DELIVERING A NAME-SERVER FOR BIODIVERSITY INFORMATION

C Hussey[1]*, S Wilkinson[2] and J Tweddle[1]

*[*1] Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom*
Email c.hussey@nhm.ac.uk, j.tweddle@nhm.ac.uk
[2] Joint Nature Conservation Committee, Monkstone House, City Road, Peterborough, PE1 1JY, United Kingdom
Email steve.wilkinson@jncc.gov.uk

ABSTRACT

The number of online resources for biodiversity information is growing. Names of organisms underpin access to information but present a number of unique problems when used as search terms. We examine these problems and assert that a taxonomic name-server or thesaurus is necessary to enable optimal retrieval of records from multiple datasets. A simple solution is presented, based upon our experience working with "real-world" data in the National Biodiversity Network (NBN) in the United Kingdom. The NBN provides access to over 18 million observational records and incorporates a nomenclator covering 198,000 names.

Keywords: Biodiversity, Natural History, Taxonomy, Names of Organisms, Database design

1 INTRODUCTION

This paper is based upon a presentation given at the 19th International CODATA Conference in Berlin, November 2004, but has been expanded to cover other aspects of using organism names in databases.

1.1 Background

A great challenge for the Information Society is the compilation of an inventory of life on earth. Even after 250 years of scientific endeavour, it is estimated that only 15 to 20 per cent of organisms have been named and described (May, 1988; Wilson, 2000). Nevertheless, a great deal of raw data have been accumulated: there are estimated 1.5-3 billion specimens in museums worldwide (Scoble, 2003) and countless observation records. The last decade has seen an acceleration in digitisation effort as collections and observation records become accessible on-line, and major EU and NSF funded projects have contributed to the development of models, technologies, standards and networks of resources. The Global Biodiversity Information Facility (GBIF) already provides access to 74.5 million records (GBIF, n.d.) and the uBio Taxonomic Name Server holds 2.3 million names (uBio, n.d.). We are, however, still a long way from compiling a global inventory. Impediments include widely distributed information, the large volume of data that are not in electronic form, difficulty in obtaining funding to build content, an insufficient number of taxonomic experts and inadequate authority files to help with data validation.

Names of organisms, both formal scientific names and vernacular names, are key to accessing biological information in diverse disciplines such as conservation, legislation, genetics, invasive species, organisms of medical importance and trade in protected species (see Chapman, 2005a; Patterson, 2003). Advances in taxonomic understanding lead to changes in nomenclature. This, and other complexities surrounding the formation and citation of names, presents challenges to application designers and data managers, particularly when attempting to integrate disparate datasets.

1.2 The National Biodiversity Network

Biological recording has had a long history in the United Kingdom and is still actively pursued. It is estimated that more than 60,000 individuals are involved and over 60 million species-based records exist (Burnett, Copp & Harding, 1995). The National Biodiversity Network (NBN) was formed in 2000 to provide easy access to this data and share it between users (www.nbn.org.uk). The NBN maintains a repository of observation records, which are freely accessible through a web portal known as the NBN Gateway (www.searchnbn.net). At the time of writing, there are 18.7 million records available, from 145 different datasets. The NBN provides the Recorder family of software tools for biological recorders and collections managers. Nomenclatural services underpinning the Gateway and Recorder, and the taxonomic needs of NBN partners are provided by the Species Dictionary project which is managed by Natural History Museum and covers 198,000 names compiled from 269 checklists.

A guide to British wildlife for non-specialist users is provided by the Nature Navigator website. The products and services produced by the NBN are driven by user needs. The experience gained over 5 years has provided a keen appreciation of the problems inherent in dealing with "real-world" data.

2 RECORDING THE IDENTITY OF AN ORGANISM

The identity of an exemplar of a type of organism is fixed by association with a formal scientific name, formed subject to defined rules (International Commission on Zoological Nomenclature, 1999; Greuter, McNeill, Barrie, Burdett, Demoulin, Figuerias, et al., 2000; Trehane, Brickell, Baum, Hetterscheid, Leslie, McNeill, et al., 1995; Lapage, Sneath, Lessel, Skerman, Seeliger & Clark, 1992; Franki, Fauquet, Knudson & Brown, 1990), together with the citation of the naming authority. The formal names, based upon the Latin and Classical Greek vocabularies, have a world-wide currency amongst scientific users. On the other hand, vernacular names, where they exist, will be different in each country in which the organism occurs. Vernacular names should not be overlooked, as they provide a position of familiarity (and often stability) for non-scientific users. Such users may be disempowered if a biodiversity database only allows access through scientific names.

The circumscription of a species may change over time. Different authors may hold different opinions about what constitutes a particular species, or may apply a name to organisms that differ in some respect from the original description. This makes it important that the authority for the name is always recorded. Ideally the source used for an identification should also be cited. The term 'taxon concept' is now in current usage for a name in a particular context, according to a particular source (see Geoffroy & Berendsohn, 2003). As understanding of a taxonomic group advances, species may be reassigned to different genera. One species may be subsumed into another or, alternatively, what used to be regarded as a single species may be split into two or more species. Aside from taxonomic revisions, it should be recognised that individual specimens, or observations, may be subject to redetermination, so that the names attached to them may change over time.

2.1 Issues and Challenges

Names of organisms present unique challenges, both for the design of individual databases and in the integration or federation of databases.

