Research Ideas and Outcomes 9: e113378
doi: 10.3897/rio.9.e113378

OPEN ACCESS

Project Report

Understanding the users and uses of UK Natural History Collections

Helen Hardy[‡], Laurence Livermore[‡], Paul Kersey[§], Ken Norris[‡], Vincent S. Smith[‡]

[‡] The Natural History Museum, London, United Kingdom
[§] Royal Botanic Gardens, Kew, Richmond, United Kingdom

Corresponding author: Helen Hardy (h.hardy@nhm.ac.uk)

Reviewable v 1

Received: 27 Sep 2023 | Published: 02 Oct 2023

Citation: Hardy H, Livermore L, Kersey P, Norris K, Smith VS (2023) Understanding the users and uses of UK Natural History Collections. Research Ideas and Outcomes 9: e113378. https://doi.org/10.3897/rio.9.e113378

Abstract

UK natural science collections hold over 137 million items, an unrivalled source of data about 4.56 billion years of planetary development and hundreds of years of biological change, including the differences made by humans — but the scientific, commercial, and societal benefits of these collections are constrained by the limits of physical access, and by highly fragmented digitisation efforts with less than 10% digitally available. Following work with Frontier Economics in 2021, which showed potential for £2 billion in benefits to the UK economy from digitising all UK natural science collections, in 2022–23 the Natural History Museum London worked, with analytical support from McKinsey and Company, to understand the impact of what has already been digitised and shared by UK natural science collections — what is the demand for these data, what are they used for, and how does this deliver efficient, effective and impactful research?

This study focuses on usage via the Global Biodiversity Information Facility, the largest source of relevant usage data, examining 7.6 million records from twelve UK institutions. While these UK collections data are just 0.3% of total GBIF occurrences, they are cited in 12% of peer reviewed publications citing GBIF data, showing the disproportionate impact of UK collections data and the historical, geographical, and taxonomic richness that they bring. Researchers have already benefited from more than £18 million of efficiency savings from digital UK specimen data. Data from natural science collections held in the UK are uniquely impactful resources, vital to a future in which people and planet thrive, and a step

change in the pace of digitisation is needed to unlock their potential for researchers, policymakers, and society.

Keywords

natural history collections, natural science collections, value of research, value of collections, collections users, economic benefits, digitisation, digitization, collections impact, conservation, biodiversity

1. Introduction

The Distributed System of Scientific Collections UK (DiSSCO UK, https://www.dissco-uk.org) is a partnership of UK natural science collections, led by the Natural History Museum, London. It aims to unlock and harness the power of natural science collections data as vital infrastructure for research into the key challenges facing humanity and the planet.

Supported by the Arts & Humanities Research Council (AHRC — who are responsible for UK heritage collections as research infrastructures), DiSSCo UK surveyed institutions holding natural science collections in 2022 (Smith et al. 2022), identifying over 137 million items, an unrivalled source of data about 4.56 billion years of planetary development and hundreds of years of biological change, including the impacts of human activity. The scientific, commercial, and societal benefits of these collections are constrained by the limits of physical access, and by highly fragmented digitisation efforts: over 60% are not digitised in any way, and only 12% are currently digitised to a level that is likely to meet scientific research needs (Smith et al. 2022).

A key part of making the case to unlock the potential of UK natural science collection data is to understand the benefits in more detail. Following previous work (Popov et al. 2021) which showed potential for £2 billion of economic benefits to the UK over 30 years if all relevant collections were digitised (a seven- to ten- times return on investment), we wanted to understand in more detail the current position. In particular, this study set out to explore:

- The volume of data currently available from UK natural science collections;
- The demand for these data — who is using them and what research topics they are being used for; and
- Their value — in terms of research efficiency, research effectiveness, and wider impact.

Analysis was conducted between November 2022 and January 2023, combining quantitative and qualitative approaches to investigate the characteristics of uploaded data, users and uses of data, and the value created by that usage. This is a rapidly evolving area, with more data being released every week and new uses developed.

We looked for data sources that were openly accessible; in possession of consistent data; offering good metadata and/or access to data for analysis; and recognised as preeminent data sources within their domain. While we considered a range of sources, including for example institutional data portals, by far the largest-scale and most consistent source of FAIR (findable, accessible, interoperable and resuable) collections data with consistent quantitative evidence about users and usage is the Global Biodiversity Information Facility (GBIF — www.gbif.org), so GBIF data are the primary basis for this study. GBIF is a globally recognised resource, designated a global core biodata resource in December 2022 by the Global Biodata Coalition (which includes UK Research and Investment) (GBIF Secretariat 2022). UK institutuions submit data to GBIF for redistribution through this aggregation platform. As GBIF covers biodiversity data about life on earth, however, this analysis excludes the impact of the UK's important geo-science collections, which would add to the beneficial impacts described in this paper. In addition to GBIF and supplementary sources of quantitative data as documented in the methodology, we also undertook qualitative interviews with a small number of current research data users, to gain deeper insight into how they are using UK natural science collections data.

Recent research (GBIF Secretariat and Deloitte Access Economics 2023) looking at the economic value of global GBIF data as a whole shows that nearly 50% of users would not have been able to achieve their research outcomes without GBIF data, and another 41% could only have done so with significantly increased time and effort. Over 90% of users link their use of GBIF-mediated data to advancing the UN Sustainable Development Goals. Overall, this research showed that every €1 invested in GBIF provides €3 in direct benefits to users and up to €12 in societal benefits. Key insights from our study show a similar pattern of high demand, value and impact from UK collections data.

1.1 Key Insights

The key insights generated from this study are as follows:

- 7.6 million specimens, from 248 territories and countries, are freely accessible on GBIF from the 12 UK institutions investigated — these represent less than 6% of total UK natural science collections. [see Section 3.1]
- 39 billion individual specimen records from UK institutions have been downloaded from GBIF since 2015, and 2,710 publications cite these data. [see Section 3.2.1]
- 12% of the total peer-reviewed journal articles citing GBIF data cite UK natural science collections — these data currently make up just 0.3% of total occurrences on GBIF, meaning they punch some 40 times above their weight. [see Section3.2.1]
- In 2022, there was a download event of the Natural History Museum's data from GBIF on average every 3 minutes 24 seconds, and 2.2 publications per day on average cited UK institutions' data. [see Section 3.2.1]
- More than 250 publications on each of the themes of climate change, invasives and conservation cite UK institutions' uploads — research areas that are key to a future in which people and planet thrive. [see Section 3.3.3]

- ~1,200 UK-affiliated researchers are authors of publications that use UK collections data, among 13,000 researchers from at least 160 countries who have cited UK institutions' GBIF data. [see Section 3.2.3]
- £18m in research efficiencies have already been realised, assuming a single physical visit saved per citation, of which £1.4 million can be attributed to UK-affiliated researchers. [see Section 3.3.2]
- These savings can be reinvested in additional research. Interviews with researchers show that real savings can be many times higher, particularly when digital collections data are combined with AI analysis techniques. [see Section3.3.4]

1.2 Project context

A short, summary paper of the key insights and context around this study is also available (Hardy et al. 2023). This longer version contains the methodology and a fuller description of the findings, as well as supplementary material to support the methodology and expand on the results.

2. Methodology

We undertook both a quantitative and qualitative analysis, of which the quantitative was much more extensive. The queries used to generate the source data are provided in Jupyter notebooks (Suppl. material 1); an overview of the method is provided at Suppl. material 2; and data used to generate the charts with some additional data are linked within the text and in chart captions. Calculations are provided for key insights including efficiency savings (Suppl. material 13). We have not provided the full aggregated dataset from GBIF owing to size. The interview guide for the qualitative analysis is also provided (Suppl. material 3). The code and calculations provided in the supplementary materials were developed by McKinsey & Company, and data extracted by them, under the supervision of the Natural History Museum. These materials are provided at the discretion of the Museum for the benefit of anyone wishing to replicate this methodology.

2.1 Selection of relevant institutions

In order to identify relevant UK collections data for this analysis, we needed to identify UK institutions who currently publish collections data. There are more than 90 institutions in the UK who hold natural science collections, however the majority of these do not yet publish specimen data to GBIF.

The NHM provided a list of relevant institutions who hold collections in the UK, including museums, botanic gardens, universities and specialist centres and societies. A subset of 22 institutions who have published data to GBIF were examined in more detail. GBIF holds occurrence data of two main types — observation data (such as a record of a human or sensor observation of a bird), and specimen data (i.e., data 'vouchered' by a link to a physical object such as the specimens in collections). This study is focused on specimen data from collections, so ten of the 22 institutions were excluded from the analysis because they publish a majority of observation data rather than specimen data.

The remaining twelve institutions considered in the analysis were as follows: they have published data at various times since 2015, with total uploads containing over 99% specimen data, and no institution on this list uploading less than 50% specimen data:

- Cumbria Biodiversity Data Centre
- Department of Zoology, Cambridge
- Leeds Museums and Galleries
- Manchester Museum, The University of Manchester
- National Museums Scotland
- Natural History Museum, London
- Nottinghamshire Biological and Geological Records Centre
- Royal Botanic Garden Edinburgh
- Royal Botanic Gardens, Kew
- The University Museum of Zoology, Cambridge
- UK Polar Data Centre (British Antarctic Survey)
- World Museum, National Museums Liverpool

2.2 Sources for quantitative data

Having identified the relevant institutions, the core dataset for this study was the collections data that those twelve institutions had published to GBIF at the time of this work.

