Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System

Tony Rees[1] and Y. Zhang[2]

[1]CSIRO Marine Research, GPO Box 1538, Hobart 7001, Australia
E-mail: Tony.Rees@csiro.au

[2]Institute of Marine and Coastal Sciences, Rutgers, The State University of New Jersey, 71 Dudley Road, New Brunswick, NJ 08901-8521, USA
E-mail: phoebe@marine.rutgers.edu

Abstract

The initial release of OBIS, the Ocean Biogeographic Information System, provided a distributed search mechanism to retrieve marine species distribution records from a range of remote data providers in real time, based on a match on species scientific name and other parameters if specified. This 'fully distributed' version 1 of OBIS was upgraded in 2004 to provide improved functionality, system response times, and metadata-level information on available data via the OBIS system, by the introduction of two new components, an 'OBIS Index' comprising a species name index and a spatial index, and a local cache of commonly queried attributes of OBIS data items, refreshed on a rolling basis from the remote data providers. The conceptual, implementation and performance aspects of these developments are described in the present paper.

Keywords: Biological information systems; Biogeography; Databases; Indexing / Spatial indexing; Distributed searching.

Introduction and OBIS version 1

OBIS, the Ocean Biogeographic Information System, is conceived as a two-, three- and ultimately four-dimensional atlas of marine species distributions based on globally distributed data holdings accessed via a central portal (Grassle and Stocks, 1999; Zhang and Grassle, 2003), and is also the designated information and data management component of the Census of Marine Life (for information on the latter, see www.coml.org/). Functionally, OBIS comprises a central portal – presently located at Rutgers University, New Jersey, and accessible via www.iobis.org – which communicates with the various remote data providers via standard web protocols (XML over HTTP), while the inevitable heterogeneity of database or file structures at the provider end is standardised using 'wrapper' or translation software which enables the portal to issue common requests to, and receive back data in a common format from, any provider connected to the system.

Version 1 of OBIS was constructed in late 2001 and went live on the Rutgers site in January 2002, using a fairly standard architecture for what is effectively a fully distributed system, as illustrated in Fig. 1.

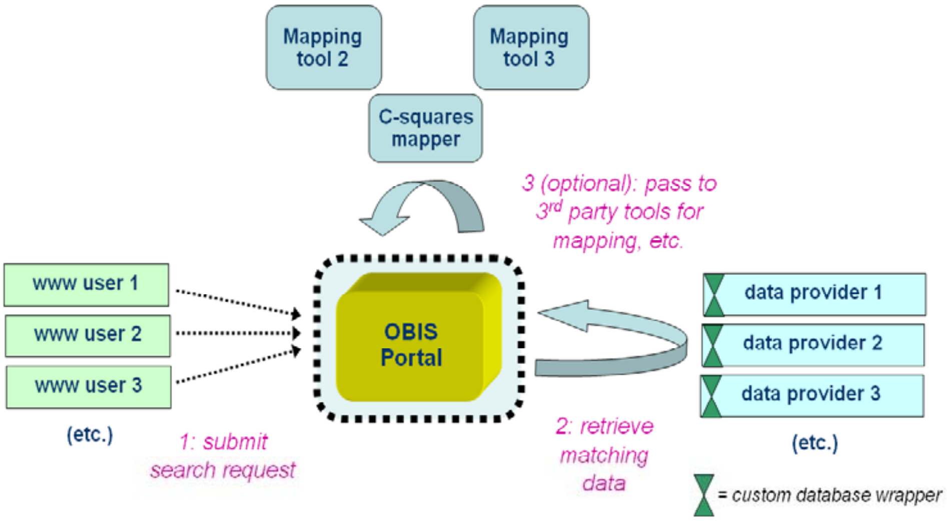


Fig. 1. Architecture of OBIS version 1, January 2002-February 2004.

In this architecture, the types of queries to be supported by the system are first designed and reflected in an XML data schema, then a custom wrapper is designed and installed at each remote data provider which will support queries on these terms and parameters (such as species scientific name, date and time of collection, water depth, and location by latitude and longitude) and return matching data to the portal upon request. Once this architecture is in place and connected via the internet, a user can submit a request for data (*e.g.* by species name or locality – labelled 1 in diagram), the portal issues the appropriate request to the remote data providers and merges any matching results returned into a single result set (2 in diagram), and the user then has further options (3 in diagram) to pass this result set to one of a number of available mapping or modelling tools as desired, or simply view or download the data to their own system for further investigation and user-specific operations.

