Global Biodiversity Sub-Committee (GBSC)

Meeting papers

Creating Electronic Information Resources - Activities in UK Taxonomic Institutions

February 2007

For other documents from
Global Biodiversity Sub-Committee (GBSC)
Visit: http://www.jncc.gov.uk/page-4628

CREATING ELECTRONIC INFORMATION RESOURCES - ACTIVITIES IN UK TAXONOMIC INSTITUTIONS

1. Executive Summary

Through capturing data on species names and specimen-based records in digital form, institutions and individuals around the world are creating the basis of an electronic information resource for use globally.

Digitization activities focused on taxon name data and specimen data are widespread in the UK's major repositories for natural history specimens. Many of these institutions are also focusing effort on creating databases of biological data at the taxon level. All of these efforts are necessary to enable user-defined needs to be met both within the UK and elsewhere.

The UK is a world leader in the development and maintenance of authoritative lists of taxon names for living organisms, resources of central importance for a wide variety of endeavours in science, education, conservation and sustainable use.

The UK is second only to the US in the volume of digital records relating to biodiversity which it disseminates through the Global Biodiversity Information Facility (GBIF) reflecting the strong UK tradition in collection of biodiversity observation records.

UK diversity scientists and information specialists play a significant role in global and EU organisations focused on increasing access to biological data e.g. the Taxonomic Databases Working Group (TDWG), the Global Biodiversity Information Facility (GBIF) and network projects such as EDIT, SYNTHESYS and ENSCONET.

UK institutions hold a large volume of specimens documenting the biological diversity of many species-rich, developing countries in the tropics. These include a large proportion of the nomenclatural types relating to the world's diversity, essential points of reference in identifying and classifying biodiversity.

Each specimen (types and others) represents the occurrence of a particular species at a particular point in space and time, thus collectively these resources provide a unique record of changes in the distribution of organisms through space and time. Digitization of label data from these collections enables powerful analyses which can offer insights into current issues e.g. rates of biodiversity loss, potential impacts of climate change.

Creation and dissemination of high quality digital images of specimens of particular interest can facilitate use by a wide range of stakeholders, including those in countries from which the specimens originate. Such data-sharing is one of the most important contributions which the UK can make to assist species-rich countries to fulfill their obligations under the Convention on Biological Diversity.

However, fewer than one per cent of the UK's international natural history specimen records are widely available in digital form.

In terms of volume and proportion of important collections available in digital form, UK institutions have fallen behind other major institutions of comparable international standing e.g. Smithsonian Institution, National Herbarium of the Netherlands, New York Botanical Garden, Missouri Botanical Gardens.

UK efforts are also outstripped by those of major institutes in megadiverse countries including Australia, Brazil (especially Amazonia), China, Mexico and South Africa.

As major holders of biodiversity material originating from megadiverse countries (e.g. Kew holds 20,000 type specimens from South Africa alone), UK institutions must redouble their efforts to mobilise these data and make them available to countries of origin. Failure to repatriate the data relating to these collections in a timely fashion in response to growing expectations in countries of origin is likely to lead to increased requests for repatriation of the material itself and greater difficulties in scientific collaboration with these important partners.

Accelerated digital data capture efforts are desirable in order to realise the full potential of the existing investment in our natural history collections with the necessary urgency resulting from immediate threats to global biodiversity and world climate. Such effort has the added potential to benefit priority areas such as health and governmental obligations resulting from the CBD, GSPC, CITES and GBIF.

Digital collections management structures and information services tailored to user-needs are also essential.

The UK's natural history institutions lack the resources to tackle this Digitization challenge at a scale proportional to the magnitude of the task. Digitization is just one of a number of major new areas of activity that they have undertaken over the past two decades in response to government initiatives and commitments but without any corresponding increase in grant-in-aid in real terms (others include Access and Benefit sharing negotiations and greatly increased international capacity-building programmes).

Much of the resource currently devoted to natural history Digitization projects in the UK is from sources outside the UK, for example US-based Foundations. Thus the focus and rate of growth of the digital resources derived from UK natural history institutes are influenced in large part by needs and priorities determined outside the UK. In many cases these priorities are internationally agreed and closely aligned with UK commitments and interests. However, as the proportion of the total collections available in digital form increases, the priorities of different stakeholders are increasingly likely to diverge.

A UK policy framework, co-ordination and prioritisation mechanism is desirable in this area in order to maximise synergy, plan more effectively at a national level, and enable better integration.

A clear business model to support data capture and maintenance must be developed based on a more complete understanding of the needs and constraints of all stakeholders.

A UK fund to support Digitization of biodiversity data is also desirable in order to ensure that the growing digital collections respond to UK needs and priorities, both domestic and international.

