Envisaging a global infrastructure to exploit the potential of digitised collections

Quentin Groom

Mathias Dillen

Wouter Addink

Arturo H. Ariño

Christian Bölling

Pierre Bonnet

Lorenzo Cecchi

Elizabeth R. Ellwood

Rui Figueira

Pierre-Yves Gagnier

Olwen Grace

Anton Güntsch

Helen Hardy

Pieter Huybrechts

Roger Hyam

Alexis Joly

Isabel Larridon

Vamsi Krishna Kommineni

Laurence Livermore

Ricardo Jorge Lopes

Jeremy Miller

Sofie Meeus

Kenzo Milleville

Marc Pignal

Renato Panda

Jorrit H. Poelen

Blagoj Ristevski

Tim Robertson

Cristina Rufino

Joaquim Santos

Maarten Schermer

Katja Seltmann

Ben Scott

Heliana Teixeira

Maarten Trekels

Jitendra Gaikwad

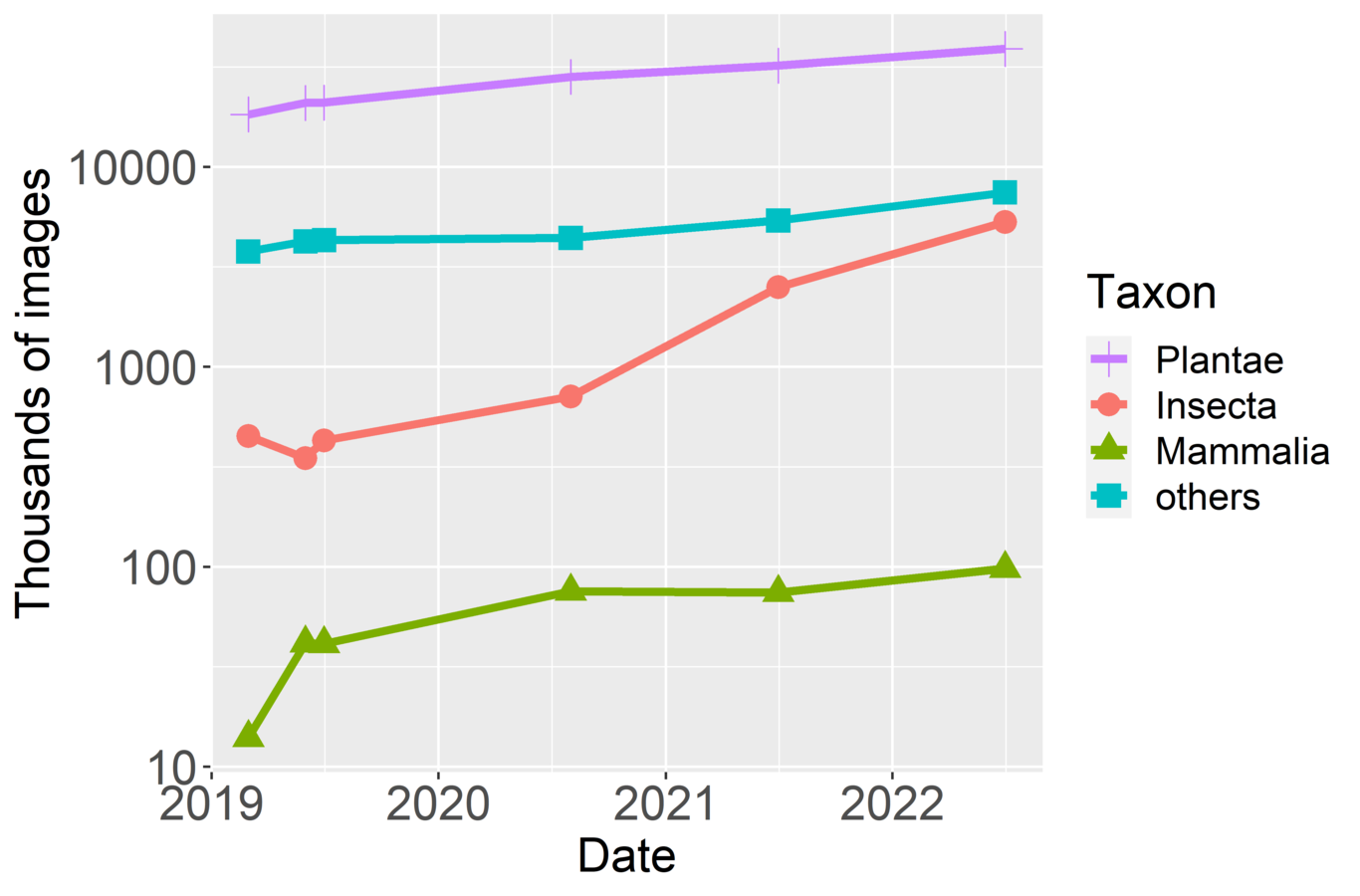
# Abstract

Tens of millions of images from biological collections have become available online in the last two decades. In parallel, there has been a dramatic increase in the capabilities of image analysis technologies, especially those involving machine learning and computer vision. Whilst image analysis has become mainstream in consumer applications, it is still only used on an artisanal basis in the biological collections community, largely because the image corpora are dispersed. Yet, there is massive untapped potential for novel applications and research if the images of collection objects could be made accessible as a single corpus. In this paper, we make the case for building infrastructure that could support image analysis of collection objects. We show that such an infrastructure is entirely feasible and well worth the investment.

# Introduction

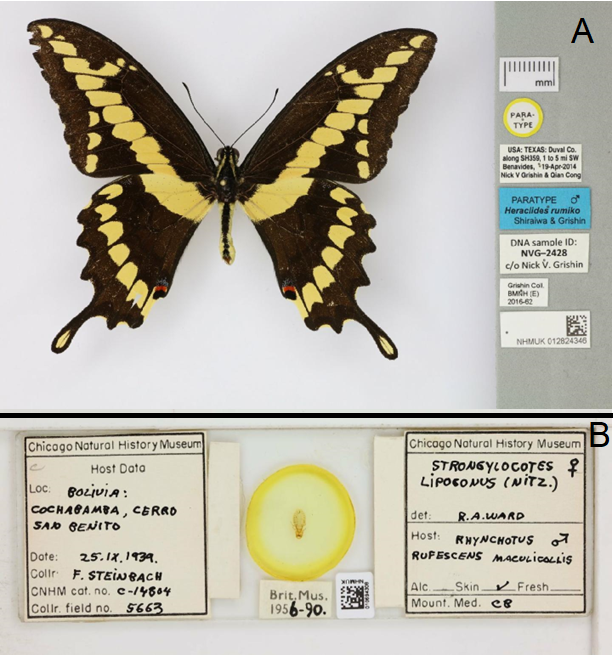
Owing to their central role in cataloguing the world’s biodiversity, global biological collections likely hold samples of most known macro-biodiversity. As such, they are an irreplaceable asset for research of all kinds, including ecology, conservation, natural history and epidemiology  (Bradley et al., 2014; Cook et al., 2014; Davis et al., 2019; Antonelli et al., 2020). They are also seen as an important and underused resource to address numerous questions in the context of biodiversity under global change (Soltis, 2017; Meineke et al., 2018; Hussein et al., 2022). Thus ensuring access to, and integrating data from these collections is globally important for the future. Conservation and sustainable use of biodiversity are fundamental to the 2030 Agenda of the (Secretariat of the Convention on Biological Diversity, 2016) and achieving its sustainable development goals is only realistic with the collections that underpin accurate naming and knowledge of biodiversity.

To keep pace with the demand for access to collections, digital imaging of biological collections has progressed at pace (Fig. 1). As of September 2022, the Global Biodiversity Information Facility (GBIF) has more than 49 million preserved or fossil specimens with an image. For just the nearly 400 million specimens of plants held in collections globally (Thiers, 2020), there are almost 38 million (9%) occurrences with images on GBIF. This number is expected to grow substantially. For example, the digitisation of the Kew herbarium, which holds over 7 million specimens will add to already major digitization programs in Australia, China, Europe and the United States among others (Willis et al., 2018; Nelson and Ellis, 2018; Borsch et al., 2020; Chinese Virtual Herbarium, 2021).??????



Progress in digitising natural history collections held by museums and herbaria. Many of these are accessible from either the Global Biodiversity Information Facility, iDigBio or BioCaSE. To examine the rate and volume of these specimen images, we used six snapshots of these databases taken since 2019. This was achieved using Preston, a biodiversity dataset tracker (Elliott et al., 2022; Poelen, 2022; Poelen and Groom, 2022). Although likely to be an underestimate of the specimen images that exist, because not all images are linked to the snapshot datasets, trends give an indication of progress towards digitising the world’s natural history collections. The number of available images is increasing approximately exponentially. There are seven times more plant specimens than insects in our most recent snapshot, even though insects are far more numerous in nature, an estimated 5.5 million species of insects (Stork 2018) vs 350,000 plants (Cheek et al., 2020). Nevertheless, the rate of increase of insect images is faster and if one extrapolates the curves it is easy to imagine that insect images will surpass plant specimens in a few years. Imaging of far less numerous but well-represented taxa in natural history collections, such as mammalia (~6,400 species; (Burgin et al., 2018)), while increasing, is not doing so as rapidly as insects.

With so many digital images coming online, it is not surprising that computer vision techniques are now being applied to specimen images. These techniques include segmentation (Allan et al., 2019; Goëau et al., 2020; White et al., 2020; de la Hidalga et al., 2022), object detection (Champ et al., 2020; Ott et al., 2020; Triki et al., 2020) and object recognition specifically to identify taxa (Carranza-Rojas et al., 2017; Earl et al., 2019; Valan et al., 2019; Little et al., 2020). In recent years, machine learning in particular has become mainstream and has been built into workflows that start with digital images and their metadata and result in statements about the image and what it illustrates. Such workflows can be used to extract information about a biological specimen from the typed or handwritten text on them (Allan et al., 2019). Yet, there are many other uses for image analysis of specimens as we elaborate on below (Pearson et al., 2020; Soltis et al., 2020).



Zoological specimen labels contain rich secondary data. (a) Paratype of Heraclides rumiko, showing information encoded on multiple labels. Catalogue number NHMUK012824346 by The Trustees of the Natural History Museum, London (CC-By) (b) Specimen of a chewing lice (Philopteridae): Strongylocotes lipogonus, a parasitic species including host information on the label. Catalogue number NHMUK010694309 by The Trustees of the Natural History Museum, London (CC-By).

Improving online access to collections is important because collections are physically dispersed, yet still interconnected through their origins and exchanges (Nicolson et al., 2018). Researchers are rarely able to obtain a full set of specimens for a single taxon or single collector from a single institution. Most are scattered across tens or even hundreds of different collections. Digitization and digital access can break down physical barriers between collections and make them accessible as a unified research tool (Hardisty et al., 2020). Online access is also fundamental to addressing historic imbalances in the amassing of collections in the northern hemisphere, from high-biodiversity regions (Grace et al., 2021).

Unified access to specimen images is particularly important because image files are comparatively large and image analysis pipelines are therefore demanding on processor time. Furthermore, current internet bandwidth makes transferring large numbers of files a bottleneck, particularly if those files need to be moved multiple times. Therefore, it makes sense to store large numbers of images close to where the processing is going to occur. While such infrastructures exist for other data types (e.g., *Copernicus*for remote sensing and *WLCG* for the Large Hadron Collider), no such support exists for biological collections-based image processing. All image analysis pipelines have so far built their own corpus of images and processed them independently. This approach is not scalable, it is wasteful of human time and effort, not to mention the internet bandwidth that would be required to do this on the scale of global collections. It is also unsuitable for dynamic image corpora and workflows that are intended to be run multiple times.

### The Vision

We envisage a data space for biological collections with a centrally accessible image corpus with built-in processing. This will allow anyone to access digitised images of specimens, without having to concentrate on the logistics of corpus creation and maintenance. By building accessible interfaces, it would also make it possible to remove technological barriers that prevent taxonomists and ecologists, among other users, from using advanced image analysis tools. Through supervised expert contributions the system could be further advanced with the integration of knowledge from many scientific disciplines. Such a corpus would be constantly furnished with new images from publishing collections and would support both the citation and reproducibility of the workflows, and their underlying collections, in alignment with FAIR Data Principles (Wilkinson et al., 2016). It would also make it easier to curate image datasets and use them for research (e.g. for benchmarking and challenges for machine learning) and for activities like teaching species identification from digitised specimens.

## Scope

With such an infrastructure, we aim to increase the use and improve the usability of biological collections for research. The initial focus would be to support two-dimensional images from preserved, fossilised or geological specimens. This could later be extended to other types of specimen image, such as to three-dimensional images and X-ray computed tomography (CT scanning).

Images from living organisms are not considered here, nor other media, such as sounds, though they are undoubtedly useful and deserve attention. However, the challenges of pictures of living organisms are different, their numbers are at least two orders of magnitude larger and increasing more rapidly than digitised preserved specimens and dedicated infrastructures already exist to process them, such as Pl@ntnet and iNaturalist. The creators of such images are also more varied, as are the licensing requirements placed upon images. An exception might be those pictures of living organisms *in situ* before they were preserved. Such pictures give additional context to the specimen and can potentially be used together with the preserved specimen both for human and computational comparison (Goëau et al., 2021).

In this paper, we answer the questions: what research could be done with such an infrastructure, who would use it, what functionality would be needed and what are the architectural requirements? First, we present the purposes for such a unified corpus of specimen images, and secondly we envisage what such an infrastructure might look like. In total we imagine a future where we can search across global collections for such things as the pattern of a butterfly’s wing, the shape of a leaf, the logo of a specific collection, or for examples of someone’s handwriting.



Labels of specimens from Meise Botanic Garden contain secondary data features, such as the handwriting, the colour of the inks, shape of the label and printed label decorations. (A) Specimen labelled Potentilla recta collected by amateur botanist Camille Francotte with distinctive label decorations (<https://www.botanicalcollections.be/specimen/BR0000009398214>) CC-By-SA (B) Specimen labelled Eriophorum angustifolium where the collector can be recognised from their signature (<https://www.botanicalcollections.be/specimen/BR0000005134137>) CC-By-SA (C) Specimen labelled Agathosma villosum showing distinct cup-shaped label (<https://www.botanicalcollections.be/#/en/specimen/BR0000015671271>) CC-By-SA (D) Specimen label of Alyssum calycinum collected by the former director of Meise Botanic Garden François Crépin, notorious for his hard-to-read but easily recognised handwriting; also showing his signature (<https://www.botanicalcollections.be/specimen/BR0000010426135>) CC-By-SA

# Purposes

Infrastructure costs money, it needs to justify the cost through the benefits, not just for science, but society in general. We also need to understand who the users will be and the beneficiaries. Below we have outlined below some uses and users for an imaging infrastructure for collections. Though there are undoubtedly uses we have yet to imagine.

