



FUNCTIONALITY FOR SATISFYING USER DEMAND

NICOLAS BAILLY

The purpose of this chapter is to consider the general specifications of the *ENHSIN* methodology. It forms part of the technical implementation and assessment of the Network and is related to Chapters 2, 3, and 7.

133

ASPECTS OF THE VISION

IMPROVING ACCESS TO SPECIMEN INFORMATION IN MUSEUMS ACROSS EUROPE

The data associated with specimens held in Natural Science Collections represent an enormous amount of information on past and present biodiversity. Until now, access to this information has been restricted to specialists, because it was stored on labels attached to specimens, in manuscript catalogues, and in manuscript or type-written card index systems. These sources were maintained by the curators inside the institutions holding the collections, and were accessible only through a written enquiry or a personal visit.

A small fraction of this information has also been published, mostly as scientific publications such as type catalogues and taxonomic revisions, and usually in specialised journals with a restricted dissemination. Thus, the information was (and, largely, still is) scattered throughout an abundant literature, which makes extensive analyses of large datasets difficult.

While personal visits to institutions have the advantage of allowing access to the specimens as well as the associated information, they are costly to undertake. Good database systems will, at the very least, allow a researcher to judge whether a visit will be worthwhile. Once information is transferred

to computer systems that can be interrogated remotely, the burden on curators of answering enquiries will be removed. This benefit, however, has to be balanced against the cost of keyboarding the information in the first place.

CATERING FOR A DIVERSITY OF USERS

With the coming of the informatics revolution, curators have, since the early sixties, become keen to computerise their collections. Their original objective was to ease the internal management of the lots and specimens and transactions such as external loans of specimens for research and exhibition. Collection data were considered as necessary adjuncts to the specimens, but were not considered to have an independent use. When the Internet was developed, the potential for wide dissemination of data became available, and the production of large datasets was stimulated. An outcome was that the data themselves suddenly acquired a value independent of the specimens to which they applied. Soon, not only the specialist but also the public at large were the target of this dissemination, and it became apparent that data associated with biological and mineralogical specimens could serve the needs of a number of sectors, including education, government (federal and local), environmental conservation and industry.

134

We are still in the process of moving gradually from a position of local (institutional) accessibility of scarce information by specialists, towards a situation where massive datasets are being made widely available to the public and other users. This chapter aims to anticipate the further progress of this (r)evolution and to point out some features, trends, and difficulties of the metamorphosis.

It should be noted that the expression 'digitise their collections' is sometimes interpreted by the public as meaning the production and dissemination of images of specimens. Although certainly an important part of the revolution, it is beyond the scope of *ENHSIN* to tackle collection imaging – and is a topic that will need to be addressed by other projects.

ACCESS CRITERIA

Five issues are considered under the topic of access criteria, the first three being more relevant to all users, the last two particularly to taxonomists. These are:

1. Controlled access to sensitive data, such as those on endangered species and exploitable natural resources.
2. Controlled access to imprecise, uncertain, and unreliable data, such as identification and location; also of importance here is the question of how to proceed with missing data and warn users.
3. Sustainability of the portal to collection databases: is it possible to establish a payment system, catering for different types of users? Although the trend is to encourage the free exchange of data, there are

more restrictions apparent in Europe compared with in the US (see Owens & Pryor, 2000, and Owens, this volume).

4. Special specimens: e.g. types, historically important material. It is a convention of the codes of nomenclature that types 'are the property of science', which implies that access to data about them should always be free to bona fide researchers.
5. Specimens in the process of being described as new taxa: a specialised taxonomic type of sensitive data.

Points 1, 3 and 5 address the type of data that are to be made accessible to whom, while points 2 and 4 refer to the means by which data are made accessible. By whom the decision to make data available is taken remains the authority of the data providers although the portal may be used secondarily as a filter.