Much of the above summary has been true in qualitative terms for a decade or more. However, global trends such as

(i) growing urgency to tackle biodiversity issues relating to global environmental change
(ii) the growth of the Internet;
(iii) increased ease and reduced cost of digital data capture and
(iv) increasing expectations of electronic access on the part of actual and potential users accentuate the need for a new UK-wide approach to the biodiversity Digitization challenge.

Beyond these general trends, a number of specific, time-bound opportunities are outlined in the final section of this review, to support our conclusion that a new approach in this area is not only necessary but urgently required.

2. Introduction

Recent years have seen rapid and significant developments in our ability to store information of different types digitally (including images) and in our ability to analyse, disseminate and use these data. The taxonomic institutions are heavily engaged and there are increasing calls by non-taxonomists, the CBD, as well as taxonomists themselves, for further initiatives within such bodies. UK taxonomic institutions hold many millions of biological specimens from around the world, as well as from the UK, making the country a major potential source of data. We also employ a still-significant number of taxonomists, capable of creating an electronic information resource of immense practical value. This report examines the needs for such information, current activities, and blocks and barriers to greater activity.

3. Scope of data collection for the information resource

Many different types of data are collected in taxonomic and related institutions, but this report will cover four types, with a particular emphasis on the second:

a) Taxon names – the scientific and informal names applied to species of animals, plants and microorganisms. These may be considered as two broad types: nomenclators, which focus on taxon names and associated data (e.g. the International Plant Names Index – IPNI; Index Fungorum; ZooBank), and taxonomic checklists and catalogues, which hold in addition a classification, synonyms and some aspects of taxonomic concepts (e.g. International Legume Database and Informatrion Service – ILDIS; the Tineid moth database, the Universal Chalcidoidea Database; Wtaxa (weevils); the aggregated Species 2000 system). All of the above examples are based in or led from the UK.

b) Specimen data – the data associated with biological specimens, including, for example, collector name, collection locality, date, altitude, associated species and morphological details lost on preservation. Geographical and habitat data can be of particular value in conservation studies. The level of detail is variable across specimens and is often particularly patchy for historical collections. Increasingly, high quality digital images of the specimens are created in the same workflow as the database record, both for archival purposes and for dissemination along with the specimen details. Such digitization also serves to provide a substitute for loans of physical specimens between institutions.

c) Observational data – of similar content to specimen data but lacking a voucher and typically collected with a different sampling bias. Observational data can be highly variable in quality. A considerable contribution to the data collection effort is provided by amateur volunteers.

d) Biological data – miscellaneous data associated with taxon names rather than specimens, for example taxon descriptions and distribution information and taxon-level conservation assessments.

Other databases are also being created, such as gene sequence data, use and management information, literature and other information associated with natural history collections, but these are beyond the scope of this document.

4. Why are the data collected?

Data may be collected for a wide variety of reasons, but it is important to distinguish between the possible uses of the data and the reasons why an individual or institute may spend time and resources creating and making data available. UK policy, as expressed through its statements on the GBIF Governing Board, is to ensure that data provision is targeted at user needs, particularly in line with CBD implementation. The UK has been very effective in carrying this point.

4.1. Uses of data.

4.1.1. Taxon names

The names of species are a basic tool for communication. Differences in taxonomic understanding between disciplines, and between different countries, and changes made in the course of taxonomic research, can limit this communication. An ideal response to this would be access to lists of names (checklists) which provide all synonyms and an authoritative statement as to which name should be used. In cases where there are competing taxonomic opinions, the facility to map these alternative opinions should be available. Ideally, both scientific names and vernacular names (where they exist) should be included. The CBD (e.g. COP IV/1.D; VI/8) has called for the digitization and increased availability of such data, as part of a global taxonomic information system Demand for taxon name data also comes from a diversity of user communities, for example:

- Conservation, to clarify the name to be applied to taxa under threat, and its hierarchical rank (species may have a different legal status to subspecies, varieties etc). Protected Area Management and Environmental Impact Assessment, to build up species lists for assessment and monitoring.
- Quarantine / pest interceptions, to ensure names on manifests and from identification sources can be meaningfully compared to black lists etc.
- Sustainable development requires reliable access to all information published on use and management of particular species.
- Academic research, for example into the nutritional status or potential medicinal value of natural products. Some aspects of research have potential commercial applications (e.g. pharmaceutical companies engaged in herbal medicine research).
- Formal and informal educational resources require scientific and/or vernacular names.