To address our research questions about data availability and use, we needed to understand the volume of relevant data; its characteristics (e.g., specimen taxonomy and geography); and data use via downloads and citations.

To understand value and impact, we also needed to understand the topics of data use and citation, and how these translate to economic value; information about the users (e.g., their association with publications/citations and their institutional or geographic affiliation); and information about comparative costs particularly costs of physical visits.

2.2.1 GBIF data

GBIF data were used to determine:

1. details of the specimen data uploaded by UK institutions to GBIF, in relation to the wider body of all the specimen data available on GBIF:

- Volume
- Occurrence type
- Specimen type
- Taxonomy
- Country of collection

2. Usage of the data uploaded by the UK, in relation to the wider body of all the data on GBIF:

- Volume of publications citing the data
- Topic tags of the publications
- Affiliated institution geography of publications' authors
- Download data

2.2.2 Additional sources of quantitative data

Additional sources were used to enhance our analysis (see Fig. 1 and supplementary material).Crossref are a not for profit membership organisation who aim to make research objects easy to find, cite, link, assess, and reuse, sharing metadata to reveal relationships between research outputs*[1]. Crossref data were used to determine the geographic location of the affiliated institutions for authors of publications that have cited relevant UK data. This helped us to understand where in the world data are being used and having impact.

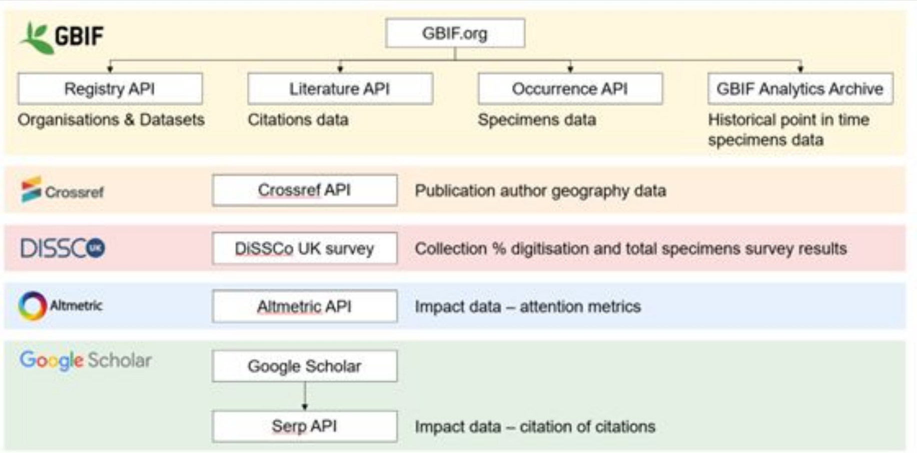

Figure 1. doi

Five sources of data (GBIF, Crossref, DiSSCo UK, Altmetric, Google Scholar) and the types of data obtained for the analysis.

DiSSCo UK survey data (Smith et al. 2022) were used to determine overall (self-declared) levels of digitisation and volumes of specimens for UK scientific institutions that were part of the DiSSCo UK survey, to be able to compare this with the set of data currently published to GBIF.

Google Scholar is intended to provide a simple way to broadly search for scholarly literature. Google Scholar data were used to determine onward citations of the publications that cited UK institutions' data uploaded to GBIF, using the Search Engine Results Page (SERP) API with query parameters described in "03_Google_scholar_data_extraction.ipynb" (Suppl. material 1). This gives us additional insight into the impact of data use, by showing how impactful publications go on to be — it should be noted that this continues to increase over time, so a point in time analysis such as this one is inherently limited.

Altmetric aims to broaden and deepen understanding of the value of research by tracking online engagement from a range of sources. Altmetric data were used to determine the Altmetric Attention Scores for publications citing UK institutions' data uploaded to GBIF (Williams 2016) — again, this provides additional information about the onward impact of those publications.

In addition, the NHM supplied unpublished data from the SYNTHESYS programme (Smith et al. 2019) about the costs of researcher visits to collections. These were used to calculate the cost per physical visit to an institution for a researcher including travel, accommodation, per diems and to calculate the cost to a hosting Institution from hosting a researcher. These data focus only on European institutions and travel, however they are the most detailed and consistent visit cost data that we are aware of. This allowed us to estimate savings from access to the digital data, assuming one physical visit per relevant publication.

The NHM also supplied an unpublished estimate of total spend on its core digitisation team between the start of the Digital Collections Programme and this piece of research — some £5.4 million not including the numerous other resources involved in delivering digitised collections, such as curatorial staff, data managers, data portal developers and others. This figure gave a very conservative baseline of benefit provided to researchers accessing UK collections data that are free at the point of use.

Frontier Economics' analysis (Popov et al. 2021) describes five thematic areas that derive value from research using digitised natural history collection data. These areas were mapped to GBIF topic tags of current citations, to confirm likely economic value and to provide assumptions that support the calculation on translation to value for a researcher.

2.3 Sources for qualitative insights

Case studies and qualitative insights drawn on in this study have been sourced from:

- Discussion with the research team, including representatives of NHM and Kew;
- UK institutions' published information e.g., blogs and press releases;
- Publications that have cited UK institutions' uploaded data;
- Interviews with researchers that have used cited UK institutions data uploaded to GBIF (Suppl. material 3);
- Discussion with the GBIF team.

Interviews were conducted to get feedback from a small subset of end users (scientific researchers) on the impact of UK institutions' digitised data for them (e.g., ability to conduct their research, quality of research possible, efficiencies created), and to test assumptions used in the translation to value methodology.

There was limited time and resource to arrange and conduct interviews within the scope of this study — five interviewees were selected based on high frequency of data use; high impact of their research; and use of novel research techniques on the data. The majority of interviewees specialise in botanic material — this reflects the long standing availability of workflows to digitise pressed plant specimens, meaning that they are strongly represented in the currently available UK collections dataset. While this is a limitation, we do not believe that it materially affects the points made about data value — a range of research topics and techniques were discussed, and the benefits and opportunities identified appear highly transferable to other types of research and specimens.

2.4 Data analysis

2.4.1 Data engineering

A repeatable data engineering process was followed:

1. **Inputs sourcing**: Identify the open access data sources and documentation necessary for the analysis.
2. **Data extraction**: Extract raw data from the different data sources using API endpoints using the relevant parameters defined in the API documentation for the specific point in time, or via direct downloads.
3. **Data cleaning**: Clean the extracted data and complete the necessary joins between the datasets.
4. **Data analysis**: Process and analyse the cleaned, extracted data in the defined data platform to showcase the analysis using a python compatible data analysis tool, e.g., Jupyter notebook. Data can also be exported as a CSV for upload to Excel.
5. **Output generation**: Save analysed data in CSV format and store in the defined object storage, e.g., Amazon Web Services (AWS) S3, Azure Data Lake, Google Cloud Platform (GCP) cloud storage.
6. **Data visualisation**: Fetch the analysed data from the object storage and then visualise using VizX, a python-based visualisation package based on Plotly.

Further details can be found in the supplementary materials (Suppl. materials 1, 2). Details of the code used to extract the data have been documented, but it should be noted that the code used is not production ready and should not be expected to follow best software engineering principles like modularity and unit testing. The code is meant only for re-running the analysis.

2.4.2 Data considerations and limitations

It should be noted that, owing to the volume of GBIF occurrences, raw data could not be downloaded at the occurrence level for this study. Analyses have been conducted on aggregated dataset information or extracting counts from GBIF API end points. Further considerations and constraints in relation to particular areas of our analysis are set out below.

2.4.3 Citations & publications

GBIF provides data on publications citing GBIF datasets, used in this study to address our questions about how UK collections data are currently being used in research. Query parameters used for this study can be found in "01_Overall_GBIF_data_extraction.ipynb" and "00_GBIF_UK_data_extraction.ipynb" in Suppl. material 1. For general use, it is possible to search for particular institutions under 'publishers' from the GBIF homepage and to see citations of their dataset(s), or to search for literature under 'resources'. GBIF also summarise key research uses and citations in their Science Review publications (Secretariat 2021).

There is of course a time lag between data upload and citation in publications. Publications refers to all relevant forms of literature, including journal articles, books, conference proceedings, preprints, reports, and others. Where relevant, our findings specify whether we were looking at all publications or specifically at peer-reviewed publications. Peer-reviewed publication data was taken from the GBIF web portal by setting both the “Peer-reviewed” filter to “Yes”, and the “Literature type” to “Journal article”.

Publication citations are made at the dataset level, not to individual specimen occurrences. This means that it is not possible from analysis of GBIF data to do an analysis of citations to individual specimens and to determine which, or how many, individual specimen records were cited. If a dataset contains both observations and specimens, publications cannot be split into those with citations to observations and those with citations to specimens — however the aggregate data uploaded by the twelve UK institutions considered in this report is 99% specimen data, indicating that publications with citations to these UK institutions’ datasets are citing specimens.

Where publications cited multiple institutions or multiple datasets within the same institution collection, these publications have been de-duplicated as necessary to show accurate numbers of publications.

A publication can have multiple topic tags. Therefore, a single publication may contribute to the count of several different topics. 99% of publications on GBIF have at least one topic tag.

