New approaches to creating global species databases in entomology

Malcolm J. Scoble (7)

Abstract

Global species databases are, broadly speaking, computerised taxonomic catalogues. Databases have, however, the capacity to be more extensive than catalogues, and they are much more effectively searched. They can also be networked. It is increasingly evident that the kind of information inherent in traditional taxonomic catalogues is of value beyond the systematics community. In particular, it forms the basis for such products as life-lists, biodiversity surveys and inventories, which are needed to meet certain requirements under the Convention on Biological Diversity. The main difference between creating global species databases for insects and most other groups of organisms is that of the size of the task. In this paper I give an example of the steps in creating one such database (on geometrid moths) and the hardcopy catalogue that was derived from it. Although a great deal of effort was required to complete the database in a timely fashion, such large compilations are quite possible given appropriate facilities and the right people. Key features in the production of the work, both material and in terms of human effort, will be discussed. Attempts are now being made by the University of Essex and The Natural History Museum, London, to build a Versatile Interactive Archive Document System (VIADOCS). The project will use modified optical character recognition (OCR) software to convert species data on index card archives. A demonstration system utilizing a particular Lepidoptera index-card archive (on Pyraloidea) will be developed and evaluated against current manual conversion methods. The data will be made accessible, interactively, over the Internet. The aim of producing this system is to provide a means of making extensive quantities of data available, which are currently trapped in typed and hand-written archives. All these efforts should be seen in the broader context of computerising biological data typically associated with biological collections.

Keywords: taxonomic information, catalogues, archival data.

Introduction

Entomological catalogues have been printed for a wide variety of taxa. They are durable products based on data that form the bedrock of museum biodiversity information. They have been built from details gleaned from card indexes, from descriptions in publications, from centralized compilations, such as the *Zoological Record*, and from information in collections of specimens themselves. Traditionally they have been published on paper. But while the value of that medium arguably has still an important role in conveying fundamental information, the advent of desktop computers has changed our view of these traditional products for two main reasons. First, computers can be loaded with software for constructing relational databases thus enabling data to be searched flexibly. Second, the Internet allows (actually or potentially) both shared access to live data-sets and, increasingly, added value through interoperability across distributed data-sets.

Taxonomists are now open to new conceptual levels. They realize that while the net is a wonderful, if flawed, mechanism for communication its value rests on the quality of the input. They know that this situation will persist. But they appreciate also that flexible searching methods across any single database provides a new dimension to using data, for now questions can be asked that are impossible of a simple index in a printed book. Furthermore, they see

(7)   Department of Entomology, The Natural History Museum, Cromwell Road, London,UK, SW7 5BD,
       E-mail: m.scoble@nhm.ac.uk

that taxonomic names can be used, in effect, as hypertext links. At present these links mainly allow access to species web-pages, but interoperable systems are evolving to provide instantaneous access to data across multiple servers. This, in turn, creates the possibility of new opportunities for analysis of these data.

Suddenly taxonomic catalogues, which were never noted, nor indeed ever compiled, for their capacity to excite, have gained, potentially, a wider role. They act as a source for biodiversity inventories, required by most governments as national contributions towards the Convention on Biological Diversity. Futhermore, lists of names can act as thesauri permitting access to a growing body of information in global digital space.

The present account is written with this wider perception of the value of databases in mind, although it deals specifically with examining the process of data input. Two projects are featured, both involving lepidopteran data stored on archives of index cards. In the first of these, an account is given of the logistics of how a particular database (and the catalogue derived from it) was produced for a group of moths, the resources on which it was based, and the value of teamwork in its production. In the second I outline a recent collaborative initiative to computerise, semi-automatically, taxonomic data stored in index card archives using another large group of moths as a demonstrator of the system. In both cases, the broader aim has been to make taxonomic information trapped in museum archives more widely available.

A global taxonomic facility for geometrid moths

The first of these databases is to a species-rich family of Lepidoptera – the Geometridae, which are the moths with 'looper' caterpillars. This project involved the production of a printed catalogue (Parsons *et al.*, *in* Scoble, 1999) and a computerised database to all available species- and genus-group names of Geometridae.

The Geometridae were selected for several reasons. Geometrid moths are currently being studied in The Natural History Museum, London (BMNH). The resources for this group of insects within the institution are the most comprehensive in the world, which makes the location the most appropriate one in which to compile the work. The museum houses the most extensive collection of these insects taxonomically and geographically, and it has a comprehensive library with access to virtually all original descriptions for the species. Associated with the collection is a card index to all genus- and species-group names of the family within which is contained the kind of data that typically goes into creating taxonomic catalogues. Furthermore, the Geometridae are species-rich, so the group provided a good basis for testing our ability to create large databases to a timetable. From the wider standpoint, lepidopteran caterpillars are primary consumers of plants and, as a species-rich group, have a significant, although unmeasured, impact in ecosystems. This is true particularly of tropical ecosystems, which is where most geometrid species occur.

The participants preferred to refer to this project as a *global taxonomic facility* rather than simply a catalogue. In this way the dual catalogue/database output of the work was emphasized, as was the fact that fundamental taxonomic information on a world basis was included, and that the compilation would act as a tool for collections management and research. Moreover, given the flexibility of modern relational databases, the capacity for expansion with the addition of images and specimen records was made possible.

The card archive

The basic information on each of the 35 000 cards in the archive, two examples of which are shown in Figure 1 on the following page. The cards are divided first by subfamily. Within each subfamily they are indexed alphabetically by genus, within each genus alphabetically by species, and within each species alphabetically by subspecies. Junior synonyms and homonyms are arranged alphabetically following the senior name. Many of the genus/species combinations are unpublished having been incorporated during detailed curation of the collection. As no indication was given when new combinations or synonymys were made, it

was considered wiser not to flag any of them as new since, without searching all the literature subsequent to the original descriptions, we could not be sure that they were, in fact, unpublished.

The index also contains many infrasubspecific names. These were not added to the database on a comprehensive basis since they fall outside the regulations of the *International Code of Zoological Nomenclature* and are of very limited value biologically. Time was better spent concentrating on the available names.

It should be emphasized, therefore, that the catalogue was not constructed directly from original descriptions, for much of the information needed already existed in the card archive. The geometrid moth archive forms a part of a much larger set of cards, which covers all the Lepidoptera, and which were compiled over many decades. For the Geometridae, L.B. Prout (1864-1943), who made such an impact on the classification of the family, was responsible for the early development of the archive. This work was continued, expanded and refined by D.S. Fletcher and his collaborators who curated the archive alongside the collection of an estimated one million specimens of the family. So although the current collaborators have made the archival data more accessible by means of the published catalogue, much of the original information was incorporated in the index by the earlier curators. As with so much of taxonomy, our knowledge has developed accumulatively.

| Figure 1. | Two index cards from the geometrid moth archive at The Natural History Museum, London |
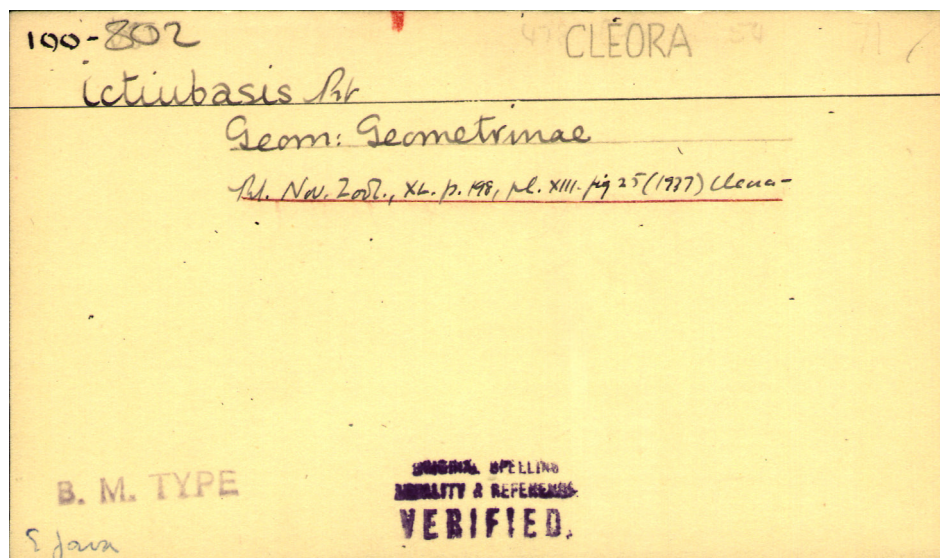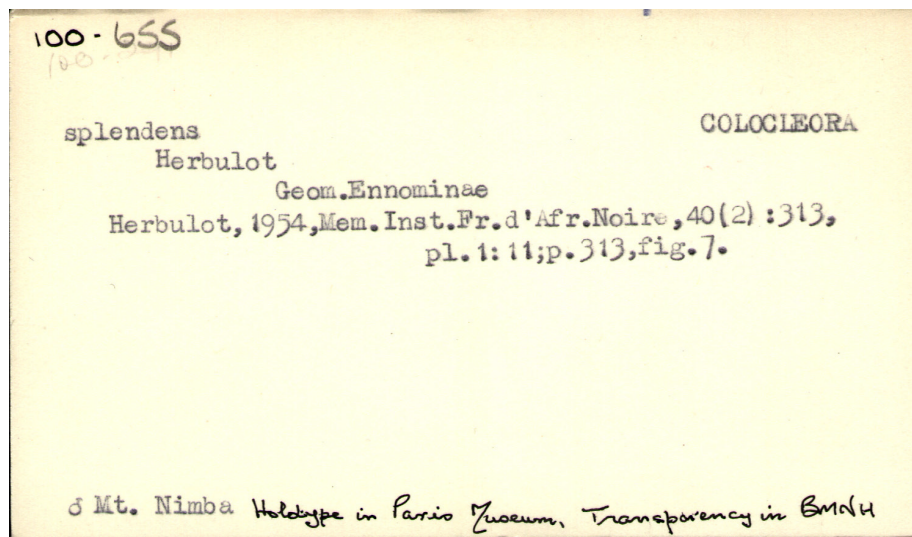
Nevertheless, the project did not involve simply converting the card archive to a database. Many errors were corrected and additional information was added, particularly by reference to original species-group descriptions, nearly all of which were examined during the course of the study. Deciding on whether names were subspecific (and thus available) or infrasubspecific (and thus unavailable) was a particularly time-consuming process. Efforts were made to resolve as many as possible of the nomenclatural issues that became evident during the project. Numerous taxonomic changes that were not in the index, but that were present in the literature subsequent to the original description (i.e., in revisions), were incorporated in the database. A considerable research effort was therefore required.

Purpose

The purpose of the project was to produce both a computerised database for research and collections management and a hardcopy version for publication. Although databases are more flexible for searching different fields, paper versions of catalogues provide users with certain advantages. The pages of books (personal copies at least) can be annotated, and volumes are rather more easily moved around collections even than laptop computers. The further advantage of a hardcopy publication is that it provides a fixed reference – a brick in the taxonomic literature. With the growing impact of dynamic information on the Internet, it is by no means clear just what the future of hardcopy publications for large taxonomic data-sets will be, but we are not quite at the stage where a complete switch from static hardcopy to dynamic electronic information has been made. The dual media of hardcopy and database seem highly complementary each having their own virtues.

Large taxonomic catalogues are compiled typically by individuals, or perhaps two collaborators, often over a long period whereas a further aim of the present project was to see if we were able to produce a catalogue to the Geometridae in a relatively short period. Speed should not, ideally, be the primary criterion in assembling taxonomic works, but taxonomists are, reasonably enough, under pressure to produce their work in a timely fashion, notably for demands to inventory the species of the world.

The database

The fields in the database (Table 1) include information typically found in entomological taxonomic catalogues. For nearly all species-group names we were able to give original and current genus, the author and date and the reference and page. Infrasubspecific names were omitted and several other names in the card archive were excluded because they were deemed likely to be manuscript names, no literature source having been traced. Besides the basic fields, we also added type status, the depository of the primary type (or of syntypes if no primary type had been designated) and the type locality. Foodplant information was provided where we were able to gain access to it, the most comprehensive source being *HOSTS - a database of World Lepidoptera hostplants* (Robinson *et al.*, in prep, - NHM internal research document - see also http://www.nhm.ac.uk/entomology/hostplants).

Table 1.	Fields in Geometridae database

Subfamily
Genus
Original genus
Species
Subspecies [if appropriate]
infrasubspecific [some]
Author
Year of publication
Journal title [in standardized abbreviated form]
Journal series [if existing]
Journal part [if existing]
Volume
Page + plate + figure
Comments [for notes on nomenclatural issues]
Junior synonym? [Yes or No]
Synonym of genus [if Yes]
Synonym of species [if Yes]
Synonym of subspecies [if Yes]
Original reference checked? [Yes or No]
Type depository
Type status [holotype, lectotype, neotype, syntype(s)] and sex
Type locality: country
Type locality: state
Type locality [place]
Zoogeographical area
Drawer number [in BMNH collection]

Project logistics

The process of keyboarding the data, checking original references to species-group descriptions, resolving nomenclatural issues, editing the final manuscript derived from the original database, and dealing with our publishers took about five years with the catalogue being published in 1999. This can be accounted for as follows:

- 20 months to keyboard the data in the archive
- 16 months to check references to original descriptions
- 24 months to edit and publish

These figures do not represent person-months, but rather the approximate periods taken for aspects of the work to be completed. The compilation took about four and a half years of staff time, which may be divided thus:

- 3 years keyboarding and checking references by one person working full-time
- advisers (on nomenclatural issues) and people checking references: 6 months
- editing: 1 year.

