



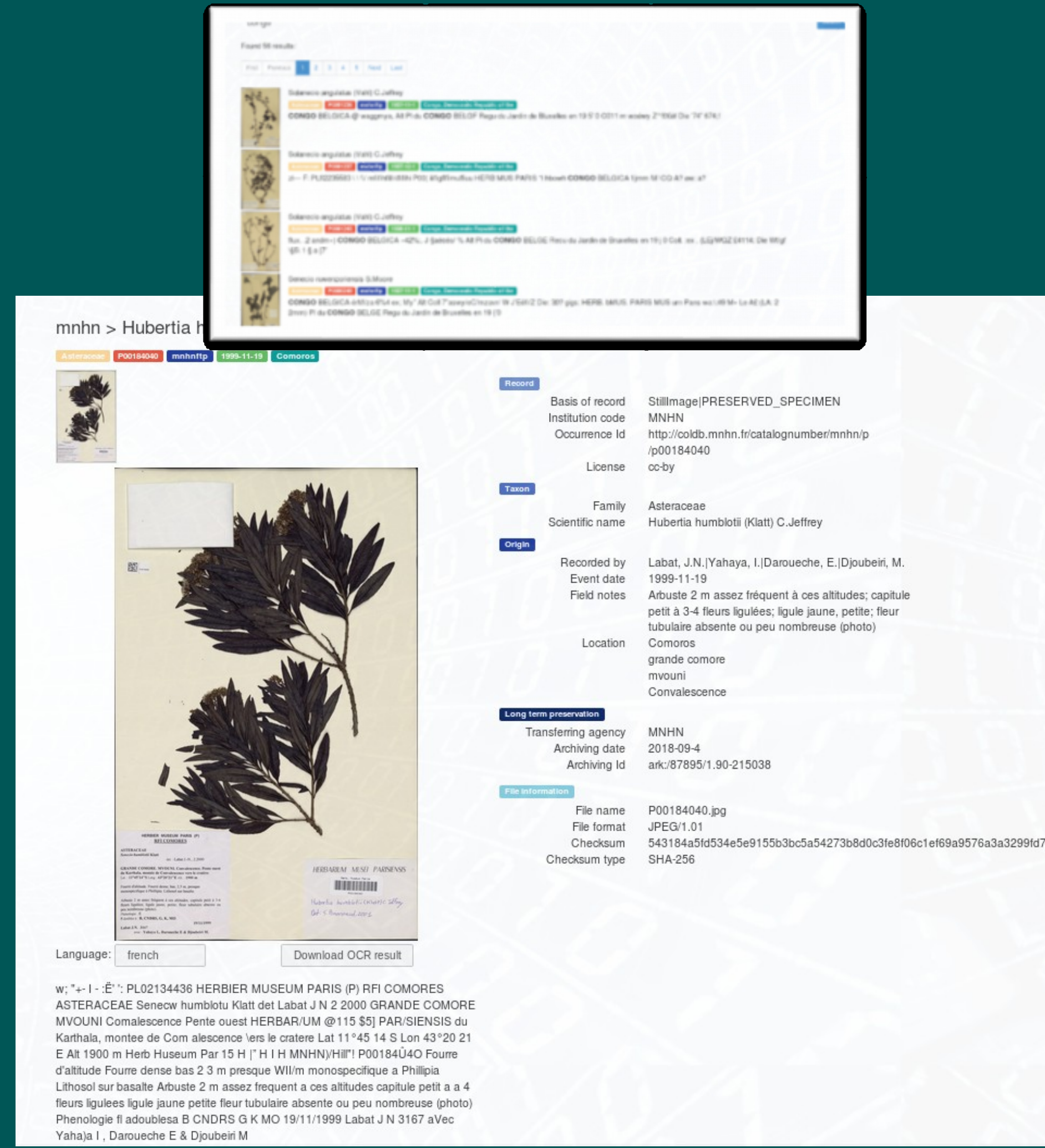
INFRASTRUCTURE FOR LONG-TERM PRESERVATION AND OCR ANALYSIS OF HERBARIUM IMAGES

Innovation and consolidation for large scale digitisation of natural heritage

OPENDATA

Full text search portal

The open data portal CINES enables a full text search on metadata and OCR results to display a list of matching results and then to access the herbarium sheet.
URL : <https://opendata.cines.fr>



CHALLENGES

500 million specimens

in herbaria all over the world. Most scientific corpora is available in analog form but physical copies are fragile. Digital copy must be preserved without loss of information in future.

- global trend to industrial digitizing
- data difficult to handle even for medium size institutes
- same challenges being faced by potentially hundreds of herbaria in Europe
- makes sense to work together to develop a solution

impacts on research activity

- full text search on OCR results
- sharing knowledge on image analysis, specimens will be discoverable by the entire scientific community
- characterisation of features, annotation, meta data extraction
- organisation of the data and metadata in certain formats
- data curation to ensure that meta data and data formats remain meaningful in future



ACHIEVEMENTS

> 4 million specimens

- have been ingested at CINES computing center based in Montpellier, France
- the whole processing chain including, quality control, OCR analysis on HPC facilities and indexing on a full text searching using Elastic Search engine has been implemented.

metadata

- Darwin Core and Dublin Core based
- have been imported in the Trusted Digital Repository (TDR) and mapping has been done on the long-term archival system from the Global Biodiversity Information Facility



PERSPECTIVES

enhanced data analysis

- plant morphology recognition
- Big Data mining
- crowdsourcing, collaborative knowledge-building

eTDR European Trustworthy Digital Repository

- securing and certifying long-term archiving
- long distance replication with EUDAT partners.



scientific use-cases

- mining for duplicates in collections located in different museums
- exploring the data using deep learning algorithms
- identifying seldom species



Workflow and infrastructure design EUDAT-CINES