SENSITIVE DATA

Field research on biodiversity provides huge amounts of information. This may reveal new data, especially from exploration of areas such as the deep sea, tropical forests, polar and sub-polar regions, and marine continental shelves; or during monitoring surveys, such as for natural resources stock or pollution assessments. Once the resulting voucher specimens have been deposited in museum collections, and the associated data recorded in collection databases, wide access to this information becomes possible for all. However, this very information may be exploited by companies or collectors, leading to threats to vulnerable populations and damage to the environment.

Information on collecting sites may be precise, allowing accurate location of individual populations of a given species, or even single individuals in the case of colonial aquatic animals or trees. Threatened and/or endemic species with restricted distribution and of commercial value, especially those that are particularly sought after by collectors (e.g. beetles, butterflies, shells and orchids), may be in further jeopardy if collectors have complete access to precise location data, not only through geographic co-ordinates, but also by detailed description of the exact site and the habitat (e.g. 'Under the bridge, Km 213 on the road from Rio de Janeiro to São Paulo, in the moss on the biggest rock.').

In the case of endangered and/or protected species, managers of biodiversity may also wish to restrict access to detailed locality data on populations, not just to protect them from collecting or harvesting, but to prevent too many members of the public from visiting areas where the species lives, particularly breeding sites.

For exploited and potentially exploitable species, commercial and industrial companies may use information to harvest populations soon after its publication on the web, and even before controlling legislation has been enacted and put into effect. Examples of populations that may be depleted in

a few years include deep-sea fishes such as Grenadiers (Macrouridae like the Roundnose Grenadier, *Coryphaenoides rupestris* Gunnerus) and Orange Roughy (*Hoplostethus atlanticus* Collett). Illegal fisheries activities may be facilitated by the dissemination of such information, e.g. on fishes, abalones, sea-cucumbers and other marine organisms.

Thus opening a portal on the web may require some attention to the type of data that should be made accessible to different types of users. At least, the creators of the portal should warn the collection database holders about possible misuse of their data, for, obviously, data access has to be managed first at the local database level. At the portal level, the database wrappers must include the possibility of managing accessibility flags that may operate at taxon, location or specimen level, or combinations of these. Depending on the structure of the database, flags may be defined in different tables in different ways (common access issue).

RAW DATA, INTERPRETATION, QUALITY ISSUES

Collection information represents raw data, whereas the interpretation of these data may require particular scientific skill and knowledge. Certainly, the direct use of raw data may lead to incorrect conclusions. For example:

136

- ◆ Depending on the complexity of the data structure and knowledge representation – and also on the level of controls provided in the software during the data entry – maps are perfectly capable of being produced from the uninterpreted collection data showing the presence of cod in Paris! In an actual case, further research showed that the specimen was bought in a fish market!
- ◆ Sometimes the absence of co-ordinates is interpreted by default as '0, 0'. This practice gives the apparent hot spot with the highest marine biodiversity as occurring in the Gulf of Guinea, precisely at the co-ordinates 0 degrees Latitude, 0 degrees Longitude!
- ◆ Endemic Antarctic species may be shown, falsely, as present in the Arctic as a result of a mistake during data entry: South and North are coded – and + respectively in many systems. Similar errors may arise from the confusion of East and West.
- ◆ Old records often come only with imprecise locality information (West, North, East Africa, '*Afrique Occidentale Française* (AOF)'), or archaic names that may cover several modern countries (e.g. Indochina, Yugoslavia, Russia), or worse, that may be confusing (the area named Sudan in the nineteenth century was actually Mauritania and a part of Senegal, not Sudan as it is known today).

While database owners should strive to clean their data, this is often a long, difficult and sometimes impossible process. We should view *ENHSIN* and

related collection database networks, not only in terms of a completed facility for information dissemination, but also as the process of building the network and as a mechanism for enhancing the quality of its basic components. It is naive to design systems on the assumption that they will contain only completely clean and verified data.

