



OSCARS

Open Science Clusters' Action
for Research & Society

Funded Project

AMBCAT - Digital Amber Catalogue

Presenter: Jörg U. Hammel, Helmholtz-Zentrum Hereon, ORCID 0000-0002-6744-6811
Neele Rahmlow, Deutsches Elektronen-Synchrotron (IT)
Frank Schlünzen, Deutsches Elektronen-Synchrotron (IT)

• Implemented
by



Museum of
Natural History



CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

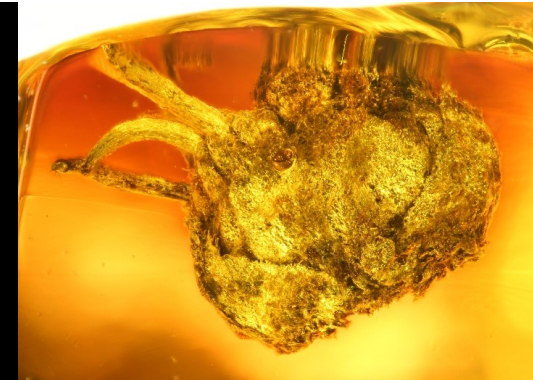
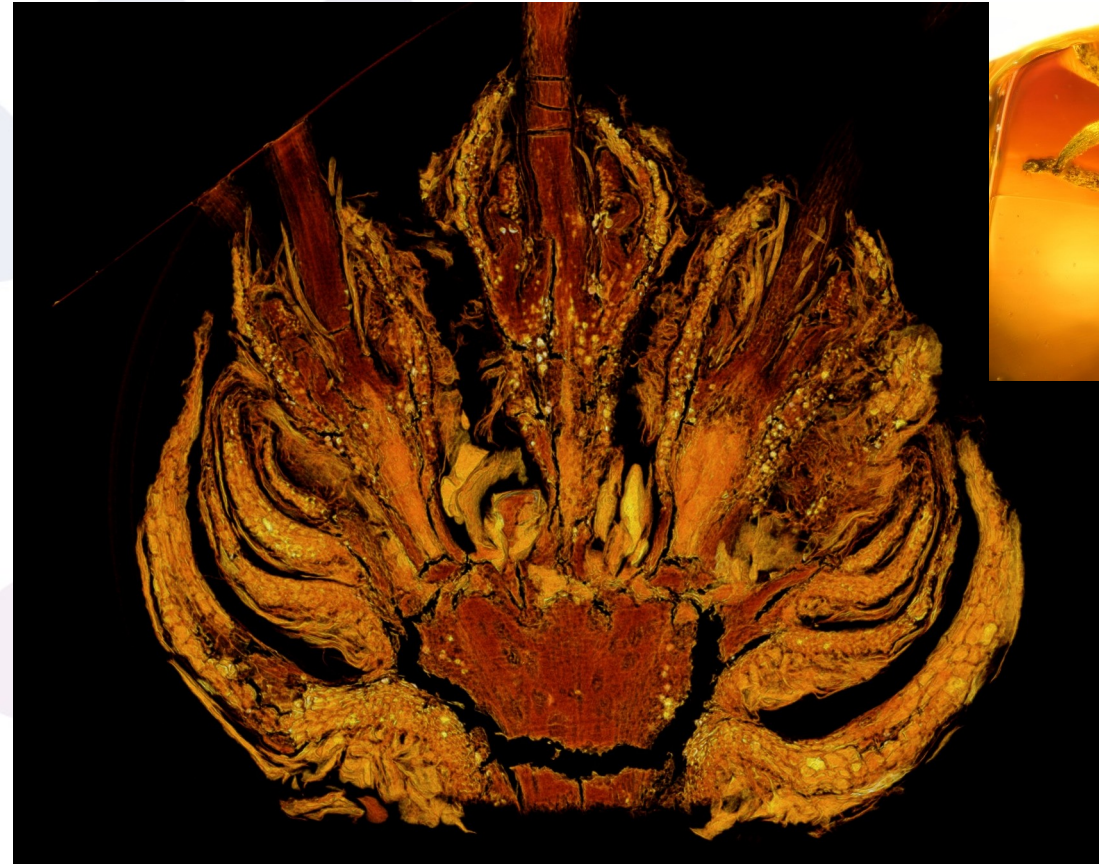


Funded by
the European Union

What problem(s) are you going to solve?

