

7

TECHNICAL PARAMETERS: HOW TO MAKE ENHSIN WORKABLE

CHARLES G. HUSSEY

102

INTRODUCTION

This chapter discusses some of the technical issues involved in meeting the aims of *ENHSIN*. The *ENHSIN* project sets out to create a shared interoperable infrastructure. To be a success the resulting network must be both usable and sustainable. That is, it must be easy for users to access, support a range of features to query and retrieve information that meet the needs of users, and deliver responses in an acceptable time. If it fails in any of these, it will not get used! It should also be able to be implemented and maintained without the use of extensive resources, either by the data provider or by a co-ordinating institution.

There are in fact a number of methodologies that could be employed to deliver the network. In order to understand which might be the most appropriate, it is first necessary to look at the nature of the datasets that might make up the network.

FIRST CHALLENGE: INTEGRATING DISPARATE SOURCES

Data compiled by museums from their natural sciences holdings take many forms. Not only will data dictionaries be different but also the data may be structured in various ways. Databases may exist as flat files, spreadsheets, relational tables, nested relational, hierarchic or object databases. Some of the larger museums maintain multiple systems, within different disciplines. Most databases will be actively maintained but some are 'legacy systems', perhaps compiled during a fixed-term project, and no longer being added to.

The different software packages used for the databases, whether single-user bespoke systems or full-blown commercial collections management systems, will vary in their ability to export data or accept odbc and sql queries. Both the database software and the underlying operating system will affect the accessibility of the data and may constrain the software and protocols that the network can use to interact with the datasets. The accessibility will also depend on the level and flexibility of security that can be provided both by the databases and operating systems.

A survey conducted in 2000 by the Natural History Museum of databases used by UK museums with natural history collections received replies from 87 institutions. A total of 33 different products have been used and, in addition, 33 institutions (38% of total) used bespoke systems that they had created themselves (or had created for them), mainly in 'desktop' database applications such as Microsoft Access, Filemaker Pro, Paradox and Cardbox. Five were storing specimen data in a spreadsheet application. Another fact to emerge from this survey is that most museums change their systems over time: *ENHSIN* will need to look at how to handle transitions between systems.

The *BioCISE* project also surveyed database systems in use in European institutions and found a very similar situation (www.bgbm.fu-berlin.de/BioCISE/Publications/Results/4.htm). They reported that 292 institutions that maintained collections databases were employing 60 different applications. Two-thirds of the applications were developed in-house. At the time of the survey in 1998/99, only 8% of institutions provided Internet access to unit-level collections data.

The technical demands upon data providers need to be taken into account. Do they have the resources to:

- i) Implement and maintain a local Internet Server providing 24-hour-a-day access.
- ii) Compile metadata (collections level or record level), such as that required to indicate quality of data.
- iii) Supply additional data (such as resolving localities to provide coordinate data, or providing elements of the higher taxonomy of specimens) to bring their datasets to an agreed standard.
- iv) Maintain the quality of their datasets, perhaps by employing controlled terminology.
- v) Construct views of their data, or wrappers (which might involve CORBA orbs or cgi scripting) to map their data to a common data dictionary.
- vi) Handle version control if a copy of their data is made periodically to an ISP or central node.

SECOND CHALLENGE: COMPARING LIKE WITH LIKE

For the network to allow successful searching across different datasets it will be necessary either to ensure that contributing datasets conform to precise rules regarding syntax and terminology; or to consult a thesaurus that is able to recognise and map between equivalences.

For instance, consider the following: a complete citation of the scientific name of an organism should include the authority (which for botanical species is abbreviated but for zoological species is not). In many specimen databases the authority may be missing for some or all names. If an authority comprises two authors it could be recorded in several variations: Smith and Jones, 1957; Smith et Jones, 1957; Smith & Jones, 1957. Specific epithets may be common to more than one species so searching using just a species name will yield records from different genera. Synonyms need to be mapped (museum specimens are often recorded under old names) and homonyms (particularly of genera) need to be recognised either through including an authority or name of the parent: e.g. *Morus* is both a gannet (Zool.: Sulidae) and a mulberry (Bot.: Moraceae).

Personal names can be stored in a single field (often in an inverted form) as e.g. 'Name: Smith, A.C., Prof.' or the elements may be separated, e.g. 'Surname: Smith, Initials: A.C., Title: Prof.'. Initials may be expanded to include forenames and the title may be given in full or abbreviated/contracted. And then there are such names to contend with as 'Barbara, Marchioness of Hastings', or 'The Maharajah of Kush Behar', which may have to be recorded verbatim.

The geographic co-ordinates (latitude and longitude) can also pose difficulties for storage and extraction, since they may be recorded as degrees, minutes and seconds, decimal degrees, or degrees and decimal minutes. Poles and Direction (North, South, East, West) are usually abbreviated, but South and East are represented as negative values when degrees are decimal. Co-ordinates can be stored in a single field or the elements can be separated (when stored in a single field, diverse characters may be used to denote the degree symbol). Specimens collected during cruises may have two co-ordinates recorded: for the start and the end of the trawl.

Name of places present numerous variations, both in syntax (North East Atlantic, North eastern Atlantic, N.E. Atlantic, North-east Atlantic), and in foreign names (United Kingdom, Royaume Uni, Reino Unido).

When seeking matches to search strings, one also has to take account of variants in spelling and language (including accented characters, transliteration, foreign scripts, multi-byte character encoding). In due course, Unicode values will help with foreign characters and scripts but otherwise this is best handled by creating thesauri that map between related words and terms. Thesauri can also deal with synonyms as related terms, and can order terms in a hierarchy of broader and narrower search terms. Names of higher taxonomic rank are often not recorded in collections systems (but are useful for searching) and because of both the effort and expertise required, are

unlikely to be created retrospectively by curators. Here, then, lies a challenge. A thesaurus is best built and maintained at a central point in the network. The obvious place would be as part of a common access system (CAS), but it would also be possible for the CAS to access a nameserver site elsewhere to validate terms and return related values. It would also be possible to produce a programme that checks for new terms as they appear in contributing databases and automatically add them into the central thesaurus. Wherever it may be constructed, much work is required to construct a thesaurus and an added complexity, as far as taxonomic names are concerned, is that alternate classification schemes are current for some groups. A consensus on terminology would be of great benefit and help and should be encouraged among the data providers and, through the work of standards groups such as the Taxonomic Database Working Group (www.tdwg.org) and CODATA (www.bgbm.fu-berlin.de/TDWG/codata/), among the wider community.

EXPANDING THE PILOT

The pilot common access system produced as part of the *ENHSIN* project must be viewed as being merely a proof of concept. It was originally hoped that two iterations would be possible but this was excluded when funding was reduced. It was therefore necessary to be entirely pragmatic and produce a working system within the available resources, and not indulge in a research exercise to explore different options. The pilot had to be produced within 50 working days.

The selection of data sources to contribute to the pilot was determined as much by practical as technical requirements. The political aspect of obtaining agreement from data owners was perhaps the main determiner. At a technical level, the ease of providing connection was important as it was seen not to be feasible for members of the Technical Implementation Group to visit individual data providers to assist in connecting up their databases. Data providers were selected who could build views of their data using sql. The element set defined for use in the pilot does, however, provide more general elements to accommodate data from systems that do not have atomic fields.

One possible disadvantage of the approach adopted for the pilot is that it required Microsoft technology (NT server + IIS + ASP) to be used both for the CAS and by the providers. Many data providers may have placed their data on servers with UNIX or LINUX operating systems and there are known security issues with ASP.

Some constraints within the element set are already becoming apparent; for instance the altitude element needs to be expanded to handle ranges and the way that geographical co-ordinates are recorded may make it difficult to analyse data extracted from *ENHSIN* in Geographic Information and mapping systems. It may also be desirable to specify a precise syntax to be used for personal names.

The initial indications are that performance is adequate although it is not possible to properly test scalability with the amount of data available to the pilot.

It can be assumed that a fully operational network will provide additional features. These should be determined following analysis of user requirements. They will need to be developed and tested before an operational network goes live. There is no doubt that other databases can be located that could be wrapped to the data dictionary of the pilot and lead to an expanded network within a relatively short space of time. However, a large-scale sustainable network might require additional or alternative access methods, as discussed in the next section.