There are a number of ways in which names can be stored in a database. These vary from a single field for the entire name string and authority, through an atomisation of the name with separate fields for each component (genus, species, infraspecific name, etc.), to full normalisation of each element. Chapman (2005b) recommends atomisation. Certainly, it is easier to concatenate separate elements to recreate a complete name than it is to parse a name into its component parts. If data are entered by keyboarding, then filling separate fields is relatively straightforward, but it becomes challenging if data are to be imported from a file where the names are not already atomised. It is our experience that names can be complex, consisting of up to 14 words in the case of some hybrid formulae. Having separate fields for each possible element may lead to redundancy in the database, with some fields seldom filled. However, if insufficient fields are provided, there is a danger that data become shoe-horned into fields in an inconsistent manner, which creates anomalies in sorting and searching.

The basis of interrogating datasets relies upon string matching. A mismatch of only a single character means that a record is not returned. Soundex and fuzzy-matching algorithms can certainly be employed, but the results of these techniques can be confusing to users who may not be familiar with the names returned. An important question is: if a user fails to get any hits from a search, can they be sure that the database does not contain any records relating to what they wanted? For example, a negative result could arise due to a misspelling in the search term, or within the database itself. In respect to user's misspelling search terms, a log of the Nature Navigator website records terms entered that were unmatched and shows alarmingly poor spelling - in this case of vernacular names (mammels, mamals, hampster, wosps).

Examples of failed matches can occur within scientific name strings, authorities and vernacular names. All can be subject to errors of transcription. Endings of species names may be variable, as in the case of 'brandti/brandtii', or where the ending depends on the gender of the genus 'splendidus/splendida'. Hybrids are denoted by an 'x', but this can be entered as an alphabetic character or as a multiplication symbol. In the case of infraspecific names, these may be written with the infraspecific rank name spelled out in full 'variety', contracted 'var.', or omitted entirely. Authorities are particularly problematic. The nomenclatural codes lay down rules for the formation of authorities but these are not always followed. Errors include: non-standard abbreviations of authors for plant names, incorrect punctuation or spacing following initials, wrong use of parentheses, and transliteration errors. Accents may also be omitted, or wrongly displayed upon import due to incorrect use of code pages or Unicode. In some cases, names may be provided without any authority information at all. Variation in the name string also

occurs in vernacular names - consider the case of 'bumblebee', 'bumble bee', 'bumble-bee'. If queries are case sensitive, then errors may also occur due to inconsistent capitalisation.

Another challenge is that the same name may be applied to different organisms; such names are termed 'homonyms'. Generic homonyms are more common than might be expected and must be taken into account whenever datasets span more than one Kingdom. There are even a few homonyms at species level. At the time of writing, the NBN Species Dictionary contains 132 instances of genus-level homonyms and we check for over 5,400 known homonyms as part of our data import routines. There are also instances of the same vernacular names referring to more than one species, either within or across Kingdoms. Examples include Elk (see Table 1.) and Sweet William (*Mustelus mustelus* (Linnaeus, 1758) - a shark, *Galeorhinus galeus* (Linnaeus, 1758) - a shark, *Dianthus barbatus* L. - a flowering plant). This may create problems for Data Managers, who will need to distinguish between the common names at a structural level and also ensure that users can pick the name that is appropriate for their needs. A similar situation applies where the name string of a genus is also in use as a vernacular name (e.g. Anemone, Hydrangea, Petunia).

Table 1. Different application of vernacular names in different countries

North America		Europe
Moose	*Alces alces* (Linnaeus, 1758)	Elk
Elk	*Cervus elaphus* Linnaeus, 1758	Wapiti

On the other hand, several names may apply to the same organism. A good example, where taxonomic opinion has changed, is the Rainbow Trout; formerly known as *Salmo gairdneri* Richardson, 1836, this is now placed as *Oncorhynchus mykiss* (Walbaum, 1792). Accepted names may vary by region. For example, *Lacerta vivipara* Jacquin, 1787 is current in the United Kingdom, but is referred to as *Zootoca vivipara* (Jacquin, 1787) in continental Europe. An additional factor is that specimens can be subject to re-identification events. These may confirm or change a previous determination, or result in multiple determinations (according to different experts). Berendsohn, Anagnostopoulos, Hagedorn, Jakupovic, Nimis, Valdés, et al. (1999) list nine types of identification event, including a negative determination (the specimen does not belong to a particular taxon) and similarly, for observational data, where a species was not found during a survey. These last two identification types would need to be properly distinguished in search results; something that is possible within a single database, but more difficult when integrating results from dispersed databases.

A further complexity is that naming resources for the biological recording community will need to contain names of species aggregates (species groups). These are used for species that are difficult to identify in the field and pose two problems for database designers. Firstly, records of aggregates need to be returned in searches on any of their component species – assuming that these are known and recorded in the database. Secondly, there are instances where species aggregates contain species from more than one genus, which makes it difficult to assign a parent in a placement in a hierarchical classification.

Lastly, it should be remembered that many datasets will lack the information necessary to properly handle taxon concepts and that most user interfaces are not sufficiently refined to search at a level of taxon concept, rather than just a name. When searching across distributed databases, what can be presented will be determined by the lowest common denominator. When names lack authorities, they cannot be assigned with certainty to a taxon concept. When species have been split, but one part retains the original name, it still forms a new taxon concept. However, unless a date of determination is available, what is meant by such a name cannot be known.