- Amber fossils are rare and unique objects -> access is limited
- Microscopic inspection is limited to externally visible characters
- X-ray computed tomography allows to study hidden characters and share data easily

Flowers of *Castanopsis kaulii* (Fagaceae) from 34-38 million-years-old Baltic amber



für Natur
MUSEUM FÜR
NATURKUNDE
BERLIN

Published in:

Sadowski, E.-M., J. U. Hammel, and T. Denk. 2018. Synchrotron X-ray imaging of a dichasium cupule of *Castanopsis* from Eocene Baltic amber. *American Journal of Botany* 105(12): 2025–2036.

What are you planning to do to solve the problem?

- More than 1500 unique amber fossils FAIR & OPEN available
- SciCat catalogue
- Userfriendly web frontend
- Crosslink metadata from related databases e.g. collection data from museum
- Interactive online visualization and inspection service included in the web portal



What will be the results and how do you plan to make them available to the broader community?

- Webservice providing access to CT data and additional meta data
- Online visualization tool implemented in webservice
- Free and open accessible as a service to the community and public



What risks could limit the success of the project, and how can they be mitigated

- All required data already available
- Use of established tools to implement the service
- Online visualization tools not yet tested – performance?



Who is doing it

- Dr. Frank Schluenzen (DESY)
- Neele Ramlow (DESY)
- Dr. Jörg U. Hammel (Hereon)

Consortium Use Case Partners

- Dr. Ulrich Kotthoff (LIB) - Germany
 - Dr. Danilo Harms (LIB) - Germany
 - Dr. Benjamin Wipfler (LIB) - Germany
 - Dr. Eva-Maria Sadowski (MfN) - Germany
 - Dr. Ricardo Pérez-de la Fuente (OUM) - UK
 - Dr. David Peris (CSIC-CMCNB) - Spain
 - Dr. Viktor Baranov (CSIC) - Spain
 - Dr. Brendon Boudinot (SMF) - Germany
 - Dr. Monica Solorzano-Kraemer (SMF) - Germany
 - Dr. Hans Pohl (FSU) - Germany
 - Dr. Joachim Haug (LMU) - Germany
-

What steps are to be taken?

1.) Data curation

- Ingest existing P05 data into SciCat
- Automate continuous ingestion for future experiments

2.) Tools for data maintenance

- Improve data curation with user-friendly tools
- Interface with external sources (e.g. museums or universities)

3.) Implementing AMBCAT

- Create a new Streamlit web application
- Architecture similar to the 'Human Organ Atlas'

