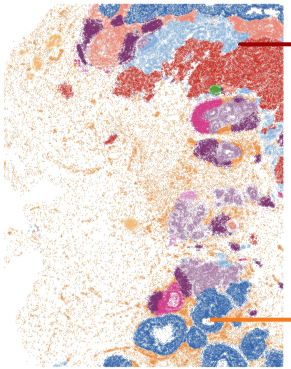
Biological insights generation



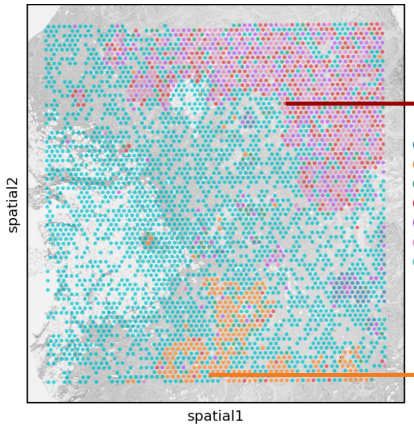
single cell atlas mapping

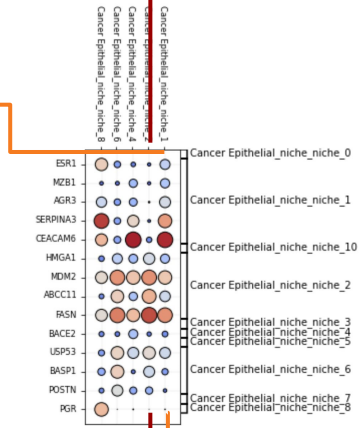


Automatic niche detection



Cancer clone mapping





Spatial hierarchical differential expression

Project 1: #xenium_omni_pipeline

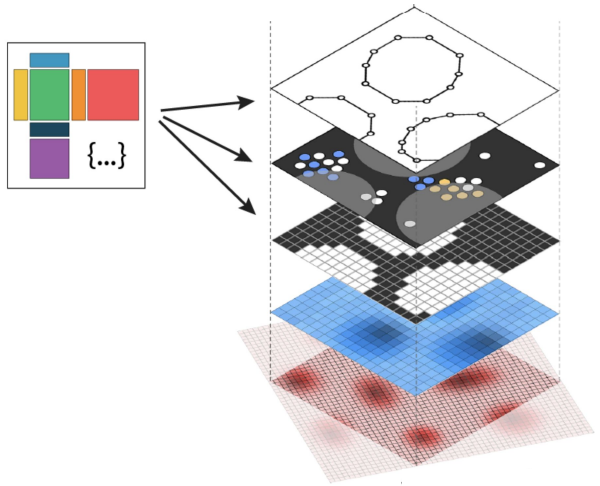
Follow up and lessons:

1) Reimplement the steps in a more reproducible environment
2) Extend clonal annotation to Xenium
3) Think more about QC metrics (hopefully run the txsim)
4) Wait for the new Xenium cancer data
5) **Get slides from people before they enter the party mode (although it still worked out)**

Project 2: #zarr_for_data_and_metadata

- **Project motivation**
  - Analysis of multiple spatial datasets with all their metadata is

    laborious, varies in quality and lacks biological meaning
- **Our solution**
  - Increase **interoperability and performance** using a new standard: `SpatialData`
  - Add metadata for **dimensionality, physical distances and MITI**
    - useful for upcoming methods e.g. *Project 4: finding overlapping cells in z*
  - Document and make **reproducible** using `spatial-sandbox`



Project 2: #zarr_for_data_and_metadata

Heterogeneous, technology-dependent data storage (both raw and processed)

```
heart
├── in_situ_sequencing
│   ├── counts
│   │   └── Heart_gene_spot_coord.csv
│   └── DAPI
│       └── nuclei_heart.tif
│   └── README.md
└── scRNAseq
    ├── README.md
    └── TabulaSapiens_adult_heart.h5ad
```

```
ResolveMouse
├── 20272_slide1_A1-1_DAPI.tiff
├── 20272_slide1_A1-1_raw.tiff
├── 20272_slide1_A1-1_results.txt
├── 20272_slide1_A1-2_DAPI.tiff
├── 20272_slide1_A1-2_raw.tiff
├── 20272_slide1_A1-2_results.txt
├── 20272_slide1_C2_DAPI.tiff
├── 20272_slide1_C2_raw.tiff
├── 20272_slide1_C2_results.txt
├── 20272_slide1_D2-1_DAPI.tiff
├── 20272_slide1_D2-1_raw.tiff
├── 20272_slide1_D2-1_results.txt
├── 20272_slide1_D2-2_DAPI.tiff
├── 20272_slide1_D2-2_raw.tiff
└── 20272_slide1_D2-2_results.txt
```

```
human_brain
├── 202201271624_H1930002Cx46MTG202007105_vmsc00401
│   ├── images
│   │   ├── micron_to_mosaic_pixel_transform.csv
│   │   ├── mosaic_DAPI_z3_ds8.tif
│   │   └── mosaic_PolyT_z3_ds8.tif
│   └── TXsim_detected_transcript.csv
├── 202201281113_H1930002Cx46MTG202007104_vmsc00401
│   ├── images
│   │   ├── micron_to_mosaic_pixel_transform.csv
│   │   ├── mosaic_DAPI_z3_ds8.tif
│   │   └── mosaic_PolyT_z3_ds8.tif
│   └── TXsim_detected_transcript.csv
├── human_brain_README.txt
├── human_snRNAseq
│   ├── anno.feather
│   ├── data.feather
│   ├── desc.feather
│   ├── human_snRNAseq.h5ad
│   └── medians.feather
└── human_snRNAseq.zip
```

```
├── fixed_1001844875.csv
├── fixed_1001844875.1.csv
├── MERFISH_genes.csv
├── MERFISH_genes.1.csv
├── Allen_MERFISH_spots_with_anatomy.csv
├── Allen_MERFISH_spots_with_anatomy.1.csv
├── Allen_MERFISH_Layers.geojson
└── Allen_MERFISH_Layers.1.geojson
```

```
spleen
├── scRNAseq
│   ├── README.md
│   └── TabulaSapiens_adult_spleen.h5ad
├── seqFISH_spleen_central
│   ├── 5 positions each.pos
│   ├── counts
│   │   ├── high_threshhold
│   │   └── low_threshold
│   ├── DAPI
│   │   ├── dapi-hyb0-pos10.tif
│   │   ├── dapi-hyb0-pos11.tif
│   │   ├── dapi-hyb0-pos12.tif
│   │   ├── dapi-hyb0-pos13.tif
│   │   ├── dapi-hyb0-pos14.tif
│   │   ├── dapi-hyb0-pos15.tif
│   │   ├── dapi-hyb0-pos16.tif
│   │   ├── dapi-hyb0-pos17.tif
│   │   ├── dapi-hyb0-pos18.tif
│   │   └── dapi-hyb0-pos19.tif
│   ├── full_slide.tif
│   ├── gene-list-seqFISH-spleen.csv
│   ├── README.md
│   └── segmentation_mask
│       ├── segmentation_mask_fov10.tif
│       ├── segmentation_mask_fov11.tif
│       ├── segmentation_mask_fov12.tif
│       ├── segmentation_mask_fov13.tif
│       ├── segmentation_mask_fov14.tif
│       ├── segmentation_mask_fov15.tif
│       ├── segmentation_mask_fov16.tif
│       ├── segmentation_mask_fov17.tif
│       ├── segmentation_mask_fov18.tif
│       └── segmentation_mask_fov19.tif
```

```
data
├── visium
│   ├── tissue_image.tif
│   ├── spatial.tar.gz
│   ├── spatial
│   │   ├── tissue_positions_list.csv
│   │   ├── tissue_lowres_image.png
│   │   ├── tissue_hires_image.png
│   │   ├── spatial_enrichment.csv
│   │   ├── scalefactors_json.json
│   │   ├── detected_tissue_image.jpg
│   │   ├── cytassist_image.tiff
│   │   ├── aligned_tissue_image.jpg
│   │   └── aligned_fiducials.jpg
│   ├── probe_set.csv
│   ├── molecule_info.h5
│   ├── image.tif
│   ├── filtered_feature_bc_matrix.h5
│   ├── analysis.tar.gz
│   └── analysis
│       ├── umap
│       ├── tsne
│       ├── pca
│       ├── diffexp
│       └── clustering
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_transcripts.zarr.zip
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_transcripts.parquet
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_transcripts.csv.gz
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_nucleus_boundaries.parquet
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_nucleus_boundaries.csv.gz
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_morphology_mip.ome.tif
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_morphology_focus.ome.tif
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_morphology.ome.tif
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_experiment.xenium
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cells.zarr.zip
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cells.parquet
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cells.csv.gz
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cell_feature_matrix.zarr.zip
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cell_feature_matrix.tar.gz
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cell_feature_matrix.h5
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cell_boundaries.parquet
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_cell_boundaries.csv.gz
├── Xenium_FFPE_Human_Breast_Cancer_Rep1_analysis.zarr.zip
└── Xenium_FFPE_Human_Breast_Cancer_Rep1_analysis.tar.gz
```

Project 2: #zarr_for_data_and_metadata

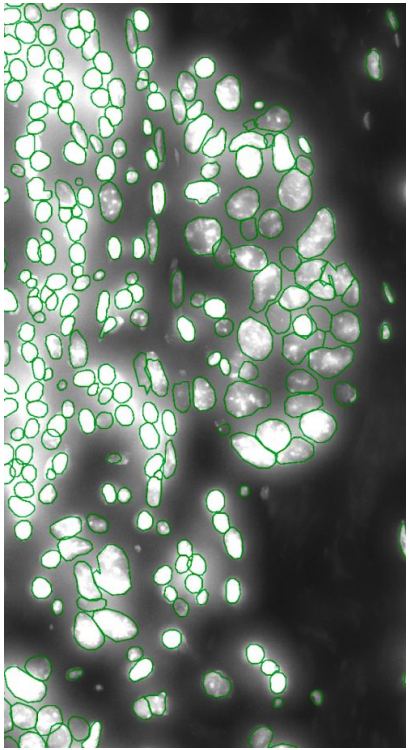
SpatialData allows for **efficient, modular** storage in .Zarr