In addition to direct citations, we also looked at onward citations of papers that cite GBIF data, to consider their ongoing impact. Numbers of onwards citations were taken from Google Scholar on 09/01/2023 using publication DOIs taken from the GBIF web portal on the same date. 77% of publications taken from GBIF have affiliated Google Scholar data, and 74% of these have a number of onward citations greater than zero.

2.4.4 Downloads of UK collections data from GBIF

Publications and citations do not fully capture the use of the UK digitised dataset on GBIF — the first step in use is usually downloads, not all of which result in publications. Download event numbers are therefore much higher than publication figures; for example, the number of downloads events is 254 times larger than the number of citations for Natural History Museum data. GBIF download data were therefore also analysed to consider the wider importance of UK collections data. Download data were taken from the GBIF API Endpoint on 12/12/2022.

Download events are by dataset. This is not representative of the number of occurrence records that have been downloaded, since any number of occurrence records can be downloaded in one download event. The Natural History Museum has one dataset in its collection on GBIF (all data uploaded are treated as a single set representing the NHM collection), therefore, the number of download events for this dataset is the same as the number of download events for the NHM collection as a whole. Other institutions have more than one dataset, which can result in higher numbers of download events linked to their collection.

2.4.5 Researchers and affiliation

We analysed researcher affiliation to understand the geography of collections data usage and to be able to understand and assign benefit to UK-affiliated researchers. GBIF and Crossref data were used for this.

Researcher geography data on GBIF were available for 87% of publications citing UK institutions. GBIF only provides the set of countries from which a publication's researchers were affiliated; however, it does not provide the number of researchers affiliated with each country. Researcher geography data, where aggregated, has been de-duplicated such that publications with multiple researchers affiliated to countries in the same region are only counted once. For example, a publication with researchers from France and Spain will only be counted a single time in the number of publications with researchers affiliated to Europe. These data have not, however, been deduplicated between regions. For example, a publication with researchers from France and Brazil will be counted once in the number of publications with researchers affiliated to Europe, and once for South America. This means that the sum of publications across regions is greater than the true number of publications (2,710 publications).

Crossref provides the number of unique researchers affiliated with each country, for a given publication (for query details see "02_Crossref_data_extraction.ipynb" in Suppl. material 1). 84% of publication DOIs taken from GBIF have affiliated researcher data on Crossref. 33% of these have data on country of researcher affiliation.

For analyses showing country of researcher affiliation, data were taken from Crossref on 12/12/2022 using publication DOIs taken from GBIF API Endpoint on the same date.

2.4.6 Calculations of value

Simple arithmetic calculations were performed to arrive at key insights (for example, rates of citations per day), and estimated savings and benefits (Suppl. materials 13, 14). As a conservative approach to quantify estimated efficiency benefits, we assumed that each citation of the relevant UK collections datasets would require a single visit to a relevant institution if the data were not already available. Visit costs were estimated based on unpublished data from the Natural History Museum's leadership of the EU SYNTHESYS programme (Smith et al. 2019) — this takes account of the cost for a researcher to make visits averaging eleven days within Europe, and for relevant institutions to host them. Further details are included at section 3.3.2 below.

3. Results

3.1 UK Institutional data representation in GBIF

The twelve institutions covered in this analysis have together uploaded at total of 7.6 million occurrences to GBIF. This equates to less than 6% of the 137 million specimens identified in the DiSSCo UK surveys (a greater percentage of these specimens have some form of digital record, however these data are not published to GBIF and the majority are not available as FAIR data). Three institutions (The Natural History Museum, Royal Botanic Garden Edinburgh, and Royal Botanic Gardens, Kew) have contributed 95% of total UK specimens uploaded to date, with over five million specimens (66% of all UK specimen uploads to date) from the Natural History Museum alone.

By uploading specimens, UK institutions are increasing the diversity of GBIF. 88% of all GBIF uploads are observations (with the remaining 12% being specimens). More than 99% of the uploads from the twelve UK institutions are specimens. Occurrences uploaded to GBIF by UK institutions make up only 0.3% of all occurrences on GBIF, but 3% of all specimens uploaded to GBIF.

3.1.1 Geographic Diversity

UK institutions upload specimens with a different and more diverse geographical makeup than GBIF's occurrence uploads as a whole. Specimens have been collected from 7 continents, and 248 territories and countries (including regions like the Vatican City, Greenland, Antarctic regions and other small territories like Sint Maarten). UK institutions provide a significantly higher percentage of specimens from South America, Antarctica, Asia and Africa (11%, 23% and 16% respectively) than wider GBIF occurrence data (4%, 5% and 3% respectively) (Fig. 2).*[2] This coverage reflects the historical legacies of collecting including colonialism; however, making these global data available to communities of origin and to wider global users is an important part of addressing these legacies (The Natural History Museum's principles for understanding and sharing the collection can be found at https://www.nhm.ac.uk/about-us/governance/understanding-and-sharing-the-collection.html. There is a very substantial and rapidly developing body of discussion and research in this area, but see also for example recent work by Nicolson et al. 2023).

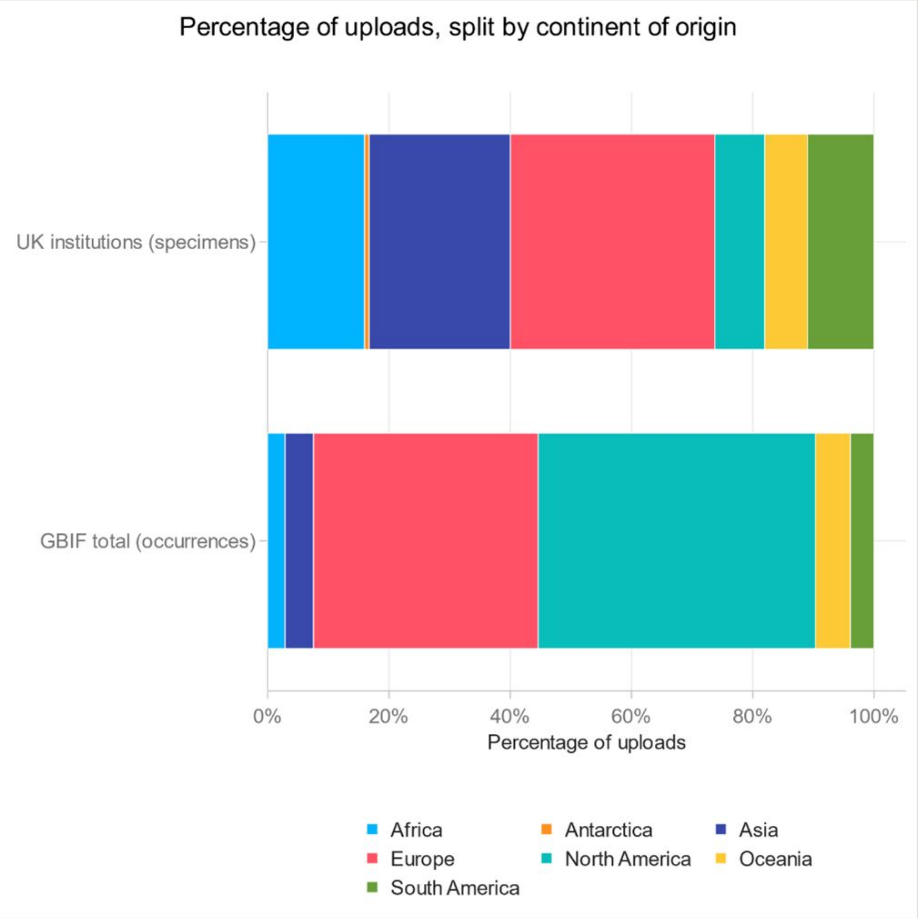

Figure 2. doi

GBIF uploads by continent of origin showing that UK collections have a relatively high proportion of African, Antarctic, Asian and South American specimens compared to GBIF occurrence data overall. *Note: Incertae sedis (Latin for 'of uncertain placement') is a taxonomic grouping used when a specimen's broader relationship to another taxonomic group is unknown.* See Suppl. materials 4, 5.

3.1.2 Taxonomic Diversity

UK institutions currently upload a higher percentage of occurrences than GBIF as a whole in two main kingdoms: Plantae (24 percentage points more), and Chromista (2 percentage points more), adding to the taxonomic diversity of occurrences (Fig. 3). Once again, this reflects biases and practices over time in both collecting and digitisation, including long-term focus on workflows and community collaboration around digitisation of pressed plant specimens (see for example Ryan 2018), and there are opportunities to understand these better and contribute to dialogue by making these data globally accessible.

