Resource Identification for a Biological Collection Information Service in Europe

Botanic Garden and Botanical Museum Berlin-Dahlem
Resource Identification

for a

Biological Collection Information Service in Europe (BioCISE)

Results of the Concerted Action “BioCISE Resource Identification”, funded by the European Commission, DG XII, within the EU Fourth Framework's Biotechnology Programme, August 1, 1997 to December 31, 1999

Edited by Walter G. Berendsohn

Botanic Garden and Botanical Museum Berlin-Dahlem
Department of Biodiversity Informatics
The Authors:
Andrea Hahn, Anton Güntsch, Birgit Felinks, Linda Olsvig-Whittaker, Louis Rechaussat, Pedro Fernandes, Pier-Luigi Nimis, Richard White, Walter Berendsohn, and Wouter Los were members of the Concerted Action or belonged to the BioCISE secretariat (see p. iv).
David Lazarus, Institut fuer Paläontologie, Museum für Naturkunde, Zentralinstitut der Humbolt-Universität zu Berlin, Invalidenstraße 43, D-10115 Berlin, Germany
Mark J. Costello and Chris Emblow, Ecological Consultancy Services Ltd (EcoServe), 17 Rathfarnham Road, Terenure, Dublin 6W, Ireland
Jürgen Koenemann and Christoph Thomas, humanIT Human Information Technologies GmbH, Rathausallee 10, D-53757 Sankt Augustin, Germany
Neil Thomson, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom

© 2000 Botanic Garden and Botanical Museum Berlin-Dahlem
Individual chapters of this publication may be reproduced in any form without prior permission, provided that the source is acknowledged by citation of the Author(s), Editor, and Publisher.

Printed in Germany.

ISBN 3-921800-44-7

Published by the Botanic Garden and Botanical Museum Berlin-Dahlem,
Freie Universität Berlin
Königin-Luise-Straße 6-8, 14191 Berlin, Germany
www.bgbm.fu-berlin.de

# Contents

I. The BioCISE project  
   Project methods and summary of results  1  
II. Computerizing and networking biological collection data  2  
   Capture of biological collection data  5  
   Promises of networking  8  
   Problems arising from combining information from different sources  9  
III. The complexity of collection information  13  
IV. Users and uses of biological collections  19  
   Major categories of users  21  
   Access to biological collection information  25  
   Desired output  28  
   User interface  31  
V. On the value of palaeontological collections  33  
   Major uses of palaeontological collections  34  
VI. CABRI: access and quality control for biotechnological collection resources  37  
VII. Information Resources - The BioCISE Survey  39  
   Scope and execution  39  
   Results  42  
   Conclusions  48  
VIII. Networking collections: BioCISE national workshops  49  
   The BioCISE project in Portugal  49  
   Networking biological databases: the Israeli experience  50  
   Italy: the struggle between Titian and the beetle  51  
IX. Prototyping unit-level access  52  
X. The BioCISE collection catalogue  53  
   Prototype networking in the BioCISE collection catalogue  55  
XI. Concepts for a European Portal to Biological Collections  59  
   Achieving breadth and scalability: Meta-information to the rescue!  60  
   Issues for informatics research  63  
   Basic organization of the Service  66  
   Summary  70  
Acknowledgements  71  
References  72
Members of the Concerted Action

Anastasios Anagnostopoulos, The Goulandris Natural History Museum, 14562 Athens, Greece
Prof. Dr. Walter Berendsohn (Project co-ordinator), Freie Universität Berlin, ZE Botanischer Garten und Botanisches Museum Berlin-Dahlem, 14191 Berlin, Germany
Pedro Fernandes, Instituto Gulbenkian de Ciencia, 2781 Oeiras Codex, Portugal
Gregor Hagedorn, Biologische Bundesanstalt für Land- und Forstwirtschaft, Institut für Mikrobiologie, 14195 Berlin, Germany,
Dr. Jasmin Jakupovic, (formerly at) Technische Universität Berlin, Germany
Prof. Dr. Jacques Lebbe †, Université Paris VI, Laboratoire Organisation & Evolution des Systèmes, 75252 Paris Cedex 5, France
Dr. Wouter Los, Institute for Systematics and Population Biology / Zoological Museum Amsterdam, Amsterdam, The Netherlands
Prof. Dr. Jos van der Maesen, Agricultural University Wageningen, Dept. of Plant Taxonomy, 6700 ED Wageningen, The Netherlands
Prof. Dr. Pier Luigi Nimis, University of Trieste, Department of Biology, 34127 Trieste, Italy
Dr. Richard J. Pankhurst, Royal Botanic Garden Edinburgh, Taxonomic Computing, Edinburgh EH3 5LR, United Kingdom
Dr. Louis Réchaussat, INSERM, 75654 Paris Cedex 13, France
Dr. Jarmo Saarikko, Finnish Forest Research Institute, 00170 Helsinki, Finland
Dr. Karsten Siems, Analyticon AG, Research Department, 13355 Berlin, Germany
Prof. Dr. Benito Valdés, University of Sevilla, Depto. de Biologia Vegetal y Ecologia, 41012 Sevilla, Spain
Besher Wattar, (formerly at) Novo Nordisk A/S, Scientific Computing, 2880 Bagsvaerd, Denmark
Dr. Richard White, University of Southampton, Biodiversity & Ecology Research Division, Southampton SO16 7PX, United Kingdom
Dr. Linda Olsvig-Whittaker, Nature Reserves Authority, Science and Management Division, Jerusalem 94467, Israel, linda@bgumail.bgu.ac.il

Project Secretariat

I. The BioCISE project

Walter G. Berendsohn

Objective

The near-exponential accumulation of information in biological collections and databases, development of international interests in biological questions, and the expansion of computer literacy among scientific professionals in recent years have combined to generate a demand for easy and rapid computer-based access to biological information. BioCISE and preceding projects and initiatives brought together researchers from different disciplines and professional backgrounds who share a common vision: An electronic access system facilitating queries across the hundreds of millions of specimens and monitoring or mapping records held by institutions, projects and individual researchers in the EU and partner countries.

This publication reports on the concluded concerted action project, which set out to identify and analyse biological collection information and its environs with the aim to prepare a sound collaborative technical and structural base for a Biological Collection Information Service in Europe and a strategy for its implementation.

Justification

Tasks like the acquisition, cultivation, preservation, and storage of objects in biological collections are an integral part of biological research in many sub-disciplines.

**Biological Collections:**

The term “Biological collection” is here understood to include the following main categories:

- Living collections (e.g. botanical and zoological gardens, microbial strain collections)
- Natural history collections (mainly in museums and universities)
- Data collections used in faunistic and floristic mapping projects and surveys.

Biological collections include microbial and tissue culture collections, plant genetic resources, natural history museums, botanical and zoological gardens, natural substance...
collections, as well as observation data (surveys, mapping projects) and multimedia data such as animal sounds and pictures of organisms. In most fields, Europe houses the most extensive living and natural history collections worldwide. Taken together, this represents an immense knowledge base on global biodiversity. Field and research notes and specimen labels contain valuable and detailed data; and the object itself can be a physical resource for research and industry. In contrast to mere observations, the object also presents a falsifiable source of information, i.e. it can be re-observed to verify a scientific hypothesis based on it.

Currently, this knowledge base is largely under-utilised, because its highly distributed, heterogeneous, and complex scientific nature obstructs efficient retrieval. Efforts to network the resources exist, but these are restricted to the local or national level or to often narrowly defined biological sub-disciplines. Transfer of technology and co-ordination on a European level is lacking.

Project methods and summary of results

A previous EU project (CDEFD, see Berendsohn 1997a) has produced a detailed information model of biological collection information. The CDEFD model was further extended and finally published under BioCISE (Berendsohn et al. 1999a).

The model depicts in detail the information structure of biological collection units, ranging from natural history specimens to strains in culture collections to occurrence records of organisms in biological surveys. On the one hand, it documents the complexity of collection information (see Chapter III). On the other hand, the model
The project indicates that biological collection data covering a wide range of traditionally separated biological sub disciplines could and should be united under a common access system, because their information structures are so similar. The approach of the BioCISE system was based on this conclusion.

In a large-scale survey of collections in the countries covered by the EU's 5th Framework Programme, questionnaires were sent by mail to 2550 laboratories (institutes and other collection holders). In addition, 413 professional societies throughout Europe were asked to support the survey. 483 laboratories (or 19 % of those contacted) replied to the survey. Of the respondents, 292 (61%) do maintain one or more biological collection databases. The total number of collection units (including survey records) catalogued in these 448 databases exceeds 42 million. For further details on the survey's result see Chapter VII.

The BioCISE World Wide Web Collection Catalogue provides access to the survey's results; for 60 % of the laboratories, which responded to the survey it is the first representation of their collections on the World Wide Web. In addition to the detailed survey results of respondents, all corroborated institutional addresses are accessible. The

---

**European collection holders – a fragmented community**

Fragmentation occurs mainly along five lines:

(i) Taxonomic boundaries (botanical gardens, zoological gardens, microbial culture collections, floristic mapping projects, faunistic monitoring, bird sightings)

(ii) Species survey vs. specimen collection

(iii) Collections as archives or as a means to an end. There are numerous surveys and collections effected e.g. in environmental impact analysis projects. The objects and the data generated by such projects may be highly valuable, but they are usually not accessible using the normal pathways of information retrieval, because either their peers do not recognize their value for the community at large, or they regard them as confidential research information

(iv) Along national lines, mainly due to differences in the organisation of research collections. e.g. largely centralised national natural history collections vs. a number of distributed smaller institutions

(v) According to size; large and small institutions may have different interests.
Collection Catalogue also provides access to a set of linked networks to demonstrate the idea of interoperability. See Chapter X for details. The WWW site and the database will be maintained by the BGBM at least until December 2000.

The collection meta-information system

The simplest useful definition of metadata is “structured data about data.” This very general definition includes an almost limitless spectrum of possibilities ranging from human-generated textual description of a resource to machine-generated data that may be useful only to software applications (Anon. 2000).

A collection meta-information system must be capable to process information on all levels, from entire collections to single units, from very detailed data to the most general, from simple text descriptions to highly atomised data structures.

Several subgroups within the project investigated user needs, i.e. the question if and how collection information is or may be put to use (see Chapters IV and V). In this context, several meetings were organized to incorporate potential users in the design process of a Biological Collection Information Service. The meetings also served to publicise the Survey, and to foment collaboration between collections on a national level (see Chapter VIII). Participation of BioCISE members in several international congresses and scientific meetings served the same purpose.

The results were used in several funding applications, where new partners contributed to the discussion. As a synthesis of the resource identification process, a two-tiered strategy for the creation of a European Collection Information Service emerged. The strongly varying level of technical possibilities of collection holders, the relatively low availability of computerized records (as a percentage of the entire holdings), as well as the differing willingness to share information over computerized network makes it necessary to design a central system to provide collection-level data and thus the overall cover of collections. Users can use this system to gain access to the information (most of them) really demand: unit-level data. Initiatives to provide common access to unit-level information must be fomented (see Chapter IX), and be it restricted to specific thematic areas or to the national level. However, participation in ongoing international standardisation efforts should be secured. The system should eventually become a “portal” to collection information (see Chapter XI), by linking unit-level information, where available, to well-structured descriptions of entire collections.
II. Computerizing and networking biological collection data

Linda Olsvig-Whittaker and Walter G. Berendsohn

Capture of biological collection data

Over the last two decades, curators in Natural History Museums as well as those responsible for ecological mapping or monitoring projects started databasing their collection and observation information. While databases were first used as a tool for collection management, over the last years awareness grew that this digitised information can be regarded as a most valuable source for biodiversity research. From the data provider’s point of view a tendency “from IT for audit and accountability to IT for access” (MDA 1997) became obvious. In view of accelerating environmental changes, loss of biodiversity, and a pressing need for fast decisions in environmental politics, this information base must be utilized to the greatest possible extent. Collection information is of significant importance not only for the scientific community, but also for the whole of society (see Lane 1998). Databasing and facilitating access to collection and observation data has been recognized as one of the priorities of the OECD Megascience Forum GBIF initiative (Edwards 1999).

However, most biological and ecological information is not yet accessible on-line and therefore not fully useful for the community (OECD 1999), and in many cases not even available in electronic form at all. Right now, a variety of programs and initiatives form distributed networks of biological information, and show the degree to which information technology can serve as a tool for mastering the information mountain (Anderson 1998, Scott 1998); for an overview, see BIOSIS (1999).

BioCISE: The Vision

A common electronic access system facilitating queries across the hundreds of millions of specimens and monitoring or mapping records held by institutions, projects and individual researchers in the EU and partner countries.

Data capture in existing collections

Due to technical advances, the digitisation of even huge collections and printed survey data becomes feasible in principle, while the overall cost of digitisation is relatively small compared to the costs of gathering and maintenance of specimens. The
collections-community generally agrees in principle that computerisation can and should be achieved. However, laboratories often presently lack the resources even to digitise the data for newly acquired units, let alone updating their databases for older parts of their collections.

Additional difficulties arising in the digitisation of older stock relate to aspects of data quality. Historical collections may carry very little accurate geospatial information, while survey data records are rarely vouchered, thus not allowing verification of taxonomic identifications. Standard geospatial information for historical (and even some recent) collections is necessary to analyze these important information resources with modern methods. However, vague locations and casual observations often characterize older collection data, whereas records resulting from modern official surveys or comprehensive scientific research programs usually provide clear polygon or point-based data. The conversion is error-prone for various reasons. Location names may no longer exist, have changed their circumscription, or have been applied to more

### Collection Information System

Biological collection information systems provide information on the single “unit”, i.e. the individual specimen or observation. Not directly part of such a system is synthesised information such as species-related information, e.g. information on migratory species, resistance to heavy metals, etc. The main data areas of a biological collection information system thus consist of:

- Gathering event (who collected, when, in what context)
- Gathering site (location, geographical and ecological features)
- Determination (taxon, who did it)
- Unit information (age, stage, sex, description, states as nomenclatural type, etc.)
- Unit (specimen) location (laboratory, duplicate distribution)
- Unit (mostly specimen) management (catalogue and accession number, preservation, storage, amount of material)

than one area in a certain region; national map grids have been used that do not correspond with the general system of latitudes and longitudes, or geographical information is very general or even lacking.

Generally, raw data from historic and many recent collections and surveys may be of low accuracy and thus cannot be used directly for analysis. Evaluating and – to some
Networking biological collections

extent – standardizing the data greatly enhances their usability, but it must be done with experts and may require further checks and research, especially with respect to historic units. It is obviously not feasible to check each record individually; hence techniques must be developed to automatically identify “suspect” record sets (e.g. outliers) for expert re-examination (Chapman 1992). For natural history specimens, the original information should be made accessible together with the interpreted data to allow re-examination of labels, for example. Digital pictures of specimens, labels, collector's field books, accession ledgers, etc. provide a new possibility to achieve this, with the advantageous possibility of de-coupling the capture of quality data from the location and handling of the specimen itself (Berendsohn 1999b).