2.2 Needs

Much excellent work has gone into developing standards for data exchange and wrapping datasets (ABCD, n.d.; BioCASE, n.d.; Darwin Core, n.d.; DiGIR, 2005.; SPICE, n.d. - see references), with most of the work being led through the Taxonomic Databases Working Group (TDWG, n.d.). The community is moving from theoretical modelling and pilot studies into operational systems and networks. These now need to be proved with real-world data of varying degrees of quality and completeness. A user expects that a search will return all relevant names and a system of mapping names – a query expansion tool – will be needed to make this possible.

Having dwelt upon some intricacies of taxonomic names, it must be recognised that databases will be consulted by different classes of users, with quite different needs. Some users will only want to see a list of current scientific names, whilst others will want to see a full listing of synonyms. Although superseded, such synonymous names will allow them to uncover research in older publications and specimens in museums and herbaria. Legislators will expect to work with names, often outdated, that appear in directives and conventions.

Non-expert users will want to use only vernacular names. For the very successful FishBase website, which holds 207,900 common names relating to 28,900 species, Froese & Pauly (2000) comment: "Claiming that the common names of fish are one of their most important attributes is an understatement. In fact, common names are all that most people know about most fish as shown by the fact that most people accessing FishBase on the Internet do so by common name". Mapping will be needed between common names in different languages, and between common names and scientific names. Ideally this will extend beyond names for species to include informal names for higher groupings.

Systems that are built to provide access to naming resources will need to be both scalable and sustainable. This relates as much to service provision as to technical solutions. Systems will need to cope not only with changing nomenclature in the face on ongoing taxonomic research, but also provide for error correction through feedback mechanisms. Systems also need to support searching on an element of a name - such as a word used as a genus or subgenus name. This poses a considerable challenge if the name is stored as a single string.

3 SOLUTIONS ADOPTED BY NATIONAL BIODIVERSITY NETWORK PROJECTS

The NBN has adopted a data warehouse principle where copies of datasets provided by data owners are pooled in a single structure. Signed agreements with the data providers allow them to retain ownership but licence the use of data to the data collators (the NBN) and end users. Data owners retain control over access to their datasets.

3.1 The Species Dictionary project

The Species Dictionary (www.nhm.ac.uk/nbn) implements part of the original version of a highly-developed data model designed for the NBN by Charles Copp (Copp, 2000), which he has subsequently reworked to include a multilingual thesaurus (Copp, 2004a & 2004b). The TAXON table stores name strings, both scientific and vernacular, with authorities contained in a separate field. The model allows for different versions of names, with qualifier details such as ' *nec* Smith' and *'sensu lato'* stored in the ATTRIBUTE field of the TAXON_VERSION table. This table also records the rank and a higher group name for each taxon version. It should be noted that whilst the Dictionary can handle taxon concepts, it does not follow a rigorous model. Names are associated with checklists, which may themselves have versions (editions). Details of these checklists are stored in the TAXON_LIST and TAXON_LIST_VERSION tables and the association between TAXON_VERSION and TAXON_LIST_VERSION is made in the TAXON_LIST_ITEM table. This latter table can also store row pointers for a parent and preferred name, which allows for a limited mapping between vernacular names or synonyms to valid scientific names. Mappings may vary between lists and the model can handle multiple classifications.

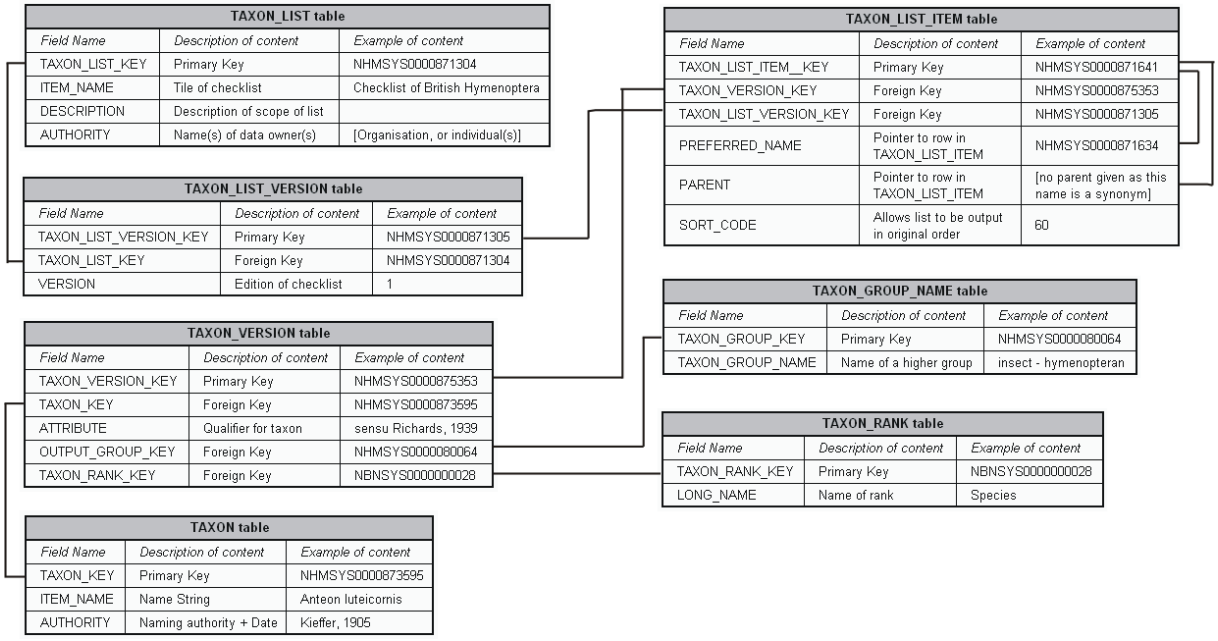


Figure 1. Simplified relationship diagram for core of Species Dictionary (not all fields or tables are shown)