It follows, therefore, that users might reasonably expect any portal to provide an assessment of the quality of the data stored locally in the contributing collection databases (see Quality assessment, below). For example, how should it be indicated during the extraction of data that some records should be ignored? How might users be permitted to delete these records for certain purposes, and how might the delivery of such records to particular users be prevented? How, furthermore, is it possible to distinguish erroneous records, showing species as occurring outside their published distribution range, from those records that genuinely extend the range of a species?

Systems, at local and portal levels, need to provide features to tackle dynamic and evolving data with imprecise, uncertain, and unreliable characteristics. Ideally, the portal should cater for users of varying needs and skills, and provide the choice of excluding access to data that require interpretation or further research. Requirements range from providing full datasets to specialists, to secure and high quality data for the public at large, acknowledging that sectoral users may be able to manage data of intermediate quality.

Another important issue is how to handle specimens with no directly associated data or with incomplete data. Should records of these specimens be disseminated to all users? Should they be included in all analyses?

- ♦ Some localities are unknown, especially for older material. Comparing original expedition reports with such specimens sometimes allows us to gather data, but the process is time consuming. Nevertheless, if properly managed, specimens representing several taxa can often be dealt with concurrently.
- ♦ Lack of any identification at the level of species or above also constitutes missing data.
- ♦ A distribution map cannot be derived from records extracted from just one or a few collection databases. Thus conclusions drawn from analyses based on raw collection data may have to be severely restricted because of limitations of the component datasets. In effect, the problem here is one of missing data since it results from non-dedicated sampling.

A broad question is whether or not we should provide all data, or only those that have been interpreted and used to address certain research questions? Do raw observations represent valuable information in their own right, independently of any conceptual framework? Epistemologists have said no to the latter question for a long time in other frameworks, suggesting that all observations/data are theory-laden. If this view is accepted, it follows that we should filter data and knowledge before disseminating them to users. We

have to ensure that we deliver data to given users at a relevant level, or at least that we minimise the risk that data are incorrectly interpreted.

To take a recent example (September 2001), a species of Lionfish, belonging to the Indo-Pacific genus *Pterois*, was caught in Atlantic waters off the shores of Long Island (New York state, USA). From external knowledge of the natural distribution, it was established that an aquarium keeper released his specimen in Long Island waters. Without the interpretation, the specimen would be recorded as representing a species new to the US ichthyofauna.

FINANCIAL SUSTAINABILITY – FREELY AVAILABLE DATA VERSUS COST OF CAPTURE AND MAINTENANCE

Currently, among the scientific community there exists a wide preference, and often an obligation (see, for example, the National Science Foundation in the USA), to make data freely accessible, especially when they are assembled from public funds. In comparison, many literature databases are difficult to access for low-funded researchers because, since they are commercially managed, they are extremely expensive to request online or to buy on CD-ROMs. If one of the main objectives of collection information databases (and associated portals) is to disseminate data more widely, imposing a charging system may have the reverse effect.

138

Although difficult, it is currently possible to find funds to create databases. It remains, however, extremely difficult to obtain funds to maintain them and enhance their quality (even if database maintenance is an ongoing and essential activity). The recent establishment of the Global Biodiversity Information Facility (*GBIF*) may address this issue at a global level, but it is important to continue to explore the possibility of other funding streams.

In situations where the management and maintenance of these databases are intended to be financially self-sustainable (wholly or in part), different categories of users might have access under different fee levels, ranging from free to commercial rates.

Several categories of the user community (Table 1) were identified in the *ENHSIN* survey (see Calabuig *et al*, this volume), and terms of their access vary. For instance, researchers from public institutions might expect free access to collection databases, while researchers of the same skill level, but from privately funded or commercial institutions, might be expected to contribute financially.

The following user categories are modified from those given by Felinks *et al*. (2000: 19–25):

- ◆ Public research and education.
- ◆ Public services and Administration bodies (including NGOs, international research and education, organisations, environmental agencies and parks).
- ◆ Private research and industry.

- ♦ Commercial services (including entertainment) in environmental sector.
- ♦ General public.

Table 1.