## Species identification

Machine learning applications for the identification of organisms mostly use digitised photographs of living organisms (e.g. (Wäldchen and Mäder, 2018; Bonnet et al., 2020)).

Most experiments with species identification from digitised specimen images have focused on herbarium specimens (e.g. (Carranza-Rojas et al., 2017; Kho et al., 2017; Pryer et al., 2020; Hussein et al., 2022)). This is because they are two-dimensional, they follow a fairly standardised format and are highly available. Herbaria have preceded the digitization of animal collections that tend to be more three-dimensional and, in the case of insects, are much larger (Fig. 1). Nevertheless, because insect specimens are, in general, much more abundant there is clear demand for automated identification of these specimens too (Earl et al., 2019; Valan et al., 2019; Høye et al., 2021). A clear advantage of insects is that their colour and morphology are well preserved in specimens. This means that automatic identification trained on specimens may work on living specimens and vice versa, having the possibility to create training datasets for rarely seen organisms (Goëau et al., 2021). Specimens from natural history collections have been used successfully to train models that assist in sorting images from camera traps, thus greatly facilitating the monitoring process (Høye et al., 2021). Camera traps are routinely deployed in ecological monitoring and have been advocated as a method of global biodiversity monitoring (Wearn and Glover-Kapfer, 2019).

Similarly to insects, the state of preservation, uniformity and distinctiveness of pollen grains also makes them good targets for automated identification whether they are from preserved collections or fresh. Indeed, pollen is well preserved as fossils and sub-fossils making them useful targets to analyse evolutionary and ecological change (Romero et al., 2020; Hornick et al., 2022). Manual identification of pollen grains by experts is slow and laborious, which machine learning could transform into a much more routine process (Bourel et al., 2020), with potential applications in environmental monitoring, archaeology and forensics.

Work on ichthyological collections (Elhamod et al., 2022) has demonstrated that the addition of phylogenetic information can strengthen neural network models and improve the identification of specimen images. This Hierarchy-Guided Neural Network allows for imperfect, yet realistic scenarios such as damaged specimens or limited training data while incorporating and potentially improving knowledge of taxonomic relationships.

The main advantage of automated identification of digital images of preserved specimens is not the accuracy, but the potential for high throughput. Accessing large numbers of images in a suitable computational environment remains a critical factor to mainstreaming automatic specimen identification across collections.

## Extracting trait data

Morphological, phenological and colourimetric traits are often clearly visible on images of specimens (e.g. Fig ???A). Such traits might be diagnostic for the identification of the organism, but they are also used to understand how traits evolve and what the traits tell us about the evolutionary history of a taxon.

### Functional traits

Traits are interesting from the perspective of the functions they have evolved to perform. Morphological functional traits have been used to predict impacts of climate change on ecosystem functioning (Pigot et al., 2020), species distributions (e.g. (Pollock et al., 2012; Regos et al., 2019)), community structure (Li et al., 2015), and how these traits fit into the land surface component of climate models (Kala et al., 2016). Functional traits recorded from preserved specimens supplement field recorded data filling geographic and temporal gaps and providing legacy data (Heberling and Isaac, 2017; Bauters et al., 2020; Kommineni et al., 2021), as well as potentially enabling discovery of newly-relevant morphological traits. Examining such traits in preserved specimens in collections is also considerably cheaper than fieldwork.

Leaf morphological traits are particularly amenable to extraction from herbarium sheets, both because they are laid flat and because they do not necessarily require magnification (Heberling, 2022). Their size, dimensions, arrangement, dentation and venation are all possible targets for machine learning and experiments with extracting these parameters have shown it to be feasible and reliable (Heberling and Isaac, 2017; Triki et al., 2020; Weaver et al., 2020). The extraction of traits from collections of insects has great potential, particularly as the state of preservation of insects in collections is high (Høye et al., 2021).

In the case of fish, due to the large number of species globally, the enormous number of morphological traits and large amount of variation, we can only hope to fill the gaps in our knowledge of traits if preserved specimens are used (Hay et al., 2020; Kattge et al., 2020). Furthermore, specimens have the advantage that there is a voucher where the measurements can be verified and new measurements can be taken.

Using well-documented machine learning algorithms for extracting traits from specimens would mean much greater efficiency if a single large corpus were available for analysis, but also measurements could be less prone to error and more reproducible if the source code and training data are open and shared (Meeus et al., 2020). Digital specimens share similar pitfalls as their physical counterparts, such as missing metadata from specimen labels.

Further, collection practices have changed considerably over the more than four centuries they have been amassed (Kozlov et al., 2021). Also, characters of specimens can change upon preservation, for instance, shrinkage associated with drying (Tomaszewski and Górzkowska, 2016). Yet, with suitable awareness and controls, there is much to be learned from trait data gathered from digital specimens.

### Phenology

A trait of particular interest for climate change impact studies is phenology. Changes in seasonal temperatures and rainfall patterns affect the hatching or emergence of dormant animals and the maturation of leaves, flowers and fruits. Such changes may lead to a mismatch in seasonality among organisms (Renner and Zohner, 2018). Detecting the phenological state of an organism is possible through machine learning (Lorieul et al., 2019; Davis et al., 2020; Triki et al., 2021; Goëau et al., 2022) though not to the level of accuracy achieved manually. Nevertheless, the obvious advantage of machine learning is the potential for high throughput processing of images to track phenological shifts (Pearson et al., 2020).

### Colour analysis

Imaging of specimens is almost always done with colour cameras, even though some organisms, such as plants, change colour when they are preserved (Davis et al., 2013). Nevertheless, there are animals, such as insects and birds, that maintain their colour well and may be interesting targets for research (Hoyal Cuthill et al., 2019). Among other avenues of research, studies have shown that colour is an important factor in climate change adaptation of insects (Zeuss et al., 2014).

### Species interactions

Organisms are in constant conflict with their predators, parasites and pathogens. Specimens provide a record of this and have been shown to reveal long-term changes related to environmental change, such as the introduction of non-native species (Vega et al., 2019), pollution and climate change (Lang et al., 2019). For example, manually extracted changes in leaf herbivory of herbarium specimens were correlated with climate change and urbanisation in the north-east of the United States of America (Meineke et al., 2019). Indeed, (Meineke et al., 2020) took this further and investigated the potential for extracting leaf damage data from herbarium specimens of two species through a process of detection and classification of images split into grid cells. Although in this instance image analysis was less accurate than human classification, the possibility remains of applying such an analysis to many more species over a much larger geographic area that would be possible only with automation using images from multiple collections.

## Collections care, curation and management

The preceding use cases extract data from specimens for research, but information is also needed for curation, organisation, storage and management of collections. A pertinent example is the need to identify specimens treated with toxic substances, suc as mercuric chloride used historically on herbarium specimens to prevent insect damage. Over time, mercuric chloride leaves stains on the specimen mounting paper that image classification can be used to distinguish. (Schuettpelz et al., 2017) used a convolutional neural network trained to detect such stained sheets. It has a false-negative rate of 8%, which is comparatively high error for a situation related to toxicity, yet could likely be improved, particularly if provenance information is combined. Similarly, one might use a similar approach to detect pests in collections, such as *Lasioderma serricorne* (J.Fabr., 1792).

One could even imagine image analysis workflows that detect the type of mounting strategy and preservation state of the specimen. This would help curators triage specimens that need remounting or some other form of curational care.

Numerous other checks and controls can be performed on images of specimens. For example, quality control of the images themselves, such as lighting, colour, cropping, orientation and focus. Additionally, the presence and accuracy of image components, such as the barcode, ruler and colour chart, can be verified and is a useful check to the integrity of a corpus.

## Visual features of the specimen

### Image segmentation and object separation

Image segmentation is a fundamental low-level image processing task to facilitate higher-level tasks such as object detection and recognition (de la Hidalga et al., 2022). In preparation for image analysis, such as searching for signatures, or to support further digitisation with a human-in-the-loop, it is often more efficient to recognise the individual objects in an image, classify them and separate them into multiple images. For example, if the image contains multiple specimens, or if a label needs to be extracted from the image to present to humans for transcribing or to do further image analysis on. Biological specimens from different collections representing the same species often show a large variety in backgrounds, caused by different mounting techniques and different digitisation processes. Separating the object and repositioning it in preparation for further image analysis may help in establishing training sets that ignore the differences in background and positioning.

In an infrastructure built for image analysis, standard segmentation workflows could be run and optimised to avoid every researcher repeating the segmentation step and users of the infrastructure could choose themselves whether they want to analyse the whole image, all the segments, or specific classes of segment.

### Labels

Specimens are usually annotated with information on their labels. In the case of plants, these labels are on the mounting paper, for insects they are on the mounting pin, while for larger zoological and plant specimens labels might be tied to the specimen, or on, or in, specimen jars. These labels document characteristic data from the collecting event (Fig. ??? & ???). Therefore, as images of mounted specimens often contain text it is useful to provide printed and handwritten text recognition output as part of an image processing pipeline. If this text can be recognized, this additional metadata can be used to enrich the items of the collection and automatically perform cross-collection linking. Furthermore, the recognized text can aid in the digitization process and validation of the extracted metadata, reducing the amount of manual input required and improving the quality of the data being transcribed (Drinkwater et al., 2014).

Although state-of-the-art text recognition systems perform well on printed text, accurately recognizing handwritten text is still a challenge. Older handwritten text might contain unique writing styles and may have deteriorated. Nevertheless, such cases can still provide valuable information relating to the writing style. Text written by the same author could be automatically clustered based on visual similarity and used to identify the collection and reduce manual validation.

Besides text, secondary data hidden in the handwriting, ink colour, mounting paper, label shape and printed label decorations (Fig. ???, ??? & ???) can be used to determine their origins and history. Image analysis by itself can be enough to make clusters of specimens for particular purposes, for example, a group of specimens from a particular expedition. These clusters can also help to do further image analysis on images that share some common characteristics.

### Rulers and colour checkers

Another element often seen on digitised images of collection objects are rulers, scale bars and colour checkers. These come in many different types and sizes, as different institutions often customise them based on the requirements of the imaging campaign. Colour checkers are used to validate the fidelity of the colours of the specimen image, while a ruler provides a reference to the actual size of the specimen with regards to the image size. Especially when digitising with a digital camera, it can be complex to calculate the actual dimensions of the image, as it depends on the camera lens and individual camera parameters. As it is time-consuming to measure each specimen manually, specimen dimensions are often not included as metadata. Therefore, the detection of rulers and colour checkers on digital images can prove useful to estimate the actual specimen size and correct colour balance. A generic object detection or instance segmentation model can be trained to detect these common objects. If all the rulers in a collection are of a fixed size, the length of the detected ruler can be used to calculate a transformation from pixels to the ruler’s unit of measurement (e.g. cm, mm). This transformation can then be combined with specimen segmentation models, to automatically extract the dimensions and specimen traits (Triki et al., 2021). However, when rulers are not of a uniform size, the distance transformation needs to be estimated by calculating the pixel distance between the measurement stripes or bars on the ruler (Bhalerao and Reynolds, 2014). To extract the specific unit of measurement, the text denoting the unit on the ruler can be recognized or additional metadata about the specimen can be used to infer it.

### Finding stamps and signatures

Specimens are often labelled with rubber stamps and occasionally printed or embossed with crests that indicate provenance or ownership (Fig. ???). For instance, the stamps of botanical exchange clubs (Fig. ???C, ???E), which operated in Europe, and particularly the United Kingdom, from the middle of the 19th century into the 1930s (Groom et al., 2014). Tens of thousands of specimens were exchanged this way and found their way into collections around the world. If a specimen was part of a botanical exchange club, it implies that duplicates of this specimen existed and it circumscribes the dates within which a specimen was collected. Although stamps usually contain some text, they are often circular or oval, making them intractable to standard OCR engines.