4.1.2. Specimen-level data.

Data are associated with all of the many millions of specimens held in UK collections. For animal, plant, fungal and many microorganism collections these data will usually include collector and date collected, locality, and, possibly, associated organisms or species (e.g. host of a parasite, plant being fed on by an insect herbivore), a description of the live specimen and

the habitat in which it was found. Specimens will also be associated with a name, often with information as to who applied the name, and some represent the nomenclatural type specimens of their species which are of special importance to taxonomists as the reference standards for taxon names. Representing a unique, verifiable, record of changes in distribution of taxa through space and time, such data offer opportunities for research which is highly pertinent to current biodiversity issues. There are an increasing array of tools used to model such data, for example environmental niche modelling, which can be used to predict species distributions. The uses of such specimen-level data are manifold[1], and include:

- Management of endangered species (e.g. use of herbarium specimen records to assess conservation status and target collection of priority species for seed banking)
- Management and prediction of the spread of invasive species
- National and regional planning studies
- Natural resource management
- Conservation planning, including biodiversity assessment
- Health and public safety, including mapping and prediction of disease vectors and spread of disease.
- Mining, where species indicating high mineral concentrations have been used.
- Biosafety (e.g. in Mexico such data are used regularly to study potential distributions and likelihood of genetic transfer and inform government decisions)
- Indicator information for 2010 target. The Sampled Red List Index seeks to record changes in conservation status of a representative sample of taxa in order to assess progress towards the 2010 target. For plants, Kew and collaborators are using specimen data to estimate population size and range changes over time. An initial digitization phase will provide the baseline data against which to compare records collected in the future.
- Studies of life histories, phenology, biogeography, species diversity and populations.
- Impact of climate change.
- Taxonomic research. In debates on taxonomy within the context of the CBD there have been calls by countries of origin for specimens to be returned to them from collections in the North. These calls have been partially satisfied by the provision of data from specimens, and in some cases images of the specimens.
- There have been calls through the CBD (CBD COP III/10; IV/1.D; VI/8) to digitize specimen data to facilitate access and use by countries of origin of the material. Such uses include knowledge of the specimens collected within their borders thereby supporting national taxonomic and related research (e.g. finding where specimens are

[1] See Chapman, 2005, Uses of primary species-occurrence data, http://www.gbif.org/prog/digit/data_quality/UsesPrimaryData.pdf

distributed throughout the world's museums and herbaria; finding where type specimens are held, and what species are based on types collected within the country), for basic analysis (e.g. compiling a taxon list for the country, based on what is known, identifying areas for protected status, identification of potential invasive species) or more advanced analysis (e.g. assessing global conservation status for species with ranges that extend across nations, identifying biodiversity hotspots).

- The concurrent creation and dissemination of high resolution digital images of selected specimens often broadens the value of digitization efforts to external audiences, including the countries of origin of the specimen material. Such images have been called for by many participants in meetings on the GTI around the world.

4.1.3. Observational data

Observational data usually comprise sighting records for a given taxon, and are likely to include observer locality, date, observer name and perhaps other details such as behaviour. Observations may equally be based on recognition of animal sounds, or secondary evidence such as nest sites or spoor. In some instances observations may be backed up by supplementary evidence such as recordings or photographs. Such data may be used in the same ways as specimen data. However, whereas specimen-level data tend to be biased towards rare species, be good for numerous point sources but weaker on quantifiable observations, observational data, in contrast, are weak for rare species, may be good for numerous point sources, and good for non-rare species. Together, observational and specimen data combined may provide a good picture of abundance and distribution across a group (e.g. birds) (Guralnick & Van Cleve, 2005[2]). The UK has made extensive use of observational data in particular, for example in the production of the BSBI New Atlas of the British and Irish Flora and the annual Breeding Bird Survey, led by the BTO, RSPB and JNCC. In both these examples, as for many observation records, substantial input is provided at the data collection phase by amateur volunteers.

4.1.4. Biological data

This is a much wider and more diffuse concept, and relates to detail that does not get fully associated with specimen label data or, if it does, includes associations at specimen-taxon or taxon-taxon level. There are issues of species concept to be taken into account and effective production and interpretation of such datasets is dependent on authoritative lists of accepted taxon names and synonyms. There is a necessity to clarify the sources of the information; often bibliographic resources are simply re-statements of earlier published statements. That said, such sources often contain valuable information, which is important to capture digitally. For example, a Pollinator database currently being planned by FAO under the International Pollinators Initiative, and with the involvement of GBIF, includes a component where information on pollination will be included. The data will be drawn from specimens, observations and interpretations, and will include information about the confidence with which

[2] Guralnick, R. & Van Cleve, J., 2005, Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. *Diversity & Distributions*, **11** (4), 349-359.

a pollination association can be asserted, the context in which such pollination takes place, and the relative success of pollination by different pollinators with single plant species. These data will assist in managing agro-ecosystems for pollinator maintenance, and also have the potential to influence conservation management plans for protected areas and species.