In user workshops and interviews, opinions differed widely on whether public access should be given to raw data. On one hand, raw data may be misleading and may even damage the reputation of the data provider; on the other hand, the decision of which data to use should be left to the users. Raw and validated data should be marked to be clearly distinguishable. Additional information on performed data validation becomes crucial, especially where datasets from a number of different sources are combined in a common access system. However, in many cases certain items in the raw data (such as the name of the collector and / or identifier) may help the informed user to assess data quality, without further effort on the provider's side. In survey records, the possibility to make raw data such as pictures or sound records available to the user will allow for the checking of taxonomical identifications, in some ways analogous (though not equal) to the citation of specimens.

Ease of capture vs. data quality

Progress in data capture was greatly impeded by a lack of adequate software. The complexity of collection information (see Chapter III) was often underestimated, leading to much duplication of efforts and inadequate software solutions which are often being maintained by research scientists and curators. Those not involved in the development of the software often regarded databasing as an additional complication of curator's task, sometimes even as an impediment for research, competing for scarce resources. An often-heard argument against some of the advanced systems, which are only now becoming available, is also that they are “too complicated”. There is some truth in this point of view: a database application will always be less straightforward to handle than a word processor, since the information must be structured and data quality is usually at least partly checked at the point of entry. However, this is also the strength of data capture in a database: information quality is usually higher, and the information can be linked to other sources. The more atomised the data areas are and the more “fields” a user has had to fill in, the easier is it to pursue aims like standardization, error tracking, and linking. Fortunately, the power of today's desktop computers increasingly allows implementing interactive functions to ease the task of user input. Locality data capture from on-screen maps is an example. Input of scientific names etc. as text in a
single “field”, which is parsed into its atomic data element, processed in the background and (if necessary) corrected by feedback mechanisms is another.

At least for new collections, the problem of geographical locality input is increasingly eased because of the availability of accurate co-ordinate data captured in the field using Global Positioning Systems. Where co-ordinates are not available, on-line gazetteers are evolving into valuable aids for data input (see Berendsohn 1999c). The input of taxonomic data (names) also becomes less error prone due to the availability of catalogues on the WWW (e.g. The Plant Names Project 1999, Farr & Zijlstra 1999, IOPI 1999).

**Promises of networking**

When talking about collection information systems, we start by thinking of an individual database of a floristic mapping project or a museum’s collection. Even within a single database, applying the right tools to the data collection can give new insights. Simple analyses can provide geographical distribution patterns of species and population dynamics from original data. Sometimes biological data can be combined with records of environmental parameters, climate data, and taxon information for more complex statistics. Very often, though, databases are restricted to a special field of application. Mapping project databases seldom include specimen information but usually just observation records, museums collections are typically restricted to physical object data.

With queries being aimed across wider organism groups, time, and geo-ecological criteria, answers are sought not only combining information from different fields within one database, but also integrating contents from a number of distinct databases in an information system. Interoperability of different collection information systems, and possibly a common search interface, are the natural extensions of combined queries.

Drawing together information from different sources is well worthwhile: A common access to information now stored in dispersed, autonomous, and heterogeneous databases could considerably enlarge the information content, not only by expanding accessible records, but adding value by giving new insights. Some examples shall demonstrate this:

**Tracing of toxic agents.** Decreases in some bird populations, caused by a thinning of eggshells and first observed in the field, could be traced back to the wide spread use of DDT by analyses on natural science collections of eggs (Duckworth et al. 1993). Similarly, dynamics in the concentration of heavy metals in the environment can be inferred from analyses of hair samples taken from museum objects. Given sufficient material to establish a time series, such data can be used to reconstruct the history and causes of species decline.

**Biodiversity assessments in developing countries.** Natural history collections in Europe are holding huge amounts of data on the biodiversity of developing countries in
Networking biological collections

the form of specimens and associated data (labels etc.). In these countries, which often count with high alpha and beta diversity, the combination of such data with local information and expertise can be put to work for a variety of tasks, ranging from defining priority areas for conservation measures to research planning (Soberón & al. 1996).

**Analysis of environmental change.** Since higher plants adapt to different atmospheric concentrations of CO₂ by adjusting the density of stomata on their leaves, this feature may be exploited in the reconstruction of changes in the atmospheric composition. Some caution applied in the interpretation, taking into account other possible influences, data and material collected over a time period at a given location allow conclusions on the development of the CO₂ concentration (Sieders 1998, Wagner 1998).

Another example of environmental parameter reconstruction comes from the aquatic sector: Analysing the community composition of diatoms, a group of unicellular, shell bearing algae, is a well established tool in the inference of a number of hydrochemical and hydrophysical properties of water bodies, including acidity and nutrient availability. Due to the highly solution resistant silica shells, a long timeline can be assembled from recent water samples combined with analyses of (sub-)fossil communities in sediment cores (Battarbee 1981, van Dam & Beljaars 1984, van Dam 1996, Juggins et al. 1996). Samples to be included into such an investigation can also be gathered off aquatic macrophytes in historical or recent past herbarium collections (van Dam & Mertens 1993, ter Braak & van Dam 1989).

Chapman (1992) points out that a visualisation of environmental parameters (e.g. climatic profiles) may help to indicate apparently suitable locations for a species. Though this approach alone does not allow accurate predictions on the distribution of species, in a combination with actual observation records it could be used to focus survey efforts on areas revealing a higher probability for finding additional populations or to confirm predicted distribution boundaries. Information combined from several databases might form the basis to model biological interactions as, e.g., how changes in bird distribution are likely to affect their role in pollination or seed dispersal, and what the consequences will be for the vegetation. This leads up to the use of a computerised information system in biodiversity research: “The more detailed knowledge is available, the more we can begin to ask complex questions” (Blackmore 1998).

**Problems arising from combining information from different sources**

With drawing together datasets from different databases, some problems arise apart from the purely technical difficulties in interconnecting highly heterogeneous sources. We are facing four major complexes:
Heterogeneous data definitions
How can datasets be combined which use a variety of different reference systems in geography (point data, several grid nets, polygon data,..), taxonomy (classification schemes and taxon names highly varying with time of identification and opinion), to name but the two most important? These aspects will have to be solved in a general approach while setting up a collection information service: Different reference systems in collection databases will have to be taken as they are, and cross-referenced in a background structure (thesaurus and other representations and mapping of differing concepts, see Chapter XI).

Data quality
How can data quality be measured and indicated? This is a multi-facetted and multi-layered problem. Ideally, every data item should be traceable as to its source, changes made, and as to the people who handled these processes. However, following this through would probably overburden any system, let alone a large network of systems. For catalogues of collections providing commercial materials strict standardization of procedures may lead to far reaching quality control and ensuing reliability (see Chapter VI). However, for most non-commercial collections, assembled in the process of scientific research in a variety of fields and numerous subjects, this type of quality control would certainly not be achievable. Moreover, such measures cannot be applied to historical materials, which are undoubtedly a most valuable resource of natural history collections. Traditionally, personal data such as the collector’s and the identifier’s name are often the expert's criterion to assess data reliability. However, large-scale publication of personal data, from which, e.g., itineraries can be reconstructed, may not be liked by all persons in question, and may even be illegal. A kind of peer review process of collection data, i.e. having the data scrutinized before publishing it is also not feasible, because the amount of expert personnel resources this would take are completely out of question. However, the current mechanism of annotating specimens in natural history collections could be extended to annotate data on the network, thus providing additional information to the user.

In the short term, we need to use mechanisms to describe entire datasets as accurately as possible, so to hand on information to the user from which they may be able to judge. Very useful first steps in this direction are standardized content descriptions of “documents” (in this case: database) content. A set of such attributes has been published as the “Dublin Core” (Anon. 1998), which has been achieved by broad international consensus. Further standardization of the structure of individual databases participating in the network is also significant for data quality, because an exact definition of the information content of individual attributes is facilitated. Defining a set of attributes as a common denominator for access to data on individual specimens in natural history collections is a question that has been tackled by the ZBIG project in the United States (Vieglas 1999) and, more recently, by the ENHSIN project in Europe.
The European Natural History Specimen Information Network is an EU financed project where major European collection holders united to solve problems related to common data access. ENHSIN goes back to an initiative by CETAF – the Consortium of European Large Scale Taxonomic Facilities – and BioCISE.) These efforts will hopefully ensue in a common convention, probably under the auspices of the IUBS Commission for Taxonomic Databases (TDWG).

**Sensitive data**

Advantages of access to a high number of detailed datasets notwithstanding, there are also some sensitive data to which access restrictions have to be closely observed. This includes the personal data of collectors, identifiers and other people associated with collection and evaluation, but also information on, e.g., endangered species or research in progress. Again, this touches aspects of general policy decisions to be made in the installation of an information service (who shall be given access to what data?), and autonomous decisions of the data owners, to hold back certain information (datasets or parts of them) from publication.

**Intellectual property rights**

Above all with survey data, one reservation is often stated more or less openly: If we make our observation data publicly available, there is nothing left – our data are all we have. Apart from the aspect that most field surveys are publicly funded, a contribution to a common knowledge base could be far outweighed by the increase in information to be derived from it, providing dense time-series through the constant accumulation of new data – an activity which will still be an essential of environmental research. Protection of the (often considerable) investment information providers made in their database was discussed along several lines during the BioCISE workshops: Outright sale of the data to a national agency is one possibility, a “Join the Club” approach could be another, with all members profiting from the access to the others’ data. The idea of federated databases, i.e. several databases continuing to be held and maintained by the provider, but accessible in a common system, is another option.

In any case, a common access system must adequately recognize the respective origin of all data, and make it compelling for any user to correctly cite the sources he exploited. Collections represent an enormous knowledge base on global biodiversity. This knowledge has been produced by human beings, who work for or in research institutes or companies, or by motivated individuals. In natural history collections, the materials and data gathered stem from lands belonging to people under a certain local and national jurisdiction. The scope of problems that may relate to collections being published on networks is enormous. For example, how can benefits derived from using the system be properly shared with the countries of origin (Biodiversity Convention)? How can authors of individual data items, such as the person identifying a certain
specimen, be properly credited for their work? How can collection institutions ensure their stake in the information made available?

These are largely unresolved questions, paralleled in many other fields where knowledge is being made accessible over the network. The ENHSIN (2000) project contains a work package dealing with these subjects. Finding pragmatic solutions is a priority, because IPR problems could considerably impede progress towards a common information service.
III. The complexity of collection information

Walter G. Berendsohn and Pier Luigi Nimis

Living materials for biotechnology such as microbial strains address a number of industrial and high priority research areas. Therefore, computerization of catalogues and standardization in these collections is much further advanced as compared to other types of collections (see Chapter VI). But also non-commercial living collections are usually further advanced in their computerized management than natural history collections – simply because inventories are not static and must be managed, and also because the number of units for example in a botanical garden is usually much smaller than in a herbarium.

Entry of materials into natural history collections is more of a museal task. Once the appropriate conditions for storage are achieved, the material documents itself; information regarding the specimen is often stored with the material. Access to the information traditionally meant accessing the material. The aim of a collection information service is on the one hand to help to localise appropriate materials, on the other hand to uncouple at least some of the needed information from the object itself.

“Units” in the BioCISE Model

Any object containing, being, or being part of a living, petrified, or conserved organism is considered a unit as soon as it appears in the system. The unit may be gathered (observed or collected) in the field and derived units may recursively emerge from it through specimen processing, breeding or cultivation. In addition, Units may form Associations (e.g. host/parasite), Ensembles (lichen on a rock with fossils), and Assemblages (herd, artificial grouping). Gathering events, specimen management (acquisition, accession, storage, preservation, exchange, ownership), and taxonomic or other identifications relate to the Unit.

The term “specimen” can often be used as a synonym of unit, however, it lacks a precise, context independent definition, it is often perceived in a narrower sense compared to the unit, but as the example shows, a single specimen can represent a number of units.

To reach the latter aim, the information structures present in biological collection units were to be investigated and several information models were published.
Fig. 1: Lichen envelope (above) and specimen (below) from Trieste
A network diagram showing the accommodation of the individual data items in the model. The oval shapes depict instances of entity types, with their name given in parentheses, followed by a selection of attribute values. A vertical bar indicates that attributes from related entity types have been included (e.g. the Gathering Site represents the entire geo-ecological reference system). The connecting arrows represent instances of relationships. The arrowhead designates the sending end of the relationship. The color of the arrowhead denotes the unit’s taxonomic identity (field data, Calypso thalassina commutata, Bagletto purpurascens, Bagletto christiansenii).

A dotted outline indicates unit ensembles, i.e., units that are normally handled and stored together. Dotted arrows represent links with instances from other entity types that refer to every individual unit in the ensemble.

Throughout the diagram, the only attributes cited for derived units are the availability (followed by the reason, if not available), and (occasionally) the essential category. For Bremondhede et al. 1995a for the entire information model.

The diagram is intended to show the relationship between the different units and their attributes. It is not exhaustive of all possible relationships, but rather a selection that illustrates the main concepts.

**Fig. 2. Instance Diagram (Rumbaugh 1991, adapted)**

The Gathering or Field Unit is in this case of completely virtual nature: it serves only to connect the field data to the other units and to document the original storage of the Material in Eino.
Intentionally left blank
The first information model becoming widely available was the result of a workshop of the Association of Systematics Collections in the United States, the so-called ASC model (ASC 1993). CDEFD and finally the BioCISE model (Berendsohn et al. 1999a) are later and broader approaches. The latter is thought of as a reference model, i.e. a tool to interpret data rather than a model that should be turned directly into a database.

We have used a set of units represented by a specimen in the lichen collection of the herbarium in Trieste (TSB) to illustrate the degree of complexity which (at least in some cases) has to be documented in natural history collections, and to illustrate how that information is handled in the BioCISE model.

The example. The envelope (fig. 1, top) is part of the Lichenes Selecti Exsiccata, distributed by Prof. Antonin Vezda of the Botanical Institute of the Czech Academy of Sciences in Brno. The original material was collected by P.L. Nimis, J. Poelt and A. Vezda in Sicily, and was brought to Brno by the last collector, for later distribution. Meanwhile, Poelt and Nimis studied the material, and decided that it belonged to a new species; the formal description was sent to Vezda, in order to have it on the exsiccatum itself. The holotype (i.e. the specimen the scientific name is based on) is the specimen conserved in Graz (GZU); all other specimens distributed within the exsiccatum are isotypes. The information printed on the original label (fig. 1) is valid for all specimens distributed within the exsiccatum.

Further information was added on the label and inside the envelope during the history of this specific envelope. The history is as follows: GZU (Graz) received two pieces of the exsiccatum, both were accessioned. One was retained in GZU as the holotype, the other was sent to Trieste (TSB) as a gift.

In Trieste the sample received a new accession number, and was placed in the reference herbarium, which has a different location than the main herbarium. In Trieste, a lichenologist (M. Tretiach) made a note on the label, stating that the new name is probably a synonym. A chemical analysis of the secondary, accompanying species was carried out (by Nimis) and a parasite was discovered on the main species (by M. Castello). Nimis, furthermore, made a slide of the apothecia and a drawing of the spores, which were placed inside the envelope (left picture).

