Research Ideas and Outcomes 8: e79187
doi: 10.3897/rio.8.e79187

Research Idea

Sharing taxonomic expertise between natural history collections using image recognition

Michael Greeff[‡], Max Caspers[§], Vincent Kalkman[§], Luc Willemse[§], Barry Dermot Sunderland[|], Olaf Bánki[§], Laurens Hogeweg[§]

‡ Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland
§ Naturalis Biodiversity Center, Leiden, Netherlands
| ETH Library Lab, Zürich, Switzerland

Corresponding author: Michael Greeff (greeffm@ethz.ch)

Reviewed v 1

Academic editor: Editorial Secretary

Received: 10 Dec 2021 | Accepted: 25 Jan 2022 | Published: 01 Mar 2022

Citation: Greeff M, Caspers M, Kalkman V, Willemse L, Sunderland BD, Bánki O, Hogeweg L (2022) Sharing taxonomic expertise between natural history collections using image recognition. Research Ideas and Outcomes 8: e79187. https://doi.org/10.3897/rio.8.e79187

Abstract

Natural history collections play a vital role in biodiversity research and conservation by providing a window to the past. The usefulness of the vast amount of historical data depends on their quality, with correct taxonomic identifications being the most critical. The identification of many of the objects of natural history collections, however, is wanting, doubtful or outdated. Providing correct identifications is difficult given the sheer number of objects and the scarcity of expertise. Here we outline the construction of an ecosystem for the collaborative development and exchange of image recognition algorithms designed to support the identification of objects. Such an ecosystem will facilitate sharing taxonomic expertise among institutions by offering image datasets that are correctly identified by their in-house taxonomic experts. Together with openly accessible machine learning algorithms and easy to use workbenches, this will allow other institutes to train image recognition algorithms and thereby compensate for the lacking expertise.

Keywords

Digitization, image recognition, taxonomic expertise, herbaria, natural history collections

Overview and background

Worldwide there are thousands of repositories housing natural history collections (Hobern et al. 2020) which are aggregations of preserved (parts of) biological objects. Repositories range from large national institutes with millions of specimens stored in multistory warehouses to smaller, sometimes privately owned collections. The importance of natural history collections has been highlighted from different angles in numerous papers, editorials or book chapters, for example by Suarez and Tsutsui (2004), Bakker et al. (2020), National Academies of Sciences, Engineering, and Medicine (2020), or Raes et al. (2020). The primary types housed in these collections together with published descriptions form the foundation of binomial nomenclature, enabling the anchoring of scientific names to verifiable evidence. Natural history collections also form the foundation for taxonomic classification to determine taxonomic units, such as species and higher taxon boundaries and circumscription. The collections themselves become increasingly important as windows to the past which allow us to study the impact of the Anthropocene on biodiversity (Meineke et al. 2018). The critical role natural history collections play in our society also becomes evident from the steady flow of researchers visiting repositories and the large number of research papers which use natural history collections as a primary source of data with at present nearly three peer-reviewed articles relying on data from the Global Biodiversity Information Facility GBIF being published every day (www.gbif.org/literature-tracking).

Taxonomic identifications guarantee collection accessibility

To make full use of natural history collections, both their physical and digital visibility and accessibility are crucial. Physical accessibility is linked to the degree of management applied to collections (for an overview of collection management levels, see McGinley (1993) or Woodburn et al. (2019)). Besides basic requirements such as climatic and sanitary conditions and ensuring minimal risks of damage, key requirements for physical accessibility are the level of identification and a transparent classification system. The Linnaean style scientific name is a key element for both physical access to specimens and online searches in biodiversity related research. Not surprisingly, the scientific name, often through the accepted taxon name, forms the central entity in most data models of biodiversity information systems to which all other information is linked. Because of the crucial role a taxon name plays in providing access to biodiversity related information, the quality of identifications that lead up to a taxon name is, or at least should be, equally important.

In most repositories, collections cover large parts of the biodiversity often from all bioregions of the world. The larger the taxonomic and geographic scope of a collection the more taxonomic expertise and working time is required for its identification. For quite a while, however, there has been a trend for taxonomy to receive less and less attention in the curricula of universities, and positions in public institutions incorporating traditional taxonomy were filled with staff with no or only little taxonomic expertise. This trend, coined the taxonomic impediment (Hoagland 1996, Hopkins and Freckleton 2002), led to a

diminution of taxonomic expertise available at repositories with natural history collections and a decline in their capacity to properly identify newly acquired specimens and update identification of existing specimens.

Likewise, the degree of digital data capturing not only depends on capacity and funding but to a large degree also on the systematic organization of a collection, which can only be done if specimens have proper taxonomic identifications. In line with this, the Minimum Standard for Digital Specimens (MIDS), which was developed for the Distributed System of Scientific Collections DiSSCo (www.dissco.eu; Hardisty et al. 2020), considers the taxonomic identification a basic requirement with regard to digitization priorities (Hardisty 2019). Digitizing unidentified specimens seems hardly useful for most biodiversity information usages other than taxonomy itself, and digitizing wrongly identified specimens carries risk. As a result, digitization projects tend to focus on the well-sorted and thus well-known organisms, thereby neglecting large parts of collections and creating significant shortcomings and biases in our understanding of the past and present biodiversity (Troudet et al. 2017).

Image recognition to the rescue

As discussed in the previous sections, taxonomic knowledge is distributed very unevenly and resources for taxonomic work are scarce. For many years, there have been calls for collaboration between taxonomists and specialists in artificial intelligence, machine learning, and pattern recognition to develop automated systems capable of conducting high-throughput identification of biological specimens (Gaston and O'Neill 2004, MacLeod et al. 2010, Wäldchen et al. 2018, Høye et al. 2021). Once trained, these systems learn to distinguish objects and correctly classify them by deducing rules from a set of training data, analogous to a human brain (Mitchell 1997). Image recognition is a powerful tool to reduce the manual taxonomic workload and could be part of the solution. In this way, taxonomists will be freed from repetitive work concerning common species and put their expertise to optimal use. By using automated taxonomic identification systems, collections can upscale both their available knowledge and the range of potential staff working in collections, be it paid employees, untrained students or volunteers.