4.) Data visualisation

- Investigate ParaView
- Integrate features into AMBCAT



Tasks

- Ingest existing P05 data into SciCat
- Automate continuous ingestion for future experiments




```
def main() -> None:
    # Get the data path given as an argument by the user
    1 path = ArgumentParser().get_path_from_argument

    # Create a scicat metadata class that reads files containing metadata
    2 metadata_creator = SciCatMetadata(path)
    # Check if files could be read and metadata can be created for scicat
    error = metadata_creator.can_create()

    if not error:
        # Create common metadata
        metadata_common = metadata_creator.retrieve_common_metadata()
        3 # Create scientific metadata
        metadata_scientific = metadata_creator.retrieve_scientific_metadata()

        # Combine metadata
        metadata = {
            **metadata_common,
            **{"scientificMetadata": metadata_scientific},
        }

        # Read ingestor credentials
        ingestor_username, ingestor_password = get_ingestor_credentials()

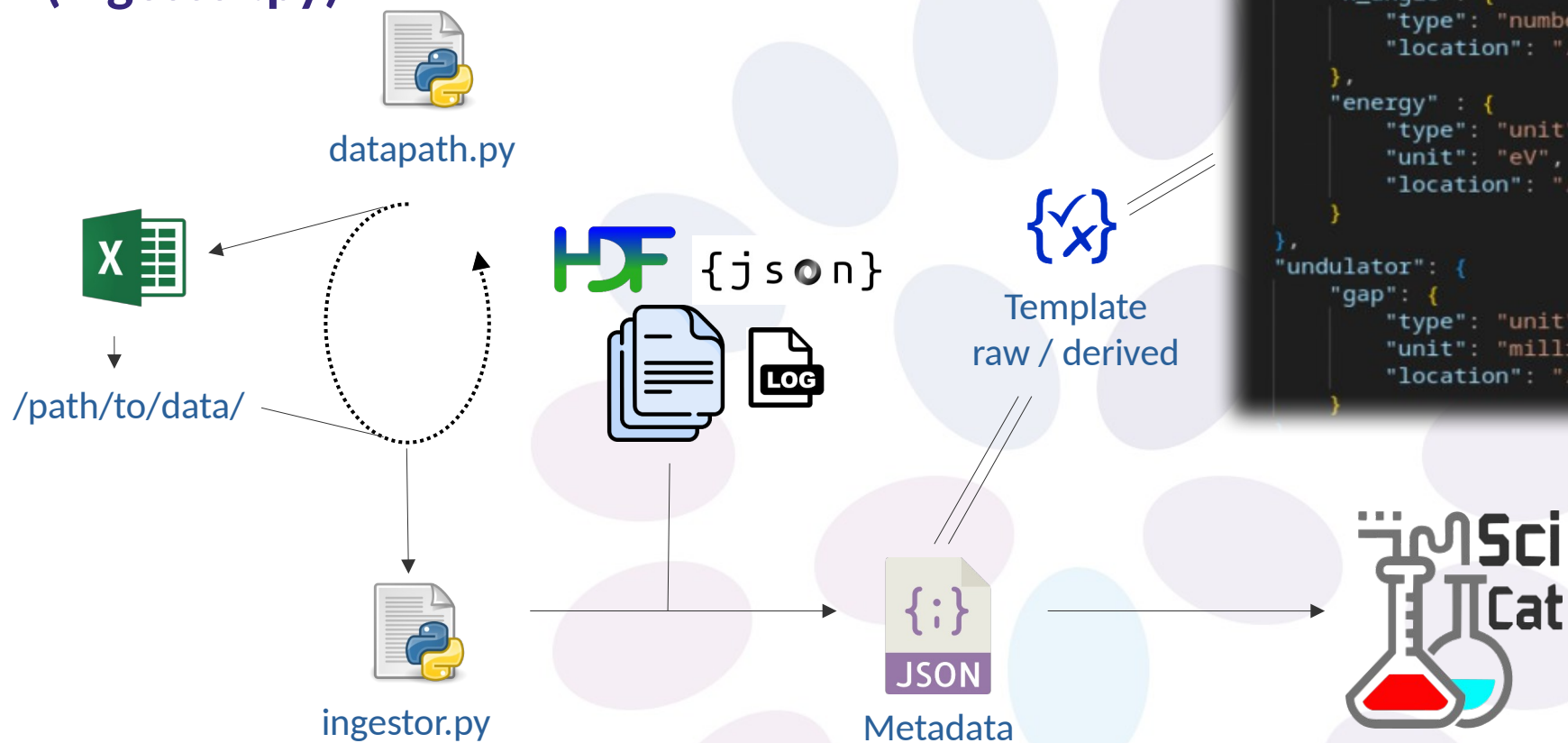
        # Ingest metadata to scicat
        scicat = SciCatRequestor(ingestor_username, ingestor_password)
        error = scicat.can_ingest(metadata)
        4 if not error:
            scicat.ingest_dataset(metadata)

    # Print an error message (if it occurs) and save it in a file
    5 if error:
        ErrorPrinter().print_error(path, error)
```