	Organisation type
1	Administrative body
2	Botanical garden
3	Non-governmental organisation (NGO)
4	Other type of organisations
5	Private company
6	Private museum
7	Public museum
8	Research institution
9	University or other educational body
10	Zoos

Users willing to pay for access to specimen information are likely to be restricted mainly or exclusively to the private sector, although some others might accept low levels of charging. It is expected that access by specialists would be free. Charging for access by the general public is still a question without any clear answer. Other than the skilled amateurs who are often members of taxonomic learned societies, it is unclear to what extent the general public will actually wish to access such data. Further evaluation on the need of a payment system for the general public is clearly needed.

A further consideration is that, within a single institution, individuals should not necessarily expect equal levels of access. Those persons who, for example, have provided data would expect free access to the whole database to which they contributed. That means that an individual access system may be required.

The approach of classifying users of collection databases by interest allows us to define further categories.

Table 2.

Interests types
Exhibition/education
Taxonomic research
Nature conservation
Natural resources management
Industrial/commercial use of organisms
Other interests

Another issue is related to the extent and nature of the data downloaded. A simple text output on the screen may be made freely accessible, while downloading a large structured dataset might deserve to be charged. Curators will need also to determine which datasets are suitable for immediate access

through the web, and which might require further work, such as filtering data according to quality, creating or formatting outputs, and/or agreeing what might be delivered. Making just a simple output available to the public may avoid the problem of charging them, whereas commercial researchers, in carrying out further analyses, need complex requests returning large datasets as answers that could be charged.

The Western Australian Museum has developed a pricing system where, for some categories of user, there is a virtual 'charge' for access. Usage is recorded within the museum for monitoring purposes, but charges are not actually passed on to these users, to whom access remains free. This system allows a quantified case to be made to potential funding bodies as to the value of the collections and the data access service provided.

If a system of charging users for access via the portal is to be developed, a preliminary market study will need to be carried out. This activity itself leads to administrative work that should also be covered by the fees. Such a system does not provide net added value to the portal.

SPECIAL SPECIMENS: TYPE SPECIMENS, HISTORICAL MATERIAL

As name bearers, type specimens are fundamental to a stable taxonomic nomenclature. Data linked to these specimens have high research value, and the accessibility of these data, together with their quality, is a priority in taxonomic systems. As such, issues of charging would appear irrelevant. Nevertheless, certain data may remain sensitive.

It is recommended that database managers should provide free access to this information to specialists in museums, universities, and research institutes, other than for the restriction given in the next section.

WORK IN PROGRESS: SPECIMENS USED TO DESCRIBE NEW TAXA, AND DATA UNDERGOING REVISION

This topic applies to taxonomic research, when specimens are deposited as types of new taxa (species and infraspecific levels). Such research is undertaken mainly by specialists or skilled amateurs. Two situations need to be considered; the first is where the author has already informed the curator of the name of the new taxon, but where the name awaits publication. The second arises when the taxon has not been named, or when the specimens of the new taxon are still in the author's institution, but the author asks for a museum registration number.

This may lead to a conflicting situation. Collection managers are likely to prefer to database the information as soon as possible since they rely on the information system to manage the collection – from the receipt of a specimen to its final inclusion in the collection. Curators may be disinclined to release registration numbers, for taxonomists to include them in their publications until they are in possession of the specimens (and type specimens may have to be sent to journal referees for examination).

By contrast, the taxonomist is likely to prefer to wait until the publication appears before specimen information is released by curators, so as to discourage requests for the loan of specimens before publication. One possible solution to this problem is to allow the researcher a period of confidentiality, such as one or two years, before releasing the data so as to prevent specimens from being kept for decades awaiting descriptions. Although the problem might appear to be one for local management, the organiser of the portal should make stakeholders aware of the possible restriction.

The outcome of this section suggests that good practice in collection management is not independent of portal features or dissemination of collection data. As a rule, portals should define specifications or standards that candidate databases should conform to (in respect of data consistency, terminology and collection management) in order to be included in the network.