Embossed crests and stamps on herbarium specimens provide a rich secondary data source. Secondary data like this forms an important part of the provenance of specimens that is often not directly transcribed. (A) Lion and crown signifying ownership by the botanical garden of Brussels, as found on specimen BR0000013433048 of Meise Botanic Garden Herbarium. (F) Stamp with handwriting is evidence of a loan from the national botanical garden of Belgium to the herbarium of Paris (Herbarium Musei Parisiensis, P), on specimen BR0000017682725 of Meise Botanic Garden Herbarium. (B) Stamp of the A. C. Moore Herbarium at the University of South Carolina as on specimen USCH0030719 (image in public domain) (G) printed crest on label of specimen P00605317 held by Museum National d’Histoire Naturelle (licensed under http://creativecommons.org/licenses/by/4.0/). Stamps used by botanical exchange clubs connect thousands of specimens across collections in the world, such as this stamp (C) of the Watson Botanical Exchange Club on specimen E00809288 of the Royal Botanic Garden Edinburgh Herbarium (image in public domain) and (E) Stamp of the Botanical Exchange Club of the British Isles on specimen E00919066 of the Royal Botanic Garden Edinburgh Herbarium (image in public domain). Stamps are often unique to a time period and specific institution such as (H) designating the collections the specimen belonged to on specimen LISC036829 held by the LISC herbarium of the Instituto de Investigação Científica Tropical. The meaning of stamps is sometimes lost, such as the stamped star (J) with unknown meaning on the same specimen as (B), if meaning is derived for a single specimen, this information can be linked to all specimens with the same stamp. The meaning of a stamp is not always immediately clear, such as (K) a stamp designating the specimen belonging to the Herbarium I. Thériot, on specimen PC0702930 in the cryptogamy collection (PC) at the Herbarium of the Muséum National d’Histoire Naturelle. (licensed under CC-By), whereas others convey a clear meaning, such as (L) a stamp designating the specimen was once in the collection of the Universidad Estatal Amazónica, now housed in the Missouri Botanical Garden herbarium under catalogue number 101178648 (licensed under CC-By-SA). Stamps, imprints and embossings change over time, such as this example (D) from the A. C. Moore Herbarium of the University of South Carolina on specimen number USCH0061250 (image in public domain), a practice carried over from (B). (l) is another more classical example of a crest used by a natural history collection, in this case the Muséum National d’Histoire Naturelle (MNHN - Paris), as found on specimen with catalogue number PC0702930. (licensed under CC-By)   

Embossed crests and stamps on herbarium specimens provide a rich secondary data source. Secondary data like this forms an important part of the provenance of specimens that is often not directly transcribed. (A) Lion and crown signifying ownership by the botanical garden of Brussels, as found on specimen BR0000013433048 of Meise Botanic Garden Herbarium. (F) Stamp with handwriting is evidence of a loan from the national botanical garden of Belgium to the herbarium of Paris (Herbarium Musei Parisiensis, P), on specimen BR0000017682725 of Meise Botanic Garden Herbarium. (B) Stamp of the A. C. Moore Herbarium at the University of South Carolina as on specimen USCH0030719 (image in public domain) (G) printed crest on label of specimen P00605317 held by Museum National d’Histoire Naturelle (licensed under <http://creativecommons.org/licenses/by/4.0/>). Stamps used by botanical exchange clubs connect thousands of specimens across collections in the world, such as this stamp (C) of the Watson Botanical Exchange Club on specimen E00809288 of the Royal Botanic Garden Edinburgh Herbarium (image in public domain) and (E) Stamp of the Botanical Exchange Club of the British Isles on specimen E00919066 of the Royal Botanic Garden Edinburgh Herbarium (image in public domain). Stamps are often unique to a time period and specific institution such as (H) designating the collections the specimen belonged to on specimen LISC036829 held by the LISC herbarium of the Instituto de Investigação Científica Tropical. The meaning of stamps is sometimes lost, such as the stamped star (J) with unknown meaning on the same specimen as (B), if meaning is derived for a single specimen, this information can be linked to all specimens with the same stamp. The meaning of a stamp is not always immediately clear, such as (K) a stamp designating the specimen belonging to the Herbarium I. Thériot, on specimen PC0702930 in the cryptogamy collection (PC) at the Herbarium of the Muséum National d’Histoire Naturelle. (licensed under CC-By), whereas others convey a clear meaning, such as (L) a stamp designating the specimen was once in the collection of the Universidad Estatal Amazónica, now housed in the Missouri Botanical Garden herbarium under catalogue number 101178648 (licensed under CC-By-SA). Stamps, imprints and embossings change over time, such as this example (D) from the A. C. Moore Herbarium of the University of South Carolina on specimen number USCH0061250 (image in public domain), a practice carried over from (B). (l) is another more classical example of a crest used by a natural history collection, in this case the Muséum National d’Histoire Naturelle (MNHN - Paris), as found on specimen with catalogue number PC0702930. (licensed under CC-By)

Many specimens are signed, either by their collector, determiner, or both (Fig. ???). Expert curators within an institution learn to recognise the signature of prolific collectors, but they are often illegible without that knowledge. Yet, it is common practice to use the name of a collector, together with their collecting number to identify a gathering (collection event) uniquely. Furthermore, due to exchanges, loans and gifts, a collector’s specimens may be spread between a number of institutions. If the name is not distinct enough to be transcribed accurately, finding the specimens from a specific collector across the whole corpus of global collections would be an impossible task without some automated process.

## Unsupervised learning

The stacked layers of deep neural networks can be regarded as a set of transformations that learn useful representations of the starting data. Using representations of specimen images learned by neural networks, rather than extracted metadata, would allow content-based interaction with, and comparison between, images. Such interaction is useful for tasks where a high-quality labelled dataset does not currently exist or where the characteristics of a specimen that are important to a task are not well-defined. For instance, (White et al., 2019) used representations of specimen images learned by a neural network trained to classify fern genera to directly compare specimen morphology and test biogeographic hypotheses. Similarly, (Hoyal Cuthill et al., 2019) trained a network to estimate the similarity of two sets of butterfly specimen images and used the learned representations to test mimicry hypotheses.

Furthermore, some tasks require researchers to inspect and compare specimen images individually. The reduced dimensionality of deep representations in combination with scalable nearest-neighbour search (e.g. (Johnson et al., 2021)) makes direct comparison of images very efficient, opening up opportunities to explore collections through image content rather than through the metadata. Tasks like searching a collection for similar specimens during identification, and identifying misidentified or poor-quality specimens, become much more efficient in a digital setting.

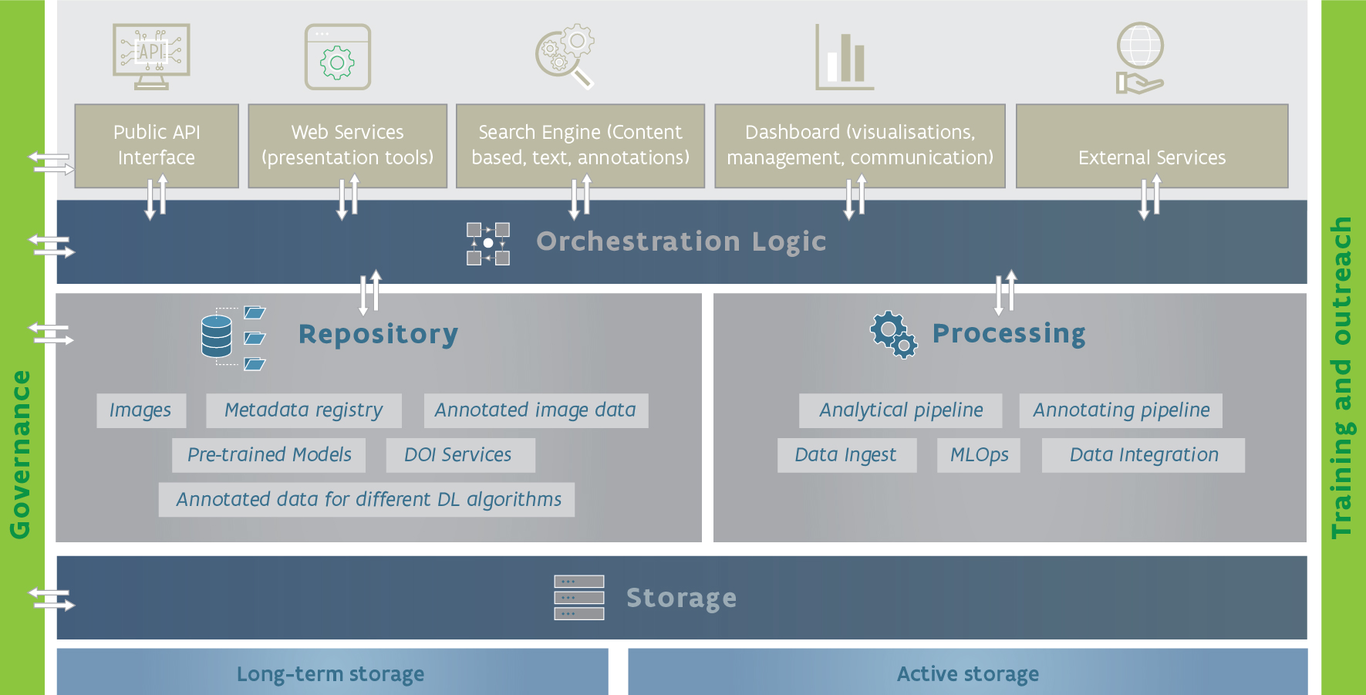
Recently, interest in learning useful representations from unlabelled data has surged (Rives et al., 2021) in the field of unsupervised (or self-supervised) representation learning. These studies have shown that large numbers of unlabelled images (millions to billions) can be used to learn representations that work well as a starting point for supervised classification tasks, such as species identification. There is exciting potential to apply these methods to herbarium specimens (Walker et al., 2022). A large, centralised repository of specimen images would further enable this research by allowing the development and curation of the two types of dataset necessary for self-supervised representation learning: large training corpora and smaller, task-specific benchmarking datasets (Van Horn et al., 2021).

# Conceptual framework of the infrastructure

Unlocking the potential for machine learning in natural history collections is contingent on technical infrastructure which is easy-to-use, interoperable with regional and global biodiversity data platforms, and accessible to the global scientific community. To build the infrastructure will require extensive consultations with the scientific community and funding agencies. It is imperative that investments in the infrastructure are scientifically, economically and socially justified, as well as sustainable. Here, we present a conceptual framework which is conceived as a roadmap for building the envisioned infrastructure. Although the infrastructure could be implemented in different ways (e.g. distributed or centralised) with advanced components depending on the scope and requirements of the research community, there are essential components that form the foundation of this proposed infrastructure. In the following section, we describe the three core technical components of the infrastructure, coordinated by the orchestration logic: (1) the repository to index data and metadata, (2) the storage of images, models and data and (3) the processing of images to generate new data and annotations, as well as train new models (Fig. ???). The orchestration logic will consist of components such as technical workflows, security protocols and application integrations that enable implementation of business logic and access to services. In addition to the technical components, the infrastructure will require a governance structure and set of protocols, as well as training and outreach to reach the intended audience.

## Component 1: The Repository

A dedicated repository is needed which will reference and index information such as specimen metadata, image metadata and annotations alongside machine learning models with their performance metrics and outputs (Fig. ???). Some infrastructures already exist, or are in development, to accommodate some of these data types, such as GBIF for specimen data, but none integrate the full spectrum of specimens, images, models and model outputs. These existing infrastructures can be reused, either by integrating or connecting with the repository or becoming it by extending their own capabilities. The repository should operate on the FAIR principles, facilitating data discovery and reuse. This includes the support for, or provision of, persistent identifiers for the different types of content, as well as different data standards.



Framework of an infrastructure for analysis of specimen images showing the services, storage and relationships between them

The image metadata in the repository will include a reference to the image object located in the storage layer (Component 2), along with annotated training image data. Different kinds of image annotations will be supported, including geometric-based regions of interest (ROI), taxonomic or ecological traits and textual representations of label data. For interoperability, data standards supporting the machine readability of these annotations are required. As different standards exist for these annotations and not all are equally suitable for any model, the platform should ensure support for multiple standards, such as *COCO* (JSON), *Pascal VOC* (XML) and image masks (rasterized or vectorized images). Multiple annotations can be made on a single specimen record, making persistent identifiers for these specimen records vital. The metadata indexed in the repository will facilitate the findability of suitable annotations, for instance, to serve as training data. A feedback mechanism may be implemented to correct and/or update annotations.

The pre-trained machine learning models will be stored in the repository and made available for reuse, along with accuracy metrics and the model output, such as the segmented features or species metadata. To ensure findability, models should be classified by use-case through the use of keywords, since they are often trained for very specific use-cases but could later be reused in other contexts. As part of the metadata, suitability scores will facilitate comparison of models in terms of their efficacy, possibly through community feedback or by analytics scores that take standardised model performance metrics into account. These results should be linked to the original images used in the training of the model (on the platform) and also to the images that were analysed in the use case.

Persistent identifiers such as Digital Object Identifiers (DOIs) or hash-based content identification (e.g., *Software Heritage PIDs* for code or simple SHA-256 hashes for images) will be assigned to the digital objects produced during the use of the infrastructure to make them citable. It will also be possible to assign persistent identifiers to different versions, reflecting any subsequent updates the submitter makes to the digital objects. The repository will display citations of the persistent identifiers, including links to publications in which they are included, as well as any instances of their reuse in other projects within the repository. It is not only important to make the digital objects or outcomes openly available, but also under appropriate licences *(e.g., Creative Commons)* as indicated by the *FAIR for research software (FAIR4RS) working group* and (Labastida and Margoni, 2020).

Managed through the orchestration logic, the repository is connected to a storage system and the processing unit, while having features such as a content-based search engine to browse the content not only on the traditional humanly-annotated metadata, (e.g., date and place of observation, taxonomy, and others), but also on information extracted from the images themselves. Advanced features can be built into the system, such as the ability for users to upload an image and search the catalogue by similarity (e.g. similar handwritten signatures), or query and filter the collections of data using the indexed metadata extracted from the observations, either humanly or automatically annotated. In general terms, such functionality can be summarised as the ability to aggregate to each specimen media record entry all the information that is extracted from it either manually or automatically, and indexed making it available to query. Some good examples of similar content-based systems exist in production today. [*Pl@ntNet*](https://plantnet.org/en/) and [*iNaturalist*](https://www.inaturalist.org/home) provide species identification of organisms from photographs. Results can be refined by providing the user’s location, thus limiting the possible results to the most likely matches, boosting accuracy. A more general example is Google Image Search, where anyone can search images using either a keyword (e.g., dog), or using an image as the search term’. This function is also available on Google Photos (web or mobile), where a user can search their personal photos for specific people, different types of objects, places, ceremonies, events, and so on. Although different, all those systems share similar logic: (1) they include machine learning models trained for specific tasks (e.g. object detection) that have been created offline using massive datasets in large GPU clusters (e.g., TensorFlow [*Model Zoo*](https://modelzoo.co/) and [*COCO dataset*](https://cocodataset.org/); (2) when a new image is added to the collection (or possibly all, when new models are deployed), in addition to the submitted user tags, the images are processed with these models (inference/prediction pipeline) and tags are extracted; (3) the extracted information is saved and indexed, and made available as searchable data. The envisioned system should provide similar functionality, although with the added complexity of supporting a myriad of different models and images, as illustrated by the use cases listed in the previous section, such as searching for colour bars, rulers, the institutional stamp or a specific trait.

## Component 2: The Storage

The storage component (Fig. ???) encompasses all physical storage that is a local part of the platform, and on which images, models, metadata and results are stored. It also includes functions, managed via orchestration logic, required to manage that data as far as access control (e.g. governance) and low-level file management is concerned (such as back-ups). Higher level management, such as handling uploads, selection of specific images and the moving of images to processing, is the responsibility of other components. The storage component is divided into two areas, archive and regular (active) storage. This distinction is primarily a technical one, separating high-performance storage required for accessing images while training models, from less advanced storage for other purposes.

