Big data integration: Pan-European fungal species observations’ assembly for addressing contemporary questions in ecology and global change biology

Carrie ANDREW, Einar HEEGAARD, Paul M. KIRK, Claus BÄSSLER, Jacob HEILMANN-CLAUSEN, Irmgard KRISAI-GREILHUBER, Thomas W. KUYPER, Beatrice SENN-IRLET, Ulf BÜNTGEN, Jeffrey DIEZ, Simon EGLI, Alan C. GANGE, Rune HALVORSEN, Klaus HØILAND, Jenni NORDÉN, Fredrik RUSTØEN, Lynne BODDY, Håvard KAUSERUD

Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo, Blindernveien 31, 0316 Oslo, Norway
Forestry and Forest Resources, Norwegian Institute of Bioeconomy Research, Fanafletaten 4, N-5244 Fana, Norway
Cardiff School of Biosciences, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, UK
School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey TW20 OEX, UK
Department of Research and Collections, Natural History Museum, University of Oslo, NO-0318 Oslo, Norway
Norwegian Institute for Nature Research, Gausvadsallén 21, NO-0349 Oslo, Norway
Centre for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, DK-2100 Copenhagen, Denmark
Swiss Federal Research Institute WSL, CH-8933 Birmensdorf, Switzerland
University of Cambridge, Department of Geography, CB2 3EN, UK
Global Change Research Centre and Masaryk University, 613 00 Brno, Czech Republic
Department of Botany and Biodiversity Research, University of Vienna, A-1090 Vienna, Austria
Bavarian Forest National Park, Freyunger Str. 2, D-94481 Grafenau, Germany
Department of Soil Quality, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands
Mycology Section, Jodrell Laboratory, Royal Botanic Garden, Keu, Surrey TW9 3DS, UK
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Abstract

Species occurrence observations are increasingly available for scientific analyses through citizen science projects and digitization of museum records, representing a largely untapped ecological resource. When combined with open-source data, there is unparalleled potential for understanding many aspects of the ecology and biogeography of organisms. Here we describe the process of assembling a pan-European mycological meta-database (ClimFun) and integrating it with open-source data to advance the fields of macroecology and biogeography against a backdrop of global change. Initially 7.3 million unique fungal species fruit body records, spanning nine countries, were processed and assembled into
6 million records of more than 10,000 species. This is an extraordinary amount of fungal data to address macro-ecological questions. We provide two examples of fungal species with different life histories, one ectomycorrhizal and one wood decaying, to demonstrate how such continental-scale meta-databases can offer unique insights into climate change effects on fungal phenology and fruiting patterns in recent decades.

1. Introduction and aims

National-scale datasets documenting species occurrence data and biogeographic patterns are increasingly common through the digitization of records from museums, educational institutions, natural history and mycological societies, and from citizen-science projects (Schmeller et al., 2009; Henderson, 2012; Tulloch et al., 2013; Isaac et al., 2014). Publicly available databases (such as GBIF (Global Biodiversity Information Facility, http://www.gbif.org)) are simultaneously pioneering open-source data repositories, providing massive amounts of complementary metadata with which to answer the leading questions in global change biology (Mills et al., 2015). However, global-scale questions in ecology require data at spatial scales surpassing national-scale boundaries and dataset sources. This scaling up necessitates not only the unification of multiple national-scale datasets into a single, comprehensively analyzable unit, but further requires the integration of datasets from multiple, partly open-sourced, repositories.

Developing processes that stitch together datasets across multiple countries and sources into a 'meta-database' (Table 1), with corresponding metadata, offers previously unprecedented opportunities for the exploration of continental-to global-scale biogeographic patterns, including the responses of organisms to global change. For example, such analyses enable the tracking of latitudinal shifts of many species in response to climate change (Chen et al., 2011). Fungi are conspicuous in their absences throughout all these analyses, with only one notable exception: Bebber et al. (2013) used large-scale datasets to record the northward spread of plant pathogens attacking crop plants.