Python programme (ingestor.py)

- 1) Takes a data path as an argument
- 2) Reads HDF5, JSON and log files for one dataset
- 3) Creates general and scientific metadata from these files automatically
- 4) Uploads the metadata to SciCat, or updates the data if PID already exists
- 5) If an error occurs, an error message is saved

Python programme (ingestor.py)

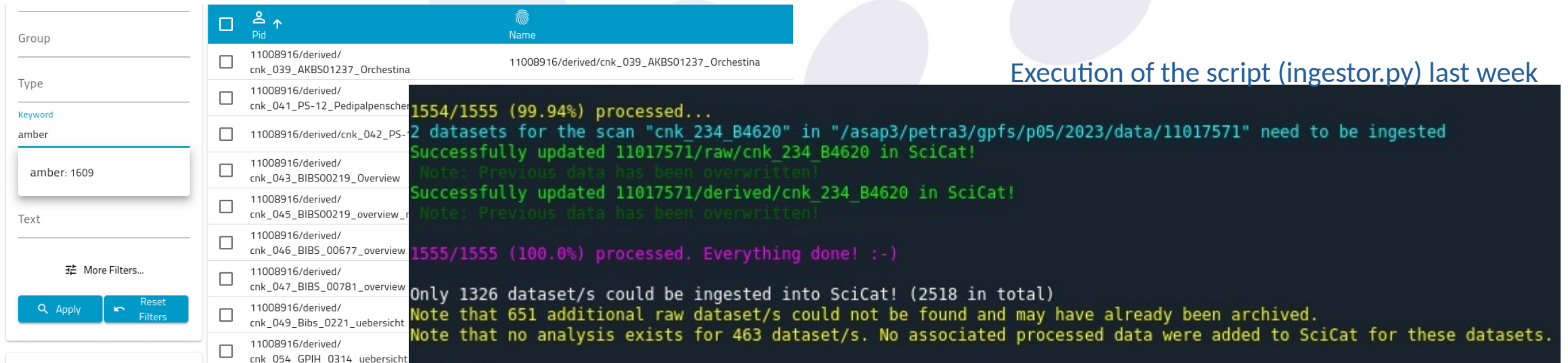


```
schema" : {  
  "scan" : {  
    "mode": {  
      "type": "string",  
      "location": "/entry/scan/mode"  
    },  
    "n_angle": {  
      "type": "number",  
      "location": "/entry/scan/n_angle"  
    },  
    "energy" : {  
      "type": "unit",  
      "unit": "eV",  
      "location": "/entry/scan/setup/pos_p05_energy"  
    }  
  },  
  "undulator": {  
    "gap": {  
      "type": "unit",  
      "unit": "millimeters",  
      "location": "/entry/hardware/undulator/gap/position/value"  
    }  
  }  
}
```

Python programme (ingestor.py)

- 762 raw and 847 derived datasets (**1,609 in total**) already added to SciCat
- Some of the data has already been archived and must be retrieved
- Code and templates can be extended and used for future datasets
- Further metadata will be added, especially attachments

Execution of the script (ingestor.py) last week



The screenshot shows a web interface on the left with filters for 'Group', 'Type', and 'Keyword'. The 'Keyword' filter is set to 'amber' with a count of 1609. Below the filters are 'Apply' and 'Reset Filters' buttons. To the right is a table with columns 'Pid' and 'Name'. The table lists several datasets, including '11008916/derived/cnk_039_AKBS01237_Orchestina' and '11008916/derived/cnk_041_PS-12_Pedipalpenschel'. Overlaid on the right is a terminal window showing the output of the ingestor.py script. The terminal output indicates that 1554/1555 (99.94%) datasets were processed, 2 datasets for the scan 'cnk_234_B4620' need to be ingested, and 1326 dataset/s could be ingested into SciCat (2518 in total). It also notes that 651 additional raw dataset/s could not be found and may have already been archived, and that no analysis exists for 463 dataset/s.


```
1554/1555 (99.94%) processed...
2 datasets for the scan "cnk_234_B4620" in "/asap3/petra3/gpfs/p05/2023/data/11017571" need to be ingested
Successfully updated 11017571/raw/cnk_234_B4620 in SciCat!
Note: Previous data has been overwritten!
Successfully updated 11017571/derived/cnk_234_B4620 in SciCat!
Note: Previous data has been overwritten!
1555/1555 (100.0%) processed. Everything done! :-)
```