Especially in the context of national and international digitization initiatives such as DiSSCo, the Integrated Digitized Biocollections iDigBio (www.idigbio.org, Matsunaga et al. 2013) or the Swiss natural history collections network SwissCollNet (Frick et al. 2019), millions of images will be created on the one hand, and taxonomic expertise will be necessary on the other hand. The ideal solution would be a machine learning solution capable of identifying all known species of the world at a high accuracy. Existing solutions such as iNaturalist (www.inaturalist.org) or Observation.org (https://waarneming.nl/apps/obsidentify) are aiming to identify all current living organisms, but they still have a strong bias for certain organismal groups and rely on quality checks by a community of human experts (Unger et al. 2020). Collection staff, however, have different needs. They want a solution focused on specimens mounted in a fixed position often already organized in taxonomic groups, for example by class or order, or originating from a geographically limited area. Compared to the current practice of sending specimens to specialists residing

in institutes around the world, identification by iNaturalist would already be a much faster solution. Yet, iNaturalist still relies on a community of other users for verification which can delay the final identification by hours or days. For efficient sorting of large numbers of specimens, collection staff therefore need identifications at a very high accuracy and within seconds. In addition, collection staff often face opposite scenarios from uninformed nature lovers in the field: they do not need identifications of the commonly observed taxa, but of taxa that are less prominent in our everyday life, be it because of their lack of "beauty" or their secretive lifestyle, etc. In natural history collections, such taxa might exist in substantial numbers as collectors prefer the rare and hard to find objects. Image recognition tools trained on preferences of the average nature lover can be expected to have a strong bias for the common and to score badly on the groups found in collections (Valan 2021).

Although machine learning solutions are getting ever more powerful and capable of identifying diverse objects, a single universal machine learning model for all known biological taxa is still technically challenging and costly. As a reasonable solution for the time being, collection staff therefore need machine learning tools focusing on subsets of biodiversity such as organisms from limited geographical areas and/or limited taxonomic groups. For instance, machine learning models have been developed for British ground beetle species (Hansen et al. 2019), for Palearctic butterfly species (Dhall et al. 2020, Sunderland 2020), or for closely related families of mosses (Schuettpelz et al. 2017). The authors stress that AI solutions are currently still in their initial stages and need to be treated with caution. There are many challenges such as incomplete sets of training data, geographic biases in collections or other defects. Nevertheless, as AI solutions are currently improving at a tremendous rate, the authors believe this is the right time to start integrating AI in collection management procedures and to learn from any initial obstacles.

Automated identifications are transparent and reproducible

Recent studies proved that machine identifications have become almost as accurate as identifications done by human experts in quite a few groups (in benthic macroinvertebrates (Ärje et al. 2020), in Diptera: Chironomidae (Milošević et al. 2020), in Diptera and Coleoptera (Valan et al. 2019), in dinoflagellates (Culverhouse et al. 2003)). Taxonomic identifications by human experts do not necessarily need to be better than machine identifications. Diverse processes and a wide range of people may be involved in the identification of each specimen in a natural history collection. Given the crucial role correct taxonomic identifications play in providing access in biodiversity related research, one should assume that a transparent evaluation system for identifications exists. Traditionally, the quality of identifications is deduced from the name of the person who performed the identification which is mentioned on an identification label. However, attaching identification labels stating details on the determiner and the taxon is time consuming and often this is omitted rendering the person and the provenance of the identification process obscure. As identifications have been carried out routinely by collection staff ranging from technicians to curators and by visiting naturalists ranging from early-stage novices to world specialists, interpreting the quality of previous identifications is far from easy. A study by Freitas et al. (2020) showed more than 23% of the Auchenipteridae fish records in GBIF (www.gbif.org;

Edwards 2004) and in Brazil's SpeciesLink Network (www.splink.org.br) to have inaccurate taxonomic information (this can include everything from outdated information to misidentifications). Similarly, Goodwin et al. (2015) evaluated 4,500 specimens of African gingers from 40 herbaria in 21 countries and found 58% of the specimens to have wrong names.

In contrast to identifications done by human experts, machine identifications not only deliver taxonomic names, but also metadata about the probability of the determination, the range of taxa considered, the version of the application, and other parameters. Machine determinations therefore are quantifiable, transparent, and reproducible by anyone (the data management techniques involved fall under the term provenance which help reproduce, trace, assess, understand, and explain models and how they were constructed). As natural history collections data are increasingly used in statistical modeling of environmental changes and large datasets are assembled from different repositories, transparent identifications become ever more important (Souza et al. 2021).

Objectives

An automated image recognition ecosystem

The authors envision the establishment of a machine learning ecosystem for natural history collections which allows the sharing of existing models, image datasets and know-how between institutions and collection personnel. An avant-garde of a few experienced institutions shall develop the necessary core modules in machine learning, which can easily be re-trained by other institutions to serve their individual needs. This ecosystem should rest on four pillars:

1. a central library of machine learning algorithms and associated applications (e.g. mobile apps)
2. a central library of available expert validated training datasets, be it the images themselves or simply the information where to find these images
3. a digital workbench that allows even inexperienced users to customize existing machine learning solutions to their individual needs
4. a user forum for the discussion of problems and the coordination of next steps, for the evaluation, testing and implementation of novel technologies, etc.

Deep learning

Feature extractor. Deep learning models (Szegedy et al. 2015, Guo et al. 2016), the most popular in machine learning nowadays, consist of several parts, two of which are particularly important in the present context: the feature extraction network (short: feature extractor), which is sometimes also referred to as backbone, and the classifier. Well-known examples of feature extractor networks are VGG (Simonyan and Zisserman 2014), Inception (Szegedy et al. 2016) and ResNet (He et al. 2016). The feature extractor is the core of any deep learning model as it recognizes features (properties) in the signal it

analyses. In case of images, the feature extractor would recognize shapes, colors, patterns etc. Feature extractors can readily be adapted to analyze other classes of objects as long as these show similar features (Tajbakhsh et al. 2016). For instance, a feature extractor trained on images of beetles can be used as well to identify images showing true bugs, cockroaches, and other morphologically similar insects. Feature extractors will therefore rarely be trained de novo, but rather be recycled in various similar contexts. The training of the feature extractor requires expert IT-knowledge, computing facilities and time and is usually done by bioinformaticians at larger institutions. An important development in recent years represents the so-called task-independent feature extractor training, in which also unlabeled images (e.g., images with unknown taxonomy) are used to extract useful features. Learning with unlabeled images is part of the field of unsupervised machine learning. A surge of recent papers (Chen et al. 2020, Caron et al. 2020) have shown that using a large number (billions) of unlabeled images can reduce the number of labeled images needed to achieve high recognition accuracy. This opens the possibility of making use of the large volumes of unlabeled material that are present in museum collections.

Classifier. The feature extractor does not relate the resulting categories to explicit human concepts such as animals, plants, or cars. For this, the machine learning model relies on a classifier network, which associates the output of the feature extractor with names and concepts (i.e., "classes"). In the natural history context, for instance, the classifier would associate certain features with a family of plants, a species of beetle etc. Classifiers can be easily (re)trained, with regard to time, computing power and experience of the user (e.g., see Valan et al. (2019) for a study in insect recognition). If for example a model existed for the Brassicaceae of Northern America, a simple retraining of the classifier might suffice to adapt this model to the Brassicaceae of Europe. This so-called transfer learning is a central deep learning concept that dramatically speeds up the training of new models and often leads to performance improvements, such as higher identification accuracy (Yosinski et al. 2014).