4.2. Individual and institutional reasons for data collection

The potential benefits to be derived from the use of digitized data often differ from the immediate drivers for data capture. Common drivers include:

- Internal institutional policy and priorities including statutory obligations. These may include, for example, databasing at specimen or lot level at object entry; of type specimens or other historically-important collections. Databasing may be driven by audit and transaction management needs rather than scientific requirements. While many institutions seek to maximize the benefits of all such data capture exercises, the data collected for such purposes may not be prioritised to meet other user needs, either in terms of data elements, taxa covered, or geographic areas covered. Where digital images are captured alongside data, this provides the opportunity to capture additional data at a later date in an efficient manner.
- Availability of grants. This is more likely to lead to accessible data that match user-defined needs. Examples of this type of work include Sri Lankan snails (co-funded by the Darwin Initiative and NHM), African and Latin American plant type specimens (funded by the Andrew W. Mellon Foundation), fungi data (co-funded by GBIF) and Moss types (co-funded by GBIF and NHM).
- Research needs of individual researchers. In the past, the scope of such databasing efforts will often have been limited, and not designed to match needs of users other than the individual concerned. In such cases there is a risk that data may be retained only for the length of a project, may be in non-standardised formats, and may not be made accessible outside the institution where they are captured. However, institutes are increasingly working to overcome these limitations, by working for example, to migrate ‘private’ legacy datasets into their institutional databases and to ensure that future research-driven databasing efforts by individuals are compliant with data standards so that they can more easily be migrated and made more generally accessible.
- Intent to meet a user-defined need, as indicated above. This is most likely to apply to bodies whose primary mission is aligned with one of those needs, such as JNCC, but also applies in cases where institutions have commitments to meet strategic needs identified by a broader community e.g. many Kew digitization projects explicitly address targets identified within the CBD’s Global Strategy for Plant Conservation
- Educational objectives including production of images and data-sets for inclusion in educational resources.

The interaction of these drivers combined with the lack of an overall strategic direction has resulted in the following outcomes:

- Data captured as a result of personal or institutional internal priorities may not be maintained in standard formats and therefore not amenable to analysis or use by other users.
- The match between data coverage at taxon level and priorities of other data users may be poor.
- The match between the data elements captured for each specimen / taxon and those required by other data users may be poor.
- A short-term approach to data capture efforts reduces the prospects for long-term sustainability of datasets.
- Data captured may not be made available to users outside the data-generating institution.
- Where efforts are made to deliver external access, diversity in data standardization and in data elements captured create obstacles to data-sharing across institutions.

4.3. User demand and current use

There has been a rapid growth of data availability via the Web in the past 5 years. This growth has been coupled to aggregations of data from different sources, so that single portals to data from different sites have been created (e.g. Species 2000, GBIF). However, many of the users of data have built their operations either on data they collect themselves, or on pre-existing data suppliers. Even now, the data they might use are either not currently available, not sufficiently comprehensive or of insufficient quality to make it worthwhile to change their current data-gathering or analytical methods. With proliferation of datasets it becomes increasingly difficult for potential users to identify authoritative sources. Consequently, while potential user demand for data is high, actual levels of use of existing resources have not yet realised this potential because the work patterns to make use of data aggregated and available on the web have not been developed and implemented.

Another issue is that users may employ the information without detailed cognizance of its origin. For example, considering the development of international policies to manage spread of invasive species, many of the implementation issues require the provision of taxonomic names, and the sharing of these names among signatories to the policies. The information requires involvement of taxonomists and, ideally, aggregating initiatives such as GBIF. Responses have included ITIS (Integrated Taxonomic Information System) and Fauna Europaea. Both are affiliated to both Species 2000 and GBIF, and both are dependent on taxonomists in the UK and elsewhere.

While the UK is making an important contribution to global information services, it must be recalled that these services may be of importance not only to other countries but also to the UK itself. In an age where organisms are regularly found far outside their country of origin, and we may be seeing a response to climate change in distributions, UK scientists, quarantine

officers, environmental managers and change modelers, to name but a few, need the information provided by global information services, just as those services need the data that the UK can provide. Species 2000 and GBIF are not simply initiatives that UK science contributes to, they are resources that UK science needs (and uses) today.