Whilst this model handles variations between lists it does not provide the flexibility required to map all of the uses of a name, regardless of which list, or lists, the name appears in. The NBN Gateway, in particular, requires a query expansion tool that returns observation data for any associated names (vernacular, scientific, synonyms and spelling variants). The latest NBN data model (Copp, 2004a) provides this functionality, but we have been constrained by the need to support existing applications (NBN Gateway and Recorder software) and the Dictionary website, although we do not preclude the possibility of moving to the new model in the future. Instead, we implemented a very simple solution that has proved its worth in a short time. This involved the construction of a new table (NAMESERVER table) that maps every name in a taxonomic grouping to a recommended scientific name. Three flag fields are also provided which record the form of a name, the status of a name and the type of a name. The TAXON_FORM field is used to indicate whether a name has a complete authority and is spelled correctly. The values for this field are well-formed. ill-formed or unverified. The TAXON_STATUS field indicates whether the name is a recommended name, a synonym or unverified (i.e. requiring confirmation from a taxonomic expert). The TAXON_TYPE field shows the name as scientific or vernacular. These three fields are useful for filtering data, especially in the production of lists of approved names. Having made a link between a name and a recommended name, it is then possible to find all other names mapped to that recommended name. The name-server thus acts as a query expansion tool by returning information attached to all equivalent name strings.

NAMESERVER table		
Field Name	*Description of content*	*Example of content*
NAMESERVER_ID_KEY	Primary Key	NHMSYS0000882968
INPUT_TAXON_VERSION_KEY	Foreign Key	NHMSYS0000875353
TAXON_VERSION_FORM	Flag for completeness/correctness of names	[**W**]ell-formed, [**I**]ll-formed, [**U**]nverified
TAXON_VERSION_STATUS	Flag for preferred/non-preferred names	[**R**]ecommended name, [**S**]ynonym, [**U**]nverified
TAXON_TYPE	Flag for scientific or vernacular names	[**S**]cientific, [**V**]ernacular
RECOMMENDED_TAXON_VERSION_KEY	Foreign Key	NHMSYS0000875346

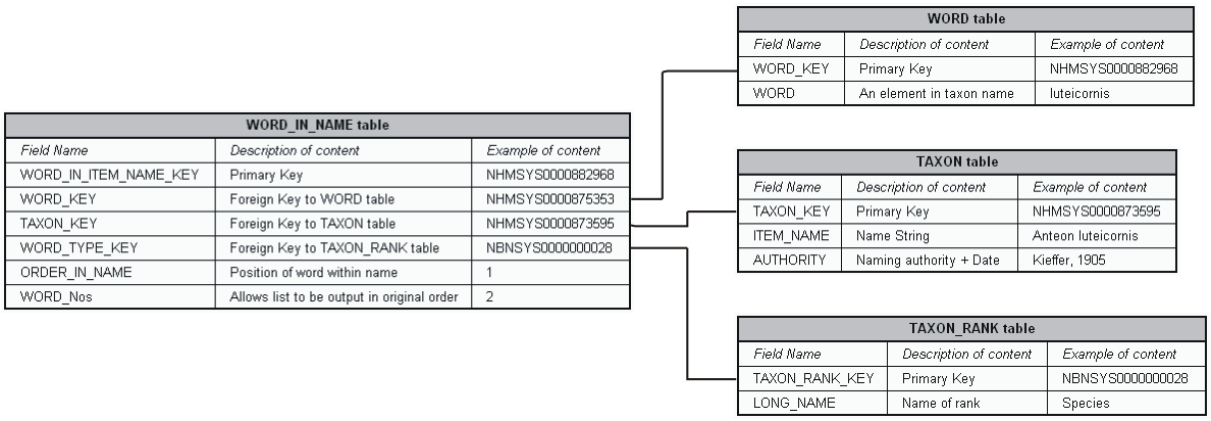
Figure 2 Structure of NAMESERVER table.

The name-server is being built up group by group. Many of the mappings are obvious and easily constructed but some cases need to be checked with taxonomic experts. Construction of the name-server is time-consuming and, to date the facility includes 43% (84,100 rows) of names in the Dictionary. In the beginning it was a challenge to discover all of the names in the Dictionary belonging to a particular taxonomic group. A scavenging tool was constructed to extract names. This explored existing links between parents and children, and synonyms and preferred names in each list We are now using a system of named higher groupings to provide a browse interface (in addition to a search interface) and to help users readily distinguish different types of organism in lists of species returned as a result of a search. An example of the use of these group names can be seen in the NBN Gateway. We have now assigned every Taxon Version record to a group, which means that we no longer have to scavenge the database for names.

The Species Dictionary records names as they appear in published checklists and will retain spelling errors and badly-formed names. The reasoning behind this decision is so that we are able to reproduce checklists verbatim. Curators will also wish for names from specimen labels to be preserved exactly as written. Chapman (2005b) advocates the cleaning up of database entries. We endorse this view, but a problem for search applications is that misspellings do arise in publications and databases and, unless you know about them, they will get omitted from search results. Simple lists of authoritative names are not sufficient. Our policy would be to detect errors and inform data providers so that they may correct them in future versions. The NAMESERVER table, together with comments fields in the TAXON_VERSION and TAXON_LIST_ITEM tables (not shown in Figure 1) at least allow known errors to be marked as such. The original Dictionary data model was designed to accommodate published checklists, which might go through several editions or amended versions. We are now moving towards designating certain lists as 'recommended lists', which will be dynamically maintained by the data providers. There will be feedback forms that encourage users to report errors for correction. Maintained lists will also facilitate the rapid incorporation of new records and nomenclatural changes and speed up the cycle of passing on new research to the Recorder software and other end users.