INTERFACES

Everybody has experienced websites with empty boxes to be completed with search terms to perform a request, and yet not knowing what to enter. For requests where users know what they are looking for, a form-based approach can work well. But it can be discouraging for more general users who have no precise search word, do not know what is contained in the database, or who wish simply to browse the database. In other instances, users may know that the information that they want is in the database, but may be unable to understand how to retrieve it.

One solution is to use pick-lists. Another approach is used in DB2WEB (<http://lis.snv.jussieu.fr/productions>), which is a type of navigation that does not rely on an assumed skill of the user to enter the request. Other systems have developed a tree approach similar to the Explorer component of Windows. Scalability can be a limitation of these types of systems, which are difficult to handle when the database reaches a certain size, not only for technical reasons, but also from a user point of view. It is difficult to manage several screens of results displaying large quantities of information.

USER EXPECTATIONS

DATABASE SELECTION. Users must be able to select databases to be searched according to different criteria, such as: taxonomic (e.g. Phanerogams); geographic origin of the specimen (e.g. from South America); location of the hosting institute (e.g. Scandinavia); and type of collection and/or institutes (e.g. living collections such as botanical garden). Any such filter should be considered, especially those created for the European collections survey during the *BioCISE* project.

SEARCH PARAMETERS. Any piece of information stored in collection databases should be treated as a possible search parameter, which means that, in effect, every field should be searchable. This is because, apart from the most used request of 'where and what', depicted by *BioCISE* and *ENHSIN* inquiries,

when a specimen is difficult to find, every small clue may be of use to specialists. The portal must answer both in 'breadth and depth', for finding a critical piece of information on a unique specimen (depth) may be for a specialist as valuable as extracting a large dataset (breadth) for another user.

The result is that at least both a simple request system and an advanced one must be provided for general users and specialists. The former must be designed from the results of the end-user inquiries, the latter must be designed, not as a pre-selection of parameters, but as a general philosophy for data navigation, which implies successive refinements, different request and navigating systems that are interconnected, and allowing the exploration of as many fields as possible.

ASSISTANCE WITH TERMINOLOGY. Glossaries must be established to facilitate user sessions. Possibly separate lists of terms by field should be displayed for each of the contributing databases (dictionaries), rather than having a single term-list serving the whole portal.

Explanations of technical taxonomic and nomenclatural terms should be provided, principally to users, not to aid in searching, but to explain the output. Such glossary terms could appear as hyperlinks in the HTML output, so that with a simple click, the user has the definition in a pop-up window.

142

LANGUAGE ISSUES

Issues concerning user interfaces, common access and interoperability need consideration against the background of the following paragraphs.

AFFECTING NAVIGATION AND CONTENT DISPLAYED. Most users prefer to request and to obtain information in their own language. The cultural richness within Europe becomes a restriction on the development of common information systems throughout the continent, especially if it is extended to the Urals. This situation contrasts with the relative linguistic unity within the USA and most of Canada, although multi-language initiatives to gather data from Canada to Mexico may change this situation (see *ITIS*, Integrated Taxonomic Information System, www.itis.usda.org). Moreover, if a world portal, rather than a European one, is an ultimate goal, language issues will become further pronounced.

The problems may be exemplified simply by the names of countries. If the portal is used to search for France, records may be recorded under Franca, Francia, Frankreich, Frankrike, etc. We might employ wild cards, like Fran*, but problems would remain for, say, the United Kingdom, Royaume-Uni, etc. One solution would be to use numeric codes for countries, such as the ISO standard (n°3166).

While the use of unique codes (alphabetic or numeric) would greatly facilitate searching for countries, thus easing the use of look-up lists for requests, this solution does present some disadvantages. It requires that all local databases that do not already record ISO country codes should slightly modify their structure, and retrospectively enter the country codes (or make the link between names and their codes). Archaic names of countries and names of

large areas are not covered by the standard. The interface must provide a means to select easily a given country by its local name, linked to an ISO code to perform the request. There is the further requirement that local data need a field for the codes, for otherwise, the users will need to know the relevant codes.

