Long-term preservation of herbarium specimen images

Herbadrop is both an archival service for long-term preservation of herbarium specimen images and a tool for extracting information by image analysis. Developed by five institutes from Finland, France, Germany, Netherlands and Scotland it aims to be available to other herbaria in the future.

Making the specimen images and data available online from different institutes allows cross-domain research and data analysis for botanists and researchers with diverse interests (e.g. ecology, social and cultural history, climate change).

The Scientific Challenge

Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. High-resolution images of these specimens require substantial bandwidth and disk space. New methods of extracting information from the specimen labels have been developed using Optical Character Recognition (OCR) but using this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts.

Much of the information is available only using handwritten text recognition or botanical pattern recognition which is less mature technology than OCR.

Who benefits and how?

The information on the specimens can serve as basis for diverse scientific disciplines. Specimens are primarily used for taxonomic and systematic research to identify, describe, classify, and name species. Plant checklists and floras comprise the species range of a certain area. Occurrence information derived from specimens is important for the compilation of checklists. But also other fields make use of Herbarium data including Paleobotany, Phylogeneny, or research of Ecosystem dynamics.

Making the specimen and data online available and merging data from different institutions allows cross domain research and data analysis. Undescribed species stored in herbaria can be discovered more easily.

Distribution information about species can be evaluated along past periods. Herbarium data can give proof of the point of time when an invasive species occurred for the first time in a certain area. Historians can analyze herbarium data tracing back Itineraries from historical characters. The data can be used to calibrate predictive models of the oncoming changes in biodiversity patterns under global threats. This diverse information will be useful for a wide user community including conservationists, policy makers, and politicians.

Safeguarding long-term data storage is an important precondition for a reliable accessibility of herbarium information. Thanks to this pilot, long-term storage for herbarium images will be guaranteed. The specimen will be thus discoverable for the entire scientific community.

Technical Implementation

The Herbadrop architecture is divided into a sequence of functions that process one-step of the workflow.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

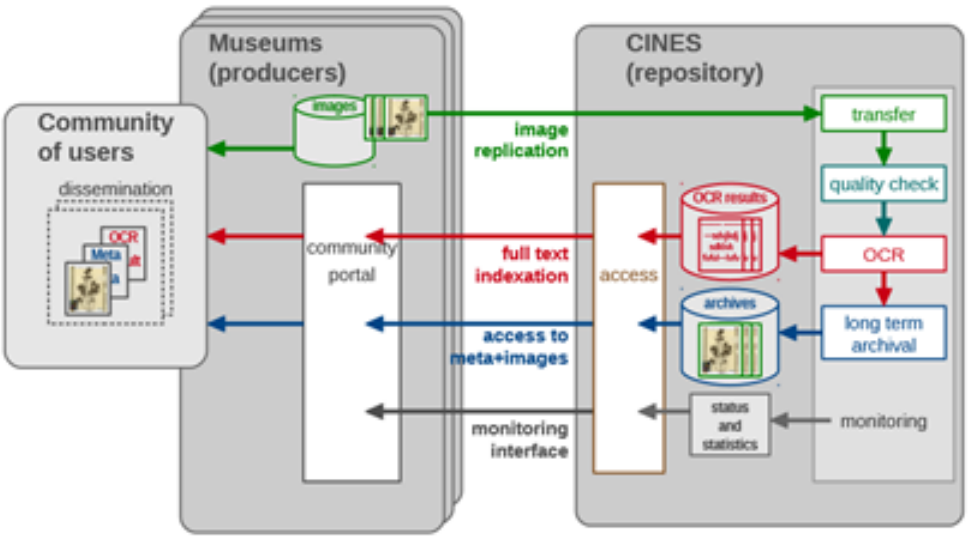

Figure 1 - Herbadrop architecture

Preliminary Results

The pilot use case has been refined, especially the workflows between the B2SAFE submission and the process of data & metadata. It has currently performed the ingestion of more than 2 Million images representing 11 M OCR files and 1,78 M images processed, equalling more than 15 TB of volume and 180 000 hours of computation power.

The access function to the objects is under testing while the archive function still requires a review of the common metadata to be used. Current specifications allow only harvesting of specimen catalog number.

The core of the concept of Herbadrop is to harvest metadata from OCR analysis from text written in herbarium images. The choice has been to proceed to a full text analysis using a Lucene engine Elastic search. The objective of this approach is to provide a powerful interface for further data curation as part of the preservation process (identifying duplicates, or inducing new taxonomic relations, etc.).

Contacts

- Elspeth Haston, RBGE, e.haston(at)rbge.org.uk
- Simon Chagnoux, MNHN, chagnoux(at)mnhn.fr
- Pascal Dugenie, CINES, dugenie(at)cines.fr

Further Information

Scientific Challenges behind the Pilot
Learn more about the EUDAT/Herbadrop pilot Collaboration here

Read more

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.