Whether images are mirrored from their original source onto the platform, or if they are only downloaded temporarily onto the platform when needed for training, is a technical design question that should be answered during implementation. While this choice has no functional impact, it does have profound technical implications, as well as budgetary consequences. Locally mirroring all images referenced in the repository guarantees availability and predictable speed of access, but will also require extensive management to accurately reflect changes made to the source material, and will take up an increasingly large storage volume. On the other hand, while downloading images on-the-fly greatly diminishes the required storage volume, it implies less control over availability and carries the risk of images becoming unavailable over time.

### Storage of training images

Images to use in training are discovered through the repository component, which functions as a central index of images, metadata, models and results. Actual image files might be hosted on the platform, or remotely, on servers of associated parties. In case of the latter, because of the technical requirements (i.e., high throughput, guaranteed availability, low latency), these images must be downloaded to the platform, and be made available locally to be used in the training of models. Selection of these images is done in the repository, and the orchestration logic functions as a broker between the repository and remote hosting facilities, taking care of downloading of images. The storage component is responsible for the local storage of these files. This includes facilitating access control (i.e. keeping track of what images belong with which training jobs), and making images available to the processing component, where the actual training takes place. In the scenario where the local storage of training images is temporary, the images will be deleted once the training cycle of a model has been completed, while only the references in the repository to those images are retained with the resulting model. The handling of images while stored in the system, including their accessibility and deletion policies, is subordinate to the platform’s governance policies.

### Storage of models

Once a model is deemed finished or suitable for use, it may be published as such in the repository, and thus become available for researchers to use. Again, the repository functions as a central index that allows researchers to find suitable models, while the actual code that makes up a model will be stored in the storage component. Once a model has been selected by a researcher for use (see also next section), it is retrieved from storage and copied to the processing component for use. A similar scenario applies when a stored model is used as the basis from which to further train a new model, or a new version of the same model (transfer learning). Since there are no specific performance requirements for storing a model, they will be stored in the archive section of the media storage component. Besides models that have been trained locally, the platform can also host and publish models that were trained elsewhere. From the point of view of storage, these models are treated as identical as ones trained locally. As with images, availability of and access to models stored on the platform is subject to governance policies.

### Storage of images for analysis

Another function of the processing component is using ‘finished’ machine learning models for the analysis of images, resulting in the annotation of newly uploaded images with or without metadata (such as classification or identified regions of interest). For this purpose, images will be uploaded by researchers, after having selected a model or models from the repository to run on the images. Uploaded images will be stored in the storage component, and kept there for the duration of the experiment running the selected models. Responsibility for running these experiments, including the loading and execution of the selected models, lies with the processing component. Actively making available the images to the models is facilitated by orchestration logic.

Once experiments have been completed, these images will be moved to a low-performance part of the media storage component (archive storage), where they are stored with the newly acquired metadata, in line with relevant governance policies. These archived images and their annotations are registered in the repository component, so as to make them findable by other researchers. If, at a later stage, someone wants to perform further image analysis on images that were analysed previously, these images can be moved back to the active storage area for further analysis.

The technical requirements for analysis processes are far less demanding than those of training processes, especially with regards to the need for constant high throughput. It is therefore conceivable that the platform will allow researchers to access stored models remotely through an API, in which case no images are stored locally.

### Storage of model results

Value for researchers is to be gained from access to results derived from the models on the platform. These results might be produced by analysis processes as described hitherto, or by use of a model remotely, either via API access or even by entirely running a model remotely. The form of these results can be manifold; besides previously mentioned examples such as classification or the identification of regions of interest, they can also include more generalised performance characteristics of a model, such as the average recall and precision for a given set of images in case of a classification experiment. Uploading such results, in whatever format they might take, and associating them with the models that generated them is the responsibility of the repository component, while the physical storage of data is taken care of by the storage component. Negotiation between the two components, both when storing and when retrieving, is performed by the orchestration logic. Again, all handling of these results follows the platform’s governance policies.

## Component 3: The Processing

The processing component encompasses all the services and pipelines whose focus is to compute tasks on batches of incoming data into infrastructure or already existing in the system, such as those stored in the repository and storage components (Fig. ???). In other words, it supports a myriad of computational-intensive tasks, from ingesting new data, to the automated extraction of information from media, as well as exporting new datasets or scheduling the training of new models or the retraining of old ones.

This component requires a considerable amount of computing power to handle all the scheduled tasks in the system, which can even be elastic (i.e. cloud principles) given the fluctuating demand. These are delegated by the orchestration logic component, a set of services that are responsible for handling the external requests, such as those from users through frontend applications, or other external services using one of the public APIs, serving as both gateway and manager to the main internal components – repository, storage, and processing (Fig. ???). The biggest computational demand comes from tasks related to the creation of machine learning and deep learning models periodically, updating the existing services or adding new ones. For these, specific hardware capabilities such as several GPU/TPU instances may be required from time to time.

The processing component, and the tasks and services supporting it, should be able to scale vertically, that is, handle more tasks by adding more RAM, more CPU cores or a better GPU to a cluster node, but preferentially able to scale horizontally, namely, by adding more nodes, hence able to process multiple independent tasks in parallel.

The processing component can be organised into sub-components, among which are: (1) Data ingestion, (2) Machine learning models and analytics services (such as image segmentation, objection detection, and image classification) (3) Analytics pipelines (processes or programming scripts built to provide analytical services), (4) Data integration and (5) Data export; which helps to deal with any given use case such as depositing new images and metadata, annotating the images, and depositing trained deep learning models.

## Data ingestion

Data ingestion is the process of adding new data to the system, encompassing tasks such as crawling, parsing, validating and transforming information to be indexed. This process includes several types of data, including metadata, images, annotations, analytics pipelines, which includes services and models, and so on. To this end, specific tools should handle the incoming data to the infrastructure, following different paths depending on the data’s source and type. Some brief examples are given below.

### Specimen datasets

When a new dataset, such as a metadata set, is submitted, each entry of the set undergoes a series of tasks to parse, validate and transform the information to facilitate a standardised entry. This may include crawling additional data from external services like GBIF and Wikidata, or to compute metrics, validate geographic coordinates and map them to locations. Additionally, this process will check for duplicate entries based on the existing data in the infrastructure.

### Specimen images

Following the above-mentioned logic, if a dataset contains an associated image, the file needs to be processed before being added to the system. This might include validating the metadata of the file, transformations and even triggering the analytics pipelines to automatically infer new data about the image content itself.

### Image annotations

One of the key features of the system will be the ability to provide annotations for the existing images. When a set of annotations is supplied, these need to be validated and ingested into the system in a series of steps, depending on the data type. This includes ingesting, validating and transforming them into standard data types and structures, depending on the problem (e.g., classification, object detection, natural language processing and optical character recognition). After preprocessing, the set of annotations will be additionally validated to find: if they duplicate existing annotations or if the attached labels make sense, if the tagged region falls inside of the image and so on. This information will then be indexed and provided by the repository component and can be included in datasets, which will serve to improve existing inference tools and develop new ones.

## Machine learning models and analytics services

The same applies to other tasks such as submitting a new image analysis pipeline. As explained below, analytics pipelines provide services to extract information from images and other sources. These sets of services can be added over time and should be open, well documented and reproducible. This means that the new pipeline, such as one to identify the species of a specimen, would include data and metadata; machine learning models; source code; service containers; automated workflow and service provisioning information as code; results and others. Each of these must be verified and tested, before being included as part of the analytics toolset.

Eventually, the ingestion of data of one type will trigger other sub-components of the system such as the analytics pipelines, to infer data from images, or the data integration components, to invoke other parts of the system, as the repository to index the parsed information, or the storage component, in order to store the image and its derivatives if needed.

### Analytics pipelines

This sub-component encompasses the set of services and functionalities responsible for processing images or other media, to automatically infer information that would otherwise be manually tagged, e.g., identifying a specific trait. To this end, each service provides specific functionality, and encompasses a sequence of instructions, from using multiple pre-trained machine learning and deep learning models, to image transformations or other solutions, depending on the problem at hand. For instance, when ingesting a new dataset, for each given specimen image, various analytics pipelines will be scheduled to run, each made of different steps and deep learning models trained for specific tasks (e.g., detect mercuric chloride stains, identify specific botany traits, extract collector/label information). As a result, the output might include the predicted species of the image, the presence of stains, particular signs of the images, text from handwritten and computerised labels, specific traits and their values, and so on.??????

## Build machine learning models and services

The analytics pipelines are built of pre-trained models, as well as containerized applications and services which have been previously built. The most computationally intensive part of the envisioned system will be training and building these, either to add novel analytics pipelines or to update existing ones. Hence, it should be possible to schedule the execution of these heavy tasks, which will include preparing the data to be used (e.g., resize, augmentation), configuring the environment and parameters to use, training the models and assessing the performance, building, testing, and packaging the services. Moreover, if selected, these should be deployed to production as part of the existing analytics pipelines.

The system must allow the definition of the service workflow as code, from the infrastructure, to model training and application packaging. This requires two parts. First, fully documenting the modelling experiments guaranteeing its reproducibility, which includes providing the data (i.e., link to the exact dataset) and code with the exact environment (e.g., by using conda and [venv](https://docs.python.org/3/library/venv.html) under Python, or [renv](https://rstudio.github.io/renv/articles/renv.html)in R), the pre-trained models and all the required parameters, hyperparameters and similar, as well as controlling the randomness of such models (e.g., initialising seed state). Such data should be indexed by the system and allow anyone to rerun the experiment and obtain the exact model and results.

Secondly, the entire analytics pipeline should be documented as code, from infrastructure to the application level. This allows for the exact replication of the build, test, package and deployment. Over the last decade several technologies and sets of practices have appeared to attain such goals, normally linked to software development concepts such as DevOps, MLOps and GitOps. GitHub provides[Actions](https://github.com/features/actions)  to attain continuous integration and deployment, allowing the automation of the entire workflow of a software service, from building to testing and deploying, based on simple text files (YAML). On the other hand, Docker images and similar solutions allow services to be containerized using similar simple definitions and shared. Going a step further, it is nowadays possible to define both the infrastructure and how services interact as code too (e.g., used under Docker compose or with Terraform and Kubernetes).

Thus, such concepts must be exploited by the processing component, allowing the submission of novel analytics pipelines fully documented as code. As the number of annotated datasets grows over time, the system might schedule the retraining of models and associated pipelines, reporting the results and, if desired, replacing the existing analytics pipelines. Moreover, all the details, code and pre-trained models can be provided, so anyone can reuse the models and code anywhere. Given the computation power needed, possibly requiring several GPUs for bursts of work, hybrid solutions offloading part of this work to cloud providers could be implemented, as an alternative to hosting and managing GPU clusters.

### Data integration

Data integration will push the data extracted or generated by the above mentioned sub-components and push it to the respective parts of the system, that is the repository (e.g., metadata registry of the trained models and images, datasets, the annotated data, and so on) and the storage (e.g,. Image files and their derivatives, pre-trained models, metadata packages, among others).

### Data export

The system will catalogue millions of specimens, each with variable amounts of metadata. These data can be filtered with complex queries, based on several parameters and fields. As an example, a user might want to search for specimen images of a specific species, containing images and having annotations regarding the presence of signatures between a specific timespan. Requesting the generation of an image dataset based on the result of such query requires several processing tasks that need to be scheduled, from the extraction and merging of the relevant metadata into the desired format, to resizing images if needed, assigning a persistent identifier so it can be uniquely identified, generating a dataset page and notifying the user. Moreover, if new images falling on the same search criteria are annotated in the following months, the user might request the dataset to be updated, generating a second version and assigning a new or versioned persistent identifier such as the [DOI](https://support.datacite.org/docs/versioning). As an example of the feasibility of this, part of this functionality is already demonstrated by GBIF, which uses background jobs to export datasets on user request (excluding images and DOIs, but allowing the export of metadata based on queries). Moreover, this sub-component may also be responsible for exporting machine learning datasets to public platforms such as the Registry of Open Data on [AWS](https://registry.opendata.aws/) or [Google Datasets](https://cloud.google.com/solutions/datasets), allowing users to easily mount them on external cloud solutions.

# Discussion

As the decades of the 21st century proceed, we anticipate important changes in global biodiversity. The resource needed to adequately monitor this change is far greater than the cadre of professional ecologists and taxonomists can provide. Machine learning offers one of the promising technologies to dramatically increase our collective capacity to provide the data and identify the taxa, and in complementary fashion prioritise the attention of human taxonomists where they are most needed. In table 1 we have listed the direct benefits to biodiversity and research into artificial intelligence, but there are also positive impacts for society, the economy, the environment and for the collections holding institutions (also see (Popov et al., 2021)). Making images accessible in a common infrastructure is an opportunity for small collections to gain access to tools that would otherwise be unavailable to them given their limited resources. Indeed, Open Access for all researchers including those from the Global South is critical to ensure that collections fulfil their obligations to access and benefit sharing. Making an infrastructure accessible will require a commitment to ease of use, good tutorials, a user focused design and capacity building.

Such an infrastructure aligns with the European Strategy for Data (European Commission, 2020), which aims to overcome challenges related to fragmentation, data availability and reuse, data quality and interoperability, and dissolve barriers across sectors. Having a global infrastructure in place will incentivize natural history collections and their funders to digitise their specimens, and attract funding to do so.

Infrastructures that aggregate images of biological specimens from different sources do exist. For example, [Europeana](https://www.europeana.eu/)bridges the historical divide between natural history and cultural history collections and it can be searched as a whole (Petras et al., 2017), but all images have to be downloaded before they can be used in image analysis. GBIF aggregates specimen data from many different providers, which can also supply links to images of these specimens. GBIF maintains a temporary *cache*of these images, using [*Thumbor*](http://www.thumbor.org/), but they are not readily available for processing.

In the USA, [*iDigBio*](https://www.idigbio.org/) (Integrated Digitized Biocollections) was established to coordinate a nationwide digitization effort, developing the infrastructure to standardise and preserve specimen data. Currently it holds almost 43 million media records. Although it has now been phased out, an experimental pipeline had been set up for users to apply image processing algorithms to subsets of the hosted media (Collins et al., 2018).

In Europe, [EUDAT](https://eudat.eu/)allows data access, services and storage to support the scientific community. Its development is based on a network of more than 20 European research organisations and data centres, based in 14 different European countries. It builds the backbone of the European Open Science Cloud ([EOSC](https://eosc.eu/)), which aims to offer open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines. In the context of the *Herbadrop*project, EUDAT has managed and processed more than 2 million specimen images representing, equalling more than 15 TB of data and 180,000 hours of computation power.