Another means of navigation is by means of multi-lingual thesauri. The request system is certainly more difficult to build, both with respect to the interface and the internal routines, but it avoids having to make modifications to the local databases. The most difficult part in that system is to avoid generating noise, i.e. outputting information not requested.

It would be impractical for each local database to maintain several fields in different languages, for there would be too many entries to make and fields to correct. So standards must be used and defined whenever possible, and tools to help local standardisation should be developed. Note that in many systems, for the same information there is a 'verbatim field' (i.e. an exact copy, even with misspellings, of the labels or the manuscript catalogues), and a standardised (shadow) field. The latter could be the one used for common access and interoperability.

Another issue concerns the use of accented letters. Accents are important for native-language readers. Sometimes, the word with or without an accent has a different meaning, and inaccuracy or omission may lead to confusion. A practical, and very real, problem may arise with differences in language-dedicated keyboards. As for requests, a good compromise is to use systems that (mainly through the use of indexing) allow the entry for a search to be made with or without accents. Thus, for example, requests with Sénégal or Senegal will deliver the same results. The general use of Unicode as a means of handling foreign characters needs further investigation.

143

AFFECTING INSTRUCTIONS/HELP INFORMATION. Help online, including contextual help, is a powerful aid for the user. But like the instructions, button names, links, and field names displayed in web pages, it must be in the user's native language to be most effective. For all these words and texts, it might seem desirable to create a static translation for each language, because changes are infrequent. Yet this would be a poor solution for two reasons. First, however infrequent, changes would be inevitable, which would mean incorporating them into the help texts, of all languages. Second, several versions of the portal would be required – one in each language. This might be a workable solution if it is decided that the portal should be mirrored in different countries. The problem is that if changes do not concern the text displayed, but rather technical features in the web page related, for example, to scripts, then all the pages in other language versions would have to be modified and tested as well. This would require, for adequate maintenance, a great deal of organisation, and a large set of predefined procedures.

The outcome of this section is that the only realistic solution at present is to use automatic translation software 'on the fly'. A good start may be to adapt the tool developed by the EU to translate official documents into eight European languages.

In fact, only a few fields in collection databases will present translation difficulties, once dedicated standardised dictionaries are developed. But those that may be natural language texts (description of location, habitat, and historical and biological remarks on specimens) are still difficult to tackle, especially because the proper names (location, persons) must be kept as they are and not translated. This could be achieved by the means of 'anti-dictionaries' (exclusion lists), or by tagging the words not to be translated, although this would present a huge task.

These automatic programmes are for the future, at least for the content with which we are concerned, together with the increasing use of XML.

OUTPUT

DISPLAY (WHAT GETS RETURNED AND HOW IS IT DISPLAYED). In terms of what should be extracted and displayed as outputs from a request, there are two, non-exclusive, possibilities. One way is to provide a standard and simple output with a minimum set of common collection information. The other is to let the user choose what information they want to be returned. Between these extremes, can be added several predefined output formats.

144

In terms of how the information should be displayed, long, non-structured lists may be difficult to read, but the fact that old specimens may present more blank fields than filled fields should be taken into consideration. Here again, predefined formats, and formats constructed by the user, may both be provided.

PRINTING/EXPORT. Another interface issue concerns scalability, and the questions as to the management of hundreds, even thousands, of records in an answer and in filtering information received from hundreds, or thousands, of databases. The Species Analyst initiative (<http://speciesanalyst.net/zportal/tsasimple.asp>) has begun to tackle the problem, but at present it seems to target mainly trained users.

When accessing information in databases, we may be seeking answers to a focused question, or require extraction from datasets for further analyses. For the former, and assuming that the number of answers is not too large, screen outputs in HTML format are the simplest means of access. Using client 'find text in page' facility allows the location of precise information or extraction of relevant material from within the output of a database search.