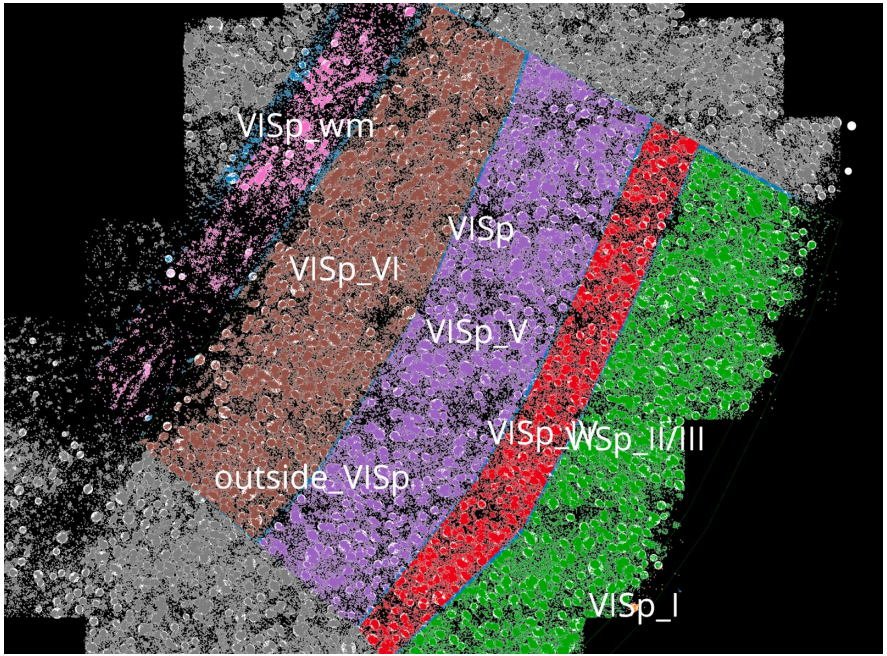
```
SpatialData object with:
├── Images
│     ├── 'morphology_focus': MultiscaleSpatialImage[cyx] (1, 25779, 35416), (1, 12889, 17708), (1, 6444, 8854), (1, 3222, 4427), (1, 1611, 2213), (1, 805, 1106), (1, 402, 553), (1, 201, 276), (1, 100, 138)
│     └── 'morphology_mip': MultiscaleSpatialImage[cyx] (1, 25779, 35416), (1, 12889, 17708), (1, 6444, 8854), (1, 3222, 4427), (1, 1611, 2213), (1, 805, 1106), (1, 402, 553), (1, 201, 276), (1, 100, 138)
├── Points
│     └── 'transcripts': pyarrow.Table shape: (100000, 3) (3D points)
├── Polygons
│     ├── 'cell_boundaries': GeoDataFrame shape: (999, 1) (2D polygons)
│     └── 'nucleus_boundaries': GeoDataFrame shape: (999, 1) (2D polygons)
├── Shapes
│     ├── 'cells': AnnData with osbm.spatial (167782, 2)
│     └── 'nuclei': AnnData with osbm.spatial (167782, 2)
└── Table
      └── 'AnnData object with n_obs × n_vars = 167782 × 313
    obs: 'cell_id', 'transcript_counts', 'control_probe_counts', 'control_codeword_counts', 'total_counts', 'cell_area', 'nucleus_area'
    var: 'gene_ids', 'feature_types', 'genome'
    uns: 'spatialdata_attrs'': AnnData (167782, 313)
```

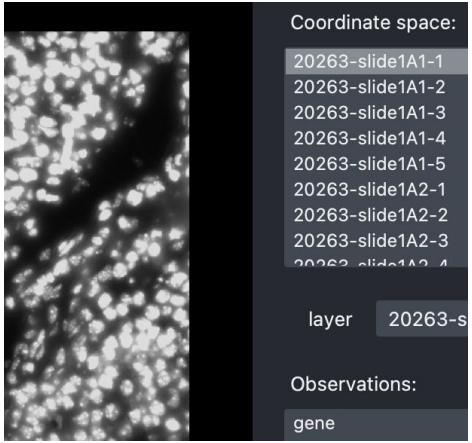
```
SpatialData object with:
├── Images
│     └── 'rasterized': SpatialImage[cyx] (1, 522, 575)
├── Points
│     ├── 'Points2_cyx': pyarrow.Table shape: (20, 2) (2D points)
│     ├── 'Points_cyx': pyarrow.Table shape: (3, 2) (2D points)
│     ├── 'cells_cyx': pyarrow.Table shape: (2399, 2) (2D points)
│     └── 'single_molecule': pyarrow.Table shape: (3714642, 3) (2D points)
├── Polygons
│     └── 'anatomical': GeoDataFrame shape: (6, 1) (2D polygons)
├── Shapes
│     └── 'cells': AnnData with osbm.spatial (2399, 2)
└── Table
      └── 'AnnData object with n_obs × n_vars = 2399 × 268
    obs: 'cell_id'
    uns: 'spatialdata_attrs'': AnnData (2399, 268)
```

```
SpatialData object with:
├── Images
│     ├── '20272_slide1_A1-1_DAPI': MultiscaleSpatialImage[cyx] (1, 12864, 10720), (1, 6432, 5360), (1, 1608, 1340), (1, 201, 167), (1, 12, 10)
│     ├── '20272_slide1_A1-1_raw': MultiscaleSpatialImage[cyx] (1, 12864, 10720), (1, 6432, 5360), (1, 1608, 1340), (1, 201, 167), (1, 12, 10)
│     ├── '20272_slide1_A1-2_DAPI': MultiscaleSpatialImage[cyx] (1, 10720, 8576), (1, 5360, 4288), (1, 1340, 1072), (1, 167, 134), (1, 10, 8)
│     ├── '20272_slide1_A1-2_raw': MultiscaleSpatialImage[cyx] (1, 10720, 8576), (1, 5360, 4288), (1, 1340, 1072), (1, 167, 134), (1, 10, 8)
│     ├── '20272_slide1_C2_DAPI': MultiscaleSpatialImage[cyx] (1, 21440, 12864), (1, 10720, 6432), (1, 2680, 1608), (1, 335, 201), (1, 20, 12)
│     ├── '20272_slide1_C2_raw': MultiscaleSpatialImage[cyx] (1, 21440, 12864), (1, 10720, 6432), (1, 2680, 1608), (1, 335, 201), (1, 20, 12)
│     ├── '20272_slide1_D2-1_DAPI': MultiscaleSpatialImage[cyx] (1, 19296, 8576), (1, 9648, 4288), (1, 2412, 1072), (1, 301, 134), (1, 18, 8)
│     ├── '20272_slide1_D2-1_raw': MultiscaleSpatialImage[cyx] (1, 19296, 8576), (1, 9648, 4288), (1, 2412, 1072), (1, 301, 134), (1, 18, 8)
│     ├── '20272_slide1_D2-2_DAPI': MultiscaleSpatialImage[cyx] (1, 19296, 2144), (1, 9648, 1072), (1, 2412, 268), (1, 301, 33), (1, 18, 2)
│     └── '20272_slide1_D2-2_raw': MultiscaleSpatialImage[cyx] (1, 19296, 2144), (1, 9648, 1072), (1, 2412, 268), (1, 301, 33), (1, 18, 2)
└── Points
      ├── '20272_slide1_A1-1_results': pyarrow.Table shape: (4754932, 4) (3D points)
      ├── '20272_slide1_A1-2_results': pyarrow.Table shape: (3481604, 4) (3D points)
      ├── '20272_slide1_C2_results': pyarrow.Table shape: (9358543, 4) (3D points)
      ├── '20272_slide1_D2-1_results': pyarrow.Table shape: (5287534, 4) (3D points)
      └── '20272_slide1_D2-2_results': pyarrow.Table shape: (1083300, 4) (3D points)
```