The size of the task may be understood by appreciating that for each of the 35 000 species-group names, there were data on an index card to be converted, and that maximally 26 fields were filled on the database. The six month period of work estimated under 'advisers' (see immediately above), involved the sum of efforts by at least seven people. It is impossible to give an accurate measure of this time since much of the work took place by the research assistant, whose time was entirely dedicated to the project, discussing problems with colleagues and eliciting advice from them on demand.

The museum team included a supervisor responsible for the project funding, management and editing; a research assistant working full-time for three years; and six others who variously provided advice, helped with checking original references and also in editing. Those involved in the project were from both research and collections management divisions of the Entomology department at the BMNH.

In addition to the in-house staff, we received a great deal of help from key colleagues at other museums, notably the American Museum of Entomology, New York, where a specialist checked entries to all the New World taxa and provided a great deal of extra advice. Nomenclatural problems on certain European species were discussed by e-mail with colleagues at other museums.

Funds to cover the costs of the full-time research assistant for three years were granted by The Leverhulme Trust, and this support enabled the project to be carried out.

Limiting factors

For a project of this size inevitably there were limiting factors. The first was the validity of the taxonomic system. A great deal has been achieved in constructing a classification of the Geometridae, but much has yet to be done. Many of the genera still require revision and we can expect there to be numerous new species combinations, species and genus synonymies, and descriptions of new species.

Second, we were intent on making the catalogue available within a reasonable period of time. Resolving complex issues of nomenclature can take a considerable amount of work, so a balance had to be achieved between resolution and completion. Finally, in a data-set of this size, there remain errors.

Conclusions

Despite the demands of producing global species databases, production of the geometrid moth catalogue demonstrated that compilations of this magnitude really can be achieved in a timely way, provided that access is available to the basic resources of information, staff time and team-work. The card archives, collections and library resources at the BMNH made the project one that could be carried out on a single site with little in the way of *material* outside sources (i.e., discounting advice from willing colleagues). Undoubtedly without the card archive the task would have been much more time-consuming: information would have to have been extracted from original descriptions and modified by assessing changes in revisions published subsequently.

Nevertheless, for most taxa there are likely to be published checklists or catalogues at least for parts of the groups. There are, furthermore, many card archives in museums around the world, so, with better networking, compilations of a similar size should not be impossible to achieve given the will, sufficient finance and time. A much larger initiative, Fauna Europaeae, which is already in progress, and which aims by means of pan-European collaboration to list all species of animals in Europe, is a good example of how a managed approach can make an impact in recording a significant number of species across a large geographical area. Certainly if we are to help overcome the taxonomic impediment in a reasonably timely way, taxonomic information is likely to have to be made more accessible more quickly.

Although it is obvious that catalogues can be as good only as the existing taxonomy from which they are derived, the great advantage of computer databases is that they are relatively easily to update. So although hardcopy publications start to become outdated as soon as they are published, modification and addition can still continue electronically. The combination of hardcopy with computerised databases provides therefore a complementary system combining the permanence of publication with a flexible (computerised) means of updating. It is desirable, therefore, that at some stage the database will be made accessible on the Internet so that static and dynamic versions of the data are accessible. Since we have excellent working relationships with our publishers, who have taken a considerable commercial risk in publishing a two-volume work of over 1000 pages, Internet access will need to be agreed with them.

Versatile Interactive Archive Document Conversion System (VIADOCS)

Compiling the geometrid database was achieved by manually keyboarding the card archive. With IT collaborators from the University of Essex, we are currently exploring ways of converting card archives from digital images using modified OCR software. The exemplar for the project is an archive to moths of the superfamily Pyraloidea, which is roughly equivalent in size to that of the Geometridae. The archive of approximately 30 000 cards has been digitised using a SEAC BANCHE RDS 3000 document reader scanner. There are a series of problems to be addressed if the data are to be transferred within the card archive to a database containing several fields, but good quality images have already been derived from the scanner. Major aims of the work will be to recognize the characters and parse the data into fields for incorporation to a database. The type-written script on the cards takes the form of courier font from old mechanical typewriters, for, as with the geometrid archive, that to the pyraloids was compiled prior either to the use of electric typewriters or to modern printers linked to computers.

Although OCR of type-written script might be expected to be a relatively straightforward process, problems arise with touching characters. Off-the-shelf OCR software fails to give an accurate conversion, and modifications are required. The problem of character recognition is at its most extreme for dealing with handwriting on the cards. In some cases, characters are difficult to interpret even by a human observer. Nevertheless, we expect to make progress with interpreting some of the handwritten emendations to the archive.

Dictionaries (partial or relatively complete) for comparing text are already available from various sources. For example, a set of authors names for many Lepidoptera is available from the HOSTPLANTS database (see above). A computerised list of many journal titles in standard abbreviated form were compiled during the geometrid database project and during work on providing digital access to the generic names of moths of the world (B.R. Pitkin & P. Jenkins, *in prep.*). These dictionaries should allow inconsistencies in archive citations to be resolved and enable the standardization and recognition of textual elements.

Perfect conversion is not expected and many corrections will need to be made manually. But through a series of iterations we aim both to deal with editorial matters and to improve the archive conversion system. Furthermore, we intend to upgrade the data in the pyralid moth card archive where possible by checking original references to key sections, by incorporating element of the latest research on the group, and by resolving nomenclatural issues. By providing Internet access to the images of the cards, pyralid moth taxonomists will be in a position to feed information back to the compilers enabling constant upgrading to occur.

The archive to the Pyralidae (including Pyraliformes and Crambiformes) was targeted for the VIADOCS project both because it provides a substantial demonstrator for the system and also because it is the main lepidopteran taxon the names of which have yet to be digitised. Two other large families of Lepidoptera have been catalogued recently, the Geometridae, as described above, and the Noctuidae (Poole, 1989). The fact that the names to parts or the whole of many other families are already digitised means that once the Pyralidae are catalogued, we shall be close to having a digitised databank to all the names of the order Lepidoptera, which is itself, a major insect group.

Given that the broader aim of VIADOCS is to create a system by which card archives in general can be converted, it is hoped that the system will be of value to paper archives in natural history collections and also to those institutions housing cultural artefacts.

Wider issues

The data being gathered for catalogues/databases to the groups of Lepidoptera discussed above, are of the kind that will eventually populate broader initiatives to catalogue the living resources of the world. Notable is 'Species 2000' (Bisby, 1994; http://www.sp2000.org), which provides interoperability software to query various databases on different servers (see also

http://litchi.biol.soton.ac.uk and http://www.systematics.reading.ac.uk/spice/ for associated technical developments).

Other programmes are in various stages of development, but while the development of interoperability software is clearly being made increasingly effective, the limiting factor appears to be that of getting the Global Species Databases to populate the systems. Gaining funding to compile these databases is more difficult than obtaining support for innovative software development.

Epilogue

Global Species Databases should be perceived as one part of a biological information system constructed from resources existing within collections-based institutions. The other major component of such a system is specimen data. Ways of digitising text and images of specimens are in a state of evolution, and management systems for prioritizing the selection of data to be digitised are unsophisticated. Although total digitisation is a worthy eventual goal, the huge number of specimens makes prioritization essential. One approach to providing access to specimen (unit) information is via descriptions of collections (metadata), as suggested by the BioCISE project (the Biological Collections Information Service for Europe, http://www.bgbm.fu-berlin.de/biocise/). Descriptions of collections are easier to compile in a more comprehensive way and in a much shorter time than unit data, and access to information about collections provides users with a means of knowing where to seek the latter. For a comprehensive model addressing questions of access to biological collections see Berendsohn et al. (1999). Access to specimen data is, however, being addressed through ENHSIN (the European Natural History Specimen Information Network, http://www.nhm.ac.uk/science/rco/enhsin/

The digitisation of museum resources, whether from archival data on species of the kind described above, or from the labels of specimens, already shows signs of adding a further dimension to natural history museums and other physical 'memory institutions' (Dempsey, http://www.ariadne.ac.uk/issue22/dempsey). This new dimension is not simply the sum of information made accessible on the Internet. Rather, it is the creation of new and complementary places existing in digital space. The coexistence of the physical and the digital, so elegantly described by Dempsey (loc. cit.), provides an enormous opportunity for natural history museums if they will reaffirm their position as custodians (collectively) of the best stored sample we have of biodiversity, while at the same time creating dynamic access to their very considerable resources.

Acknowledgements

This paper is based on a paper delivered at the XXI International Congress of Entomology held in Iguassu, Brazil in August 2000. I am very grateful to Hannu Saarenmaa and the late Ebbe Nielsen for organizing the relevant symposium. This manuscript was being finalized around the time of Ebbe Nielsen's untimely death and becomes, thus, a very small memorial to his enormous achievements.

References

Berendsohn, W.G., Anagnostopoulos, A., Hagedorn, G., Jakupovic, J., Nimis, P.L., Valdés, B., Güntsch, A., Pankhurst, R.J., White, R.J., 1999. A comprehensive reference model for biological collections and surveys. *Taxon* 48: 511-562.

Bisby, F.A., 1994. Global master species databases and biodiversity. *Biology International* 29: 33-40

Poole, R.W., 1989. Noctuidae. Parts 1-3, xii+1314 pp., *in* Heppner, J.B. (ed.), *Lepidopterorum Catalogus 118 (New Series)*, E.J. Brill, Leiden; Flora & Flora Publications, Gainesville.

Scoble, M.J. (Ed.), 1999. A taxonomic catalogue to the Geometridae of the world (Insecta: Lepidoptera). 2 vols. CSIRO Publications.

An information infrastructure for German insect collections including multimedia and GIS tools

Karl-Heinz Lampe [8] and Klaus Riede [8]

Abstract

The German Ministry of Science and Education has launched the EDIS-project (Entomological Data and Information System) to digitise and harmonise the rich, but scattered entomological collections housed at various German institutions. The concept is illustrated for the DORSA-subproject, which will integrate German Orthoptera collections within one 'Virtual Museum', accessible by an internet-based user interface. DORSA is a network project, connecting expertise in data-basing, collection management, systematics, geographical information systems and neuroinformatics. The core of DORSA is a specimen-based database of important grasshoppers and crickets in German collections. The taxonomic backbone will be the 'Orthoptera Species File', a global species register already available on the World Wide Web. DORSA integrates specimen-based pictures and sound recordings. The species-specific songs will be used as a knowledge base for the development of song recognition algorithms and bio-acoustic 'Rapid assessment tools'. In addition, all localities will be geo-referenced, resulting in a huge data-set of point data, which can be intersected with other GIS-maps (e.g. on rainforest distribution). A customised Java-tool allows geographic depiction and retrieval of taxonomic data.

Keywords: Orthoptera, species database, specimen database, collection management, GIS, song recognition

Introduction

The number of insect species on earth and their actual extinction rates is a matter of speculation for several years already, but exact numbers are still not available (Stork et al. 1997). This is mainly due to the lack of an efficient information infrastructure. Complete registers of valid taxa only exist for few insect groups, and much of the information stored within museum collections is not readily accessible, because most of it is not digitised. In many institutions it is common notion that 'computerisation' of collections might be possible for the vertebrate departments, but that the mission will be impossible for invertebrate sections due to the overwhelming number of species and specimens, compared to the lack of staff and money. Further difficulties for insect collection managers are the high number of undetermined specimens or undescribed (new) species ('taxonomic impediment').

Nevertheless, an impressive demonstration of feasibility has recently been accomplished by the Insect@thon project (Komen & Marais 2000), where around 21 000 insect inventory records of the Namibian National Museum have been entered by 92 schoolkids on 1 weekend. The project managers stress the need for digital access to the hand-written catalogues of huge first-world collections such as the Natural History Museum London, with 65 million insect specimens: 'We estimate that some 70% of these collections originate in the third world. Inasmuch, we strongly believe that first-world museums are urgently accountable to us…' (http://www.natmus.cul.na/biodive/insectresults.html). Other initiatives such as CONABIO or InBio show that biodiversity-information management in developing countries is much more advanced than in many developed countries. Especially European institutions seem to have severe difficulties in adopting the new information technologies.

---

(8)   Zoologisches Forschungsinstitut und Museum Alexander Koenig (ZFMK), Adenauerallee 150-164, D-51113 Bonn, Germany, E-Mail: k.lampe.zfmk@uni-bonn.de; k.riede.zfmk@uni-bonn.de.

This asymmetry was part of the rationale for the establishment of the Global Biodiversity Information Facility (GBIF) (Edwards et al. 2000). As part of this initiative, the German Ministry of Science and Education (BMBF) has launched the EDIS-project (Entomological Data Information System) to digitise and harmonise the rich, but scattered entomological collections housed at various German institutions within one specimen-based collection database (http://www.insects-online.de/).

The EDIS-project consists of separate and self-dependent subprojects including 'global species registers' (cf. GLOBIS, this volume) and collection databases of specimens with a connection to geographical information systems (GIS). Further projects are rapid assessment tools for automatic identification at the molecular level, by optical analysis of bee's wing venation and sound recognition of crickets and grasshoppers (http://www.insects-online.de). The respective databases will be pooled by an Oracle-based database, which will provide Internet-access (SYSTAX: see http://www.biologie.uni-ulm.de/systax).

This paper deals with the subproject 'German Orthoptera collections' (DORSA, according to the German acronym). Germany harbours rich collections including material from tropical countries dating back to the 19th century (Figure 1). Most of this material is not data-based. The DORSA project will digitise specimen-specific information for crickets and grasshoppers and integrate them within one 'Virtual Museum', accessible by an internet-based user interface.