The recent (2021) funded US National Science Foundation [Imageomics Institute](https://imageomics.osu.edu/)will establish infrastructure for biologists to use machine learning algorithms to analyse existing image data from publicly funded digital collections, including natural history collections. Also, Google’s AutoML provides a commercial platform in the cloud for training custom models using cloud services. Such developments may mean that the whole infrastructure does not need to be built from scratch, indeed if it can be built upon existing systems it will be less expensive and more reliable.

Opportunity, obstacles and risks to realising a shared infrastructure for natural history collections

Given the many use cases, the large number and diversity of stakeholders, and the potential for innovative services and research, what is holding us back from creating the proposed infrastructure? One clear issue is that the experts in machine learning are not always aware of the needs of biological collections. More needs to be done to bring these communities together, but perhaps also to find the areas of more general interest where collections can benefit from generalised approaches. A lack of standardisation and consequent lack of interoperability further impedes progress (Lannom et al., 2020).

We suggest that the most intractable obstacles to a shared, global infrastructure are socio-political. We envisage an infrastructure without institutional and national borders, in which people, organisations and nations are co-beneficiaries of a system, in which knowledge, skills, financing and other resourcing are acknowledged (Pearce et al., 2020). Furthermore, tracking the provenance of resources is also needed to ensure reproducibility and replicability of the system (Goodman et al., 2014).

Experiments so far lack scalability, often have manual bottlenecks in the workflows, and there is a significant time lag in the production of results due to limited access to computational and physical resources, but also to human resources to create and curate training datasets (Wäldchen and Mäder, 2018).

Data access is especially important to ensure researchers in places where biodiversity is especially rich, and threatened with extinction, including tropical countries in the Global South (Fazey et al., 2005). A large percentage of the world’s natural history specimens are housed in collections in global institutions in the Global North (Thiers, 2020). This undoubtedly contributes to the exclusion of local scientists from research on their own countries (Dahdouh-Guebas et al., 2003).

The establishment of a new paradigm in research on collections impacts the frameworks and the workflows currently used in collection curation and the research based on them and can therefore be disruptive. One of the largest risks is introducing inherent errors and biases that are derived from the algorithms and prejudices that may be embedded unknowingly in training data (Boakes et al., 2010; Osoba and Welser, 2017).

The institutions that hold collections have safeguarded this rich resource of information about biodiversity and natural history. They are the main stakeholders for these materials to be preserved and associated data to become available for researchers and society. Paradoxically, making the data accessible digitally might create the illusion that there is no need to maintain the collections physically any more. We must resist any impression that physical specimens are not less valuable, because in fact the more information we can extract and link to them, the more valuable they become for any future technology that can be applied to them. It is therefore critical to guarantee the link between the digital specimen and the physical specimen to ensure neither becomes obsolete, risking the real value attached to both.

Finally, the long-term sustainability of an infrastructure should be considered. Infrastructures consume resources, need maintenance and replacement. Software should be updated periodically to keep up with the latest technology releases. Infrastructure needs to justify their maintenance costs and they must have the means to monitor and quantify their impact on science, technology and society. Unstable funding and short-term prioritisation might undermine the potential of such a resource for the future.

## Table 1: Potential positive impacts of an infrastructure to support image analysis of biodiversity specimens.

### Research and Investigation

* An increase in the ease and speed at which specimens can be found
* An increase in the speed of digitization of specimen label data
* An improvement in the value of data by increasing its quality and accuracy
* A force multiplier to further innovation into new use cases Increased rate of species discovery and description
* An increase in historical baseline data for biodiversity
* Development of new approaches to specimen data analysis, such as network analysis
* More sensitive detection of environmental changes
* A liberation of data to support research into the history of science and scientometrics
* Empowerment of taxonomists and conservationists in biodiverse countries

### Society

* A reduction in travel to visit collections
* An increase in access to biological collections and the knowledge therein
* More equitable access to collections globally
* More evidence-based environmental policy
* Improvement in, and wider access to pest identification and wildlife management tools
* Greater use of biological specimens in education

### Economy & Environment

* A reduction in the costs of digitalization and research
* A reduction in the costs of access to biodiversity data
* Improved species identification, providing benefits in:
* pest detection methods and a decrease in the use of pesticides
* faster and more accurate detection of invasive species and reduction of economic losses from threats
* Identification and conservation of threatened species
* Improved access to data (including distribution and trait data) and awareness of samples, providing benefits in many sectors including agricultural research & development and medicines discovery
* Improved environmental forecasting
* Improved mitigation of the impact of biodiversity loss, in turn supporting the preservation of vital ecosystem services
* Filling of missing data for models of climate change impacts
* Reduced economic costs of mitigating the negative impacts of climate and biodiversity change
* Better informed environmental planning

### Museums

* Reduced risk of losing historical data
* A reduction in the damage to collections through reduced handling
* Spin-offs that can be applied to other types of collections
* Greater engagement of scientists directly with collections

## The future

Objects in natural history collections represent one of the most important tools to understand life on our planet. Mobilising the capacity to analyse billions of objects with the help of machine learning is essential to meet the challenge of conserving and sustainably using biodiversity in the alarmingly short time-frames. This paper is written to emphasise the huge potential and the challenges. The main limitation to achieving our vision is not the software for machine learning, nor the ideas for using it, but it is the accessibility of data and images of specimens in a computational environment where they can be processed efficiently. The potential applications of machine learning to specimen images can be divided into those of the specimen object itself and those related to the associated labels. Future expansion of the approach will see further extraction of traits from the specimen images include the 1) quantitative and qualitative analysis of organisms structure (e.g., relative proportions, topological arrangement of organs, symmetry); 2) actual type-based taxonomy, i.e., clustering morphological groups of specimens, assessing the consistency of current hierarchies in taxonomy and building automatic identification tools through a direct link to the traits retrievable from specimens and 3) solving metadata limitations through the analytical and comparative analysis of preparation and mounting “styles”, even when their identity is not explicitly linked to the specimen itself.

Many additional uses can be imagined for the analysis of non-specimen data, that is the additional information that is linked to the physical object, either when directly written on attached labels or linked inventories, catalogues, or spreadsheets (Hardisty et al., 2022). These include extracting information on: 1) the interaction between species and the abiotic elements of their environment; 2) collection data expressed by cryptic textual elements, such as *idem* and *ibidem* , that imply a link to other text, and 3) tracing nomenclatural type material, by linking data elements on type specimens. There is also enormous potential for biological collections that have so far not been the main focus of digitization, including microscope slides of thin sections; histological; or other extractions (Fig. 2B).

The uses of machine learning on collection images are numerous, but as we have shown the real benefits come from scaling up the approach and being able to combine images of many collections. One can imagine research into fields ranging from morphometry, evolution, environmental change to biomimicry and subjects in the humanities. Although imagination is the ultimate limit, we are currently limited by the availability of infrastructure to conduct such research.

# Acknowledgements

This work was supported by European Cooperation in Science and Technology (COST) as part of the Mobilise Action CA17106 on Mobilising Data, Experts and Policies in Scientific Collections. Heliana Teixeira thanks FCT/MCTES for the financial support to the host institution CESAM (UIDB/50017/2020+UIDP/50017/2020). Renato Panda was supported by Ci2 - FCT/MCTES UIDP/05567/2020. This work was also facilitated by the Research Foundation – Flanders research infrastructure under grant number FWO I001721N, the BiCIKL (grant agreement No 101007492) and SYNTHESYS+ (grant agreement No 823827) projects of the European Union’s Horizon 2020 Research and Innovation action under grant agreement.

# References

**Allan**, **E L, Livermore, L, Price, B, Shchedrina, O, and Smith, V 2019 A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. *Biodiversity Data Journal*, 7: e32342. DOI:** [**https://doi.org/10.3897/BDJ.7.e32342**](https://doi.org/10.3897/BDJ.7.e32342)

**Antonelli**, **A, Hiscock, S, Lennon, S, Simmonds, M, Smith, R J, and Young, B 2020 Protecting and sustainably using the world’s plants and fungi. *PLANTS PEOPLE PLANET*, 2(5): 368–370. DOI:** [**https://doi.org/10.1002/ppp3.10150**](https://doi.org/10.1002/ppp3.10150)

**Bauters**, **M, Meeus, S, Barthel, M, Stoffelen, P, De Deurwaerder, H P T, Meunier, F, Drake, T W, Ponette, Q, Ebuy, J, Vermeir, P, Beeckman, H, wyffels, F, Bodé, S, Verbeeck, H, Vandelook, F, and Boeckx, P 2020 Century‐long apparent decrease in intrinsic water‐use efficiency with no evidence of progressive nutrient limitation in African tropical forests. *Global Change Biology*, 26(8): 4449–4461. DOI:** [**https://doi.org/10.1111/gcb.15145**](https://doi.org/10.1111/gcb.15145)

**Bhalerao**, **A and Reynolds, G 2014 Ruler Detection for Autoscaling Forensic Images:. *International Journal of Digital Crime and Forensics*, 6(1): 9–27. DOI:** [**https://doi.org/10.4018/ijdcf.2014010102**](https://doi.org/10.4018/ijdcf.2014010102)

**Boakes**, **E H, McGowan, P J K, Fuller, R A, Chang-qing, D, Clark, N E, O’Connor, K, and Mace, G M 2010 Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLoS Biology*, 8(6): e1000385. DOI:** [**https://doi.org/10.1371/journal.pbio.1000385**](https://doi.org/10.1371/journal.pbio.1000385)

**Bonnet**, **P, Joly, A, Faton, J M, Brown, S, Kimiti, D, Deneu, B, Servajean, M, Affouard, A, Lombardo, J C, Mary, L, Vignau, C, and Munoz, F 2020 How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence*, 1(2). DOI:** [**https://doi.org/10.1002/2688-8319.12023**](https://doi.org/10.1002/2688-8319.12023)

**Borsch**, **T, Stevens, A-D, Häffner, E, Güntsch, A, Berendsohn, W G, Appelhans, M, Barilaro, C, Beszteri, B, Blattner, F, Bossdorf, O, Dalitz, H, Dressler, S, Duque-Thüs, R, Esser, H-J, Franzke, A, Goetze, D, Grein, M, Grünert, U, Hellwig, F, Hentschel, J, Hörandl, E, Janßen, T, Jürgens, N, Kadereit, G, Karisch, T, Koch, M, Müller, F, Müller, J, Ober, D, Porembski, S, Poschlod, P, Printzen, C, Röser, M, Sack, P, Schlüter, P, Schmidt, M, Schnittler, M, Scholler, M, Schultz, M, Seeber, E, Simmel, J, Stiller, M, Thiv, M, Thüs, H, Tkach, N, Triebel, D, Warnke, U, Weibulat, T, Wesche, K, Yurkov, A, and Zizka, G 2020 A complete digitization of German herbaria is possible sensible and should be started now. *Research Ideas and Outcomes*, 6. DOI:** [**https://doi.org/10.3897/rio.6.e50675**](https://doi.org/10.3897/rio.6.e50675)

**Bourel**, **B, Marchant, R, de Garidel-Thoron, T, Tetard, M, Barboni, D, Gally, Y, and Beaufort, L 2020 Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains. *Computers & Geosciences*, 140: 104498. DOI:** [**https://doi.org/10.1016/j.cageo.2020.104498**](https://doi.org/10.1016/j.cageo.2020.104498)

**Bradley**, **R D, Bradley, L C, Garner, H J, and Baker, R J 2014 Assessing the Value of Natural History Collections and Addressing Issues Regarding Long-Term Growth and Care. *BioScience*, 64(12): 1150–1158. DOI:** [**https://doi.org/10.1093/biosci/biu166**](https://doi.org/10.1093/biosci/biu166)

**Burgin**, **C J, Colella, J P, Kahn, P L, and Upham, N S 2018 How many species of mammals are there?. *Journal of Mammalogy*, 99(1): 1–14. DOI:** [**https://doi.org/10.1093/jmammal/gyx147**](https://doi.org/10.1093/jmammal/gyx147)

**Carranza-Rojas**, **J, Goeau, H, Bonnet, P, Mata-Montero, E, and Joly, A 2017 Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology*, 17(1): 181. DOI:** [**https://doi.org/10.1186/s12862-017-1014-z**](https://doi.org/10.1186/s12862-017-1014-z)

**Carranza-Rojas**, **J, Goeau, H, Bonnet, P, Mata-Montero, E, and Joly, A 2017 Going deeper in the automated identification of Herbarium specimens.. *BMC Evol Biol*, 17: 181**.