A fungal perspective on biodiversity can provide novel insights into ecosystem functioning, as saprotrophic (decomposer) and biotrophic (mycorrhizal and parasitic) fungi are crucial components of intact systems (Boddy et al., 2014; Heilmann-Clausen et al., 2015). The ClimFun (‘Climate Change Impacts on the Fungal Ecosystem Component’) project combines many large datasets into a meta-database suited to global-change topics (Table 1). Multi-source fungal fruiting datasets from nine European countries were merged to create a unique meta-database capable of addressing broad-scale questions, such as the effects of climate and land-use change, on fungi.

Here we, first, emphasize recognition of meta-database analyses for global change ecology with the use of fungal data. Our focus on fungi helps alleviate their under-representation in large-scale scientific research. The processes and concepts provided, such as effective data assembly with relatively low loss of data, are transferable to and demonstrate promise for other multi-source dataset assemblies. To extend meta-database research further, we integrate the data assembled with those of open-source data and give examples of research topics that can be addressed with such biological meta-databases. Our examples demonstrate how the fusing of meta-databases with open-source data offers unique potential for answering impending questions in ecology and global change biology.

2. The data assembly process

Over 7.3 million national-scale fruit body records from nine European countries were compiled from multiple sources including museums, educational institutions, mycological societies, and citizen-science projects. Such a large, diverse dataset presents challenges for the assembly process (Table S1). For example, discrepancy in scientific names and underlying taxonomic differences among different datasets, and spanning decades to centuries of time, is a ubiquitous challenge for all biological datasets (e.g., Primack and Miller-Rushing, 2012). Often databases and literature are available to help with nomenclatural harmonization to a single standard, i.e., Index Fungorum (http://www.indexfungorum.org) incl. Species Fungorum (http://www.speciesfungorum.org) for the data we assembled (taxonomy updated December 2015). The datasets were thus formatted and filtered in many steps by sequentially addressing issues concerning taxonomic and nomenclatural synonymisation between datasets, date formatting, geographical reliability, habitat and host translations, and other quality control checks (Tables 1, S1).

The resulting meta-database comprises over 6 million records and more than 10,000 species, spanning the years of 1723–2010, with the greatest accumulation of records after 1960 (Table 1, Fig. 1). On average, 77 % (±7 % standard error) of records and 70 % (±4 % standard error) of species from the original dataset remained after processing (Tables 1, S1). The proportion of records retained post-processing was similar across all national datasets (χ² = 6.25, df = 7, p > 0.05) with the exception of one smaller, unique dataset (Estonia; test of significance when included: χ² = 52.17, df = 8, p ≤ 0.01), mainly attributable to records removal which lacked species-level taxonomic resolution, georeferenced coordinates, or were out-of-country.

3. Spatiotemporal distribution of records

Given big data, assessing the spatiotemporal distribution of records can determine any potential analytic problems with large-scale analysis. Despite differences in the numbers and timespans of fungal records across countries, the relative
Table 1 – Processing steps (a) for the entire ClimFun meta-database and descriptive statistics plus contact information (b) for the national scale datasets within. For the processing steps by dataset and further information, see Table S1.

(a) Processing steps and amounts of all records

<table>
<thead>
<tr>
<th>Step</th>
<th>% Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original records</td>
<td>97.8%</td>
</tr>
<tr>
<td>Taxonomy updated</td>
<td>97.0%</td>
</tr>
<tr>
<td>Georeferences &amp; altitude formatted</td>
<td>96.0%</td>
</tr>
<tr>
<td>Out-of-country records removed</td>
<td>96.0%</td>
</tr>
<tr>
<td>Discrepancies, species NA values removed</td>
<td>64.5%</td>
</tr>
<tr>
<td>Georeference NA values removed</td>
<td>82.2%</td>
</tr>
<tr>
<td>Duplicate records removed</td>
<td></td>
</tr>
<tr>
<td>Preliminary taxonomy limitation</td>
<td>82.2%</td>
</tr>
<tr>
<td>Final records</td>
<td>6,025,149</td>
</tr>
</tbody>
</table>