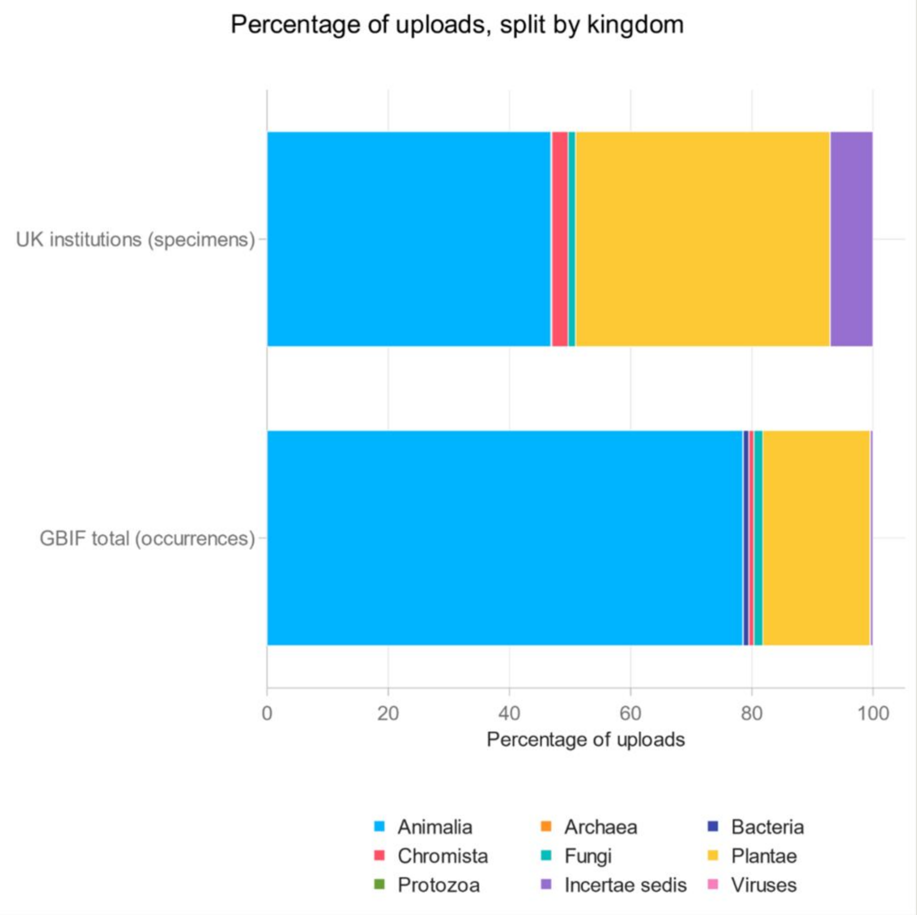

Figure 3. doi

GBIF uploads by taxonomic kingdom, showing that UK collections have a relatively higher proportion of Plantae and Chromista compared to GBIF occurence data overall. See Suppl. material 6.

3.2 Usage of UK Collections Data

3.2.1 Citations

Both uploads to GBIF from institutions around the world, and the number of publications citing those data, have increased over time (Fig. 4). This aligns with the hypothesis that increasing uploads leads to increasing usage of digitised natural history data in scientific research. The number of publications each year using GBIF data has increased significantly between 2015 and 2020, growing by 866% over this period. The number of publications each year citing the UK datasets within the scope of this study has also seen a steep rise, growing from 10 publications in 2015 to 799 in 2022 — an average of 2.2 publications every day of that year, including 587 peer-reviewed journal articles (an average of 1.6 per day).

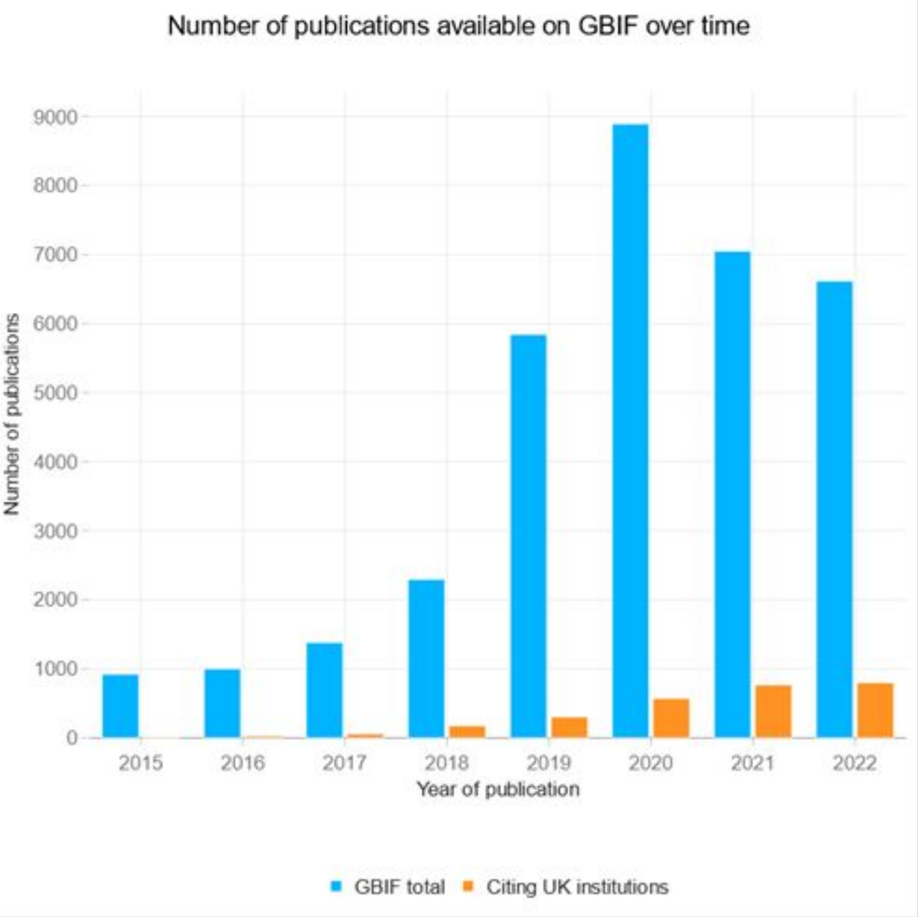

Figure 4. doi

Number of publications citing GBIF data over time, and number of those citing UK collections data. *Note: This chart shows the number of new publications uploaded each year available on GBIF or with citations to UK institution data, it is not cumulative over time.* See *Suppl. materials* 7, 8.

Despite a decline in the total number of publications citing GBIF data from 2020 to 2022, the number of publications citing UK institution uploaded data has increased by 40% over this time period. This may indicate that UK institution uploaded data provides additional value relative to the overall dataset available on GBIF.

The ratio of records uploaded to citations provides an approximate metric of “usefulness” of the data which has been uploaded to GBIF. It shows, on average, the number of additional specimens which had to be uploaded to gain each additional publication with a citation to that dataset. GBIF’s total collection has 7,411 uploads per publication, while the UK institutions have only 2,816 specimens per publication, some 2.63 times fewer. GBIF’s total collection has 17,248 records uploaded per peer reviewed journal article, while the UK institutions have only 3,943 specimens per peer reviewed journal article, over four times fewer. These proxy measures again suggest the high relative value of UK collections data.

Looking at the number of citations for each of the UK institutions in this study, the order is the same as that for volume of uploads, strongly indicating that there is demand for UK specimen data such that as UK institutions upload more specimen data, they also see more publications citing those data (Fig. 5). Data from The Natural History Museum has been cited by 2,253 publications to date, 86% of the total publications citing UK data.

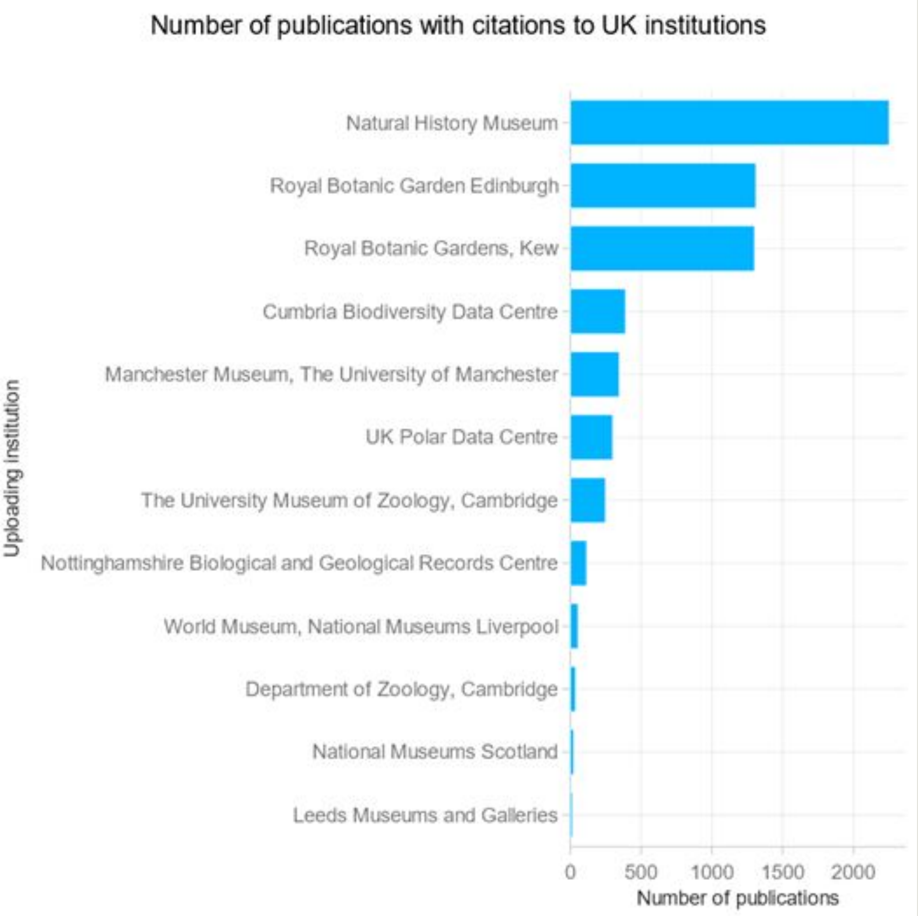

Figure 5. doi

The number of publications per UK institution in this study that cite their GBIF dataset(s). See *Suppl. material* 9.