ARCHITECTURES

The pilot produced as part of the *ENHSIN* project allowed the exploration of some of the issues involved in setting up a network, but time and resources permitted only a single solution to be developed. In fact there are a number of ways that shared access to datasets may be organised. It would be possible to develop a network using any of the following architectures, each of which has its own advantages and disadvantages.

POSSIBLE APPROACHES

106

1) SINGLE CLIENT/SERVER DATABASE USED BY ALL PROVIDERS AND USERS. In this model, institutions contributing to the network would be required to use common software. A single database would be sited on a server at a central institution with clients installed at provider institutions. General users would query the database using an Internet browser interface, but high-need, paying customers could also be supplied with suitable client software.

At first sight, this approach may seem unwieldy and costly, as a special application would need to be built. However, given that so many institutions appear to be using what should be regarded as interim solutions for databasing their collections, the availability of a fully featured collections management package might be welcomed. If momentum could be built up within the museum community, this solution would offer a high degree of standardisation together with benefits such as shared authority files.

This model might be the one of choice in an environment where many users were expected to contribute data and verify data – such as a network for recording field observations.

This model might also be adopted if surveys of user needs indicated that a more detailed interaction with data was required than could be readily delivered to a user's web browser. The specially written client software that could be distributed to selected users could allow sophisticated querying and reporting functions, such as sorting on grouped fields. The client could also provide an interface to the data, identical or similar to that used by staff at the contributing institutions, which would allow sight of the data in full detail (which, as mentioned in the previous section, might comprise over 400 fields).

A variant of this would be where providers each had a copy of the server database software, but posted their data to a central server for public access. This would give participating institutions greater ownership of their data.

2) **CENTRAL SUMMARY SYSTEM.** This differs from the first approach, in that contributing institutions maintain their own (different) collections management systems and send copies of core data to a centrally maintained database that services Internet queries. There is no attempt to connect to (and query directly) the providers' databases. The onus is on the providers to export their data and, if necessary, transform them to comply with the schema of the central database. This approach is the least demanding technically. However, providers would require mechanisms to track versioning so that, if new records were added to their databases, or if existing records were amended, these could be identified and copied to the central system. This approach is called a summary system because it is assumed that, in order to be able to accept data from many different sources, only a subset of core fields (common to all systems) would be maintained in the central database. It would be possible for the central system to provide onward links (at a record level) to the providers' databases – provided that these were web-enabled.

3) **CENTRAL GATEWAY TO DISTRIBUTED DATABASES.** In this approach there is no central database. There is, however, a centrally maintained Common Access System (CAS), which sends queries to the providers' separate databases and collates and displays the data returned in response to the query. Technically, this approach is more challenging since wrappers must be constructed for each contributing database to relate the query to the underlying data structure and return a response in a form that the CAS can handle. The CAS may be given additional functionality – for instance, it could store metadata describing the contents of the source databases; which would make it possible to direct queries only to sources that contain records for the group of organism requested.

Again, there is a possible variant to this approach. In this case, rather than require wrappers to ensure the provider databases conform to the requirements of the CAS, the CAS would store metadata on each contributing data source to know how to form queries that comply with the requirements of the data sources. Here, the CAS is limited to directing queries to data sources that already mount direct Internet access.

4) **PEER-TO-PEER DATABASES (MULTIPLE Z39.50 SERVERS AND CLIENTS).** Here, providers could, potentially, have access to other providers' data. Each provider's database would be a Z39.50 server and they could also act as Z39.50 clients. As in the first approach, it should be possible for providers to work with shared authority files. There would probably also be a separate site acting as a Z39.50 client to provide public access to all participating databases, although, in theory, the network could be set up so that each node

could act as a public gateway to other nodes. Z39.50 is a client/server-based protocol for information retrieval in a distributed environment (further discussed in Sections 5.3 and 6 below). Access may be limited to databases whose vendors provide a Z39.50 connection.

5) **WEB DIRECTORY POINTING TO DATA SOURCES.** This is essentially a portal: a resource discovery service that categorises the data sources and provides a link to the providers' websites. However, if it contained sufficient metadata about the data sources, and maintained authority files and an efficient navigation system, it would be possible to provide direct links, at a unit level, to the distributed data sources. This approach might be regarded as a fallback position, allowing users to identify, and make contact with, data sources that (for various technical or political reasons) it was not possible to connect to in any other way.

ISSUES TO BE ADDRESSED

The decision on which of the above approaches (or indeed, a combination of approaches) should be employed in setting up a network will be influenced by a number of issues, some of which are discussed here. Again, we need to keep in mind the likely nature of the available data sources.

108

- ♦ Wherever possible, solutions should be vendor independent, comply with existing standards and be platform independent. For instance, CORBA products from different vendors exhibit incompatibilities (Xu *et al.*, 2001) and Java code is often browser specific. Solutions should be able to work against data sources mounted on Windows, Unix and Linux operating systems and client software should support both Windows and Apple Mac users.
- ♦ In building a system that will communicate with a whole range of types of data source, there is a danger that it will be constrained by the lowest common denominator – so that one ends up with a system that supports only the most basic functionality. User needs will determine how many of the features of a full Collections Management System will be required to be supported by the network. Provision of features must be balanced against complexity of the system. The *BioCISE* user survey reports that users want many features and high quality, complete and up-to-date data. In reality, it will not be possible to satisfy all these requirements – principally due to gaps in recorded data.
- ♦ Where processing is to be done: (a) on the server holding the source data, (b) by the Common Access System, or (c) by the client (browser). This decision will be influenced by relative computing power at each site, and network bandwidth. In a traditional client/server system, it is more efficient to carry out processing on the server since this is usually

a more powerful machine than the client workstation, and this also serves to reduce the amount of data sent over the network. However, when a server is running an Internet service then it is best configured as a file server, rather than as an application server, particularly when a great deal of Internet requests need to be handled: this would indicate that processing should be passed to the client. Since the power of modern desktop computers has increased dramatically in the last few years, it is now quite feasible to shift much of the processing load to the client. Where processing is carried out on the client side then applets or scripts will need to be passed to, and probably stored on, the users' computers. Potentially, large amounts of data would also need to be passed for processing on the clients. With processing handled at the data source or by the CAS, only screen dumps (HTTP or XML pages) would need to be sent to the users. When XML (www.w3.org/XML/) is transmitted, it may be more efficient to first compress it, in order to reduce Internet traffic, but this requires data provider and user to have compatible compression software. It may be expected that the bandwidth available to institutional users will grow significantly in the future.

- ♦ As the network grows, issues of scalability will need to be addressed. Firstly there is the ease (or otherwise) of adding additional data sources to the network. This could impact on the need to grow authority files and possibly also require structural changes if new data sets bring in new data elements. Scaling up to handle large amounts of data will depend both on the number of data sources that are available for querying, and on the number of records that might be returned. Both can affect performance, but they will also impact on the user interface, particularly in the displaying of results. It then becomes beneficial to resort to caching information in the CAS, both in terms of metadata that determines which data sources are relevant to a given query, and in terms of data held for further transaction processing (rather than re-transmitting a query). Concurrency in querying (processing queries in parallel rather than serially) will also help to speed up response times. Where metadata is held centrally, this raises the issue of version control and, potentially, there may be associated IPR issues.
- ♦ Few systems managers would permit public Internet access to operational Collections Management Systems. Although there are Library systems (notably *OPACS* – Online Public Access catalogues) that are designed to operate on both sides of a firewall, the solution of choice will be to make a copy of the original data source that can be treated as 'sacrificial data'. This mirror of the original data will be mounted on a server outside of an organisation's firewall. It should be borne in mind that whenever a copy is made of the data, particularly when this is done by reporting or export tools (rather than through a

straight copying of files), there is the possibility of transforming the structure of the data to conform with requirements of the CAS.