Algorithms. Machine learning models make predictions and are trained in a particular way and with a particular dataset as described above. Using models in practice often involves additional functionality. The complete process from image(s) to identifications can generally be described as an *algorithm*. Besides the models themselves, algorithms contain pre- and post-processing functionality that cannot be easily fitted into the model formalism of a feature extractor and a classifier. An example of pre-processing is explicitly localizing the organism in the picture before identification. Examples of post-processing are combining multiple predictions into one and combining image recognition models with species distribution models.

Central Library of Algorithms. To facilitate the exchange of these models and algorithms, the authors suggest setting up a Central Library of Algorithms (Fig. 1). This library would follow open data and open-source policies, provide search functionality, and allow other institutions to use the models for automatic identification after they are made available (deployed) through identification web services. For efficient utilization, the library should offer discovery as well as access and download services, a performant scalable infrastructure, an API (application programming interface) supporting machine to machine

communication, and some tracking of use and accreditation services (DOIs). A comparable library or repository has been established under the name BioImage Model Zoo (www.bioimage.io), which offers community driven AI models for the analysis of mostly cellular images. Setting up a leaderboard for best performing feature extractors and algorithms will create an incentive for the generation of better training datasets and for technological improvements. For machine learning models to be useful in the daily collection work, further accompanying applications and web services are necessary, which could be shared on the central library as well. They offer user interfaces to apply the models and allow, for instance, accessing the camera and image gallery of the mobile device, cropping the images, and uploading them to the machine learning model as well as displaying the results.

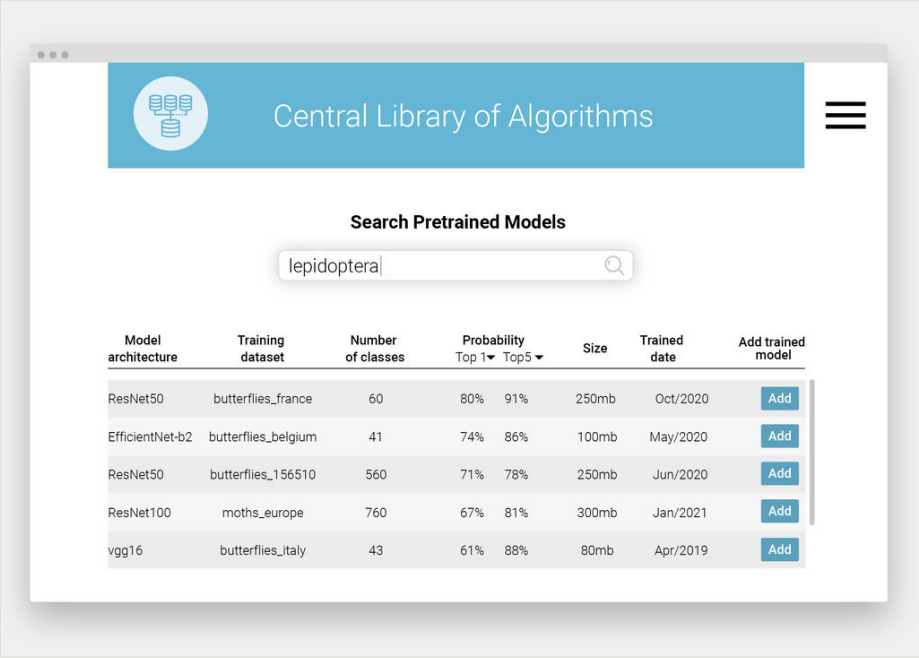

Figure 1. doi

In the Central Library of Algorithms, natural history collection staff will select algorithms (feature extractors, models, etc.) that are most appropriate for the identification of their target organisms and add them to the workbench. The current figure shows a mock-up.

Central Library of Datasets

Further sharing of taxonomic knowledge would be provided through a Central Library of Datasets. This library would be a system to access a collection of public image datasets for images that are suitable for supporting large-scale centralized training of feature extractors and local training of classifiers at the individual institutions (Fig. 2). As in the Central Library of Algorithms, this library should offer various services for users, a scalable infrastructure with interfaces and tracking functionalities. For biodiversity images, datasets may be found

on GBIF and/or iDigBio. Repositories with a more general focus might be the Research Data Alliance (Parsons 2013) or the European Open Science Cloud (www.eosc-portal.eu). In the area of machine learning, the online community of data scientists often share their datasets on Kaggle (www.kaggle.com). With so many collections digitizing their holdings worldwide, the number of images of specimens is growing at an impressive rate (Tegelberg et al. 2014). However, not all images are appropriate for training as many collections digitize their holdings without prior verification of the taxonomic identifications. Using these images could decrease the quality of the model but can still be used in unsupervised learning (Chen et al. 2020, Caron et al. 2020). The authors therefore suggest establishing a Central Library of Datasets to access specimen images with high confidence identifications.

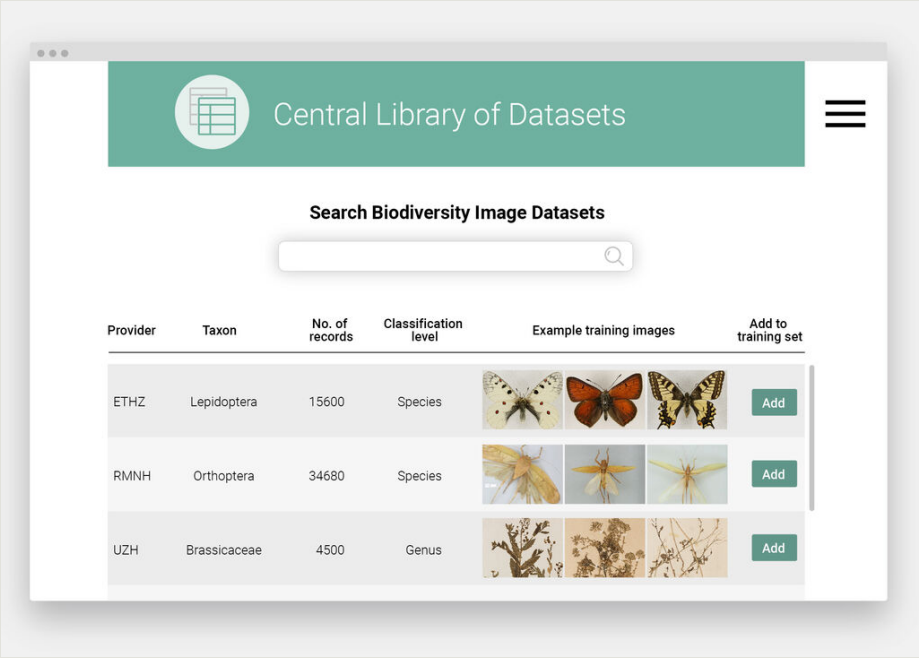

Figure 2. doi

In the Central Library of Datasets, natural history collection staff will find correctly identified images of their target organisms and download the data for training of an individually customized classifier (photos: Lepidoptera by Entomological Collection of ETH Zürich; Orthoptera by Naturalis Biodiversity Center; Brassicaceae by United Herbaria Z+ZT, ZT-00164967, ZT-00167494, ZT-00171530, CC BY-SA 4.0). The current figure shows a mock-up.