(b) Number of records

<table>
<thead>
<tr>
<th>Dataset&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Final ClimFun database</th>
<th>Original datasets</th>
<th>Proportion of original retained with ClimFun database</th>
<th>Final ClimFun database</th>
<th>Original datasets</th>
<th>Access information</th>
<th>Contact person(s) and/or associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>398274</td>
<td>433042</td>
<td>0.92</td>
<td>3988</td>
<td>5211</td>
<td><a href="http://www.univie.ac.at/oemykges/pilzdatenbank/">http://www.univie.ac.at/oemykges/pilzdatenbank/</a></td>
<td>I. Krisai-Greilhuber, Austrian Mycological Society</td>
</tr>
<tr>
<td>CH</td>
<td>283232</td>
<td>311041</td>
<td>0.91</td>
<td>4974</td>
<td>6120</td>
<td><a href="http://www.swissfungi.ch/">http://www.swissfungi.ch/</a></td>
<td>B. Senn-Irlet, P. Jakob, C. Bässler, German Mycological Society</td>
</tr>
<tr>
<td>DK</td>
<td>475737</td>
<td>534239</td>
<td>0.89</td>
<td>4007</td>
<td>5185</td>
<td><a href="http://svampe.databases.org/">http://svampe.databases.org/</a></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>11051</td>
<td>52873</td>
<td>0.21</td>
<td>2114</td>
<td>3931</td>
<td><a href="http://kogud.emu.ee/?do=coll&amp;id=9&amp;lang=eng">http://kogud.emu.ee/?do=coll&amp;id=9&amp;lang=eng</a></td>
<td>T. Kuyper, Netherlands Mycological Society</td>
</tr>
<tr>
<td>NL</td>
<td>1795551</td>
<td>2094920</td>
<td>0.86</td>
<td>3694</td>
<td>4607</td>
<td><a href="https://www.verspreidingsatlas.nl/paddenstoelen">https://www.verspreidingsatlas.nl/paddenstoelen</a></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>170084</td>
<td>245382</td>
<td>0.69</td>
<td>4128</td>
<td>5939</td>
<td><a href="http://www.gbif.no/datasets/">http://www.gbif.no/datasets/</a></td>
<td>N. Ogrić, Slovenian Mycological Association</td>
</tr>
<tr>
<td>SI</td>
<td>164269</td>
<td>199441</td>
<td>0.82</td>
<td>2663</td>
<td>3676</td>
<td><a href="http://www.zdravgozd.si/mikoteka_index_en.aspx">http://www.zdravgozd.si/mikoteka_index_en.aspx</a></td>
<td>P.M. Kirk, British Mycological Society</td>
</tr>
<tr>
<td>UK</td>
<td>1402424</td>
<td>1937526</td>
<td>0.72</td>
<td>6453</td>
<td>14875</td>
<td><a href="http://www.fieldmycology.net/">http://www.fieldmycology.net/</a></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> AT: Austria, CH: Switzerland, DE: Germany, DK: Denmark, EE: Estonia, NL: Netherlands, NO: Norway, SI: Slovenia, UK: United Kingdom.

<sup>b</sup> Of those with an updated taxonomy.
Fig. 1 – Number of records and species in each country-level dataset in the ClimFun meta-database. (a) Number of records from each country, years 1960–2010. (b) Species accumulation curves by observation numbers. Confidence intervals shade median species number estimate, best noted by unequal line widths. Total numbers of species are provided in parentheses. (c) Phenological distribution of dataset records by fruiting day. Note that record numbers are provided on a log-scale to highlight two main fruiting seasons: spring (April–June) and fall (August–October). NO, Norway; EE, Estonia; DK, Denmark; UK, United Kingdom; DE, Germany; NL, Netherlands; AT, Austria; CH, Switzerland; SI, Slovenia.

contributions of records are remarkably spatially similar (Fig. 1). The geographical coverage also is very comprehensive, although variation in sampling intensity leads to uneven geographic densities of records, especially with finer temporal and spatial resolution (Figs 2, S1). Notably, the meta-database provides a continuous north–south latitudinal gradient in fungal records spanning 71–45 °N, i.e., from north of the Arctic Circle to the European Alps and Adriatic Sea (Fig. 2). The sampling intensity has increased over time, reflecting increased activities by amateur mycologists as part of the larger practice of citizen science, the latter of which can be biased by population and infrastructure densities (Geldmann et al., 2016). Geographical coverage especially has become more complete since the 1960s and 1970s (Figs 1, S1) so that we suggest limiting the data to 1970–2010 for climate-relevant studies, thereby reducing the data to 88–90 % of all original records and to 96–97 % of all species.

