Hespi

Herbarium specimen sheet pipeline

The world is facing a biodiversity crisis.

Ecosystems are under pressure and species are disappearing at alarming rates. Researchers urgently need data to understand what is being lost and how to protect what remains.

But much of the world's biodiversity data is locked away.

For centuries, scientists and collectors have preserved plant and fungal specimens on herbarium sheets (Fig. 1).

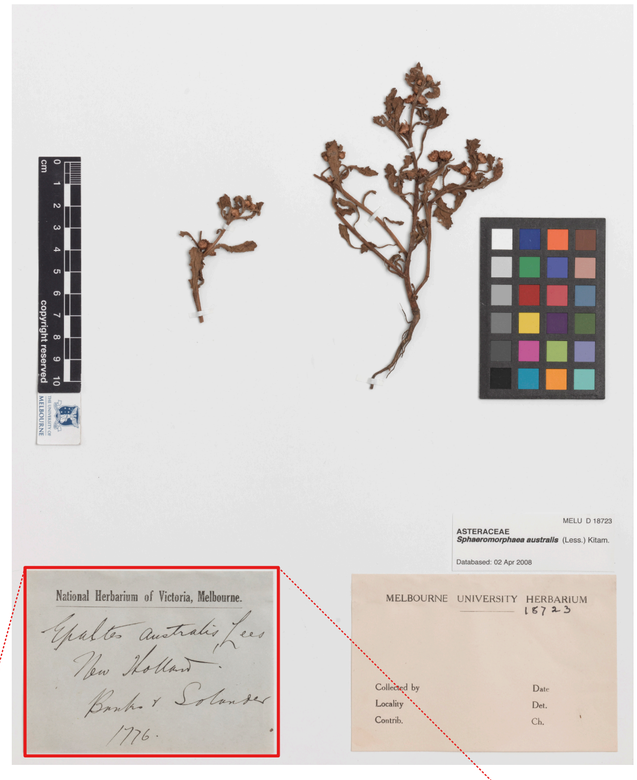

Fig. 1. The specimen sheet for spreading nut-heads (Epaltes australis), collected by Joseph Banks and Daniel Solander in 1770. Now at the University of Melbourne Herbarium (MELU).

NB. the collection date was historically incorrectly written as 1776 on the specimen label).

There are an estimated 395 million specimens held in over 4,000 herbaria around the world (Fig. 2).

They are a treasure trove of biodiversity data.

For researchers to access this information, these specimen sheets need to be photographed and the relevant text entered into fields in a database.

At the present rate, it will take decades for all specimens in the world's herbaria to be fully digitised.

Fig. 2. A map of the world's herbaria: https://sweetgum.nybg.org/science/ih/map/

We need to rapidly accelerate this process.

Our software Hespi rapidly extracts text from specimen images using multiple AI models.

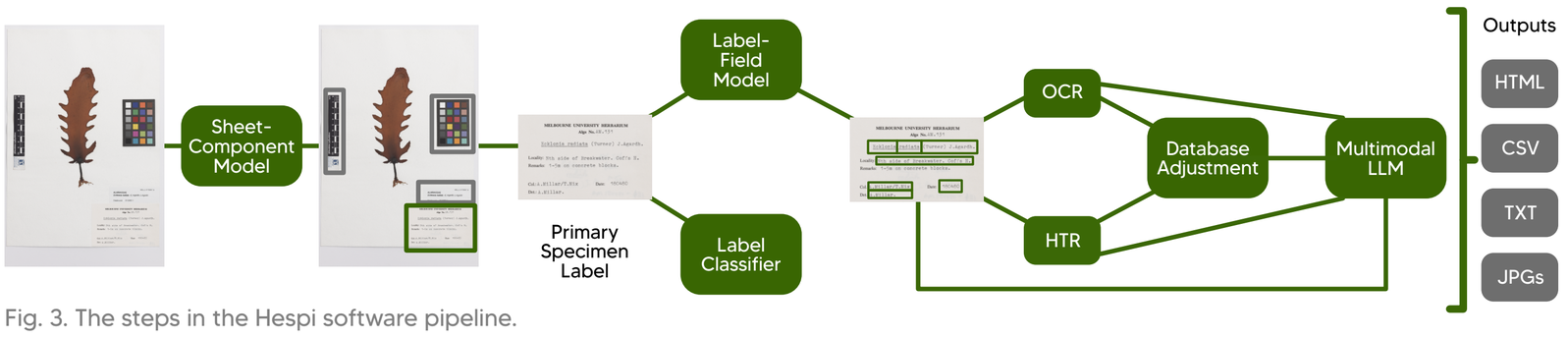

Fig. 3. The steps in the Hespi software pipeline.