Later in this booklet (chapter IV) it will be shown that potential users of a collection information service will want to use several criteria to query for this specimen. Almost all the information stated here is relevant for research purposes: the presence of the different species determined at a defined locality at the time of gathering, the host-parasite relationship, the new name based on one of the lichens present in the specimen, etc.

Representation in the model. To be able to access this information efficiently in a database eventually representing holdings of many millions of specimens, this information has to be digitised in a highly structured and standardized form. Fig. 2 shows how the BioCISE information model accommodates all details given. As said before, this does not mean that the database must be organised along this line, but it
should be possible to extract data represented in this information structure from it accurately by appropriate means.

The diagram depicts, in a rather simplified form, that the information contained in the specimen in Trieste represents a high number of units in the database, some of them only of an intermediate, “virtual” nature, some of them representing materials (the specimens or parts of the specimens) at hand in a collection. The single “Gathering or Field Unit” corresponds to the materials as found and taken in the field. It serves as an interface between the gathering's circumstances (the who, how, where, and when) and all “Derived Units” stemming from the material. With the identification of the material as belonging to two different species, Nimis and Poelt created two further “virtual” units, because by definition a unit has to be homogeneous as to its taxonomic identification. They formed an Ensemble, because they continued to be handled together. We can assume that Vezda, while creating 25 specimens from the original material, ensured that all of them contained both species, because he stated so on the label. In effect, he thus created 50 new derived units by sending the material to 25 herbaria around the world. One, designated as the holotype of Caloplaca thamnoblasta by Nimis and Poelt ended up in the Herbarium in Graz (GZU), of course still forming an Ensemble with the other species. We presently do not know if any further work was done on the specimen there, nor what happened with 23 of the other specimens. However, we do know that several additional “Derived Unit Creation Events” took place in Trieste: Tretiach's identification did not change the unit's circumscription nor content, so it remained unchanged. However, the identification of yet another species (a parasite) created a new Ensemble of three units, two of which (the host and the parasite) form an association. Two further events, the addition of a microscopic slide sample, and of a spore drawing brought the final number of units in the specimen to 5, its present state.

At first this looks like an enormous complication to the seemingly simple form of traditional data storage (“it's all there on the specimen!”). However, considering the possibilities of networking the results and retrieving all results obtained from the entire set of specimens already shows some of the benefits a highly structured system could bring about. To be realistic – we will not have this degree of detail for all collections anytime soon, if ever (see previous chapter). However, although biodiversity research has many aspects, one of its main information sources are collections. As will be shown in chapter IV, data to be used in research applications have to be of a high quality and exact description. The BioCISE model demonstrates that biological collections have a common information structure, which – if properly implemented – could enhance the value of collection information systems for present and future research applications. As will be shown in the last two chapters, one of the basic prerequisites for the functioning of a European collection service has to be its scalability, its ability to integrate highly heterogeneous degrees of data atomisation, structuring and quality control.
IV. Users and uses of biological collections

Birgit Felinks, Andrea Hahn, Linda Olsvig-Whittaker and Wouter Los

Introduction
Throughout the project, BioCISE collected information on users and potential users of biological collections. Because of the broad, interdisciplinary approach of the project the identification of user groups and user interests had to follow an informal procedure – BioCISE used meetings, discussions and interviews rather than questionnaires to gather information. The needs of users, especially of those from the environmental sector have been in the focus of two workshops. Various interviews and discussions were conducted during national and international ecological congresses. Most helpful have also been responses from Taxacom and ZBIG-L e-mail lists to a query asking for “most useful queries” to a common collection access system (Beach 1998). The broad approach of BioCISE, originally based on the analysis of the underlying information structures of biological collections, was justified once more by the results with respect to user communities: users are only rarely interested in a single category of collections.

To provide a framework for the results, the project focussed on the following questions:
Can we categorise users into broad groups according to their primary interests in collection information? Can we categorize uses of collections? What are the common denominators among different user groups and different uses of collections, i.e. what are the most important tasks a collection information service should tackle?

Four major groups of users may be distinguished:

• Users from the fields of biotechnological research and industry (applied research and production), e.g. the biotechnology company looking for a particular bacterial strain or for a plant producing a specific natural substance.

• Users from biological systematics and university research (basic research) e.g. the zoologist looking for type specimens within a taxonomic group to finish her revision of the group.

• Users working in the environmental sector (monitoring and impact studies, development of guidelines) e.g. the environmentalist compiling information on past changes in the flora and fauna of a particular area to assess the possible impact of building activities or invasive species (but also: the customs officer seeking information on the legal status of plant or animal specimens to be imported).

• The general public (education and entertainment), e.g. the teacher inquiring for the whereabouts of oak species in the Botanic Garden to prepare a visit with his second grade, or the child interested in large dinosaur skeletons exhibited in a particular city.
Uses of biological collections fall into the following categories:

- Supply of properly identified raw material (for the extraction of natural substances, DNA, parent plants for crosses, or as educational materials for university courses).
- Supply of research material (taxonomic investigations, molecular and genetic research, global change measurements in preserved materials, etc.).
- Supply of variability data (morphological and genetic variability can - to a certain extent - be preserved and investigated in biological collections).
- Supply of ecological data (label and ledger data, as well as site sampling can provide synecological information, water samples provide pollution data, etc.).
- Voucher and reference purposes (taxonomic identification by comparison, nomenclatural types, and research vouchers ensuring the replicability of research results).
- Supply of occurrence data (where and when was the organism collected or observed? This may help indirectly to supply materials, and it may serve as ecological data, e.g. in monitoring invasive species).

This introduction already makes clear that many different communities can use collection information in variable ways. “Traditional means of access, – personal visits to collections and long diligent searches of paper records – guarantee that most [potential users of biological collections] will find other means of acquiring the information they need, or act with insufficient information” (ASC 1993). This means that they will not utilise the primary sources, collection objects, but sources like secondary literature or “expert opinion”, perpetuating earlier, sometimes erroneous interpretations instead of adding to knowledge (OECD 1999).

Knowledge of particular holdings of a collection (e.g. of algal strains) may be quite common within a certain interest group (e.g. systematic phycologists). However, this does not mean that the hydrobiologist who looks into the causes for a heavy increase in biogenic toxin concentration in the local pond is aware of the existence of a reference collection to identify the culprit, and possibly learn more about it. That information may already be available, perhaps even in an on-line database. A common access system should help to access information either by pointing to relevant resources or – better still – by directly providing the needed information. It is obviously important to consider user needs from the beginning in the design process of such a system. The four identified user groups shall be introduced in more detail, and an attempt at summarising their needs and wishes will be made. Some of those may seem obvious; not all of them will be easily satisfied.
Major categories of users

Users from the fields of biotechnological research and industry

Biotechnology. Research areas where collections are used include pharmaceutics, natural products research, medical research on pathogens, vectors, tissue cultures, plant and animal breeding (including genetic engineering), research on plant pathogens and host-parasite relationships, and weed research. Collections here serve mainly to provide raw material: plant or other material either for direct use, or to serve as the research object in chemical compound analysis and synthesis. Less frequently, collections are directly used as a reference for identification by comparison (e.g. seed collections in weed research or forestry), a use that is normally mediated by systematists (see below). In view of discussions on the “intellectual property rights” on active substances isolated from indigenous plants (Gollin 1999), and considering more and more restrictions to the exploitation of wild plants, research on cultivated material will increase in importance. A further field of application is that of test organisms from different groups (test animals, cell cultures, etc.) for the indication of potential toxicity effects of products like medicines or cosmetics, or for bioassay indication for waste products in routine applications such as the screening of water quality.

Consequently, information requested is usually to enable the user to access physical objects of a certain kind, e.g. belonging to a taxon or microbial strain with specific properties. The system should supply addresses and links to Internet sites with on-line databases or provide direct access to catalogues of commercial products. All data should be accessible fast, be regularly updated and presented in a unified form, and the materials from different sources should adhere to common quality standards. In pharmaceutics, the need for an effective way to extract data from a variety of different database systems has led to a move to link databases by drug firms by establishing a common standard (Williams 1997). The CABRI (Common Access to Biological Resources and Information) service is a working system providing such an access to commercial collection materials, such as microbial strains and tissue cultures (see Chapter VI).

Users from biological systematics and university research

Systematic research and taxonomy play a key role in the access to all organism-related data. Research institutes in this field provide the basis for information exchange in all the other fields discussed here by providing a common reference system (scientific names), keeping track of the history of name changes, by enforcing the preservation of voucher specimens for the reference system (nomenclatural types) and by offering long-term preservation of vouchers in natural history collections. Systematic research is mainly conducted in systematics departments of universities and natural history museums, but often it is also an integral part of other biodiversity related activities, for
example when organism inventories (checklists) are compiled in ecology. As the common reference system, organism names link the results from different areas of biological research, such as molecular sequence data or ecological observations.

Taxonomy is based on collection objects, and collection objects supplement taxonomic data as a reference system of vouchers. Nomenclature in zoology and botany depends on type specimens to preserve the original objects to which names are attached. Living objects are used to examine the variability of organisms.

Systematists often request highly specific collection information, e.g. on the presence of a single specimen in the collection (one that has been cited in a publication, for example). For taxonomic revisions – the treatment of a specific group of organisms – loans of specimens belonging to the taxon under investigation are requested for further examination. Historic collections are often of particular importance because the concept of the taxon a previous investigator formed is becoming clear by examining the material seen at that time. In addition, in larger collections it may include hitherto unrecognised type specimens. Data on collectors and the date of collection may be valuable to pre-select specimens, as the data on the collection site may be, because most revisions are restricted in their geographical scope. The name of the identifier may give an indication of the reliability of the identification to the informed user. A picture of the specimen or its label can in most groups of organism provide valuable (original!) information in a preliminary phase of research, to decide if the specimen may be of interest for closer examination. Experienced specialists in specific groups are scarce; this already constitutes the “taxonomic impediment” in biodiversity research (ABRS 1998). A collection information service should thus aid systematists to gain access to needed research materials with as little as possible waste of precious time. In appropriate groups, digitised images may also aid in the training of a new generation of taxonomists who are able to recognize taxa by looking at specimens or the organism itself.

Research in molecular systematics usually relies on a supply of fresh material of secure identification. Ideally, material from the type locality of the relevant (name giving) species of the group should be investigated, or at least the type species of the examined group. Since this can rarely be achieved, the quality of the identification and the selection of the material attained using “traditional” taxonomic techniques, become even more important.

Users from the environmental sector

Environmental users belong to three main categories: (i) the domain of public service observation, monitoring and preservation of the environment, including related administration; (ii) commercial services (landscape planning, monitoring and mapping projects, and impact assessment studies); and (iii) the sector of environmental education (teachers, public relations staff, and officials in contact with the general public).
Public services and administration. Users from the sectors of public services, science, and administration in environmental observation, monitoring, and preservation belong to, e.g., environment agencies on different national to supra-national levels, commissions concerned with nature protection (e.g. the IUCN Species Survival Commission for the protection of endangered species), nature reserves and national parks. They carry responsibility for nature conservation, are concerned with ecological research in universities and other projects (e.g. the long time ecological research network – LTER 1999), and work in administrative functions responsible for decisions regarding environmental protection. The scope of their work and responsibilities reaches from basic research to legislation, and from global scale (as in global change research) to concerns of an individual nature reserve.

What is expected from them (and for which to deliver they expect support from collection information systems) are background research results on the functioning of ecosystems and long-term as well as short-term ecological phenomena; the provision and publication of data and information on, e.g., cumulative effects of human activities and natural processes on the environment; the identification and effective protection of areas with high conservation value and of endangered species; and the fulfilment of legal obligations like those arising from the Rio Convention, compelling all signing states to provide information on specimens stored in museums or records held by institutions specialised on observations.

These users need data: as the base for their own information evaluation and publication. They need to present suitably processed information to inform the general public. They need data in the form of digitised species lists and databased inventories in order to outline areas with high conservation value and to answer specific questions (for example, for the distribution and occurrence of endemic, endangered, or invasive species). In species protection, they need as much species related data as possible to establish successful management plans and protection measures. Research on and understanding of long-term ecological phenomena across national and regional boundaries, contributing to the comparability of ecological information and indirect measurements of environmental changes, requires access to complete time series covering a vast spatial range to enable effective pattern analysis.

Commercial services. Like the public services concerned with the environment at large, enterprises working in this field usually are suppliers as well as users of collection information. As private or public companies, they work in landscape planning, in monitoring and mapping projects, and as consultants in impact assessment studies. While contributing to the data “pool” by own field studies in principle, evaluation results and decisions often have to be reached fast: Contracts for impact assessments usually require almost immediate results, and there may be no time for collecting comprehensive field data by surveys and research. In this case, the professional has to rely on already published data.
**Environmental education.** The term “education” is here understood to include teachers at schools and universities, but also, in a broader sense, PR staff of environmental agencies and nature reserves or, e.g., rangers in national parks who are in contact with visitors. Their task, in general, is to rise public awareness of environmental and nature conservation issues at different levels, and to generate an interest in and better acceptance of environmental goals in the public.

What they require, above all, are the tools to provide motivation: Comprehensible, pre-processed data integrated with new electronic media, pictures and animated visualisations, demonstration objects, and the general possibility to involve their clientele in interactive processes like, e.g., the participation of school classes in an internet-based observation project.

**In summary,** users from the environmental sector want access to a completely digitized and databased inventory of biological object collections as well as survey data. Data should be unit-level if possible, though information describing collection contents and pointing the way to additional sources is highly appreciated where unit-level data are not available. The presentation of search results should provide the user with choices regarding the level of detail he wants to review, should offer means to flexibly switch between different visualization modes, and enable the user to download datasets into his own files and applications.

All three categories of users from the environmental sector have in common that – when querying a database – they primarily focus on the geo-ecological, the temporal, and the taxonomic domain. Key questions users from the environmental sector want to pose to a database address the “where”, “when”, “what”, “who”, and “how many” of observations and collection objects. Discussions during the user workshops showed that access to information by a defined location is often top priority. Primary access questions for users from the environmental domain to query a collection database are geo-ecological and temporal criteria, while taxonomy is often used as a kind of “connector” to add organism-related information.

**Users from education and entertainment**

The “general public” forms a very heterogeneous group. It combines pupils and teachers from schools and other educational institutions (high schools, universities, adult education), visitors of museums, zoological and botanical gardens, “interested laypersons” and hobby researchers, but also customers of commercial services. In collections, they expect to find recreation value (a botanical garden to jog in, going to looking at the exhibitions of the natural history museum), demonstration objects (especially in education, using living organisms as well as preserved specimens), or a selection of goods for sale (nurseries or seed catalogue). In education, collections largely serve to provide motivation and visualize the subject. Projects stimulating observations, e.g. taking note of the first song of a blackbird heard in spring, or the first fruit of an acorn found, also lead to a closer scrutiny and increased use of observational
data already available – see, for example, the project Nature Detectives ("Naturdetektive", Freiberg 1999) of the German Clearing House Mechanism. Easy access to quantities of high quality (microscopic) images, video clips and audiovisual data can also help the student to become familiar with basic structures and biological processes (Schumann 1998). It should be added that „Biology with its combination of different concepts, different kinds of information, and ‘attractive’ pictorial information provides an ideal experimental platform to develop new learning tools” (Schalk & Los 1998).