Interactive visualization with **napari**

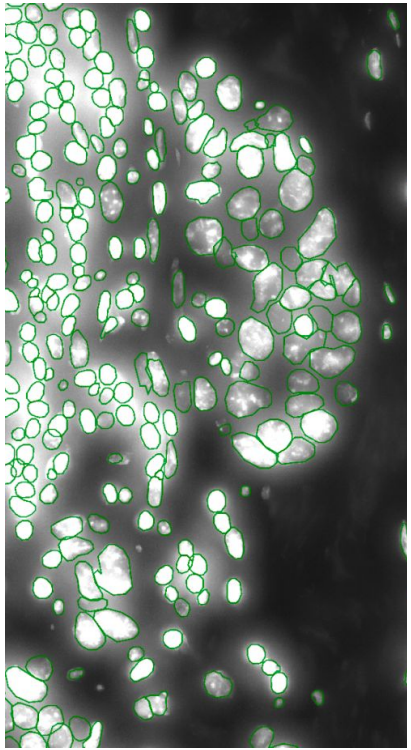


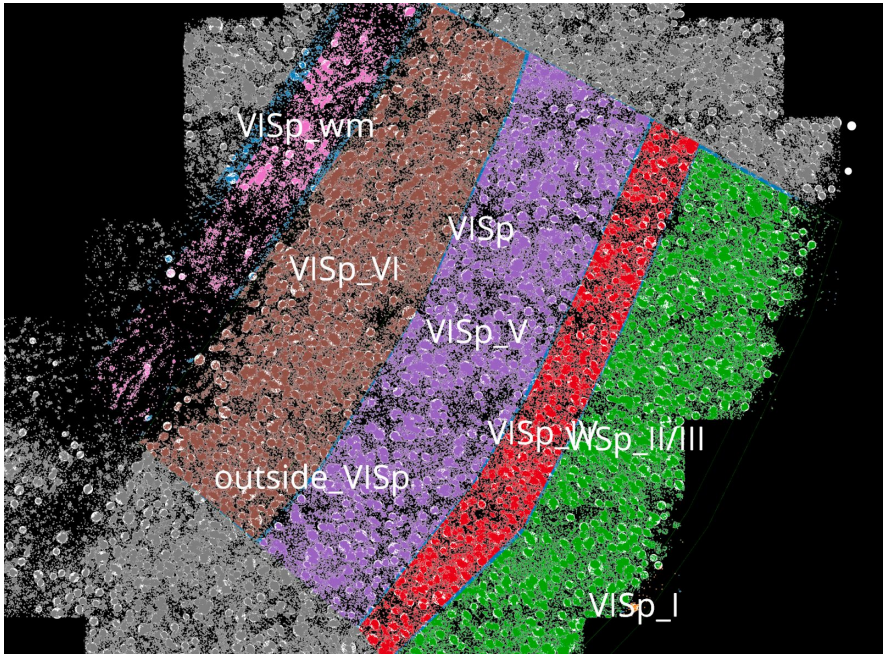


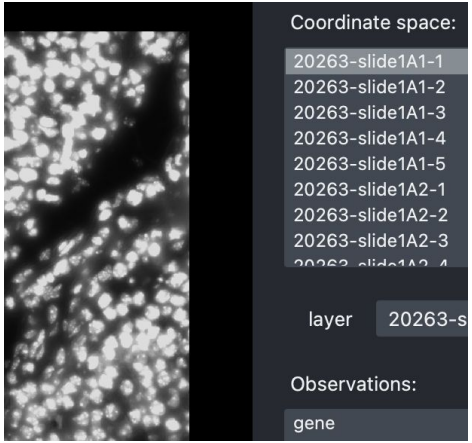


Project 2: #zarr_for_data_and_metadata

Interactive visualization with **napari**





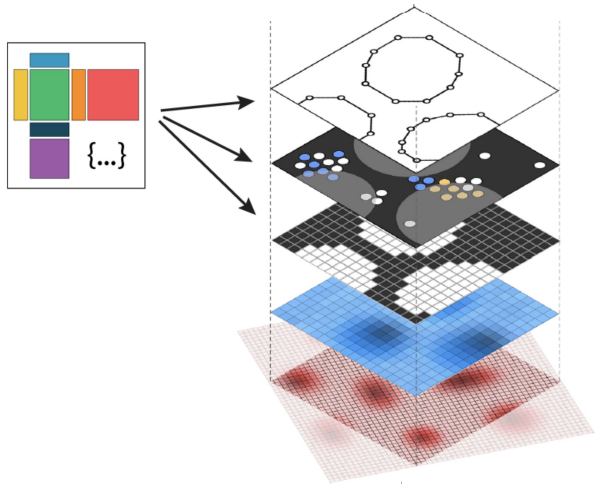


Project 2: #zarr_for_data_and_metadata

Key Takeaways

- Done for Xenium (breast) and Molecular Cartography (liver/melanoma)
- New datasets inform **downstream debugging** e.g. visualizations
- In cases when stitching is required, adding transforms is challenging
- WIP: lung (Wouter-Michiel), mouse brain (Luca), heart (Benjamin)
- WIP: Not all metadata is covered in the OME-NGFF specification
    - investigating **AnnData** for this purpose
- WIP: **TissUUmaps** for associating quality control data readily with imaging
    - meeting with developers in January

Follow-up in [scverse SpatialData meetings]



Project 3: #segmentation-x

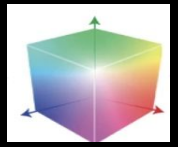
Mesmer

Greenwald et al.
2021 Nat. Biotech.

State of the art (SOTA) CNN based models

Cellpose 2

Stringer et al. 2021
Nature Methods



Baysor

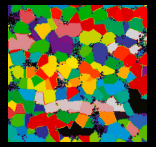
Petukhov et al.
2021 Nat. Biotech.

Probabilistic spot based segmentation

Ilastik

Berg et al. 2019
Nature Methods

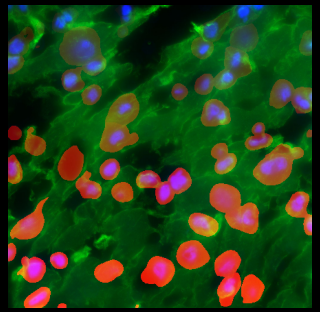
Random forest pixel classification method

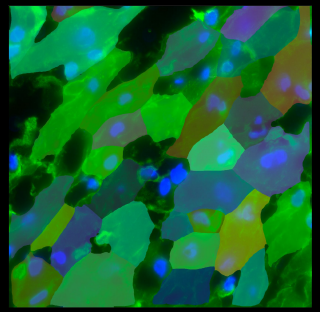


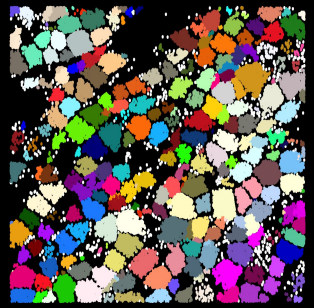
JSTA

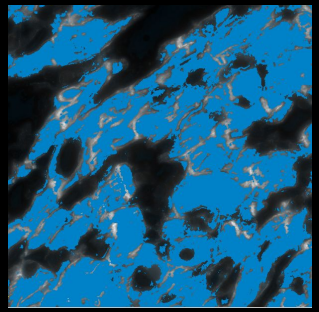
Littman et al. 2021 Molecular
Systems Biology

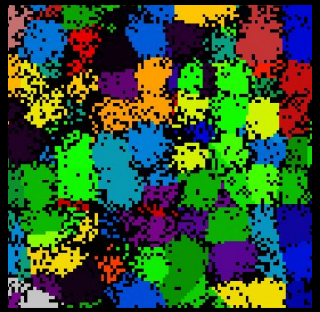
Joint cell segmentation and cell type annotation











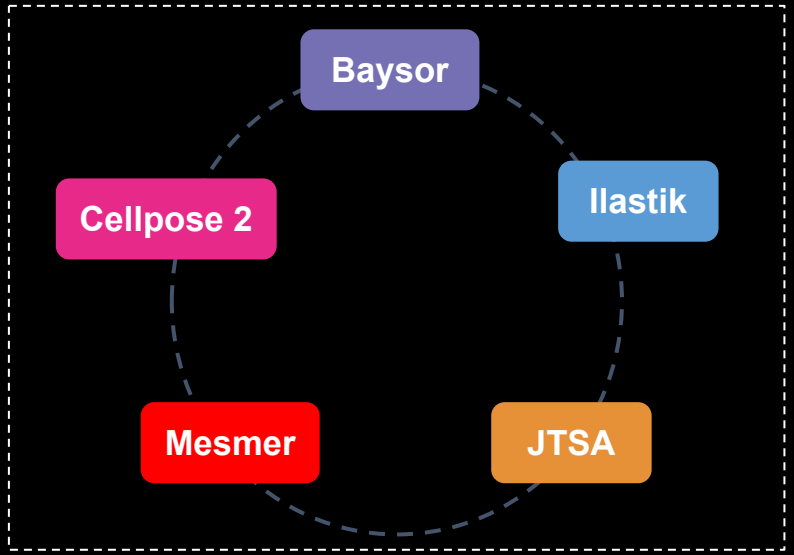
Takeaways

<u>There is no out-of-the-box solution</u>

Cellpose 2 worked best at recognizing cell shape in the heart

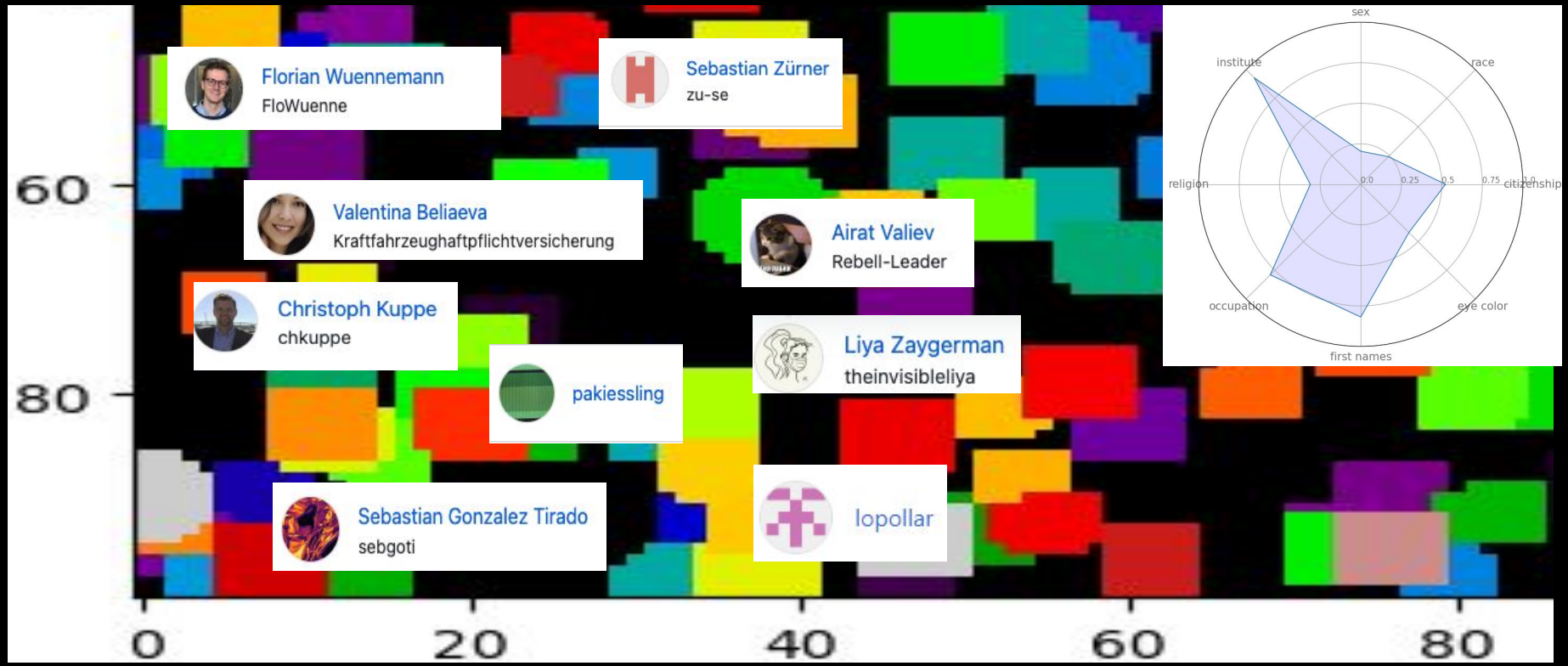
New approaches or a combination of methods is likely needed:

- Image pre-processing
- Additional stains for specific celltypes, subcelullar locations
- Training / rewriting tools to be more applicable to different cell shapes



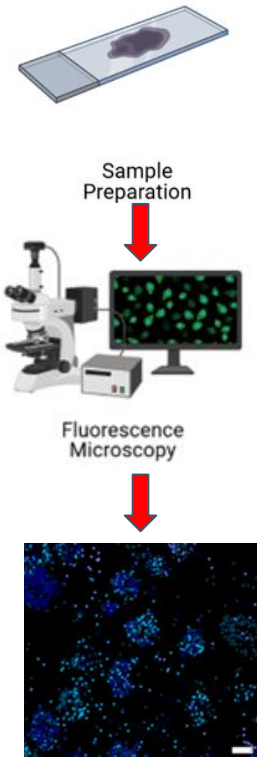
Project 3: #segmentation-x

Our segmentation-x team



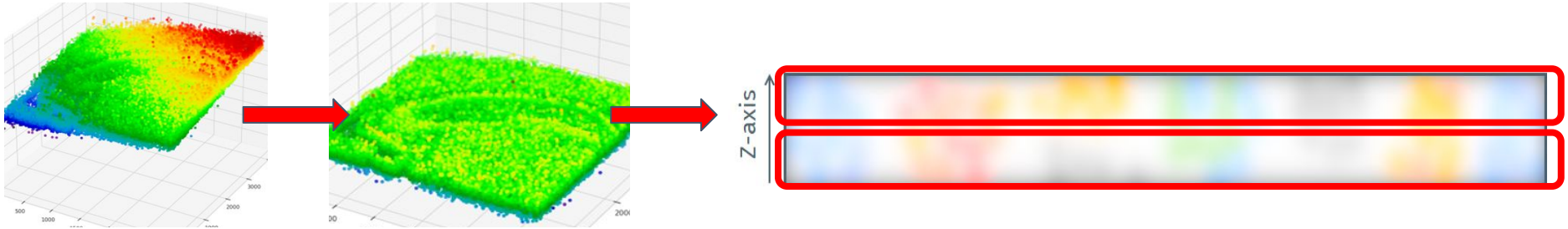
Project 4: #eval-overlapping-cells

- single-molecule-resolution spatial mRNA detection methods output a z-coordinate
  - Oftentimes ignored when analyzing tissue slices…
- Idea: Produce a light-weight algorithm to specifically detect cases of vertical signal incoherence (overlapping cells in the z-axis)

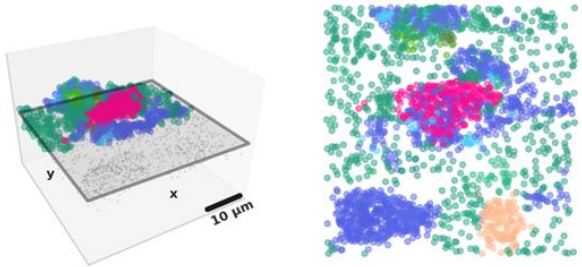


Project 4: #eval-overlapping-cells

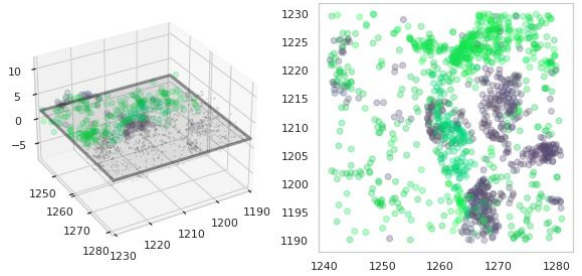
- centralize, analyze top/bottom signal individually



Results:



Xenium



MERFISH

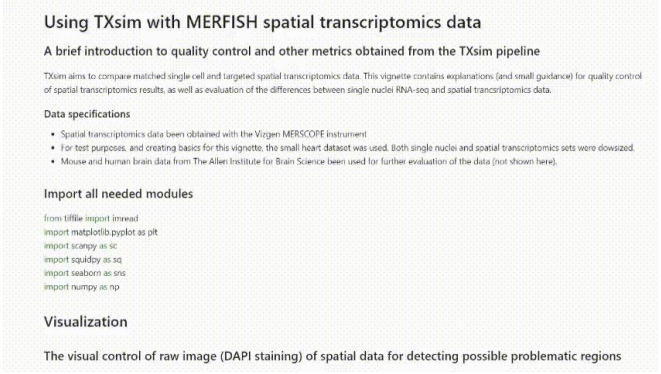
Project 5: #brain-evaluation

- **Project motivation - what is the problem?**
  - Quantitative differences between single-cell data and spatial data
  - Interpretation of quality metrics for mapping single-cell and spatial gene expression is not easy due to lack of general understanding of these metrics
- **Our solution**
  - Evaluate single cells of the brain tissue from mouse and/or human
  - Create vignettes for metric measures in the `TXsim` pipeline with extra visualizations for a better interpretation.

Project 5: #brain-evaluation

Results: successfully runned TXsim pipeline on multiple datasets

Vignette for selected metrics



Using TXsim with MERFISH spatial transcriptomics data

A brief introduction to quality control and other metrics obtained from the TXsim pipeline

TXsim aims to compare matched single cell and targeted spatial transcriptomics data. This vignette contains explanations (and small guidance) for quality control of spatial transcriptomics results, as well as evaluation of the differences between single nuclei RNA-seq and spatial transcriptomics data.

Data specifications

- Spatial transcriptomics data been obtained with the Vizgen MERSCOPE instrument
- For test purposes, and creating basics for this vignette, the small heart dataset was used. Both single nuclei and spatial transcriptomics sets were downsized.
- Mouse and human brain data from The Allen Institute for Brain Science been used for further evaluation of the data (not shown here).

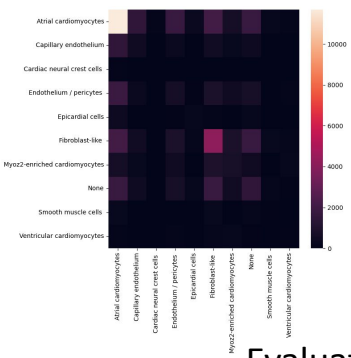
Import all needed modules

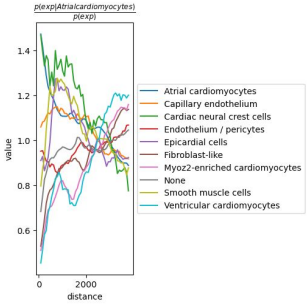
```
from tifffile import imread
import matplotlib.pyplot as plt
import scanpy as sc
import squidpy as sq
import seaborn as sns
import numpy as np
```

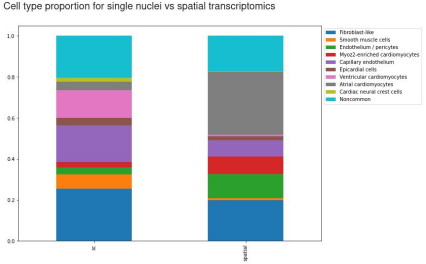
Visualization

The visual control of raw image (DAPI staining) of spatial data for detecting possible problematic regions