- ◆ Ideally, every data provider should be able to offer 24-hour, seven day per week, access. As things stand at present, many will not be able to do so. Indeed, many collections databases are not web-enabled and are designed for internal use only. While Internet Service Providers (isps) generally provide for storage of files, it may be more difficult to find isps who will support databasing activities. In such cases it may be best for a data provider to copy their data to another partner in *ENHSIN* who is better equipped to provide Internet access. This could be a National Node, as proposed by the *BioCASE* project. Again, in copying data there is the possibility of transforming it by importing it into a different (perhaps central) database system.
- ◆ In architectures where data, or metadata, is copied from a provider's database there must be a mechanism to keep the copy abreast of additions or emendations made to the original. The copy could be kept up to date either by total replacement or by selective editing. In either case there should be a means of marking changed records with the date of the update. Some classes of user might wish to see previous versions of changed items: this would require maintaining an audit trail. Version tracking may be handled better by marking updates with a generation (edition) number rather than using date- and time-stamping, because of time differences across Europe.
- ◆ In any case, users would need to be alerted both to new records and changed records. To indicate to a user changes that have occurred since their last visit would require some sort of user profiling, which in turn would require the user to log in and/or the use of 'cookies'. An alternative would be to alert users to updates by email.
- ◆ The choice of architecture will also be influenced by the management model. Such as the resources required to sustain a CAS or National Node.
- ◆ It is useful, and in some cases important, for institutions providing data sources to receive proper acknowledgement of their contribution. Logos from the contributors can be positioned against returned records and, if the logo image files are stored on the contributor's server, a hit against their site will be recorded every time a query references their records. A legitimate concern of contributors is that *ENHSIN* will be cited as the source of data rather than the contributors themselves, who are the owners of the data. Copyright ownership and IPR must be made explicit to users.
- ◆ Another way to help contributors is to provide regular statistics and other feedback on who is using the network and the amount of traffic

that is generated. Although contributors can monitor calls to their own databases, the CAS should also have monitoring software installed.

- ♦ Whatever architecture is employed, it must recover from failures gracefully. If no hits are returned, or the CAS fails to connect with a data source, or if a user or provider closes a session abnormally, then the user should be alerted, but should be able to continue working. In other words, the network must be fault-tolerant and robust. Mirror copies of data sources may be maintained in the case of persistent connection problems.
- ♦ Metadata has already been mentioned, mainly with reference to information needed by the CAS to perform efficiently. Metadata should also be provided to aid in Resource Discovery. Search engines and agents need to identify content. Dublin Core can be used to good effect but cannot provide more than high-level terms. It is not obvious how comprehensive details of taxa, localities and people that can be searched for using the system, can be communicated to search engines. One way to advertise the capabilities of the network is to obtain entries on relevant link sites, particularly portal sites, such as biome (<http://biome.ac.uk>).

OTHER PROJECTS

111

A number of initiatives already employ distributed querying of biological databases. The following list is not exhaustive, but is illustrative of some of the different architectures mentioned earlier.

SPECIES ANALYST. The Species Analyst (<http://habanero.nhm.ukans.edu>) is a research project developing standards and software tools for access to the world's natural history collection and observation databases. It is based at the University of Kansas Natural History Museum and Biodiversity Research Center. According to the classification of architectures adopted above, it is a Type 4 with CAS.

The Species Analyst has served as an excellent demonstrator of how to retrieve data from distributed data sources, not only for display but also for analysis. It has been innovative in approach and it is to be hoped that a worthwhile network will become established.

The techniques employed to query distributed databases have relied heavily upon the fusion of the ansi/niso Z39.50 standard for information retrieval (ISO 23950) and xml. Z39.50 was considered to provide an excellent framework for distributed query and retrieval of information both within and across information domains although it was also thought that its use was restrictive because of the somewhat obscure nature of its implementation. All the tools used by the Species Analyst transform Z39.50 result sets into an xml format for further processing, either for viewing or data extraction. There was a conscious decision to adopt a standards-based approach.

The project has gone on to produce a Z39.50 client (known as zx), which is a Z39.50 protocol handler for Internet Explorer version 5.0 and later. It provides a mechanism for a full Z39.50 search and retrieval to be specified in a single url.

Other tools developed by the project include:

- ♦ ZASP, which is a set of asp pages that offer an alternative to a Z39.50 server for providing access to natural history specimens and observation data. The pages generate XML output in a format expected by Species Analyst clients in response to a pqn encoded query, IIS4.0 or later, or MS.
- ♦ The ZPortal is a set of Active Server Pages that provides a general purpose HTTP-Z39.50 gateway that provides access to Z39.50 resources and encapsulates the results in an XML document of the same format produced by ZX.

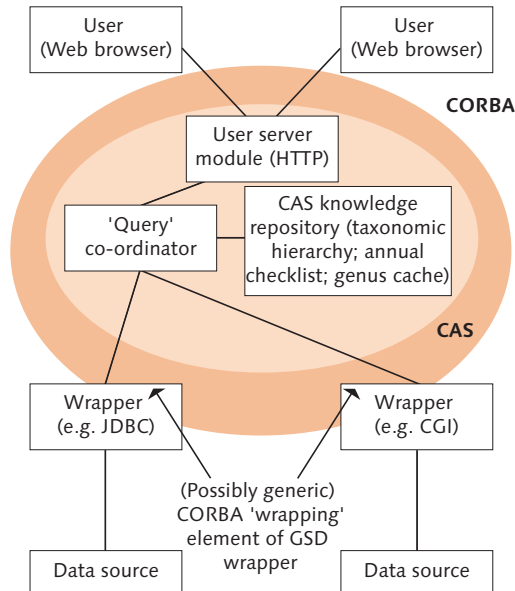
One very useful feature provided by Species Analyst is the ability to return data from multiple databases as a single Microsoft Excel spreadsheet. A mapping function has also been demonstrated, although this involves a real-time link to a third-party site.

112

ARTS & HUMANITIES DATA SERVICE (AHDS). The AHDS Gateway (http://prospero.ahds.ac.uk:8080/ahds_live/) is given as another example of Z39.50 usage. The Gateway, physically based in London, submits queries to five totally different databases containing information on archaeology (York), history (Colchester), the performing arts (Glasgow), the visual arts (Newcastle), and textual studies (Oxford). The databases describe different data types according to different cataloguing standards. They are driven by different database management software and run on a variety of hardware platforms. Nevertheless, despite these differences, the combination of Z39.50 and Dublin Core metadata elements is sufficient to enable meaningful searches across the five sites.

SPECIES 2000. Species 2000 (www.sp2000.org/) will deliver a Common Access System allowing users to query a large number of distributed databases that hold global species databases, each one covering a discrete group of organisms. Seven core fields have been defined, although individual databases may contain (and deliver) richer data. The programme is essentially still in a prototype stage and a research project (SPICE for Species 2000, www.systematics.reading.ac.uk/spice/) is investigating the methodologies and technologies that will underpin the delivery of the service. This project has already produced much useful information that is relevant to *ENHSIN*, as it sets out to investigate many of the same issues. The SPICE project involves the universities of Reading, Southampton and Cardiff, with test data being supplied by the Royal Botanic Gardens, Kew, and the Natural History Museum, London.

Figure 1. The SPICE architecture.

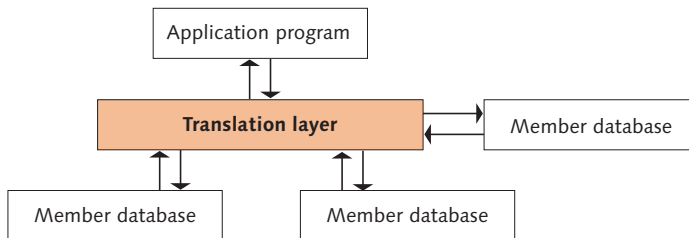


Species 2000 will work by building a single browser-based interface that will post queries to remote databases that could vary greatly in structure and software. SPICE has adopted a distributed object approach rather than a gateway-based approach (such as Z39.50 systems would use); it conforms to the third type of architecture see p.107, above. A Common Data Model (of only seven defined fields) has been developed and this defines what information may be searched (and returned) in the contributing databases. The general SPICE architecture (Fig. 1) is based upon CORBA (Common Object Request Broker Architecture) and all communication between the SPICE Common Access System and the individual database wrappers must comply with this standard. CORBA was chosen as the means of providing interoperation across platforms and databases. Each contributing database must be wrapped to conform to the Common Data Model. In the experimental environment of the SPICE project, two types of database wrapper were investigated; one used full CORBA-based wrappers and the other was based upon the CGI/XML protocol. The CORBA client object communicates with server objects using the Internet Inter-ORB Protocol (IIOP). Where client-side firewalls only enable outgoing communication via the HTTP protocol, HTTP tunnelling is performed where the IIOP packet is transformed into HTTP so that it may pass through the firewall. In the CGI/XML approach, a CGI form is sent by the CAS to the source database and the database wrapper returns an XML document. The XML Document is generated with reference to a document type definition (DTD). The XML document is then processed and localised by the CAS as CORBA objects. Wrapper building will be under the control of the data providers, although a wrapper-building toolkit is under development.