Figure 1.	German research collections with estimated numbers of Orthoptera holotypes. Most of the material is not data-based

Orthoptera collections in Germany
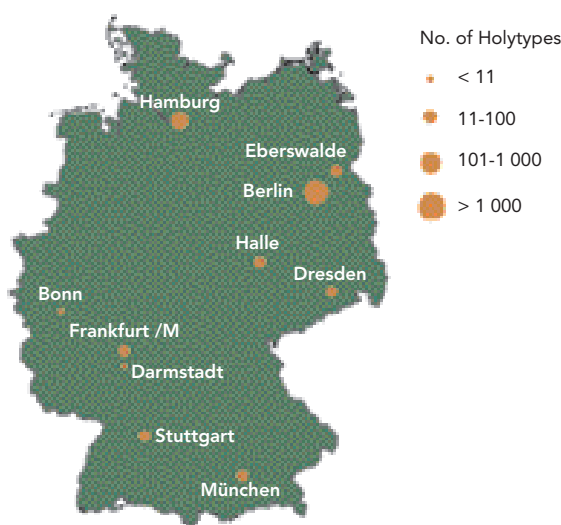
Databasing and collection management

In practice the overall efficiency of data-basing the inventory of traditional entomological collections depends on two factors: suitable software, and management measures to ensure the highest possible data quality already during the input process. All data entry is based on determined specimens (systematic label information), but 'determination' is not limited to the species level - it can be any higher taxon (e.g. 'Acrididae' for an unknown grasshopper). In our institution we introduced a lock-step programme for ergonomic and efficient data entry consisting of the following steps:

1. primary data capture of systematic label information
2. validity check of the systematic information against a current catalogue
3. set up of a collection based catalogue of taxa (consisting of the updated systematic information including synonyms, hierarchy of taxa, authors, year, etc.)

4. secondary data capture of sampling information (such as locations, collectors, determinators etc.)
5. validity check of the geographic information against a current gazetteer (by adding geographical attributes such as latitude, longitude and in a hierarchy such as province/area, state, country, continent/ocean and a link to the zoogeographical region)
6. set up of a collection based catalogue & completion of lists (e.g. collectors, determinators)
7. final data entry of existing specimens into the database

Each step of the procedure is clearly separated from the next. Therefore, everyone involved in this work can clearly see his own area of responsibility as well as the progress of his work. A nice side effect is the allocation of the various jobs involved where they are most welcome. Someone who is interested in working with catalogues by looking for further geographical or systematic information can work easily together with someone who is more interested in doing an accurate data capture. That means a single person is no longer forced to complete all the various tasks alone. Yet another advantage is that any of these procedures can be stopped or interrupted and even taken over by a third party with very little extra effort.

Systematic backbone

The 'Orthoptera Species File' (OSF) will be used as a taxonomic backbone. The OSF is an electronic catalogue of named grasshoppers and crickets, including pictures and sounds, and is available on the WorldWide Web (OSF: Otte and Naskrecki 2000). The OSF is one of the few fully functional global species registers. Queries allow searching for valid names as well as synonyms, their taxonomic reference and the depository place of the holotype. This means that one can ask for all holotypes for a certain museum, as known from the type descriptions. DORSA will realise the next step: a link between species names and existing specimens in German collections. In the future, a simple mouse click on a taxon should produce a list of specimens, together with a map of point data.

Geographical Information System (GIS)

All data sets refer to a locality. To connect them to a GIS, they have to be geo-referenced by their geographic coordinates. In the ideal case, they have been determined exactly, for example by 'Global Positioning System' (GPS). In most cases, localities are given as geographic names, and coordinates have to be determined afterwards by searching gazetteers or atlasses. In many cases, locality information is vague, which requires coding of imprecision. Geo-referencing is a time-consuming process, but can be speeded up during the project by building up a thesaurus of specific collection sites and major collector's routes.

Once geo-referenced coordinates can easily be exported into a GIS and plotted as a distribution map or analysed by geo-processing. For example, the intersection with borders of states or provinces produces calculated species lists for administrative units. These lists are useful for conservationists and decision makers, but their maintenance is a time-consuming task. There are numerous additional applications for biodiversity maps in GIS-format, among them: Comparison of maps from different sources and different projections; Calculation of biodiversity hotspots; Intersection with other GIS-layers such as eco-regions, land use, population pressure or climate change predictions, to name just a few.

Figure 2 shows an example for the potential of GIS analysis of collection data. Point data based on 3.578 data-sets of a ZFMK Homoptera collection (with 7.969 specimens) are plotted on a satellite view of the world (Figure 2). The original data for each specimen can be requested by simple mouse-click at the locality point.

Figure 2.   World map with localities of the ZFMK Homoptera collection, superimposed on a satellite picture
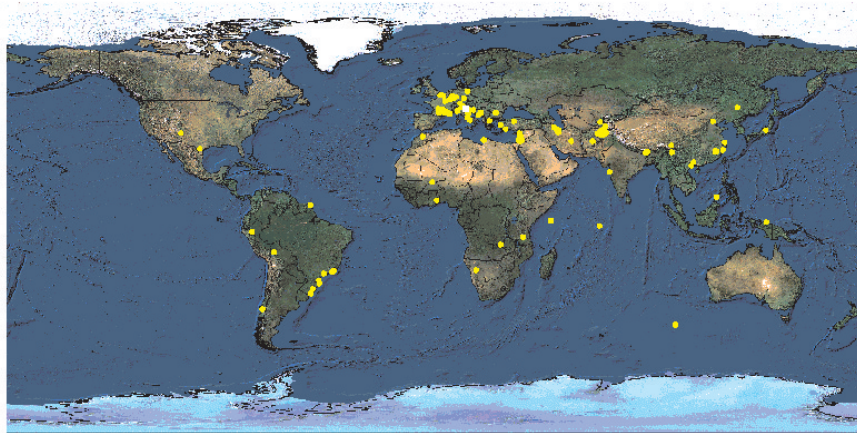

Figure 3.   Ecoregions of Central Asia and collection sites (red dots: origin of type material)

With a GIS one can easily zoom into the world map and enlarge special areas and/or change the background information. Figure 3 shows parts of Afghanistan and Pakistan structured by ecological attributes such as the vegetation type. Dots represents collection sites. GIS analysis can be greatly enhanced by pooling data sets from different collections (see Spatial Analyst, this volume).

In spite of these advantages, biologists are still reluctant to use GIS tools. One of the reasons is the user-unfriendlness, such as lack of a standardised query language. The introduction of desktop GIS improved this slightly, but there are still many problems. Even simple questions such as: 'How many species occur within a certain area?' require a number of complex operations. Therefore, a new Graphical User Interface (GUI) has been developed to publish interactive maps on the World Wide Web in cooperation with the Geography Department, Bonn (Fitzke & Friebe 1999). It consists of a platform independent, Java- based information desk for a combined display of geographic and database information (see a pilot version used by the 'Global Register of migratory species': http://www.groms.de). The information desk will be adapted for DORSA to allow queries such as: Show distribution map of a given species! Show maps of non-described species within a higher systematic taxon such as a family! Show

material collected during the Kaiserin-Augusta-Fluss-Expedition (Papua-New Guinea 1920). Or even strange sounding queries such as: Show all species singing with a carrier frequency lower than 2 kHz! Results will be depicted as point data maps.

Multimedia and neuroinformatics

A special feature of crickets and grasshoppers are their species-specific songs which can be used for rapid species identification (cf. Riede 1993). Germany has a long tradition in Orthopteran bio-acoustics and neuroethology, which resulted in several important sound archives ('phonotheks') at universities and in private hands (for reviews on Orthopteran communication see Ragge 1998, Riede 1998). At present, the majority of these data is stored on analog media such as magnetic tapes, film or videos. These recordings are now digitised and stored in a standardised format (e.g. wav-files for acoustic data). File names together with data on the original analog data source (tape number, deposit, etc.), recordist and localities are entered into DORSA. Voucher specimens exist for some, but not all of the recordings. Sound files could be used either to analyse inter- and intraspecific song variation, or to provide input for automatic song recognition algorithms. Neural networks for song classification and identification at the species level are presently developed in close cooperation with the Neuroinformatics Department at Ulm University.

The aim is a bioacoustic 'rapid assessment tool' for non-invasive mapping and identification of Orthoptera in the field.

Perspectives

DORSA will be accessible by Internet from any part of the world as one 'Virtual Museum Collection', which is important for potential users in species-rich, but resource-poor developing countries with incipient biodiversity infrastructure. The 'Virtual Museum Collection' will help to improve classical taxonomic work such as description of new taxa. Further important functions of this database such as distribution map generation and retrieval of pictures and songs from determined specimens, will be especially useful for ecologists, conservationists and applied entomologists.

At present, types and paratypes are entered into DORSA. The process of databasing is an excellent opportunity for type revision, lectotype designations and eventually repatriation of secondary types. In the case of type loss, re-collection of topotypes should be initiated. Topotypes should also be collected and designated for the country representing the 'terra typica', in particular for endemic species. The database will reveal information about historic species distributions which can be compared with actual distributions. Especially in rainforests, such a comparison will form the base for estimates of the actual conservation status and insect extinction rates, which at present are not even informed guesses.

Acknowledgements

References

Edwards, J.L., Lane, M.A. & Nielsen, E.S. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289, 2312–2314.

Fitzke, J. und Müller, M.(2000): Simple Features in der Praxis:

OpenGIS-Strukturen in Auskunftssystemen für Umwelt- und Naturschutz. In: Cremers, A.B. und Greve,K.(Eds.): Umweltinformation für Planung, Politik und Öffentlichkeit (Environmental Information for Planning, Politics and the Public)., Beiträge zum 14. Internationalen Symposium 'Informatik für den Umweltschutz', 'Umweltinformatik aktuell', Band 26, Marburg 2000, pp. 321–329.

Komen, J., Marais, E. 2000. The Insect@thon project. http://www.natmus.cul.na/biodive/insectresults.html

Otte, D., Naskrecki, P. 2000. Orthoptera Species Online. http://viceroy.eeb.uconn.edu/Orthoptera.

Ragge, D.R., Reynolds, W.J. 1998. The songs of the grasshoppers and crickets of Western Europe, Harley Books in association with the Natural History Museum, London.
Riede, K. 1993. Monitoring biodiversity: Analysis of Amazonian rainforest sounds.
Ambio 22, 546–548.

Riede, K. 1998. Acoustic monitoring of Orthoptera and its potential for conservation. Journal of Insect Conservation 2, 217–223.

Stork, N. E., Adis, J., Didham, R. K. (Eds.) 1997. Canopy arthropods. London: Chapman & Hall.

Engineering considerations for biodiversity software

Robert A. Morris ([9]), Mathew Passell ([9]), Jun Wan ([9]), Robert D. Stevenson ([10]) and William Haber ([11])

Abstract

We describe how object oriented design and programming, together with object databases, support applications that require diversity in their data structures reflecting diversity in the description of the data. A three tier web-based architecture permits flexible multiple views on the data. With working instances of an electronic field guide and of applications that federate distributed biodiversity data, we show how XML and object technologies can ease the burden of biologists who must prepare descriptive and diagnostic data for inclusion in web-accessible database.

Introduction

Biodiversity software encompasses a wide range of applications including the maintenance of specimen records, analysis of phylogenies, examination of biogeographic relationships, and recording of ecological observations (see Biosis, 2000, Geocities, 2000, and Lampinen, 2000 for lists of free and commercial packages). The rapid development in power and sophistication of biodiversity software in the last 15 years mirrors progress in the broader software industry. All software is moving from single user platforms to Web based tools, from text to graphical user interfaces and from custom made software built from the ground up to applications layered on top of commercial or open source components.

Biodiversity software concentrates at the species level because species are the leaves of the hierarchical grouping system called the 'taxonomic tree' used by scientists to classify life forms. Despite limitations to the species concept (Futuyma, 1998), the classification system is well established in the scientific community, having been used since Carl Linnaeus invented it over 250 years ago. Therefore is it not surprising that a significant portion of biodiversity software deals with the management of taxonomic information. Examples include programs that help manage collections of specimens such as Biota (Colwell, 1996) and Biolink (Shattuck and Fitzsimmons, 2000) from which taxonomists describe and name species, construct keys to differentiate species (Dallwitz, 2000) or document the tree itself (Maddison, 2000). Common to all of these programs and to many other efforts to share biodiversity information is the species page (also called the homepage, species summary or species treatment) in which biologists present basic information about a species.

Below we describe a biodiversity software application we are developing called the Electronic Field Guide (EFG). The EFG has elements common to biodiversity software listed above but different goals (see below). There are many efforts within scientific and environmental communities to establish standards and way of linking biodiversity information across multiple databases. Among these efforts are those of the Taxonomic Database Working Group http://www.tdwg.org/, the U.S. National Science Foundation's Long Term Ecological Research Sites http://www.lternet.edu/informatics/, those of governmental agencies such as the National Biological Information Infrastructure, NBII http://www.nbii.gov/ and http://www.nbii.gov/home/partner/bioeco/index.html; the Global Biodiversity Informatics Facility (GBIF) http://www.gbif.org/); Cornell's Laboratory of Ornithology Citizen Science projects http://birds.cornell.edu/citsci/, the Nature Mapping Program http://

(9) University of Massachusetts, Department of Computer Science, 100 Morrissey Boulevard, Boston MA 02125, USA, ram@cs.umb.edu
(10) University of Massachusetts, Department of Biology, 100 Morrissey Boulevard, Boston MA 02125, USA
(11) Missouri Botanical Garden, P.O. Box 299, St. Louis, Missouri 63166, USA

www.fish.washington.edu/naturemapping/; and Discover Life in America http://www.discoverlife.org. In the second part of the paper we discuss how XML (the eXtensible Markup Language) now being adopted in the Web community as a replacement to HTML is likely to help with the federation of information across multiple heterogeneous data sources.