Overall, the UK’s digitised specimens are in demand and are highly used by researchers around the world. There are 2,710 publications citing UK institutions’ data uploaded on GBIF, 1,932 of which are peer-reviewed journal articles. While the UK specimens examined in this study make up just 0.3% of total GBIF occurrences, they are cited in 12% of the peer reviewed publications that cite GBIF, indicating that UK specimen data punches some 40 times above their weight compared to wider occurrence records.

We also examined onward citations and Altmetrics to consider the onward impact of publications citing UK collections data, however the time lag and continued growth of these metrics over time mean that a point in time analysis does not yield significant insights, particularly given that much of the growth in publication volumes themselves is comparatively recent. Of the 2,710 publications citing UK collections data, 57% had at least one onward citation that we were able to trace. 150 publications had gone on to receive over 30 onwards citations, and 21 had received over 100 citations, suggesting that research citing UK collection is having onward impact that enhances and enables further research. Altmetric scores for these publications showed no significant differences to the average overall.

3.2.2 Downloads

While publications and citations are the most reliable indicator of usage and research impact, they do not fully capture the use of the UK digitised dataset on GBIF — use typically starts with downloading data, which may or may not eventually lead to a publication. The number of download events is 254 times larger than the number of citations for Natural History Museum (572k download events versus 2,253 citations). The number of download events again appears to track closely with the number of specimens uploaded by each institution, with the order of institutions by download events being the same as that by upload.

The number of download events for UK institutions data has been growing steadily since 2014. The Natural History Museum saw an especially high increase in the number of download events recently, from 92,000 in 2021 to 154,000 in 2022, a 66% increase and an average of one download every 3 minutes and 24 seconds.

3.2.3 Research topics supported

Specimen data uploaded by UK institutions contributes to publications across a variety of highly impactful and relevant research topics like ecology, conservation and climate change (Fig. 6).

1,549 publications with citations to UK institution data have received one or more onward citations by other publications. This accounts for 57% of the total 2,710 publications with citations to UK data. 150 publications have gone on to receive over 30 onwards citations and 21 publications have gone on to receive over 100 citations. This suggests that much of the research conducted using UK institution data has been used in multiple onwards citations and has likely enhanced or enabled further research.

3.2.4 Where are UK collections data used?

UK institution data uploaded to GBIF is used to support research in the UK, on six continents and in 160 countries and territories around the world (Fig. 7). The USA is the top contributor of publications which cite UK institution data at 681 publications, followed by the UK at 336 publications (Fig. 8). Similarly, the USA has the highest number of researchers citing UK institution data at 1,463 researchers, followed by the UK at 327 researchers.

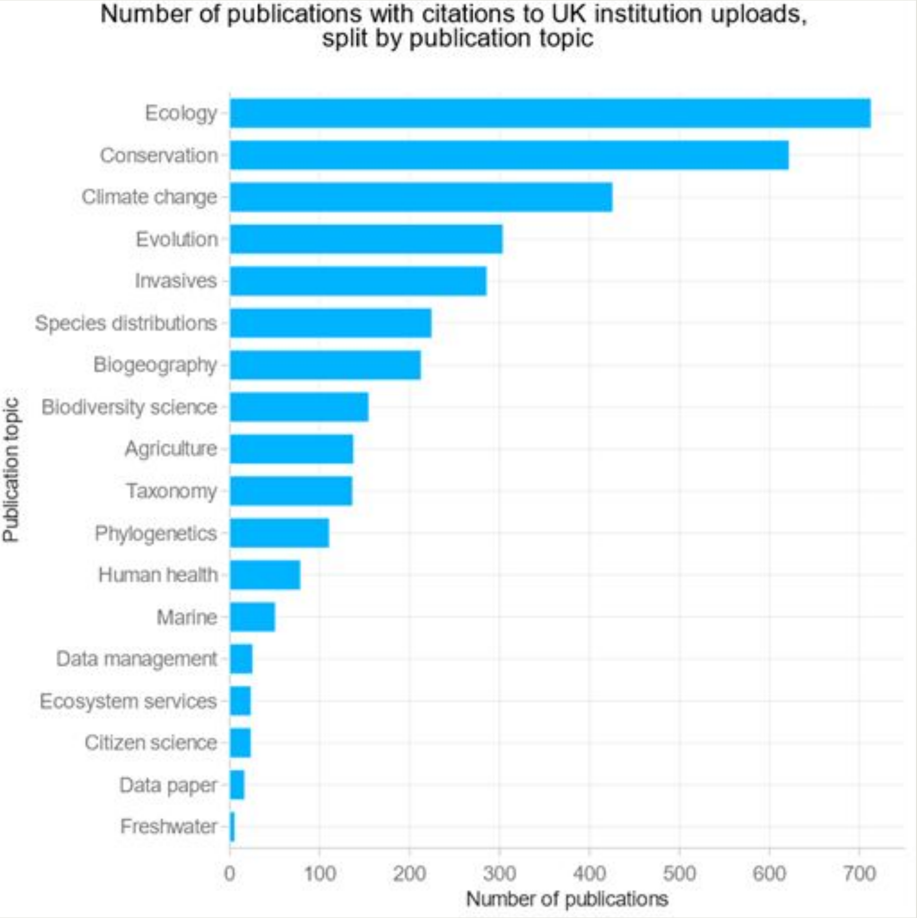

Figure 6. doi

The number of publications by topic tag (broad research area) that cite UK institution GBIF datasets, showing high numbers of publications relevant to ecology, conservation, climate, evolution, and invasive species. *Note: 99% of publications on GBIF have at least one topic tag. A publication can have multiple topic tags, so a single publication may contribute to the count of several different topics.* See *Suppl. material* 10.

Extrapolating the percentage of UK researchers where affiliation can be determined across the total number of publications, it is estimated that 1,200 UK-affiliated researchers have been supported by UK collections data.

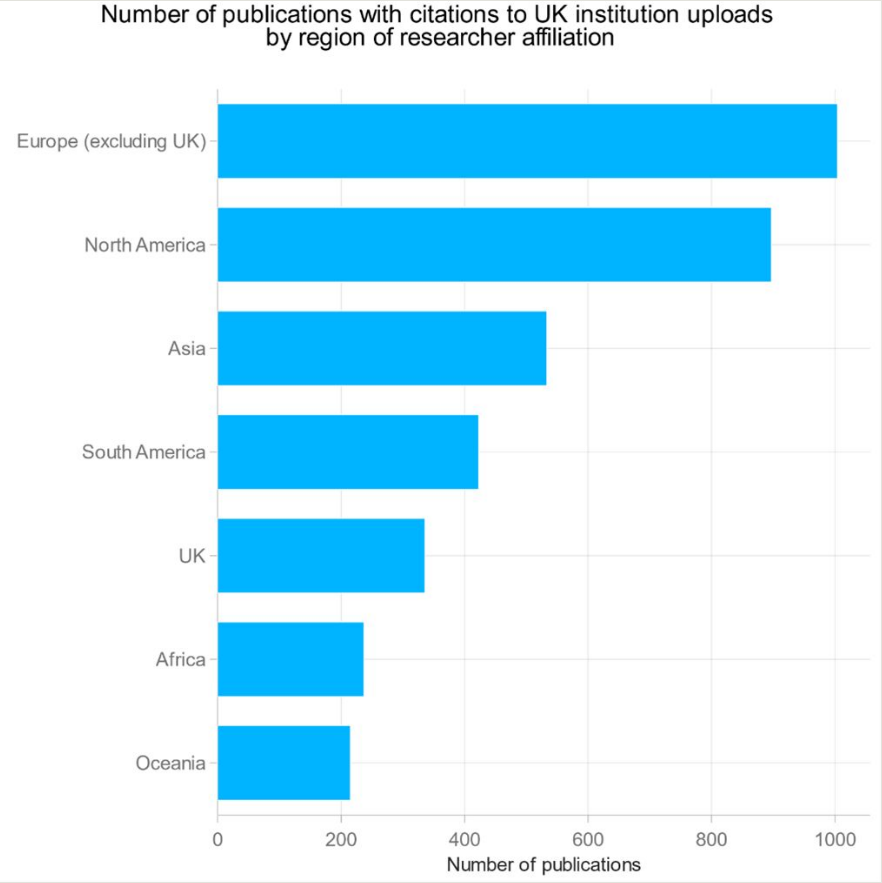

Figure 7. doi

Number of publications citing UK institution GBIF datasets by region of researcher affiliation (at continent level plus UK). See Suppl. material 11.

1,137 publications cite GBIF data from only one of the twelve UK uploading institutions; however, 1,573 publications (58%) have citations to GBIF data uploaded by more than one UK institution. This suggests that UK institution data is often used in combination for research and publications.

3.3 Value and benefits to research

In estimating the value created by digitising UK institutions' collections we considered the existing investment in digitisation by the Natural History Museum (it was not possible to source comparable figures for the other eleven UK institutions within the scope of this study). Secondly, we consider efficiencies created for researchers and institutions:

- Collective investment by institutions in digitisation is more cost-effective that efforts by individual researchers to create digitised datasets.

- Research can be conducted more efficiently because of the free access to digitised specimen data, e.g., with fewer visits required for researchers. This allows more research to be conducted and reduces hosting costs for institutions.