**Champ**, **J, Mora‐Fallas, A, Goëau, H, Mata‐Montero, E, Bonnet, P, and Joly, A 2020 Instance segmentation for the fine detection of crop and weed plants by precision agricultural robots. *Applications in Plant Sciences*, 8(7). DOI:** [**https://doi.org/10.1002/aps3.11373**](https://doi.org/10.1002/aps3.11373)

**Cheek**, **M, Nic Lughadha, E, Kirk, P, Lindon, H, Carretero, J, Looney, B, Douglas, B, Haelewaters, D, Gaya, E, Llewellyn, T, Ainsworth, A M, Gafforov, Y, Hyde, K, Crous, P, Hughes, M, Walker, B E, Campostrini Forzza, R, Wong, K M, and Niskanen, T 2020 New scientific discoveries: Plants and fungi. *Plants, People, Planet*, 2(5): 371–388. DOI:** [**https://doi.org/10.1002/ppp3.10148**](https://doi.org/10.1002/ppp3.10148)

**Collins**, **M, Yeole, G, Frandsen, P, Dikow, R, Orli, S, and Figueiredo, R 2018 A Pipeline for Deep Learning with Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens. *Biodiversity Information Science and Standards*, 2: e25699. DOI:** [**https://doi.org/10.3897/biss.2.25699**](https://doi.org/10.3897/biss.2.25699)

**Cook**, **J A, Edwards, S V, Lacey, E A, Guralnick, R P, Soltis, P S, Soltis, D E, Welch, C K, Bell, K C, Galbreath, K E, Himes, C, Allen, J M, Heath, T A, Carnaval, A C, Cooper, K L, Liu, M, Hanken, J, and Ickert-Bond, S 2014 Natural History Collections as Emerging Resources for Innovative Education. *BioScience*, 64(8): 725–734. DOI:** [**https://doi.org/10.1093/biosci/biu096**](https://doi.org/10.1093/biosci/biu096)

**Dahdouh-Guebas**, **F, Ahimbisibwe, J, Van Moll, R, and Koedam, N 2003 Neo-colonial science by the most industrialised upon the least developed countries in peer-reviewed publishing. *Scientometrics*, 56(3): 329–343. DOI:** [**https://doi.org/10.1023/A:1022374703178**](https://doi.org/10.1023/A:1022374703178)

**Davis**, **A K, Woodall, N, Moskowitz, J P, Castleberry, N, and Freeman, B J 2013 Temporal Change in Fur Color in Museum Specimens of Mammals: Reddish-Brown Species Get Redder with Storage Time. *International Journal of Zoology*, 2013: 1–6. DOI:** [**https://doi.org/10.1155/2013/876347**](https://doi.org/10.1155/2013/876347)

**Davis**, **A P, Chadburn, H, Moat, J, O’Sullivan, R, Hargreaves, S, and Lughadha, E N 2019 High extinction risk for wild coffee species and implications for coffee sector sustainability. *Science Advances*, 5(1). DOI:** [**https://doi.org/10.1126/sciadv.aav3473**](https://doi.org/10.1126/sciadv.aav3473)

**Davis**, **C C, Champ, J, Park, D S, Breckheimer, I, Lyra, G M, Xie, J, Joly, A, Tarapore, D, Ellison, A M, and Bonnet, P 2020 A New Method for Counting Reproductive Structures in Digitized Herbarium Specimens Using Mask R-CNN. *Frontiers in Plant Science*, 11: 1129. DOI:** [**https://doi.org/10.3389/fpls.2020.01129**](https://doi.org/10.3389/fpls.2020.01129)

**Drinkwater**, **R, Cubey, R, and Haston, E 2014 The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. *PhytoKeys*, 38: 15–30. DOI:** [**https://doi.org/10.3897/phytokeys.38.7168**](https://doi.org/10.3897/phytokeys.38.7168)

**Earl**, **C, White, A, Trizna, M, Frandsen, P, Kawahara, A, Brady, S, and Dikow, R 2019 Discovering Patterns of Biodiversity in Insects Using Deep Machine Learning. *Biodiversity Information Science and Standards*, 3: e37525. DOI:** [**https://doi.org/10.3897/biss.3.37525**](https://doi.org/10.3897/biss.3.37525)

**Elhamod**, **M, Diamond, K M, Maga, A M, Bakis, Y, Bart, H L, Mabee, P, Dahdul, W, Leipzig, J, Greenberg, J, Avants, B, and Karpatne, A 2022 Hierarchy‐guided neural network for species classification. *Methods in Ecology and Evolution*, 13(3): 642–652. DOI:** [**https://doi.org/10.1111/2041-210X.13768**](https://doi.org/10.1111/2041-210X.13768)

**Elliott**, **M J, Poelen, J H, and Fortes, J 2022 Signed Citations: Making Persistent and Verifiable Citations of Digital Scientific Content (preprint). MetaArXiv. DOI:** [**https://doi.org/10.31222/osf.io/wycjn**](https://doi.org/10.31222/osf.io/wycjn)

**Fazey**, **I, Fischer, J, and Lindenmayer, D B 2005 Who does all the research in conservation biology?. *Biodiversity and Conservation*, 14(4): 917–934. DOI:** [**https://doi.org/10.1007/s10531-004-7849-9**](https://doi.org/10.1007/s10531-004-7849-9)

**Goodman**, **A, Pepe, A, Blocker, A W, Borgman, C L, Cranmer, K, Crosas, M, Di Stefano, R, Gil, Y, Groth, P, Hedstrom, M, Hogg, D W, Kashyap, V, Mahabal, A, Siemiginowska, A, and Slavkovic, A 2014 Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology*, 10(4): e1003542. DOI:** [**https://doi.org/10.1371/journal.pcbi.1003542**](https://doi.org/10.1371/journal.pcbi.1003542)

**Goëau**, **H, Bonnet, P, and Joly, A 2021 Overview of PlantCLEF 2021: cross-domain plant identification. In: *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*. Bucarest, Romania. pp. 1422–1436. Available at** [**https://hal.inrae.fr/hal-03353469**](https://hal.inrae.fr/hal-03353469) **[Accessed August 1, 2022**].

**Goëau**, **H, Lorieul, T, Heuret, P, Joly, A, and Bonnet, P 2022 Can Artificial Intelligence Help in the Study of Vegetative Growth Patterns from Herbarium Collections? An Evaluation of the Tropical Flora of the French Guiana Forest. *Plants*, 11(4): 530. DOI:** [**https://doi.org/10.3390/plants11040530**](https://doi.org/10.3390/plants11040530)

**Goëau**, **H, Mora‐Fallas, A, Champ, J, Love, N L R, Mazer, S J, Mata‐Montero, E, Joly, A, and Bonnet, P 2020 A new fine‐grained method for automated visual analysis of herbarium specimens: A case study for phenological data extraction. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11368**](https://doi.org/10.1002/aps3.11368)

**Grace**, **O M, Pérez-Escobar, O A, Lucas, E J, Vorontsova, M S, Lewis, G P, Walker, B E, Lohmann, L G, Knapp, S, Wilkie, P, Sarkinen, T, Darbyshire, I, Lughadha, E N, Monro, A, Woudstra, Y, Demissew, S, Muasya, A M, Díaz, S, Baker, W J, and Antonelli, A 2021 Botanical Monography in the Anthropocene. *Trends in Plant Science*, 26(5): 433–441. DOI:** [**https://doi.org/10.1016/j.tplants.2020.12.018**](https://doi.org/10.1016/j.tplants.2020.12.018)

**Groom**, **Q J, O’Reilly, C, and Humphrey, T 2014 Herbarium specimens reveal the exchange network of British and Irish botanists, 1856–1932. *New Journal of Botany*, 4(2): 95–103. DOI:** [**https://doi.org/10.1179/2042349714Y.0000000041**](https://doi.org/10.1179/2042349714Y.0000000041)

**Hardisty**, **A, Saarenmaa, H, Casino, A, Dillen, M, Gödderz, K, Groom, Q, Hardy, H, Koureas, D, Nieva de la Hidalga, A, Paul, D, Runnel, V, Vermeersch, X, van Walsum, M, and Willemse, L 2020 Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. *Research Ideas and Outcomes*, 6: e54280. DOI:** [**https://doi.org/10.3897/rio.6.e54280**](https://doi.org/10.3897/rio.6.e54280)

**Hardisty**, **A R, Ellwood, E R, Nelson, G, Zimkus, B, Buschbom, J, Addink, W, Rabeler, R K, Bates, J, Bentley, A, Fortes, J A B, Hansen, S, Macklin, J A, Mast, A R, Miller, J T, Monfils, A K, Paul, D L, Wallis, E, and Webster, M 2022 Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. *BioScience*, biac060. DOI:** [**https://doi.org/10.1093/biosci/biac060**](https://doi.org/10.1093/biosci/biac060)

**Hay**, **A, Xian, W, Bailly, N, Liang, C, and Pauly, D 2020 The why and how of determining length‐weight relationships of fish from preserved museum specimens. *Journal of Applied Ichthyology*, 36(3): 373–379. DOI:** [**https://doi.org/10.1111/jai.14014**](https://doi.org/10.1111/jai.14014)

**Heberling**, **J M 2022 Herbaria as Big Data Sources of Plant Traits. *International Journal of Plant Sciences*, 183(2): 87–118. DOI:** [**https://doi.org/10.1086/717623**](https://doi.org/10.1086/717623)

**Heberling**, **J M and Isaac, B L 2017 Herbarium specimens as exaptations: New uses for old collections. *American Journal of Botany*, 104(7): 963–965. DOI:** [**https://doi.org/10.3732/ajb.1700125**](https://doi.org/10.3732/ajb.1700125)

**Hornick**, **T, Richter, A, Harpole, W S, Bastl, M, Bohlmann, S, Bonn, A, Bumberger, J, Dietrich, P, Gemeinholzer, B, Grote, R, Heinold, B, Keller, A, Luttkus, M L, Mäder, P, Motivans Švara, E, Passonneau, S, Punyasena, S W, Rakosy, D, Richter, R, Sickel, W, Steffan‐Dewenter, I, Theodorou, P, Treudler, R, Werchan, B, Werchan, M, Wolke, R, and Dunker, S 2022 An integrative environmental pollen diversity assessment and its importance for the Sustainable Development Goals. *Plants, People, Planet*, 4(2): 110–121. DOI:** [**https://doi.org/10.1002/ppp3.10234**](https://doi.org/10.1002/ppp3.10234)

**Hoyal Cuthill**, **J F, Guttenberg, N, Ledger, S, Crowther, R, and Huertas, B 2019 Deep learning on butterfly phenotypes tests evolution’s oldest mathematical model. *Science Advances*, 5(8): eaaw4967. DOI:** [**https://doi.org/10.1126/sciadv.aaw4967**](https://doi.org/10.1126/sciadv.aaw4967)

**Hussein**, **B R, Malik, O A, Ong, W-H, and Slik, J W F 2022 Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. *Ecological Informatics*, 69: 101641. DOI:** [**https://doi.org/10.1016/j.ecoinf.2022.101641**](https://doi.org/10.1016/j.ecoinf.2022.101641)

**Hussein**, **B R, Malik, O A, Ong, W-H, and Slik, J W F 2022 Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. *Ecological Informatics*, 69: 101641. DOI:** [**https://doi.org/10.1016/j.ecoinf.2022.101641**](https://doi.org/10.1016/j.ecoinf.2022.101641)

**Høye**, **T T, Ärje, J, Bjerge, K, Hansen, O L P, Iosifidis, A, Leese, F, Mann, H M R, Meissner, K, Melvad, C, and Raitoharju, J 2021 Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, 118(2): e2002545117. DOI:** [**https://doi.org/10.1073/pnas.2002545117**](https://doi.org/10.1073/pnas.2002545117)

**Johnson**, **J, Douze, M, and Jegou, H 2021 Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547. DOI:** [**https://doi.org/10.1109/TBDATA.2019.2921572**](https://doi.org/10.1109/TBDATA.2019.2921572)

**Kala**, **J, De Kauwe, M G, Pitman, A J, Medlyn, B E, Wang, Y-P, Lorenz, R, and Perkins-Kirkpatrick, S E 2016 Impact of the representation of stomatal conductance on model projections of heatwave intensity. *Scientific Reports*, 6(1): 23418. DOI:** [**https://doi.org/10.1038/srep23418**](https://doi.org/10.1038/srep23418)

