APPLICATION
The Global Pollen Project: a new tool for pollen identification and the dissemination of physical reference collections
Andrew C. Martin* and William J. Harvey
Long-Term Ecology Laboratory, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Summary
1. The study of fossil and modern pollen assemblages provides essential information about vegetation dynamics in space and time. A major methodological component of these studies is the identification of pollen grains to plant family, genus and species. This identification is achieved through the use of pollen keys and reference collections of physical specimens, which are regional in scope, disparate and incomplete, slowing the identification process. Reference material is also held in museums and research institutions, where access can be limited. Identification is particularly challenging for those new to the field, such as graduate students.
2. To aid the identification of pollen grains, and provide virtual access to reference material, we present a new online tool: the Global Pollen Project (https://globalpollenproject.org). The project aims to enable people to share and identify pollen grains and through this, will create an open, free and accessible reference library for pollen identification.
3. The online tool has been developed as an open, peer-reviewed database of global pollen, where content and expertise is crowdsourced from across the world. The tool enables: (i) the submission and identification of unknown pollen grains; (ii) the submission and digitisation of existing ‘physical’ reference collections and (iii) the availability of a free public database of pollen images and their metadata, for use in scientific research and education. The tool connects to external services, including the Global Biodiversity Information Facility and Neotoma Palaeoecology Database, to provide botanical descriptions and occurrence data for each taxon, alongside pollen images and metadata.
4. The database currently holds information for over 1500 species, from Europe, the Americas and Asia. As the collection grows, we envision easier pollen identification, and greater use of the database for novel research on pollen morphology and other characteristics, especially when linked to other palaeoecological databases, such as Neotoma.

Key-words: crowdsourcing, digitisation, dissemination, palynology, pollen, reference-collection

Introduction
Analyses of fossil and modern pollen assemblages provide vital evidence for determining vegetation dynamics in response to environmental change across space and through time. The task of undertaking palynological research can prove challenging, particularly when confronted with the task of identifying grains to genus or species level (Mander & Punyasena 2014). Difficulties arise from the condition and/or preservation of individual grains (Berglund 1986; Bennett & Willis 2002; Pearseall 2015), as well as the availability and accessibility of reference collections with which to compare them. Fast and accurate pollen identification requires a degree of skill that is hard for new researchers to develop if appropriate training and reference material is not at their disposal. Existing pollen reference collections are often kept as collections of slides within research institutions or private collections, which constrains accessibility. Individual researchers and institutions often collect and create their own reference collections that often fail to be shared or archived, and are forgotten, becoming so-called ‘dark data’ (Heidorn 2008). This issue is not only limited to palynology but also prevalent within other taxonomic fields, including collections held within both public and private institutions. This has led to recent efforts to digitise collections from herbaria to natural history collections (see Brooks et al. 2014; Wen et al. 2015; iDigBio 2016). We sought to develop a tool that could address the following key aims:
1. to provide virtual access to palynological reference material derived from all regions of the globe,
2. to aid in the identification of pollen grains via education and accessibility;
3. to foster scientific crowdsourcing to improve palynological studies and its use in biodiversity conservation, climate change and anthropogenic impacts on vegetation; and
4. to integrate with supportive material from other global biodiversity and palaeoecological resources.

*Correspondence author. E-mail: andrew.martin@zoo.ox.ac.uk
Here we discuss the development and release of the Global Pollen Project (GPP), a web-based tool that enables digitisation (the upload and sharing of images and metadata in a common format) of pollen reference material. It addresses the aims of access to material, aiding identification and fostering crowdsourcing, by providing an encompassing, evolving peer-reviewed database of pollen through crowdsourcing content and expertise from across the world. By being ‘global’, we aim to encompass information for all plant taxa around the world, as well as encouraging a world-wide pooling of expertise.

**Application framework**

The GPP has been designed with the purpose of aiding and disseminating palynological research from within the scientific community as well as the wider public sphere. A key drive in the development of the tool, therefore, was the creation of a flexible underlying framework and an intuitive interface that would encourage use from novices as well as experts. The framework was designed so that users can upload images of pollen grains irrespective of their identification status. If a grain is unidentified, the intention is that its identification can be crowdsourced from other users. Alternatively, users that have well-identified collections can upload their collection digitally with detailed labels. These two inputs of pollen grain images come together to form the GPP’s master reference collection.