The UMASS-Boston Electronic Field Guide Project, UMB-EFG (Stevenson and Morris, 2000) provides a web-accessible distributed object-oriented database for the identification of biological specimens from field observations. The data, including both taxonomic and environmental or ecological data, will aid in identification by building a context for each observation. As observation data accumulates, larger-scale ecological studies can be carried out using the data. UMB-EFG is being constructed and populated under The EFG project has recently been expanded to encompass investigation of a number of issues and solutions under discussion in the eco-informatics community, including the use of XML for federation of data from disparate distributed database, as well as for more common tasks such as data exchange and system configuration. This paper describes our engineering approach to the building of these systems and reports on their current status.

Field guides and descriptive data: object-oriented representations

Central to the majority of field studies in biology is the correct scientific identification of species. People learn from others or use field guides if possible, but for most groups identification is accomplished with keys constructed by specialists with knowledge and experience in a taxonomic group. The process begins with the collection of specimens in the field. Taxonomists and systematists then prepare, study and catalogue the specimens, usually in academic institutions or natural history museums. Finally, written descriptions are published and a name given to each new species discovered. Some taxonomists work only in the laboratory, obtaining the specimens from field collectors.

Most paper field guides are devoted either to a specific collection of taxa, e.g. birds, trees, wild-flowers, etc., of a specific geographic region. In most cases, a field guide user who is interested in ecological interactions will require several different field guides. For example, a guide to butterflies may have some narrative identifying the host and nectar plants of a particular butterfly species, but it would give little help in identifying that plant (which might in turn help the reader identify the butterfly). An electronic field guide on contemporary computers (or the web) can easily hold data on a wide variety of taxa, but a data representation issue immediately arises: descriptive characters appropriate to one group of taxa may have little to do with a radically different group. For example, plants have no wing spots since they have no wings, and butterflies have no leaves to be characterized as simple or compound. Therefore, to represent both organisms in a traditional relational database one must either accept large sparse tables or manage very complex joins on the ecological relationships. Essentially, the diversity of life is not amenable to a single description.

Object-oriented Database Management Systems (OODBMSs) (King, 1997) are a solution to the problem of representation of biodiversity, because their data is self-describing. (This is also true of XML, of which more later). Saarenmaa et al. (1995) have observed in detail how object-oriented techniques in general, and OODBMS in particular are well suited to taxonomic databases. However, they reported that with technologies then available, they were unable to usefully create taxa as classes, rather than instances of a single class. In a database with a huge number of taxa, class loading overhead would still make this impractical today. In any case this approach might model specimen collections well, but it does not seem appropriate to a field guide, where it would result in one instance per class. Instead, we model an author's treatments of a large group, e.g. a family, as a class, and model individual taxa as instances. The cost of this is that a species described in an electronic field guide in several different treatments, e.g. by different authors for different locations, are not in the same object-oriented class. We discuss our approach to this issue below. We have implemented the UMASS-Boston Electronic Field Guide (UMB-EFG, or just EFG) on eXcelon Corporation's Object Store product (eXcelon, 2000). This OODBMS is a persistent store for Java (or C++) classes, and we describe next how we create such classes, along with our design requirements for author-friendliness, i.e. criteria by which the authors of descriptive data and keys are kept

isolated from the technology. We note in passing that object/relational database management systems (ORDBMSs) (Grimes, 1998) may well provide the support we require, but did not have mature programming interfaces at the time we began coding.

A design requirement of the EFG is that the system should be fundamentally ignorant of the nature of descriptive data. If an author chose to offer characters of restaurants of Boston, our software would produce a meaningful and useful restaurant guide instead of a field guide to the butterflies of Costa Rica, our initial target.

Diverse data such as we described above is called *semi-structured data* in recent literature (see Abiteboul et al., 2000). This means that it need not have a precise database schema such as would be found in a relational database. Pure OODB's are a special case, but of course so is a relational database (Abiteboul et al. *op. cit.*). This property makes it particularly easy for us to import diverse descriptive data and combine it in a single database. The taxon abstraction we use is a software construction called a *JavaBean*. (Sun, 2000a). Beans differ from other Java classes in having particularly convenient *introspection*. The introspection interface allows programs to inquire of the bean — in our case a collection of taxa having a common set of descriptive characters — what its properties and behavior are. This frees us from requiring advance knowledge of that behavior. As a simple example, when the user selects a group of taxa, e.g. butterflies of Costa Rica, we create, on the fly, a biologist-friendly search facility comprising a typical web form with pull-down menus for each character, the names of the character having been gained by introspection and the possible character values by inquiry into the database. (Here we mention 'biologist-friendly' because we will also describe below how our architecture supports flexible user interfaces by separating the UI from the rest of the system. In particular, we'll describe experiments in visual keys more suitable for amateur users.) A JavaBean must be compiled and loaded into the current Java Virtual Machine, so in essence we are generating and compiling source code on-the-fly when we import biological taxon treatments and we do this with some unsupported Java compiler classes available from Sun. This being a somewhat fragile approach, we are presently designing a different mechanism based on Java Map (Sun, 2000b) interfaces, which also can describe and manipulate objects based on their properties, and so are good candidates for a software model of a taxon.

Architecture overview

We have a typical three-tier Web application with Java servlets as middleware forwarding questions to an eXcelon Corp. ObjectStore OODB. HTML forms and Java Applets pass queries to the servlets, which transform them into queries suitable for the ObjectStore backend. After retrieving the data, the servlets build html pages, or send data to the applets, for presentation to the user. Although we operate our ObjectStore and servlets on the same host, this is completely unnecessary. In addition, ObjectStore is itself a distributed client-server system and various pieces of the database could be scattered around the Internet with no change to our architecture except to add resource discovery mechanism to find the distributed data. In a symmetric fashion, because servlets accept connections on internet IP ports, other clients than our own front end can forward queries to our servlets. See Nakhimovsky and Myers (1999) for an introduction to three-tier applications.

Importing data

We do not require an author to understand object-oriented technologies. Most biologists keep their descriptive data in software that is, or rests on, standard table-based databases or in a spreadsheet. Typical systems are Excel, Microsoft Access, FileMaker, and specialized systems such as Biota (Colwell, 1996), which uses 4D ODBC support (4D, Inc., 2000). Most of these products can respond to SQL queries along an ODBC connection (Microsoft, 1999) and in turn can be accessed in Java by the JDBC-ODBC bridge (Sun, 1997). More sophisticated systems based on Oracle can use JDBC natively. Our code uses this bridge to import database field names (i.e. character names). We must also attach the data at the appropriate place in the taxonomic tree, because in most cases the author will deliver data about a group of taxa,

e.g. a family in a particular locale, without a complete taxonomy. To accomplish the attachment we require a simple XML description of the common taxonomic hierarchy above all the taxa in the data source.

This import strategy is convenient, but leads to several problems. Most notably, the order of the delivery of fields is not determined by the ODBC protocol. Typically the order is that in which the fields were created, but that is often not the order in which the biologist may wish them displayed. Our solution, not fully implemented, is to specify the character order in metadata that is also represented in XML. A second difficulty is that it is possible to specify field names in many databases that do not yield legal Java identifiers (e.g. they contain white space). We convert illegal characters to underscores, but this is not a robust solution, since it may not allow going backwards from the OODB to the original. XML metadata specifying a one-to-one mapping of illegal characters to not-necessarily easily readable Unicode characters in Java identifiers would suffice, but would probably require tools to make the Java identifiers readable for software maintenance. Table 1 shows a typical XML metadata file.

Table 1.	Sample Metadata file

<EFGMetadata MetaDataID='Test'>
<EFGField>
<name>AntennaColor</name>
<type>morphological</type>
<weight>0</weight>
<dataCount>single</dataCount>
<dataType>simple data</dataType>
<javaType>java.lang.String</javaType>
</EFGField>
<EFGField>
<name>Habitat</name>
<type>ecological</type>
<weight>0</weight>
<dataCount>multiple</dataCount>
<dataType>simple data</dataType>
<javaType>java.lang.String</javaType>
</EFGField>
<name>Similar_Species</name>
<type>ecological</type>
<weight>0</weight>
<dataCount>multiple</dataCount>
<dataType>taxonomic reference</dataType>
<javaType>java.lang.String</javaType>
</EFGField>
…
</EFGMetadata>

Sample Data File associated with the above Metadata file	Table 2.

```
<EFGImportDocument MetaDataID='Test'>
<CommonPath>
<Domain>Eukaryotes</Domain>
<Kingdom>Animalia</Kingdom>
<Phylum>Arthropoda</Phylum>
<Class>Insecta</Class>
<Order>Lepidoptera</Order>
<Family>Nymphalidae</Family>
<Subfamily>Ithomiinae</Subfamily>
</CommonPath>

<TaxonList>
<Taxon>
<EFGPath>
<Genus>Ithomia</Genus>
<Species>heraldica</Species>
</EFGPath>
<CharacterData>
<ForeWingSpot>yes</ForeWingSpot>
<AntennaColor>orange</AntennaColor>
…
</CharacterData>
</Taxon>
<Taxon>
<EFGPath>
<Genus>Mechanitis</Genus>
<Species>polymnia</Species>
</EFGPath>
<CharacterData>
…
</CharacterData>
</Taxon>
…
<Taxon>
…
</Taxon>
</TaxonList>
</EFGImportDocument>
```

To present to a user with data arising from several heterogeneous sources there are fundamentally two approaches. One may build an integrated database and import each source, or one may build a federated view in response to a particular query that is executed—possibly after rewriting—against all of the data sources. The UMB-EFG architecture supports both, but our current implementation is only of the former. Later we describe some preliminary work in federation in which the UMB-EFG could potentially be one of the data sources, rather than a federating data consumer. However, even in our current implementation, any application anywhere on the Internet could make an IP connection to our Java servlet middleware and make queries in the same way our own HTML forms and Java servlet front-ends now do.

When we import data from an author's database we produce intermediate XML with the character data and such taxonomy as is described in the data source, and we require the author to provide some simple external taxonomy in XML form, namely <CommonPath> element in the example in Table 2. This latter specifies a position in the taxonomic tree at which the group is to be attached, specified as an edge-labeled path in the taxonomic tree, with edges labeled by the taxon level and nodes by the taxon name. The edge labels, i.e. the taxonomic level, are generated automatically—the author need only supply the taxon name. Authors can use simple available tools, such as Microsoft XML Notepad, to provide this CommonPath information. Edge-labeled digraphs have been considered previously for taxonomic data representation (Zhong et al., 1996), (NCGR, 2000). All data in the source must have taxonomy specified relative to the terminal taxon of that path. For example, if the source provides character values for the Ithomid butterflies, each record would provide genus and species. Typical XML generated by the importer is illustrated briefly in Table 2. From XML representation to JavaBeans is a transformation well-supported by currently available Java classes.

In Ithomid butterfly example, the imported data would normally be expected to have a field labeled Genus and one labeled Species and the data will be created in the taxonomic tree by subtrees under the Ithomiinae node, each subtree having edges labeled *Genus* and *Species* and with nodes appropriately named from the values in the imported data. Also, where appropriate, we treat various life forms (e.g. larva, pupa, etc.), and also each sex if the data identifies them as different, as though they were at a taxanomic level below species.

In addition, we have metadata for each character that comprises:

- The character name (e.g. *similarSpecies, 'foreWingSpotPresent')*
- A character type (presently either *morphological* or *ecological*)
- A numeric weight by which the author has ranked the character as to its importance in identifying an organism (See Dallwitz (1999b) for a discussion of this issue.)
- A string describing whether there may be multiple data in the same field
- A data type, presently one of *simple data, image filename, sound filename,* or *taxon reference.* A taxon reference is a reference to another taxon already in, or to be inserted in the database
- A Java data type.

Ultimately, we intend to treat this metadata with an XML Schema (W3C, 2000a) but currently the Java support for XML Schema is not mature. However, at this writing many other XML tools are.

User interfaces

Our three-tier interface allows us to support a variety of user interfaces. We presently have three in place, requiring increasing levels of biological sophistication.

The first of these is a purely visual interface to the Ithomid butterflies of Costa Rica, designed by one of us (Haber) with extensive field experience in the subject. In this interface the butterflies are first divided into four visually distinct groups ('Tiger' 'Clearwing', 'Yellow and Black', and 'Translucent Gold'). We display a conventional tree browser, functionally identical to the file system browser found on Windows but instead of folder and file icons, the

icons are thumbnail images of actual Ithomids representative of the group. As with a file browser, the user can click to expand one of the icons to another series of 3-4 butterflies representing a more subtle visual distinction within the group. This continues recursively about four deep and usually 3-4 wide. At leaf nodes of this tree, clicking brings a descriptive page for the species identified. There may be several paths to the same species, so internally this UI is represented not by a tree but by a directed graph (digraph). Digraphs are familiar data structures that have found related application in biological keys (Beaman et al., 1999), (Zhong et al., 1996), (NCGR, 2000). Note that key digraphs are not, in general, trees. That is, there may be several paths to the same taxon. We believe that a purely visual interface corresponds closely to the way amateurs use field guides, because we find that amateurs quickly get lost in traditional dichotomous keys and their generalizations. The biggest issue with this or any key is to verify that the identification is in fact correct. As an aid to this, our species description pages also contain links to similar species so that the user may compare their chosen species to ones that the author warns may be confused with it.