Finally, we considered the economic value of the research supported by digitised data; and broader potential benefits to society of the digitised data beyond research (e.g. for education and entertainment, as well as potential to unlock further innovation and value in as yet unknown use cases).

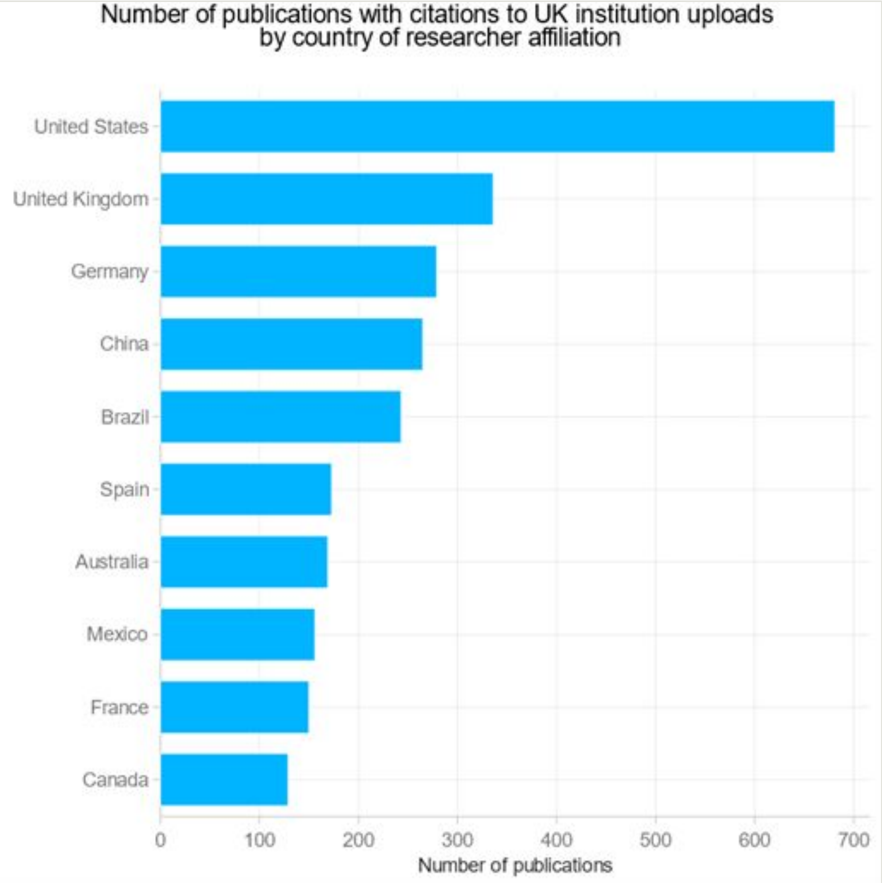

Figure 8. doi

Publications citing UK institution GBIF datasets by country of researcher affiliation. *Note: The full dataset* (Suppl. material 12) *includes 160 countries in total. There is a long tail of 150 countries not shown in this chart*.

3.3.1 Investment in digitisation

Institutions around the UK have invested to digitise their collections and make them free to access on GBIF — the Natural History Museum alone estimate that they have invested approximately £5.4 million in digitising the collection over the last nine years (between financial years 2014–15 and 2022–23). The majority of this is investment from core Museum funds in the central digitisation team, not including the work of many other staff that is key to enabling digitisation, including curators, data managers, and the teams managing the Natural History Museum data portal and other collection management systems. This also includes over £2 million which has been raised (e.g., through grants and philanthropy) and spent on smaller digitisation projects that explore a new technique or capture more data (such as 3D digitisation of fossil mammals collected by Darwin - Pavid (2018)), or very closely related projects such as innovation in machine learning and AI related to extracting data from natural science collections.

To date there has been no dedicated public investment at the national level in digitisation of UK natural science collections. The Royal Botanic Gardens, Kew, were recently awarded £10 million through DEFRA towards their herbarium digitisation programme, to be spent during 2022–23 and 2023–24 (Figg and Hirschler 2023). The DiSSCo UK initiative is making the business case for national level investment (Smith et al. 2022).

3.3.2 Efficiency benefits, including benefits to UK affiliated researchers

As set out above, researchers benefit from the investment in digitisation that creates openly available collections data — this value is not known across the twelve relevant UK institutions but is in excess of £5 million for the Natural History Museum alone.

Investment in digitisation by institutions is more cost-efficient than digitisation by individual researchers, due to best practice workflows, economies of scale, team capabilities and not needing to travel or to ‘pick and choose’ individual specimens. Moreover, the data are reusable to all scientists (and others) when institutions make them freely accessible via GBIF, rather than being individually prepared and privately held or uploaded to disparate repositories. For individual researchers to create the datasets that they have downloaded and cited from GBIF themselves would require significantly more funding, due to the scale and efficiency of institutions’ digitisation; the reusability of data; and the need in many cases to create data from multiple organisations.

In practice, some of the research citing UK collections via GBIF would not have been possible at all without these data being available — either it would not have been conceived (for instance, AI uses of collections images in Wilson et al. 2022); or would have been prohibitively expensive and/or time consuming to collect, owing to the volumes of data and the number and locations globally of institutions where relevant collections are held. It is not possible to accurately assess these proportions in relation to the UK datasets covered in this study, however research since published about the value of GBIF data as a whole (GBIF Secretariat and Deloitte Access Economics 2023) shows that nearly 50% of users would not have been able to achieve their research outcomes without GBIF data, and another 41% could only have done so with significantly increased time and effort.

To produce a publication that concerns specimens, especially in large quantities, researchers need access to information about those specimens. The needs of each publication will vary considerably, but without digital access, physical visits are often required to collect such information. While some publications would not necessitate a physical visit, some would require multiple visits to collections in different locations — and, at the very least, most would require data collection by someone working with the relevant collection. Physical visits incur significant costs in researcher time, researcher costs (e.g., travel and subsistence), and costs in time and effort to the host institution.

We therefore assumes as a conservative approach that each citation of the relevant UK collections datasets would require a single visit to a relevant institution if the data were not already available. Visit costs were estimated based on unpublished data from the Natural History Museum's leadership of the EU SYNTHESYS programme (Smith et al. 2019). We estimated an average cost per visit of £6,500 including travel and subsistence costs, researcher time, and host costs.

Based on 2,710 publications each saving £6,500, this equates to savings of £17.6 million for researchers and hosts across all the publications citing UK collections. Looking just at savings to researchers (not hosts), extrapolating publications with a UK affiliated author (see Section 3.2.3), we estimate £1.4 million in savings to UK-affiliated researchers (Suppl. material 14).

It is likely that looking only at citations underestimates the efficiency benefits created. In theory, anyone downloading the data would otherwise have had to visit or otherwise request its creation. At the point of this analysis, there had been 571,518 download events of NHM data. Taking an approximate researcher time/cost of £3,500 (excluding host costs) and applying Frontier Economics assumptions of benefits in the range of 5–12.5% (Popov et al. 2021) provides an estimate of researcher efficiencies reaching some £100–£250 million.

3.3.3 Economic and wider benefits

Time and money saved for researchers can go towards further research, with an estimated 20–40% return on investment for society (Popov et al. 2021).

In total, Popov et al. (2021) identified potential economic value from digitising all UK collections of £2 billion over 30 years, across five key themes where digitised specimens can have major impact. Many of the 2,710 publications completed over the last seven years directly address those themes, including 622 publications addressing biodiversity conservation; 286 on invasive species; 79 relevant to medicines discovery; and 138 relevant to agricultural research and development. In agriculture, for example, the number and breadth of papers indicate that these estimates are conservative, considering a single use case in relation to wild relatives of key crops to predict impact of some £20–70 million in agricultural research and development, where clearly additional potential uses cases exist which were not analysed.

Other publications also contribute less directly to these themes, and/or to related areas, e.g., those on biodiversity science and ecology. Mineral exploration is an area that is not currently represented by analysis of GBIF biodiversity data. And citations of UK collections data via GBIF also cover a range of other topics as set out above (Fig. 6), including 426 publications on climate change, where in 2016, the UN estimated the cost of adapting to climate change to be $140–300 billion per year given current trends (United Nations Environment Programme 2021). These topics also align to the UN Sustainable Development Goals (https://sdgs.un.org/goals) — natural science collections data are not only relevant to understanding Life on Land and Life Below Water, but also contribute to Climate Action and to many other goals including reducing hunger, poverty, ill health and inequality.

Wider uses and benefits beyond research are outside the scope of this study, but it can easily be seen that there is scope for digital collections to play a role in education, engagement, the arts and innovation, with further benefits to the economy and society in the UK and beyond. The full possibilities and economic impact of digitised natural science collections cannot be anticipated.

3.3.4 Interview insights

Our interviews highlighted the breadth of benefits from digitised UK collections data, both to individual researchers, their areas of research, and wider society (see interview case studies 1 and 2). We spoke to a climate change researcher in Indonesia (who did not wish to be named), who told us that digital data directly inform the conservation and policy priorities for her work, as well as saving time and money.

Alexandre Antonelli is Director of Science at the Royal Botanic Gardens, Kew, and a biodiversity researcher. He uses GBIF data almost every week, both for his research and in answering policy and media questions about biodiversity. Much of his research would not be possible without digital data. He told us that digital integration of collections, and continued effort to include smaller but locally important collections around the world, are key to having a complete picture — and the future opportunities are extraordinary, particularly when factoring in AI tools to speed the processes of recording and extracting information, and using it e.g., for species identification.

