The BODC Taxon Browser – A powerful search tool for the discovery of taxonomic information

Michael Hughes and Roy Lowry

British Oceanographic Data Centre
Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, United Kingdom.
E-mail: mhug@bodc.ac.uk

Abstract

The BODC Taxon Browser is used for the discovery of descriptive information about the biological taxa that are held in the archives of the British Oceanographic Data Centre (BODC). The main aspect of its functionality is the capability of hierarchical searching. This means that on entering a taxonomic search term, the browser will retrieve information for the term and information for all organisms taxonomically related to the term. The framework of the hierarchical search is derived from the Integrated Taxonomic Information System (ITIS). Other features include: automatic checking for synonyms where taxonomic names are not valid; search by ITIS code, BODC code, scientific name and common name; filter options to focus the search. This search tool is powerful – it goes beyond simply matching the exact term that is entered and so discovers information which users may not even realise they are looking for!

Keywords: Taxonomy; Hierarchical searching; Metadata discovery; Parameter dictionary; The Integrated Taxonomic Information System.

Introduction

The BODC Taxon Browser is a web-based search tool which enables easy navigation through the substantial amount of biological data which is managed by BODC. BODC holds nearly 400,000 biological records spanning 4,500 different taxa. Measurements range from the biomass of microscopic algae in the Indian Ocean to counts of Cetaceans in the North-East Atlantic. Each record is linked to detailed metadata (*e.g.* **what** was measured, **how** it was measured, **who** measured it, **where** and **when** it was measured) via a relational database management system (RDBMS). One branch of the RDBMS is a parameter dictionary which contains the "what" and "how" metadata. The BODC Taxon Browser allows the user to efficiently access this metadata for every taxonomic entity that is described within this parameter dictionary.

The parameter dictionary is constantly expanding and currently describes nearly 17,000 types of variables, 11,000 of which originate from taxonomic samples. On entering a taxonomic search term, the browser will retrieve metadata for the term and the metadata for all organisms taxonomically related to the term. So, if you search for "birds" you'll also get "penguins". This works even if the actual search term is not present in the dictionary. For example, a search for "*Cetacea*" (an Order of marine mammals) will return metadata for every genus and species of dolphin that is held in the dictionary even

though the word "*Cetacea*" does not appear in any part of the dictionary. This is particularly helpful if users have only a general idea of the taxonomic information they are looking for.

Where possible, every taxonomic entity in the parameter dictionary is mapped to an entry in the Integrated Taxonomic Information System (ITIS). This gives credibility to the taxonomic information at BODC as ITIS is an authoritative taxonomic resource and it also enables the Taxon Browser to use the ITIS taxonomic hierarchy as a framework for the "intelligent" searching described above.

Other Taxon Browser features built on the ITIS framework are:

- Automatic checking for synonyms where taxonomic names are no longer considered to be valid
- Searching by common names.

To make the Taxon Browser functional, a number of stages of development were required, namely:

- Downloading a local copy of the complete ITIS database
- Mapping all BODC taxonomic entities to their equivalent taxa in ITIS
- Generating a taxonomic hierarchy from the ITIS tables
- Creating a web-based search interface to enable dynamic interaction between the user and the database.

This paper will outline these aspects of development and give some working examples of using the Taxon Browser.

Incorporating ITIS

ITIS provides reliable information on species names and their hierarchical classification. The database is regularly reviewed to add newly described species and to keep up-to-date with the validity of taxonomic classifications. Every scientific name in ITIS is accompanied by the author and date, taxonomic rank, associated synonyms and vernacular names where appropriate, data source information, data quality indicators and a unique taxonomic serial number (TSN). The TSN becomes the label for what is known as a "Taxonomic Unit". In order to integrate the BODC taxonomic data with the ITIS taxonomic units, every taxonomic entity at BODC is assigned a TSN where available (see section 1.2).

Downloading ITIS

The ITIS database is updated on a monthly basis and the latest version can be freely downloaded from: ftp://ftp-fc.sc.egov.usda.gov/ITIS/. The folder at this location contains text-only versions of all the tables used in the ITIS database and an SQL file with the necessary code to set up the tables locally. The ITIS tables were downloaded according

to the instructions at: http://www.itis.usda.gov/ftp_download.html. This process is repeated every two months to keep up with the latest updates to the ITIS database.