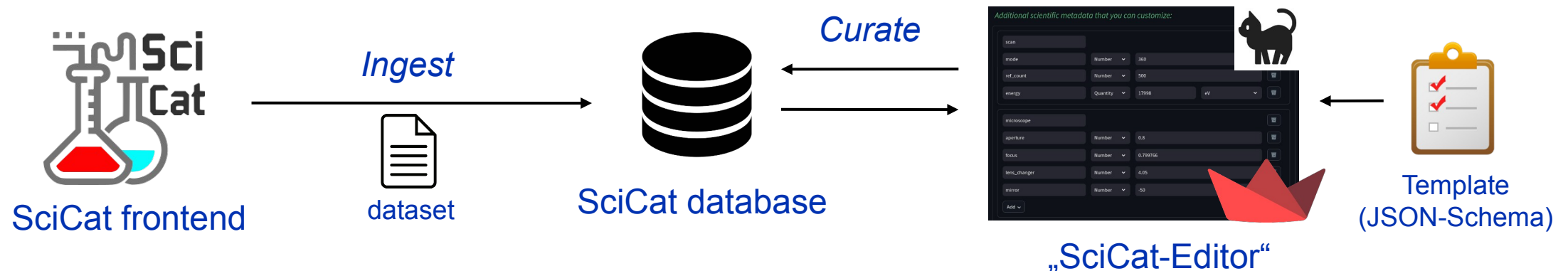
Only 1326 dataset/s could be ingested into SciCat! (2518 in total)
Note that 651 additional raw dataset/s could not be found and may have already been archived.
Note that no analysis exists for 463 dataset/s. No associated processed data were added to SciCat for these datasets.

Tasks

- Improve data curation with user-friendly tools
- Interface with external sources (e.g. museums or universities)

SciCat-Editor

- Web application created with the Streamlit framework 
- Allows manual input of scientific metadata for datasets in SciCat
- Structure and validation of metadata can be specified via a template (e.g. data types, mandatory fields, ...)



SciCat-Editor

Datasets

A total of 12 datasets were found!

Dataset ID ↑	Dataset Name	Type	Ownergroup
cenak_134_2041	cenak_134_2041	raw	11016664-part
cenak_134_2041_reconstruction	cenak_134_2041	derived	11016664-part
randomPID	anotherTest	raw	string
string	string	raw	string
undefined/3161a6a3-5597-4737-a89e-050d401ae211	nfs/neelee	raw	it
undefined/671e7351-7e82-4d1d-a6fa-2261948061ce	First Test Dataset	derived	admin

Metadata found in the p05 schema:

mandatoryString	String		✱
mandatoryNumber	Number		✱
optionalBoolean	Boolean	Choose an option	✱
optionalValue	Date	2024/08/26	✱
mandatoryValue	Datatype		✱
mandatoryObject			✱
mandatoryChild	String	testValue	✱
optionalChild	Quantity	Choose an option	✱

=> Curation is simplified

=> Control, guidance and validation

Website is currently unavailable, but will be relaunched in the future!