An architecture such as this has advantages with respect to being relatively simple and rapid to design and implement, having a fairly ‘light’ footprint at the portal end so far as storage and maintenance are concerned, freeing the portal from any data custodianship issues, and no currency issues (information is always as up-to-date as at the remote providers). On the other hand, there are also problems with such systems which quickly become apparent in practice, and can result in a less than ideal user experience. Specific problems can be identified as follows:

- Reliability. The system is only as reliable as the ‘up time’ of all the contributing databases allows. If a remote data source is down, it cannot be searched.

- Speed. The system is only as fast as its slowest contributor to respond, and / or the bandwidth of the physical link to the same. Often the wait for a response may be set to a default timeout (*e.g.* 2 minutes), which means in practice that many searches take this long even if they return no data.
- User information. There is no information presented to the user in advance as to what is the coverage of the system – which species have associated data and are therefore worth searching on, and the size of the resulting dataset which will be returned (which in the OBIS case, can vary from 0 to over 40,000 records for a single species).
- Value adding. The system returns matches by scientific name, but with no added value such as an associated common name or 'organism type' (taxonomic category), and no synonym resolution since this information is not available from the remote data sources in any consistent manner, if at all.
- Serial versus parallel searching. Searches are undertaken serially, *e.g.* to discover OBIS information on the 42 known species of whales one has to undertake 42 separate searches (at up to 2 minutes per species), and searches on larger groups rapidly become impracticable (*e.g.* the 16,000+ marine fishes, or the subset of the latter beginning with 'A', etc.).
- Service chaining. In order to map (for example) the distribution of a marine species – for example *Balaenoptera physalus*, the fin whale – first, all 43,000 records must be retrieved, and only then can be sent to a mapper, *e.g.* as an XML file (which may or may not be able to cope with such a quantity of data).
- Spatial searching. Search for (for example) all data items within a given region defined as a bounding box can be slow, on account of the large quantities of data to be parsed at remote locations and returned in real time.
- Need for correct spelling. If a user enters an incorrectly spelled name, no data are returned (unless by chance a similar erroneous name exists in a contributing database), and there is no indication to the user of applicable 'near matches' owing to the nature of the query method (which searches for an exact match).

Drawing on those aspects of OBIS version 1 which had already proved successful, including building a community of remote data providers and implementing a common search protocol, planning for a 'new, improved' version 2 of OBIS started in March 2003, which would address the issues identified above with the goal of significantly upgrading usability and performance of the system.

The OBIS Index

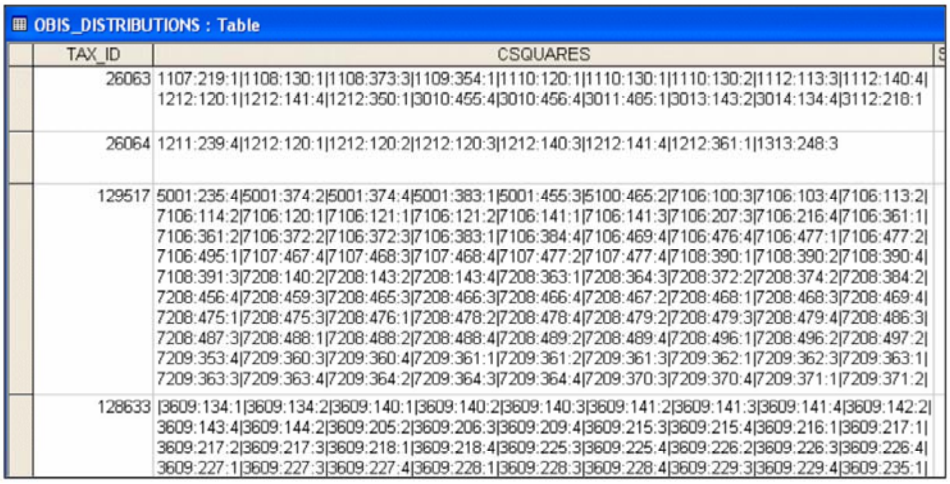
It was realised at the start of the upgrade process that incorporation of a central 'OBIS Index', to reside on the portal, would be a key concept in addressing the majority of the issues identified above, in other words, moving from a fully distributed system to a hybrid approach based on crawling the remote data providers and holding a set of summary-level information or metadata regarding each species on the portal. Such an index would then allow user searches to be split into a two stage operation: 'stage 1' searches would operate on the index and provide metadata level information on available OBIS content very rapidly, while 'stage 2' searches would retrieve actual item-level content once the user had identified exactly what data should be retrieved. In the initial prototype constructed, these ‘stage 2’ searches were fully distributed queries to the remote data providers as described above for OBIS version 1, while by the time the system was ready to deploy, these had been replaced by queries to a locally held data cache (see next section).