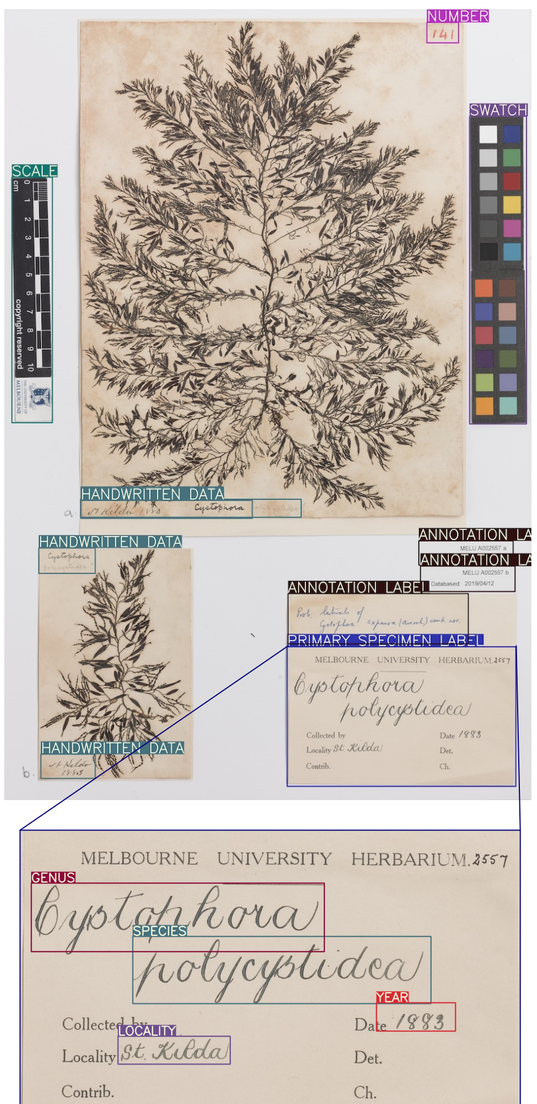

Fig. 4. Components of a specimen sheet of large brown algae (MELUA002557a) at MELU.

Step 1

Hespi finds the important components and labels using a YOLO object detection model (Fig. 4).

This model finds the primary label with a mAP50 score of 99.5%.

Step 2

On the primary label, it locates the important text fields (Fig. 5).

This model finds these fields with an F1 score of 91.1%.

Step 3

It uses OCR, handwritten text recognition, database lookup tables and Large Language Models (LLMs) to read the text in the fields (Fig. 6).

Step 4

It outputs the extracted information into an HTML report (Fig. 7) and a CSV to import into the herbarium database.

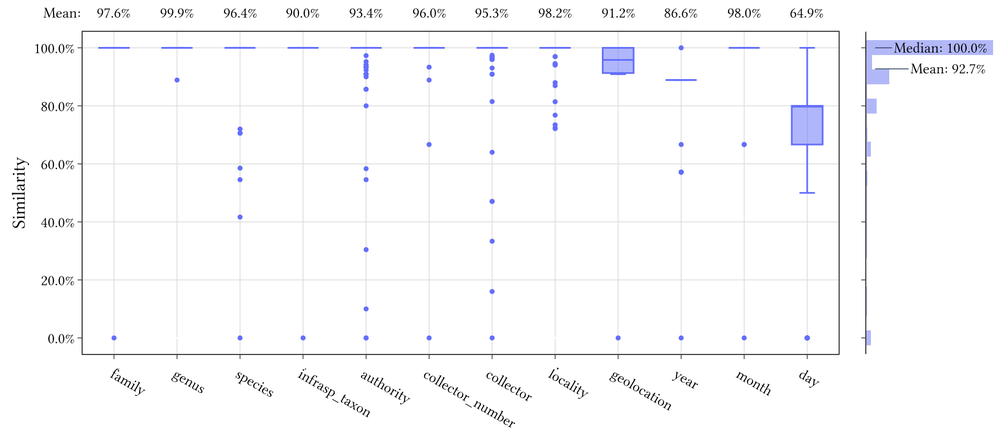

Fig. 6. Results from Hespi for labels which were printed or typewritten at MELU on the test dataset showing a mean accuracy of 92.7% and a median accuracy of 100% showing that most text fields were entirely correct.

Fig. 5. The text fields identified by Hespi on the primary label.

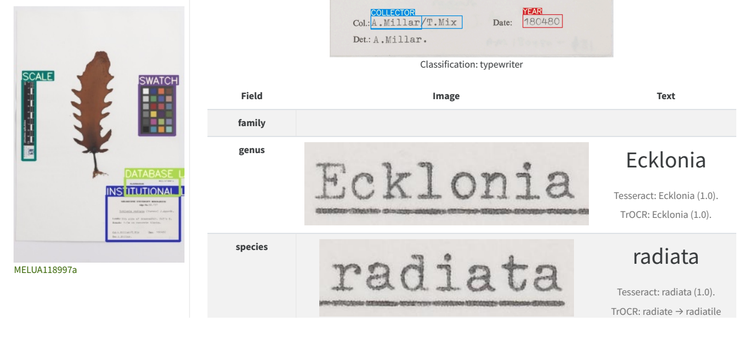

Fig. 7. The HTML report generated by Hespi.

Robert Turnbull[1], Emily Fitzgerald[1], Karen Thompson[1], Joanne Birch[2]

[1] Melbourne Data Analytics Platform, The University of Melbourne

[2] School of BioSciences at the University of Melbourne

'Hespi: a pipeline for automatically detecting information from herbarium specimen sheets'

BioScience, Aug 2025

DOI: 10.1093/biosci/biaf042

https://bit.ly/hespi2025

Also read the article in The Conversation:

'Botanical time machines'

https://bit.ly/3Hydpyh