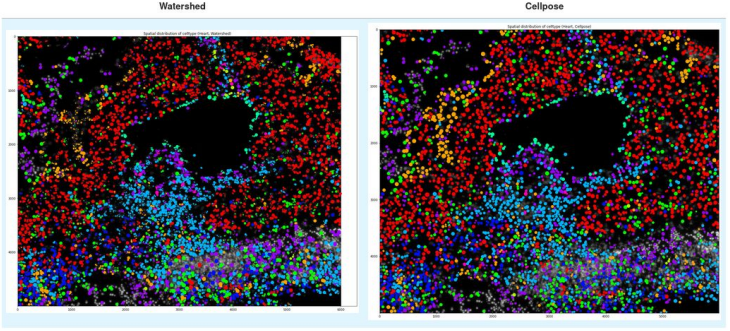
Plotting functions for selected metrics



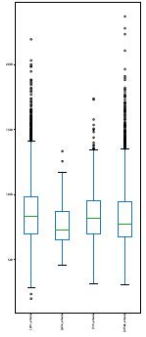




`Watershed` vs `cellpose`



Evaluate cell size differences between cell types

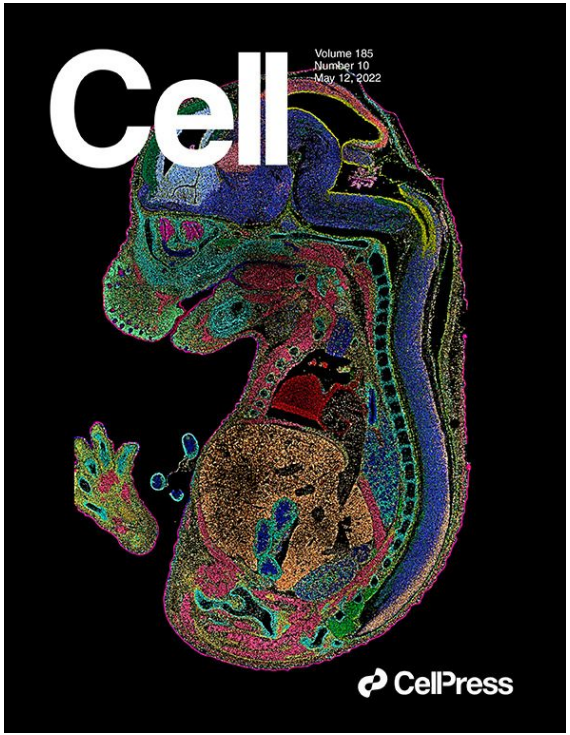


Project 5: #brain-evaluation

- Follow up:
  - Implement further metrics (e.g. distribution divergence)
  - Incorporate plotting functions in the snakemake
  - Cell type specific segmentation
  - Compare different tissues and different species

Project 6: #stereoseq

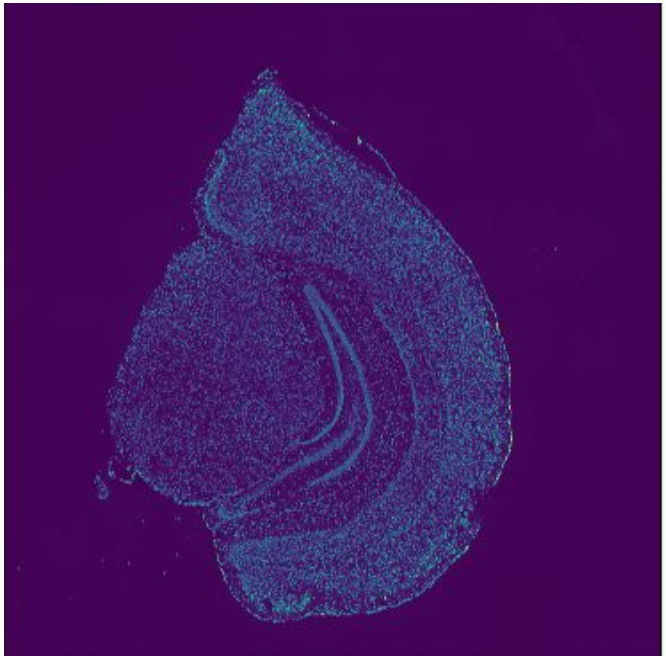
- **The team:**
  - Ian Dirk Fichtner
  - Shashwat Sahay
  - Lotte Pollaris
- **Goal: Explore STEREO-SEQ**
  - New promising spatial technology [Chen et al. 2021] [Xia et al. 2022]
  - Transcriptome-wide, sub-cellular spatially resolved data
- **Mid-hackathon status**
  - Curated some datasets
  - Define data structures
  - Troubleshooting



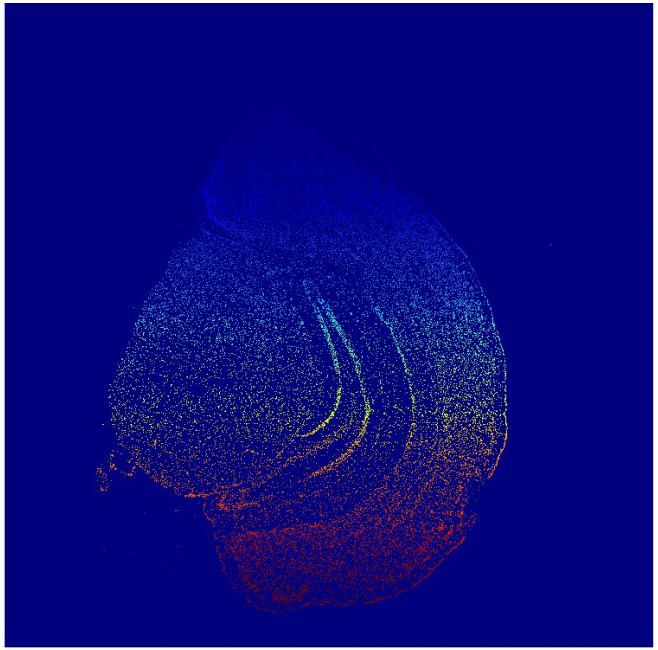
Project 6: #stereoseq

- **Results**

  8119736 cells **x 22413** transcripts

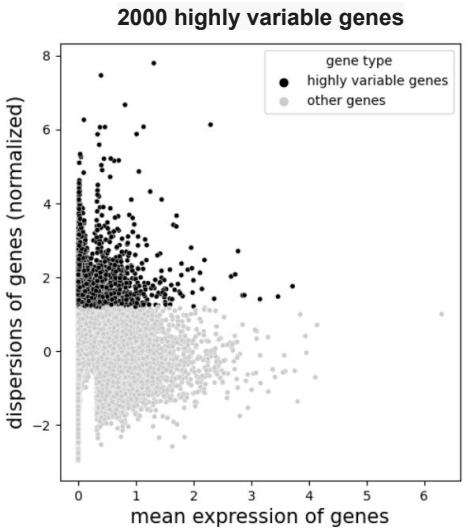


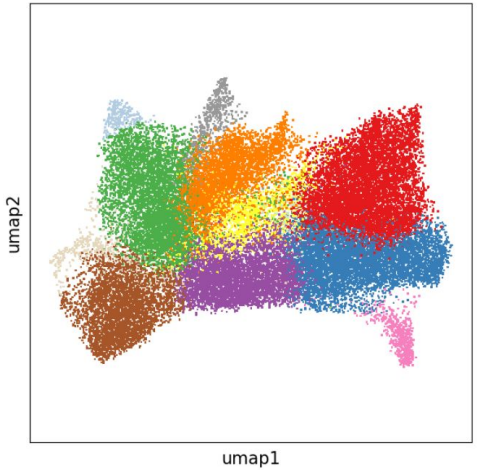
Segmented Watershed

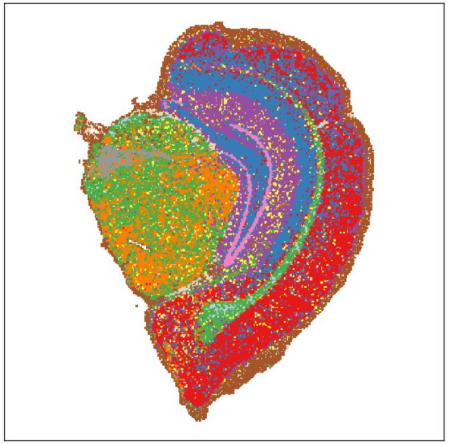


Project 6: #stereoseq

- **Results**

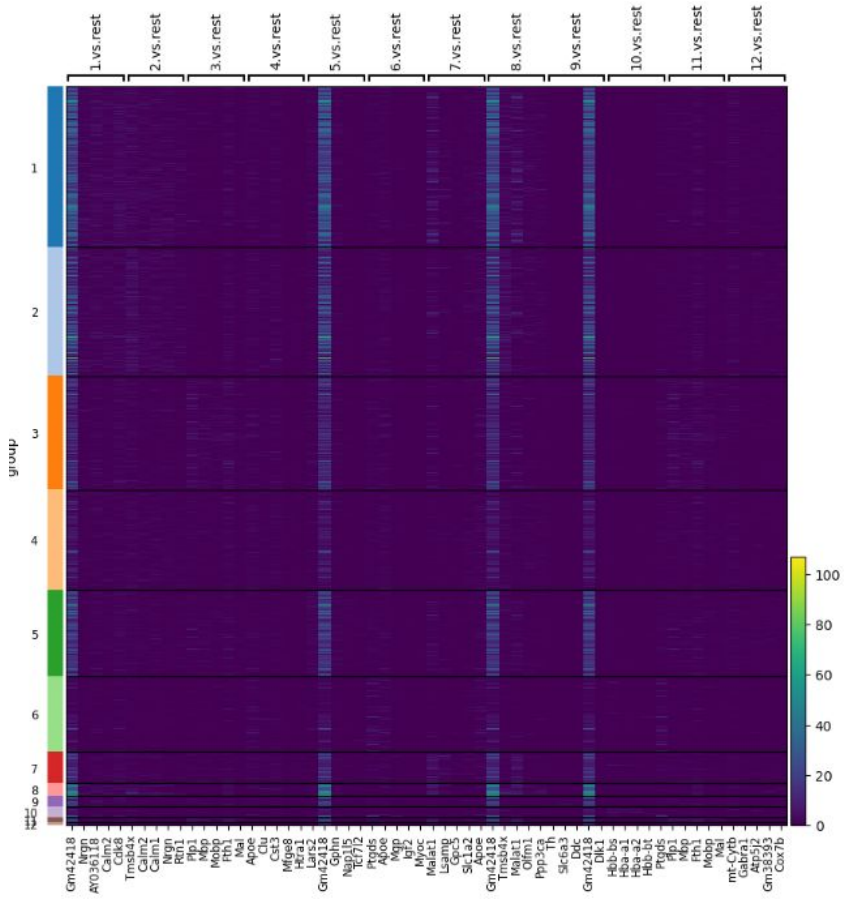






Project 6: #stereoseq

- **Results**



Project 6: #stereoseq

- **Outlook**
  - Set the groundwork for further downstream analysis
  - More segmentation methods
  - Allows for integration into `TXsim` pipeline
  - Evaluation of data quality