A characteristic of the CAS developed for the SPICE project is that it performs functions other than just passing queries to source databases. A knowledge repository caches metadata, which, among other things, allows the CAS to effectively route queries only to databases covering the group requested. The knowledge repository also includes internal databases that will hold a copy of an annual checklist (that can act as a mirror for providers' databases that are currently offline), and a taxonomic hierarchy (that will act as a navigational aid).

BIODIVERSITY ON THE WEB (www.biodiversity.org.uk/ibs/; www.biodiversity.org.uk/ibs/globalsearcher/nojava/research.htm). This is a project run by the University of East London to publish their fossil records database. However it also provides facilities for searching and integrating data from databases available on the Internet. The model conforms to the variant on the type 3 architecture mentioned above.

Figure 2. The multiple database application model.



In the multiple database application model implemented in this project (Fig. 2), each member database is registered to the translation layer with its metadata information, and the translation layer translates a global query to a set of native queries according to the metadata of each member database. The translation layer also receives the data returned from the member databases in response to each native query and again references the metadata to integrate the results. Since the results returned are mostly in HTML/XML, a multi-state parser (controlled by the metadata) is employed to break the returned content into records and fields. In the model proposed, database providers can register (and de-register) their database with the translation layer. The process of translating a global query into a native query depends entirely on the metadata information provided by the member database. The implication here is that contributing databases already have their own web interfaces and it is the information on how these operate that is included in the metadata. The databases remain autonomous and no wrapper needs to be installed by the data provider.

All the work of interacting with the contributing databases is handled within the CAS, but only once data has been returned to the CAS in response to a query. The project has also produced applications that, by means of

applets, can be passed to the user's workstation. These integrate the returned data for further processing and analysis. One such application is an impressive Global Plotter, which is a multiple database application program that can be used to plot and manipulate geographical data.

BIODIV. Biodiversity Resources in Belgium (www.br.fgov.be/biodiv/) is a far-seeing and highly developed initiative of the Belgian Federal Office for Scientific, Technical and Cultural Affairs, with the objective of delivering an inventory of resources concerning biodiversity research in Belgium. This includes projects, programmes, research institutes, biological collections, surveys, and experts. It conforms to a type 5 architecture (p.108). A website gives access to catalogues of collections and collection information systems. For some of the collections cited, it is possible to query unit-level information. This has been handled in a variety of ways. Where collections mentioned already maintain an online database, the entry in *BIODIV* will either provide a hyperlink to the provider's site or a search box that will pass a query direct to the provider database. In some cases, copies of data are held on the *BIODIV* server, either as a searchable database or as lists to the *BIODIV* server. This hybrid approach is a good example of how an overview of resources can be provided through collections level indexing together with direct links when available. It does not, however, permit simultaneous querying of multiple databases.

115

PALAEONTOLOGY COLLECTIONS MANAGEMENT SYSTEM AT THE NATURAL HISTORY MUSEUM, LONDON. This has been included simply because it illustrates the feasibility of the first type of architecture. It has been developed as a client/server system, using a nested-relational database at the back-end and clients running on desktop computers that communicate with the server via a telnet session (but with a Microsoft Windows interface). Connections between client and server have been tested offsite and shown to perform well. It is a 'thin client' system, with only data needed to populate a screen being passed across the network. Whenever a client starts a session, the version of each screen loaded is checked against the version stored on the server and, if the server version is later, it is automatically downloaded to the client. This allows client installations to be automatically updated. The bespoke application was written for Unidata database software in the SB+ application language. Both products are marketed by ibm. Another product, called Redback, could be used to produce a Web interface to the application (not, as yet, implemented in this application).

NATIONAL BIODIVERSITY NETWORK. The National Biodiversity Network (NBN, www.ukbiodiversity.net/) is a federation of bodies that hold data on the biodiversity of the uk. There is a large biological recording community in the uk, and one of the services delivered by the NBN is a software package called Recorder 2000 that is designed to collate field observations, together with associated locality, biotope, taxonomic and bibliographic data. The relevance to

the issues being discussed is that this represents a model of how a scientific community is agreeing to use a single software package (as proposed in the first architecture in Section 5). Recorder 2000 has several features that assist in its deployment in a network. While there was a significant cost in developing the software, it is being sold at a very affordable price. Technical support for users of the product has been well organised. A data exchange mechanism, employing XML is provided. Records in authority files (taxonomy, biotope, locality) are identified with codes, which distinguish the source of the record. Thus if a user adds a new name to a taxonomic list in their copy of Recorder 2000 and, later, this name gets incorporated in regional or national lists, then it can be traced to a particular copy of Recorder 2000.

DATA RETRIEVAL

Having discussed possible architectures, the next two sections focus on the transport mechanisms (protocols) suitable for delivering data and on how the user interacts with the system. It has been assumed that the methodology adopted will be based upon having a common data model with defined fields. Contributing data sources would either build views of their data to conform to the model, or would build wrappers to interpret their data, or would copy and transform their data.

116

Interaction with remote databases and the passing of data across the network is underpinned by protocols and transport mechanisms. Some candidate protocols are outlined below. Some of the protocols are still emerging as standards and a choice must be made on whether to adopt tried and tested technology, or whether to guess which are likely to become the method of choice. There is, however, no doubt that XML will come to play a major role in dealing with the sort of rich data held in collections systems, whilst HTTP remains the transport mechanism of choice. Most of the protocols still rely on the source databases complying with SQL (Structured Query Language) – in all its variants – and ODBC (Open Database Connectivity) standards. Each of the protocols mentioned could be used within *ENHSIN*. Whatever protocol is chosen, it must effectively handle session management and error handling.

Z39.50. Z39.50 (Z39.50 Standard, www.loc.gov/z3950/agency/); Miller, 1999; Z39.50 in Europe, www.ukoln.ac.uk/dlis/z3950/) is an ANSI/ISO standard for computer to computer information retrieval that allows client applications to query databases on remote servers, retrieve results, and perform other retrieval related functions. The protocol is stateful and connection-oriented. The standard includes a query language, record syntax options for transferring data records, a language for constructing records to be transferred, and several defined query types (Lynch, 1997). It is widely used in the Library Community and by the Government Information Locator Service (GILS) in the USA (www.gils.net/index.html). Z39.50 operates against 'profiles' that support a particular application or community. Examples of profiles are the Bath Profile

for interoperability between library and cross-domain applications and Darwin Core (http://tsadev.speciesanalyst.net/DarwinCore/darwin_core.asp); developed by the Z39.50 Biology Implementers Group for specimen and species data classes. Z39.50 is particularly good at utilising metadata.

CORBA. The Common Object Request Broker Architecture (CORBA, cgi.omg.org/library/corbaiop.html) is an emerging open distributed object-computing infrastructure being standardised by the Object Management Group (which comprises about 800 companies). Other standards for distributed objects are DCOM (which is specific to Microsoft) and RMI (which is specific to Java), whereas CORBA provides interoperation across hardware, operating systems and programming languages. In practical terms, when building an *ENHSIN* network, CORBA could be useful middleware, but could also make demands on the data providers who would have to build wrappers to their databases. The SPICE project has encountered problems in interoperability of ORB (Object Request Broker) software supplied by different vendors, which typically only implement a subset of the CORBA standard.