**TECHNICAL IMPLEMENTATION**

The GPP is a web application built on Microsoft’s .NET Core 1.0 platform with the ASP.NET framework (Microsoft, 2016a), using Entity Framework Core (Microsoft 2016b) with a Microsoft SQL database (Fig. 1). It is currently deployed on Microsoft’s Azure cloud infrastructure (Microsoft 2016c) enabling high availability, routine backup and upscaling of resources as the database and image library grows. The application was developed in a decoupled way, such that the web user interface is interchangeable, and the core is not reliant on a specific technical infrastructure.

**TAXONOMIC SYSTEM AND TAXONOMIC BACKBONE**

Taxonomic organisation of modern and fossil pollen types has been implemented through a hierarchical system relating submitted images to their corresponding plant families, genera and species. New taxa are created automatically when a new user-submitted grain becomes identified, or a slide is digitised for a new taxon. To overcome taxonomic conflicts between collections, we automatically constrain the GPP taxonomy using a dynamic taxonomic backbone. This taxonomy provides an authority on accepted names and synonyms. Any new taxon generated in the GPP must exist in the backbone as an accepted taxon. Synonyms will not be accepted and must be entered in accordance with current convention. As changes are made to the dynamic backbone, these are reflected in the GPP taxonomy. The Plant List (2013), the only complete, working list of vascular plant names, is utilised by the GPP as its default taxonomic backbone. The Plant List contains 1 064 035 scientific species names, of which 350 699 are accepted (The Plant List 2013). Palynological morphotypes are often used as a proxy for the presence and abundance of plant taxa. It is important for the integrity of any palynological backbone to be firmly connected to a botanical list, so that changes and amendments made by the scientific community can be tracked through time. Other current palynological lists lack such tracking, thus cannot be used in this way.

![Fig. 1. Integrated framework of the Global Pollen Project. The main infrastructure and functionality is outlined, alongside connections to external services.](image)
USER SYSTEM, PERMISSIONS AND DATA INTEGRITY

The GPP handles users and their permissions to promote data integrity, and limit abuse. Users must provide and validate an email to create an account, and malicious users can be banned, with their contributions subsequently removed. The submission of unknown grains, and their identification, is open to any registered user. If a user disagrees with past identifications of a grain, whether it has gained a confirmed identity or not, they can add another identification and the identity will be recalculated. In this way, the quality of the data increases as more people use it. As the tool is further developed, an improved rating system will be implemented for both users and submissions.

For reference material, the ability to digitise collections is restricted, and rights must be requested using the online form. These requests are handled by database moderators. The quality of digitised material is validated by end users: an image quality rating mechanism is in testing that allows images to be filtered by perceived quality. There is also a reporting mechanism, through which a digitised slide can be flagged as having an identification problem.

User use cases

The web interface of the GPP supports three major use-cases of the tool:
(a) the submission and identification of unknown pollen grains via crowdsourcing;
(b) the submission, digitisation and dissemination of images derived from existing ‘physical’ reference collections, both modern and fossil; and
(c) the public availability of an integrated global database of pollen images with associated metadata, for use in scientific research and education.

SUBMISSION AND IDENTIFICATION OF UNKNOWN POLLEN GRAINS VIA CROWDSOURCING

Crowdsourcing – the process of obtaining ideas from a wide community – has been implemented within the tool for the identification of pollen grains. Such approaches are already used for the identification of plant, mammal and insect species, e.g. iSpot (Silvertown et al. 2015), as well as data collection in the environmental sciences e.g. Zooniverse projects, such as the tropical cyclone project (Hennon et al. 2015). To enable crowdsourcing of pollen identification, we implemented mechanisms for uploading unidentified pollen grain images with metadata, group-based identification and community-based competition.