**Kattge**, **J, Bönisch, G, Díaz, S, Lavorel, S, Prentice, I C, Leadley, P, Tautenhahn, S, Werner, G D A, Aakala, T, Abedi, M, Acosta, A T R, Adamidis, G C, Adamson, K, Aiba, M, Albert, C H, Alcántara, J M, Alcázar C, C, Aleixo, I, Ali, H, Amiaud, B, Ammer, C, Amoroso, M M, Anand, M, Anderson, C, Anten, N, Antos, J, Apgaua, D M G, Ashman, T-L, Asmara, D H, Asner, G P, Aspinwall, M, Atkin, O, Aubin, I, Baastrup-Spohr, L, Bahalkeh, K, Bahn, M, Baker, T, Baker, W J, Bakker, J P, Baldocchi, D, Baltzer, J, Banerjee, A, Baranger, A, Barlow, J, Barneche, D R, Baruch, Z, Bastianelli, D, Battles, J, Bauerle, W, Bauters, M, Bazzato, E, Beckmann, M, Beeckman, H, Beierkuhnlein, C, Bekker, R, Belfry, G, Belluau, M, Beloiu, M, Benavides, R, Benomar, L, Berdugo-Lattke, M L, Berenguer, E, Bergamin, R, Bergmann, J, Bergmann Carlucci, M, Berner, L, Bernhardt-Römermann, M, Bigler, C, Bjorkman, A D, Blackman, C, Blanco, C, Blonder, B, Blumenthal, D, Bocanegra-González, K T, Boeckx, P, Bohlman, S, Böhning-Gaese, K, Boisvert-Marsh, L, Bond, W, Bond-Lamberty, B, Boom, A, Boonman, C C F, Bordin, K, Boughton, E H, Boukili, V, Bowman, D M J S, Bravo, S, Brendel, M R, Broadley, M R, Brown, K A, Bruelheide, H, Brumnich, F, Bruun, H H, Bruy, D, Buchanan, S W, Bucher, S F, Buchmann, N, Buitenwerf, R, Bunker, D E, Bürger, J, Burrascano, S, Burslem, D F R P, Butterfield, B J, Byun, C, Marques, M, Scalon, M C, Caccianiga, M, Cadotte, M, Cailleret, M, Camac, J, Camarero, J J, Campany, C, Campetella, G, Campos, J A, Cano-Arboleda, L, Canullo, R, Carbognani, M, Carvalho, F, Casanoves, F, Castagneyrol, B, Catford, J A, Cavender-Bares, J, Cerabolini, B E L, Cervellini, M, Chacón-Madrigal, E, Chapin, K, Chapin, F S, Chelli, S, Chen, S-C, Chen, A, Cherubini, P, Chianucci, F, Choat, B, Chung, K-S, Chytrý, M, Ciccarelli, D, Coll, L, Collins, C G, Conti, L, Coomes, D, Cornelissen, J H C, Cornwell, W K, Corona, P, Coyea, M, Craine, J, Craven, D, Cromsigt, J P G M, Csecserits, A, Cufar, K, Cuntz, M, da Silva, A C, Dahlin, K M, Dainese, M, Dalke, I, Dalle Fratte, M, Dang-Le, A T, Danihelka, J, Dannoura, M, Dawson, S, de Beer, A J, De Frutos, A, De Long, J R, Dechant, B, Delagrange, S, Delpierre, N, Derroire, G, Dias, A S, Diaz-Toribio, M H, Dimitrakopoulos, P G, Dobrowolski, M, Doktor, D, Dřevojan, P, Dong, N, Dransfield, J, Dressler, S, Duarte, L, Ducouret, E, Dullinger, S, Durka, W, Duursma, R, Dymova, O, E-Vojtkó, A, Eckstein, R L, Ejtehadi, H, Elser, J, Emilio, T, Engemann, K, Erfanian, M B, Erfmeier, A, Esquivel-Muelbert, A, Esser, G, Estiarte, M, Domingues, T F, Fagan, W F, Fagúndez, J, Falster, D S, Fan, Y, Fang, J, Farris, E, Fazlioglu, F, Feng, Y, Fernandez-Mendez, F, Ferrara, C, Ferreira, J, Fidelis, A, Finegan, B, Firn, J, Flowers, T J, Flynn, D F B, Fontana, V, Forey, E, Forgiarini, C, François, L, Frangipani, M, Frank, D, Frenette-Dussault, C, Freschet, G T, Fry, E L, Fyllas, N M, Mazzochini, G G, Gachet, S, Gallagher, R, Ganade, G, Ganga, F, García-Palacios, P, Gargaglione, V, Garnier, E, Garrido, J L, de Gasper, A L, Gea-Izquierdo, G, Gibson, D, Gillison, A N, Giroldo, A, Glasenhardt, M-C, Gleason, S, Gliesch, M, Goldberg, E, Göldel, B, Gonzalez-Akre, E, Gonzalez-Andujar, J L, González-Melo, A, González-Robles, A, Graae, B J, Granda, E, Graves, S, Green, W A, Gregor, T, Gross, N, Guerin, G R, Günther, A, Gutiérrez, A G, Haddock, L, Haines, A, Hall, J, Hambuckers, A, Han, W, Harrison, S P, Hattingh, W, Hawes, J E, He, T, He, P, Heberling, J M, Helm, A, Hempel, S, Hentschel, J, Hérault, B, Hereş, A-M, Herz, K, Heuertz, M, Hickler, T, Hietz, P, Higuchi, P, Hipp, A L, Hirons, A, Hock, M, Hogan, J A, Holl, K, Honnay, O, Hornstein, D, Hou, E, Hough-Snee, N, Hovstad, K A, Ichie, T, Igić, B, Illa, E, Isaac, M, Ishihara, M, Ivanov, L, Ivanova, L, Iversen, C M, Izquierdo, J, Jackson, R B, Jackson, B, Jactel, H, Jagodzinski, A M, Jandt, U, Jansen, S, Jenkins, T, Jentsch, A, Jespersen, J R P, Jiang, G-F, Johansen, J L, Johnson, D, Jokela, E J, Joly, C A, Jordan, G J, Joseph, G S, Junaedi, D, Junker, R R, Justes, E, Kabzems, R, Kane, J, Kaplan, Z, Kattenborn, T, Kavelenova, L, Kearsley, E, Kempel, A, Kenzo, T, Kerkhoff, A, Khalil, M I, Kinlock, N L, Kissling, W D, Kitajima, K, Kitzberger, T, Kjøller, R, Klein, T, Kleyer, M, Klimešová, J, Klipel, J, Kloeppel, B, Klotz, S, Knops, J M H, Kohyama, T, Koike, F, Kollmann, J, Komac, B, Komatsu, K, König, C, Kraft, N J B, Kramer, K, Kreft, H, Kühn, I, Kumarathunge, D, Kuppler, J, Kurokawa, H, Kurosawa, Y, Kuyah, S, Laclau, J-P, Lafleur, B, Lallai, E, Lamb, E, Lamprecht, A, Larkin, D J, Laughlin, D, Le Bagousse-Pinguet, Y, le Maire, G, le Roux, P C, le Roux, E, Lee, T, Lens, F, Lewis, S L, Lhotsky, B, Li, Y, Li, X, Lichstein, J W, Liebergesell, M, Lim, J Y, Lin, Y-S, Linares, J C, Liu, C, Liu, D, Liu, U, Livingstone, S, Llusià, J, Lohbeck, M, López-García, Á, Lopez-Gonzalez, G, Lososová, Z, Louault, F, Lukács, B A, Lukeš, P, Luo, Y, Lussu, M, Ma, S, Maciel Rabelo Pereira, C, Mack, M, Maire, V, Mäkelä, A, Mäkinen, H, Malhado, A C M, Mallik, A, Manning, P, Manzoni, S, Marchetti, Z, Marchino, L, Marcilio-Silva, V, Marcon, E, Marignani, M, Markesteijn, L, Martin, A, Martínez-Garza, C, Martínez-Vilalta, J, Mašková, T, Mason, K, Mason, N, Massad, T J, Masse, J, Mayrose, I, McCarthy, J, McCormack, M L, McCulloh, K, McFadden, I R, McGill, B J, McPartland, M Y, Medeiros, J S, Medlyn, B, Meerts, P, Mehrabi, Z, Meir, P, Melo, F P L, Mencuccini, M, Meredieu, C, Messier, J, Mészáros, I, Metsaranta, J, Michaletz, S T, Michelaki, C, Migalina, S, Milla, R, Miller, J E D, Minden, V, Ming, R, Mokany, K, Moles, A T, Molnár V, A, Molofsky, J, Molz, M, Montgomery, R A, Monty, A, Moravcová, L, Moreno-Martínez, A, Moretti, M, Mori, A S, Mori, S, Morris, D, Morrison, J, Mucina, L, Mueller, S, Muir, C D, Müller, S C, Munoz, F, Myers-Smith, I H, Myster, R W, Nagano, M, Naidu, S, Narayanan, A, Natesan, B, Negoita, L, Nelson, A S, Neuschulz, E L, Ni, J, Niedrist, G, Nieto, J, Niinemets, Ü, Nolan, R, Nottebrock, H, Nouvellon, Y, Novakovskiy, A, Network, T N, Nystuen, K O, O’Grady, A, O’Hara, K, O’Reilly-Nugent, A, Oakley, S, Oberhuber, W, Ohtsuka, T, Oliveira, R, Öllerer, K, Olson, M E, Onipchenko, V, Onoda, Y, Onstein, R E, Ordonez, J C, Osada, N, Ostonen, I, Ottaviani, G, Otto, S, Overbeck, G E, Ozinga, W A, Pahl, A T, Paine, C E T, Pakeman, R J, Papageorgiou, A C, Parfionova, E, Pärtel, M, Patacca, M, Paula, S, Paule, J, Pauli, H, Pausas, J G, Peco, B, Penuelas, J, Perea, A, Peri, P L, Petisco-Souza, A C, Petraglia, A, Petritan, A M, Phillips, O L, Pierce, S, Pillar, V D, Pisek, J, Pomogaybin, A, Poorter, H, Portsmuth, A, Poschlod, P, Potvin, C, Pounds, D, Powell, A S, Power, S A, Prinzing, A, Puglielli, G, Pyšek, P, Raevel, V, Rammig, A, Ransijn, J, Ray, C A, Reich, P B, Reichstein, M, Reid, D E B, Réjou-Méchain, M, de Dios, V R, Ribeiro, S, Richardson, S, Riibak, K, Rillig, M C, Riviera, F, Robert, E M R, Roberts, S, Robroek, B, Roddy, A, Rodrigues, A V, Rogers, A, Rollinson, E, Rolo, V, Römermann, C, Ronzhina, D, Roscher, C, Rosell, J A, Rosenfield, M F, Rossi, C, Roy, D B, Royer-Tardif, S, Rüger, N, Ruiz-Peinado, R, Rumpf, S B, Rusch, G M, Ryo, M, Sack, L, Saldaña, A, Salgado-Negret, B, Salguero-Gomez, R, Santa-Regina, I, Santacruz-García, A C, Santos, J, Sardans, J, Schamp, B, Scherer-Lorenzen, M, Schleuning, M, Schmid, B, Schmidt, M, Schmitt, S, Schneider, J V, Schowanek, S D, Schrader, J, Schrodt, F, Schuldt, B, Schurr, F, Selaya Garvizu, G, Semchenko, M, Seymour, C, Sfair, J C, Sharpe, J M, Sheppard, C S, Sheremetiev, S, Shiodera, S, Shipley, B, Shovon, T A, Siebenkäs, A, Sierra, C, Silva, V, Silva, M, Sitzia, T, Sjöman, H, Slot, M, Smith, N G, Sodhi, D, Soltis, P, Soltis, D, Somers, B, Sonnier, G, Sørensen, M V, Sosinski Jr, E E, Soudzilovskaia, N A, Souza, A F, Spasojevic, M, Sperandii, M G, Stan, A B, Stegen, J, Steinbauer, K, Stephan, J G, Sterck, F, Stojanovic, D B, Strydom, T, Suarez, M L, Svenning, J-C, Svitková, I, Svitok, M, Svoboda, M, Swaine, E, Swenson, N, Tabarelli, M, Takagi, K, Tappeiner, U, Tarifa, R, Tauugourdeau, S, Tavsanoglu, C, te Beest, M, Tedersoo, L, Thiffault, N, Thom, D, Thomas, E, Thompson, K, Thornton, P E, Thuiller, W, Tichý, L, Tissue, D, Tjoelker, M G, Tng, D Y P, Tobias, J, Török, P, Tarin, T, Torres-Ruiz, J M, Tóthmérész, B, Treurnicht, M, Trivellone, V, Trolliet, F, Trotsiuk, V, Tsakalos, J L, Tsiripidis, I, Tysklind, N, Umehara, T, Usoltsev, V, Vadeboncoeur, M, Vaezi, J, Valladares, F, Vamosi, J, van Bodegom, P M, van Breugel, M, Van Cleemput, E, van de Weg, M, van der Merwe, S, van der Plas, F, van der Sande, M T, van Kleunen, M, Van Meerbeek, K, Vanderwel, M, Vanselow, K A, Vårhammar, A, Varone, L, Vasquez Valderrama, M Y, Vassilev, K, Vellend, M, Veneklaas, E J, Verbeeck, H, Verheyen, K, Vibrans, A, Vieira, I, Villacís, J, Violle, C, Vivek, P, Wagner, K, Waldram, M, Waldron, A, Walker, A P, Waller, M, Walther, G, Wang, H, Wang, F, Wang, W, Watkins, H, Watkins, J, Weber, U, Weedon, J T, Wei, L, Weigelt, P, Weiher, E, Wells, A W, Wellstein, C, Wenk, E, Westoby, M, Westwood, A, White, P J, Whitten, M, Williams, M, Winkler, D E, Winter, K, Womack, C, Wright, I J, Wright, S J, Wright, J, Pinho, B X, Ximenes, F, Yamada, T, Yamaji, K, Yanai, R, Yankov, N, Yguel, B, Zanini, K J, Zanne, A E, Zelený, D, Zhao, Y-P, Zheng, J, Zheng, J, Ziemińska, K, Zirbel, C R, Zizka, G, Zo-Bi, I C, Zotz, G, and Wirth, C 2020 TRY plant trait database – enhanced coverage and open access. *Global Change Biology*, 26(1): 119–188. DOI:** [**https://doi.org/https://doi.org/10.1111/gcb.14904**](https://doi.org/https://doi.org/10.1111/gcb.14904)

**Kho**, **S J, Manickam, S, Malek, S, Mosleh, M, and Dhillon, S K 2017 Automated plant identification using artificial neural network and support vector machine. *Frontiers in Life Science*, 10(1): 98–107. DOI:** [**https://doi.org/10.1080/21553769.2017.1412361**](https://doi.org/10.1080/21553769.2017.1412361)

**Kommineni**, **V K, Tautenhahn, S, Baddam, P, Gaikwad, J, Wieczorek, B, Triki, A, and Kattge, J 2021 Comprehensive leaf size traits dataset for seven plant species from digitised herbarium specimen images covering more than two centuries. *Biodiversity Data Journal*, 9: e69806. DOI:** [**https://doi.org/10.3897/BDJ.9.e69806**](https://doi.org/10.3897/BDJ.9.e69806)

**Kozlov**, **M V, Sokolova, I V, Zverev, V, and Zvereva, E L 2021 Changes in plant collection practices from the 16th to 21st centuries: implications for the use of herbarium specimens in global change research. *Annals of Botany*, 127(7): 865–873. DOI:** [**https://doi.org/10.1093/aob/mcab016**](https://doi.org/10.1093/aob/mcab016)

**Labastida**, **I and Margoni, T 2020 Licensing FAIR Data for Reuse. *Data Intelligence*, 2(1-2): 199–207. DOI:** [**https://doi.org/10.1162/dint\_a\_00042**](https://doi.org/10.1162/dint_a_00042)

**Lang**, **P L M, Willems, F M, Scheepens, J F, Burbano, H A, and Bossdorf, O 2019 Using herbaria to study global environmental change. *New Phytologist*, 221(1): 110–122. DOI:** [**https://doi.org/10.1111/nph.15401**](https://doi.org/10.1111/nph.15401)

**Lannom**, **L, Koureas, D, and Hardisty, A R 2020 FAIR Data and Services in Biodiversity Science and Geoscience. *Data Intelligence*, 2(1-2): 122–130. DOI:** [**https://doi.org/10.1162/dint\_a\_00034**](https://doi.org/10.1162/dint_a_00034)

**Li**, **R, Zhu, S, Chen, H Y H, John, R, Zhou, G, Zhang, D, Zhang, Q, and Ye, Q 2015 Are functional traits a good predictor of global change impacts on tree species abundance dynamics in a subtropical forest?. *Ecology Letters*, 18(11): 1181–1189. DOI:** [**https://doi.org/10.1111/ele.12497**](https://doi.org/10.1111/ele.12497)

**Little**, **D P, Tulig, M, Tan, K C, Liu, Y, Belongie, S, Kaeser‐Chen, C, Michelangeli, F A, Panesar, K, Guha, R V, and Ambrose, B A 2020 An algorithm competition for automatic species identification from herbarium specimens. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11365**](https://doi.org/10.1002/aps3.11365)

**Lorieul**, **T, Pearson, K D, Ellwood, E R, Goëau, H, Molino, J F, Sweeney, P W, Yost, J M, Sachs, J, Mata‐Montero, E, Nelson, G, Soltis, P S, Bonnet, P, and Joly, A 2019 Toward a large‐scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras. *Applications in Plant Sciences*, 7(3). DOI:** [**https://doi.org/10.1002/aps3.1233**](https://doi.org/10.1002/aps3.1233)

**Meeus**, **S, Van den Bulcke, J, and wyffels, F 2020 From leaf to label: A robust automated workflow for stomata detection. *Ecology and Evolution*, 10(17): 9178–9191. DOI:** [**https://doi.org/10.1002/ece3.6571**](https://doi.org/10.1002/ece3.6571)

**Meineke**, **E K, Classen, A T, Sanders, N J, and Jonathan Davies, T 2019 Herbarium specimens reveal increasing herbivory over the past century. *Journal of Ecology*, 107(1): 105–117. DOI:** [**https://doi.org/10.1111/1365-2745.13057**](https://doi.org/10.1111/1365-2745.13057)

**Meineke**, **E K, Davies, T J, Daru, B H, and Davis, C C 2018 Biological collections for understanding biodiversity in the Anthropocene. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763): 20170386. DOI:** [**https://doi.org/10.1098/rstb.2017.0386**](https://doi.org/10.1098/rstb.2017.0386)

**Meineke**, **E K, Tomasi, C, Yuan, S, and Pryer, K M 2020 Applying machine learning to investigate long‐term insect–plant interactions preserved on digitized herbarium specimens. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11369**](https://doi.org/10.1002/aps3.11369)

**Nelson**, **G and Ellis, S 2018 The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763): 20170391. DOI:** [**https://doi.org/10.1098/rstb.2017.0391**](https://doi.org/10.1098/rstb.2017.0391)

**Nicolson**, **N, Paton, A, Phillips, S, and Tucker, A 2018 Specimens as Research Objects: Reconciliation Across Distributed Repositories to Enable Metadata Propagation. In: *2018 IEEE 14th International Conference on e-Science (e-Science)*. Amsterdam: IEEE. pp. 125–135. DOI:** [**https://doi.org/10.1109/eScience.2018.00028**](https://doi.org/10.1109/eScience.2018.00028)

**Osoba**, **O and Welser, W 2017 *An intelligence in our image: the risks of bias and errors in artificial intelligence*. Santa Monica, Calif: RAND Corporation**.