The uploaded images shall be collected in a dataset, in this context defined as a fixed curated list of images with additional metadata such as the name of the taxon, geographic coordinates and information on the probability of the identification. The Central Library of Datasets will reference existing public datasets such as GBIF and/or iDigBio. Over time, this can encourage collection staff and collection users to generate and publish their own

datasets on public portals, possibly remedying biases and shortcomings in existing datasets (this could be done as 'data papers', see Chavan and Penev (2011)or Costello et al. (2013)). To this end, it is necessary that relevant criteria for images to qualify for training data are defined. At the BioDiversity_Next conference in Leiden in 2019 for instance, the association 'Biodiversity Information Standards TDWG' (www.tdwg.org) initiated the discussion on establishing a 'Deep Learning Standards Interest Group', which could take over this task. In general, the generation of interoperable training datasets will be greatly facilitated if collections start to adopt standards in all areas. The new Catalogue of Life, for example, aims to provide an authoritative nomenclature and taxonomic foundation that could function as a clearinghouse covering all scientific names of biological taxa worldwide and allows for seamless data exchange between institutions following this standard (www.catalogueoflife.org, Bánki et al. 2018). With regard to the accuracy of taxonomic metadata, the quality of initial identifications done by human experts will be relevant. The quantification of expert knowledge as suggested by Caley et al. (2013) could prove to be a feasible solution and be used for algorithms that aggregate information and annotations (Simpson and Roberts 2015) or simply as part of one single set of criteria to define the confidence level of identifications.

Digital workbench

Retraining an existing model to a new group of organisms is easy – for IT specialists. The average collection manager would most likely struggle with the necessary procedures. The authors therefore propose the establishment of a digital workbench for machine learning (e.g., Google AutoML, Microsoft Azure), which allows non-experts to curate datasets (e.g., completing taxonomic or geographic information) and retrain existing models for their individual purposes. Ideally, the workbench should have a graphical user interface. Users could import existing feature extractors and further algorithms from the Central Library of Algorithms, and training data from the Central Library of Datasets (Fig. 3). The training of the model could be started with a few clicks, and in the end the workbench would provide a standardized evaluation of the new model informing about the accuracy and about further relevant performance indicators. When the performance is sufficient, the collection manager imports the algorithm into a mobile app or a web service, and finally may even publish it again to the Central Library of Algorithms for others to use.

User forum

Critical readers might consider this vision too idealistic. And it is true, for everything to work properly, many prerequisites just need to be right: a feature extractor needs to be available, appropriate images need to exist, the workbench and the applications need to work flawlessly. The authors therefore propose a further measure: the establishment of a user forum. On this forum, users can post their wishes, discuss shortcomings, and interact with more experienced institutions and providers of machine learning solutions. The user forum should thus serve as a marketplace where collection managers search for technological expertise and assistance and in return offer image datasets and taxonomic expertise. As a result, this user forum should guarantee that over time well identified image datasets and machine learning models become available for most groups of organisms, as well those

that have been neglected so far. In addition, this will be the place to discuss and find strategies for shortcomings of the AI solutions related to inherent collection biases, be they geographical, cultural, taxonomical or other.

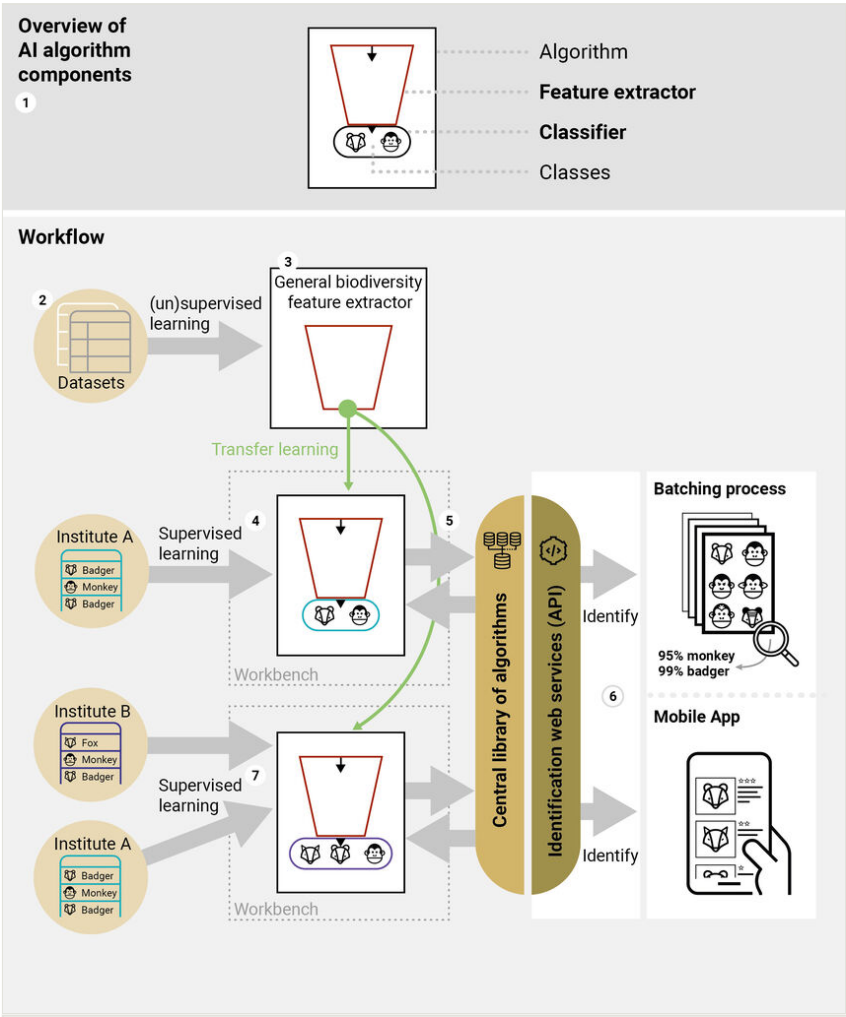

Figure 3. doi

Sharing of taxonomic knowledge between institutes. (1) Each algorithm contains two basic components: the feature extractor and the classifier. (2) The Central Library of Datasets allows the user to browse through all available images of collection objects; (3) based on all available images, a regularly updated central feature extractor is created and published; (4) custom made algorithms can relatively easily be created by building a classifier based on a selection of taxa from the central library and combining this with the central feature extractor; (5) newly created algorithms together with their metadata (probability & information on content) are published through a web service in the Central Library of Algorithms (6) and can be used through the Identification web services (API) either for batch processing of images or through a mobile app. Models can be easily extended by other institutions by combining data sources (7).