The process of uploading unidentified grains is designed to be fast and thus requires minimal information. The service requires between one and four images of the grain in question, which can be cropped and rotated during the upload. Other information required are the maximum grain diameter (micrometres), and the location from which the grain was collected. Grain diameter provides essential context to grain identification because morphologically similar grains may vary significantly in size (e.g. in the Compositae family). Similarly, spatial context (latitude and longitude) is important for the exclusion of certain taxa during the identification process. The temporal context is currently optional, but in future we envisage this may also be an important filter for excluding certain groups (e.g. extinct species), and for studies of modern non-fossilised pollen (e.g. in pollination research). If the user has a date for the deposition/collection of a given grain, they may input this as additional data in any commonly used convention (e.g. years before present for radiocarbon-dated fossil material, or calendar year for modern samples). Once a grain is uploaded, it is listed on the GPP as unidentified, and is available to the community for identification.

The crowdsourcing grain identification option in the tool was designed to quickly determine confidence the user has in the identification. Identifications can occur at family, genus and species level. For speed of entry, we integrated The Plant List taxonomy so that plant taxa are suggested while typing, and family is auto-filled when a genus is selected (Fig. 2). We defined three requirements to confirm the identity of a grain through crowdsourcing: (i) there are at least three identifications by different users, (ii) there is a minimum agreement of 70% and (iii) the identification is validated by the taxonomic backbone.

To encourage identifications as well as submissions, a scoring system has been included under the header ‘Leaderboard’. Unidentified grains are given a base score, which builds over time. This score is given for each taxonomic level identified, such that an identification to genus level (e.g. Betulaceae – Betula) would gain double points compared with just a family level identification (e.g. Betulaceae). All users get points. The leaderboard is based on individuals as well as their organisational affiliation, which enables a cross-institutional competitive incentive to the identification process.

Fig. 2. Identification of an unknown pollen grain, demonstrating the auto-suggestion mechanism of valid taxonomic names, which uses The Plant List as the data source.
DIGITISATION AND DISSEMINATION OF EXISTING ‘PHYSICAL’ REFERENCE MATERIAL

Pollen reference slides provide an invaluable source of information for pollen grain identification. Many laboratories and funding agencies are now interested in the dissemination of this information to researchers and to the general public. The GPP allows any individual or institution to request digitisation rights, with which they can digitise their offline reference collections and upload onto this platform for dissemination. Any slides digitised in this way appear in the GPP reference collection alongside the user-submitted grains. All reference material is fully attributed to individuals and their institutions, alongside contact information for the associated party. Submitted material is currently under a non-commercial use licence (CC BY-NC 4.0 - https://creativecommons.org/licenses/by-nc/4.0/).

As before, the only information required to upload images is the maximum grain diameter. To aid speed of entry, we also integrated The Plant List (2013) matching in the digitisation interface. To encapsulate the different views of a pollen grain, specifically the rotation and focus that one could obtain from looking down a microscope (Peck 1974), we have developed a ‘focus image’ function. A z-stack comprising five images of sequential focus levels, taken using a fixed camera, are displayed as a single image that the user can focus through, emulating a light microscope experience (see http://globalpollenproject.org/Reference/Grain/835 as an example of *Agrostemma githago* L.).

A PUBLIC, GLOBAL, INTEGRATED DATABASE OF POLLEN IMAGES AND METADATA

The combination of user-submitted and identified grains along with physical reference collections are collectively presented in the form of a master digital pollen reference collection. These are grains with a ‘confirmed’ identification and can then be used as both a key for comparative identification, and as a teaching tool (Fig. 3). The database provides various browsing and searching features, and will soon be accessible programmatically. Incorporated into the web interface is data from the Global Biodiversity Information Facility (GBIF) Application Programming Interface (API) to visually depict current distributions of records of individual taxa along with brief descriptions of their habitat and ecology (GBIF 2016). For a given taxon, the user can therefore navigate directly from the GPP to Fig. 3. The web interface for the GPP Master Reference Collection (MRC), organised by genus. The collection can be grouped by family, genus or species. Each taxon is displayed with information generated recursively (for itself and all sub-taxa). Person icon: number of unknown grains that have gained this taxon as their confirmed identification. Globe icon: number of digitised reference slides. List icon: number of taxa directly under this in the hierarchy (species in this case).