4. Taxonomic distributions of the national datasets

The percent contribution of fungal orders is highly consistent across the datasets (Fig. 3) and not statistically different ($\chi^2 = 98.5$, df = 80, $p > 0.05$), despite variation in collecting traditions, taxonomic specializations (exemplified by the emphasis on fungi belonging to the Pezizales in the Estonian (EE) dataset (Fig. 3b)), and the form of data provision among countries. Gillied mushrooms dominate all datasets (Fig. 3a). A preponderance of the records is from taxa within the main order of gilled mushrooms, the Agaricales (54.5%), ranging from 49 to 57 % among countries. Agarics in total (adding the Cantharellales and Russulales) comprise, on average, 71.8 % of records, ranging from 63% to 76 % (Fig. 3b). This high percentage identifies the Agarics as a universally represented group well suited for further studies.

In contrast, the percentage of unique species is not consistent among datasets ($\chi^2 = 203.7$, df = 8, $p < 0.01$; Fig. 3c) nor is the number of records (Fig. 1c). More important is that relative differences in monthly numbers of records are consistent among datasets (e.g., Estonian record amounts by month are always the lowest, while differences among countries are consistent across months; $\chi^2 = 5.15$, df = 88, $p > 0.05$).

5. Regional similarity of species composition

Among regions, shared and unique species offer different insights into a meta-database’s potential uses. For example, unique species can be used to examine the heterogeneity, distributions and amounts of different habitat types across regional climatic and altitudinal gradients. From a conservation perspective, it is very interesting to note the hundreds of species unique to each country (Fig. 3c). Some of these may be consequences of varied gathering strategies or collector biases (preceding the meticulous data assembly process), especially if not associated with unique or rare habitats. However, others are likely to be accurately quantified as rare. It has long been known that range-restricted species are at greatest risk from climate change (Parmesan, 2006). Therefore, these species may be at greatest risk from extinction, if they are not able to track changes in climate and their habitats with adequate speed (Senn-Iriet et al., 2007). This would be exacerbated by them having small ranges and populations. These data certainly provide the greatest potential for climate change impact and conservation studies of rare species (Heilmann-Clausen et al., 2015).
Shared species, on the other hand, are optimal candidates for studies of broader-scale global change processes (see below for examples of two species). A total of 947 species are shared across all countries (Fig. 3c). The size of this pool of shared species is largely determined by the number of species in the smaller datasets (e.g., Slovenia (SI) and Estonia (EE)), which highlights both the benefits and challenges associated with the incorporation of additional datasets. The benefits of incorporating countries with smaller datasets include greater geographical coverage and taxonomic representation within the meta-database, but inclusion could constrain analyses that depend on consistent taxonomic representation across countries.

6. Phenological fruiting patterns

Describing the timing of life-cycle events is critical to determine species interactions with climate and one of the most widely investigated indicators of climate change across the globe (Feehan et al., 2009; Parmesan and Hanley, 2015). The current meta-database allows an unprecedented examination of continental-scale fungal phenology that can extend the spatiotemporal scales and investigative depth of previous studies within individual countries (Gange et al., 2007; Kauserud et al., 2008, 2010, 2012; Büntgen et al., 2012; Diez et al., 2013). Based upon the ClimFun data, two main fruiting periods can be identified: a minor spring period and a much larger autumnal fruiting period (Figs 1c, 5).