Colin Khoury is the Senior Director of Science and Conservation at San Diego Botanic Garden, and a plant and conservation scientist who focuses on food crop diversity. He has used digitised data throughout his career, including to understand the distribution and conservation status of wild relatives of crops (e.g., Castañeda-Álvarez et al. 2016,Khoury et al. 2020), as well as to form global conservation indicators about this biodiversity (Khoury et al. 2019) — having data available can save years of effort on international studies. The availability of data continues to improve but more resources are needed to fill gaps and to include key data such as accurate geo-referencing. Collaboration to release specimen data can make a huge positive impact.

Interview case study 1: UK collections data, innovation and climate change

Phillip Fenberg, Researcher at the University of Southampton and Science Associate at the Natural History Museum, uses collections data, combined with occurrence records from monitoring and other key datasets such as temperature, to ask questions such as how organisms respond to climate change. Natural science collections enable these questions to be studied over periods of many decades.

While Phillip's original PhD research involved him visiting museum collections to gather specimen data in person (e.g., body size measurements, occurrence records), digital collections data have transformed the efficiency and scope of what is possible. The combination of digital collections images and new computer vision techniques for analysis is incredibly powerful, allowing for previous hypotheses to be tested at scale. For example, Phillip and his team used the NHM iCollections dataset of over 180,000 butterfly specimens (Paterson et al. 2016) to show how the adult body size of UK butterflies responds to warmer temperatures (Wilson et al. 2022).

Use of an innovative computer vision pipeline — 'Mothra' (https://github.com/machine-shop/mothra) — showed that it was possible to accurately detect specimens in images, set the scale, measure wing features such as forewing length, and identify the sex. Not only that but like for like comparison of forewing length measurements showed that Mothra could complete work in a week (or less, if more than 10 analyses had been run in parallel on a computer cluster) that would take a human some 3,000 hours, or around two years (assuming eight hours a day with no breaks, and only one measurement (forewing length) per specimen).

Phillip is looking forward to the expansion of digital collections image data, particularly the possibilities that will come with increased linkage between genetic and image datasets; the greater integration of AI into taxonomic work; and useful metadata such as information on what proportion of any particular collection set has been digitised.

Interview case study 2: collections data and conservation

Conservation scientist James Westrip (interviewed 6th January 2023) is a 'superuser'; an author of some 117 papers citing UK collections data, owing to his work with the International Union for Conservation of Nature (IUCN), assessing species for the 'Red List' on species conservation status.

Red List assessment demands good data about species distribution over time — many species are data deficient and cannot be assessed, meaning that risks of biodiversity loss are greater than reported, and key conservation actions may be missed.

Since 2019, GBIF data have been transforming how James (and his colleagues) do their work. Geographical data is the most critical for them — ideally in the form of a fully geo-referenced latitude and longitude for specimen collection, but descriptions from labels can be sufficient. This enables species distribution and prevalence to be examined over time, based on different collecting events. Habitat data can also be helpful — one of the benefits of collections data is their coverage of rarer species that are not often observed by humans otherwise.

These data are helping to make the Red List more comprehensive and in particular more representative of species diversity, covering for example more insects, plants and fungi as well as vertebrates which were traditionally well-represented.

And they make the work much more efficient — combining digital specimen data with mapping tools reduces the time taken for many species assessments from weeks to just a day or two. While data quality isn't always perfect, James has processes to identify and remove outliers. Digital data have reduced the checks needed with collections staff. This work directly informs policy decisions, so the more data are available, and the more species covered, the more impactful it will be.

4. Discussion

Going back to our research questions around the volume of data currently available from UK natural science collections; the uses of these data; and their value in terms of research efficiency, research effectiveness, and wider impact; the demand for and potential of UK natural science collections data are very clear.

We can see major usage, research impact, and benefits for research efficiency, the economy and society even from less than 6% of relevant collections' data, which make up only 0.3% of total occurrences on GBIF. Even at this small percentage, UK collections data are contributing to the historic, geographic and taxonomic diversity of GBIF; being cited at a rate 40 times that of other GBIF occurrence data, in thousands of publications (2.2 per day on average for the Natural History Museum alone), across topics that reflect the key challenges facing humanity and the planet; and yielding more than £18 million of efficiency savings for researchers.

It is estimated that $44 trillion of economic value generation (or over 50% of the world's GDP) is moderately or highly dependent on nature, with biodiversity loss and ecosystem collapse among the key challenges that the planet faces (World Economic Forum 2020), further illustrating the economic potential and benefit of research underpinned by these collections for the UK and globally.

Understanding what is in collections now, in the UK and globally, is also key to understanding what is needed as we collect for the future, to underpin policy and investment decisions in future centuries (Johnson et al. 2023).

Digitisation also brings wider benefits than those examined in this study. As the custodians of collections from around the globe, digitisation of natural science collections held in the UK supports the involvement of communities of origin and the enrichment of collections through the knowledge and experience of these communities and of experts from the global network, including the opportunity to understand and address biases and the legacies of colonialism. The broader significance of these collections for education, the arts and humanities, and of course leisure and wellbeing can also only be enhanced by the availability of digital collections data for discovery and access.

While this study has yielded useful insights into the current use and impact of UK collections data, it does face limitations. GBIF data do not reflect the full breadth of UK collections in the geo-sciences, underestimating benefits in this important area. It has not yet been possible to consistently track the use of particular specimens (or groups of specimens), or to reach granular insights about the usefulness of particular data fields, although we know from qualitative discussions that, for example, geo-referencing is frequently of high value. Insights therefore cannot yet be used to inform detailed prioritisation of digitisation activities.

Stakeholders also expressed interest in better and wider metadata, for example to understand the percentage of a particular collection type that has been digitised (thus understanding not only the data available but the data gap, and what might become available in future). Usage of collections outside those published on GBIF is much harder to quantify, with high variety in availability, quantity and measurement — this also applies to physical use of specimens, where visits are recorded differently by different institutions or sections, and it can be hard to trace citations of physical material.

The Global Registry of Scientific Collections (https://www.gbif.org/grscicoll) combined with the nascent Latimer Core data standard (https://github.com/tdwg/ltc) for collections descriptions are promising community developments that aim to support the representation and discovery of natural science collections prior to full digitisation, by structuring data about higher level groups of objects within those collections, allowing a line of sight for discovery and use that can start at the level of a whole collection and link through to subgroups and individual items when data about these becomes available (Hobern et al. 2022, Woodburn et al. 2022).

It is also a welcome development that, while the SYNTHESYS access programmes for natural science collections across Europe are now complete, DiSSCo EU (https://www.dissco.eu/dissco/timeline/) and the UK AHRC Research infrastructure for conservation and heritage science (RICHeS) programme continue to develop avenues that can associate the digital discovery of specimens with targetted routes to physical access and enhanced analyses, associated with access to relevant facilities and labs (UKRI 2023)*[3].

UK Natural Science collections are joining forces through the Distributed System of Scientific Collections UK (DiSSCo UK) to set the vision and make the business case for investment in these collections as a distributed research infrastructure (Smith et al. 2022). Digitisation of UK natural science collections through DiSSCo UK will act as a ‘pathfinder’ for wider cultural heritage collections — both through the wide variety of natural science object types, and the fact that these are frequently held as part of multidisciplinary collections. Both the methods of digitisation and data mobilisation, the demonstration of impact from heritage collections, and the continued evolution of underpinning technological infrastructures and linkages, will have relevance to the wider movement represented by the AHRC's Towards a National Collection programme (https://www.nationalcollection.org.uk).

The UK has set itself the ambition to be a science and technology superpower (Council for Science and Technology 2021) and natural science collections present an opportunity for the UK to be at the forefront — but while the UK are thought leaders in collections digitisation e.g., the development of data standards and digitisation workflows, we are falling behind in the investment needed to unlock these incredible assets and the value that they can generate, both directly for the UK, and to underpin a future in which both people and planet thrive. The investment sought by DiSSCo UK, of the order of £155 million to digitise critical mass of UK natural science collections, will unlock at least a seven- to ten-fold economic return on investment, as well as efficiency savings for UK and global researchers and the potential for research innovation and studies that have not previously been possible.

Author contributions

Contribution types are drawn from the CRediT Contributor Roles Taxonomy.

- Helen Hardy: Conceptualization; methodology; supervision; writing (original draft); writing (review & editing).
- Laurence Livermore: Conceptualization; investigation; methodology; writing (review & editing).
- Paul Kersey: Conceptualization; methodology; writing (review & editing).
- Ken Norris: Conceptualization; methodology; writing (review & editing).
- Vincent Smith: Conceptualization; methodology; writing (review & editing).

Acknowledgements

The Natural History Museum would like to thank McKinsey and Company for analytical support. The Museum are responsible for the conclusions and recommendations of the research.

Thank you to our Natural History Museum colleagues who supported this work in a variety of ways including attending steering meetings, and signposting domain-specific data resources: Lisa French (https://orcid.org/0000-0001-7279-8582), Al Turner (https://orcid.org/0009-0009-4911-1696), Sarah Vincent (https://orcid.org/0000-0002-4012-0571), and Tara Wainwright (https://orcid.org/0000-0001-5816-2529).