5. Current digitization activities in the UK

5.1. Taxon names

The UK is a world leader in producing and making available authoritative lists of taxon names, almost all focused on taxa rather than a user-defined or environmentally-defined system. Many of these are multi-institutional and multinational efforts, and identifying the precise contribution of the UK is not simple. Of the 38 datasets currently accessible through the Species 2000 annual checklist nine originate in UK institutions or initiatives based in the UK (NHM, Kew, Reading University, CABI). In addition there are a number of large databases accessible directly from institutional sites, such as the International Plant Name Index based at Kew (and Harvard), the World Checklist of Monocotyledons (Kew), the Flora Europaea database (Edinburgh), the Dipterocarpaceae database (Edinburgh), Butterflies and Moths of the World (NHM), Bumblebees of the World (NHM), Index Fungorum (CABI, CBS, LCR-NZ) and numerous others.

The drivers behind the creation of these authoritative name lists vary and include personal interests of their creators. However, there is a degree of response to external needs, such as the checklists of selected plant families, a response to the Global Strategy for Plant Conservation (Target 1) and checklists of orchids, aloes etc. produced to support implementation of CITES. Funding for list development has come from numerous sources, although primarily the institutions at which the lists are compiled and support from other institutions in global consortia. However, funds have been secured from GBIF, the Darwin Initiative, the Leverhulme Trust, the EU, the Headley Trust, the Reuters Foundation, the USNSF, the USGS, CSIC (Spain) and others.

One source of names and other data on the web is the ION database of Thomson Zoological (publishers of *Zoological Record*). Although freely available, it is sourced from a commercial company and indexes back to their more extensive records (which can be accessed through payment). This database is being made available to GBIF and, perhaps more significantly, Thomson is working with the International Commission on Zoological Nomenclature to produce an authoritative database 'ZooBank' in a public-private partnership. Index Fungorum – Index of Fungi has a similar relationship between the Index Fungorum partnership and CABI.

5.2. Specimens

The UK holds a very large number of specimens, not only from the UK but also from other countries, particularly from species-rich, developing countries in the tropics. These specimens are of importance not only because they provide a huge quantity of point-source data that can

be used to investigate distributions, occurrences etc., but also because they include a very high proportion of the nomenclatural type specimens which form the basis for identifications around the world. The actual number of specimen records in the UK is difficult to assess, but as a rough guide the NHM holds some 71 million specimens in total[3], the Kew Herbarium some 7 million plant specimens and 800,000 mycological specimens, University of Oxford some 3.25 million zoological and entomological specimens, RBG Edinburgh some 2.5 million herbarium specimens, National Museums of Wales ca. 3 million biological specimens (of which some 600,000 are digitized in one format or another), Liverpool Museum ca. 1.6 million biological specimens and CABI ca 400,000 fungal specimens. In only one case, CABI, is more than a small fraction of these databased. Of these UK specimen records, 1,330,199 are currently served through GBIF (primarily from NHM, but also RBG Kew, British Antarctic Survey, University of Reading, Oxford University, CABI). In terms of number and proportion of total records available in digital form, the main natural history institutions of the UK have fallen behind globally important collections elsewhere both in developed countries and in certain southern, megadiverse countries.

Examples of databasing projects on these international collections include:

- Stag beetles (NHM, following internal curatorial priority) (35,649 records)
- Butterflies (NHM, following internal curatorial policy) (69,703 records)
- Linnaean butterfly specimens (NHM, with financial support from Leverhulme Trust, created for taxonomist use) (305 taxa)
- Orchids (NHM) (39,000 records)
- Paraguayan plant specimens (NHM, linked to Missouri Botanical Gardens, arising from Darwin Initiative project)
- Fish specimens (NHM) (128,000 records)
- Mammal type specimens (NHM)
- Microbiology slide collection (NHM) (28,000 slides)
- Kew dried herbarium specimens, incl. Monocot Types, African types etc. (136,000 – 1.9%)
- Kew herbarium spirit collections (70,000 – 99%)
- Kew mycological collections (115,000 – 14.4%)
- RBG Edinburgh herbarium specimens (231,000 records)
- Palearctic Diptera (Oxford University) (5,000 species)
- World birds (Oxford University) (17,000 records)
- CABI Biosciences Fungus Collection (358,349 records)
- Mollusca (National Museums of Wales, Cardiff) (130,000 records)

While this is by no means an exhaustive list, examination of the coverage, and of the records themselves, shows that:

(a) The proportion of collection contents digitized is only a small fraction of the total holdings, amounting to less than 1% of the national collections.

[3] Including minerals and fossils

(b) Principal users of the digitized data are expected to be taxonomists, either in the UK or overseas.

(c) Many of the data that have been digitized are accessible only through the institutional web interface, and not through a single multi-institutional portal. GBIF serves approximately 1.3 million specimen records from UK institutions (excluding NBN); the NHM, for instance, holds at least half a million digitized records that are not yet available through GBIF. A positive note is that in the very near future the "Fungal Records Database of Britain and Ireland" (FRDBI) will be serving some 1.2 million records of British fungi, which will be made available to NBN and consequently to GBIF[4].