Recently, we have borrowed from a new part of the data model (Copp, 2004b) to produce a Word-in-Name table. The problem of atomising complex names has been mentioned in Section 2.1 above and, from the outset, the Dictionary was designed to store taxonomic names as a single string. However, many portals expect to be able to send queries based upon a species, genus or family name. All elements of a name are parsed and the model employed can handle even the most complex name. Fields for 'number of words in name' and 'order in name' allow extraction of individual genus names and reconstitution of name strings. The 'Word Type' will usually be a

name of a taxonomic rank (e.g. 'family' or 'species'). Instead of creating a new WORD_TYPE table, we have used the existing TAXON_RANK table to fulfil this role and have added three new entries: 'vernacular name', 'excluded word' (for elements of a name such as *'sensu'* or 'near') and 'rank name' (for infraspecific rank names such as 'var.').



WORD_IN_NAME table

Field Name	*Description of content*	*Example of content*
WORD_IN_ITEM_NAME_KEY	Primary Key	NHMSYS0000882968
WORD_KEY	Foreign Key to WORD table	NHMSYS0000875353
TAXON_KEY	Foreign Key to TAXON table	NHMSYS0000873595
WORD_TYPE_KEY	Foreign Key to TAXON_RANK table	NBNSYS0000000028
ORDER_IN_NAME	Position of word within name	1
WORD_Nos	Allows list to be output in original order	2

WORD table

Field Name	*Description of content*	*Example of content*
WORD_KEY	Primary Key	NHMSYS0000882968
WORD	An element in taxon name	luteicornis

TAXON table

Field Name	*Description of content*	*Example of content*
TAXON_KEY	Primary Key	NHMSYS0000873595
ITEM_NAME	Name String	Anteon luteicornis
AUTHORITY	Naming authority + Date	Kieffer, 1905

TAXON_RANK table

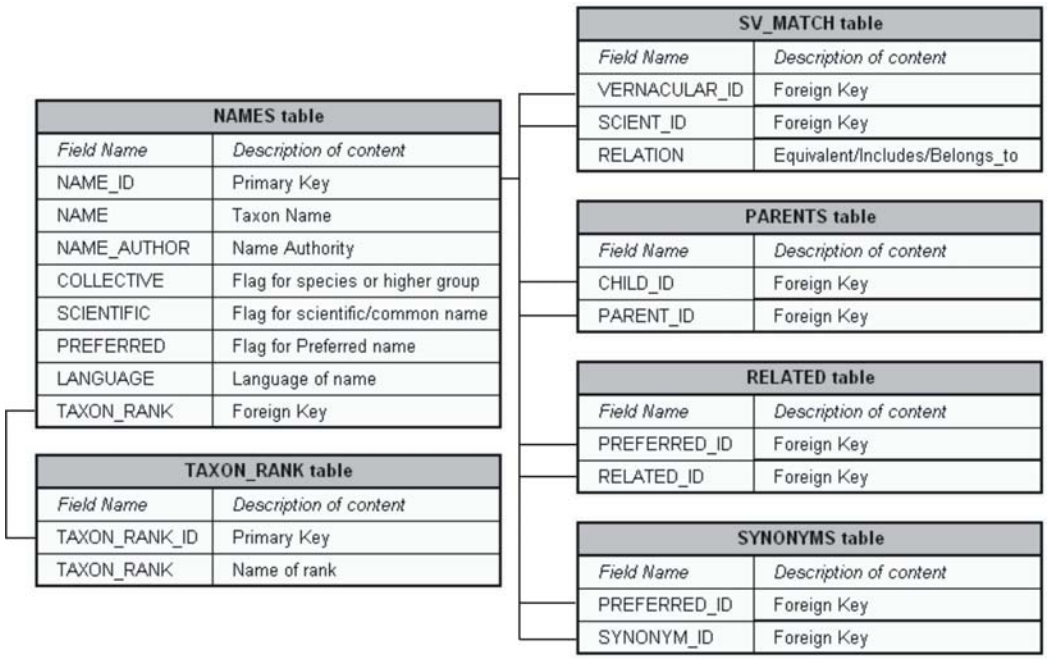
Field Name	*Description of content*	*Example of content*
TAXON_RANK_KEY	Primary Key	NBNSYS0000000028
LONG_NAME	Name of rank	Species

Figure 3. Model for parsing elements in a name

The mappings that we have undertaken have proved invaluable but it should be noted that they operate at the level of names and not at the level of individual identifications. Within the Dictionary, which deals with checklists, this is not a problem. Other NBN products, which deal with observational or collections data, such as the Gateway and Recorder do need to handle misidentifications.