CGI. Common Gateway Interface (CGI) is the original method for server side programming and is identified in the Hypertext Transfer Protocol (HTTP) specification. CGI programs provided a relatively simple way to create a web application that accepts user input, queries a database, and returns some results back to the browser. A CGI program can be written in just about any language, C++ and Perl (Practical Extraction and Report Language) probably being the most popular. In a web application infrastructure, the web server plays the role of a gateway between the web browser and the CGI application. The biggest disadvantage of CGI programming is that it does not scale well.

117

JSP (JAVA SERVER PAGES). Java Server Pages are an alternative to CGI for developing interactive web pages. JSP makes it possible to embed Java code fragments into special html tags. A JSP engine automatically creates, compiles, and executes servlets to implement the behaviour of the combined HTML and embedded Java code. It can communicate with any ODBC (or JDBC) compliant database through the JDBC API. Servlets facilitate the development of database applications by providing a set of useful mechanisms such as connection pools, session objects to keep state, and request/response model of HTTP. Database access with JSP is achieved either by writing specific code, or by using one of the available database connectivity tags.

JSP is attractive as it provides a more flexible solution than Microsoft ASP (Active Server Pages). While JSP technology is designed to be both platform and server independent, ASP, because it uses ActiveX controls for its components, cannot work on platforms other than Windows, or web servers other than Microsoft IIS & PWS, without third-party porting products.

SOAP (SIMPLE OBJECT ACCESS PROTOCOL). SOAP (current version 1.2 is a W3C working draft, www.w3.org/TR/soap12/) is a lightweight protocol for exchange of information in a decentralised, distributed environment. It is an xml-based protocol that consists of four parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined data types, a convention for representing remote procedure calls and responses, and a binding convention for exchanging messages using an underlying protocol. SOAP can potentially be used in combination with a variety of other protocols; although it will mainly be used in combination with HTTP. SOAP therefore provides a way of transmitting XML over HTTP. SOAP offers several benefits over a proprietary xml vocabulary, as SOAP is an open standard with a growing body of developers and vendors supporting it. As more vendors offer SOAP products and services, the advantages of using SOAP will become more pronounced and it may well turn out to be the preferred protocol for an *ENHSIN* network.

USER INTERFACE – TECHNICAL ISSUES

USER EXPECTATION

118

The features to be provided in the user interface to the CAS will (and should) respond to user requirements, and these may be expected to evolve over time. Taking the prime criterion as 'usability', the system must provide a simple and intuitive way to navigate through the features, and the results sets returned by queries. The CAS will also have to cope with scaling up as more and more databases become available.

There is a problem in matching user expectations (they appear to want so much!) with the current variable state of specimen databases. Undoubtedly, it will be best to keep the system as simple as possible to start with: if it is too difficult or too slow to respond, then it will not get used.

It will be essential to list the different contributing datasets, together with information (metadata) describing the content of each dataset, and administrative details concerning the provider institution. This amounts to collection level descriptions and could be expanded to cover other known collections: marking the ones that may be queried through *ENHSIN*, those that provide their own Internet access, and those which may only be accessed by a personal visit. While *ENHSIN* exists to prove the ability to query distributed specimen databases directly, there is no reason why an operational version of the *ENHSIN* network, should not also provide this collections level information (such as the *BIODIV* project does) and, indeed also supply interpreted data and general biological facts (as requested in the user survey). It must however be clear to users when they are searching for datasets and when they are searching for information contained within datasets.

There should, ideally, be a way of tailoring information to the user's needs. This would be achieved through a mixture of deploying 'cookies' and having

the user register and log onto the site. Besides assisting with customising information, logging in would also facilitate access control (which is covered below, p.127) and evaluation of use of the site. It is not envisaged that the site is set up, at least initially, to provide a chargeable service. If a login system is instituted, there should still be a basic service offered to all-comers. A well-organised site providing user registration can be found at Florabase established by the Western Australian Herbarium (florabase.calm.wa.gov.au/).

As indicated under 'Architectures' above, there should be a 'What's New' Section (which could be individualised if there was user login).

SEARCH FACILITIES

There is a balance to be struck between empowering the user, giving them freedom to search at will, or constraining them to available choices and presenting default options. Once again, the guiding principle should be simplicity. The key to successful searching is knowing what there is to search (having knowledge of the content of data sources) and having a good online help system to guide and prompt at relevant points. Some degree of automation is possible (such as the 'intelligent' CAS in the SPICE project that 'knows' to which data sources to direct a taxonomic query).

The most basic query form presents the user with empty boxes in which the user enters the text string (or partial string) to search on. Clear instruction must be given on the scope of the search: whether the programme can only handle an exact match, whether a partial string is acceptable and, if so, whether this is taken to be a 'starts with' or 'contains' and whether partial strings need to be terminated with wildcards. Worked examples will help with this. 'Fuzzy' searching may be employed to widen the scope and trap terms that differ only slightly (such as through transliteration or with accented/non-accented characters).

Misspelled names, whether in a query string or in the data, will cause the search to fail. There are various options for validating the user entry:

- 1) **RADIO BUTTONS:** for very limited choices, i.e. gender.
- 2) **DROP-DOWN LISTS.** A very effective way of choosing from available choices. It is even more so if it incorporates incremental searching where each letter keyed in moves the user to the point in the list corresponding to that letter combination. With very large lists there should be a means to pre-filter: for instance by having radio buttons or hyperlinks against sections of the alphabet (A-E, F-H, etc.) – downloading small lists is faster (although the overhead of calculating the values in the lists remains).
- 3) **HIERARCHIES.** These are another way of filtering terms; suitable for taxonomic and geographical names.

- 4) **BACKGROUND LISTS.** These are not visible to the user but help by mapping non-preferred terms to preferred terms and also help formulate the search parameters. They can also act as a query expansion tool – adding synonyms or foreign-language equivalents to the search automatically.
- 5) **MAP-BASED SEARCHING.** Useful for gathering area rather than point-based data, but only effective when the underlying datasets have full and precise co-ordinate information.

An entire (and effective) navigation system, based on available choices has been developed at the Laboratoire Informatique et Systématique at the University of Paris 6 for the Computer-aided Identification of Phlebotomine sandflies of America (CIPA) project (<http://cipa.snv.jussieu.fr/>).

While options 3, 4 and 5 provide an effective means of navigation for the user, they require a database application to be maintained centrally (normally within the CAS, but conceivably by an external site that acts as a taxonomic nameserver or a gazetteer). A continual effort is required to keep the lists up to date as additions to the contributing databases may contain new terms. However, if proper handling of synonyms and translation between different terminologies and between languages is to be effected, such a system of lists and thesauri must be built up and maintained.

120

It is probable that a combination of these search options will appear on the site and indeed it could be worth giving the user a choice of method (as people do have preferred methods – some would rather type text, others prefer picking from a list). The end result of search construction is usually an SQL statement. This is usually hidden from the user, but it could be made accessible to advanced users to edit by hand. If this is allowed, then care must be taken both to parse the statement for accuracy and to ensure that it is handled by the application, as some software only implements a subset of SQL.

A final feature that might be offered to regular users (and it would depend on there being user registration and login) is the ability to save queries for re-use. This is only really worth the effort when the underlying databases are expected to change in content fairly continuously.

Whatever search construction options are employed, it would be prudent to build in safeguards against a user performing a global search or one that requests such large numbers of records as to severely tax the computing and transmission resources. Should performance be slow (for whatever reason), there is a possibility that the user becomes dissatisfied and terminates the query (or the session). The application must be designed in such a way that it can tolerate abnormal closure.

Since any network set up as the result of *ENHSIN* is expected to be pan-European, it would be extremely useful for screens to be available in a choice of languages. This is now made easier through XML and XHTML. An example can be seen at the itis North America site (http://sis.agr.gc.ca/pls/itisca/taxaget?p_ifx=plgt), where the user can switch between English, French and Spanish. Terms in select lists would also need to be translated.

DISPLAYING RESULTS

It must be borne in mind that results may be drawn from many data sources and constitute many hundreds or records. How these are displayed and manipulated onscreen will clearly influence the ease of use of the application. As mentioned above, there must be mechanisms to limit impossibly extensive queries. The presence of a progress bar should also be of help to the user.