The second interface is a tree browser that follows the standard taxonomic tree, similarly to the Tree of Life project (Maddison, 2000). That system moves from web page to web page in the tree traversal, whereas we use the same Java tree browsing classes underlying the visual key. Hence, as for a file system browser, we keep everything on the same screen. The population of our database is presently too small for us to decide whether the attendant required scrolling would be counterproductive in a collection as large as the Tree of Life.

Finally, we have a pure html form-based character-value interface. The forms are constructed on-the-fly from the database, and the middleware builds and serves the form to the browser. Each form field offers the user the possible character states. The user can select the maximum number of species they wish to have offered that meet the current character value choices, and if the system finds more, the user is invited to limit the output by further choice from another character. In our present implementation we do not illustrate those choices as do a number of systems, but this is not precluded by our architecture. Without such illustrations, our interface is suitable mainly for users with some training. The use of characters with a finite number of states complicates naturally continuous numeric characters with ranges (Dallwitz, 1999a).

Because our species description pages are made up by the middleware in all cases, it is not difficult for us to produce XML instead of HTML for description pages as we do in the present implementation of our servlet middleware. Doing so has well-known advantages, some of which we have been exploring even though we do not yet output XML. These are discussed in subsequent sections.

Federating and transforming XML data

Here we describe our preliminary application of XML technologies to some of the issues arising in biodiversity software. Our experience as yet is too small, and the tools too new, for us to offer many engineering insights, so this section is devoted to discussing the capabilities of our prototypes and our current directions. All the applications in this section may be reached from the link 'recent XML work' on our home page http://www.cs.umb.edu/efg. This work is supported by a subcontract from the University of Kansas Biodiversity Research Center under their NSF grant KDI-9873021.

The range and quantity of biodiversity data on the web is large and rapidly growing (UNO, 2000). It is presently difficult to combine data from this large collection of sources for several reasons. First, few such sources publish or support any API to the underlying databases. Therefore, anyone who wishes to interrogate such data programmatically must reverse engineer the arguments passed by the publisher's web interface.

A second impediment to data federation, despite the explosion of XML support, is that the pages that are returned are generally HTML and so contain detailed presentation markup and little, if any, markup of structural or biologically descriptive utility. The Tree of Life project (Maddison, 2000) embeds special markup in support of its own web crawling engine,

and this can serve some of the requirement for metadata that can be addressed with XML. We are aware of a number of projects imminent or underway to serve XML, including the Integrated Taxonomic Information System (ITIS, 2000), (ITIS*ca, 2000), a joint project of the US Department of Agriculture and the Canadian Ministry of Agriculture. ITIS*ca already has experimental XML service on its site. The Biodiversity Research Center of the University of Kansas has implemented a gateway from Z39.50 servers to XML (Vieglais, 2000), about which we say more below. Indeed, anecdotally we understand that many biodiversity database maintainers and designers intend to serve XML.

XML is widely accepted as a data exchange language and for the specification of metadata and of configuration. We described above some of this use for the UMB-EFG itself. Early public uses of XML as a replacement for HTML for presentation focused on transforming XML to HTML in the browser (supported to date mainly in Microsoft Internet Explorer 5 (MSIE5), and we describe some demonstrations of that below. More importantly, XML can also be used for federating data from heterogeneous data sources (Abiteboul et al., 2000), (Baru et al., 2000). We will next sketch our experiments with these applications of XML to biodiversity software.

XML holds the promise of extremely flexible user interface configuration. Either at the server or in an XML-aware browser, a separate stylesheet is written in XSLT, the transformation language of the XSL stylesheet language (W3C, 2000a). This transforms the XML representation of, say, a descriptive taxon page, into HTML in a manner suited to the browser user or to the policies of the host that is forwarding the XML page (which may well be different from the host holding the original data). For example, among the XML demonstrations linked on the project page (Stevenson and Morris, 2000), we have one in which a number of static species description pages from various sources may be displayed with a number of sometimes widely varying styles (including one that shows the raw XML). When using MSIE5 the user need only push a button in the control panel to instantly change style sheets, the redisplay being handled in the browser itself and not requiring any return to the server for the new view.

Because XSLT is very general, it is possible to make mini-applications simply by transforming to a restricted view. For example, one of our applications accepts a list of (suitably marked) XML species description pages and produces a table of references in each of those pages to other taxa. This application looks like a database operation, but in fact is carried out entirely in whatever is executing the XSLT (in our case, the MSIE5 browser).

Another XSLT application we show is a bilingual display. With a single button click in a control frame, a species page is toggled between a Spanish and an English view. Again all transformations are done by the browser. In this application, text in both languages is kept in a single file (easing maintenance), with a language attribute on each text element. Under control of a simple Java Script program in the control frame, a single variable ('*currentLanguage*') is toggled and a single XSLT function displays an element only if its language attribute matches *currentLanguage*.

Among applications of XML, the promise of support of database federation is the most distant from the conventional applications of XML, but it is also the most challenging and interesting application. There is substantial advantage to standardized DTD's (or better, XML Schemas when they are stable) to combine data from various sources (Baru et al., 1999), (Baru et al., 2000). But recent research suggests that suitable schemas can be inferred from the sources themselves (Ludäscher et al., 1999), (Abiteboul et al., 2000). For the moment we use a fixed DTD and code static XML species description pages to that somewhat spare DTD. Those pages lack complete taxonomy, containing, as is typical, only the genus and species name. We have built a distributed federator based on (a very small part of) that known DTD, and knowledge of the query syntax of *zportal*, an XML/Z39.50 gateway built in The Species Analysis (TSA) project (Vieglais, 2000). Z39.50 is an international data exchange standard in wide use in government agencies and more recently adopted by many museums for their collection data. Because we use the zportal, details of Z39.50 are unimportant here, but they are discussed at the technical pages (on a link named 'Z.X') of the TSA site. In our federator, a JavaScript control panel on the application page looks into the static XML species page to

find the scientific name of the species described. This is sent to a mediator on our web server, a Java servlet that understands both the zportal query syntax and the location of taxonomic resources on the web. In this case, these resources comprise a small fixed collection of Z39.50 servers that offer taxonomic authority for specific collections of taxa. The mediating servlet queries each source until it finds the taxonomy for the given species, translates the XML returned by zportal into something user-friendly, and returns a combined page to the browser. (The zportal software returns XML, but based on a DTD whose element tags are simply XML rendering of the underlying numerically coded Z39.50 record tags).

The federation application described above finesses a number of difficult points that are the subject of our (and others') future work. The problem of resource discovery, e.g. finding taxonomic authority servers, is a deep and interesting one that appears in many similar contexts. So also does query rewriting in case the sources require different query syntax, and record rewriting in case the sources have different record semantics

Finally we mention an interesting issue that is the subject of work we are just starting: circumscription control—the control of the forwarding of sensitive data obtained from trusting sources. Some biodiversity data is sensitive (e.g. the precise location of specimens of endangered species). Consequently this data is often made available only to users who are trusted by the data holder not to abuse it. Unfortunately, most existing biodiversity data is held in databases with granularity of circumscription no finer than the entire database itself. No provision is in place to prevent access to sensitive data on a field-by-field basis or even a record-by-record basis. For example, specimen location data is sensitive only for endangered species, so circumscribing all location data is debilitating to many legitimate users and applications. To address this issue we are designing a *circumscription broker*. This is a set of protocols and software by which a data source can specify its circumscription policies and trust the broker to enforce them, thereby leaving the source able to serve entire records at will. Because the broker is acting on behalf of a database mediator, which is disassembling and reassembling records for the federated view in any case, the task of enforcing the circumscription belongs at the broker/mediator and mediator/query-response interfaces.

Conclusion

Biodiversity software is best implemented with software tools designed for dealing with data diversity. The engineering details of these tools can be hidden from the biologist who must use or populate the systems built around them. Java and XML technologies prove both suitable for this purpose, and highly synergistic with each other. Using them we have demonstrated an extensible web-based Electronic Field Guide and how to build applications that federate data from distributed sources around the internet.

Acknowledgements

References

4D ODBC Driver, 2000. http://www.4d.com/products/odbcdriver.html.

Abiteboul, S., Buneman, P., Suciu, 2000. Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, San Francisco.

Baru, C., Gupta, A., Ludäscher, B., Marciano, R., Papakonstantinou, Y., Velikhov, P., 1999. XML-Based Information Mediation with MIX. Exhibitions Program of ACM SIGMOD 99. Also available at http://www.db.ucsd.edu/publications/vamp.pdf

Baru, C., Ludäscher, B., Papakonstantinou, Y., Velikhov, P., Vianu,V., 2000. Features and Requirements for an XML View Definition Language: Lessons from XML Information Mediation. Principles Of Database Systems (PODS) 2000. Also available at http://www.db.ucsd.edu/publications/xmas.html.

Beaman, J., Luo, Y., Pramanik, S., Zhong, Y., February, 1999. HICLAS: A taxonomic database systems for comparing biological classification and phylogenetic trees. Bioinformatics Journal.

Biosis and Zoological Record's Internet Resource Guide for Zoology Software, 2000. http://www.york.biosis.org/zrdocs/zoolinfo/software.htm

Colwell, R., 1996. Biota: The Biodiversity Database Manager. Sinnauer Press, Sunderland Massachusetts. See also http://viceroy.eeb.uconn.edu/Biota

Dallwitz, M.J., 1999. Desirable Attributes for Interactive Identification Programs. http://biodiversity.uno.edu/delta/www/idcrit.htm

Dallwitz, M.J., 1999. Introduction to Data requirements for Natural-language Descriptions and Identification. http://biodiversity.uno.edu/delta/www/descdata.htm

Dallwitz, M.J., 2000. A Comparison of Interactive Identification Programs. http://www.biodiversity.uno.edu/delta/www/comparison.htm

Object Store Product Description, 2000. http://www.exceloncorp.com/products/objectstore.html

Futuyma, D.J., 1998. Evolutionary Biology, 3rd Edition. Sinauer Associates, Massachusetts.

Digital Taxonomy Software, 2000. http://www.geocities.com/RainForest/Vines/8695/

Grimes, S., April, 1998. Modeling Object/Relational Databases. DBMS, http://www.dbmsmag.com/9804d13.html

Integrated Taxonomic Information System, 2000. http://www.itis.usda.gov/plantproj/itis

ITIS*ca - Canadian version of ITIS Integrated Taxonomic Information System, 2000. http://res.agr.ca/itis/

King, N.H., June, 1997. Object DBMSs: Now or Never. DBMS, http://www.dbmsmag.com/9707d13.html

Lampinen, R, October, 2000. Cartographic Links For Botanists. http://www.helsinki.fi/~rlampine/cartogr.html

Ludäscher, B., Papakonstantinou, Y., Velikhov, P., Vianu, V., 1999. View Definition and DTD Inference for XML. Post-ICDT Workshop on Query Processing for Semistructured Data and

Non-Standard Data Formats. Also available at http://www.db.ucsd.edu/publications/icdt-ws-99.pdf

Maddison, D. (Coordinator and Ed.), 2000. The Tree of Life, http://phylogeny.arizona.edu/tree/phylogeny.html

Microsoft Corporation, 1999. Microsoft ODBC. http://www.microsoft.com/data/odbc/

Nakhimovsky, A., Myers, T., 1999. Professional Java XML Programming with Servlets and JSP. Wrox Press, Birmingham, UK.

National Center for Genome Resources NCGR Taxonomy Project, 2000. http://www.ncgr.org/research/sequence/taxonomy.html

Papakonstantinou, Y., Velikhov, P., 1999. Enhancing Semistructured Data Mediators with Document Type Definitions. Data Engineering 99. Also available at http://www.db.ucsd.edu/publications/icde99.pdf

Saarenmaa, H., Leppäjärvi, S., Perttunen, J., Saarikko, J., 1995. Object-oriented taxonomic biodiversity databases on the World Wide Web. In: Kempf, A., Saarenmaa, H. (Eds). Internet Applications and Electronic Information Resources in Forestry and Environmental Sciences. Workshop at the European Forest Institute, Joensuu, Finland, August 1-5, 1995. EFI Proceedings 3. Also available at http://www.efi.fi/~saarenma/oobdwww-nature-latest.htm

Shattuck, S., Fitzsimmons, N.J., 2000. BioLink®The Biodiversity Information Management System. CSIRO Publishing, Collingwood Victoria 3066, Australia. See also http://www.ento.csiro.au/biolink/index.html

Stevenson, R.D., Morris, R.A. (PIs). The UMASS-Boston Electronic Field Guide Project. http://www.cs.umb.edu/efg

Sun Microsystems, 1997. JDBC-ODBC Bridge Enhancements. http://java.sun.com/products/jdk/1.2/docs/guide/jdbc/bridge.html

Sun Microsystems, 2000a. JavaBeans Tutorial , Part 1. http://developer.java.sun.com/developer/onlineTraining/Beans/Beans1/index.html

Sun Microsystems, 2000b Java Map Interface API, 2000. http://java.sun.com/products/jdk/1.2/docs/api/java/util/Map.html

The Biodiversity and Biological Collections Web Server, 2000. http://biodiversity.uno.edu

Vieglais, D., 2000. The Species Analyst. http://habanero.nhm.ukans.edu/. For description of the zportal interface, go to the Z.X section of the site.