(d) Each institution has databased specimen-level records that are not yet accessible to external users.

The last two points require some additional comment. Data may not be made available outside an institution for several reasons. There are potential problems with data quality[5] (e.g. the accuracy of georeference points, or names of associated taxa), and with data cleanliness[6] (e.g. with data entered incorrectly or in the wrong fields, or data which are effectively uninterpretable). Some records held by institutions may refer to multiple specimens from the same locality rather than single specimens, and some specimens (e.g. subdivided plant specimens) may have been databased several times. Data collected for audit purposes may not be of sufficient breadth to be of value outside the institution; data may be stored in databases that are not suitable for direct linkage to the web. All of these reasons and others may lead institutions not to release access to data; all can be resolved, but in each case resources – sometimes significant resources, may be needed.

The amount of external funding that has supported these databasing efforts is difficult to assess, but funders outside the data-holding institutions themselves include GBIF, the Darwin Initiative, the Leverhulme Trust, the Andrew W. Mellon Foundation, and the Gordon and Betty Moore Foundation. Notably many of these funding sources are based outside the UK, and in consequence the policies which influence the focus of projects they fund may not align entirely with UK policies.

5.3. Observations

The UK is rich in digitized observational data covering its own biota. An example is the provision through the NBN of over 15 million UK records[7] to GBIF – the second-largest individual contribution of the 191 GBIF data providers of these data classes. All or almost all of these NBN records are observation records. These records come from datasets managed by

[4] See http://www.fieldmycology.net The database is managed by the British Mycological Society, hosted by CABI, and includes data from Kew, CABI and many individuals.
[5] See Chapman, 2005, Principles of data quality, http://www.gbif.org/prog/digit/data_quality/DataQuality.pdf
[6] See Chapman, 2005, Principles and methods of data cleaning, http://www.gbif.org/prog/digit/data_quality/DataCleaning.pdf
[7] There are currently **20,823,239** species records available on the NBN Gateway from **197** different datasets.

some 30 different bodies[8]. Whilst UK institutions do hold observational data from other countries, these are far fewer in number than UK records.

5.4. Biological Records

Such records are accessible on the websites of many individual organizations, including all of the major taxonomic institutes. They are, however, not accessible through a single portal. Generally their focus reflects the interests of their creators, although they often have a wider value to users.

6. Access to data.

Of data which have been collected, accessibility may be of the following types:

- Through personnel at the data-holding institute only (i.e. it is either not digitized or, if digitized, has not been made available on the web).
- Through the web site of the data-holding institute.
- Through a third-party portal (e.g. BioCase, Species 2000, GBIF, NBN)

Of these alternatives, the final one provides the greatest degree of access to data, since such aggregators act as a 'one-stop-shop' for data users, and may support tools or standard export protocols to allow users to access data in a manner best suited for their purposes. These mechanisms offer the most effective way of making data available. Moreover, the initiatives in the third category are all exploring or have implemented interoperability and to a greater or lesser extent provide access to the same data, according to their remits, although there is a cross-dependence involved, so that GBIF depends for its authoritative name content on the data delivered through Species 2000 from its own data providers.

There is UK involvement in each of these, NBN being, of course, a fully UK operation. BioCase is one of a series of EU-funded projects designed to increase data availability. Currently SYNTHESYS and EDIT, among others, are taking this process forward in different ways. UK scientists have been very active in these projects, as they have in other initiatives to make data available, such as the Taxonomic Databases Working Group (TDWG), GBIF and ENSCONET (European Seed Conservation Network).

The funding for EU projects has been a series of separate grants. Sustainability is up to the partner organizations, and dependent on international uptake of the data standards developed. Species 2000 is a global partnership with strong leadership from within the UK, but has no

[8] Aquatic Heteroptera Recording Scheme, BATS & the Millennium Link, Countryside Council for Wales, Dorset Environmental Records Centre, Glasgow Museums BRC, Herts Bird Club, Lothian Wildlife Information Centre, National Biodiversity Network Trust, Natural England, Northumberland Wildlife Trust, Rossendale Ornithologists' Club, RSPB, Scottish Borders Biological Records Centre, Suffolk Biological Records Centre, Take a Pride in Fife Environmental Information Centre, the Balfour-Browne Club, The Bees, Wasp and Ants Recording Society, The Botanical Society of the British Isles, the BRC and various UK recording schemes managed by the BRC and partners, the Bristol Regional Environmental Records Centre, the British Bryological Society, the Conchological Society of Great Britain and Ireland, the Environment and Heritage Service, The Herpetological Conservation Trust, the Highland Biological Recording Group, the JNCC, the Marine Biological Association [MarLIN (Marine Life Information Network)], the Marine Conservation Society, the Scottish Environment Protection Agency, Wiltshire and Swindon Biological Records Centre.