© 2017 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society, *Methods in Ecology and Evolution*, 8, 892–897
GBIF to examine additional information regarding current distribution, classification and description. Taxa in the GPP are also linked directly to corresponding taxon IDs in the Neotoma Palaeoecological Database (Neotoma 2016). Neotoma’s API is similarly used to visualise distributions derived from palaeoecological data up to 50,000 years before present, with user-selectable time windows (Fig. 4).

Discussion

There is inherent value in physical palynological collections as well as in their GPP digitised form. Each form has a number of strengths and weaknesses lending to preferential use in education and research. While physical slides may preserve an array of grains from different angles and orientations, and may be physically handled under a microscope, they degrade over time, which could, and has, rendered many collections defunct. Through the process of digitising these physical collections using the GPP protocol, we can preserve and make readily available palynological collections to anyone with a computer and an internet connection. Further added value of the GPP collections compared to physical collections and/or other digital collections is the unique meta-data associated with each GPP slide/grain, which due to its common, indexed, format is inherently search- and sort-able, enabling integration with other databases and platforms such as the TRY plant traits database (Kattge et al. 2011). The core framework of the GPP is an open source modular platform that can also be utilised to host and organise other digital collections in other taxonomic groups such as diatoms, Chironomidae, Ostracoda and invertebrates.

The GPP creates new opportunities for research, for example automatic recognition. The image library will be a valuable resource for pollen identification using machine learning approaches. The GPP currently contains over 6000 unique images (when focus images are split) linked directly to botanical taxa. A truly global database of plant pollen images will enable greater power within deep neural network image recognition of pollen grains. Specifically, we suggest that the database promotes image replicates within taxa, which will be crucial to developing morphological pollen recognition using light microscope images.

![Fig. 4. Ericaceae as an example taxon page in GPP.](image)

(a) The pollen traits panel displays recursive trait information. Currently, only maximum grain diameter is collected and displayed as a histogram, alongside the mean value. (b) The distribution panel displaying palaeoecological occurrence data (obtained from the Neotoma Palaeoecological Database) from 1000 to 10,000 years before present. The slider allows user visualisation of changes in the distribution with different time windows. (c) Digitised reference slides are displayed recursively (in this case for all genera and species within Ericaceae). Unknown grains that have gained this identification are shown below digitised reference slides. (d) Botanical information obtained from the Global Biodiversity Information Facility (GBIF). In this case, GBIF returned an image and text summary obtained from Wikipedia. The ‘View at GBIF’ link connects to Ericaceae on the GBIF website.
Use of the GPP has been growing. Since the GPP was launched in March 2016, 1697 unique users from 86 countries have used the master reference collection (to 15th November 2016), with major use in Scandinavia, Russia, the USA, Canada, Mexico and Brazil. The GPP has also been used in teaching in a university setting, incorporating the tool into lectures and laboratory sessions.

In attempting to maximise the effectiveness of this tool, we have addressed four overarching themes. First, the standardisation of palynological practices, specifically relating to taxonomic nomenclature. Second, the assimilation of resources and collections, often in the form of data that is ‘locked away’ in physical collections. Third, access for all, providing free tools for the dissemination of collections and data to encourage wider scientific and public interest and use across the globe, particularly in less well-known regions such as the tropics and neotropics. Fourth, use of this collection as a teaching tool. Each aspect of the GPP can be used collectively or individually. The three functions of the GPP: (i) identification of pollen grains, (ii) dissemination of reference collections and (iii) provision of an accessible, global and searchable database, highlight the major areas of expected use within the palynological discipline. Future development of this tool will include the addition of functionality for pollen traits (macro/micro), the ability to rate quality of images, and further development work for eased data access and interoperability.

Authors’ contributions

W.J.H. and A.C.M. developed the concept for the GPP; A.C.M. implemented the ideas into a database and web app. Continued revision of the framework was conducted by W.J.H. and A.C.M. The authors contributed equally to the manuscript.