We would like to thank our interviewees for their enthusiasm and insights into using digital collection data — thank you Alexandre Antonelli (https://orcid.org/0000-0003-1842-9297), Phillip Fenberg (https://scholar.google.co.uk/citations?hl=en&user=-CSvSUQAAAAJ&view_op=list_works), Colin Khoury (https://orcid.org/0000-0001-7893-5744), and James Westrip.

We would like to thank our colleagues at the Global Biodiversity Information Facility for their support with extracting and interpreting some of the relevant data.

Hosting institution

The Natural History Museum, London

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Castañeda-Álvarez N, Khoury C, Achicanoy H, Bernau V, Dempewolf H, Eastwood R, Guarino L, Harker R, Jarvis A, Maxted N, Müller J, Ramirez-Villegas J, Sosa C, Struik P, Vincent H, Toll J (2016) Global conservation priorities for crop wild relatives. Nature Plants 2 (4). https://doi.org/10.1038/nplants.2016.22
- Council for Science and Technology (2021) Strengthening the UK's Position as a Global Science and Technology Superpower. Council for Science and Technology. A letter sent to the Prime Minister sent in June 2021.. URL: https://www.gov.uk/government/publications/the-uk-as-a-science-and-technology-superpower
- Crossref (2021) About us. https://www.crossref.org/community/about/. Accessed on: 2023-8-09.
- Figg P, Hirschler B (2023) 5 ways digitising Kew's specimens can help save the world. https://www.kew.org/read-and-watch/5-ways-digitisation-save-the-world. Accessed on: 2023-8-09.
- GBIF Secretariat (2022) GBIF named a Global Core Biodata Resource. https://www.gbif.org/news/6PHdgoyIF6RmI7u4VOouuD/gbif-named-a-global-core-biodata-resource. Accessed on: 2023-8-09.
- GBIF Secretariat, Deloitte Access Economics (2023) Economic valuation and assessment of the impact of the GBIF network. Deloitte Access Economics. URL: https://www.gbif.org/news/5WZThcL928vmPnSvrGhZfE/report-reveals-return-on-investments-in-gbif
- Hardy H, Livermore L, Kersey P, Norris K, Pullar J, Smith V (2023) Users and uses of UK Natural History Collections – a Summary. https://doi.org/10.5281/zenodo.8403318
- Hobern D, Livermore L, Vincent S, Robertson T, Miller J, Groom Q, Grosjean M (2022) Towards a Roadmap for Advancing the Catalogue of the World's Natural History Collections. Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e98593
- Johnson K, Owens IP, The Global Collection Group (2023) A global approach for natural history museum collections. Science 379 (6638): 1192-1194. https://doi.org/10.1126/science.adf6434
- Khoury C, Amariles D, Soto JS, Diaz MV, Sotelo S, Sosa C, Ramírez-Villegas J, Achicanoy H, Velásquez-Tibatá J, Guarino L, León B, Navarro-Racines C, Castañeda-Álvarez N, Dempewolf H, Wiersema J, Jarvis A (2019) Comprehensiveness of

conservation of useful wild plants: An operational indicator for biodiversity and sustainable development targets. Ecological Indicators 98: 420-429. https://doi.org/10.1016/j.ecolind.2018.11.016

- Khoury C, Carver D, Greene S, Williams K, Achicanoy H, Schori M, León B, Wiersema J, Frances A (2020) Crop wild relatives of the United States require urgent conservation action. Proceedings of the National Academy of Sciences 117 (52): 33351-33357. https://doi.org/10.1073/pnas.2007029117
- Nicolson N, Trekels M, Groom Q, Knapp S, Paton A (2023) Global access to nomenclatural botanical resources: Evaluating open access availability. PLANTS, PEOPLE, PLANET https://doi.org/10.1002/ppp3.10438
- Paterson GLJ, Albuquerque S, Blagoderov V, Brooks SJ, Cafferty S, Cane E, Carter V, Chainey J, Crowther R, Douglas L, Durant J, Duffle L, Hine A, Honey M, Huertas B, Howard T, Huxley R, Kitching I, Ledger S, McLaughlin C, Martin G, Mazzetta G, Penn MG, Perera J, Sadka M, Scialabba E, Siebert D, Sleep C, Toloni F, Wing P (2016) iCollections. Natural History Museum https://doi.org/10.5519/0038559
- Pavid K (2018) The giant fossil mammals that inspired Charles Darwin's theory of evolution. https://www.nhm.ac.uk/discover/news/2018/april/giant-fossil-mammals-inspired-charles-darwin-theory-evolution.html. Accessed on: 2023-8-09.
- Popov D, Roychoudhury P, Hardy H, Livermore L, Norris K (2021) The Value of Digitising Natural History Collections. Research Ideas and Outcomes 7 https://doi.org/10.3897/rio.7.e78844
- Ryan D (2018) Global Plants: A Model of International Collaboration. Biodiversity Information Science and Standards 2: 28233. https://doi.org/10.3897/biss.2.28233
- Secretariat GB (2021) GBIF Science Review. GBIF Science Review https://doi.org/10.35035/w3p0-8729
- Smith V, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston E, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e46404
- Smith VS, Hardy H, Wainwright T, Livermore L, Fraser N, Horák J, Aspinall J, Howe M (2022) Harnessing the power of natural science collections: a blueprint for the UK. Zenodo https://doi.org/10.5281/zenodo.6472238
- UKRI (2023) Host facilities as part of our heritage science infrastructure. https://www.ukri.org/opportunity/host-facilities-as-part-of-our-heritage-science-infrastructure/. Accessed on: 2023-8-09.
- United Nations Environment Programme (2021) Adaptation Gap Report 2020. United Nations Environment Programme. https://doi.org/10.18356/9789280738346
- Williams C (2016) The Altmetric score is now the Altmetric Attention Score. https://www.altmetric.com/blog/the-altmetric-score-is-now-the-altmetric-attention-score/. Accessed on: 2023-8-09.
- Wilson R, de Siqueira AF, Brooks S, Price B, Simon L, van der Walt S, Fenberg P (2022) Applying computer vision to digitised natural history collections for climate change research: Temperature-size responses in British butterflies. Methods in Ecology and Evolution 14 (2): 372-384. https://doi.org/10.1111/2041-210x.13844
- Woodburn M, Buschbom J, Droege G, Grant S, Groom Q, Jones J, Trekels M, Vincent S, Webbink K (2022) Latimer Core: A new data standard for collection descriptions. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91159

- World Economic Forum (2020) Nature Risk Rising: Why the Crisis Engulfing Nature Matters for Business and the Economy. World Economic Forum. URL: https://www.weforum.org/reports/nature-risk-rising-why-the-crisis-engulfing-nature-matters-for-business-and-the-economy/

Supplementary materials

Suppl. material 1: Jupyter Notebooks doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: Zip file (.zip) containing Jupyter notebooks (.ipynb)
Brief description: Thirteen Jupyter notebooks (*.ipynb files) for the quantitative data extraction, cleaning, analysis and visualisation.
The code used is not production ready and should not be expected to follow best software engineering principles like modularity and unit testing. The code is meant only for re-running the analysis with underlying data.
Download file (188.82 kb)

Suppl. material 2: Data engineering user guide doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .docx (Word document)
Brief description: Document describing the data extraction framework, data landscape for analysis, data pipeline architecture, and infrastructure setup (compute & storage, Conda environment, Jupyter and Kernel, libraries used).
Download file (969.84 kb)

Suppl. material 3: Users and usage of UK collections data – researcher interview guide doi

Authors: Helen Hardy
Data type: .docx (Word document)
Brief description: Document with questions used for qualitative insights on researchers' experience working with digitised specimens.
Download file (23.71 kb)

Suppl. material 4: GBIF occurrences by continent of origin doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (44.55 kb)

Suppl. material 5: UK institution specimens by country of origin doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (44.55 kb)

Suppl. material 6: UK and GBIF specimen count by kingdom doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (60.31 kb)

Suppl. material 7: GBIF publications over time doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (110.00 bytes)

Suppl. material 8: UK peer reviewed journal articles over time doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (96.00 bytes)

Suppl. material 9: Publications for each UK institution doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (508.00 bytes)

Suppl. material 10: Publications by topic doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (329.00 bytes)

Suppl. material 11: Publications by region of researcher doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (1.19 kb)

Suppl. material 12: Researchers by country doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .csv
Download file (1.19 kb)

Suppl. material 13: Calculations for key insights doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .xlsx (Excel)
Download file (21.41 kb)

Suppl. material 14: Calculations for translation to value doi

Authors: Helen Hardy (with McKinsey & Company)
Data type: .xlsx (Excel)
Download file (22.22 kb)

Endnotes

[*1] "We’re a not-for-profit membership organization that exists to make scholarly communications better. We rally the community; tag and share metadata; run an open infrastructure; play with technology; and make tools and services—all to help put research in context." - Crossref (2021)

[*2] Country of specimen origin data can be established for 66% of specimens uploaded to GBIF by the 12 UK institutions considered, while continent of origin data can be established for 95% of all occurrence uploads to GBIF.

[*3] As of 2023-08-09 the funding call for "Host facilities as part of our heritage science infrastructure" had a total fund of £15,700,000 and a maximum award of £1,000,000. The fund scope was to "[...] enable you to purchase or build equipment and upgrade facilities that complement your existing research strengths, and ongoing funding to recruit and retain staff to enable access to your research facilities and collections " and to "[enable] access to heritage science facilities, collections and expertise for a wide range of users, to catalyse new collaborative research projects and amplify the impact of heritage science research "