7. Scientific potential of the meta-database

The inherent scientific potential of the meta-database is highlighted by the phenological characteristics of two widespread and commonly occurring species (Figs 4 and 5). Analysis of temporal and spatial patterns of the ectomycorrhizal Amanita...
rubescens (the blusher; Fig. 4e) and the saprotrophic Hypholoma fasciculare (the sulfur tuft; Fig. 4f) show that climate-related shifts in their mean and median fruiting dates (1970–2010) have occurred across latitudes (Fig. 4c and d). Fruiting seasons by altitude highlight species-specific responses with time, where A. rubescens typically fruits earlier in the season and H. fasciculare later in the season than in past decades, particularly at high elevations (Fig. 4a and b).

The meta-database reveals interesting large-scale geographic patterns in fruiting phenology of A. rubescens and H. fasciculare that are related to effects of the oceanic or continental climates as well as geographical latitude, longitude and altitude ranges (Fig. 5). Proximity to oceans and temperature effects (as derived from latitude and altitude) both influence fungal fruiting, and give rise to clear patterns despite geographical differences in the density of records. For the two species in oceanic regions, mean fruiting has changed more markedly in the season for recent years (1990–2010) compared with previously (1970–1990) (Fig. 5). With increased altitude, the mean fruiting time becomes earlier, demonstrating the interactive effects of latitude and altitude for fungal fruiting (Figure S3, S4). It will be important to distinguish, however, the first, last and mean fruiting patterns (including a possible increased tendency towards a bimodal pattern in fruiting) prior to accurately determining potentially climate-driven consequences to fungal phenology.

8. **Open-source data integration**

The integration of meta-databases with open-source data is key for elucidating the responses of organisms to global change (Fig. 6). For example, normalized difference vegetation index (NDVI) not only captures the seasonal ‘wave’ of primary productivity in non-equatorial regions (Bischof et al., 2012), but is also a proxy for understanding carbon dynamics and fungal correlations with plant biomass (e.g., Timling et al., 2014). When combined with soil organic carbon, these data might help elucidate the role of fungi in carbon decomposition and sequestration (e.g., Kuyper, 2017). Any response of climate effects on fungi will also require simultaneous investigations into other, equally important global-change components such as nitrogen deposition (Jarvis et al., 2013). The meta-database that we have compiled can also be mapped onto changes in land-use cover and habitat fragmentation (Nordén et al., 2013; Heilmann-Clausen et al., 2014).

9. **Value of continental-scale meta-databases**

Meta-databases are relevant for investigating an array of ecological questions at spatiotemporal scales previously unattainable. The ClimFun project demonstrates how the collaborative effort between a team of scientists and the general public may result in unique and powerful meta-databases to combine with open-source data. The results provided here use only a small subset of the data available in the ClimFun meta-database to ask a few initial questions. Nonetheless they extend scales from earlier studies at more local-scales (e.g., Gange et al., 2011; Heilmann-Clausen et al., 2015). Ongoing research will further examine climate-related effects on fungi and macroecological patterns through projects examining how species’ range shifts, overall distributions, and host-substrate shifts are related to climate and land use changes. We urge other scientists to follow suit with such types of collaborative projects.

Scientists face significant challenges in trying to understand biological responses to ongoing climate change, and
Fig. 4 – Phenological distribution of (a, c, e) Amanita rubescens and (b, d, f) Hypholoma fasciculare fruiting days for two time intervals, 1970–1990 (upper, lighter) and 1990–2010 (lower, darker). The boxplots provide median values with the bars and mean values with shaded diamonds. Datasets are arranged by altitude (c, d) and latitude (e, f). Estonia in this case was removed due to low record numbers. Photos courtesy of Alan Gange (Amanita rubescens) & Per Marstad (Hypholoma fasciculare).
Fig. 5 – Geographical distribution and phenology of fruiting day, based on a Gaussian distribution of a general additive model, by spline functions of the interaction of latitude and longitude plus altitude, for (a, b) Amanita rubescens and (c, d) Hypholoma fasciculare. Lighter-colored symbols indicate lower altitudes and grade into higher altitude values with darker shading. Geographical changes in the mean fruiting phenology dates are depicted with the isolines with corrections by altitude to fruiting day inset. Records are divided into two time periods: 1970–1990 (a, c) and 1990–2010 (b, d). Mean day always refers to the mean fruiting day at 0 altitude. Estonia in this case was removed due to low record numbers.