The OBIS Index was constructed to be both a name index and a spatial index. The name index holds summary information by species, such as total number of accessible records, contributing data sources, date range (earliest, latest years represented in the data set for any species), a selected common name for initial rapid display, and any synonym resolution as required, the latter drawn from recent work of the ‘Catalogue of Life’ project (Bisby *et al.*, 2004), together with allocation to a custom taxonomic hierarchy to support searching and grouping by taxonomic category as required, as well as presentation of an ‘organism type’ – examples ‘a fish’, ‘a whale’ – alongside every returned species name. An additional feature of the name index is to provide support for a ‘fuzzy name matching’ function via a modified version of the species name stored alongside the original. This information, together with inventory-level data such as which species names originate from which contributing databases, is stored in a small relational database which resides at the portal and provides support for ‘stage 1’ queries as defined above, which basically return lists of relevant species names and associated metadata in response to a user’s query. Every unique scientific name available via the system is also allocated a unique (internal) numeric identifier which links the various tables together.

A supplementary component of the name index is the addition of names of species considered to be marine in the Catalogue of Life, but presently unrepresented by distribution data among OBIS’ current data contributors; this allows a degree of gap analysis (assessing the percentage of known species in any particular group for which OBIS data are available over time), at least to the extent that Catalogue of Life coverage is itself complete, and also allows users to check the spelling of entered marine species names whether or not OBIS has matching species distribution data at the present time.

The spatial index forms a separate table within the ‘OBIS Index’ database and comprises a list of species identifiers, each associated with a set of codes which represent the spatial distribution of the available data within a set of global 0.5 x 0.5 degree squares, labelled using the ‘c-squares’ hierarchical notation (Rees, 2003, 2004), as shown in Fig. 2. In the current implementation, multiple c-squares codes (examples: 1107:219:1, 1108:130:1) are held as a concatenated text string using a separator character (vertical bar or ‘pipe’) between each code, and spatial search is by matching the code for any square entered by the user to any position in the c-squares string, in order for a ‘hit’ to be detected. For example, in the c-squares notation, the ten degree square extending from 10º to 20º N and 70º to 80º E is represented by the code 1107, the five degree square from 10º to 15º N and 75º to 80º E is represented by the code 1107:2, the one degree square from 11º to 12º N and 79º to 80º E is represented by the code 1107:219, and the 0.5 degree square from 11º to 11.5º N and 79º to 79.5º E is represented by the code 1107:219:1, and a search for data in any of this nested set of squares will return a ‘hit’ for the first species indicated (species id = 26063, which in fact corresponds to *Suggrundus macracanthus*, a species of fish). By this means, the spatial index (when cross referenced to the name index) supports queries of the type 'list all the species with data in a selected X degree square', where X belongs to the set ten, five, one or 0.5 degrees (at present, only the ten degree option is offered, in order to avoid too many queries returning no data), or optionally, a number of other spatial queries can be constructed, such as constrain by taxonomic group, etc.

Of further interest is that this example species (selected at random) is associated with 25 unique data records, but only 18 squares (at half degree resolution), in other words a degree of information compression is incorporated into the spatial index when multiple records occur in close proximity, which leads to additional efficiency for data storage and transfer (*e.g.* to relevant mappers). For example, the map shown in Fig. 6 for the minke whale, *Balaenoptera acutorostrata* (2,131 records), is generated from a list of 593 relevant squares, while that for the fin whale, *Balaenoptera physalus* (43,435 records) requires only 1,678 squares at the same resolution, a saving of over 96% in (meta-) data storage, data transmission time and required bandwidth via the web, and mapper processing time to generate the relevant 'quick map'.