For a number of users, the key information is access information: the visiting address of a zoological garden, the address of where to send the order or request a catalogue, or where to find a contact person for a specific question. Descriptive information on collection holdings is equally important, for example commercial catalogues of rose plants. Media sources (pictures, sounds) are also in high demand, preferably being accessible via the Internet.

Access to biological collection information

In this section, we will detail the main criteria by which users want to search a database, and the manner in which they would like to pose their query.

Requests by geo-ecological criteria

With some exceptions, specimens and observations in biological collections carry information as to their provenance, i.e. where the unit was originally collected or observed. Spatial search criteria cover geographical aspects (“Mediterranean coast”; “the Pyrenees”), political entities (“Italy”; “Burgundy”, “Sevilla”) and ecological designations (“moorlands”, “deciduous woods”).

Access tools for a database quest are inter-active maps and lists of location names, combined with free-text searches. Map access in particular is most attractive. The possibilities to select an area by “clicking” or circling it on the computer screen, by selecting co-ordinates or positions in grid or polygons, by entering point data and specifying a radius, by zooming in, and by the selection from a subset after entering habitat specifications (“heather”) are an intuitive as well as effective ways of searching by geo-ecological criteria. Supplementary lists of location and habitat names (detailing, e.g., country, province, state, floristic regions, nature reserves, parks, aquatic and terrestrial habitats) allow searches in cases where exact locations are less obvious on a map. They offer a pre-selection to choose from, and ideally also visualize the corresponding locations on a map. An additional free text search provides a means for less foreseeable quests and more detailed specifications, also comparing them to the thesaurus data used for the lists to suggest or use cross-links where possible.

Access to collection information by means of geo-ecological criteria is clearly most prevalent for the users from the environmental sector. Systematists working on the flora
or fauna of a specific region, or investigating the geographical distribution or ecological range of a species also use it. Educational and general public use is often focused on the local resources. For biotechnology, locality data are only important in the context of gaining access to materials (e.g. natural substances), where gathering the material in the wild is the only means to obtain it.

Requests for a defined time period or time sequence

Temporal criteria consist of roughly four categories: “the present”, “the (near) past”, historical, and pre-historical. Questions for the present ask for up-to-date records, mainly of observation data, and possibly on vouchers: What is there to be had at present? What is the present state of knowledge? The past, further subdivided into the “near” past or short-term (up to about 20 years), and long-term (~ 100 years), is mainly addressed for retrieving time series: What observations have been made, what specimens collected over the past 20 / the last 100 years? Quests along the historical timescale ask, e.g., for herbarium specimens or natural history collection objects collected within a given period, while pre-historic criteria concern material of (sub-) fossil origin.

Correspondingly, specifying a time period is the most common approach. A user may ask what records are available from 1980 to 1990. Other approaches include asking for an exact date, specifying an end or a start date (“give me all entries before May, 1920” or, in its extreme form, “show me the first / the last record ever made”), and requesting a seasonal time series (“all records ever made in June”). Time-related questions therefore ask for a much higher degree of “tolerance” on the part of the user interface than might be expected. The interface must be able to handle specific dates and periods defined by specific start and end dates, but also be able to process requests just stating a month for seasonal series, or allow searches for “the most recent” entry without any date specified at all.

Requests for a given taxon

Taxon names are often regarded as some kind of “handle” used to retrieve organism related information – the taxonomic details being of secondary importance. Queries are posed either by the scientific name of an organism, or by its trivial name. Preferred ways of access are selecting a taxon name from a menu, browsing through a hierarchy, or entering a name in a free-text search. Most common queries are for names on hierarchical levels from genus level downwards (genus, species, subspecies), although questions for higher ranks (e.g. family) are not unusual. For users in contexts like education or environmental campaigning, enhancing the hierarchical and list selection search by offering pictures of representative members for the selected group would further contribute to a useful search interface.

In general, entering a search term in any of the three specified ways (list selection, hierarchical browsing, free-text search) should hide taxonomic details from the user as
much as possible. The main objective is “to use the taxonomic framework to establish a connection with the available knowledge accurately accessed under the ‘correct’ species. This means that the users in applied biology prefer the taxonomic indexing to be completely invisible and yet perfect in its functioning, comprising automated synonymic indexing as well as precise handling of the ambiguities caused by homonyms, misapplied names and pro-parte synonyms” (Bisby 1998). In contrast, a user from systematics will be interested in details such as the identification history of a specific specimen.

Additional search criteria

In addition to these three main access criteria in searches on biological databases, there are some others deserving mention. First of all, the name of the collector or observer appears as a search criterion: What objects are there collected by xyz? Where has xyz collected, when have expeditions taken place? Incomplete specimen data, especially in historic specimens, can often be reconstructed by such questions. Systematists looking for type material often request a specific specimen, identified by collector and collection number. A second regular type of question addresses organism characteristics: What data on host-parasite relationships are available? What observations / specimens are there of flowering stages? of larval stages? However, information requests may also refer to a specific institution: What does it have, is it worth a visit? The location of the institution may also be important – what natural history collections can I find in Edinburgh?

Combined search criteria

In a real-world information system working with large data resources, users are normally posing their questions as a combination of criteria. Typical combined requests in collection database searches are:

- Specimens or observations of a certain taxon from a defined area.
- Distribution of taxa within a defined location at a defined time period. On a national level this may be interesting for decision-makers in the field of landscape planning or nature conservation, on an international level rather for research and education.
- Comparison of distribution maps of different taxa in order to detect and represent spatial patterns.
- Correlation of collection data with environmental parameters. The resulting distribution maps might help to elucidate species variation along spatial or environmental gradients.
- Distribution maps of endangered or indicator species, and vegetation types. This access will mostly be used for the establishment of management plans and to
answer questions e.g. whether endangered species occur within a protected area or not. For the purpose of nature conservation, some data will have only restricted access.

- Access to information on autecology and synecology, e.g. life forms, functional morphology, sociological affinity, “Red List” status, symbiotic relationships, parasitic/non parasitic or host associations (Smith 1998), pollination indices etc.
- Distribution maps and patterns of migratory species.
- Biogeographical spectra.
- Biomonitoring (e.g. Mediterranean lichen database, Grube & Nimis 1997).

![Diagram of access to collection information in the environmental sector]

**Figure 3: Access to collection information in the environmental sector**

**Desired output**

The following sections will specify the users’ expectations as to what the database should deliver, and in which form. What kind of information do they want to retrieve (content) and in which form should this information be delivered (presentation)? Once
more, the account somewhat focuses on users from the environmental sector, as these seem to have the most complex requirements.

**Data content**

Answers to the question of “what kind of data do users want to get” are as diverse as the user groups, and differ with the motivation the user had to pose the query in the first place. They may best be typified by breaking them down from the most general to the most specific:

- **General orientation**, stating what is there (in a given region, at a given time). This kind of very general approach is usually associated with education and administration.

- **Statistical overview**, based on detailed presence/absence data. What observations/collections have been made, and where/when have species been reported missing? The purpose is often a general evaluation, as used in nature conservation and environmental protection. Information on gaps is of interest, to identify under-sampled areas both in taxonomical and geo-ecological aspects and to define “interesting” areas for further assessment.

In the previous two points, taxon-related information may be of a general nature, stopping at genus level or summarizing on the rank of order, family or even higher taxa (“plants, animals, and micro-organisms”). In this case, though, at least an indication of the associated species numbers is requested as a measure of biodiversity. In the following categories, the focus changes to species level or below (e.g. population level in research on genetic resources).

- **Locations or time regime of previous gatherings**: these questions also aim at the identification of hot spots and under-evaluated areas or periods (e.g. seasons), but aim rather at planning individual field research than evaluating the general status. The information retrieved has to be detailed enough to allow the user to assess who has previously collected in the region of interest, which organisms have been sampled or observed where and when, providing additional information on environmental conditions or other influences at the time of the event.

- **Changes over a time period**, usually within a given region. The expected result from such a query is a report on qualitative and quantitative changes observed and documented, e.g. for a reconstruction of population dynamics. In this context, a clear distinction between documented absence of a species and unavailability of data is of special importance to avoid misinterpretation.

- **Qualitative characteristics** may be used to further restrict the result returned for a given taxon or organism group. Endangered species of medical plants, protagonists in host-parasite relationships, type specimens, or fruit bearing specimens only. This kind of requests often is connected with the planning of field trips and must
therefore deliver location data and, in searches for certain life stages, calendar dates as accurately as possible.

In contrast to the previous categories and like the following one, the request for qualitative characteristics is expecting specimen information rather than the summaries on species level or above, which are associated with the more general requests.

- Additional specimen information. There are several kinds of additional data users want to attain referring to individual specimens as the smallest collection or observation units. This includes detailed data on the specimens in their role as organisms (gender, morphometrics, age, life stage, images, sounds), and data on the specimens as collection objects (collection holder, owner, accession number, mode of preservation, available derivates like skins, tissue, or seeds, conditions for loans). In research based on surveys, the inclusion of specimens provides a reliable scientific basis for observation records, and broadens the available pool of data. A combination of organism-related information from a sufficient number of specimens permits the reconstruction of ecological characteristics such as the phenology of a species at a given site.

- Linked information: Environmental parameters at the time of collection or observation are often an integral part of biological collection information. In contrast, taxon-related information is usually not included. However, such information is usually of high interest to users querying a collection database: the variability of a given species, its geographical distribution, tolerance to environmental factors, weediness, migration patterns, etc. Many users request the integration of taxonomic and collection information systems.

Data quality

Three points are top on the list of user’s wishes regarding the data they receive – up-to-date, high quality, complete. These demands appear as justified as they seem obvious. In environmental research, a constantly updated base of data, especially of observations, is most desirable, provided the data are also reliable. “Accurate with respect to geocode validation and taxonomical identification; complete on a variety of temporal and spatial scales” (Alkin 1998), such data could form a most useful base for further research as well as in more administrative functions.

Aspects of data reliability and “correctness” include a demand for markers of data quality and a proper treatment of possible inconsistencies arising from the merging of different datasets. Users want access to all information they need to assess the quality of a given dataset, while on the other hand they wish not to be swamped with details. A database system should offer the possibility to trace the history of a record, including information on the motive of the gathering (scientific research, environmental impact assessment, hobby collection), the collector’s name and field number, the identifier’s name, and information on procedures of data evaluation and revision. All of these make
Users and uses

it easier for the user to judge the reliability of the data (Diederich et al. 1998). In cases of merged datasets from different sources, inconsistencies or differences between the datasets must be detected and resolved (Alkin 1998), less the resulting dataset loose data integrity and be far less useful. Should discrepancies arising in the process be irresolvable, the least demand is for a flagging of suspicious data to make the user aware of the fact.

User interface

The wish list. The query interface should be interactive and graphic as possible. The output should on the one hand be as detailed as possible, consist of primary data and be supplied with an additional measure of quality. On the other hand the presentation of results should provide users with a choice regarding the level of detail. A choice between visualization modes should be available, additional processing tools should be integrated, and export datasets in different formats should be available for local use.

Users ask for a multi-facetted user interface, allowing selection between several possibilities of query definition, which must include list selection, hierarchical browsing, and free-text search. Additional visual help in the process of query formulation, like map visualisation or pictures of organisms, is highly welcome or even, as in geographical contexts, a prerequisite.

Three priorities rule the users’ wish list for data delivery: Direct, fast, and free of charge. Results from a query should be delivered by the queried access system in a compiled form, without compelling the user to collect results packages from different databases. Ideally, the information should be delivered electronically, promptly following sending off a query (via the Internet), and the service should be for free.

Where appropriate, the presentation of search results should have ample support through visualisations besides the supply of original data. Users ask for a highly interactive user interface, providing a choice of presentation modes. They should be able to switch between details and data quality information, summarise large datasets (e.g. provide indicative numbers of species for higher taxon map visualisations and still be able to access the single datasets on selection), and superimpose self-defined filters on a pre-selection (e.g., focus on ever narrowing time-frames). Additional wishes include statistical tools and data processing such as used in predictive mapping.

Initially, users want to be presented with a list or a map showing the distribution of collections and observations retrieved from a specific query. Since Geographic Information Systems (GIS) provide tools for the presentation of digital information in the form of maps to which layers of additional information may be superimposed, GIS applications are presented as examples for correlating collection and observation data with environmental variables. Many users look for uncomplicated ways to import query
results (lists or maps) into their own spreadsheet files or databases to combine them with their own GIS application.

The requirements of users regarding query interface and result presentation are often contradictory – intuitive simplicity with a huge range of functionality. Design and implementation of a functional user interface, therefore, will pose a major challenge. The interface will have to offer different levels of access (e.g. simple and expert search) always trying to morph into a self-explanatory interface that is characterised by minimal usage of technical language.
V. On the value of palaeontological collections

Dave Lazarus

The science of paleontology
Palaeontology is a highly interdisciplinary discipline, standing at the interface of the biological and physical sciences. The application of palaeontological knowledge can be broadly divided into 4 main areas:

- Understanding man’s place in nature
- A historical framework for understanding current biodiversity and biologic processes
- Basic chronologic and environmental data for geologic research
- A source of past “experiments” on the interaction of the biosphere and environment.

The first of these themes was also the first to be developed, and for most people is still the most significant aspect even today. Whether profound or ridiculous – from Darwin to “Jurassic Park” – fossils have fascinated and informed us of life’s vast history; and through extinction, of its inherent fragility. Within the biological sciences, palaeontology also has provided much of the historical background needed to understand the opportunistically evolved characteristics and functions of currently living organisms.

Complementing these biologic applications, palaeontology has also contributed extensively to the development of the earth sciences. Palaeontology has long been the backbone on which the divisions of geologic time have been built (only in the last century has it been possible to accurately assign absolute ages to the previously largely palaeontologically established geologic time intervals). Fossils, via their implied environmental tolerances, have also been a basic tool for reconstructing past physical conditions on the earth.

In more recent years these two aspects of palaeontology – the biologic and geologic – have increasingly begun to merge in highly integrated studies of the earth and its biota. In fields such as Global Change research or Earth System science, palaeontology is one of several vital disciplinary components. Understanding climate change, or the response of the earth’s biotic system to mass extinctions are examples of current significant contributions.

Palaeontology, with about 5,000 practitioners worldwide, has historically been divided into several sub-disciplines. Micropaleontology, the study of small (mostly protist-created) fossils is the largest branch and has concentrated on geologic applications such as biostratigraphy and palaeoenvironments. Other sub-disciplines include invertebrate
and vertebrate palaeontology, and palaeobotany. These disciplines have contributed more strongly to biology via the study of evolution.