When the query is sent to a number of data sources and one or more take too long to return data, there should be a timeout mechanism that closes those connections and displays the data only from the remaining sources.

It would be a useful feature to display the number of hits (from each data source + total number of hits) returned by a query, before the full results are displayed. This gives the user the opportunity to refine their query to further filter the response, in the case of an overwhelming number of hits or, if hardly any hits are reported, the opportunity to broaden the search parameters. It is good practice to perform a series of simple queries rather than one complex one, and it could be useful to give the ability to refine the query parameters by iteration, and to back up a level when a query fails.

Where large sets of data are returned, it will be necessary to sub-divide them in some way. This may be limited at the outset by restricting the number of hits returned. However, if the user is given a selection of say 10, 50 or 100 records, this could be frustrating to those who need to see a whole dataset and is only really appropriate where results are sorted and ranked by relevance. The normal way of handling the display of large data sets is to divide them into separate pages and provide 'Next' – 'Previous' navigation buttons. Sometimes the user can be allowed to specify the number of records to be displayed per page.

Data will normally be displayed in simple tabular form with columns generated with html. Data may also be returned with XML mark-up. There are difficulties in displaying data returned from sources with richer structure – such as those that handle multiple values within a field. It may be necessary to present only summary data and refer the user to the provider's own site for the complete picture.

While it is possible to have background translation, through the use of thesauri, to help search across data sources in foreign languages, it is anticipated that results will be displayed in their native language.

FURTHER MANIPULATION OF RESULTS

It is possible to sort results by column and this could be a desirable feature. Thought needs to be given as to what the appropriate default sort order should be.

Another feature that can easily be provided is check boxes against each record returned that allows the user to mark records, by visual inspection, that they may want to filter for further action, such as printing.

In addition to displaying results on screen, there should be facilities for saving the results in a variety of formats (such as XML, RTF, CSV). Providers may, however, wish to place constraints on the downloading of complete datasets (although this may be provided as a facility for certain classes of registered users). Complete datasets may legitimately be required in order to perform further analysis.

Searches will only operate against the elements defined in the Core Data Dictionary, which also specifies which elements are returned. To go beyond these core elements will require access to the providers' databases. Many of the protocols proposed would allow parameters to be passed to open a direct link to a web-enabled database and find more details about a specific specimen record.

QUALITY CONTROL

From our own experience with migrating data, we have found that it is usual for there to be errors and inconsistencies in most data sets. Systems developed in-house, in particular, often lack proper data validation mechanisms. Natural Science collections fall behind many other museum disciplines when it comes to standards for terminology control. Quality management is an exacting task; the best-maintained systems are often found in smaller institutions where databasing is the responsibility of only one or two individuals. Larger institutions may fail to properly assign responsibilities for regular checks and maintenance of data quality.

Many of the issues of data quality are reviewed by Olsvig-Whittaker & Berendsohn (2000), especially the effort to verify existing data. They state that: 'additional information on performed data validation becomes crucial, especially where datasets from a number of different sources are combined in a common access system'.

If an entry differs from a search term by only one character, then it will fail to be picked up. While 'fuzzy' search techniques can help in this area, consistency of data is vital.

There is a general lack of terminology control within and amongst specimen databases. Partly this is due to the lack of suitable standards and greater effort (at a European and international level) is needed. Bodies such as The Taxonomic Databases Working Group (TDWG) and CODATA have an important role to play. There have already been some useful initiatives, such as the Taxonomic Authority Files workshop held in 1998 (<http://research.calacademy.org/taf/>). Standards developed in other disciplines, such as GML – the Geography mark-up language may be relevant (www.opengis.org/techno/specs/00-029/GML.html).

It is a fact of life that most specimen databases will contain gaps in their data. These can take various forms:

- ♦ There are still parts of the collection that have not been databased.

- ♦ The collection is not suitable for databasing at specimen level (some large entomological collections).
- ♦ Insufficient foresight has led to the design of a database with an inadequate number of elements (e.g. Collection date, which is in the *ENHSIN* element set, may not have been recorded by a provider). Alternatively, information may have been recorded but placed in a 'Remarks' or 'Notes' field, which is not easily searchable.
- ♦ Data is not available for specimens (such as date of acquisition).
- ♦ Information needs interpretation. Examples are geographical co-ordinates and higher taxonomy. A named locality may be given, but co-ordinates would need to be determined by reference to maps and gazetteers: this can require considerable effort and is unlikely to be undertaken retrospectively by curators. Similarly, unless a collections management system is programmed to 'backfill' higher taxonomic ranks once a genus name has been entered in a record, further effort is required to supply this information. Since classifications are changeable there is also a maintenance effort required.

Terminology control and validation against authority files have already been mentioned in Section 7. The issue here is whether consistency can be achieved across providers, either by co-ordinating authority files used with providers' own systems, or through referencing external authority lists.

It would be particularly useful to have a system of scoring both records and datasets for quality (completeness and accuracy). This requires consensus on what is a workable system. As a fallback position, indication of when a record was entered, or last updated, may assist in determining quality. Expert users can be enlisted to help spot errors in data, and suggest corrections. This is usually achieved by providing a feedback form on the website. Fishbase (www.fishbase.org/), IPNI, (the International Plant Names Index (www.ipni.org/)) and Florabase provide this facility. Another way of assessing quality (also through user feedback) is to ask, 'Did you find this resource useful?' It may also become possible to open up the network to become a truly collaborative venture with different providers, and users, assisting each other in supplying 'missing' data, correcting errors and building authority files. In this scenario, but also as general good practice, data within records should be attributable; with both the author and date of entry/change noted.

It is also good practice (but needs a sophisticated system to achieve) to preserve changed data and maintain an audit trail. This is seen more as a matter for individual providers than for the network as a whole.

If a scoring system for data quality is in place, users can be given the option of choosing to include or exclude data below a certain quality. Some data providers may wish to supply only high-quality data sets, but users may wish to have access to all available data and decide, for themselves, on its relevance.

STANDARDS

This section reviews existing standards that might be adopted by an *ENHSIN* network. It is however one thing to apply standards to a new dataset and quite another to apply them retrospectively to an existing dataset. Different contributing datasets may adhere to different standards, or different versions (editions) of the same standard. Data providers may allege that they follow a given standard whereas, in practice, they have introduced local modifications. Some standards are better (better fit for purpose or offer tangible benefits for interoperability) than others. Standards can exist but not be taken up by a community. They are of particular use when thesauri and authority files are constructed to map between different configurations of data. They are most likely to be employed if the adopted architecture has a highly developed CAS or where large amounts of data are stored in a central resource. Given the nature of the data associated with biological specimens, it is doubtful whether many existing standards will completely satisfy requirements, but the *ENHSIN* project (and its successors) provide an opportunity to promote the use of standards and, where suitable standards are lacking, to develop new ones (it should be noted, however, that the standards mentioned below have not been specifically tested or endorsed by *ENHSIN*).

124

Candidate areas for adoption or development of standards include:

TAXONOMIC SERIAL NUMBERS

Systems of codes (usually numerical) designating taxonomic names have been developed in a number of projects. The system used by ITIS (www.itis.usda.gov/) is probably the best known. Within the United Kingdom, other examples include codes for taxa developed by the National Biodiversity Network, which allows the source of the data to be tracked, and codes for taxa within The Species Directory of the marine fauna and flora of the British Isles and surrounding seas (Howson & Picton, 1997). There have been suggestions that codes provide a more efficient way of referring to taxa and that it would be an advance to use codes to replace binomial nomenclature; although in practice this is not a feasible proposition. They could, however, offer a way of unambiguous linking between different taxonomic databases.

NOMENCLATURE STANDARDS

Biological names are governed by five sets of codes: International Code of Zoological Nomenclature (ICZN, 2000), International Code of Botanical Nomenclature (Greuter *et al.*, 2000), International Code of Nomenclature for Cultivated Plants (Trehane *et al.*, 1995), International Code of Nomenclature of Bacteria (Sneath, 1992) and International Code of Virus Classification and Nomenclature (Francki *et al.*, 1990). Details can be found at www.biosis.org/zrdocs/codes/codes.htm. An inter-union committee has been established to promote the harmonisation of botanical, zoological and microbiological

codes of nomenclature. A draft for a unified biocode is being developed through the Bionomenclature programme of the International Union of Biological Sciences (IUBS).