The ITIS table of relevance to the Taxon Browser are:

- ITIS Taxonomic Units – includes scientific names, usage information, hierarchy information, links to references, credibility information
- ITIS Vernaculars – common names in various languages which are linked via a TSN to their relevant taxonomic units
- ITIS Synonym Links – where taxonomic name usage is considered invalid (for animals and bacteria) or not accepted (for plants and fungi), the valid or accepted alternative is contained here.

The BODC-ITIS map

Once local copies of the ITIS tables were generated, the scientific taxon names at BODC could be automatically mapped to the ITIS taxonomic units. To do this, a series of SQL statements in Oracle were used to identify matches between the BODC scientific names and the ITIS scientific names. The relevant ITIS TSN was then appended to all BODC parameter dictionary entries describing sampling events for that taxon.

Occasionally, automatic mapping was not successful and records needed to be manually checked. The two main reasons for this were:

- Spellings differed between BODC and ITIS (usually due to spelling mistakes)
- A scientific name did not appear in ITIS.

Where spellings differed slightly (*e.g. Chaetoceros pelagicum* vs. *Chaetoceros pelagicus*), the BODC name was altered to comply with ITIS. Where scientific names appeared in BODC but not in ITIS, the names were submitted to ITIS according to the guidelines at http://www.itis.usda.gov/submit_guidlines.html. The web resources used to find information for the submission of species names to ITIS were: Algaebase (Guiry and Nic Dhonncha, 2005), The Ciliate Resource Archive (Lynn, 2003), The World of Protozoa, Rotifera, Nematoda and Oligochaeta (Inamori, 2003) and the Check-list of Turkish Seas Microplankton (Koray *et al*, 1999).

The BODC-ITIS map fulfils a number of roles:

- It provides a link to ITIS taxonomic units, common names, synonyms and any other useful information provided by ITIS
- The columns of the table are arranged into "semantic elements" which are the building blocks that are used to automatically generate descriptive full titles in the parameter dictionary.

An example of a parameter dictionary full title is:

Carbon biomass of Bacillariophyta (ITIS 2286) centric 30um per unit volume of the water column by optical microscopy and abundance to carbon conversion by unspecified algorithm

The semantic elements for this title are:

- Parameter: Carbon biomass
- Taxon name: Bacillariophyta
- Taxon code: 2286
- Taxon class: centric 30um
- Parameter compartment: per unit volume of the
- Compartment: water column
- Compartment class: not specified (hidden)
- Sample preparation: not specified (hidden)
- Analysis: optical microscopy
- Data processing: abundance to carbon conversion by unspecified algorithm

The full title was created by the concatenation of the fields in the BODC-ITIS map to produce a humanly readable sentence. The fields of the BODC-ITIS map contain the semantic elements. Every full title generated has the same arrangement of semantic elements to maintain consistency and accuracy when creating parameter definitions. Each semantic element is part of a controlled vocabulary meaning that there are a limited number of options for the words that can be used for any particular semantic element. For example, "Parameter" can have values including, "Carbon biomass", "Abundance" and "Count"; "Taxon name" must be the same as the name published in ITIS where available; "Compartment" can be "bed", "sediment", "water column" or "suspended particulate matter". The elements are joined by linking words such as "of" or "in the". This system permits a rich vocabulary of definitions for the description of the many forms of data at BODC and is also easily machine readable.

Building the taxon tree

After populating the BODC-ITIS map, a program was written to generate a hierarchy of all the ITIS taxonomic units. This forms the framework of the Taxon Browser's hierarchical search. It enables the Taxon Browser to navigate the ITIS hierarchy and extract every taxonomic entry in the BODC dictionary at the same taxonomic level as the search term and also all the related taxa at lower levels.