**Types, sizes and costs of palaeontological collections**

There are no really reliable estimates for the number or size of palaeontological collections worldwide. A recent number (Allmon, 1997) of about 300 million specimens is probably a substantial underestimate. Most of this material is microfossil (where only bulk samples, not the thousands of individual specimens within them, are tallied) or invertebrate material. Although the largest single type of repository are museums, academic departments, government agencies, and industrial concerns (particularly oil companies) together hold nearly equivalent numbers of specimens. While museum collections are largely concentrated in North America and Europe, many large industrial and government collections are found in developing countries as well. These numbers do not include many other extensive collections of highly fossiliferous geologic materials, such as deep-sea sediment cores (which consist largely of microfossils), even though sediments in such core repositories are often sampled for palaeontological purposes. There are no statistics at all on how much these collections cost to maintain, but, based on a subjective estimate of the percentage of overall palaeontological (and support staff) manpower devoted to collection work, the cost, worldwide, is perhaps on the order of US$ 100 million per year. Whether this is a justifiable expense depends on many things, not least the sheer intellectual value these collections provide to humanity’s perception of nature. However, in a more technical, dollars and cents evaluation, one can well justify this kind of expenditure. For one, simply the replacement cost of such large collections – certainly in the US$ tens of billions range – would justify their continued maintenance, assuming that the collections still have some practical uses as well. And they indeed do.

**Major uses of palaeontological collections**

Palaeontological collections underpin most of the activities of palaeontological science. A small fraction of the world’s collections are primarily used to teach about the history of life on earth, either via museum displays or as special “teaching collections” in university departments. Most fossil material is however collected for research, initially by industry, government or academic palaeontologists. Older materials, often stored in museums, provide the primary data documentation for published research. These older materials are often also used again as the basis of new research, particularly in larger synthetic analyses of biostratigraphy, evolution and paleoenvironments. The type specimens in these collections are particularly important, as they provide the unique, irreplaceable physical reference standard for the calibration of taxonomic names. Surveys of museums in the USA indicate that re-examination of type material is the most common way palaeontological collections are used. Much of this work is for individual basic taxonomic studies, but these investigations are often also part of
broader research programs. The International Geological Correlation Program (IGCP) for example is the primary focus of the international earth science community’s effort to improve the basis of geological correlation worldwide. A substantial part of this work involves re-examining fossil type material in museum collections in order to better cross-correlate biostratigraphies from different parts of the globe.

**Three case studies**

To illustrate the use of palaeontological collections, one specific example each is given of their application to biologic, environmental, and industrial problems.

**Understanding evolution.** One of the more important developments in evolution research in the last 20 years has been an increased awareness of the major role of mass-extinctions in the development of life on earth. Major episodes of extinction have long been known by palaeontologists - indeed, the major divisions of the geologic time scale, defined more than 100 years ago, are mostly based on such extinction episodes. However, only in recent years has it been possible to move beyond the anecdotal and begin to quantitatively measure and understand both mass extinction events and the evolutionary recovery that occurred after them. This analytic work, which has relevance as well to understanding modern threats to biodiversity, has been based on major taxonomic syntheses, most importantly the multivolume “Treatise of Invertebrate Paleontology”. And these syntheses have been created, through the work of hundreds of specialists over a period of 30 years, by the standardisation of taxonomic concepts, in large part by comparison of tens of thousands of specimens held in museum collections.

**Understanding climate change.** Concern about global warming has recently elevated academic research on past climate changes into a high-priority program in the earth sciences. One of the most fundamental pieces of knowledge on what controls climate change to have been discovered in the last several decades comes from such work, specifically from the study of deep-sea sediments. In the late 1970’s, studies of these sediments proved that changes in the earth’s orbital parameters control the timing and relative severity of the earth’s dominant climate change signal – the ice ages. This work was in turned based on archives of deep-sea sediment cores and the numerous microfossils contained in them. These collections, often consisting of cores from thousands of locations around the world, have been built up over a period of many decades in the years after the Second World War at major marine science institutes, and form the basis even today for much climate change research. For example, one integral part of the global climate change research programme involves testing computer models of global climate against reconstructed actual past configurations of climate. These reconstructions (such as those of the CLIMAP project) are largely derived from microfossils, sampled from these extensive collections of deep-sea sediments.

**Finding oil.** The petroleum industry has long relied on microfossils in sediment cores obtained while drilling – for geologic age determination, and for determining the palaeoenvironmental conditions (hopefully favourable to oil formation) prevailing
when the sediment was originally deposited. Most companies have had large staffs of micropalaeontologists, and exploration wells usually had a micropalaeontologist “on-site” during drilling operations. In the last decades micropaleontology has been increasingly supplemented by geochemical and particularly geophysical techniques. However, as the search for oil focuses ever more on smaller, more palaeoenvironmentally determined deposits, there has been a renewed need for the unique data provided by microfossils. In many cases, this new work is being done at least partly by re-examining older collections of microfossil material obtained during drilling from previous decades. That this can bring economic benefits has been dramatically demonstrated recently in the North Sea. There, a team of micropalaeontologists, stimulated by academic advances in microfossil taxonomy, and by the development of new preparation and imaging techniques, decided to re-examine the industry’s extensive collections of microfossil material obtained from several decades of drilling for oil in the North Sea. Based on this re-examination of collection material, the scientists were able to improve the temporal resolution of the biostratigraphy used in the North Sea by a factor of 3, which was then in turn used to identify numerous new drilling targets that had not been visible to geochemical and geophysical methods alone. Although many of these targets have yet to be drilled, industry specialists estimate that up to 1 Billion barrels of new oil may have been located by this work.

Concluding remarks
The links between palaeontological collections and valuable contributions to science and society are not always visible to the public, and indeed are often made only through a long chain of intermediate steps. Yet the links are nonetheless extremely strong. Although it is possible – at least in Hollywood – to make money purely out of palaeontological imagery, real world scientists need real-world materials to support their work. Without palaeontological collections, palaeontology would be a crippled discipline, and the users of palaeontological data – geologists, biologists, climatologists, and the oil industry, among others – would be hurt as well. Whether for academic knowledge, or applied research into the environment, or in searching for natural resources, palaeontological collections are an essential, and highly cost effective research tool.
VI. CABRI: common access and quality control for biotechnological collection resources

Louis Réchaussat

Europe has a wide variety of biological resource centres acting as supply and service organisations for the scientific research community and biotech industries. CABRI, the Common Access to Biological Resources and Information service, offers access to some of these centres. This common interest gateway offers many advantages to both the centres and their user communities. Instead of having to scrutinise a large number of databases, catalogues and other sources of information, CABRI offers world-wide access to these databases and allows one to simultaneously check on the availability of a particular type of organism or genetic resource and to order the required items once located. Those include: Bacteria and Archaea, Fungi, Yeasts, Plasmids, Phages, Animal and Human Cells, DNA probes, Plant cells, and Plant viruses. CABRI also acts as a node in the BioCISE network, representing an example for a degree of standardization that other biological collections are still lacking.

The CABRI service has been built around quality and each member resource centre has therefore contributed to defining the set of technical specifications and procedures, which define the handling of each resource type. These procedures are primarily based upon the centres own procedures but they have been peer reviewed and approved before the catalogue has been mounted online. In this way users of CABRI are guaranteed the highest quality materials and services. It is our intention to expand the range of materials on offer by admitting other centres with acceptable quality standards.

Assuring quality

Each of the collections within CABRI has a wealth of specialist expertise that backs up the storage and delivery of cultures. A range of contract services is therefore also available and the centres can provide information on these on request. The adherence to agreed quality standards is assured through testing. The CABRI accreditation scheme for culture collections has been developed by reviewing and collating two principal sources: current methodologies used for accessioning material into member and other collections, and quality control standards already in use and appropriate for each biological resource category. The scheme has been checked by examining samples from each of the collections to compare methods and results.

In future, random checks will be carried out to test delivery performances, and other customer responses will be collected to measure customer satisfaction.

Quality Policy

It is the goal of the CABRI member collections to provide customers with a high standard of service. This includes biological resources, information and related services.
All of these items must meet the specifications laid down in the quality control guidelines, or the guidelines for catalogue production, which include the instructions for the submission of a flat file for indexing to CABRI.

A CABRI Technical Committee has been established to implement and enforce CABRI standards. This Committee also regularly audits both laboratory practices and delivery standards and thus ensures that the CABRI information procedures as well as content are kept current and abreast of the latest developments. CABRI aims to provide the most user-friendly approach to biological resourcing anywhere in the world.

**The CABRI quality guidelines**

The guidelines consist of four parts. (1) General guidelines for culture collections, (2) Request and handling deposits, (3) Maintenance of the deposits, and (4) Delivery to the customer. They cover procedures that as far as possible guarantee:

- adherence of CABRI to international, European or national regulations as well as to ethical and safety standards in the field of biotechnology;
- authenticity of biological materials;
- purity of cultures or absence of contaminants;
- quality-controlled processing of cultures;
- accuracy of data collected and supplied
- punctuality and adherence to delivery standards.

The guidelines are available over the World Wide Web (www.cabri.org). Part 1 provides a common view for all CABRI resources. Thereafter, information is provided at different levels of detail starting with the basic core procedures common to all collections, followed by specific protocols that provide basic information on standard techniques (as far as standardization has been achieved). Finally, there is detailed information on relevant local procedures that might be unique to a member collection. These local procedures have been independently refereed to ensure their scientific soundness.

**Networking biological resource centres**

The role of biological resource centres (BRC) is now recognised by national and international organisations as essential for the development of biotechnology. Networking the BRCs is becoming a priority for an appropriate access to biological and genetic resources. The quality control of biological material and data resources are the key issue of such a network.

CABRI is a first attempt and an original European contribution to the future "virtual" worldwide biological resource centre.
VII. Information Resources – The BioCISE Survey

Andrea Hahn

Scope and execution

Scope of the survey. The first general meeting of the Concerted Action decided to conduct a survey of collection databases by means of questionnaires. In the course of the project it became clear that the survey database of collections was to become a base for the envisioned European access system and for the project proposals to be submitted to the 5th Framework Programme. Thus it had to include all collections (i.e. also those not yet computerised at all) in the entire region (i.e. the countries covered by the 5th framework programme). The scope of the survey was extended accordingly, and collections were alerted to the fact that they could participate even without holding computerized databases.

Execution of the survey. Devising the questionnaires (in English, French, and German), implementation of on-line forms, address list creation, the distribution of the questionnaires, data entry in the collection database, and the ensuing communication with data providers to clarify data content presented a major workload for the project secretariat, which was aided by project partners who organized meetings or individually distributed questionnaires. In some cases (principally major natural history museums), a personal visit by the co-ordinator was necessary to clarify the aims of the project and to obtain the data. As far as possible, collections that were known to already participate in functioning European networks of collection information were exempted from the circulation of questionnaires. This refers chiefly to zoological gardens and to collections of plant and animal genetic resources.

Mailing of questionnaires started in February 1998, and continued until July 1999, with a total of over 2500 sets of questionnaires mailed directly to collection holders. Electronic communication was gaining importance throughout the project period; about 70% of the responding collections provided an email address, which greatly eased follow-up inquiries. The updated questionnaires continue to be available through the WWW site in English, French and German. Care was taken to obtain explicit consent of the data provider for the publication of information on the WWW – most collections gave permission to fully publish their data.

A total of 413 professional societies and organisations were contacted and asked to support the survey by notifying their membership, by linking to the project's site through the WWW, or by publishing the call for co-operation in their newsletter or journal. This represented part of a continuous process of alerting the communities of collection holders to the presence of the project, to inform about its aims, and to iron out misconceptions. Oral presentations, an extensive correspondence and numerous
personal contacts to promote the idea of the survey and raise awareness of its importance were other approaches taken by project members and the secretariat.

**Impediments.** The survey had to overcome several hurdles to get a response from collection holders. Many research collections, particularly those held in conjunction with ecological, chemical, or pharmaceutical research, are very difficult to identify and contact in the first place. In identified collection-holding organizations, the questionnaires often did not reach the appropriate person due to administrative obstacles. Also, potential respondents may not have realized the change in the scope of the survey mentioned above. The questionnaires were available only in English, French, and German, so that language problems may have de-motivated some potential respondents. Due to the fragmentation of the community of collection holders (see box on p. 3), the usefulness of a common access service is not necessarily obvious to all. Last not least, a general weariness of questionnaires had a negative impact.

![Figure 4: Collections in the BioCISE catalogue](image)

**Comprehensiveness.** Our current rough estimate is that the project has contacted (directly or indirectly) about 75% of the collections which eventually could become part of a European service, estimated at a total of above 4000 (only collections accessible to the public, not including commercial nurseries and pet breeders). In spite
of the difficulties mentioned above, the initial response rate of about 11% was increased
to a final result of 19%, i.e. 484 replies.

Figure 5: Geographical distribution of respondents

Figure 4 depicts the total representation of collections in the BioCISE collection
catalogue, i.e. including verified address data and collections in linked networks (see
Chapter X). Apparent differences in the number of collection-holding institutions per
country (fig. 5) in the BioCISE catalogue may not always reflect the actual situation.
Most non EU-members are not properly represented, because they have been included
at a rather late stage of the survey (Poland is an exception due to the workshop held
there). In some cases, like Iceland or Israel, the given numbers are probably
representative, while for example in the case of Italy, with its long tradition in natural history research, language problems may have caused a relatively low impact of the questionnaires (though partly compensated for by a national workshop in Italian language). The comparatively high response rate in Germany demonstrates the importance of the immediate contact, which was confirmed by the increase of responses attained by means of national workshops in Israel, Portugal, Italy and Poland. Institutes are much more likely to respond to a survey conducted on the national level, and non-respondents can be traced more easily when they are located in the same country (however, in some cases the background of a European initiative certainly helped to attract response). One of the implications was the incorporation of a system of National Nodes at the core of the concept for the Information Service (see Chapter XI).

![Figure 6: Collection units in respondents' databases](image)

**Results**

**Collections and units**

484 laboratories answered the questionnaires in detail. Of the respondents, 292 (60%) do maintain one or more biological collection databases. The total number of collection units (including survey records) catalogued in these 448 databases exceeds 42 million.
Fig. 6 gives an idea of the subject areas covered. The bulk of units are records in floristic and faunistic mapping projects – these are mostly electronic records to start with. In contrast, most natural history collections are only starting to register their objects in electronic inventories. Although 42 million units looks like an impressive number, it must be compared to the total number held.

The actual number of objects and observations held by European collections is unknown. The results of the survey, together with some other studies, can be used to attempt some rough estimates. There are approximately 620 zoological living collections in Europe, including zoological gardens, aquaria, and animal genetic resource collections. According to our results and the information gathered from ISIS (ISIS 1999) and the Global Zoo Directory (Swengal undated), a total of about 800,000 units seems a reasonable estimate. Our idea of the number of objects in natural history collections is less exact. Global holdings of natural history specimens have been estimated at 2.5 billion (Duckworth & al. 1993), a number that may be realistic when all private holdings are considered. European facilities should hold a substantial, if not the major part of these. The very large facilities often have only a rather vague idea about their holdings. According to Naumann & Greuter (1997), the estimated number of zoological samples in natural history collections in Germany ranges between 50 and 80 million for invertebrates alone; the natural history museums in London, Vienna, and Brussels combine holdings of more than 50 million invertebrate samples; for vertebrates, numbers are given with 2.7 million for the decentralised German collections, while in Britain and France the large taxonomic facilities (Paris: 1.5 Mio, London: 5.5 Mio specimens) are likely to own the bulk of the respective national holdings.