Use Cases

Accessing unsorted collection holdings. Most collections accumulate considerable holdings of biological specimens which remain unidentified due to a lack of time or in-house taxonomic expertise. These specimens may be stored as singletons or as groups in boxes, either preliminarily sorted by higher taxonomic groupings (order, family) or by geographic region, or they may be completely mixed. In recent years, especially larger institutions have therefore started to database their holdings at the storage unit level (i.e., by the units in which specimens are stored, like drawers, jars, or boxes). In insect collections, for instance, whole drawers are being imaged and published online to be browsed through by the entomological community (Olsen 2015, Mantle et al. 2012). In this setting, machine learning applications could follow the image capturing step by splitting the image into segments, each of which features an individual specimen (Fig. 4).

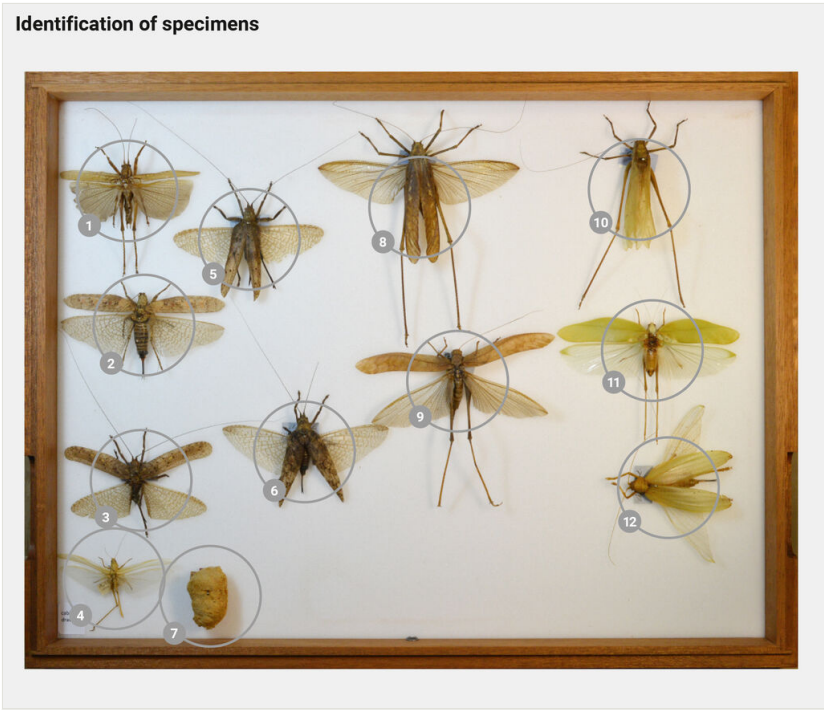

Figure 4. doi

Algorithms recognize and number individual specimens in a drawer of unsorted items. The insect drawer is from the Oxford University Museum of Natural History. The current figure shows a mock-up.

Machine learning applications would then recognize the taxonomic identity of each specimen (Table 1). And finally, the collection staff can add this information to the corresponding specimen on the image, either physically using identification labels and/or digitally in the object level registration. As a result, users will be able to search for

taxonomic information of individual specimens rather than of whole drawers, and collection staff will be able to efficiently sort and integrate these specimens into their main collection (Fig. 5).

Table 1.

Automated recognition applications identify the specimens to lower taxonomic levels and inform about the probability of the identifications.

Drawer number	Specimen number	Family	Subfamily	Probability
BE.2286032	1	Tettigoniidae	Conocephalinae	95%
BE.2286032	2	Tettigoniidae	Pseudophyllinae	85%
BE.2286032	3	Tettigoniidae	Pseudophyllinae	95%
...	...	...	...	...

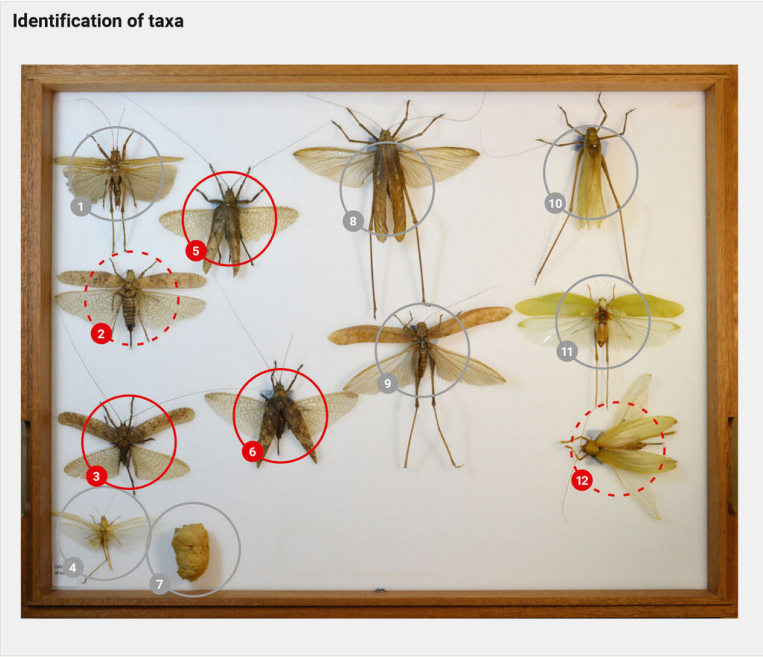

Figure 5. doi

Non-expert collection staff easily find and afterwards sort specimens by taxon (line color) and by accuracy of the identification (line type). The insect drawer is from the Oxford University Museum of Natural History. The current figure shows a mock-up.

Transparent identifications in mass digitization. Bringing down costs and time spent per treated item is of paramount importance when digitizing natural history collections (Blagoderov et al. 2012). In addition to introducing industrial style processes such as conveyor belts or division of labor, and recruiting volunteer workers, costs are often cut by omitting expensive work steps. In particular, natural history institutions rarely verify the

taxonomic identifications of specimens prior to databasing (Scoble 2010, Oever and Gofferje 2012). If specimens are imaged, image recognition applications offer cheap and scalable means to improve data quality (Fig. 6). Importantly, the recognition step can be repeated at any given point in time, thus allowing not only for verification of the past identification by humans, but also for regular updates of machine identifications in the future. When a taxon is split or synonymized, for instance, the algorithm would change the taxonomic identity of the specimen in the collection management system and inform the collection manager about necessary changes in the physical collection.