sustainable funding and is dependant on securing funds from governments and others. GBIF is an IGO, and dependant on funding from participant countries (their contribution being according to an agreed scale linked to GDP). Data provision is not strongly tied to this funding, but is provided as in-kind support from UK systematists and information specialists based in natural history institutions. Funding for the UK's contribution to GBIF is uncertain from year to year, comprising voluntary contributions from stakeholder institutions.

A key point is that there is not clear sustainability for data provision, maintenance or access, and no generally accepted financial model that has been applied to these components.

7. Measures of success and performance indicators

As noted, data collected and made accessible have been gathered for a variety of reasons. As such, the measures of success that are applied to the databases are often internal to institutions and may not reflect outside use. Thus for the databasing of collection contents undertaken within the Natural History Museum performance indicators, in so far as they have been finalized, concern numbers of records created, amount of the collection databased, ease of access to specimen data, efficiency in creating loan records (which use specimen databases) etc. They do not involve accuracy of names, application of georeferences, or use of the data outside the Museum. For databases made available through the NHM web site, the Museum records numbers of hits at database level, but not of individual records, and not broken down into user types. Kew measures use of its web resources as one of its key performance indicators and tracks trends in usage of strategically important databases. GBIF is able to return records of numbers of hits to data providers, but there are no standard agreed formats.

Given the variety of reasons for creating the databases, and the gaps between these drivers and the uses to which the data may be put, there are no clear and agreed criteria for success. The lack of these, and of overarching priorities, hinders development of any coherent policies across institutions. One notable exception to this is the Global Strategy for Plant Conservation, which has led to institutional partnerships in data capture and availability with clear targets. Another is the International Pollinators Initiative where, in response to external drivers, a full list of bee names is being developed by a number of individuals and institutions in collaboration.

8. Blocks and barriers

A number of hindrances to data collection and accessibility have been mentioned in section 4 above. It is clear that there are uses for data outside the data-holding institutions, and that these uses include activities such as those in support of CBD implementation, understanding and planning for the impact of climate change, and management of invasive species and pests. These uses are consistent with UK domestic policies, and with the needs of many other countries expressed through the CBD. What, then, is preventing full use of data held within the UK, including that not yet collected from sources in collections?

Data collection, data curation and sustainable management, making data accessible outside the data-owning institute, and building and maintaining access to multiple data sets simultaneously

all require financial support. Within the largest collection-holding institutes in the UK, overall staffing levels present a major issue, and prioritization of staff activities is governed by internal policies which include maintenance of the collections and generation of external funds. These internal policies often reflect legal responsibilities. Digitizing all specimens could take up 100% of staff time for several decades, so has to be managed against other priorities and needs. However, even to deliver information and data in response to identified external needs is beyond the resources of UK natural history institutions. The UK, through its support of the CBD, has agreed and helped craft COP Decisions that call for or require natural history institutions to take action in support of Access and Benefit-sharing issues, institutional and individual capacity building, data and information provision and other partnership programmes. Some funding has been provided through the Darwin Initiative, but the ability of UK institutions to respond is compromised by the lack of commensurate growth in core funds.

In brief, the blocks and barriers are:

(a) Understanding of potential uses of data not yet complete among stakeholders.

(b) Lack of communication to potential users of value/relevance of reliable data.

(c) Tools to analyse data not yet integrated into decision-making process of many potential users.

(d) Lack of communication of priorities in data collection and standard content between users, providers and funding bodies.

(e) Policy framework for data availability and collection not in place in UK.

(f) Concerns about providing access to data that can be considered sensitive, either for conservation reasons (e.g. providing information on localities of species under threat from poaching or over-collection), or academic reasons such as pre-empting unpublished research.

(g) Data-holding institutions following internal policies rather than external needs.

(h) Insufficient staffing available to digitize data.

(i) Insufficient funds within institutions to employ staff with prime responsibility of digitizing data.

(j) Unavailability of grant funding to enter data into extant database structures[9]

(k) Lack of sustainable support for initiatives providing access to distributed data through data portals.

(l) Lack of financial model to support data curation and sustainable management.