The number of zoological observation records is even harder to assess. A number of zoological observation databases cover from 1 to more than 2.5 million records each, giving an indication of the soaring total to be expected from all over Europe. Examples include the Austrian invertebrate survey database ZOBODAT at the Biologiezentrum des Oberösterreichischen Landesmuseums, Linz/Austria, the database on migratory birds at the Bird Migration Research Station, Choczewo/Poland, the data collections on ringing recoveries and vertebrate mappings at the Zoological Museum of the University of Copenhagen/Denmark, and the butterfly observation database of De Vlinderstichting - Dutch Butterfly Conservation Wageningen/The Netherlands. The central database of plant observations in Germany alone holds more than 15 million entries, similar databases exist in other European countries, feeding their results (as presence/absence data) into the Atlas Flora Europaea project located in Helsinki.

Collection categories

Figure 7 attempts a rough categorization of the collections accessible through the BioCISE catalogue (which includes data from linked networks and verified addresses of collection holders). The survey’s supplementary function with relation to existing
community networking activities and in preparation of inter-community collaboration must be considered when interpreting the results.

The coverage of herbaria should be more or less complete, thanks to the combination of survey data and linking through to the Index Herbariorum database (Holmgren & Holmgren 2000). Botanical garden coverage is also rather comprehensive, due to the excellent base offered by the Index compiled by Heywood et al. (1990). In contrast, only about half of the zoological gardens existing in Europe can be found in the collection catalogue, because those covered by the International Species Information System (ISIS 1999) are not included.

![Bar chart showing collection categories in the BioCISE collection catalogue](image)

**Figure 7: Collection categories in the BioCISE collection catalogue**

We assume that the representation of plant and animal genetic resources (agriculture, horticulture, silviculture, breeding and fisheries) is less complete. Commercial nurseries have been excluded, but the distinction between, say, a commercial nursery and a
horticultural research collection is arbitrary. For example, there are many commercial nurseries (as well as private persons) among the 425 collection holders recently listed in the British National Plant Collections Directory (Cook 2000) as owners of exemplary collections of garden plants. We also believe that many university departments hold reference collections, but we have had few contacts. However, joining forces with a well-organised international network (International Plant Genetic Resources Institute, IPGRI, see Chapter X) has increased the cover considerably, at least for plants.

![Figure 8: Database management system representation in the BioCISE survey](image)

Collection holding facilities, which would call themselves a “culture collection” or a “Natural history museum”, are well covered. However, the collection’s content does not differ if the holder is a private person or a university department, both of which would not be called so. Especially with culture collections, there is the additional complication of biotechnology companies holding collections that are not advertised. Nevertheless, we think that many collections that fulfil the criterion to be publicly accessible in some way are not yet represented in the catalogue. This is also true for the remaining two categories. Many ecological surveys and reference collections are hidden in government agencies and private consultancies, and many university departments hold a variety of
collections for teaching or reference purposes (e.g. oceanographic institutes holding microbial strains from deep-sea explorations, algae and collections of aquatic animals).

**Collection information systems**

The results of the survey confirm that an increasing number of collections are computerised, with a widely varying degree of sophistication. The high number of different software solutions in use for the capture of biological collection data – more than 60 different applications were named for the management of collections in just about 300 institutions - reflects the heterogeneity of the biological community, the fragmented institutional base, and the lack of commercial solutions. Only 12% of the databases were developed in some kind of collaboration with other collection holders, but about 27% of the institutes reported to have some kind of internal IT co-ordinating body. About two thirds of database owners reported their solution to be developed in-house. Stand-alone (more or less) relational systems are clearly favoured, as shown in fig. 8, with MS Access (44%) most prevalent, followed by dBase (25%). More basic systems include “flat” structures using word processing files and spreadsheet tables. Larger databases are using client-server applications, here the majority (70%) named Oracle as their database management system. Of the 390 databases for which the relevant survey question was answered, 62% were developed in-house, 24% by external service providers, the rest in co-operation with other collection holders.

Fig. 9 shows the main features of collection database systems in use. The presence or absence of these features was explicitly asked for in the database questionnaire, so that the numbers should be representative for the overall functionality of programs in use. As apparent from the lack of administrative features (label printing, loan and exchange management), more than half of the databases encountered are more or less restricted to the task of simple cataloguing of existing collections. Among the more elaborate systems, the automated printing of labels from the information system and the documentation of a specimen’s identification history were the most common features. Those with additional geographical information processing (GIS integration, mapping tools, point location) are fewer.

The questionnaires also asked for features missed by users, and these geo-referencing tools were topmost on the wish list; followed by on-line accessibility and interconnection with other databases. Complaints about missing administrative features (loan, exchange, and sales management) and label printing were less frequent.

**On-line accessibility.** Among the 483 collection-holding facilities responding to the questionnaire, 60 % had no on-line presence whatsoever before BioCISE published their data in the catalogue. Only 8 % offer unit-level information (i.e. data about individual specimens or observations of plants or animals in the field), while the remainder publish more or less detailed descriptive metadata on the content of their holdings. Since the survey was biased towards databased collections, and since the addresses of collections present on the web are much easier to access, the percentage of
collections offering unit data is likely to be lower if all collections are considered. In animal collections, general web representation is slightly more common, largely caused by the web presence of natural history museums.

Figure 9: Principal features of respondents’ collection databases

**Expertise**

The questionnaire contained a subset of questions regarding expertise and willingness to co-operate. More than 10% of the participants (52 institutions) have offered to share their professional expertise with others, e.g. by reviewing funding proposals or by offering advice or practical support in the building of collection information systems. Fields of expertise named include WWW design, database programming, and geographic information systems. This is an encouraging sign for cross-community collaboration and mutual support, which will be drawn upon in the design phase of the common collection information service (Chapter XI).
Conclusions

Despite the difficulties encountered, the execution of the survey has been a central element in the BioCISE resource identification process. Without the contribution of BioCISE’s own data, collaboration with networks (see Chapter X) would not have been possible. The survey has clearly demonstrated the importance of taking a community-oriented approach, be it oriented along thematic or national lines. It has led to the incorporation of a strong network of national nodes in the concept of the future service (see Chapter XI). The data gathered in the survey depict the heterogeneous state of biological collection databasing in Europe. The number of collection holders who were willing to fill in the rather tedious questionnaires are an encouraging sign for the strong interest in collaboration among European biological collections.
VIII. Networking collections: BioCISE national workshops

The BioCISE project in Portugal

Pedro Fernandes

The community of Portuguese collection curators (data providers in terms of the project) is very scattered. The initial survey identified about 40 collections; 30 more were added at a later time. Representatives from 50 collections attended a workshop held in Lisbon in February 1999. Due to BioCISE’s broad definition of “biological collection”, the participants were of differing orientations and most had never actually met. In the meeting very soon a sense of being on common ground for the discussion of shared problems was revealed. There was also agreement on the need of hosting a provisional forum for discussion and information interchange. This was implemented as a web site presenting the minutes of the workshop and additional contact data. These pages are now rather frequently visited (about 2,500 hits per month), especially after a button producing a query to the Berlin server selecting Portuguese entries was added. This denotes the increased value of the localization component offered by the BioCISE catalogue. The system is now available to the international scientific, educational and even commercial sectors from which the hits are quite noticeable. The scientific and technical value of the system as a locator of resources is a reality and the data consumer community is rapidly responding.

The experience shows that an information system that locates and adds value to biological collection data has the inherent benefit of reuniting an otherwise scattered community of data providers around the common problem of registering their data in an interoperable way. Taken to a European scale, capitalizing on this and similar experiences in other member countries, the system is bound to greatly enhance the value of collection data across the continent and worldwide. The implementation phase (see Chapter XI) can be viewed as an extension of this experiment, providing a much wider coverage.

Another obvious conclusion is that the existence of a national “node” greatly helps to integrate and reunite local efforts and represents a step towards the creation of a coherent user community. Users respect stability and reliability and these are better kept by an appropriate reference site.

The experience of the Instituto Gulbenkian de Ciência as the national node for EMBnet may serve as a model of how the existence of a national focal point can be essential to assemble resources and communities, harmonize methods and further interoperability. In the area of molecular bioinformatics, EMBnet regularly provides updated data, well installed and tested programs, help desk support, and training activities to about 35,000 users in 35 countries. The biological information service could pave the way for a European Network of Biological Collection Data Resources, if the effort is properly coordinated, financed and integrated in wider community efforts.
Networking biological databases: the Israeli experience

Linda Olsvig-Whittaker

The importance of the Israeli collections. The region now called Israel has been an area of interest for travellers and naturalists ever since the Middle Ages. Apart from its historical and cultural importance, Israel is geographically unique: the meeting point of three continents, representing the eastern arid extreme of the Mediterranean and its transition into the great central Asian steppes and deserts. Its flora and fauna are remarkably rich, with contributions from five biogeographic regions. Many species (wild wheat, barley, oats, legumes, etc.) are important as progenitors of domestic crops. The influx of many scientists among both visitors and Jewish immigrants since the mid-19th century provided a substantial foundation of information about the biota of the region. Today the biological collections of Israel are centralized, well organized, and often digitised, thus providing an easily accessible source of information on biota of the Eastern Mediterranean.

The BioCISE workshop in Israel: bringing together sub-disciplines. On 30 August 1988, the first public workshop in the BioCISE project was held at the Israel Academy of Sciences, attended by about 30 people invited from both users and producers of biological collection and survey databases in Israel. Brief presentations were made for the major databases in Israel, followed by discussion of the needs of database users. Until then, there had been only limited communication among users and producers of data, so the discussions were both lively and educational. We learned that users were most interested in electronically available primary data. In the absence of easily accessible data, the users were prone to substitute “expert opinion”.

The role of a future BioCISE from the Israeli viewpoint. Israel has a relatively computer-literate and sophisticated scientific community with professional interests and connections throughout the world, especially in the EU and North America. There is strong interest in networking of the type represented by BioCISE, and the capability to use it well if available. Environmental professionals need the ability to compare Israeli data to similar information in the region; plant breeders are concerned with the distribution of genetic resources, and taxonomists have the obvious need to know the distribution of study taxa and their relatives. As Israel continues to develop into a world leader in high tech and the information sciences, these needs are likely to grow even more acute.

Prospects for national organization of Israeli collections. The BioCISE workshop provided the stimulus for the main developers of biological databases to meet and discuss the possibilities for affiliation. One database group, that of Hebrew University, clearly provided the most sophisticated and interesting paradigm for connecting biological databases in Israel into a single functional network. The workshop revived a moribund agreement to join the major biological databases into a single-access national biological database. Since Israel is a party to the Biodiversity Convention, that would be under the umbrella of Israeli Clearinghouse Mechanism (CHM).
Italy: the struggle between Titian and the beetle

Pier-Luigi Nimis

The Italian BioCISE meeting took place at the Natural History Museum of Verona in June 1999, supported by a generous grant of the Region of Veneto. More than 100 participants presented a fair and decidedly unusual mix of zoologists, botanists, mycologists and palaeontologists. This diversity of academic disciplines was mirrored by a corresponding variety of institutions: universities, national, regional and very local museums, botanical gardens, various research centres, and holders of private collections. Especially important was the presence of Italian Regional authorities. Recent legislation has put the “Environmental and Cultural Heritage” of the country under one umbrella, which is supposed to cover a painting of Titian in the Accademia of Venice as well as a beetle preserved in the local museum. Obviously, attention will not primarily focus on the beetle. However, Italy – especially due to the activity of scholars in the Nineteenth century – hosts some of the most important biological collections in the World. Many nomenclatural types are preserved here, access to which is fundamental for evolutionary biology but still all too difficult for the international scientific community.

National legislation has largely assigned the responsibility for these matters to the 21 Italian Regions. Some of them (e.g. Emilia-Romagna, Lazio, Lombardia, Veneto, all represented at the meeting) have already started cataloguing projects. In some cases, however, systems used for decades to catalogue paintings, sculptures and monuments were proposed as general guidelines. It is excellent that Regional Authorities are willing to spend substantial amounts of money, but guidelines clearly have to be brought to a state adequate for computerization and networking of biological research collections.

The spectrum of initiatives presented at the meeting gave testimony to genuine efforts for computerising biological collections by the holders themselves, including many private or “local” collections, which are very important in Italy. While national – and especially some regional – authorities have correctly tried to impose some general standards (all of which require discussion), during the meeting we have enjoyed a nice example of Perfect Anarchy. All participants were proudly showing “their” database, “their” data structure, and “their” preferred system. Especially those with project funded by the Regions often concluded with something like “my database is the best”. Nevertheless, several talks showed interesting and important examples, e.g. the checklist of Italian Animals (Italy is the only country in the World to have completed such an Herculean task), the links between collection data and GIS systems, a project for computerizing living collections in Botanic gardens, and several beautiful ways to depict biological information is a user-friendly way. At least 50% of the talks referred to data already available on-line.

A very lively discussion followed the presentations. General agreement was reached that international standards are fundamental to avoid a duplication of efforts leading to a heap of isolated, albeit bold endeavours.
IX. Prototyping unit-level access

Pier-Luigi Nimis and Linda Olsvig-Whittaker

An essential feature of the planned meta-information system is its scalability. It should support access on all levels, from the most general (e.g. only the address and name) to the level of the individual unit. The first extreme has already been demonstrated (see Chapter X), but the possibility to access unit-level information must also be demonstrated from the outset of the Service’s implementation. Ideas from two initiatives originally planned as separate proposals (Mediterranean BioClISE, European Lichen Database) were presented as a project proposal able to present an advanced stage of the Service.

To be feasible at reasonable cost, the topic of the prototype has to be restricted both geographically and taxonomically, so that it can be developed in parallel with the larger metadata thesauri (see Chapter XI). It was decided to use the “Lichens of the Mediterranean” as the model group because a widely accepted standard taxonomy exists, circumventing the complication of the taxonomic thesaurus.

The Mediterranean region lends itself to demonstration of a unit-level access prototype, since it comprises a natural biological unit. Thus, it is logical to pool the fragmented national information sources to provide an overview of the biological status of the Mediterranean. The Mediterranean community of lichen specialists already forms an integrated network, so the question of a working geo-ecological thesaurus can be solved pragmatically (Grube & Nimis 1997). The proposed prototype is also restricted in its representation of user interests. Unit-level data on Mediterranean lichens would generate an information system highly relevant for environmental management (lichens are important indicators for air quality and other factors) as well as for systematics and ecology.

The prototype will concatenate and maintain species and location information contributed from existing biological collection and observational databases around the Mediterranean and relevant holdings elsewhere. The proposed database network would be based on existing digitised biological databases, which would be organized around national clearinghouse computers for the participating nations. The database centre would include a GIS-based Internet service, enabling graphic selection of regions of interest and simple online queries for species and areas. The characterization of units in the service would use the metadata attributes provided by the Service and would provide low-level metadata values to the system for the geo-ecological and taxonomic attributes.
X. The BioCISE collection catalogue

Anton Güntsch

Unit-level collection data (see box in Chapter III) are a primary source for the derivation of new knowledge in life sciences, and the unit itself provides the possibility to verify existing scientific statements (Cotterill 1999, Lane 1996). Unfortunately, a common access system for collection information on the World Wide Web is far from being implemented. Over 90% of Europe’s collections still do not publish unit-level data, in most cases because the collection is not electronically catalogued (see Chapter VII). The design of a unified interface for those collection information systems willing to publish their information through data networks is hindered by the diversity of information models and data standards in use.