also in communicating their findings to the general public. Fungi have been largely invisible in these efforts, both because most fungi are literally hidden within soils or substrates for most of their life cycle, and because large datasets have not previously been compiled for fungi as they have for plants. At the forefront of a climate awareness shift in the general public, we are primed to understand how past-to-current changes in climate have influenced not only plants and animals (e.g., Isaac and Pocock, 2015) but also the often-overlooked fungal world. The goal of the ClimFun metadata project is to bring fungi, and their responses to climate, into greater scientific prominence and more squarely in the public’s eye. A major impediment to both of these goals has been the fragmented nature of datasets documenting the distribution of fungal species. The ClimFun project demonstrates how multi-national collaborative efforts involving
Fig. 6 — A conceptualization of the scientific promise meta-databases combined with open-source data contain for global change biology research. The ClimFun fungal fruit body records amounts are shown on a 50 × 50 km grid scale also utilized for the other, open-source data layers. See acknowledgments section for further information regarding the open-source data references.
citizen science and multi-source data can effectively address scientific questions with microorganisms previously impossible to answer over regional to continental scales.

Acknowledgements

We thank the Research Council of Norway for financial support, and Dr. Dag Endresen for dataset processing aide. For providing digitized data: AT, The Austrian Mycological Society and Wolfgang Damon; CH, www.swissfungi.ch and Peter Jakob; DE, Deutsche Gesellschaft für Mykologie (German Mycological Society) and Dr. Martin Schmidt; DK, The Danish Fungal Atlas team and Tobias Frøslev, Thomas Læssøe, Jens. H. Petersen and Jan Vesterholt; EE, Estonian eBiodiversity database, Dr. Kadri Partel and Kessy Abarenkov; NL, The Netherlands Mycological Society (NMV) and A. van den Berg; NO, The Mycological Herbarium at the Natural History Museum (University of Oslo); SI, The Slovenian Forestry Institute, the Central database of fungi in Slovenia, the Slovenian Mycological Association, and Dr. Nikica Ogris; UK, www.field-mycology.net. Finally, we are deeply indebted to the thousands of fungal recorders across the nations. Without their contributions, this and future research would not have been possible.

The following sources are commended for open-source data

1. NDVI
   GIMMS AVHRR Global NDVI 1/12-degree Geographic Lat/Lon, from the NOAA satellite data with a timespan of July 1981-present and a range of 8 km: http://ecocast.arc.nasa.gov/data/pub/gijmms/3g.v00/00README.geo.txt.

2. Soil organic carbon
   OCTOP: Topsoil Organic Carbon Content for Europe, from Joint Research Centre - European Soil Data Centre (ESDAC), 1 km original resolution (note references in web link): http://esdac.jrc.ec.europa.eu/content/octop-topsoil-organic-carbon-content-europe.

3. Nitrogen (NItx, NOy)
   Reduced and oxidised downward nitrogen deposition velocities, 1850–2010, obtained from Greenhouse gas management in European land use systems (GHG Europe) FP7, using NCAR CTM data: http://www.europe-fluxdata.eu/ghg-europe/data/others-data.

4. Altitude and climate
   WorldClim data representative of 1960–1990 bioclimatic variables, 2.5 min–0.5 min original resolution: http://www.worldclim.org/current.

5. Land cover

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jfrb.2017.01.001.

Co-author contribution(s)

Designed this study: HK, CA.
Dataset ‘keepers/providers: ACG, BSI, CB, IKG, JHC, PMK, SE, TK.
Taxonomic and nomenclatural synonymisation: PMK, CA, JN.
Processed & formatted data into meta-database: CA.
Conducted data analyses: CA, EH, HK.
Wrote the manuscript: CA, HK.
Edited the manuscript: LB, ACG, RH, JHC, SE, BSI, IKG, CB, TK, UB, JD, JN, EH, KH, FR.

Conflicts of interest

There are no conflicts of interest for the authors to declare.

REFERENCES

data reveal ecological, historical and evolutionary factors shaping interactions between woody hosts and wood-inhabiting fungi. New Phytol. 212, 1072–1082.