INTEGRATION WITH DESKTOP PACKAGES. The most common categories of software in use are word processors, spreadsheets and databases, and most of them have developed import/export facilities to allow sharing of information between them. However, using Comma Separated Values (CSV) text or equivalent does not allow the export of layout features such bold and italic characters. To develop dedicated import/export facilities would require a great deal of time and money, and would be possible only with the availability of substantial grant facilities.

XML. The most obvious first step to solving this problem would be to develop a full XML exportation facility. Advanced users could develop their own importation facility for particular software applications and, possibly, the portal might store them for use by the rest of the community. For less-advanced users, however, special efforts will be necessary to provide solutions to the issues discussed above.

GEOGRAPHIC ISSUES

SEARCHING. Some examples of problems with geographic names have already been mentioned. To help users match systems to needs, we might consider using maps to facilitate geographical requests, perhaps as an adjunct to other approaches. Such an example is explored in the Alexandria gazetteer project: (<http://fat-albert.alexandria.ucsb.edu:8827/gazetteer>).

When areas are well-defined, such as countries, administrative districts, islands, rivers and lakes, simple search facilities are usually adequate. But this is not the case for 'fuzzy' areas, such as mountains, plains, oceans, open seas, where standards are not yet defined or agreed, and which obviously need further work, even in specialised geographic domains. Moreover, allowing users to draw areas defining parameters for a search on a map seems not to be workable at present as it requires a world gis system (such as is used in the Alexandria project). Nevertheless, this restriction may be resolved in the future and could be the ultimate solution in that domain for the portal. Exploration and collaboration with geographers would be of benefit.

145

MAPPING. Displaying results on a map is not strictly a fundamental feature of the portal, the basic one being to export geo-referenced datasets subsequently to be exploited with dedicated cartography software packages. Yet mapping is certainly a facility often valued by users. At least as a first step, the use of free mapping software such as MapServer (<http://mapserver.gis.umn.edu/>) or similar applications that are easy to connect, can be recommended. Nevertheless, the access and the development of these software may be stopped at once, like Xeroxmap (<http://pubweb.parc.Xerox.com/map>).

Using XSLT (www.w3.org/TR/xslt) can help in the mapping process.

QUALITY ASSESSMENT

As noted in the first section of the chapter, a way of indicating quality is needed to allow the user to select the relevant data for a given problem, and to present the data in a way that matches the level of skill of the user. What also matters to the user is having confidence in the accuracy of data returned by a search – or at least to be shown that the different datasets contributing to the search results are of varying quality. Note that this data accuracy does not concern the quality of the specimen preservation and storage (which has been addressed by several evaluation systems), but of the quality of the information linked to the specimen record.

Quality is the sum of reliability and accuracy (including completeness) for several categories of information, for example:

- ♦ Names *versus* checklist (authority file).
- ♦ Identification credibility *versus* skill level of identifier.
- ♦ Location (co-ordinates, country, general area).
- ♦ Literature publication and linkage.
- ♦ Illustrations.

In general, an index to quality may be calculated if reference standard datasets are available, which is again a matter for standards groups such as *TDWG* or *CODATA* or *GBIF*.

Several attempts to assess quality have been made, but the subject remains at an early stage. One example is published in the framework of FishBase, called NIACC (Froese *et al*, 1999), a multiple digit index (example: 11141), each digit being related to one feature (identification, location, identifier, etc.) of which reliability is evaluated on a scale of five levels, from good (1) to unknown (5). But it is necessary to know the significance of each digit, and requires a training phase for non-specialist users the first time they request the database. Obviously, a one-digit index would be better with three to five levels (very good, good, etc.). But the value of such an index should be able to be calculated automatically from the data, perhaps with a niacc-like step for specialists. It should not have to be derived manually. In general, codes and abbreviations are to be avoided since, it has been pointed out, they may be meaningless to users from different disciplines or countries.

146

Another means of indicating quality is by providing the name of the person who last checked the record and the date on which it was checked. A limitation here is for sectoral users, who will rarely recognise the name of the specialist and their domain of competence. As such they will be unable to assess the quality of the data.