**Ott**, **T, Palm, C, Vogt, R, and Oberprieler, C 2020 GinJinn: An object‐detection pipeline for automated feature extraction from herbarium specimens. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11351**](https://doi.org/10.1002/aps3.11351)

**Pearce**, **T R, Antonelli, A, Brearley, F Q, Couch, C, Campostrini Forzza, R, Gonçalves, S C, Magassouba, S, Morim, M P, Mueller, G M, Nic Lughadha, E, Obreza, M, Sharrock, S, Simmonds, M S J, Tambam, B B, Utteridge, T M A, and Breman, E 2020 International collaboration between collections‐based institutes for halting biodiversity loss and unlocking the useful properties of plants and fungi. *Plants, People, Planet*, 2(5): 515–534. DOI:** [**https://doi.org/10.1002/ppp3.10149**](https://doi.org/10.1002/ppp3.10149)

**Pearson**, **K D, Nelson, G, Aronson, M F J, Bonnet, P, Brenskelle, L, Davis, C C, Denny, E G, Ellwood, E R, Goëau, H, Heberling, J M, Joly, A, Lorieul, T, Mazer, S J, Meineke, E K, Stucky, B J, Sweeney, P, White, A E, and Soltis, P S 2020 Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. *BioScience*, 70(7): 610–620. DOI:** [**https://doi.org/10.1093/biosci/biaa044**](https://doi.org/10.1093/biosci/biaa044)

**Petras**, **V, Hill, T, Stiller, J, and Gäde, M 2017 Europeana – a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum*, 17(1): 41–46. DOI:** [**https://doi.org/10.1007/s13222-016-0238-1**](https://doi.org/10.1007/s13222-016-0238-1)

**Pigot**, **A L, Sheard, C, Miller, E T, Bregman, T P, Freeman, B G, Roll, U, Seddon, N, Trisos, C H, Weeks, B C, and Tobias, J A 2020 Macroevolutionary convergence connects morphological form to ecological function in birds. *Nature Ecology & Evolution*, 4(2): 230–239. DOI:** [**https://doi.org/10.1038/s41559-019-1070-4**](https://doi.org/10.1038/s41559-019-1070-4)

**Poelen**, **J 2022 jhpoelen/specimen-image-index: 0.0.1. DOI:** [**https://doi.org/10.5281/zenodo.7032668**](https://doi.org/10.5281/zenodo.7032668)

**Poelen**, **J H and Groom, Q 2022 Preserved Specimen Records with Still Images Registered Across Biodiversity Data Networks in Period 2019-2022 hash://sha256/da7450941e7179c973a2fe1127718541bca6ccafe0e4e2bfb7f7ca9dbb7adb86. DOI:** [**https://doi.org/10.5281/zenodo.7032574**](https://doi.org/10.5281/zenodo.7032574)

**Pollock**, **L J, Morris, W K, and Vesk, P A 2012 The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35(8): 716–725. DOI:** [**https://doi.org/10.1111/j.1600-0587.2011.07085.x**](https://doi.org/10.1111/j.1600-0587.2011.07085.x)

**Popov**, **D, Roychoudhury, P, Hardy, H, Livermore, L, and Norris, K 2021 The Value of Digitising Natural History Collections. *Research Ideas and Outcomes*, 7. DOI:** [**https://doi.org/10.3897/rio.7.e78844**](https://doi.org/10.3897/rio.7.e78844)

**Pryer**, **K M, Tomasi, C, Wang, X, Meineke, E K, and Windham, M D 2020 Using computer vision on herbarium specimen images to discriminate among closely related horsetails ( *Equisetum* ). *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11372**](https://doi.org/10.1002/aps3.11372)

**Regos**, **A, Gagne, L, Alcaraz-Segura, D, Honrado, J P, and Domínguez, J 2019 Effects of species traits and environmental predictors on performance and transferability of ecological niche models. *Scientific Reports*, 9(1): 4221. DOI:** [**https://doi.org/10.1038/s41598-019-40766-5**](https://doi.org/10.1038/s41598-019-40766-5)

**Renner**, **S S and Zohner, C M 2018 Climate Change and Phenological Mismatch in Trophic Interactions Among Plants, Insects, and Vertebrates. *Annual Review of Ecology, Evolution, and Systematics*, 49(1): 165–182. DOI:** [**https://doi.org/10.1146/annurev-ecolsys-110617-062535**](https://doi.org/10.1146/annurev-ecolsys-110617-062535)

**Rives**, **A, Meier, J, Sercu, T, Goyal, S, Lin, Z, Liu, J, Guo, D, Ott, M, Zitnick, C L, Ma, J, and Fergus, R 2021 Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118. DOI:** [**https://doi.org/10.1073/pnas.2016239118**](https://doi.org/10.1073/pnas.2016239118)

**Romero**, **I C, Kong, S, Fowlkes, C C, Jaramillo, C, Urban, M A, Oboh-Ikuenobe, F, D’Apolito, C, and Punyasena, S W 2020 Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proceedings of the National Academy of Sciences*, 117(45): 28496–28505. DOI:** [**https://doi.org/10.1073/pnas.2007324117**](https://doi.org/10.1073/pnas.2007324117)

**Schuettpelz**, **E, Frandsen, P, Dikow, R, Brown, A, Orli, S, Peters, M, Metallo, A, Funk, V, and Dorr, L 2017 Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal*, 5: e21139. DOI:** [**https://doi.org/10.3897/BDJ.5.e21139**](https://doi.org/10.3897/BDJ.5.e21139)

**Soltis**, **P S 2017 Digitization of herbaria enables novel research. *American Journal of Botany*, 104(9): 1281–1284. DOI:** [**https://doi.org/10.3732/ajb.1700281**](https://doi.org/10.3732/ajb.1700281)

**Soltis**, **P S, Nelson, G, Zare, A, and Meineke, E K 2020 Plants meet machines: Prospects in machine learning for plant biology. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11371**](https://doi.org/10.1002/aps3.11371)

**Thiers**, **B 2020 The World’s Herbaria 2019: A Summary Report Based on Data from Index Herbariorum. Available at** [**http://sweetgum.nybg.org/science/docs/The\_Worlds\_Herbaria\_2019.pdf**](http://sweetgum.nybg.org/science/docs/The_Worlds_Herbaria_2019.pdf) **[Accessed August 1, 2022**].

**Tomaszewski**, **D and Górzkowska, A 2016 Is Shape of a Fresh and Dried Leaf the Same?. *PLOS ONE*, 11(4): e0153071. DOI:** [**https://doi.org/10.1371/journal.pone.0153071**](https://doi.org/10.1371/journal.pone.0153071)

**Triki**, **A, Bouaziz, B, Gaikwad, J, and Mahdi, W 2021 Deep leaf: Mask R-CNN based leaf detection and segmentation from digitized herbarium specimen images. *Pattern Recognition Letters*, 150: 76–83. DOI:** [**https://doi.org/10.1016/j.patrec.2021.07.003**](https://doi.org/10.1016/j.patrec.2021.07.003)

**Triki**, **A, Bouaziz, B, Gaikwad, J, and Mahdi, W 2021 A deep learning-based approach for segmenting and counting reproductive organs from digitized herbarium specimen images using refined Mask Scoring R-CNN.. In: *TACC*. pp. 138–148**.

**Triki**, **A, Bouaziz, B, Mahdi, W, and Gaikwad, J 2020 Objects Detection from Digitized Herbarium Specimen based on Improved YOLO V3:. In: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. Valletta, Malta: SCITEPRESS - Science and Technology Publications. pp. 523–529. DOI:** [**https://doi.org/10.5220/0009170005230529**](https://doi.org/10.5220/0009170005230529)

**Valan**, **M, Makonyi, K, Maki, A, Vondráček, D, and Ronquist, F 2019 Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. *Systematic Biology*, 68(6): 876–895. DOI:** [**https://doi.org/10.1093/sysbio/syz014**](https://doi.org/10.1093/sysbio/syz014)

**Van Horn**, **G, Cole, E, Beery, S, Wilber, K, Belongie, S, and Mac Aodha, O 2021 Benchmarking Representation Learning for Natural World Image Collections. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12884–12893**.

**Vega**, **F E, Smith, L T, Davies, N M J, Moat, J, Góral, T, O’Sullivan, R, and Davis, A P 2019 Elucidation of Hosts, Native Distribution, and Habitat of the Coffee Berry Borer (Hypothenemus hampei) Using Herbaria and Other Museum Collections. *Frontiers in Plant Science*, 10: 1188. DOI:** [**https://doi.org/10.3389/fpls.2019.01188**](https://doi.org/10.3389/fpls.2019.01188)

**Walker**, **B E, Tucker, A, and Nicolson, N 2022 Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning. *Frontiers in Plant Science*, 12: 806407. DOI:** [**https://doi.org/10.3389/fpls.2021.806407**](https://doi.org/10.3389/fpls.2021.806407)

**Wearn**, **O R and Glover-Kapfer, P 2019 Snap happy: camera traps are an effective sampling tool when compared with alternative methods. *Royal Society Open Science*, 6(3): 181748. DOI:** [**https://doi.org/10.1098/rsos.181748**](https://doi.org/10.1098/rsos.181748)

**Weaver**, **W N, Ng, J, and Laport, R G 2020 LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11367**](https://doi.org/10.1002/aps3.11367)

**White**, **A, Trizna, M, Frandsen, P, Dorr, L, Dikow, R, and Schuettpelz, E 2019 Evaluating Geographic Patterns of Morphological Diversity in Ferns and Lycophytes Using Deep Neural Networks. *Biodiversity Information Science and Standards*, 3: e37559. DOI:** [**https://doi.org/10.3897/biss.3.37559**](https://doi.org/10.3897/biss.3.37559)

**White**, **A E, Dikow, R B, Baugh, M, Jenkins, A, and Frandsen, P B 2020 Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. *Applications in Plant Sciences*, 8(6). DOI:** [**https://doi.org/10.1002/aps3.11352**](https://doi.org/10.1002/aps3.11352)

**Wilkinson**, **M D, Dumontier, M, Aalbersberg, I J J, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, L B, Bourne, P E, Bouwman, J, Brookes, A J, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, C T, Finkers, R, Gonzalez-Beltran, A, Gray, A J G, Groth, P, Goble, C, Grethe, J S, Heringa, J, ’t Hoen, P A C, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, S J, Martone, M E, Mons, A, Packer, A L, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, M A, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J, and Mons, B 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018. DOI:** [**https://doi.org/10.1038/sdata.2016.18**](https://doi.org/10.1038/sdata.2016.18)

**Willis**, **K J, Paton, A J, and Smith, R J 2018 Science Collections Strategy, 2018-2028. Available at** [**https://brahmsonline.kew.org/Content/Projects/msbp/resources/Training/Kew\_Science\_Collections\_Strategy\_2018-2028.pdf**](https://brahmsonline.kew.org/Content/Projects/msbp/resources/Training/Kew_Science_Collections_Strategy_2018-2028.pdf)

**Wäldchen**, **J and Mäder, P 2018 Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11): 2216–2225. DOI:** [**https://doi.org/10.1111/2041-210X.13075**](https://doi.org/10.1111/2041-210X.13075)

**Zeuss**, **D, Brandl, R, Brändle, M, Rahbek, C, and Brunzel, S 2014 Global warming favours light-coloured insects in Europe. *Nature Communications*, 5(1): 3874. DOI:** [**https://doi.org/10.1038/ncomms4874**](https://doi.org/10.1038/ncomms4874)

**de la Hidalga**, **A N, Rosin, P L, Sun, X, Livermore, L, Durrant, J, Turner, J, Dillen, M, Musson, A, Phillips, S, Groom, Q, and Hardisty, A 2022 Cross-validation of a semantic segmentation network for natural history collection specimens. *Machine Vision and Applications*, 33(3): 39. DOI:** [**https://doi.org/10.1007/s00138-022-01276-z**](https://doi.org/10.1007/s00138-022-01276-z)

**Chinese Virtual Herbarium** 2021 299,000 new specimen data were added to the National Plant Herbarium Resource Bank.. Available at <https://www.cvh.ac.cn/articles/details.php?id=d29c0434> [Accessed December 27, 2021].

**European Commission** 2020 A European strategy for data. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066> [Accessed August 1, 2022].

**Secretariat of the Convention on Biological Diversity** 2016 Biodiversity and the 2030 agenda for sustainable development. Technical Note.. Available at <https://www.cbd.int/development/doc/biodiversity-2030-agenda-technical-note-en.pdf> [Accessed August 1, 2022].