Acknowledgements

The authors would like to give thanks to the reviewers, particularly Jack Williams for his constructive and detailed suggestions. We would also like to thank Kathy Willis and the Oxford Long-Term Ecology Laboratory for their support and contribution of the European Reference Collection; Sandra Nogué for her enthusiasm and contributions; Keith Bennett for his early support and contribution of Pollen of Britain and Ireland; Leo Petrokofsky for his work digitising collections held in Oxford; and to all others who have already contributed to the database. Funding was awarded by the Natural Environment Research Council (NERC) to digitise the Oxford collections. Additional funding was also awarded by the Oxford NERC Doctoral Training Partnership for maintenance of this database for the next 3 years.

Data accessibility

All source code for the Global Pollen Project is available at Zenodo (doi: 10.5281/zenodo.260180). The latest release is available at https://github.com/AndreWiOM/global-pollen-project.

References

Bennett, K.D. & Willis, K.J. (2002) Pollen. Tracking Environmental Change Using Lake Sediments: Terrestrial, Algal, and Siliceous Indicators (eds J.P. Smol, H.J. Birks & W.M. Last), pp. 5–32. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Berghlund, B.E. (1986) Handbook of Holocene Palaeoecology and Palaeohydrology: John Wiley & Sons, Chichester, UK.

Brooks, S.J., Sell, A., Toloni, F., & Sparks, T. (2014) Natural history museum collections provide information on phenological change in British butterflies since the late-nineteenth century. International Journal of Biometeorology, 58, 1749–1758.

diDigBio (2016) diDigBio. Available at: https://www.idigbio.org/ (accessed 15 December 2016).

GBIF (2016) The Global Biodiversity Information Facility. Available at: http://www.gbif.org/ (accessed 16 May 2016).

Heidorn, P.B. (2008) Shedding light on the dark data in the long tail of science. Library Trends, 57, 280–299.

Hennon, C.C., Knapp, K.R., Schreck, C.J. et al. (2015) Cyclone Center: can citizen scientists improve tropical cyclone intensity records? Bulletin of the American Meteorological Society, 96, 591–607.

Kattge, J., Diaz, S., Lavorel, S. et al. (2011) TRY – a global database of plant traits. Global Change Biology, 17, 2905–2935.

Mander, L. & Pumayasena, S. (2014) On the taxonomic resolution of pollen and spore records of Earth’s vegetation. International Journal of Plant Sciences, 175, 931–945.

Microsoft (2016a) ASP.NET Core 1.0.0. Microsoft Corporation, Seattle, WA, USA. Available at: https://asp.net (accessed 25 July 2016).

Microsoft (2016b) Entity Framework Core 1.0.0. Microsoft Corporation, Seattle, WA, USA. Available at: http://efproject.net (accessed 25 July 2016).

Microsoft (2016c) Microsoft Azure Cloud Platform. Microsoft Corporation, Seattle, WA, USA. Available at: https://azure.com (accessed 25 July 2016).

Neotoma (2016) The Neotoma Paleoecological Database. Available at: http://www.neotomadb.org/ (accessed 16 May 2016).

Pearssl, D.M. (2015) Preservation. Paleoethnobotany: A Handbook of Procedures, 3rd edn (ed D.M. Pearssl), pp. 201–203. Left Coast Press, Walnut Creek, CA, USA.

Peck, R.M. (1974) A comparison of four absolute pollen preparation techniques. New Phytologist, 73, 567–587.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J. & McConway, K. (2015) Crowdsourcing the identification of organisms: a case-study of iSpot. ZooKeys, 480, 125.

The Plant List (2013) Version 1.1. Available at: http://www.theplantlist.org/ (accessed 23 January 2016).

Wen, J., Ickert-Bond, S.M., Appelhans, M.S., Dorr, L.J. & Funk, V.A. (2015) Collections-based systematics: opportunities and outlook for 2050. Journal of Systematics and Evolution, 53, 477–488.

Received 15 December 2016; accepted 23 January 2017

Handling Editor: Sarah Goslee