Figure 10: User interface of the BioCISE collection catalogue
Several initiatives try to overcome these obstacles by standardising interfaces to collection databases and developing appropriate and convenient user interfaces. The European Natural History Specimen Information Network (ENHSIN 2000), the International Species Information System of zoological gardens (ISIS 1999), and the German Botanical Garden Information System SYSTAX (Hoppe et al. 1996), are examples of different approaches.

![Diagram showing collection-level vs. unit-level information](image)

*Figure 11: Collection-level vs. unit-level information*
Cataloguing biological collections

A first step on the way to a common access system for the hundreds of millions of collection objects held by European private and public institutions is the development of a catalogue providing collection information on various levels of abstraction.

The term collection information is here understood to cover information describing entire collections or sub-collections instead of single collection objects (units, fig. 11). Collection information covers categorizations and descriptive keywords (e.g. taxonomic and geographic), content and storage characteristics, as well as IPR statements and administrative properties. The Collection Description Working Group (Powell 1998) compiles a detailed list of meta information attributes to describe collections of any kind (including documents, art collections, etc.) based on the well-known Dublin Core (Anon. 1998).

The BioCISE collection catalogue (http://www.bgblm.fu-berlin.de/BioCISE/database/) is based on the results of the BioCISE survey (see Chapter VII). It realizes a simple variant of such an information system by offering data from the following information areas:

- Collection holders: Address, URL, and a free text self-portrayal
- Collections: Broad collection category and free text description
- Collection databases: Detailed description of systems in use including statements on information models, data quality, software features, status of data acquisition, and online accessibility
- Collection databasing expertise: Names and specialties of collection information system experts, contact information

A simple user interface (fig. 10) allows querying the system by country or broad collection category (plants, animals, fungi, fossils, micro-organisms, and other).

Prototype networking in the BioCISE collection catalogue

Biological collection catalogues already exist in some countries and for some thematic areas. These systems are community-driven, part of long-term projects, or supported by established organizations. A European information service should not duplicate their efforts, but rather help to enhance accessibility and feedback to them.

The BioCISE catalogue demonstrates a federation of catalogues which is based on a simple technique originating from the co-operation with the BIODIV project. (Güntsch & Vander Velde 1998). BIODIV – “Biodiversity Resources in Belgium” – is a Belgian initiative for cataloguing biodiversity resources on the World Wide Web. BIODIV cyclically transfers a file containing a core set of data fields necessary to establish the link from the BioCISE collection catalogue. In detail, these attributes are: Laboratory, department, and institution holding the collection, city and country, and collection categories derived from a defined set of keywords. As location, name and keywords are...
relatively stable, transfer intervals of 6 months are sufficient. The initial workload for the participating network consists of creating a query, which produces the transfer file; the workload for the updates is negligible.

To populate the BioCISE project database, the data in the transfer files are transformed to a SQL script, which maps the keywords used in the source (e.g. BIODIV) to those valid for the BioCISE collection catalogue, thus allowing for a common view on local and networked data. Whenever users of the BioCISE collection catalogue select a Belgian collection, they will be directed immediately to the BIODIV system to receive the latest information. The transformation process is automated by a Java application, which requires a minimum effort of reprogramming for the incorporation of new thematic or local networks. The regular incorporation of updated transfer files in the BioCISE catalogue takes a few minutes each, provided that the transfer files are “well-formed”. Fig. 12 summarizes the mechanism.

The collaboration between BIODIV and BioCISE served as a prototype for several similar co-operations, which were established in 1999 (table 1).

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Cover</th>
<th>Location</th>
<th>URL (http://)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIODIV</td>
<td>National</td>
<td>All collections</td>
<td>Meise</td>
<td><a href="http://www.br.fgov.be/biodiv/">www.br.fgov.be/biodiv/</a></td>
</tr>
<tr>
<td>NatureWeb</td>
<td>National</td>
<td>All collections</td>
<td>Dornbirn</td>
<td><a href="http://www.natureweb.at/">www.natureweb.at/</a></td>
</tr>
<tr>
<td>Polish Herbaria</td>
<td>National &amp; Thematic</td>
<td>Herbaria</td>
<td>Krakow</td>
<td><a href="http://www.ib-pan.krakow.pl/">www.ib-pan.krakow.pl/</a></td>
</tr>
<tr>
<td>Index Herbariorum</td>
<td>Thematic</td>
<td>Herbaria</td>
<td>New York</td>
<td><a href="http://www.nybg.org/bsci/ih/">www.nybg.org/bsci/ih/</a></td>
</tr>
<tr>
<td>CABRI</td>
<td>Thematic</td>
<td>Biotechnology</td>
<td>Genova</td>
<td><a href="http://www.cabi.org/">www.cabi.org/</a></td>
</tr>
<tr>
<td>IPGRI</td>
<td>Thematic</td>
<td>Plant genetic</td>
<td>Rome</td>
<td><a href="http://www.cgiar.org/ipgri/">www.cgiar.org/ipgri/</a></td>
</tr>
</tbody>
</table>

Table 1: Networks collaborating with the BioCISE collection catalogue

**BIODIV.** The national network “Biodiversity Resources in Belgium” has the aim to develop an inventory with information on Belgian institutions, specialists, collections, projects, and literature concerned with Biodiversity. The information is largely captured through online questionnaires.

**NatureWeb.** The Austrian “Networking Initiative for Natural History Collections and Observation Information Exchange” was initiated in 1998. One of the project’s aims is to compile an online collection catalogue, which currently includes information on 98 collections.
**Polish Herbaria.** The Institute of Botany of the Polish Academy of Science compiled an extensive overview of Polish herbaria. Collection-level information on herbaria are accessible through a simple clickable map. The State Committee for Scientific Research supported the development.

**Index Herbariorum.** The Herbaria of the World, Part I, last printed as the 9th edition, is a much relied upon standard reference for botanists, which is now also available as a World Wide Web database. It contains information on about 3000 public herbaria and nearly 9000 staff members world-wide. The abbreviation assigned to the collection is an acknowledged TDWG standard.

![Diagram](image)

*Figure 12: Co-operation of BIODIV and BioCISE*

**CABRI.** “Common Access to Biotechnological Resources and Information” was an EU financed project which has now been taken over by the participants. Its aim is to make catalogue databases of various organism types and genetic material accessible through a common interface. CABRI also developed standards and procedures to ensure the highest possible data quality (see Chapter VI). Although CABRI is focused on unit
related data, collection-level information on the participating collections is provided as a “side effect” and included into the BioCISE collection catalogue.

**IPGRI**: The Directory of Germplasm Collections held by the International Plant Genetic Resources Institute maintains an information system on ex situ germplasm collections, providing meta-information for more than 5 million objects worldwide.

Although the BioCISE collection catalogue was originally designed as a simple inventory of European collections it turned out to be a valuable tool especially to get updated contact information and collection descriptions. Nevertheless, this is a prototype. It does not fulfil several criteria, which are essential for a proper Information Service: scalability, wide range of query parameters, accuracy (and non-redundancy) of results, ease of maintenance, and feedback to participating networks. The willingness of networks to participate, however, was very encouraging – the networks in the Catalogue are indeed all that had been approached.
XI. Concepts for a European Portal to Biological Collections

Walter G. Berendsohn, Mark J. Costello, Chris Emblow, Anton Güntsch, Andrea Hahn, Jürgen Koenemann, Christoph Thomas, Neil Thomson and Richard White

The accessibility and thus use of biological collections would be significantly improved by a common “portal” through which information on the holdings of research institutes, museums, survey organizations etc. can be accessed. This portal may take the form of an Internet site where the reader can search for information based on biological names, taxonomic groups, habitat names (e.g. marshes), ecological relationships between species (e.g. parasitism), and geographic sources of specimens and observations of species. Unfortunately, such a search facility is far more complex than immediately apparent. One of the main obstacles is that collections use terminology going back 300 years and it is not feasible for most facilities to update their data comprehensively with changes in taxonomic nomenclature and geographic or political boundaries. However, we posit that this problem, as well as the difficulties created by the fragmentation of the collection community itself (see box on p. 3) can be overcome by a pragmatic and concerted effort of the interested parties.

Why should we try?

Channelling collection information into a common access system makes sense because

- the combination of information from thematically different data areas will enhance knowledge discovery and understanding
- users will be presented with a common interface covering a wide range of known and not yet known inquiries
- it will stimulate efforts to find agreement on good practice, standardization of data items and quality control
- a concerted approach will – to some extent - remove duplication of efforts; scarce technical resources can be put to use in a focused and collaborative way.

In the process of assessing user requirements and available resources on the provider’s side it became clear that such a service had to be very flexible; scalable on the collection owner’s (provider’s) side, simple in its internal mechanisms, broad in its cover of collections from different sub-disciplines, and providing a user interface adaptable to users’ needs. Realising that an extensive Europe-wide specimen (“unit”) - based access system is not yet within reach, but that user demand exists for concerted access to collection information right away, it was decided to focus on the creation of a collection-level information system as the kernel of “The BioCISE”. However, from the beginning a unit-level approach should be integrated (see Chapter IX). Recent initiatives, particularly the formation of – and EU support for – the European Natural History Specimen Information Network (ENHSIN 2000) are a promising proof of intent on both the information providers’ and the funding agencies’ side.
Achieving breadth and scalability: Meta-information to the rescue!

As has been shown in Chapter X, the physical as well as the information content of biological collections can be described using meta-information, i.e. information linked to sets of units. The interesting fact about such metadata items is that – with few exceptions – they are applicable to both, unit-level and collection level.

![Diagram showing a physical hierarchy of reference points for collection meta-information](image)

**Figure 13: Physical hierarchy of reference points for collection meta-information**

This presents a first important chance to achieve scalability of the system. Generally, the information referring to a high level in the physical hierarchy (fig. 13), an entire Natural History Museum for example, will be much less specific than that referring to a single unit. However, exceptions can be found e.g. in specialised collections, which consist of only one species, as well as in reference collections for a specific site.

The meta-information can be at any level of detail, from the very general (plants, Europe, 19th century) down to the very specific (the species, detailed gathering site, date). Between these extremes a fluent transition exists because most of the important data items describing collections and units are belonging to a (more or less) hierarchical classification system (e.g. geographical: country - department). This is the second important factor, which can contribute to achieving a scalable system. A system using meta-information organized into such a schema will allow processing of very detailed...
collection descriptions but still be able to provide information on collections that cannot
or do not want to supply this kind of detail. It will enable researchers to locate needed
information or materials and obtain them by conventional means, if necessary.
In addition, meta-information may serve as a valuable enrichment of unit data espe-
cially when statements about data quality and procedures are provided. However, at
present, the major advantage of metadata lies in the fact that they can provide informa-
tion about - and facilitate access to - units even where unit-level data are not available.

A knowledge based approach: Metadata ‘thesauri’
Identified priorities. The two most important data areas for meta-information about
collections are names of organism groups and named areas (the latter using geographi-
cal, biogeographical, geological, palaeontological or ecological terminology). Unfortu-
nately, these areas are not stable: terms may have different meanings depending on who
applied them and when (e.g. “Germany”, “Liliaceae”). Moreover, geocological and
taxonomic class names represent scientific concepts, thus parallel, partly overlapping
hierarchies may exist (e.g. generic and family delimitation in systematics).
Consequently, no single standard hierarchy exists for any of these information domains.
The development and application of integrated metadata ‘thesauri’ and classifications for
these data areas is a prerequisite for the functioning of an extendible collection informa-
tion service. To be clear: these will be pragmatic tools to facilitate access; they will not
attempt to redefine terms or derive new classifications. They will allow searches by key-
word and by following hierarchical links despite the underlying anarchy. Users must be
able to select and/or specify fuzzy concepts such as habitat boundaries (“Rainforest”) or
undefined geographic terms (“Central Europe”) that don’t map easily to the political
boundaries available in today’s gazetteers and geographical information systems. Taxo-
nomic concepts and names representing organism groups present similar problems of
parallel and partially overlapping hierarchies. To include them in data access interfaces,
information structures and methods which are able to accommodate and process such
complex inter-relations between individual metadata elements must be defined.
The thesaurus forms a common source for both indexing of collections and the design
of the portal’s user interface (e.g. implementation of a convenient taxonomic browser
instead of free text fields). The thesaurus to be constructed has to be powerful enough
to treat various semantic relations such as synonyms and hierarchies. It has to put a
special focus on taxonomic and geographic terminology to fulfil the special require-
ments of the biological community. National and thematic networks can derive key-
words to describe their collections from this thesaurus to provide a homogeneous data
source utilized by the central catalogue. In addition to data provided by the networks,
rule based technologies can be used to represent complex weighted relations among
thesaurus’ elements and to further enrich the set of keywords by deriving useful catego-
ries from collection descriptions. Rule based indexing will reduce the costs for the time
consuming process of human indexing.
**Geo-ecological thesaurus.** With respect to geo-ecological classifications of collections, two types of questions have to be addressed: The comparatively simple question of where a collection (institute) is located (e.g. “Is there a reference collection of microbial strains in town X?”) and the much more complicated question where specimens collected at a defined site can be found (e.g. “Who has holdings of specimens from the northern Mediterranean coast?” or “Where can I find organisms collected from Late Triassic St. Cassian Formation?”). The data for the former question can be captured from rather standardised address information and related to existing data collections on present administrative boundaries. The compilation of metadata for the second question is difficult for various reasons. In contrast to taxonomic data, no agreed systems of nomenclature for geographic, ecological, or palaeontological “areas” exist. The catalogue has to deal with very variable applications of terms (e.g. “St. Cassian”, “Sankt Cassian”, “Cassinian” in the above example), consider provisions for changes in the delimitation of areas in time (e.g. in the case of “Germany” and “Yugoslavia”), consider the problems of more or less linear references (“Mediterranean coast”, “River Guadalquivir”), vague delimitation (“northern Spain”) and the problems of scientific concepts represented by ecological and palaeontological terms. The identification of existing data collections (e.g. available gazetteers) and the contacting of geographical and ecological institutes will be prominent approaches. The decision on practical cut-off points for the hierarchical representation of the data (as opposed to a synonymised keyword list) is fundamental. Mapping the European languages into such a thesaurus (initially to be implemented largely in English) is a problem, which is tackled extensively in various projects; collaboration has to be sought here as well.

**Taxonomic thesaurus.** Rules of nomenclature (“Codes”, e.g. ICZN 1999, Greuter et al. 2000) exist for the area of taxonomic data (scientific names of organisms). However, synonyms, conceptual differences between applications of the same class name, as well as the problem of congruence of concepts with differing names also persist in taxonomy. This is the consequence of the naming system being a pragmatic approach to a rather complicated scientific problem: the classification of life on earth according to its natural evolution. Large data collections have already been identified which can be used to compile a catalogue of names down to genus level. This can be used as a backbone classification to which other terms can be associated. “Pseudotaxa”, i.e. higher-level class names for organisms that do not directly correspond to a taxonomic group (e.g. “Medicinal plants”, “Pests”, “Birds of prey”, “Trees”, “Microbes”) must also be treated.