The hierarchy was built using a field of the ITIS Taxonomic Units table called "Parent TSN". This is a direct link between every ITIS taxonomic unit and the taxonomic unit directly above it in the ITIS taxonomic hierarchy. For example, the species *Chaetoceros pelagicus* has a parent TSN of "2758". The scientific name with a TSN of 2758 is the genus *Chaetoceros*. *Chaetoceros* has a parent TSN of "572759" which is the TSN for Family Chaetocerotaceae. This path can be followed all the way to the kingdom level.The representation of the ITIS taxonomic hierarchy that is generated by the tree-building program consists of a source TSN followed by a string of 216 bits (see figure 1). Every 8 character section of the string represents a single taxonomic level. The

default setting of the hierarchy string is 216 x's. This equates to 8 x's per taxonomic level. A string is populated by placing the source TSN in the position on the hierarchy string appropriate to its taxonomic level. For example, the TSN for a genus name is placed at character positions 137-144 in the hierarchy string; the TSN for a species name is placed at positions 169-176. When the source TSN is in place, every parent TSN for this is inserted into the string in the appropriate place. A separate hierarchy string is created for every taxonomic unit in the ITIS database.

Source TSN: 2814
Hierarchy:
xx202422xx590735xxxx2286xxxxxxxxxxxxxxxxxxx590736xxxxxxxxxxxxxxxxxxxxx
xxxxx590748xxxxxxxxxxxxxxxxxxxxxxxxxxx572759xxxxxxxxxxxxxxxxxxxxxxxxx
xx2758xxxxxxxxxxxxxxxxxxxxxxxxxxxx2814xxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxx

Fig. 1. The hierarchy string for species Chaetoceros pelagicus. *The species portion of the hierarchy string is represented by "xxxx2814". Its taxonomic parent is represented by "xxxx2758", which is the TSN for genus* Chaetoceros. *Continuing a walk up the string identifies every parent taxon for* Chaetoceros pelagicus *up to the kingdom level (202422, kingdom Plantae). Every empty taxonomic position is left as 8 x's to maintain the length of the hierarchy string. This becomes useful later for checking for TSNs in specific taxonomic positions. For example, in between 572759 and 2758 are unoccupied taxonomic levels that represent subfamily, tribe and subtribe; after 2814 are unoccupied taxonomic levels that represent subspecies, variety, subvariety, form and subform.*

Continuing the example from figure 1, there are many other species of *Chaetoceros*, all having identical hierarchy strings from the kingdom to the genus level, *i.e.* from character positions 1 to 144. This means that every hierarchy string with "xxxx2758" at the genus level is exactly the same from positions 1 to 144 and also represents a hierarchy for a species, subspecies, variety, subvariety, form or subform of the genus *Chaetoceros*. Selecting the source TSN for all these partially matching hierarchies allows the Taxon Browser to retrieve all the relevant parameter dictionary records regarding members of the *Chaetoceros* genus.

From a wider perspective, any hierarchy string that has a TSN ending at character position *n* shares the same hierarchy from positions 1 to *n* as any other hierarchy string that has the same TSN ending at character position *n*. A hierarchy string can be sliced at the boundary of any 8-character section to widen or narrow the range of source TSNs that are extracted by the Taxon Browser. This allows the user to enter a search term at any taxonomic level and retrieve all of its relatives. A search at the kingdom level, for example will extract all taxa from BODC that have matching hierarchy strings for positions 1 to 8 for any particular kingdom. It is possible to type in "Animalia" and retrieve every taxon at every available taxonomic level for entities considered to be animals even though there is no entry in the parameter dictionary that is explicitly described as being an animal. Similar searches can be conducted for "Aves" or "Birds" or even "Oiseaux" or perhaps the user would like to narrow the search slightly and look

for "Gulls". Again, these are all words that do not appear anywhere in the parameter dictionary but will still yield comprehensive results.

This method works for searching all taxa that occur in the ITIS database. However, a special case arises when a taxon is considered "invalid" or "not accepted" by ITIS. For these cases, no parent TSN is provided in ITIS Taxonomic units and so there is no basis for the tree builder to generate a hierarchy string. For the hierarchical search to be fully functional, the taxonomic hierarchy must be derived from the "accepted" or "valid" synonym. Using the ITIS Synonym Links table, it is possible to find the valid TSN and superimpose its hierarchy onto the invalid TSN. An alternative strategy could be to change all invalid names at BODC into their valid synonyms but it is better to alter the data coming into BODC as little as possible. Using the ITIS Synonym Links table means that this can be avoided and the user can be informed that a particular taxonomic name is not valid and of its valid usage.