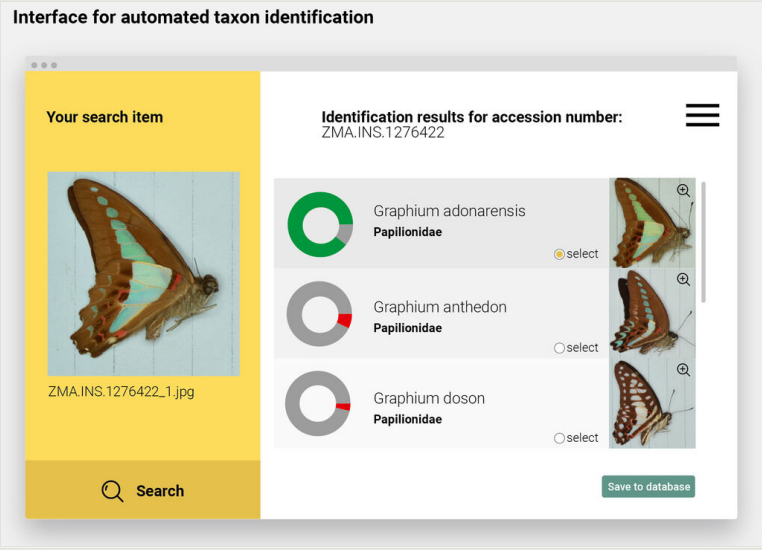

Figure 6. doi

Mock-up of an interface for automated taxon identification. Naturalis holds over 500.000 specimens of unmounted, unsorted and often unidentified, papered butterflies and moths that were collected mostly in Europe and Asia over the past 200 years. In early 2016, Naturalis embarked on a 10-year-project to digitally identify all these specimens with the help of dedicated volunteers (Caspers et al. 2019). Specimens are unpacked, photographed, had their label data registered and then repacked, still unmounted, for long-term storage. Specimen images were then dragged and dropped into a web-based interface to get a near-instant response with multiple predictions about the taxonomic identity including probability values.

Challenges ahead

A decade ago, the idea of using image recognition to share taxonomic knowledge between natural history collections would have seemed far-fetched. From a technical point of view this is no longer the case as is demonstrated by widely used field apps like iNaturalist, ObsIdentify (Schermer and Hogeweg 2018) or PlantNet (Goëau et al. 2011, Goëau et al. 2012). Based on expert validation, these apps have taken their place among traditional field guides and even started to replace the role of experts in identifying common species

observed outdoors. The challenges for the large-scale use of image recognition in collections as described in this paper are primarily organizational and concern standardization, coordination, (re-)use of existing and development of new infrastructure components, and rallying a community of contributors and users. The premise of the outlined proposal is that collections of all sorts and sizes can have a streamlined collaboration.

Algorithms. Even though most challenges ahead are organizational, machine learning still harbors some technical challenges of its own (e.g., Høye et al. 2021). One of them is that identifications will not always be correct. Incorrect identifications with a low computed probability are relatively easy to address; they can be either discarded or the probability can be recorded along with the identification in a collection management system for future reference. Incorrect identifications with a high computed probability are a bigger problem. It can have multiple causes (Nguyen et al. 2015, Hein et al. 2019), but it occurs mainly when the dataset used to train the model differs significantly from the dataset that needs to be identified. For example, the method of preparation can be different between collections (e.g., open vs closed butterfly wings), the true taxon of a specimen being identified is not part of the original training database or there are taxa in which the distinction between species is quite difficult because of the range in biological variation. Without special measures, the output of the algorithm can be unpredictable. This so-called *open world* issue is well known in machine learning (Bendale and Boult 2016, Geng et al. 2021), but further study is needed to understand the difference between aleatoric (due to noise) and epistimic (due to lack of data) uncertainty in biodiversity machine learning models (Hüllermeier and Waegeman 2021).

Standardization. One organizational endeavor is to further standardize and accelerate the digitization of natural history collections, ensuring that the images and metadata can be readily applied for image recognition. This applies to both taxonomical and geographical annotations. Even when no larger infrastructure as envisioned in this paper is built, this step is worthwhile and should be addressed by or in close collaboration with TDWG (Wieczorek et al. 2012, Morris et al. 2013). It is of equal importance that the output of the models is standardized, considering aspects of accuracy and information about taxa included. It should be legible by faunistic databases, analogous to BibTeX in libraries (www.bibtex.org), as well as by the wide variety of collection management systems used in collections. As with manual identifications, it is necessary to record the identifier. Therefore, a commonly accepted and quotable versioning and provenance system for machine learning is necessary.

Infrastructure. Another challenge is the ownership and responsibility for the proposed ecosystem. Initially, one or several larger natural history institutions will need to build a large-scale digital infrastructure to allow for the generation, exchange, and application of image recognition models, as well as to provide a platform for a community to engage with one another. The different modules of the infrastructure can be developed by different parties. In addition, the different modules could be a combination of the repurposing of existing infrastructure components and tools and newly developed ones. Recently, a landscape and gap analysis on the automated services, tools, and workflows for extracting

information from images of natural history specimens and their labels was performed (Walton et al. 2020). One could envision a similar exercise for the proposed modules. Collaborations between existing infrastructures like GBIF, Catalogue of Life, Zenodo (https://zenodo.org) amongst others, and initiatives such as DiSSCo and iDigBio could provide a framework for the repurpose of existing tooling and infrastructure components and newly developed ones. Having a standardized framework for storing and evaluating algorithms on well described datasets also provides the opportunity for the machine learning research community to compete on creating the best models in the form of challenges (Joly et al. 2020, Little et al. 2020). In the end, all modules need to fit and work together and be actively and sustainably maintained. The initial development can probably only be achieved through a grant from a national or international science foundation.

Once built, the viability of the machine learning ecosystem for collections depends on the level of contribution from its participants. Collection managers and curators would need to actively focus their capacities at collaborating with experts to identify and digitize collections, resulting in taxonomically validated and properly annotated images. Once shared, they can be used to (re)train image recognition models and benefit the entire community. Especially in the initial phase this will require a level of altruism, as contributing will take time and resources while the benefits will only become clear after a few years. The concept of *give and take* requires momentum and should be stimulated by the collections maintaining the infrastructure, ideally utilizing already existing cross-national collaborations for mobilizing collections and knowledge. Parallels of such a community-driven approach can be found in the Barcode of Life project (www.barcodinglife.org), which allows the exchange of DNA-barcodes between institutes, or OpenML (Vanschoren et al. 2014), which facilitates the exchange and analysis of large datasets.

Acknowledgements

We are especially grateful to Rod Eastwood and Samuel Glauser for their discussion of and feedback on the current text.