- **Challenges**
  - Segmentation methods are limited due to DAPI staining
  - Identification of cell boundaries with highly variable genes for segmentation is computationally challenging

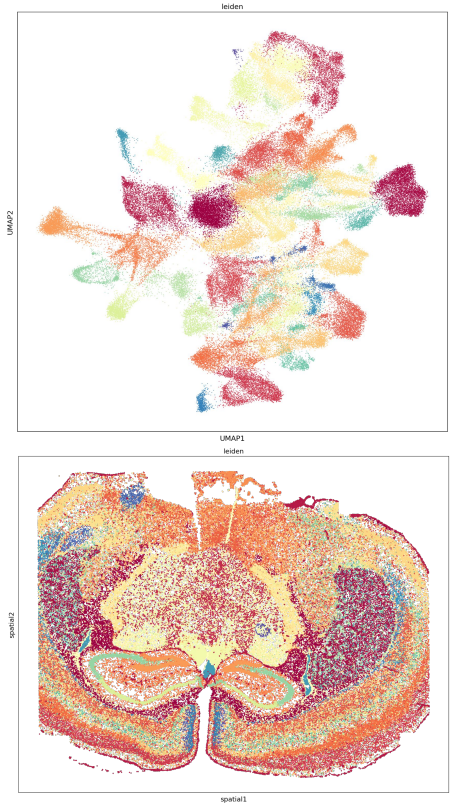
Project 7: #hyperbolic_brain

- **Project motivation - what is the problem?**
  - The cell type segmentation and annotation process can still be improved
- **Our solution**
  - Obtained Xenium Human Brain tissue dataset and applied a simple `squidpy` analysis and visualization process
  - Embed into the hyperbolic space and compare the cell clustering between the two different geometries (Euclidean and Hyperbolic)

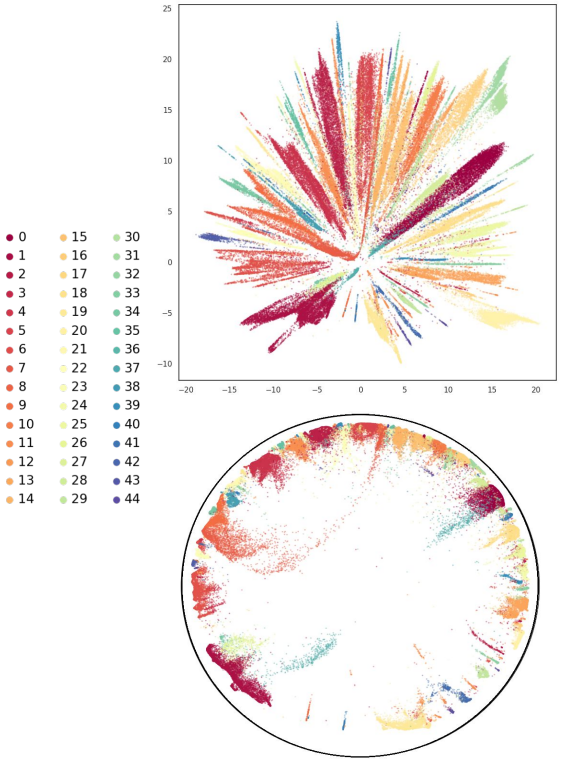
Mariane Kulik
Aimilia Christina Vagiona
Georgios Gavriilidis
Sebastian Tiesmeyer

Project 7: #hyperbolic_brain

Euclidean space



Hyperbolic space



BUT using classes from Leiden clustering (which uses euclidean distances)

:(

Next steps:

- Cluster based on the angular dimension of the Hyperbolic embedding
- Evaluate the biological significance of the “new” clustering on the spatial coordinates

Summary

Success?

- Exceeded realistic expectation of a few days of hacking
- Hypothesis generation & new ideas **>>** tasks addressed
- Ongoing collaboration planned!

A quick thank you for the unsung heros

In-person organisers:

- Louis Kuemmerle, Helmholtz Munich, Germany
- Naveed Ishaque, BIH, Germany
- George Gavriilidis, INAB/CERTH, Greece

Online organisers:

- Brian Long, Allen Brain Inst, USA
- Paulo Czarnewski, SciLifeLab, Sweden

People we didn't see but played a major role in organisation:

- **Malte Luecken, Helmholtz Munich, Germany**
- Sergio Salas, SciLifeLab, Sweden
- Fotis Psomopoulos, INAB/CERTH, Greece

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

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

Data Provider Department

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

QBiC

- Sequencing platform
- Proteomics platform
- Metabolomics platform
- Imaging platform

produce

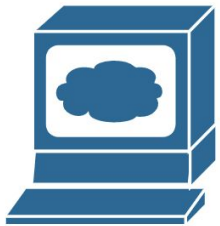
annotated datasets + studies

~ 10,000 datasets / year

Google Dataset Search

OmicsDI

access



web portal

Achievement:

- Mapping of own model to Schema.org `Dataset` and `DataCatalog`
- Registered QBiC at ROR
- First implementation draft for creation of `Dataset` pages (with Bioschemas)
- Pair-refactoring session on **MZmine3** (Java) with Olena

Lesson learned: Use the PID in the `@id` annotation

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 & Bioschemas and corresponding LinkML yaml schemas with "exact_mapping" elements.
- Developed prototype LinkML code for the translation from YAML to JSON-LD
- Deep-dive in to LinkML:
  - Discussions on importing other schemas,
  - Comparisons between the auto generated artefacts versus the Bioschemas supported ones
  - Model visualizations through Protege
- Prototype to generate GHGA JSON-LDs from YAML

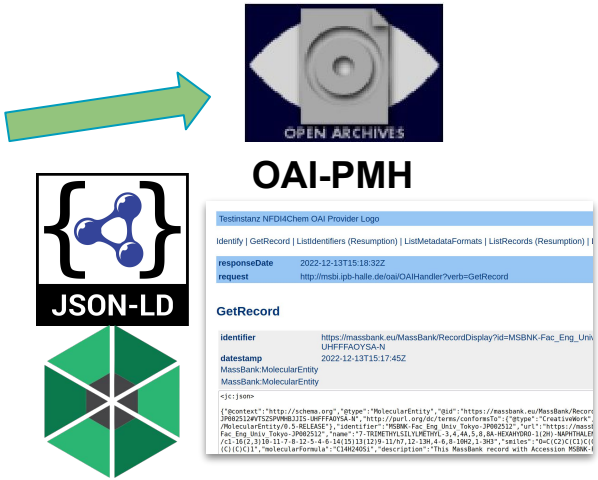
Open questions:

- Should Bioschemas offer a LinkML representation?
- How are metadata models in LinkML translated to RDF?
- How a mapping should look like?

Community department

- Supporting eDal repository (IPK) with their Bioschemas markup

Querying/Harvesting Department



Achievements:

- Successfully designed an Harvester to fetch & gather metadata from MassBank OAI Handler.

Google Dataset Search

Lessons learned:

- CKAN Provides JSON-LD script with SchemaOrg, which allows **Google Dataset Search** accessible.

Towards FAIR Computational Metabolomics Workflows - Improving Provenance Collection

Mahnoor Zulfiqar (University of Jena)
Kristian Peters (IPB Halle)
Michael R. Crusoe (ELIXIR-DE)
Simone Leo (CRS4)

Progress

The CWL workflow now takes one .mzML file, will parallelize with multiple .mzML files

Added provenance libraries in R Script (rdtLite) and Python Script (provenance)
to export “inner” provenance of those scripts

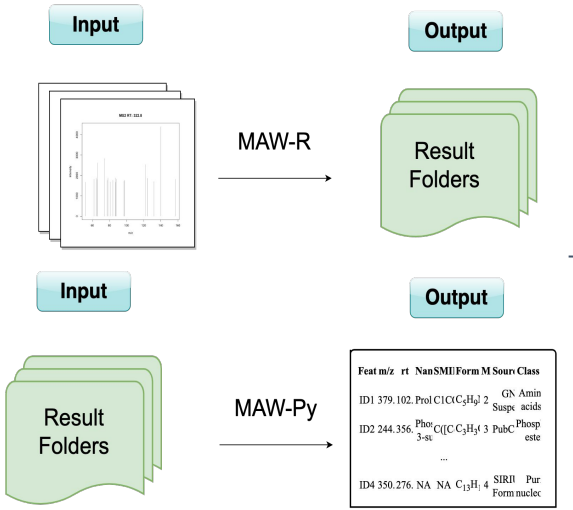
MAW is divided now into three parts:

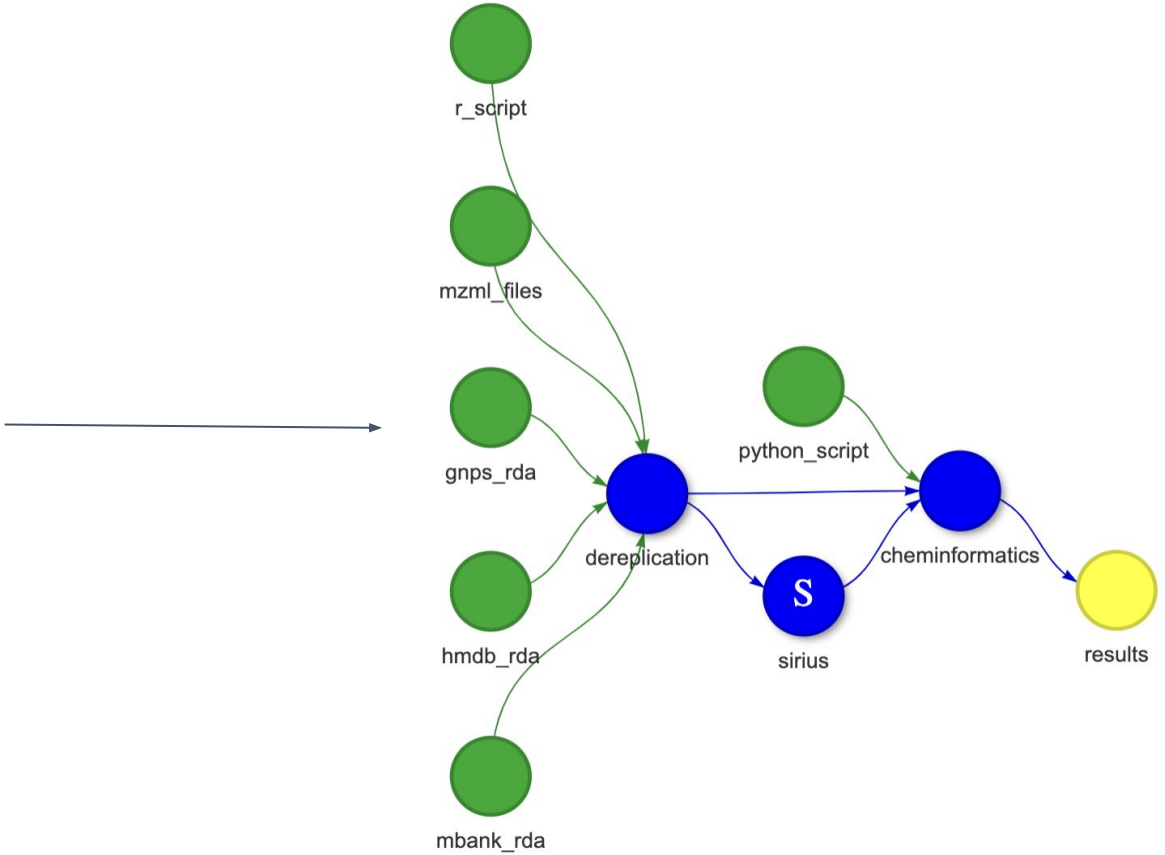
- MAW-R
- MAW-Sirius
- MAW-Py

In Progress:

- adding EDAM ontology format tag for the .mzML files
- Convert the CWLProv RO-Bundle to a Workflow Run RO-Crate using runcrate

Current CWL Workflow





Future Goals

- Explore how to pass the “inner” provenance information to CWLProv/RO-Crate
- Run the workflow using toil-cwl-runner on the ARA Slurm cluster with Singularity
- Add steps to the CWL workflow for downloading and updating the reference databases and caching
- Collect provenance questions specific to this workflow; what would we like to find out from the provenance?

  (Inspired by Renske de Wit https://doi.org/10.5281/zenodo.7113250)

Interactive data analysis and visualization in the web browser

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

Bla


de NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE


elixir
GERMANY

The ELIXIR::GA4GH Cloud

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

Mohsen’s dilemma

6+ years

vs

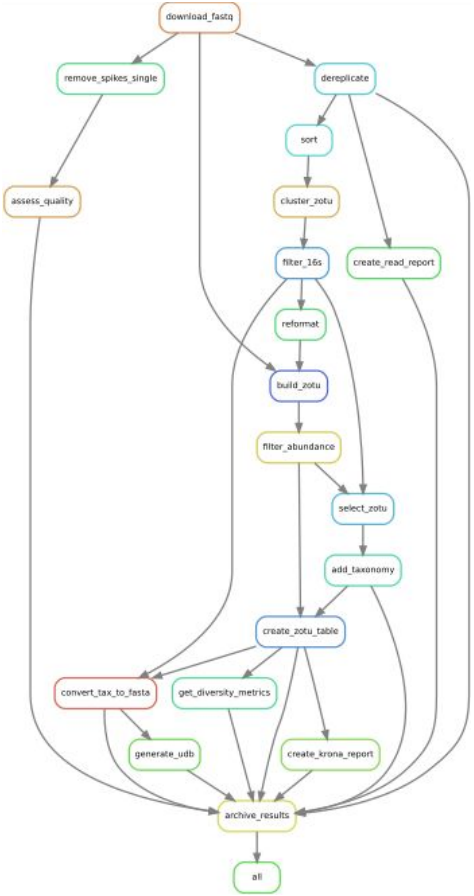
$1’000’000’000+

Snakemake workflow

- Wiring is +/- done, with dummy commands
- Completes successfully

```
[Thu Dec 15 19:48:57 2022]
Finished job 0.
211 of 211 steps (100%) done
Complete log: .snakemake/log/2022-12-15T194701.810374.snakemake.log
```

- Commands for each step +/- ready
- Testing the first few steps



Next steps

- Complete workflow
- Test cloud execution
  - File staging
  - Container caching
  - Use task federation?
  - …
- Pilot run
- Figure out costs & who will cover them :)
- Go for it :)

Lessons learned

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

GERMANY

USE FOSS!

- Use workflow languages
- Cloud is other people’s computers
- Horizontal scaling is useful, sometimes

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)

LipidXplorer to lxPostman

Jacobo

Daniel

✅**Progress:**

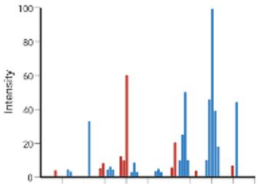
- Initial **mzTab-M output** added to **LipidXplorer**
- **Updated the GUI** to support mzTab-M output
- **Updated the CLI** to support mzTab-m output for webapp users
- Standalone executable and web version
- **Modified** LipidXplorer **MFQL** entries to prepare for **in-depth mzTab-M tables**
  - Small molecule feature table for **fragment abundances**
  - Small molecule evidence table for **fragment properties**

❌**Goal:** Glue webtools LipidXplorer and lxPostman

MZmine 3 - mzTab-M integration

Olena Mokshyna

MZmine 3



mzTab M

🎯 Goal:
- improve **export and import** support for mzTab-M
- map MZmine **annotation types** to mzTab-M & PSI-MS vocabulary

✅ Progress:
- mzTab **export module** is up and running
- MZmine annotations are successfully **exported**
- **Features** to implement and **issues** to fix are identified and noted

[Biohackathon project] Adding annotation data to mzTab-M export #1043

Draft omokshyna wants to merge 1 commit into `mzmine:master` from `omokshyna:mztab-hackathon`

Java

Collective effort! 💪

- Support and help 🤝 from other MZmine developers

Steffen Heuckeroth

Robin Schmid

- Pair-refactoring session 😎with **Sven Fillinger**

QBiC

Submission of mzTab-M to MetaboLights

🎯 Goal:

- create ISAtab files for submission to MetaboLights from mzTab-M

✅ Progress:

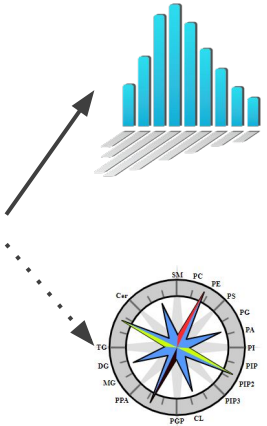
- Steffen Neumann provided code to create MAF (metabolites) file via metabolighteR
- metabolighteR was recently removed from CRAN and only supports reading of MetaboLights data
- tried Swagger codegen with MetaboLights REST API (Swagger 1.2, **EOL** ca. **2014**) =>

- Back to manual httr implementation 🚂 => create MTBLS study via REST

Steffen

Nils

rMTBLS



🥳

https://github.com/nilshoffmann/rmtbls

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

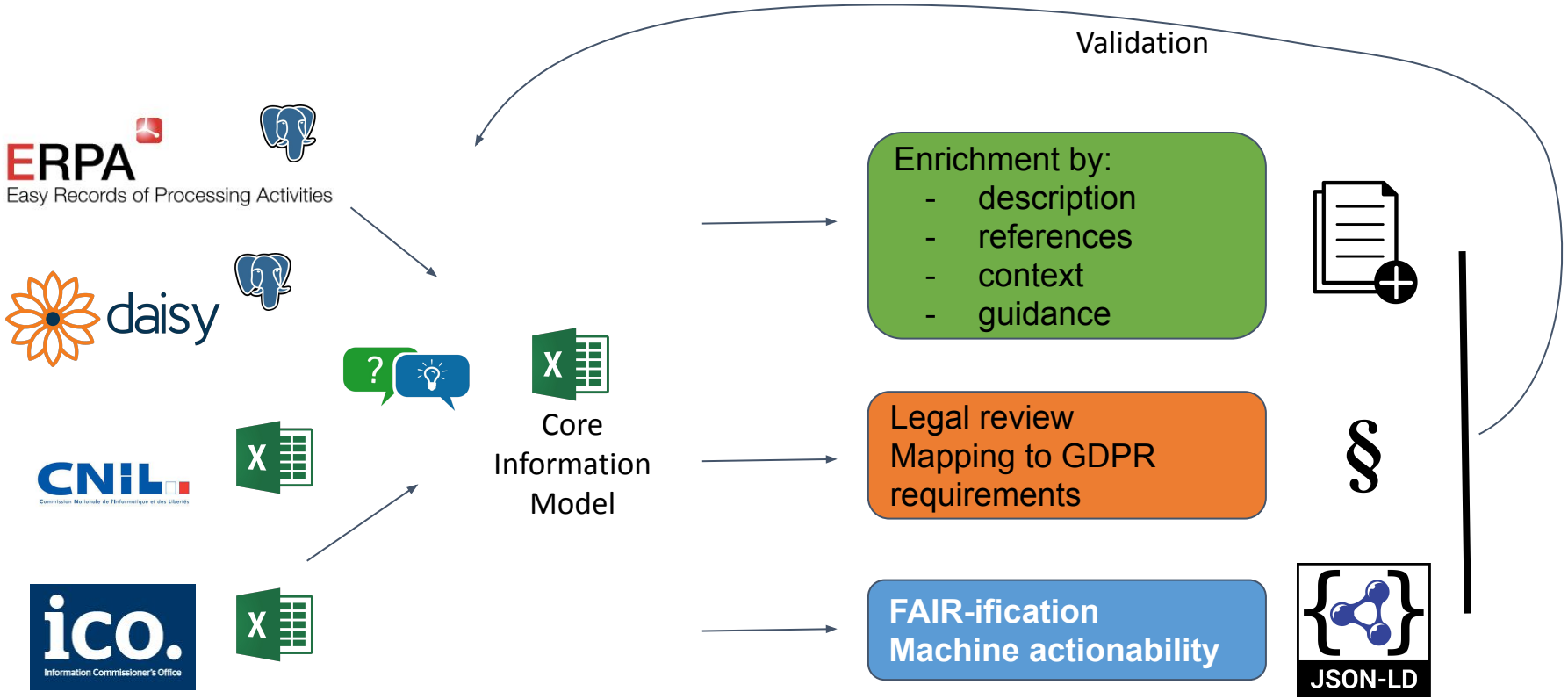
- The GDPR requires a written form for the Record of Processing Activities (Art. 30).
- The record **format can be chosen freely**, and it can be created on paper or numerically.
- Public organizations have to communicate the record to any person who demand it.