GEOGRAPHIC STANDARDS

These embrace both place names and geographic co-ordinates system and are handled by:

- ♦ **GAZETTEER APPROACHES.** Although extensive electronic gazetteers exist (such as the Getty Thesaurus for Geographic Names, www.getty.edu/research/tools/vocabulary/tgn/index.html and the Alexandria Digital Library, <http://fat-albert.alexandria.ucsb.edu:8827/gazetteer/>) museums' collection data is too diverse to be satisfied by a single reference source such as TGN or ADL and this is an area where it might be best build a gazetteer specifically to serve *ENHSIN*. Most gazetteers focus on places where people live, whereas most of the interesting biological collections come from places where people do not live (as an instance, the Botany Department of The Natural History Museum, London, have found, for current research areas, anything up to 70% of place names are not in any of the major online geographical databases). Gazetteers may not help with defining marine areas or mapping archaic place names to modern equivalents.
- ♦ **CODIFICATION SYSTEMS.** Examples include: ISO 3166 Representation of Countries <ftp://ftp.fu-berlin.de/pub/doc/iso/> and a TDWG Standard (Brummitt *et al.* 2001).
- ♦ **COORDINATE SYSTEMS.** These can include points, lines and polygons. A standard representation of latitude, longitude and altitude is defined in ISO 6709.
- ♦ **GRID-BASED SYSTEMS.** These are usually national; an exception is the Universal Transverse Mercator system (<http://mac.usgs.gov/mac/isb/pubs/factsheets/fs07701.html>). The Geography Markup Language GML v.2.0 (www.opengis.net/gml/01-029/GML2.html) provides an encoding standard for data transfer.

125

HABITAT

There are a number of classification schemes for habitat types and most of them are national. The European Topic Centre for Nature Protection and Biodiversity of the European Environment Agency has developed a common reporting system called the European Nature Information System (EUNIS, <http://mrw.wallonie.be/dgrne/sibw/EUNIS/home.html>) that is suitable for specimens and field observations throughout Europe and surrounding waters.

STRATIGRAPHY

There are stratigraphic recording schemes for a range of different methods including: Quantitative Stratigraphy, Chemostratigraphy, Cyclostratigraphy, Stable Isotopes, Magnetostratigraphy, Radiometric Stratigraphy, Sequence Stratigraphy, Biostratigraphy, Chronostratigraphy and Lithostratigraphy.

The International Union of Geological Sciences Commission on Stratigraphy (www.iugs.org) produces guidance on the principles of stratigraphic classification, terminology and rules of procedure (www.micropress.org/stratigraphy/ics.htm).

TIME

Collections databases need to record both absolute, vague and relative time periods. Standards to handle dates are incorporated in the schema being produced by CODATA. There is an ISO Standard date and time formats (Wolf & Wicksteed, 1997), and a note on this format by the W3 Consortium (www.w3.org/TR/NOTE-datetime).

PERSONAL NAMES

126

As mentioned earlier, personal names may need to be recorded verbatim. Otherwise, a good standard to follow is LCNAF (Library of Congress Names Authority), which follows the Anglo-American Cataloguing Rules format (Gorman & Winkler, 1988; Swanson, 1988).

Authorities for plant names should follow Brummitt & Powell (1992).

ORGANISATION/EXPEDITION/CRUISE NAMES

There do not seem to be any standards specifically for these items.

CURATORIAL PRACTICE

There are a number of national standards. The Canadian Heritage Information Network has produced a Natural Sciences Data Dictionary (www.chin.gc.ca/Artefacts/RULS/e_hp_ruls.html). The International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) has produced guidelines for Museum Object Information (www.cidoc.icom.org/guide/guide.htm).

CHARACTER SETS

The handling of multinational characters is nowadays best achieved using Unicode: www.unicode.org/unicode/standard/principles.html. However this is not yet fully (or consistently) implemented by software such as browsers or word processors. Cyrillic and Chinese characters may be transliterated and, if they are, the standard used within the contributed dataset should be ascertained as one of several standards that could have been used.

RESOURCE CITATIONS: BIBLIOGRAPHIC AND ELECTRONIC

An appropriate standard for bibliographic citations is the Anglo-American Cataloguing Rules, Second Edition (AACR2) (Gorman & Winkler, 1988; Swanson, 1988).

A number of cataloguing standards are being extended to cope with citations of online resources. One currently available is from the Council of Biology Editors (www.councilscienceeditors.org/pubs_ssf.shtml).

ACCESSIBILITY

These standards apply to the design of a website rather than to the data it contains. There are useful guidelines provided by W3C (www.w3.org/WAI/) and by the United States Architectural And Transportation Barriers Compliance Board (www.access-board.gov/sec508/508standards.htm).

ACCESS CONTROL

Issues of access control have been highlighted in the previous chapter. Here we are concerned with solutions. The simplest option is not to include sensitive data in the source. If certain classes of user want (and are entitled) to see everything then they should negotiate directly with providers.

There may be a need to exclude sensitive data at both field and record level. Fields are relatively easy to control as they are simply excluded from the view created onto the dataset. Sensitive records need marking in the source database and a 'NOT' statement incorporated in queries sent from the CAS.

Constraints will be needed to protect Intellectual Property Rights (IPR) and other legislation affecting data protection and personal information. This will normally require both copyright and 'Use of Data' statements to be displayed prominently on any access site and possibly by requiring users to sign a declaration before allowing registration to use the site.

Access and security issues could also affect the ability for the CAS to query providers' databases. Depending on local arrangements, there may need to be a way of authenticating the CAS (possible as a 'virtual user') to permit access.

There is another class of 'virtual user' of the network that must be taken into account. This is search engine and agent software that gather information for indexing purposes. While the best way of advertising the services offered by the network may be through placing links on selected portal sites, metadata should be provided in page headers. Dublin Core is useful in this respect, but the level of detail that would be necessary to show the full extent of data that can be provided is probably unachievable through this means. Some sort of summary pages presenting (in clear text form) the scope of the network would be useful, both to real and virtual users.

INTEROPERABILITY

Interoperability, the ability to connect across different data domains, was not investigated within the *ENHSIN* project. The desirability of accessing external

sites such as nameservers that can validate taxonomic names has already been mentioned. Undoubtedly, the ability to link to bibliographic sites, climatological sites, and others would be desirable: user surveys would give an indication of priorities.

The development of open systems and standards will be a necessary preliminary. Much will depend on the owners of the sites to be linked to providing the necessary 'hooks' to their resources.

It could be that additional capabilities are added to the CAS through collaborative effort – an example might be the provision of a component to map result sets geographically. At this stage, the most pragmatic approach would seem to be for the CAS to allow users to download the results of searches. The onus then falls upon the user, if they wish to analyse the data further.

SUSTAINABILITY

CHANGING TECHNOLOGIES

- ♦ Available technologies will change. Importantly, some existing technologies will no longer be supported. The guiding principle on when to change should be governed by the possibility of increasing accessibility.
- ♦ Advances in hardware and infrastructures should allow greater emphasis to be placed on local computing, as processor-power, disk capacity and bandwidth improve.
- ♦ Database software presents a possible problem. Manufacturers strive to continually improve their products but this sometimes leads to them changing the format (and data types). If users choose not to upgrade to the latest version they may find that the original version is no longer supported. Software products may change ownership and may in time be dropped by the acquiring company.
- ♦ The technologies incorporated into web browser software will change. This is often dictated by commercial considerations – e.g. use of Java in Microsoft products. An *ENHSIN* network will always have to deliver a service that can be accessed by a range of browser products and versions.
- ♦ In choosing which technology or software to adopt it is always better to adopt an open system rather than a proprietary standard.
- ♦ Making changes always involves a cost. It is not right to expect data contributors or users to upgrade unless there are significant benefits.
- ♦ In the longer term, the nature of computing devices and ways of accessing the Internet will change and the *ENHSIN* network must develop a strategic vision to keep abreast of the opportunities presented and of changing classes of users.