Another special case that arises from an invalid synonym being assigned the parental hierarchy of its valid counterpart is when the adopted parent is at the same taxonomic level as the invalid name. For example, **class** Solenogastres is invalid. Its valid synonym is **subclass** Chaetodermomorpha. The parent taxon for both these names is **class** Aplacophora. Since a class can not have a taxonomic parent which is also a class, the original TSN is loaded into the hierarchy string in the position that the valid synonym occupies. The program that builds the taxon tree table takes all of these issues into account and automatically adapts to the special cases.

Using the Taxon Browser

This section concerns the client-side of the Taxon Browser and describes the user interface and the various search options and features that the browser provides.

User interface

The user interface (see figure 2) was written in Perl and uses the Common Gateway Interface (CGI). CGI enables interaction between a client and a server via the World Wide Web. The program is embedded with HTML to provide the front-end graphics on the Web and SQL to fetch information from the Oracle database at BODC. The user can select different searching methods (see section 2.2), apply options to filter the search (see section 2.3) and search within BODC or ITIS.

Search options

Search by scientific name:

- The Taxon Browser matches the scientific name in ITIS Taxonomic Units and extracts its TSN
- The taxonomic hierarchy is scanned to select every source TSN with a hierarchy string that includes the search TSN

- Every TSN selected is added to an array
- The parameter dictionary metadata for each TSN in the array is extracted and displayed to the user (see figure 3).

Search by common name:

- ITIS Vernaculars is scanned for the TSN belonging to the matching common name
- The search follows the last two steps of "Search by scientific name".

Search by BODC code:

- BODC-ITIS map is scanned for the TSN belonging to the matching BODC parameter code (a unique identifier for each entry in the parameter dictionary)
- The search follows last two steps of "Search by scientific name".

Search by ITIS TSN

- The search follows last two steps of "Search by scientific name".

Features

Automatic synonym conversion:

- Where a search term is invalid or not accepted, the Taxon Browser takes the alternative accepted TSN from ITIS Synonym Links
- The taxonomic hierarchy is scanned to select every source TSN with a hierarchy string that includes the original search TSN or the accepted TSN
- The search follows steps 2-3 of "Search by scientific name".

Filter by parameter type

- The user can adapt the search to select only those dictionary records that deal with abundance or biomass data, for example.

Hierarchical / non-hierarchical searching

- The user can select whether to perform a full hierarchical search as described previously or to retrieve only the parameter dictionary metadata for the actual search term.

Advanced options

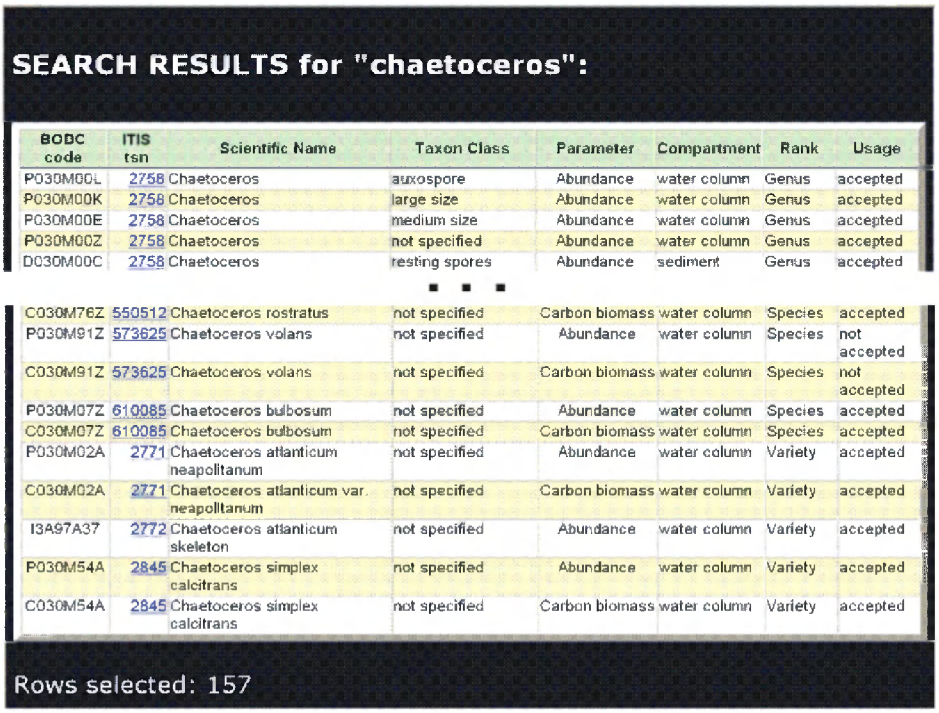
- The user can customise the items of parameter dictionary metadata to be displayed by the Taxon Browser. For example, the user may be interested in how samples

where analysed but not interested in how the data was processed and can change the settings accordingly.