World Wide Web Consortium, 2000. XML Schema. http://www.w3.org/XML/Schema

World Wide Web Consortium, 2000. Extensible Style Sheet Language. http://www.w3.org/Style/XSL/

Zhong, Y., Jung, S., Pramanik, S., Beaman, J.H., May, 1996. Data model and comparison and query methods for interacting classifications in a taxonomic database. Taxon, 45(2). Also available at ftp://ftp.cps.msu.edu/pub/hiclas/paper1/paper1.txt

Technological opportunities and challenges in building a global biological information infrastructure

Hannu Saarenmaa ([12])

Abstract

An overview is given how the newest e-business technologies can help to manage biodiversity information. The possibilities have increased dramatically just over the one or two past years. It is important to see the difference between infrastructure, meant to provide shared building blocks, and applications that are there just to solve a specific problem. Many of the new possibilities especially enable the building of infrastructure and are important for global cooperative processes such as CHM and GBIF. Many of them are based on XML that unifies presentation, data and document management. It enables information interchange with both human and computer-readable packages. Using XSL the information can be viewed from various angles. Biodiversity namespaces in XML Schemas should be standardised urgently. Examples of content that should rely on XML from this on include taxon homepages and observation data. Analogously to Internet's DNS, a new addressing system for scientific names of organisms could be designed using stable numeric identifiers. This would be used to overcome the volatility of the Linnéan names, which are not suitable for keys in information interchange. Using such global IDs for both taxa and names, biological information could be addressable from web services worldwide. How these information infrastructure components might be used in the web services of GBIF such as the Catalog of Names of Known Organisms and the SpeciesBank is discussed.

Keywords: taxonomy, biodiversity, informatics, DNS, XML, registries, web services.

The challenge

It is year 2000 and the computer industry has successfully welded off the dreaded Y2K problem that once threatened to shut down the techno-society. It is time to start focussing on the less-known Y3K bug: In the year 3000, there will not be a single species left on the planet if the current rate of extinction is linearly extrapolated. This is based on a best available guesstimates ([13]) of extinction rate 10^4 per year out of a 10^7 total species. Naturally, the development will not be linear, and the future could be different. The cause of this development is well known: Habitat destruction, which follows from low value of biodiversity, which follows from lack of knowledge about it, and poor access to information.

According to Groombridge (1992) the rate of discovery of new species happens to be of the same magnitude 10^4 per year as the extinction. This rate has remained the same over the past 30-40 years. The slowness of discovery projected against extinction means that one half of the species will not be discovered until they are lost. Again, the development is not going to be linear and it is reasonable to assume that collections already are hiding in them about a similar number ($2*10^6$) of undiscovered species as there are known organisms.

Regardless of how these projections are going to develop, the facts remain. Serious things are happening and there is poor use of knowledge and difficulty in information access. Meanwhile, taxonomic and biodiversity information is mushrooming on the Internet. Species homepages are spontaneously being created everywhere. Giving a scientific name to a general-purpose search engine returns hundreds of hits that also include serious

documentation by competent authors. However, there is almost no coordination, nor standards in this area, and certainly no quality assessment of these content is available. There is only the plain information infrastructure of Internet and variably applied scientific names.

How did it come to this? The main problem is the working habits of taxonomists – taxonomy has not been understood as a separate information science, but has been too closely been coupled with systematics. Outsiders often see taxonomy as an endless sink of resources because its returns are so slow and not shared effectively. Traditions weigh heavy and there has been very little on-line publishing of copies of new taxon descriptions. The Linnéan naming system is volatile and does not support modern information access. In order to cope with the volatility, there is built-in slowness in the Codes (International… 2000) that regulate establishment of new names. There are complex intellectual property rights issues between north and south, requiring repatriation of information ([14]).

Reflecting about the unhappiness with the situation, there has been a proliferation of initiatives and cooperative networks ([15]) and hence plenty of organisational response. However, a technological response has usually been less effective. This diversity, however, signals that some necessary ingredients have been missing from the services of all these networks. I believe what lacks is a better understanding of what constitutes an information infrastructure.

The purpose of this paper is to give an overview to the information infrastructure components that would better allow taxonomists to share and reuse information. Current developments on Internet and e-business have opened a plenty of new possibilities. Indeed, the upcoming Global Biodiversity Information Facility (GBIF) is expected to put the infrastructure components in place for the service of biodiversity informatics.

What, exactly, is infrastructure?

If I have seen further than other men, it is because I have stood on the shoulders of giants' said Isaac Newton of his works, pointing to the history of science all the way back to the ancient Greeks (Turnbull et al. 1959).

The essence of infrastructure is building on each other's work using standardised interfaces. Instead of directly working together, an intermediary shareable service is made available. Scientific publishing is one such service.

There are many kinds of infrastructure, and most people understand it as something purely physical, such as hardware and wiring. However, infrastructure can also be immaterial. Indeed, 'information infrastructure' is everything that supports the flow and processing of information ([16]). It consists of various standards, services, and support actions for representing, addressing, locating, exchanging, and securing information. Building such infrastructure does not necessarily require a mega-projects and new organisations, although they are often viewed as that way ([17]). A simple standardisation process is where to start.

The separatedness of computer software applications and the underlying information infrastructure is seldom fully understood. If you want others to build on it – then you are building infrastructure. These are characterised by the usage of open, standardized interfaces (where to go) and communication protocols (what to say). Examples include IP, Z39.50,

---

(14) http://www.biodiv.org/doc/meetings/cop/cop-05/information/cop-05-inf-03-en.pdf
(15) An incomplete, alphabetically arranged list of abbreviations to biodiversity information networks and services includes at least the following: ABREN, All Species, BCIS, BIN21, BIODI, BIOSIS, CBD, CHM, CONABIO, DIVERSITAS, ECNC, ETC/NC, EWGRB, GBIF, IABIN, IBIN, ILDIS, INBio, IOPI, IPNI, ITIS, IUCN, MAB/BIS, NBII, Species 2000, TDWG, Tree of Life, WCMC. Individual projects with a begin and an end have not been listed.
(16) http://www.whatis.com
(17) …. organizing the information from biological collections would need Museums to develop of standards for data, and database large numbers of records. Providing information in a maximally accessible form could require a distributed database system, probably integrating massive data handling environments with the web. The third aspect, application of that information to meet the needs of society requires integration with analysis and visualization capabilities…'.— *An excerpt from early NBII plans*

LDAP, IMAP, CORBA, SQL, and Posix. However, if you just want to get one job done — then you can build an application, and perhaps hope that others could use it as is. Popular examples include Microsoft Windows and Microsoft Access. Mature, successful, shareable applications often migrate to infrastructure. For instance Microsoft Office is currently found in most personal computers and one can travel with a PowerPoint file and have some confidence that the file can be opened at a remote presentation location.

Infrastructure services are typically built using a layered approach (Figure 1) where services are built on top of each other. Interoperability is achieved when an interface at one level can connect to the next and understands a protocol, which is used to express needs of requesters.

A simple example of infrastructure services is directory service. It is possible to use hierarchically arranged data on people, organisations, and their groupings with the Lightweight Directory Access Protocol (LDAP) on connecting to an appropriate port of a directory server on the Internet. Existence of such an open interface and a related protocol makes it possible for others to build applications that use directory services. These include roles and expertise, security services and accreditation mechanisms can be built on it. Also email applications increasingly build on directory services. Examples include 411.com, Infospace.com, and EIONET ([18]).

What the information infrastructure components would be for biodiversity is discussed in the rest of the article. They are tentatively identified in Figure 2. Special reference is made to the GBIF plans on these areas (OECD… 1999).

Naming and addressing biodiversity information

Telephone system is one essential community infrastructure. Its usefulness is greatly facilitated by a phonebook linking the names with unique phone numbers that can be called. Similarly, the Internet has a Domain Name System (DNS) (Albitz and Liu 2001) that maps names of computers to IP numbers. They uniquely identify computers' network interfaces, so that data packets can be sent to them. The World Wide Web has Universal Resource Locators that point to files on web servers, so that a web browser can download them.

Names of biological organisms form a comparable phonebook, in fact, a semantic hierarchically arranged network. However, there are no numbers to call, yet.

Catalog of Names of Known Organisms

The GBIF aims at compiling a global Catalog of Names of Known Organisms (CNKO). As Figure 3 illustrates, it would be used to link to all the other data. Therefore, it is clear that it is not just a database application that allows search and lookup of only the names themselves. Such a service has already been achieved for many groups by the federated databases of the Species 2000 ([19]), and there is no need for a duplicating that effort. Instead, the CNKO would have to be a resource discovery mechanism similar to the phonebook, DNS, or those currently being proposed for e-business.

---

(18) My organisation EIONET provides a directory service at ldap.eionet.eu.int, port 8983, root
     ou=users,o=eionet,l=Europe.
(19) http://www.sp2000.org/

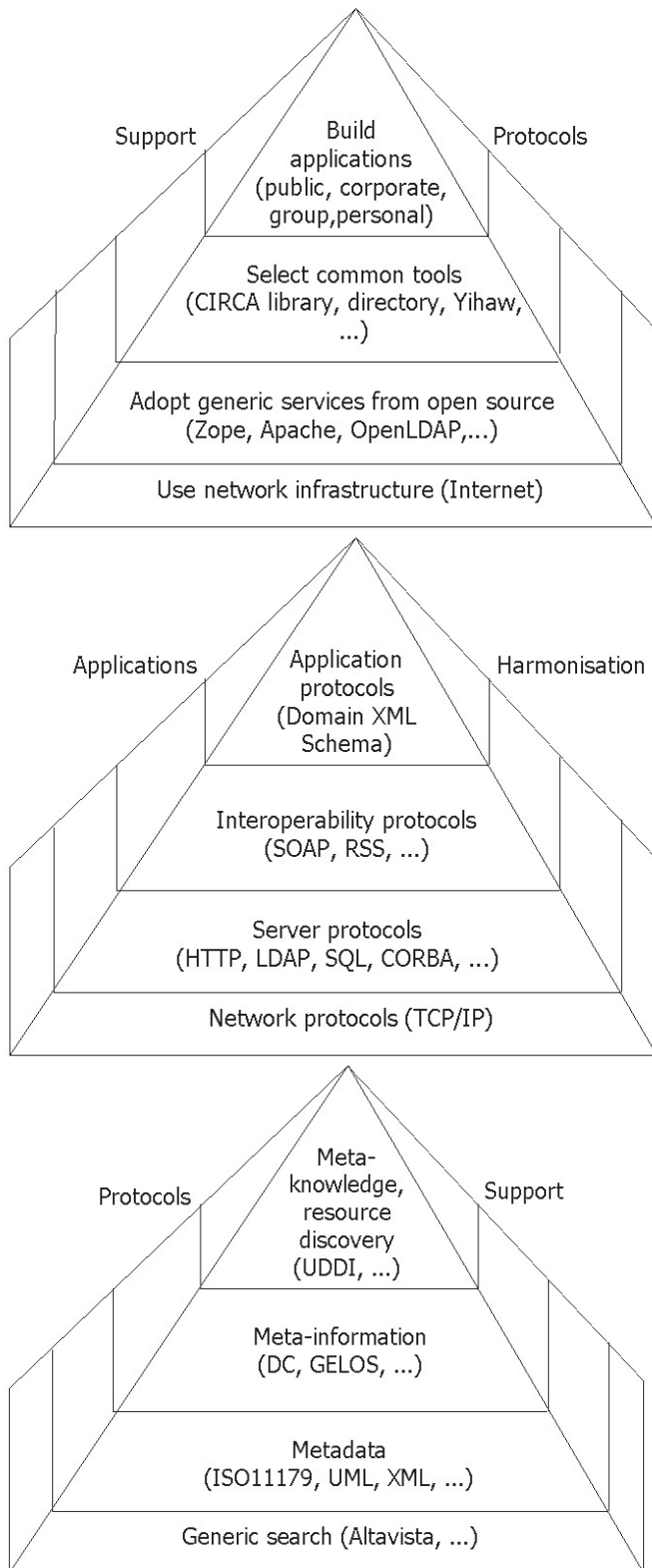| Three sides of an interoperability pyramid picturing an application, communication, and content harmonization infrastructure. Different layers of services build on each other using standardized interfaces. Examples are given in parentheses	Figure 1.

Support   Build applications (public, corporate, group,personal)   Protocols

Select common tools (CIRCA library, directory, Yihaw, ...)

Adopt generic services from open source (Zope, Apache, OpenLDAP,...)

Use network infrastructure (Internet)

Applications   Application protocols (Domain XML Schema)   Harmonisation

Interoperability protocols (SOAP, RSS, ...)

Server protocols (HTTP, LDAP, SQL, CORBA, ...)

Network protocols (TCP/IP)

Protocols   Meta-knowledge, resource discovery (UDDI, ...)   Support

Meta-information (DC, GELOS, ...)

Metadata (ISO11179, UML, XML, ...)

Generic search (Altavista, ...)