References

- Ärje J, Raitoharju J, Iosifidis A, Tirronen V, Meissner K, Gabbouj M, Kiranyaz S, Kärkkäinen S (2020) Human experts vs. machines in taxa recognition. Signal Processing: Image Communication 87: 115917. https://doi.org/10.1016/j.image.2020.115917
- Bakker FT, Antonelli A, Clarke JA, Cook JA, Edwards SV, Ericson PG, Faurby S, Ferrand N, Gelang M, Gillespie RG, Irestedt M, Lundin K, Larsson E, Matos-Maraví P, Müller J, von Proschwitz T, Roderick GK, Schliep A, Wahlberg N, Wiedenhoeft J, Källersjö M (2020) The Global Museum: natural history collections and the future of evolutionary science and public education. PeerJ 8: e8225. https://doi.org/10.7717/peerj.8225

- Bánki O, Döring M, Holleman A, Addink W (2018) Catalogue of Life Plus: innovating the CoL systems as a foundation for a clearinghouse for names and taxonomy. Biodiversity Information Science and Standards 2 https://doi.org/10.3897/biss.2.26922
- Bendale A, Boult TE (2016) Towards Open Set Deep Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) https://doi.org/10.1109/cvpr.2016.173
- Blagoderov V, Kitching I, Livermore L, Simonsen T, Smith V (2012) No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133-146. https://doi.org/10.3897/zookeys.209.3178
- Caley MJ, O'Leary RA, Fisher R, Low-Choy S, Johnson S, Mengersen K (2013) What is an expert? A systems perspective on expertise. Ecology and Evolution 4 (3): 231-242. https://doi.org/10.1002/ece3.926
- Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A (2020) Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint (arXiv: 2006.09882). URL: https://arxiv.org/pdf/2006.09882.pdf
- Caspers M, Willemse L, Miracle EG, van Nieukerken EJ (2019) Butterflies in bags: permanent storage of Lepidoptera in glassine envelopes. Nota Lepidopterologica 42 (1): 1-16. https://doi.org/10.3897/nl.42.28654
- Chavan V, Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12 https://doi.org/10.1186/1471-2105-12-s15-s2
- Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G (2020) Big self-supervised models are strong semi-supervised learners. arXiv preprint (arXiv:2006.10029). URL: https://arxiv.org/pdf/2006.10029.pdf
- Costello MJ, Michener WK, Gahegan M, Zhang Z, Bourne PE (2013) Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution 28 (8): 454-461. https://doi.org/10.1016/j.tree.2013.05.002
- Culverhouse P, Williams R, Reguera B, Herry V, González-Gil S (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Marine Ecology Progress Series 247: 17-25. https://doi.org/10.3354/meps247017
- Dhall A, Makarova A, Ganea O, Pavllo D, Greeff M, Krause A (2020) Hierarchical Image Classification using Entailment Cone Embeddings. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) https://doi.org/10.1109/cvprw50498.2020.00426
- Edwards JL (2004) Research and Societal Benefits of the Global Biodiversity Information Facility. BioScience 54 (6): 485-486. https://doi.org/10.1641/0006-3568(2004)054[0486:rasbot]2.0.co;2
- Freitas TS, Montag LA, De Marco P, Hortal J (2020) How reliable are species identifications in biodiversity big data? Evaluating the records of a neotropical fish family in online repositories. Systematics and Biodiversity 18 (2): 181-191. https://doi.org/10.1080/14772000.2020.1730473
- Frick H, Stieger P, Scheidegger C (2019) SwissCollNet – A National Initiative for Natural History Collections in Switzerland. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37188
- Gaston K, O'Neill M (2004) Automated species identification: why not? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359 (1444): 655-667. https://doi.org/10.1098/rstb.2003.1442

- Geng C, Huang S, Chen S (2021) Recent Advances in Open Set Recognition: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (10): 3614-3631. https://doi.org/10.1109/tpami.2020.2981604
- Goëau H, Boujemaa N, Joly A, Selmi S, Bonnet P, Mouysset E, Joyeux L, Molino J, Birnbaum P, Bathelemy D (2011) Visual-based plant species identification from crowdsourced data. Proceedings of the 19th ACM international conference on Multimedia - MM '11813-814. https://doi.org/10.1145/2072298.2072472
- Goëau H, Bonnet P, Barbe J, Bakic V, Joly A, Molino J, Barthelemy D, Boujemaa N (2012) Multi-organ plant identification. Proceedings of the 1st ACM international workshop on Multimedia analysis for ecological data - MAED '12 https://doi.org/10.1145/2390832.2390843
- Goodwin Z, Harris D, Filer D, Wood JI, Scotland R (2015) Widespread mistaken identity in tropical plant collections. Current Biology 25 (22). https://doi.org/10.1016/j.cub.2015.10.002
- Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew M (2016) Deep learning for visual understanding: A review. Neurocomputing 187: 27-48. https://doi.org/10.1016/j.neucom.2015.09.116
- Hansen OP, Svenning J, Olsen K, Dupont S, Garner B, Iosifidis A, Price B, Høye T (2019) Species-level image classification with convolutional neural network enables insect identification from habitus images. Ecology and Evolution 10 (2): 737-747. https://doi.org/10.1002/ece3.5921
- Hardisty A (2019) Provisional Data Management Plan for DiSSCo infrastructure. ICEDIG Deliverable D6.6 https://doi.org/10.5281/zenodo.3532937
- 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, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e54280
- Hein M, Andriushchenko M, Bitterwolf J (2019) Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR):41-50. https://doi.org/10.1109/cvpr.2019.00013
- He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR):770-778. https://doi.org/10.1109/cvpr.2016.90
- Hoagland KE (1996) The taxonomic impediment and the convention on Biodiversity. Association of Systematics Collections Newsletter 24: 61-62.
- Hobern D, Paul DL, Robertson T, Groom Q, Thiers B, Asase A, Luo M, Semal P, Woodburn M, Zschuschen E (2020) Advancing the Catalogue of the World's Natural History Collections. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59324
- Hopkins GW, Freckleton RP (2002) Declines in the numbers of amateur and professional taxonomists: implications for conservation. Animal Conservation 5 (3): 245-249. https://doi.org/10.1017/s1367943002002299
- Høye T, Ärje J, Bjerge K, Hansen OP, Iosifidis A, Leese F, Mann HR, Meissner K, Melvad C, Raitoharju J (2021) Deep learning and computer vision will transform

entomology. Proceedings of the National Academy of Sciences 118 (2). https://doi.org/10.1073/pnas.2002545117

- Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning 110 (3): 457-506. https://doi.org/10.1007/s10994-021-05946-3
- Joly A, Goëau H, Kahl S, Deneu B, Servajean M, Cole E, Picek L, Ruiz de Castañeda R, Bolon I, Durso A, Lorieul T, Botella C, Glotin H, Champ J, Eggel I, Vellinga W, Bonnet P, Müller H (2020) Overview of LifeCLEF 2020: A System-Oriented Evaluation of Automated Species Identification and Species Distribution Prediction. Lecture Notes in Computer Science342-363. https://doi.org/10.1007/978-3-030-58219-7_23
- Little DP, Tulig M, Tan KC, Liu Y, Belongie S, Kaeser-Chen C, Michelangeli FA, Panesar K, Guha RV, Ambrose BA (2020) An algorithm competition for automatic species identification from herbarium specimens. Applications in plant sciences 8 (6): e11365. https://doi.org/10.1002/aps3.11365
- MacLeod N, Benfield M, Culverhouse P (2010) Time to automate identification. Nature 467 (7312): 154-155. https://doi.org/10.1038/467154a
- Mantle B, LaSalle J, Fisher N (2012) Whole-drawer imaging for digital management and curation of a large entomological collection. ZooKeys 209: 147-163. https://doi.org/10.3897/zookeys.209.3169
- Matsunaga A, Thompson A, Figueiredo R, Germain-Aubrey C, Collins M, Beaman R, MacFadden B, Riccardi G, Soltis P, Page L, Fortes JB (2013) A Computational- and Storage-Cloud for Integration of Biodiversity Collections. Proceedings of the IEEE 9th International Conference on e-Science:78-87. https://doi.org/10.1109/escience.2013.48
- McGinley RJ (1993) Where's the management in collections management? International Symposium and First World Congress on the preservation and conservation of Natural History Collections 3: 309-338.
- Meineke E, Davies TJ, Daru B, Davis C (2018) Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B: Biological Sciences 374: 20170386. https://doi.org/10.1098/rstb.2017.0386
- Milošević D, Milosavljević A, Predić B, Medeiros A, Savić-Zdravković D, Stojković Piperac M, Kostić T, Spasić F, Leese F (2020) Application of deep learning in aquatic bioassessment: Towards automated identification of non-biting midges. Science of The Total Environment 711: 135160. https://doi.org/10.1016/j.scitotenv.2019.135160
- Mitchell T (1997) Machine Learning. McGraw-Hill Education, 414 pp.
- Morris RA, Barve V, Carausu M, Chavan V, Cuadra J, Freeland C, Hagedorn G, Leary P, Mozzherin D, Olson A, Riccardi G, Teage I, Whitbread G (2013) Discovery and publishing of primary biodiversity data associated with multimedia resources: The Audubon Core strategies and approaches. Biodiversity Informatics 8 (2): 185-197. https://doi.org/10.17161/bi.v8i2.4117
- National Academies of Sciences, Engineering, and Medicine (2020) Biological Collections: Ensuring Critical Research and Education for the 21st Century. The National Academies Press, Washington, 245 pp. [ISBN ISBN 978-0-309-49853-1] https://doi.org/10.17226/25592
- Nguyen A, Yosinski J, Clune J (2015) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR):427-436. https://doi.org/10.1109/cvpr.2015.7298640

- Oever JPvd, Gofferje M (2012) ‘From Pilot to production’: Large Scale Digitisation project at Naturalis Biodiversity Center. ZooKeys 209: 87-92. https://doi.org/10.3897/zookeys.209.3609
- Olsen E (2015) Museum specimens find new life online. New York Times.
- Parsons M (2013) The research data alliance: Implementing the technology, practice and connections of a data infrastructure. Bulletin of the American Society for Information Science and Technology 39 (6): 33-36. https://doi.org/10.1002/bult.2013.1720390611
- Raes N, Casino A, Goodson H, Islam S, Koureas D, Schiller E, Schulman L, Tilley L, Robertson T (2020) White paper on the alignment and interoperability between the Distributed System of Scientific Collections (DiSSCo) and EU infrastructures - The case of the European Environment Agency (EEA). Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e62361
- Schermer M, Hogeweg L (2018) Supporting citizen scientists with automatic species identification using deep learning image recognition models. Biodiversity Information Science and Standards 2 https://doi.org/10.3897/biss.2.25268
- Schuettpelz E, Frandsen PB, Dikow RB, Brown A, Orli S, Peters M, Metallo A, Funk VA, Dorr LJ (2017) Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity data journal 5: e21139. https://doi.org/10.3897/BDJ.5.e21139
- Scoble M (2010) Rationale and Value of Natural History Collections Digitisation. Biodiversity Informatics 7 (2): 77-80. https://doi.org/10.17161/bi.v7i2.3994
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (*arXiv:1409.1556*). URL: https://arxiv.org/pdf/1409.1556.pdf(2014.pdf
- Simpson E, Roberts S (2015) Bayesian Methods for Intelligent Task Assignment in Crowdsourcing Systems. Decision Making: Uncertainty, Imperfection, Deliberation and Scalability1-32. https://doi.org/10.1007/978-3-319-15144-1_1
- Souza R, Azevedo L, Lourenço V, Soares E, Thiago R, Brandão R, Civitarese D, Vital Brazil E, Moreno M, Valduriez P, Mattoso M, Cerqueira R, Netto MS (2021) Workflow provenance in the lifecycle of scientific machine learning. Concurrency and Computation: Practice and Experience https://doi.org/10.1002/cpe.6544
- Suarez AV, Tsutsui ND (2004) The Value of Museum Collections for Research and Society. BioScience 54 (1): 66-74. https://doi.org/10.1641/0006-3568(2004)054[0066:tvomcf]2.0.co;2
- Sunderland B (2020) BioDex: Tales from an Adventure into App Development. ETH Library Lab Blog URL: https://www.librarylab.ethz.ch/tales-from-an-adventure-into-app-development/
- Szegedy C, Wei Liu, Yangqing Jia, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR):1-9. https://doi.org/10.1109/cvpr.2015.7298594
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR):2818-2826. https://doi.org/10.1109/cvpr.2016.308
- Tajbakhsh N, Shin J, Gurudu S, Hurst RT, Kendall C, Gotway M, Liang J (2016) Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine

Tuning? IEEE Transactions on Medical Imaging 35 (5): 1299-1312. https://doi.org/10.1109/tmi.2016.2535302

- Tegelberg R, Mononen T, Saarenmaa H (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. Taxon 63 (6): 1307-1313. https://doi.org/10.12705/636.13
- Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F (2017) Taxonomic bias in biodiversity data and societal preferences. Scientific Reports 7 (1): 1-4. https://doi.org/10.1038/s41598-017-09084-6
- Unger S, Rollins M, Tietz A, Dumais H (2020) iNaturalist as an engaging tool for identifying organisms in outdoor activities. Journal of Biological Education1-11. https://doi.org/10.1080/00219266.2020.1739114
- Valan M, Makonyi K, Maki A, Vondráček D, 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. https://doi.org/10.1093/sysbio/syz014
- Valan M (2021) Automated image-based taxon identification using deep learning and citizen-science contributions. Doctoral Dissertation. Department of Zoology, Stockholm University
- Vanschoren J, van Rijn J, Bischl B, Torgo L (2014) OpenML. ACM SIGKDD Explorations Newsletter 15 (2): 49-60. https://doi.org/10.1145/2641190.2641198
- Wäldchen J, Rzanny M, Seeland M, Mäder P (2018) Automated plant species identification—Trends and future directions. PLOS Computational Biology 14 (4). https://doi.org/10.1371/journal.pcbi.1005993
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e57602
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Woodburn M, Vincent S, Hardy H, Valentine C (2019) Join the Dots: Adding collection assessment to collection descriptions. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37200
- Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? arXiv preprint (arXiv:1411.1792). URL: https://arxiv.org/pdf/1411.1792.pdf