Search ITIS

- For a direct link to the complete taxonomic information in ITIS, a search can be directed to the ITIS website.

Fig. 2. The Taxon Browser search interface.



SEARCH RESULTS for "chaetoceros":

BODC code	ITIS tsn	Scientific Name	Taxon Class	Parameter	Compartment	Rank	Usage
P030M00L	2758	Chaetoceros	auxospore	Abundance	water column	Genus	accepted
P030M00K	2758	Chaetoceros	large size	Abundance	water column	Genus	accepted
P030M00E	2758	Chaetoceros	medium size	Abundance	water column	Genus	accepted
P030M00Z	2758	Chaetoceros	not specified	Abundance	water column	Genus	accepted
D030M00C	2758	Chaetoceros	resting spores	Abundance	sediment	Genus	accepted
C030M76Z	550512	Chaetoceros rostratus	not specified	Carbon biomass	water column	Species	accepted
P030M91Z	573625	Chaetoceros volans	not specified	Abundance	water column	Species	not accepted
C030M91Z	573625	Chaetoceros volans	not specified	Carbon biomass	water column	Species	not accepted
P030M07Z	610085	Chaetoceros bulbosum	not specified	Abundance	water column	Species	accepted
C030M07Z	610085	Chaetoceros bulbosum	not specified	Carbon biomass	water column	Species	accepted
P030M02A	2771	Chaetoceros atlanticum neapolitanum	not specified	Abundance	water column	Variety	accepted
C030M02A	2771	Chaetoceros atlanticum var. neapolitanum	not specified	Carbon biomass	water column	Variety	accepted
I3A97A37	2772	Chaetoceros atlanticum skeleton	not specified	Abundance	water column	Variety	accepted
P030M54A	2845	Chaetoceros simplex calcitrans	not specified	Abundance	water column	Variety	accepted
C030M54A	2845	Chaetoceros simplex calcitrans	not specified	Carbon biomass	water column	Variety	accepted

Rows selected: 157

Fig. 3. A selection of metadata from the BODC parameter dictionary retrieved by searching for scientific name "Chaetoceros". Each column forms the semantic elements used to define a parameter.

Discussion

Taxonomic awareness

Within BODC, the linking of taxonomic entities to ITIS has forced us to become more aware of official taxonomic naming conventions and encouraged the production of a protocol for creating new taxonomic entries for the parameter dictionary. Part of this protocol is that every species name added to the parameter dictionary must comply with ITIS and must also be given an ITIS TSN wherever possible. This practice has dramatically increased the quality and credibility of the parameter dictionary definitions concerning taxonomic entities.

Semantic modelling

The semantic model was a breakthrough for the automatic mapping of BODC to ITIS. The laborious task of manually extracting and matching taxonomic names from previously semantically uncontrolled parameter definitions became an easy task via a simple SQL statement. The semantic model also has the potential to become the basis to create a web service for sharing taxonomic parameter information across the internet.

Taxonomic standards

A large part of the task when setting up the Taxon Browser was to submit species names to ITIS. This required extensive searching on the internet for accurate references and authorship information. It also highlighted the need for standardisation of scientific names within the taxonomic community. For example, a search in Google for "*Chaetoceros compressus*" gives 164 hits. A search for "*Chaetoceros compressum*" gives 119 hits. How does a non-expert user decide which name to use? Taking the name with the most hits seems an obvious strategy but this is by no means accurate. If everyone did this, an incorrect name could rise to precedence by a sort of runaway selection whereby an incorrect name with a few more hits than a correct name would be quoted more frequently. Consequently, the gap in hit numbers between the two names would widen, causing users to choose the incorrect name over the correct one with an increasingly high frequency.