3.2 The Nature Navigator project

Nature Navigator (www.nhm.ac.uk/naturenavigator/) draws upon data in the Species Dictionary but presents them to a general audience. Specifically, it has been designed to allow entry through familiar names, particularly informal names for higher groupings (trees, mammals, crabs, etc.) and to encourage users to explore relationships between organisms. It only includes species with common names and currently holds 27,200 names, as opposed to 198,000 in the Species Dictionary. The data model is considerably simpler than the Dictionary. Once again, all names are stored in a single table. Each name is associated with a source but there are no versions of sources or of names. Vernacular names are mapped to scientific equivalents in the SV_MATCH table. Parent-Child relationships are mapped for scientific names, but the hierarchy for vernacular names is built on the fly when results are displayed. Nature Navigator performs the function of a thesaurus and has tables to record synonyms (use/use-for relationships) and related names.



NAMES table

Field Name	*Description of content*
NAME_ID	Primary Key
NAME	Taxon Name
NAME_AUTHOR	Name Authority
COLLECTIVE	Flag for species or higher group
SCIENTIFIC	Flag for scientific/common name
PREFERRED	Flag for Preferred name
LANGUAGE	Language of name
TAXON_RANK	Foreign Key

TAXON_RANK table

Field Name	*Description of content*
TAXON_RANK_ID	Primary Key
TAXON_RANK	Name of rank

SV_MATCH table

Field Name	*Description of content*
VERNACULAR_ID	Foreign Key
SCIENT_ID	Foreign Key
RELATION	Equivalent/Includes/Belongs_to

PARENTS table

Field Name	*Description of content*
CHILD_ID	Foreign Key
PARENT_ID	Foreign Key

RELATED table

Field Name	*Description of content*
PREFERRED_ID	Foreign Key
RELATED_ID	Foreign Key

SYNONYMS table

Field Name	*Description of content*
PREFERRED_ID	Foreign Key
SYNONYM_ID	Foreign Key

Figure 4. Simplified representation of data structure for Nature Navigator project (not all fields or tables are shown)

4 DISCUSSION

There are a number of possible architectures for integrating disparate data sources (examined in Hussey, 2003). In all cases, the aim should be to implement systems that return all relevant data in response to a search term. Most solutions implement a Common Access System (CAS) through a web portal. If this implements searching of remote datasets, as is done by the Taxonomic Search Engine (Page, 2005) and Species 2000 (n.d.), then there remains a likelihood that variant records in the remote datasets will be missed. Only when the entire contents of the datasets are copied into a central pool can effective checks be made for variant names and these either corrected or mapped to good names. Maintaining copies of datasets can lead to version control problems, however. Applications such as BioCASE and GBIF possess automated harvesting mechanisms that can poll remote datasets at regular intervals. However frequent polling requires an increased effort in validation of the data received. Larger institutions may find that they are asked to wrap their data for several different projects (Natural History Museum, London, has wrapped data for ENHSIN, BioCASE, Species 2000 and GBIF). Emerging common schema and protocols such as TAPIR (n.d.), may help to reduce the burden on data owners. However, it is our experience that smaller data providers often lack the technical skill to wrap their data to schemas, export as XML, or even to provide 24-hour access to their data. The onus should be on the larger initiatives to support their data providers.

Should effort be directed to cleaning data or mapping variant data? This becomes a cost/benefit issue and therefore one of directing resources where they will be most effective. It can be handled at three levels:

i) Data Providers:
It is in the interest of data providers/owners to clean their data as far as possible. This is best done at the data entry stage (Chapman, 2005b) and will be helped by access to good authority files for taxonomic names. Validation of names can be helped by recording sources used for determination and by preservation of voucher specimens. Some collections and observation data holdings, however, may be too large to systematically inspect. Another potential problem is that some data owners may not have the required taxonomic expertise required to keep their data abreast of changes in nomenclature and classification, or may lack access to taxonomic literature.

ii) Data Collators:
Portal Managers may be in a position to provide the mapping between names if they have access to the total contents of datasets. As custodians of data they may not have editing rights, but they would be able to run validation checks and report errors back to the data providers for attention. They may also have responsibility for the long-term preservation of the data (Chapman, 2005c), particularly where data has resulted from fixed-term projects. Good metadata are required to establish fitness-for-purpose and the quality of the data. A Common Access System with an intelligent core may be needed; for instance to avoid sending fish queries to a plant database. Portal Managers may well find that they need to maintain lists of names, which can become useful authority files (Jones, Xuebiao Xu, Pittas, Gray, Fiddian, White, et al. 2000). Again, however, they may lack taxonomic resources to make judgements on the validity of the names.

iii) Users:
Any publicly accessible dataset should be provided with a feedback mechanism to report errors. Exposing raw data to public scrutiny is often a good way to elicit error correction (Edwards, 2004)- although it is important that users are made aware of the quality of the data (and that this may vary between different datasets). Collections databases in particular may well rely on input from visiting researchers to add value and remove errors.

Several topics merit further discussion. Firstly, authority files (whether project specific, taxon group specific, regional or global in scope) can serve as a useful resource to validate names. A potential drawback of having an authority file within a project though is that it may duplicate work elsewhere or become out of date. However, project authority files can increase their usefulness if they can feed into larger initiatives such as Species 2000 or GBIF. Taxon group lists are often maintained by experts and have a high level of reliability, but do not usually show homonyms with other groups. Regional lists will not include different names used in other parts of the world and whilst global lists hold much promise, they are thus far only complete for a small proportion of the world's biota. GBIF and uBio are compilations and may, at present, lack the rigorous checking mechanisms required to flag valid current names and map synonyms. Certainly, it is not possible to distinguish such names in the present GBIF portal. The Catalogue of Life alliance of Species 2000 and the Integrated Taxonomic Information Service (ITIS) (www.itis.usda.gov), on the other hand, do have experts validating their content. At present, the nomenclator services that are available are generally only accessible through a web interface, but it

is to be hoped that semantic web technologies and web services will soon allow automated interrogation of the underlying data to provide real-time checking of the status of names (Stein, 2002).