As often as possible, the origin of the data should be displayed explicitly, and, furthermore, there should be links to publications in which the specimen is cited as studied material. This action will help to avoid the portal being confused as the provider of data where it is, in fact, the tool used to retrieve them. That the portal guarantees the quality of data is a step further, and one that is not tackled by the present document.

Beside the assessment of the quality, it is necessary to think about ways to improve quality of the data. It can be first included in the good practice list. Then tools (or even better, web services) can be conceived to help collection database managers.

At the information level, software tools have to be developed to check data automatically against reference lists. This does not imply that specialists are no longer needed. Rather, it means that anything amenable to automatic correction should be dealt with by this route e.g. misspellings, date errors,

consistency errors (for example where an identification date precedes the capture date). At the access level, web interfaces have to be developed to allow users to report errors and send corrections. It is necessary to have a rigorous procedure in place to deal with and incorporate reported errors quickly. If a user fails to find that corrections have been made after a period of two months or so, they may well not send further corrections in future. This issue needs to be addressed at the level of data providers, but it is also one requiring good practice generally.

VERSION CONTROL

Although this issue concerns mainly the general management of metadata within the portal, it has implications for version control by the data provider. Conversely, the local version control has effects on portal features and functionalities.

Two issues affecting users must be addressed:

- ♦ A mechanism for providing updates if there is no direct link, to avoid consistency conflicts between the original living database, and a copy (cache) made at the portal level for technical access reason (speed for instance). The situation is the same in the case of mirror copies of databases (which might exist to deliver improved performance and access).
- ♦ How to cite a data source when results are to be used for published analyses.

147

It is necessary for users to be able to discover what has changed since they last looked at a dataset.

CONTINUOUSLY CHANGING DATA VERSUS PERIODIC CHANGES. Not all museums are able currently to run their own servers with a continuous connection to the Internet. Thus databases in such organisations will need to be copied to a server owned by another institute or by an Internet Service company. As for any distributed system, if the databases are kept separate, it should be possible to update individual databases as required. It will be more difficult to merge databases into a single one with more or less integration (i.e. with common tables). Nevertheless, to compare data in a distributed system requires effective management of the version information made available to the system by a local provider. For Internet-enabled databases, the version number, if any, will be managed locally, and the portal should extract this number at each request. However, the portal can hardly be expected to deliver a general version number on the data that are made accessible if local version updates occur throughout the year for all the component databases.

Systems dealing with collections data are being built to allow the querying of thousands of records, and for thousands of records to be returned in response. Given that the information in such databases has the capacity to

change on a daily basis, at least for records in large systems, results of analyses may differ slightly each time they are carried out. This can lead to tedious, time-consuming verifications and counter-verifications in the different datasets extracted.

The questions that arise, therefore, are: first, should different versions be backed up regularly and kept by all stakeholders of the databases?, and second, should the local database keep track of all modifications?

ALERTING USERS TO CHANGES AND ADDITIONS. More sophisticated systems might automatically notify users about record modification since their last querying of databases, but these may be difficult to maintain. This area is a desirable one for future research.

References

Felinks, B., Hahn, A., Olsvig-Whittaker & Los, W. 2000. Users and uses of biological collections. In Berendsohn, W.G. (Ed.), *BioCISE, Resource Identification for a Biological Collection Information Service in Europe*, pp. 19–32. Berlin.

Froese R., Bailly N., Coronado G.U., Pruvost P., Reyes R. & Hureau J.-C., 1999. A new procedure to clean up fish collection databases. In: *Proc. 5th Indo-Pac. Fish Conf.* (Sire J.-Y. & Séret B., Eds): 697–705. Paris: Société Française d'Ichtyologie (SFI), Institut de Recherche pour le Développement (IRD).

Owens, S.J. & Pryor, A. 2000. Common access to biological collection information and collection-holder's intellectual property rights – a contradiction? *Digitising Biological Collections (Taxonomic Working Group 16th Annual Meeting, Frankfurt)*. (November 2000.) See website.