**Other data areas.** Data areas that do not belong to one of the aforementioned categories must also be tackled. This covers, e.g., temporal aspects (date of collection event, dwelling time of the organisms in palaeontological contexts, etc.), representation of collection purposes (research material, archive of vouchers, exhibition, etc.) and preservation methods (often important for potential use of materials in analysis), among others. The compilation of such thesauri can be based on an analysis of the data provided by the BioCISE survey database and by looking at other similar data collections, to find out about terms applied by users in their queries, and terms used by institutes to describe their collections.
**Issues for informatics research**

Modelling and implementing catalogues and their utilization in user interfaces are general tasks of informatics research and application development. A partnership with an information technology provider should be sought to avoid duplication of efforts.

![Diagram](image)

*Figure 14: Hierarchical structure of a selection of collections at the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM)*

**A proposed representation of entire collections.** Since biological collections are generally organized hierarchically, it is straightforward to describe them as trees, the nodes representing sub collections (sets of units), and the connecting lines representing “is part of” relationships (fig. 13). Each node is linked to sets of attributes (e.g. taxonomic identification, ownership, locality) providing the sub-collections’ properties (fig. 14). These attributes are referenced in the metadata thesaurus and thus can be referred to other, more general or more specific terms. Properties can be further quantified by labelled links to express the fuzziness typical for collection descriptions (e.g. “mainly” Coleoptera, “some” Lepidoptera). Available metadata are often incomplete (for example if derived from questionnaires). Adding “dummy” nodes labelled “other” can indicate this (fig. 13).

Güntsch et al. (2000) demonstrated the use of this representation to formulate rules to derive complex concepts describing classes of collections (e.g. “Natural History Museum”, “Botanical Garden”). Future work will have to analyse the inheritance of properties within a tree representing both exact and fuzzy data to achieve a solid theoretical ground for the implementation of indexing modules, data capture tools, and user interfaces.
**Information model.** Documentation of heterogeneous information resources is a current topic e.g. in library science, museology, and environmental and medical informatics. The definition of meta-information attributes associated with collections of biological material must be based on an evaluation of existing and emerging metadata standards and work effected by current international working groups involved in the standardisation of access to museum resources. General metadata standards such as the Dublin Core definitions (Anon. 1998) must be incorporated, too.

![Diagram of attributes providing (sub-) collection properties]

*Figure 15: Sets of attributes providing (sub-) collection properties*

The representation of hierarchical structures, use of controlled vocabularies, incomplete or fuzzy data, and administrative metadata are problems extending far beyond the scope of biological collections. Similar problems are encountered in the domain of environmental information systems (e.g. in the context of the European Topic Centre on Catalogue of Data Sources, EEA 1999) or have indeed been identified as an unresolved problem for all collections (including museums, documents, archives, subject gateways, etc.) by the collection description working group of UKOLN (UK Office for Library and Information Networking; see Heaney 2000). An important component of the design process for the European Collection Information Service will be the provision of a theoretical model which helps to find a practical solution to the problems addressed.
The results influence the design of the metadatabase system directly, but also indirectly, by way of the content structure definition of the metadata catalogues.

**User interface.** There are some standard techniques to allow navigation in hierarchies. For example, in Yahoo!-style sites and most online shopping or product selection environments, there are lists of links, and following a link will result in a new page being displayed, possibly with further lists of links until the terminal nodes of the hierarchy are reached. These types of interfaces work well with small hierarchies containing well-known entries but are inappropriate for the task at hand. Direct manipulation interfaces with very fast updates of displays in reaction to user input are needed, with displays that allow users to see and select content rather than having to specify their needs formally. These interfaces have to be tailored to the particular information at hand. Furthermore, interfaces will need to be personalised based on user characteristics and current task. For example, graphic browsing of a hierarchy with Latin names of species may be appropriate for experts who are familiar with taxonomic trees and the specialist terms. Conversely, novices may require a different structure and different terminology likely to be supported by visual means such as images and symbols.

Novel interaction techniques have to be developed, because users must be able to browse and search in multiple, linked hierarchies without losing orientation, the system and its interfaces must represent missing, fuzzy, or incorrect (outdated) information, and users must be able to select and/or specify often fuzzy geocological concepts. Novel technical solutions will need to be developed to design and implement user interfaces that address these issues and at the same time support a large set of users with a diverse set of hardware (network bandwidth and processing speed) and software (browsers, Java, etc.) constraints.

**Knowledge processing.** In addition to evaluating distributed web sources, quite often new information can be extracted from the existing. This may concern the seemingly obvious one not thought to be necessary to put into (key)words: When searching for micro-organisms, looking into microbial collections seems the natural approach. But what is labelled as a “microbial collection”? Does a search for the keyword also answer with a cheese producer’s *Lactobacter* strains, or an algal reference culture collection? Are we talking prokaryotes, unicellular, or just “small”? The task of a knowledge-processing module is to apply man-made rules for such definitions to existing data and thereby, for example, generate new keywords. Through the possibilities of assigning different weightings, probabilities may be calculated: Asking for micro-organisms should deliver all bacterial collections, but also offer others lower down on the list.

One problem will be that, even with good thesauri available, most of the information used in the service will not adhere to a single structure. For example, collections may represent the data related to the gathering of a unit as a single field of text, or they may provide this in a highly structured, atomised form. Detailed knowledge about the hierarchical decomposition of such information will be very useful in the process of extracting information from text sources, especially if combined with the thesaurus.
The information provider's point of view

Synergetic effects: In the national meetings of collection holders organised or co-organised by BioCISE it became clear that the cross-subdiscipline approach was greeted with enthusiasm, but that communication even within sub-disciplines was generally wanting. Participants presented a multitude of isolated information systems developed or in preparation in their institutes (see Chapter VIII). The potential for synergetic effects was obvious. BioCISE was perceived as a possibility to focus resources and to overcome existing institutional rivalries or other political impediments, which may bear on the development of collaboration on the national level. In this context it is also important that almost all information providers would also be users of the service.

IPR and other legal concerns: Still existing impediments to the networking of unit-level data resources are the unresolved questions of IPR in databases and database networks (see Chapter II). Another important area of concern is the unresolved question of obligations imposed by the Convention on Biological Diversity (particularly in the case of living collections, which undoubtedly represent "genetic resources"). The use of metadata was greeted by some of the institutes with more advanced data holdings as a possibility to await a solution of these problems before going public with their data holdings, but at the same time being able to advertise their collection's scientific information content.

Promotion of collections. Collection holders also consider it essential to improve public understanding of the importance of natural science collections, and of the relationship between collection conservation efforts with the ability to manage, preserve, and interpret our natural heritage as well as the world we live in. One of the most important future tasks will be to show that the varieties of collections are actually daily used, and that they are of public interest. Hence the credit of the users will serve as motive to establish and to maintain a Collection Information System.

Basic organization of the Service

Data capture: As mentioned before, national meetings and spin-off activities led to the realization that the European service must rely on national networks or nodes to actually collect the information. For projects implementing such nodes national funding may be found – as has already happened in the case of Belgium, Austria, and Germany, and as it will hopefully be achieved in other European countries. The political argument in favour of such funding lies in the obligations incurred by government in the context of international conventions, the contribution to an over-all European research infrastructure, and synergetic effects to be achieved on all levels. However, governments and funding bodies do seem to believe that unit-level data capture is within the scope of individual institute's activities, so that successful attempts to find extra financing for such endeavours will probably remain the exception. For collection-level information, the BioCISE survey database can serve as an initial dataset to build upon.
**National nodes.** Regarding the access to technical innovation, Europe has grown closer over the past years. In the implementation, and above all, population, of biological collection information systems, however, it may still take a long time to reach a common level. This does not demean the value of the collections themselves but their chances to be recognised, preserved, and properly valued. The aim of the National Node set-up is to provide equal chances for all of Europe’s biological collections to be represented in a common information system. The metadata access system will help to override the inherent inequality resulting from still widely differing access to information systems, databasing expertise and staff supply. In some countries (e.g. Belgium), national information systems are already well developed, needing little initial input to adapt to contribute to a European system. In other cases providing the basic means (software and training) for the initial set-up as a National Node will be needed.

![Diagram](image)

*Figure 16: Information flow from individual collections to Central Node*

The National Nodes are to co-ordinate networking at the national level, but to provide a common access point and a conceptional framework for the European system, a central system (“Central Node”) is needed (which, however, has to be kept at a minimum to ensure sustainability). To make the information gathered by the National Nodes useful for the European system, standard protocols for metadata content and usage must be applied. This is most efficiently possible by providing the National Nodes with software which communicates with the central system’s thesauri etc. and which can be accessed by the central system for information retrieval. The software has to be designed in an easy-to-extend fashion, using widely available software as its base, so that National Nodes can easily extend their activities.

The National Nodes should be hosted by organisations committed to research or information provision on a national level, e.g. institutes being part of national
academies, or the academies themselves, organizations representing the national clearing house mechanism, or other agencies with an obligation of fostering the collection and maintenance of biological data. It is not mandatory, but of advantage, if they are information providers and/or users themselves, thus being beneficiaries of the improved information access the European Service will provide. They will take on the responsibility of hosting the national meta-database and setting up a website.

Figure 17: Feedback of information from Central Node towards National Nodes

Communication with the Central Node includes keeping up an interface to supply the necessary core data to perform searches over all connected databases (fig. 15), and to feedback enhanced quality data to the national system (fig. 16). The enhanced data result from the Central Node's application of knowledge processing and advanced indexing tools and the addition of information from different sources. Links to already operating thematic networks will be established, as demonstrated by the BioCISE project.

**Automatic extraction of keywords.** The efficiency of information access will to a large extent depend on the efficiency of linking metadata items in the thesauri to the collections or sub-collections. The central node will receive information from the national nodes in the form of core attributes and free text descriptions. One of the core attributes is the URL of the collection's website (if any). This can be used to implement web-robot techniques to analyse the websites and automatically extract keywords. For this, an advanced multilingual free text indexing tool has to be developed. Rule based techniques will then be used to work on both, the information retrieved from national nodes and that from Websites, to generate value-added indices. These can be used directly in the central user interface, but should also be communicated back as added value to the national nodes.
**User access.** Users will be able to choose between access via the European Service, National nodes, thematic networks as well as access to individual collections (fig. 17). Where available, the Service and the national nodes will offer these choices to the user.

![Diagram of access to collection information](image)

*Figure 18: Access to collection information*

**Sustainability.** Prospects for long term funding for the operation of a Biological Collection Information Service on the international scale are rather bleak. On the other hand, among participants in the various workshops and other providers and users interviewed, consensus prevailed that the service itself should be free of charge for the user. Such a facility certainly represents an infrastructure providing access to specimen data, which are the results of more than two centuries of mostly government-financed scientific endeavour. Furthermore, many of the users would be information providers as well. Taking over of marginal costs was still seen as a possibility, as was charging commercial users of the service. However, the general feeling was that the administrative and IPR-related complications inherent to that approach by far outweighed the income generated. For example, several of the thematic networks which are accessible through BioCISE provided their data with the explicit condition that they should be available free of charge.
Again, flexibility of the system is decisive: it has to be able to survive changes in its administrative set-up and location as well as changes in on the information providers side. The system software should require minimum maintenance; once installed, the maintenance and provision of base data will lie in the responsibility of national and thematic networks driven by user and provider communities. The BioCISE project has already established links to 5 national and thematic networks, and the set-up of other National Nodes has been organised in view of pending project proposals. Established contacts to European agencies and the CHM, and liaison with the Consortium of Large-scale European Taxonomic Facilities (CETAFF) further broaden the base for attempts to achieve long-term sustainability.

Summary
Collection information serves as a basis for biodiversity research, which is part of the obligations incurred by government in the context of international conventions. Interconnecting such databases of a variety of nations and scientific research topics is a justified cause to involve international research funding agencies. A biological collection information service integrating the full spectrum of resources would play an important role in the Global Biodiversity Information Facility (GBIF 2000) envisioned by the OECD Megascience Forum Working Group on Biological Informatics (Edwards 1999).

The creation of a Biological Collection Information Service in Europe is believed to be a feasible goal if an approach relying on scalable metadata provision through a network of national nodes is used. The major initial contributions an international implementation project would have to make are the following:

- Design of a model for the accommodation of metadata items and their relationships in extendible structured thesauri to be used in information access and indexing.
- Data acquisition for thesauri for taxonomic, geo-ecological and other collection-related data areas from existing sources.
- Design and implementation of a rule-processing software for the automated generation of keywords.
- Setup and development of national nodes; provision of software, where necessary.
- Standardization of core data for the harmonization of data flows between the nodes (content and protocol standardization, e.g. by means of XML data definitions).
- Technical implementation of the information service integrating the metadata model, thesauri and provided data into a user-friendly access system.
- Prototypical integration of individual collection and observation records.

A project period of 3 years with adequate resources is deemed necessary to achieve a functioning service. Questions of sustainability of a collection information service, the adequate consideration of intellectual property rights, and approaches towards data quality standards will have to undergo continued discussion.
Acknowledgements

First of all I wish to thank those who have taken the trouble to fill in the questionnaires, who attended our meetings, and participated in discussions or contributed in other ways to develop the concept of a Biological Collection Information Service.

Special thanks goes to Carlos Martínez, the Scientific Officer at DG XII responsible for the project, who helped us in various ways.

Without the steady and active support of Sylvia Steinmann we would have succumbed to the administrative overhead a project like BioCISE brings with it. In this context, the considerable support of the BGBM’s administrative staff (Regina Katlewski, Gisela Niemann-Wildung, Marlies Youssef and Egon Brüning) should not go unnoticed.

In the preparation of this publication, I wish to acknowledge the help of Andrea Hahn and Anton Gütsch gave me as well as the proofreading done by Bettina Pfeiffer-Berendsohn and several of the project’s members. Michael Rodewald and Ingo Haas assisted in the preparation of the graphics. W. Winderlich of the Photosatzwerkstatt should be mentioned for the patience he mustered.

BioCISE Resource Identification was a Concerted Action Project funded by the European Commission, DG XII, in the Biotechnology Programme, contract no. BIO4-97-2309.
References


ENHSIN 2000: ENHSIN – European Natural History Specimen Information Network. [http://www.nhm.ac.uk/science/rco/enhsin/]


ISIS 1999: International Species Information System. – [http://www.isis.org/]


Lane, M. A. 1998: Biological informatics – weaving a web of wealth. – Canberra.


Williams, N. 1997: Drug firms back move to link databases. – Science 277: 902.
The acquisition, cultivation, preservation, and storage of objects in biological collections is an integral part of biological research in many sub-disciplines. Field and research notes and specimen labels contain valuable and detailed data and the object itself can be a physical resource for research and industry.

This publication reports on the concluded concerted action project BioCISE, which set out to identify and analyse biological collection information and its environs with the aim to prepare a sound collaborative technical and structural base for a Biological Collection Information Service in Europe and a strategy for its implementation.