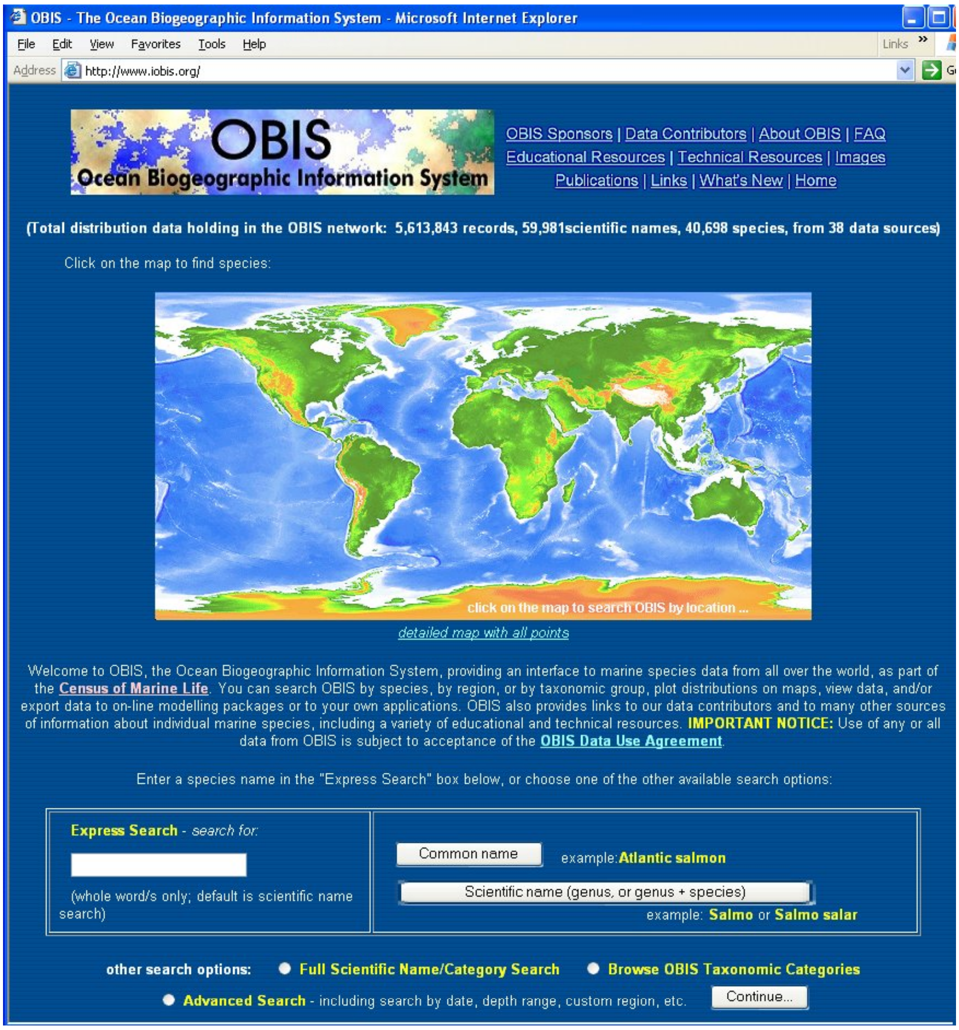


<table>
<tr><th>TAX_ID</th><th>CSQUARES</th><th>S</th></tr>
<tr><td>26063</td><td>1107:219:1|1108:130:1|1108:373:3|1109:354:1|1110:120:1|1110:130:1|1110:130:2|1112:113:3|1112:140:4|
1212:120:1|1212:141:4|1212:350:1|3010:455:4|3010:456:4|3011:485:1|3013:143:2|3014:134:4|3112:218:1</td><td></td></tr>
<tr><td>26064</td><td>1211:239:4|1212:120:1|1212:120:2|1212:120:3|1212:140:3|1212:141:4|1212:361:1|1313:248:3</td><td></td></tr>
<tr><td>129517</td><td>5001:235:4|5001:374:2|5001:374:4|5001:383:1|5001:455:3|5100:465:2|7106:100:3|7106:103:4|7106:113:2|
7106:114:2|7106:120:1|7106:121:1|7106:121:2|7106:141:1|7106:141:3|7106:207:3|7106:216:4|7106:361:1|
7106:361:2|7106:372:2|7106:372:3|7106:383:1|7106:384:4|7106:469:4|7106:476:4|7106:477:1|7106:477:2|
7106:495:1|7107:467:4|7107:468:3|7107:468:4|7107:477:2|7107:477:4|7108:390:1|7108:390:2|7108:390:4|
7108:391:3|7208:140:2|7208:143:2|7208:143:4|7208:363:1|7208:364:3|7208:372:2|7208:374:2|7208:384:2|
7208:456:4|7208:459:3|7208:465:3|7208:466:3|7208:466:4|7208:467:2|7208:468:1|7208:468:3|7208:469:4|
7208:475:1|7208:475:3|7208:476:1|7208:478:2|7208:478:4|7208:479:2|7208:479:3|7208:479:4|7208:486:3|
7208:487:3|7208:488:1|7208:488:2|7208:488:4|7208:489:2|7208:489:4|7208:496:1|7208:496:2|7208:497:2|
7209:353:4|7209:360:3|7209:360:4|7209:361:1|7209:361:2|7209:361:3|7209:362:1|7209:362:3|7209:363:1|
7209:363:3|7209:363:4|7209:364:2|7209:364:3|7209:364:4|7209:370:3|7209:370:4|7209:371:1|7209:371:2|</td><td></td></tr>
<tr><td>128633</td><td>|3609:134:1|3609:134:2|3609:140:1|3609:140:2|3609:140:3|3609:141:2|3609:141:3|3609:141:4|3609:142:2|
3609:143:4|3609:144:2|3609:205:2|3609:206:3|3609:209:4|3609:215:3|3609:215:4|3609:216:1|3609:217:1|
3609:217:2|3609:217:3|3609:218:1|3609:218:4|3609:225:3|3609:225:4|3609:226:2|3609:226:3|3609:226:4|
3609:227:1|3609:227:3|3609:227:4|3609:228:1|3609:228:3|3609:228:4|3609:229:3|3609:229:4|3609:235:1|</td><td></td></tr>
</table>

Fig. 2. Fragment of the spatial index for OBIS version 2.