SciCat-Editor

- Data curation can be improved using the SciCat-Editor
- Further data fields, such as comment, rate or linking data, need to be added

Automated ingestion (via scripts)

- Just as important, and enables metadata to be added quickly in the future too
 - Some information is still missing, such as sample information
(code of automated scripts must be extended)
 - Curation of amber data beyond P05 data is also planned (e.g. Senckenberg Search)
-

2.) Tools for data maintenance

» Identifikation

AQUILA-ID sesam-1558437
Katalognummer 3724-Be 3724.1a
Katalognummer (num.) 3724
Sammlungsname Bernstein - SMF

» Bestimmung

Taxon *Hypotrigona kleineri* Solórzano-Kraemer and Engel, 2022
Taxonomie Animalia-fossil: Arthropoda: Mandibulata: Hexapoda (Insecta): Pterygota: Hymenoptera: Apidae: Meliponini: Hypotrigona: kleineri
Typus Holotypus

» Material - tabellarisch

Erhaltung	Präparationsart	Stadium	Präparatbeschreibung	Bemerkung	Anzahl	Körperteil
Harz	Synchrotron	adult	Eingebettet in araldite	preserved in piece SMF Be-3724	1	Körper

» Fundinformationen

Fundortname Defaunation resin from Tanzania
Administrative Einheit Tansania
Kontinent Afrika

» Sammlungsobjekt – allgemein

Bemerkung Objekt Syninclusions: 1 Hypotrigona kleineri SMF-Be 3724.2a Paratype

» Verknüpfte Medien



Head in lateral view of SMF Be 3724.1a.jpg



Habitus in ventral view of SMF Be 3724.1a.jpg



Head in anterior view of SMF Be 3724.1a.jpg

Editor

Missing data, need to be added

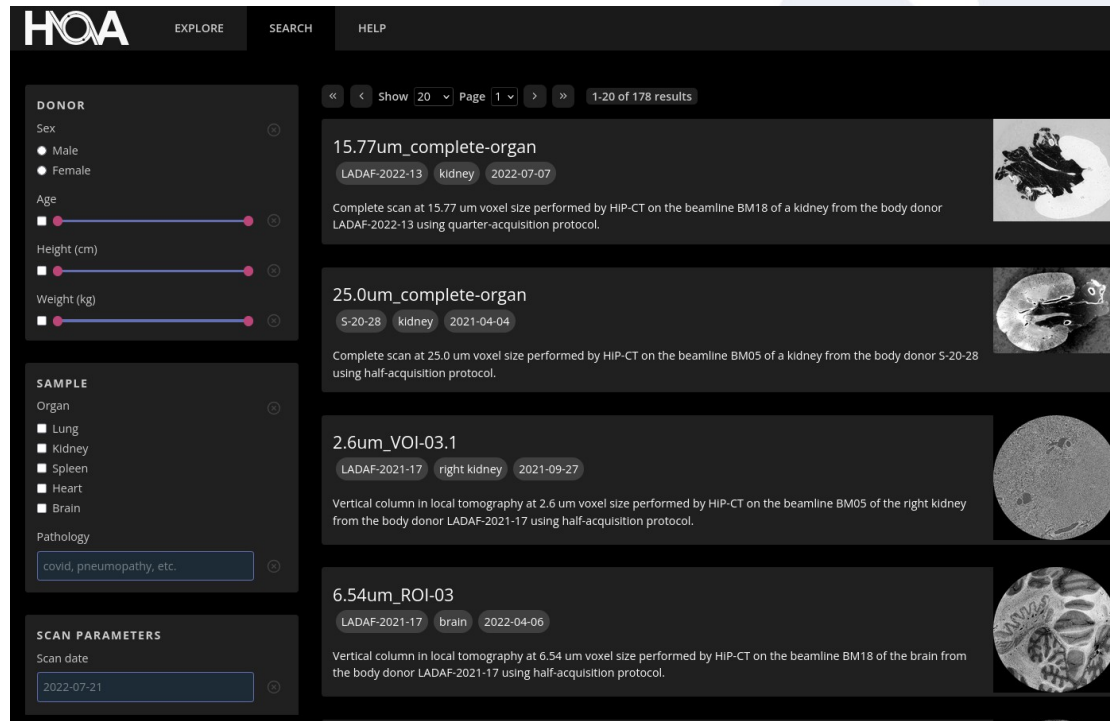
Added quickly in the future too
information

planned (e.g. Senckenberg Search)