VERSION CONTROL

This will revolve around maintaining metadata that catalogues changes made within individual datasets and data posted between systems. Version control will affect not only specimen records but also associated authority files, especially those maintained centrally. When copies are made of data, it is essential to know which is the 'master' version. Even though much of this can be automated, it will be important to establish a system of human oversight and checks to ensure that procedures are operating correctly.

SCALABILITY

As the contributing datasets grow both in size and number, steps will need to be taken to keep performance up to an acceptable standard. This could require investment in improved hardware but might also require changing database software. It is far better to plan at the outset for future growth as changing databases could lead to unforeseen problems that interrupt the service. Desktop database applications are not suitable for handling large number of simultaneous users and may have physical limits on the number of records that they can contain. Larger databases generally require more administration – both in terms of controlling quality of data and also in tuning the performance of the database. Distributed querying may become unworkable if very large numbers of remote databases are accessed and the returned result sets may become too large for the users to handle interactively.

129

CONCLUSION – PRIORITIES

The eventual aim of *ENHSIN*, and follow-on projects, will be to create a single virtual resource; with **distributed data management**, combined with the equally valuable benefits of **unified data access**.

This paper has demonstrated that there are a number of possible ways of going about things. A range of suitable technologies already exists and these have already been tested in various existing projects. Some technologies are still emerging and do not yet have defined standards.

It is possible that the best approach may be a hybrid one – and one providing substantially more than a simple gateway to collections databases: one that will include collections-level descriptions (as in the *BIODIV* project) and act as a portal to external sites (through URI links).

Whether the network operates through a single Gateway site or as a series of distributed objects will depend on how much knowledge is required to be held centrally to enable navigation of the resources. However, the basic underpinning technology is clear: users will access the system through web browsers using HTTP to transport data wrapped using XML. While it is not possible at the time of writing to be firm about the ultimate solution (partly because standards are still developing – although this is always going to be true!), a start must be made somehow. The *ENHSIN* pilot is the first step. We

must, however, be prepared for technology to evolve, and for the developing network to embrace those changes.

It is important that, regardless of where the network may head in the future, the initial solutions must take account of the current state of collections databasing activities in European museums. Therefore the possibility of developing and supplying a low-cost Collections Management System could be worth considering.

For a network to take off, it has to be made exceedingly simple for data providers to contribute their data. It may not be possible, for instance, to provide on-site technical support for data contributors. The simple expedient of mounting a copy of providers' data in a central system is achievable, but could be prone to versioning problems. The success of the network will be related to the quantity and quality of the content; and there must be perceived benefits to providers to sign up to the network.

The approach adopted by Species 2000 is probably not appropriate for an *ENHSIN* type network, as it requires much technical expertise on the part of the data providers (even with the provision of toolkits). The University of East London approach may be more suitable. Here potential providers can register their database and, provided it is already Web-enabled, do not have to build an additional wrapper. Extending the approach developed for the *ENHSIN* pilot also offers a great deal of promise.

It is expected that the final model will differ in a number of respects from the pilot (as the pilot has been determined by expediency, available skills and budget). The final model might need to link to a name-server (possibly external) to resolve taxonomy.

In the future, the network should be prepared to accept (and be able to consolidate) new datasets – data culled from different sources (say a number of Collections Management Systems, whether or not they are part of the network, and possibly including personal research data or field observations) which have been compiled/analysed/augmented by a researcher and then fed back into the system. This is already possible on the University of East London's site. Identifying the source and accurate version control then becomes very important.

Also, in the future, the network might be made accessible not just through personal computers but also to PDAS, WAP devices and home entertainment systems. This will require user-agent negotiation to handle the alternative file formats (e.g. WML – Wireless Macro language, the HTML equivalent for wireless applications). Content management should, in any case, be in place to ensure that appropriate content can be delivered to different versions of Internet browsers (and the application should degrade gracefully if the browser does not support a particular feature).

It is somewhat difficult at present to assess whole-lifetime costs associated with the different options, although this would assist decision-making and help support bids for funding. Costs that need to be addressed include:

- ♦ Commercial Software: for operating systems, databases, development languages, indexing and content management software. Costs will include both licenses support costs and the cost of upgrades).
- ♦ Software development: costs can vary greatly depending on whether is undertaken in-house or through professional programmers (who can charge between €575 and €2,000 per day).
- ♦ Training: data providers in particular may need training to build wrappers to databases.
- ♦ CAS running costs: sustainability will require adapting to new versions of (database) software and browser capabilities. Undoubtedly new features will be requested by users.

All this implies that a continuing source of funding will need to be secured to keep a network operational.

References

- Brummitt, R.K., Pando, F., Hollis, S. & Brummitt, N.A. 2001. Plant Taxonomic Database Standards No. 2. *World Geographical Scheme for Recording Plant Distributions*. ed. 2. Pittsburgh, Hunt Institute for Botanical Documentation (ISBN 0-913196-72-X).
- Brummitt, R.K. & Powell, C.E. (Eds). 1992. *Authors of Plant Names*. A list of authors of scientific names of plants, with recommended standard forms of their names, including abbreviations. Kew, Royal Botanic Gardens (ISBN 947-643-44-3).
- Francki, R.I.B., Fauquet, C.M., Knudson, D.L. & Brown, F. 1990. Classification and nomenclature of viruses. *Archives of Virology* Supplement 2: 1-445.
- Gorman, M. & Winkler, P.W. (Eds). 1988. *Anglo-American Cataloguing Rules*, Second Edition, 1988 Revision. Prepared under the direction of the Joint Steering Committee for Revision of AACR. Chicago: American Library Association.
- Greuter, W. et al. (Eds) 2000. *International Code of Botanical Nomenclature* (Saint Louis Code), adopted by the Sixteenth International Botanical Congress, St Louis, Missouri July-August 1999. Königstein : Koeltz Scientific Books.
www.bgbm.fu-berlin.de/iapt/nomenclature/code/SaintLouis/0001ICSLContents.htm
- Howson, C.M. Picton, B.E. (eds), 1997. *The Species Dictionary of the Marine Fauna and Flora of the British Isles and Surrounding Seas*. Ulster Museum & The Marine Conservation Society, 509 pp.
- ICZN, 2000. International Code for Zoological Nomenclature, Fourth Edition.
www.iczn.org/code.htm
- Lynch, C. 1997. The Z39.50 Information Retrieval Standard. D-LIB Magazine, April 1997.
www.dlib.org/dlib/april97/04lynch.html
- Miller, P. 1999. Z39.50 for All. *Ariadne*, Issue 21. www.ariadne.ac.uk/issue21/z3950/
- Olsvig-Whittaker, L. & Berendsohn, W.G. 2000. Computerizing and networking biological collection data. In Berendsohn, W.G. (Ed.), *BioCISE, Resource identification for a biological*

collection information service in Europe, pp. 5-12. Berlin.
www.bgbm.fu-berlin.de/BioCISE/Publications/Results/2.htm

Sneath, P.H.A. (Ed.). 1992. *International Code of Nomenclature of Bacteria*, 1980 Revision. Washington.

Swanson, E. 1988. *Anglo-American Cataloguing Rules*, Second Edition, 1988 Revision, Amendments 1993. Prepared under the direction of the Joint Steering Committee for Revision of AACR. Chicago: American Library Association.

Trehane, P., Brickell, C. D., Baum, B. R., Hettterscheid, W. L. A., Leslie, A. C., McNeill, J., Spongberg, S. A. & Vrugtman, F. (Eds). 1995. International Code of Nomenclature for Cultivated Plants – 1995, adopted by the IUBS Commission for the Nomenclature of Cultivated Plants. *Regnum Vegetabile* **133**. Quarterjack Publishing, Wimborne, UK.

Wolf, M. & Wicksteed, C. 1997. Specification for Representation of Dates and Times in Information Interchange ISO 8601.

Xu, X., Jones, A.C., Pittas, N., Gray, W.A., Fiddian, N.J. White, R.J., Robinson, J.S., Bisby, F.A., Brandt, S.M. 2001. Experiences with a Hybrid Implementation of a Globally Distributed Federated Database System Proc. Second International Conference on Web-Age Information Management (WAIM 2001), Springer-Verlag (Lecture Notes in Computer Science), pp 212–222.