Tracking the usage of names through taxon concepts is also of considerable importance and has been the subject of a number of research projects. As yet though, none has evolved into a sustainable operational service. Prometheus (Raguenaud, Kennedy & Barclay, 1999), Nomencurator (n.d. & Ytow, Morse & Roberts, 2001), Taxonomer (Pyle, 2004), and SEEK (2005) use different approaches to the task. A schema to handle taxonomic concepts is currently being produced as part of the SEEK project and is being proposed as a Taxonomic Databases Working Group Standard (Kennedy, Kukla & Paterson 2005).

Finally, it will be important to map equivalencies in the use of names (Patterson, 2003). One way of tying down the use of a name is through a Globally Unique Identifier (GUID). This usually takes the form of a Universal Resource Name (see W3C, 2001 for information). Life Science Identifiers (LSID Resolution Protocol Project, 2005) are being used within the Bioinformatics community and may extend to the Biodiversity Informatics community, and an alternative system of Digital Object Identifiers (Paskin, 2005) is also now being trialled with biological nomenclature (Garrity & Lyons, 2003).

5 CONCLUSION

Real-world data found in collections and observations databases demonstrate a need to be able to distinguish currently accepted scientific names and to map them to related names, including vernacular names (in multiple languages), synonyms that are in use, archaic names, different usages of a name (taxon concepts) and variants of a name (different spellings). Current applications have hardly begun to tackle this problem, but a solution can be provided by quite simple data structures. Sustainability is likely to be a major issue: taxonomic revisions, as well as the describing of new species, mean that mapping will need to be kept under constant review. Maintaining databases long-term is possible only in major institutions or government agencies. Collaboration in a global solution presents the best option and common exchange schemas and protocols, together with data that are accessible to other applications through web services, will increase the usefulness of these authority files. Finally, there is always a need for education in biological nomenclature, so as to encourage best practice in the use of names.

6 ACKNOWLEDGEMENTS

We should like to thank Charles Copp for sharing his ideas on the data modelling underlying the National Biodiversity Network data model and the BioCASE thesaurus. Nature Navigator was funded through a grant from the New Opportunities Fund and Species Dictionary has received funding from the Department for the Environment, Food and Rural Affairs, Scottish Natural Heritage and the Joint Nature Conservation Committee.

7 REFERENCES

ABCD (n.d.) ABCD Schema - Task Group on Access to Biological Collection Data. Available at: http://www.bgbm.org/TDWG/CODATA/Schema/default.htm

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

BIOCASE (n.d.) A Biological Collection Access Service for Europe. Project Home page available at: http://www.biocase.org/

Burnett, J., Copp, C. & Harding, P. (1995) *Biological recording in the United Kingdom* (Vol. 1). London: Department of the Environment.

Chapman, A. (2005a) *Uses of primary species-occurrence data*. Retrieved 20 September 2005 from *Global Biodiversity Information Facility* website: http://www.gbif.org/prog/digit/data_quality

Chapman, A. (2005b) *Principles and methods of data cleaning*. Retrieved 20 September 2005 from *Global Biodiversity Information Facility* website: http://www.gbif.org/prog/digit/data_quality

Chapman, A. (2005c) *Principles of data quality*. Retrieved 20 September 2005 from *Global Biodiversity Information Facility* website: http://www.gbif.org/prog/digit/data_quality

Copp, C. (2000) *The NBN data model and its implementation in Recorder 2000*. Newark: The NBN Trust.

Copp, C. (2004a) *The NBN Data Model*, Part 1 *Description of the Model*. Newark: The NBN Trust.

Copp, C. (2004b) *The NBN Data Model*, Part 2 *Physical Implementation of the Model*. Newark: The NBN Trust.

Darwin Core (n.d.) Taxonomic Databases Working Group: Darwin Core 2. Retrieved January 26, 2006 from: http://darwincore.calacademy.org/

DiGIR (2005.) Distributed Generic Information Retrieval (DiGIR). Homepage available at: http://digir.net/

Edwards, J.L. (2004) Research and Societal Benefits of the Global Biodiversity Information Facility. *BioScience 54*(6), 485-486.

Franki, R.I.B., Fauquet, C.M., Knudson, D.L. & Brown, F. (1990) Classification and nomenclature of viruses. *Archives of Virology Supplement 2*, 1-445

Froese, R. & Pauly, D. (Eds). (2000) *FishBase 2000: concepts, design and data sources*. Los Baños, Laguna, Philippines: ICLARM.

Garrity, G.M. & Lyons, C. (2003) Future-proofing biological nomenclature. *Omics*, *7*(1), 31-33. Retrieved 20 September 2005 from: http://www.eecs.umich.edu/~jag/wdmbio/garrity.htm

GBIF (n.d.) Homepage of Global Biodiversity Information Facility. Available at: http://www.gbif.org

Geoffroy, M. & Berendsohn, W. (2003) The concept problem in taxonomy: importance, components, approaches. *Schriftenreihe Vegetationsk. 39*, 15-42

Greuter, W., McNeill, J., Barrie, R., Burdett, H.-M., Demoulin, V., Figuerias, T.S., Nicolson, D.H., Silva, P.C., Skog, J.E., Trehane, P., Turland, N.J. & Hawksworth, D.L. (Eds.) (2000) *International code of botanical nomenclature*. Königstein: Koeltz Scientific Books

Hussey, C.G. (2003) Technical parameters: How to make ENHSIN workable. In Scoble, M.J. (Ed.) *ENHSIN The European Natural History Specimen Information Network* (pp 102-132). London: The Natural History Museum.