3.) Implementing AMBCAT (Amber Catalogue)

Tasks

- Create a new Streamlit web application
- Architecture similar to the 'Human Organ Atlas'



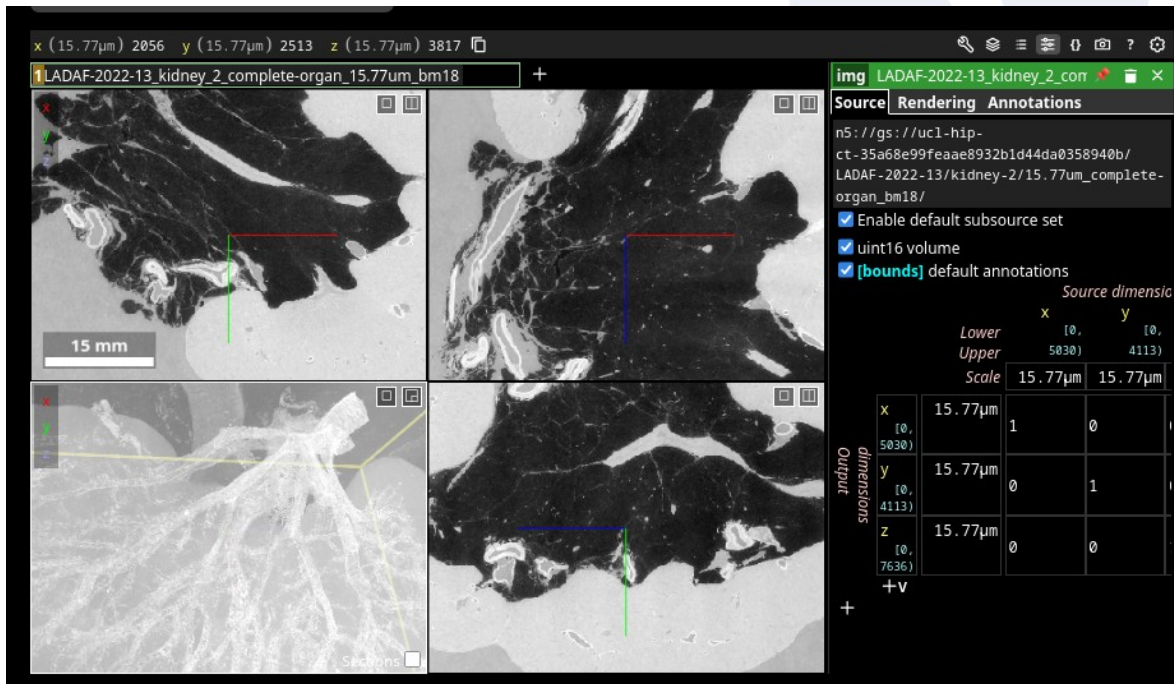
Search, Access and Explore

- Amber datasets from DESY and partner institutions are displayed
- Data is searchable and filterable
- Links provide access to the data
- Data can be viewed using 3D models

The web application "Human Organ Atlas"

Tasks

- Investigate ParaView
- Integrate features into AMBCAT



ParaView

- An application for data analysis and visualisation
- Hopefully, it will enable improved 3D visualisation compared to the Human Organ Atlas (to be tested)

3D models on the “Human Organ Atlas” website

AMBCAT (Amber Catalogue)

- AMBCAT is a Streamlit web application that enables users to **search for, access and explore** amber data
- AMBCAT serves as a collection point for amber data, providing a **higher-level context** (in terms of temporal/geographical origin, phylogenetics and ecological environment)
- AMBCAT makes a significant contribution to the **FAIR** data principle
- AMBCAT is accessible to everyone and can raise public ecological awareness





Thank You!