Reporting of sensitive data is big unknown for many stakeholders.

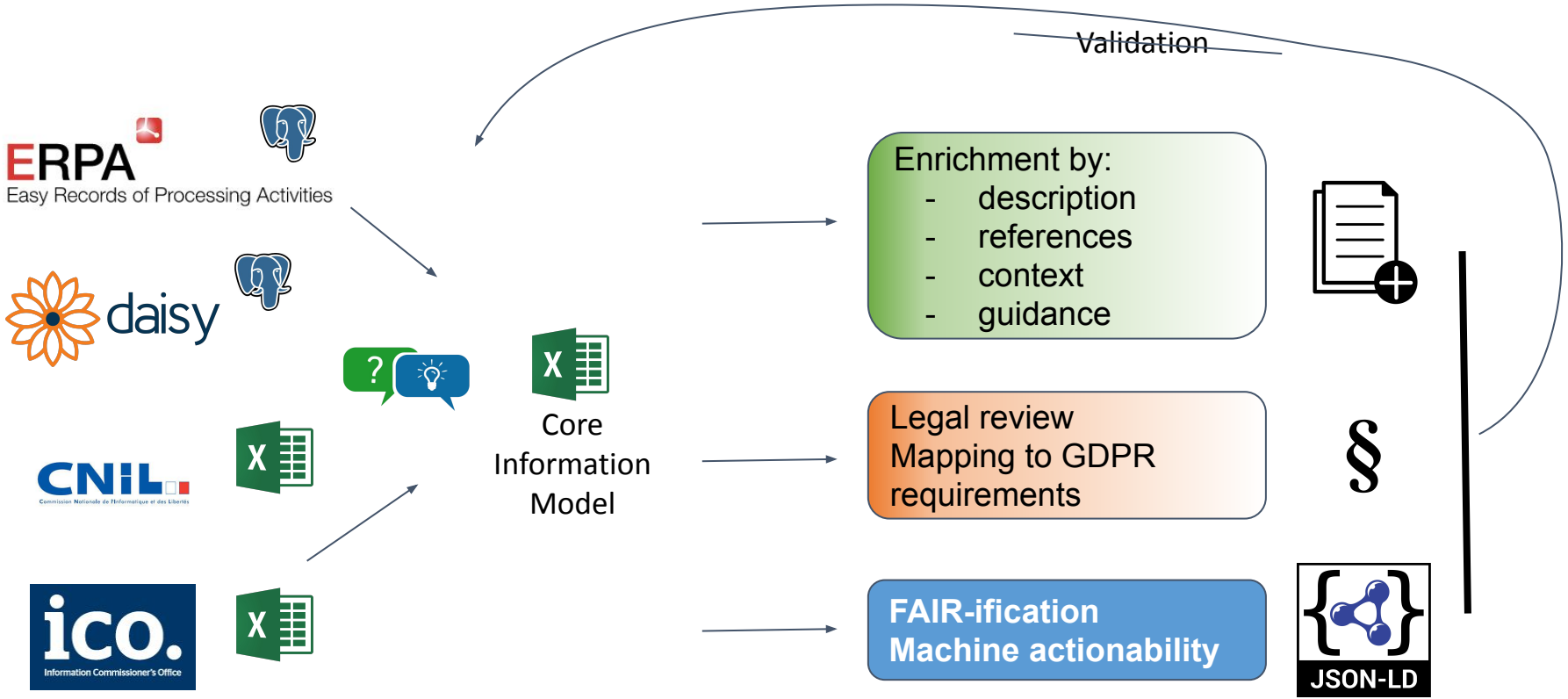
Lack of standards on format, content, scope, …

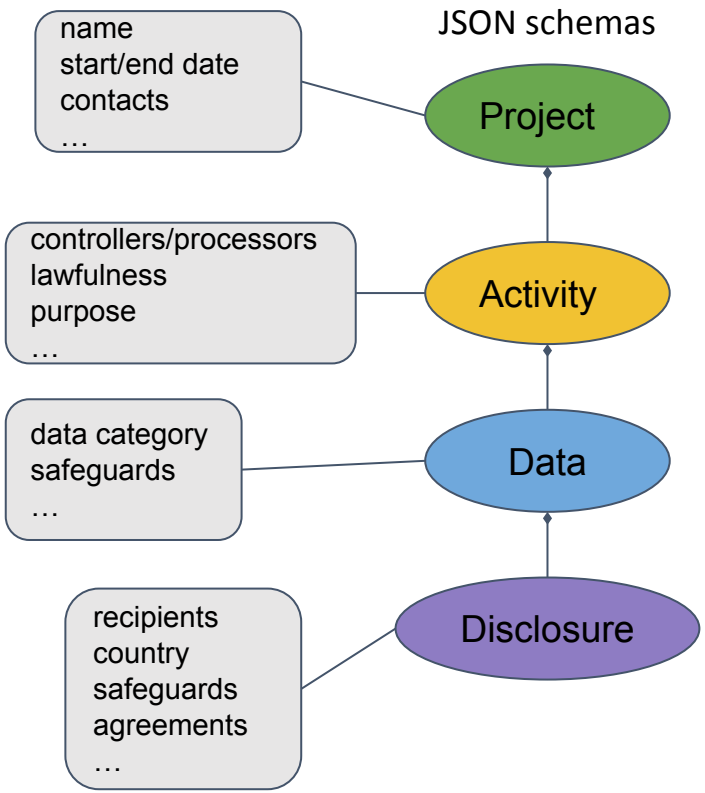
→ Minimal Information for Record Of Processing Activities (MIROPA)

Plan


Validation
ERPA
Easy Records of Processing Activities
daisy
CNIL
Commission Nationale de l'Informatique et des Libertés
ico.
Information Commissioner's Office
Core
Information
Model
Enrichment by:
- description
- references
- context
- guidance
Legal review
Mapping to GDPR
requirements
§
FAIR-ification
Machine actionability
JSON-LD

Work done


Validation
ERPA
Easy Records of Processing Activities
daisy
CNIL
Commission Nationale de l'Informatique et des Libertés
ico.
Information Commissioner's Office
Core
Information
Model
Enrichment by:
- description
- references
- context
- guidance
Legal review
Mapping to GDPR
requirements
§
FAIR-ification
Machine actionability
JSON-LD



Context mapping

Data Privacy Vocabulary (DPV)
https://w3c.github.io/dpv/dpv/

schema.org

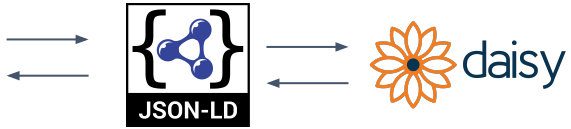
Next steps

Standalone GitHub repository

deNBI/biohackathon-2022  [elixir-luxembourg/MIROPA]

Extending exporters of existing tools

ERPA
Easy Records of Processing Activities



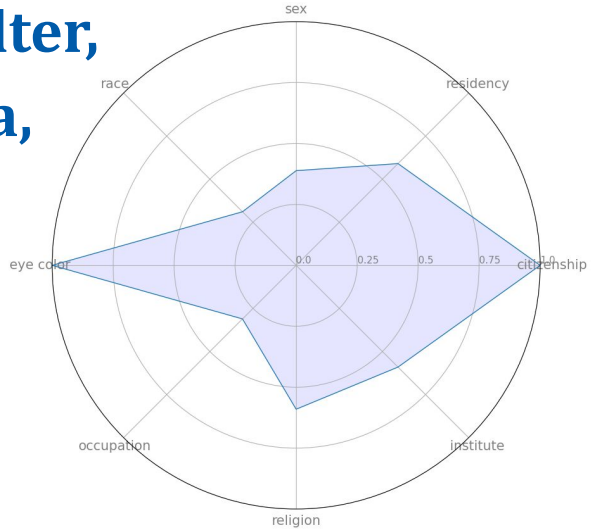
Promotion + review and validation by the community

Keep going…

- publishing (FAIRsharing.org)
- extending the context
- adding more detailed guidance and documentation
- …

Big thanks!

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



BioHackathon Germany
12.-16.12.2022

Interactive data analysis and visualization in the web browser

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

Nothing to report, but beautiful impressions

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

elixir GERMANY

Nothing to report, but beautiful impressions

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

elixir GERMANY

Nothing to report, but beautiful impressions

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

elixir GERMANY

Thank you

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

elixir
GERMANY