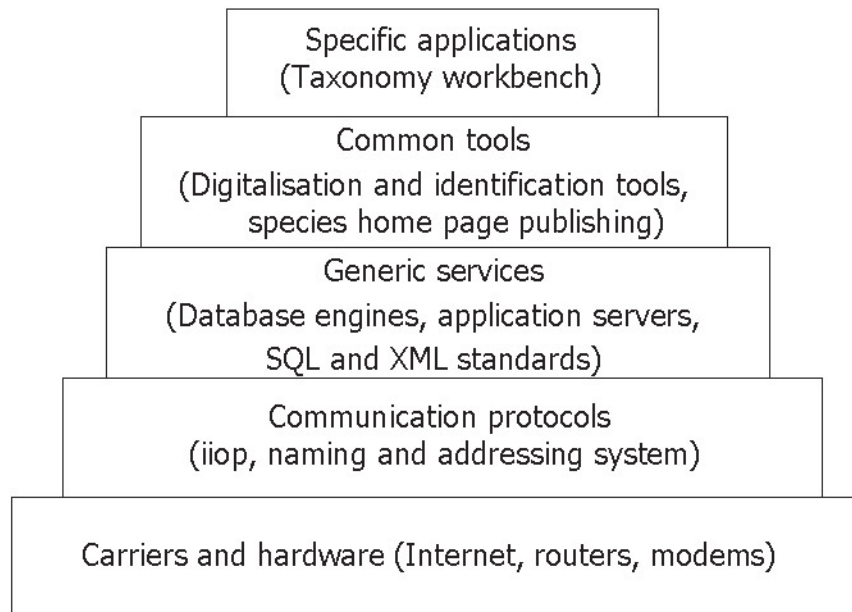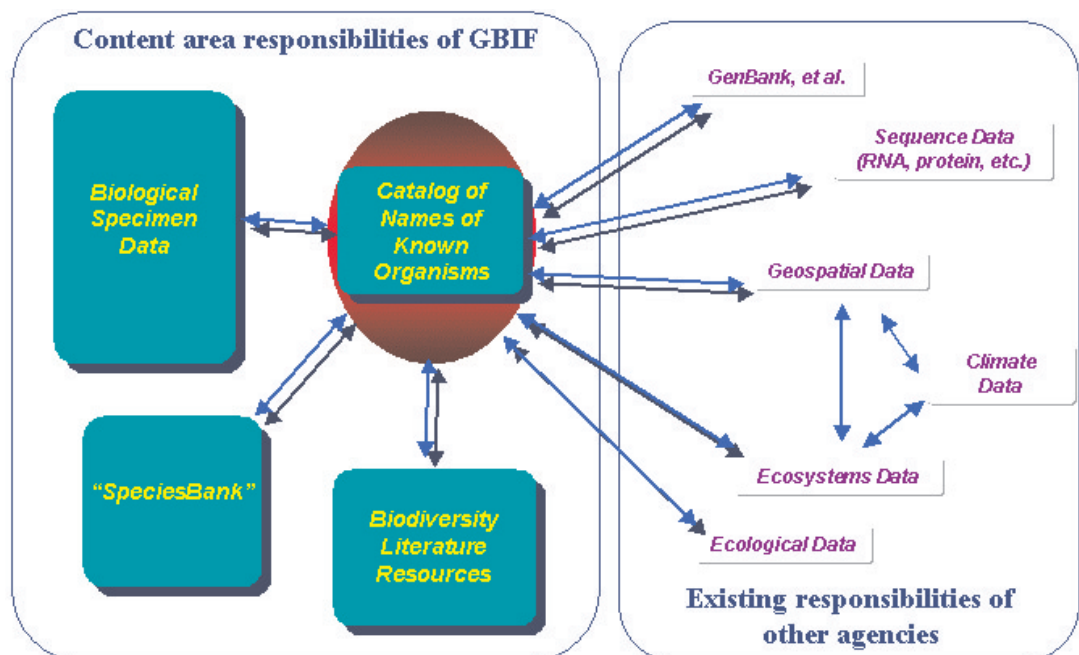| Figure 2. | An example of a biological interoperability pyramid |

| Figure 3. | The GBIF would enable synergism among existing investments that is not possible at present. The Catalog of Names of Known Organisms is the key linking component in GBIF information architecture |

If the name is used as linking mechanism it must have other qualities than just describe a placement of an organism in a certain genus or higher taxon, species and subspecies. It must be short, unique, stable, and universally accepted.

Unfortunately the Linnéan name system does not fulfil these requirements alone. The names are not short, they change their form, they are not always associated with the same taxa, and are even used variably in the different parts of the world. Association with taxa will always remain subject to change, which should not be seen as a problem, but as progress and a result

of taxonomic research. However, the other problems could probably be overcome by adoption of new globally unique identifiers for both names and taxa, such as the taxonomic serial numbers used and maintained by ITIS ([20]) and the taxon codes once created by the Nordic Code Centre ([21]).

Creating unique identifiers for names is straightforward. However, linking them to taxa is not. The only solid anchor to taxa is the name-bearing type specimen in some collection. This means that also these type specimens must bear globally unique identifiers (and be digitalised, see section 5 below). Then a linkage must be made between the type specimen identifier and the identifiers of names that have been used to describe it. It is this link, which can be called a taxon identifier that can be reused to point to other information.

Stability of names is beyond this discussion, but a few side notes could perhaps be allowed here. Much could be achieved if the concept of genus would be discontinued. Moving over to Phylocode (Cantino et al. 1999), which essentially means adoption of only single-word names would eliminate changes in the form of names and also reduce the frequency of other name changes dramatically. However, as common names in many languages will still have to be supported, there is no avoidance of synonyms of some kind. Hence, synonymy should be fully supported by any addressing system.
These requirements resemble a lot those of the Internet's DNS. So it is worthwhile to consider how far the analogy would go.

Addressing

A scientific name is unique within a kingdom. So, something like http:// sylvestris.pinus.plantae.bio/ should in principle work ([22]). What would it return? ([23]). Given to a web browser, it would first be translated by the DNS to a unique IP number, such as 130.226.11.38. Then the **one server** listening the default port of the Hypertext Transfer Protocol in that interface number would return its default content, that will be called the taxon home page (THP) in this paper. The THP would probably be an XML formatted rich homepage of the content that can be associated with that name. More about use of XML for THPs below.

The DNS routinely works with synonyms, so entering silvestris instead of sylvestris would be possible to map to the same IP number. Scots.pine is not a problem, either.

However, as there probably are dozens of equally qualified scientific laboratories that can put up authoritative content under this name, it would lead to interesting discussions at the pinus.plantae.bio root name server authority for whom to assign sylvestris. One possibility is to assign it to the institution that holds the holotype specimen, but as these may not always be capable of providing such a service, this could not be the rule. Wherever the THP is assigned, some sort of registration and broadcasting mechanism from the there on to all other sites providing related content should be devised.

One possibility is to build on the upcoming IPv6 that allows 16-byte IP numbers instead of the four of the currently used IPv4. IPv6 also comes with a broadcasting system, so that the query, which in the above scenario would go to one server only, could for instance go to all the 255 servers under 130.226.11.38.* (we don't show the 11 heading bytes here). IPv6 has so much address space that hundreds of IP numbers can be assigned for each m² of Earth's surface. It should be possible to allocate a set of 255 numbers for each known taxon. However, as an Internet Service Provider actually owns its set of numbers and handles the routing to them, using a fixed set of numbers is not straightforward, and would require elaborate reverse proxy schemes at a central location.

---

(20) http://www.itis.usda.gov/
(21) http://www.nrm.se/ncc/
(22) Of course, a top level domain .bio does not yet exist.
(23) Entering *Pinus sylvestris* to AltaVista search returns 6251 hits today. Several of these are serious, elaborate home pages to the species.

To summarise, the DNS offers interesting possibilities for linking biodiversity information. It maps changing computer names to stable IP numbers easily. However it requires additional layers of standardisation in order to support resource discovery beyond a single THP.

Representation of biodiversity content

Biodiversity information covers many levels of aggregation, such as molecular, genetic, species, population, habitat, site, and ecosystem. The two first ones have had their representation issues solved by the bioinformatics research community. Ecosystem modelling has thrived also for decades in ecological research. The biggest unresolved question is on how to represent species-level knowledge.

Earlier we identified taxon home pages (THP) as a main form of such content. These should not be understood as simple web pages in HTML, but as semantically well-organised pieces of knowledge. They need to be viewable by humans as simple or elaborate web pages, depending on the user needs. They also need to be understandable by machines for automated processing.

This can all be achieved by using the eXtended Markup Language (XML) for representing species-level biodiversity information. The XML idea is to separate content from presentation so that it can be semantically understood by machines and processed automatically (Figure 4). There are many excellent books on XML (e.g., Harold 1999) so we skip general overview to it and only discuss the basics of representing taxonomic information in XML. This is a hot issue in biodiversity informatics research currently.

| Figure 4. | The structure of XML data is defined in XML Schema and is separate from its formatting information on XSL |
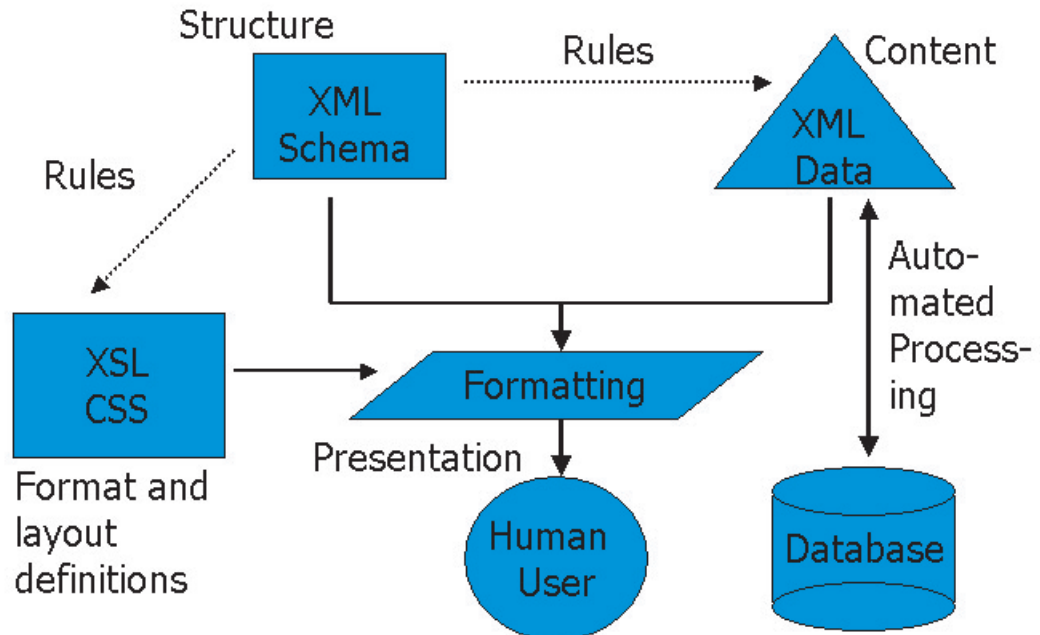
On regular web pages, the content and formatting are hard-wired. Tags such as <I>Pinus sylvestris</I> around the italicised Latin name of a species contain no hint of what has been placed inside these tags. On XML, the same could be tagged <NAME>Pinus sylvestris</NAME>. However, now there is a need to say that all <NAME> tags should be formatted using <I> when viewing on regular web browsers. This is achieved by the eXtended Stylesheet Language (XSL). When viewing an XML file, new web browsers can choose among several XSL files so that different views can be generated to the same content. However, as this requires the browser to actually understand XML, this step of XML Transformation is often processed by the server and plan HTML returned.

XML format makes it possible to analyse and process web page content automatically. XML files are therefore very popular in data exchange and are in fact replacing all other formats. However, a new problem emerges when XML files have to be understood by computers. They have to be built according to some rules so that some logical entities are formed. Otherwise the content could be misinterpreted. These rules are expressed in XML Schema language.

Table 1 illustrates how a THP would look like in HTML and XML. Table 2 shows an XML Schema that can be used to define how THPs should be formed. Of course, most THPs in XML would be formed automatically from content in databases. End user support for creation of these XML files should probably be provided by a specialised client application.

A simplified taxon home page in HTML (left) and XML (right)	Table 1.

```
<HTML>
 <HEAD>
 <TITLE>Homepage for
Cosmotriche lobulina</TITLE>
 </HEAD>
 <BODY>
 <H2>Name</H2>
 <I>Cosmotriche lobulina</I>,
 syn. <I> Selenephera lunigera</I>.
 <H2>Distribution</H2>
 Palearctic.
 <H2>Host</H2>
 <A href=http://www.examplesite.org/
?ID=27.111.1.0>Pinus</A>.
 </BODY>
</HTML>
```

```
<?XML version '1.0'>
<TAXON ID='122.120.10.1'>
 <NAMES>
 <VALID>Cosmotriche lobulina</VALID>
 <SYNONYM>Selenephera
lunigera</SYNONYM>
 </NAMES>
 <DISTRIBUTION>Palearctic
 </DISTRIBUTION>
 <HOST>
 <ID>27.111.1.0</ID>
 <NAME>Pinus</NAME>
 </HOST>
</TAXON>
```

An XML Schema that defines the XML taxon home page in Table 1	Table 2.

```
<?XML version '1.0'>
<elementType id='TAXON'/>
 <DESCRIPTION>A taxonomic unit, a homogenous group
 of organisms.
 </DESCRIPTION>
 <element type='#ID'occurs='REQUIRED'/>
 <element type='#NAMES'occurs='ZEROORMORE'>
 <element type='#VALID'occurs='REQUIRED'/>
 <element type='#SYNONYM'occurs='ZEROORMORE'/> </ELEMENT>
 <element type='#DIAGNOSTIC'occurs='ZEROORMORE'/>
 <element type='#DISTRIBUTION'occurs='ZEROORMORE'/>
 <element type='#LIFECYCLE'occurs='ZEROORMORE'/>
 <element type='#HOST'occurs='ZEROORMORE'>
 <element type='#ID'occurs='REQUIRED'/>
 <element type='#NAME'occurs='ZEROORMORE'/> </ELEMENT>
 <element type='#PROTECTION'occurs='ZEROORMORE'/>
 <element type='#REFERENCES'occurs='ZEROORMORE'/>
</elementType>
```

Most industrial and commercial sectors are now working on standardisation of their XML Schemas. The schemas can be registered on dedicated services such as OASIS ([24]) so that anyone interested in information interchange can download and use them. Recommendations already exist ([25]) for simple time and space representation, and XML records can be downloaded at least from ITIS today, but it is still a challenge for the biodiversity informatics community to agree on its own standards in this area. Such

(24) http://www.oasis-open.org/
(25) http://www.dublincore.org/documents/

comprehensive standards are called namespaces, and they can be referred to from the XML file. So it should not be necessary for every project to define its own XML Schema before being able to interchange data.