International Commission on Zoological Nomenclature, (1999) *International code of zoological nomenclature*. 4th Edition. London: The International Trust for Zoological Nomenclature

Jacquin, J.F. de (1787) *Lacerta vivipara*, observatio. *Nova Acta Helvet. 1*, 33-34.

Jones, A.C., Xuebiao Xu, Pittas, N., Gray, W.A., Fiddian, N.J., White, R.J., Robinson, J., Bisby, F.A. & Brandt , S.M. (2000) SPICE: A flexible architecture for integrating autonomous databases to comprise a distributed Catalogue of Life. In Proc. 11th International Conference on Database and Expert Systems Applications. *Lecture Notes in Computer Science*, *1873*, 981-992.

Kennedy, J., Kukla, R. & Paterson, T. (2005) Wiki for Taxonomic Concept Transfer Schema. Retrieved January 26, 2006 from http://tdwg.napier.ac.uk/.

Lapage, S.P., Sneath, P.H.A., Lessel, E.F., Skerman, V.B.D., Seeliger, H.P.R. & Clark, W.A. (Eds.) (1992) *International Code of Nomenclature of Bacteria (Bacteriological Code 1990 Revision)*. Washington: Amer. Soc. Microbiol.

Linnaeus, C.von (1758) *Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Editio decima, reformata. Holmiæ: Laurentii Salvii.

LSID Resolution Protocol Project (2005) Homepage for the LSID project. Available at: http://lsid.sourceforge.net

May, R.M. (1988) How many species are there on earth? *Science, 247*, 1441-49.

Nomencurator (n.d.) Home page for the Nomencurator project. Available at http://www.nomencurator.org/

Page, R.D.M. (2005) A taxonomic search engine: federating taxonomic databases using web services. *BMC Bioinformatics*, 6, 48. Retrieved 20 September 2005 from: http://www.biomedcentral.com/1471-2105/6/48

Paskin, N. (2005) Digital Object Identifiers for scientific data. *Data Science Journal, 4*, 12-20. Retrieved 20 September 2005 from: http://www.datasciencejournal.org/

Patterson, D.J. (2003) Progressing towards a biological names register. *Nature, 422*, 661

Pyle, R.L. (2004) Taxonomer: a relational data model for managing information relevant to taxonomic research. *Phyloinformatics, 1*, 1-54.

Raguenaud, C., Kennedy, J. & Barclay, P. (1999) Database support for taxonomy. *Prometheus Technical Report #1*, Edinburgh: School of Computing, Napier University. Retrieved 20 September 2005 from: http://www.dcs.napier.ac.uk/~osg/prometheus/prometheus_1/publications/pro99a.pdf

Richardson, J. (1836) *Fauna Boreali-Americana; or the zoology of the northern parts of British America : containing descriptions of the objects of natural history collected on the late northern land expeditions, under the command of Sir John Franklin, R.N. Third Part: The Fish*. London: Richard Bentley.

Scoble, M.J. (2003) Changing roles and perceptions in European natural history collections: from idiosyncrasy to infrastructure. In Scoble, M.J. (Ed.) *ENHSIN The European Natural History Specimen Information Network* (pp 11-20). London: The Natural History Museum.

SEEK (2005) Homepage for the Science Environment for Ecological Knowledge initiative. Available at: http://seek.ecoinformatics.org/

Species 2000 (n.d.) Home page for Species 2000 organisation. Available at: http://www.sp2000.org

SPICE (n.d.) Home page for SPICE project. Available at: http://spice.sp2000europa.org/SPICE/

Stein, L. (2002) Creating a bioinformatics nation. *Nature*, *417*,119-120

TAPIR (n.d.) FrontPage Wiki for the TAPIR [TDWG Access Protocol for Information Retrieval] Protocol. Retrieved January 26, 2006 from the Tapir Protocol Wiki website: http://ww3.bgbm.org/protocolwiki/

TDWG (n.d.) Homepage for the Taxonomic Databases Working Group. Available at: http://www.tdwg.org

Trehane, P., Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Spongberg, S.A. & Vrugtman, F. (1995) *International code for cultivated plants*. Wimborne: Quarterjack Publishing

uBio (n.d.) Universal Biological Indexer and Organiser. Home page available at: http://www.ubio.org

W3C (2001) *URIs, URLs, and URNs: Clarifications and Recommendations 1.0.* Report from the joint W3C/IETF URI Planning Interest Group. W3C Note 21 September 2001 Retrieved January 26, 2006 from: http://www.w3.org/TR/uri-clarification/

Walbaum, J.J. (1792) Ichthyologiae pars III .In Artedi, P. *Petri Artedi sueci genera piscium. In quibus systema totum ichthyologiae proponitur cum classibus, ordinibus, generum characteribus, specierum differentiis, observationibus plurimis. Redactis speciebus 242 ad genera 52*., Grypeswaldiae [Greifswald]: Ant. Ferdin. Rose

Wilson, E.O. (2000) A global biodiversity map. *Science*, *289*, 2279

Ytow, N., Morse, D.R. & Roberts, D.McL., (2001) Nomencurator: a nomenclatural history model to handle multiple taxonomic views. *Biol.J.Linn.Soc*., *73*, 81-98.