A single comprehensive database that incorporates taxonomic information from a distributed community of experts could solve this type of problem. ITIS is a big step towards this but lacks the resources to keep up with the rate at which species names are submitted. Submissions made by BODC at the end of 2003 have still not made it into the ITIS database. With limited resources, ITIS must also focus on certain species groups meaning that some are not yet included in the database. For example, only two species of *Strombidium* (a marine ciliate) are listed in ITIS whereas there are 74 in the European Register of Marine species (ERMS, http://www.marbef.org/data/erms.php). The Taxon Browser encounters a problem here because BODC has 39 species of *Strombidium* which are not in ITIS and so have not been included in the taxonomic tree that the Taxon Browser searches through. A user of the Taxon Browser will not see that there is information on these *Strombidium* species at BODC unless they use the non-hierarchical search option.

It is a huge task to collate the names of every single described species and organise them into ever-changing taxonomic hierarchies. It is inevitable that different expert opinions will arise of where a species should fit into a hierarchy or what it should be called and perhaps this is beyond the scope of a single taxonomic information centre. An alternative to a single central database is a number of expert databases with their own particular focus on selected groups of organisms. There are a number of these available but uncertainty can arise when two databases disagree. For example, ITIS considers *Emiliania huxleyi* to be a not accepted synonym for *Coccolithus huxleyi*. However, ERMS accepts the name *Emiliania huxleyi* and gives it a non-accepted synonym of *Pontosphaera huxleyi* with no mention of *Coccolithus huxleyi*. So, which information source should be used? A search in Google gives 16,700 hits for *E. huxleyi* but only 168 hits for *C. huxleyi*. This shows that the general consensus is to use *E. huxleyi* but there must be a reason why this name appears as not accepted in ITIS. Having many separate on-line databases also makes searching less efficient as the user must search each database separately for taxonomic information. A useful tool would bridge these databases and search all of them in one go. The uBio Name Mapper (The Marine Biological Laboratory, 2004) works along these lines but does not cover a wide enough range of databases.

Conclusion

The BODC Taxonomic Browser has the potential to be a very useful tool for the discovery of data in the BODC archives. Some work is required to make it fully operational but when it is released on the web, it is expected that people from around the world will access it to view the types of taxonomic data at BODC. The mapping of names to ITIS was a very useful exercise in improving the BODC parameter definitions but it also highlighted the fact that BODC should be aware of a world of taxonomic databases beyond ITIS and consider how these could be integrated.

Acknowledgements

Thanks to: The ITIS team for producing a very valuable database and providing comprehensive documentation and support to get full use out of it. Steve Loch at BODC for his ideas on generating the hierarchy strings. Richard Downer at BODC for his advice on the environmental settings for using Perl CGI on the BODC computer system. Gwen Moncoiffé at BODC, Toby Tyrell at the Southampton Oceanography Centre, Sonia Batten at Plymouth Marine Laboratory, Jeremy Young at the Natural History Museum and Mike Guiry at the National University of Ireland, Galway for taxonomy advice during the manual mapping of BODC terms to ITIS.

References

Guiry M.D. and E. Nic Dhonncha. 2005. *AlgaeBase version 2.1*. World-wide electronic publication, National University of Ireland, Galway. Available online at http://www.algaebase.org. Consulted on 25 January 2005.

Inamori Y. 2003. The World of Protozoa, Rotifera, Nematoda and Oligochaeta. National Institute for Environmental Studies, Japan Environmental Agency. Available online at http://www.nies.go.jp/ chiiki1/protoz/. Consulted on 1 December 2003.

Koray T., S. Gokpinar, L. Yurga, M. Turkoglu and S. Polat. 1999. Microplankton species of Turkish Seas. Available online at http://bornova.ege.edu.tr/~korayt/plankweb/chklists.html. Consulted on 1 December 2003.

Lynn D.H. 2003. The Ciliate Resource Archive. Available online at http://www.uoguelph.ca/~ciliates. Consulted on 1 December 2003.

Costello M.J., P. Bouchet, G. Boxshall, C. Emblow and E. Vanden Berghe. 2004. European Register of Marine Species. Available online at http://www.marbef.org/data/erms.php. Consulted on 26 January 2005.

The Integrated Taxonomic Information System on-line database. Available online at http://www.itis.usda.gov. Consulted on 9 December 2004.

The Marine Biological Laboratory, Massachusetts, USA. 2004. The Universal Biological Indexer and Organizer (uBio) Name Mapper Available online at http://uio.mbl.edu/services/pleary_working/ treeserve.php. Consulted on 26 January 2005.