Digitalisation

Biodiversity content is not just HTML and XML. There are many multimedia aspects that are important. Below we cover some of them. Going 'from atoms to bits' as conceived by Negroponte (1995), which in essence means tuning the physical object into its digital representation, brings numerous benefits. This can be done via imaging or by generating character-based descriptions of the objects for digital identification.

Imaging

Digital imaging has recently become surprisingly easy and affordable. Today's garden-variety digital cameras produce 3 megapixel images that are large enough to fill the screen of even largest monitors. This resolution is enough to document almost anything that can be shown in 2 dimensions. Such an image can be downloaded even with a regular modem. Collapse of the costs of mass storage (disk drives and writable CD-R) has also made archiving such content inexpensive.

There are many benefits in digitalisation. Loans of material become faster and risk free, if the specimens or their details are digitised by their keepers, and the results are made available over the web. Imaging also acts as backup of the specimens in case that they are accidentally destroyed or lost. Also the original colours will be saved for future. Repatriation of information to the originating country becomes feasible. In fact it should no longer matter where the specimens are physically stored. They could even be stored in fumigated vaults only accessible to a robot…

Images do not necessarily have to be true representations of individual specimens. It is possible to form arbitrary non-existing species by 'morphing', i.e., overlaying many images with each other. Such an image could represent an intermediate step in an identification process. Possibilities for automatic pattern recognition and automated diagnosis have not been studied much.

The remaining difficulties in widespread digital imaging are mainly organisational. How to set up a system such that remote material requests are processed and the digital content is made available in a demand-driven fashion. There simply is so much material to be digitised that a supply-driven approach will not work. Another related matter is creation of the appropriate storage procedures. Even though mass storage is cheap, making backups and keeping material in order is still a chore.

Digital identification

Traditionally identification of organisms has been done through literature, expert advise, and by comparing specimens with those already identified in collections. All these can be done, and also augmented with modern web services.

We have already covered how THPs could be made. Electronic field guides ([26]) consisting of collections of THPs can complement in an important way the printed media. They can include interactive features such as executable keys and links to new information that would have been released after a print date. Applied entomology, in particular is building heavily on home pages of economically important organisms and diagnostic expert systems that can reason over their content (Väkevä et al. 1996).

DELTA is a character description language that has been adopted by the international Taxonomic Databases Working Group TDWG ([27]) as a standard for data exchange. DELTA-

---

(26) http://www.cs.umb.edu/efg
(27) http://www.tdwg.org/

formatted data can be used to produce natural-language descriptions, interactive or conventional keys, cladistic or phenetic classifications, and information-retrieval systems (Dallwitz 1980, Dallwitz et al. 1993, 2001).

Remote human expert advise over the web is nowadays routinely available from several web services ([28]). It is only matter of time that similar specialised services appear for biology. Even today, without any special service a question posted on newsgroups such as sci.bio.entomology will not remain unanswered. However, the results do not accumulate and the practice is not common.

Electronic publishing

Making the original taxon descriptions available on the Internet is of course very much desirable. Only using a digital format is not possible for a publication to be recognised. However, a workable combination of printed and electronic should now be possible, as the 4th edition of the International Code of Zoological Nomenclature (1999) states: *'A work not printed on paper (e.g., on a read-only laser disk) issued after 1999 in numerous identical and durable copies may be regarded as published if supplemented by identical copies deposited in at least 5 named and publicly accessible libraries.'* So, the paper copy is only needed for the record and archive anymore.

Descriptions have already appeared as preprints on Internet. Two new species of the plant genus Tetranema were documented on the InBio web site in 1996 before the printed publication appeared (Grayum and Hammel 1996). An enhanced electronic copy of the paper is still available.

Issues that remain to be solved in electronic publishing are mainly related to copyright. Will the traditional publisher, if employed, allow an electronic copy? Standards should be established how to represent a taxon description in XML. A registry of such descriptions is needed so that they can be searched and found.

Registering and locating of biodiversity content

Above we have discussed how biodiversity content could be identified, represented, and turned into bits. The most important issue, though, is coming up with an infrastructure that motivates the users to contribute their pieces of content into the common pool. It is exciting to put digital images up on the web and create flashy taxon home pages, but if these are not contributing to a common, shared pool of knowledge, they remain just nice demonstrations. Value of such content is low.

The key here is to create a global registry of biodiversity content. It is rather surprising that such a key infrastructure component has received only little attention until lately. Examples how it might work and change how information is shared can be seen in other areas, for instance on GenBank, Napster ([29]), or any business-to-business (B2B) service that connects content providers together into a value chain.

SpeciesBank

The GBIF plans (OECD… 1999, Figure 3) identify a 'SpeciesBank'. Not much has been yet written how it might work, but the following excerpts from the GBIF plans list the requirements.

- Link to any accessible existing database that holds information about species.

---

(28) http://www.allexperts.com/ ; http://www.experts-exchange.com/;
http://www.sciam.com/askexpert/; http://experts.yahoo.com/
(29) http://www.napster.com/. When discussing Napster as a model technical solution, one should not be distracted by the unresolved IPO issues that affect that particular service. The biodiversity community has its IPO issues solved by the CHM and GBIF processes.

- Facilitate searching of Internet resources by non-specialists.
- Assist taxonomists to avoid re-naming already-described species.
- Enable rapid dissemination of information on newly discovered species.
- Speed up repatriation of information about species native to developing world.
- Increase the rate at which new species are described.
- Enhance accessibility of species information to users.

Above all, the SpeciesBank should be a central registry of distributed content, and facilitate new species discovery and description. By registry, we mean a place where meta-information of remotely held biodiversity content is stored and can be searched. This meta-information can be centralised so that many aspects of the content, such as names of taxa, locations, collections, observers, dates, etc. are held there. Alternatively it could only hold a minimum set, and the queries are forwarded to the original content providers. Probably a rich central meta-information set is better for users and robustness of the system, although it may not always be as up to date as a distributed one. Anyway, the provided meta-information fields must be soon identified. Appropriate standards such as Dublin Core ([30]) and RDF ([31]) should be built on.

An important question is how the remote content providers actually register their content such as THPs at the SpeciesBank. A manual entry using web forms is the simplest choice. However, if the content were provided in a form of rich THPs in XML, just sending an URL in an email would be sufficient.

A more sophisticated approach would be to design a special SpeciesBank protocol. A remote content provider would use a SpeciesBank tool to properly format the THP in XML and publish it on a local web server. The tool would then announce the new content to a central registry. This is how, for instance, the Napster registry of music works using its own client and protocol.

Another approach to discover distributed content is through portal web services. It should be attractive to maintain a central biodiversity portal using meta-information of the SpeciesBank. This could allow regular users to personalise the content the way they like and create an interface 'MySpeciesBank' that only shows the interesting content. Personalisation can be achieved in many ways, but one popular standard is Rich Site Summary RSS ([32]). That is an XML format for providing channels of information that can then be tiled up on the customised 'My' page.

In addition to providing access to distributed content, it could be harvested and archived into a repository. Certainly many sites with valuable content will go down for a variety of reasons. Unless explicitly forbidden by their custodians, cached copy of such content could still be made available. This concerns especially taxon descriptions and type specimen images. However, a fully central solution similar to GenBank is probably not desirable for species level biodiversity.

The link between CNKO and SpeciesBank is particularly intriguing. The CNKO forms a semantic net that is hierarchically arranged for taxonomy, but there can be many cross-links for other relations such as host-parasite. Such meta-information, if made available in SpeciesBank is very valuable when dynamically creating portals and other web services.

Resource discovery and database access

Any biodiversity web service that covers more than just basic content will need to provide its contents dynamically from a database. There are many ways to make such a connection, but the most popular ones employ a language such as Perl, PHP, Python, Java and ASP, and database connector, like JDBC. More sophisticated approaches employ an application server such as Enhydra or Zope.

---

(30) http://www.dublincore.org/
(31) http://www.w3.org/RDF/
(32) http://my.netscape.com/publish/formats/rss-spec-0.91.html

The problem with database access is that databases cannot be directly searched using general-purpose search engines. Therefore their content must be made queriable by remote programs, or at least described somehow.

Not long time ago, the obvious answer to the access problem would have been 'query using distributed objects'. The database would have to be guarded by a group of agents that receive remote requests on protocols such as CORBA or COM+ from other objects somewhere on remote sites. Sophisticated agent formalisms are available for such communication (Saarenmaa 1999). Another approach is to use the Z39.50 protocol, as the Species Analyst [33]. In order to make two databases able to interchange data such way, their data models would have to be harmonised with each other. Plenty of effort has gone into standardisation of taxonomic databases, especially in botany.

XML has changed the picture dramatically. No standardisation of data models necessary if only an XML data interchange format can be agreed on. Any taxonomic and biodiversity data model easily translates to XML Schema.

Moreover, there are other standards and protocols available from the e-business world that could be useful for resource discovery and remote access. Most of them build on XML, including ebXML [34]; Universal Description, Discovery and Integration UDDI [35]; Advertisement and Discovery of Services ADS [36]; Web Services Description Language WSDL [37]; and more. Particularly interesting is Simple Object Access Protocol SOAP [38], which can be used to invoke queries to remote databases after they have been discovered with some of the above-mentioned services. The relevance of all these new solutions should be studied urgently.

Conclusion

Above we have covered how biodiversity information could be addressed, represented, digitalised and registered. None of these services will alone scale up beyond demonstration, but when pieced together into a standardised and shared biodiversity information infrastructure, breakthrough is possible.

To summarise the priorities, also suggesting an order of importance, the following things should be achieved in near future: 1) XML namespaces for taxonomic and other biodiversity information should be standardised. 2) Central registries of this distributed content should be started. 3) Electronic publishing (online copies) of new organism descriptions should be facilitated. 4) Digitalisation of types should become the practice and museums should stop borrowing materials in any other way. 5) A distributed biodiversity addressing system, including .bio Internet top-level domain will have to be studied further.

When the above has been achieved, I believe biodiversity informatics will emerge as a new important science and will help taxonomy to gain ground again. Entomologists who are responsible of the largest chunk of biodiversity information have a key role here.

---

(33) http://habanero.nhm.ukans.edu/TSA/
(34) http://www.ebxml.org
(35) http://www.uddi.org/
(36) http://d23xapp2.cn.ibm.com/developerWorks/web/ws-ads/index_eng.shtml
(37) http://www.w3.org/TR/wsdl
(38) http://www.w3.org/TR/SOAP/

References

Albitz, P. & Liu, C. 2001. DNS and BIND, 4th Edition. 601 p. O'Reilly & Associates.

Cantino, P.D., Bryant, H.N., de Queiroz, K., Donoghue, M.J., Eriksson, T., Hillis, D.M. & Lee, M.S.Y. 1999. Species names in phylogenetic nomenclature. Systematic Biology 48(4): 790-807.

Dallwitz, M.J. 1980. A general system for coding taxonomic descriptions. Taxon 29, 41–46.

Dallwitz, M.J., Paine, T.A. & Zurcher, E.J. 1993 onwards. User's Guide to the DELTA System: a general system for processing taxonomic descriptions. 4th edition. http://biodiversity.uno.edu/delta/

Dallwitz, M.J., Paine, T.A. & Zurcher, E.J. 2001. Interactive identification using the Internet. Computers and Electronics in Agriculture (in print).

Grayum, M.H. & Hammel, B.E. 1996. The genus Tetranema (Scrophulariaceae) in Costa Rica, with two new species. Phytologia 79: 269-280. http://www.inbio.ac.cr/papers/Tetranema/tetpage.html

Groombridge, B. (Editor) 1992. Global biodiversity. World Conservation Monitoring Centre and Chapman & Hall, London.

Harold, E.R. 1999. XML Bible. 1015 p. Hungry Minds, Inc.

International Code of Zoological Nomenclature. 1999. 306 p. The International Trust for Zoological Nomenclature, c/o Natural History Museum, London. http://www.iczn.org/

Morris, R.A., Passell, M., Wan, J., Stevenson, R.D. & Haber, W. 2001. Engineering considerations for biodiversity software. Computers and Electronics in Agriculture (in print).

Negroponte, N. 1995. Being digital. 243 p. Albert A. Knopf Inc., New York.

OECD Megascience Forum. 1999. Working Group on Biological Informatics. Final report. 74 p. January 1999. http://www.oecd.org/dsti/sti/s_t/ms/prod/BIREPFIN.pdf

Saarenmaa, H. 1999. The Global Biodiversity Information Facility: Architectural and implementation issues. European Environment Agency, Technical Reports 34, 34 p. Copenhagen. http://www.eionet.eu.int/gbif/gbif-implementation-latest.html

Turnbull, H.W., Scott, J.F. & Hall, A.R. 1959. The correspondence of Isaac Newton Vol. I, 1661-1675. Cambridge University Press. http://www.newton.org.uk/books/

Väkevä, J., Perttunen, J., Saarenmaa, H. & Saarikko, J. 1996. A diagnostic information service about damaging agents of trees on the World-Wide Web. In: Kempf, A. & Saarenmaa, H. (Editors). Internet Applications and Electronic Information Resources in Forestry And Environmental Sciences. Workshop at European Forest Institute, Joensuu, Finland, 1-5 August 1995. EFI Proceedings 10: 129-137. http://www.efi.fi/