These problems are to a great extent interdependent. For example, a greater understanding of user needs, information availability and the options for data analysis might be fostered by

[9] E.g. BBSRC-supported work under the heading "Methods to create biological databases from uncomputerised or unstructured data" includes on their web-site the statement "Research proposed under this heading must have strong links to the biological data needed by the community. The feasibility of some of these techniques may require the building of exemplar databases, which should, however be of wider benefit to the community. Creation of new databases by entering data into existing database structures was not supported."

policy, lead to improved prioritization of data capture and precision in its delivery, and better feedback from data users to data providers. In turn this might lead to a better understanding of the economic and other benefits of constructing the electronic information resource, and by matching use to benefits, help the development of a business model.

9. Requirements for change

For the current situation to change, the following needs should be addressed:

(a) Development of a business model to underpin data collection, delivery and maintenance.

(b) Development of a mechanism by UK funding bodies to support data capture, delivery and on-going maintenance.

(c) Policy framework addressing priorities for future data collection efforts, accessibility and maintenance by UK institutions and other relevant bodies to be developed, aimed at maximizing synergy, more effective planning at a national level, and fostering integration between sectors. The policies should cover both UK and non-UK taxa and specimens.

(d) Work between data-developing institutions and their sponsoring bodies to develop appropriate performance indicators.

(e) Development of a strategy for on-going engagement between data holders, managers and the users of data.

(f) Greater emphasis on engaging users in understanding the digital resources and prioritising delivery services required.

10. Opportunities

Much of what has been outlined in this document has been true in qualitative terms for a decade or more. However, the need for co-ordinated action on biodiversity digitization in the UK should be high on the UK policy agenda in light of global trends including:

(i) growing urgency to tackle biodiversity issues relating to global environmental change
(ii) the growth of the Internet;
(iii) increased ease and reduced cost of digital data capture and
(iv) increasing expectations of electronic access on the part of actual and potential users

Beyond these general trends, certain specific, time-bound opportunities suggest that actions taken in this area in the coming two to three years will be critical to broader, medium- and long term outcomes of strategic significance to the UK. Such opportunities include:

Incorporating global diversity data (especially for species of montane, coastal, island and arid ecosystems) into ongoing UK work to model the impacts of climate change;

Delivering a checklist of known plant species by 2010, meeting Target 1 of the Global Strategy for Plant Conservation and thereby strengthening the credibility of the GSPC as a whole and consolidating the UK's position as a world leader in this field;

Capitalising on the increased interest in authoritative biodiversity datasets on the part of government, European and international agencies and the commercial sector in order to develop a business model for long-term sustainability in this area.

Securing ongoing funding streams for biodiversity data digitization in the UK from existing international funders by facilitating matched funding arrangements from a UK digitization fund.

Christopher H. C. Lyal (NHM)

Mary Gibby (RBGE)

Eimear Nic Lughadha and Anna Saltmarsh (RBG Kew)

With thanks to Charlotte Couch (RBG Kew), Frank Bisby (Species 2000), Paul Kirk (CABI), Mike Wilson (NMW)

Appendix 1: List of Acronyms

Acronym	Full name
BBSRC	Biotechnology and Biosciences Research Council
BRS	Biological Records Centre
BSBI	Botanical Society of the British Isles
BTO	British Trust for Ornithology
CABI	CAB International (formerly the Commonwealth Agricultural Bureau)
CBD	Convention on Biological Diversity
CBS	Centraalbureau voor Schimmelcultures
CITES	Convention on International Trade in Endangered Species of Fauna and Flora
CoP	Conference of the Parties (to CITES or CBD)
CSIC	Consejo Superior de Investigaciones Científicas
EDIT	European Distributed Institute of Taxonomy
ENSCONET	European Seed Conservation Network
EU	European Union
FAO	Food and Agriculture Organisation
FRDBI	Fungal Records Database of Britain and Ireland
GBIF	Global Biodiversity Information Facility
GSPC	Global Strategy for Plant Conservation
GTI	Global Taxonomic Initiative
IGO	Inter-Governmental Organization
ILDIS	International Legume Database and Information Service
ION	Index to Organism Names
IPNI	International Plant Names Index
ITIS	Integrated Taxonomic Information System
JNCC	Joint Nature Conservation Committee
K	Herbarium, Royal Botanic Gardens, Kew
LCR-NZ	Landcare Research New Zealand
MarLIN	Marine Life Information Network

NBN	National Biodiversity Network
NHM	Natural History Museum
NMW	National Museum of Wales
RBGE	Royal Botanic Garden, Edinburgh
RBG Kew	Royal Botanic Gardens, Kew
RSPB	Royal Society for the Protection of Birds
SYNTHESYS	Synthesis of Systematic Resources
TDWG	Taxonomic Database Working Group
USGS	United States Geological Survey
USNSF	United States National Science Foundation
Wtaxa	Electronic Catalogue of Weevil names (Curculionoidea)