As mentioned above, the spatial index also supports the production of 'quick maps' (representation of species distributions by global 0.5 degree squares) directly from the index, in other words without requiring a (potentially slower) request to retrieve the atomic level species data for this purpose. This is achieved by rapidly assembling relevant strings of codes from the spatial index into the HTML page of search results in advance, so that the user is presented with a set of pre-configured links which, when pressed, will submit the relevant list of squares to a web based utility at CSIRO Marine Research, the c-squares mapper (see www.marine.csiro.au/csquares/about-mapper.htm) which processes the list, plots the relevant squares on to one of a range of user-selectable base maps, and returns the result as a gif image to the user's web browser, as per the example in Fig. 6.

Figs. 3-6 illustrate aspects of current OBIS ‘stage 1’ searches, including two initial search pages, an example search result for information on ‘all whales’, and an example ‘quick map’ for a user’s selected species, all drawn from the holdings of the Index, that is, without requiring any connection to the item-level data at this point.


OBIS - The Ocean Biogeographic Information System - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
Address http://www.iobis.org/
OBIS
Ocean Biogeographic Information System
OBIS Sponsors | Data Contributors | About OBIS | FAQ
Educational Resources | Technical Resources | Images
Publications | Links | What's New | Home
click on the map to search OBIS by location ...
detailed map with all points
Common name
Scientific name (genus, or genus + species)
Continue...

(Total distribution data holding in the OBIS network: 5,613,843 records, 59,981scientific names, 40,698 species, from 38 data sources)

Click on the map to find species:

Welcome to OBIS, the Ocean Biogeographic Information System, providing an interface to marine species data from all over the world, as part of the **Census of Marine Life**. You can search OBIS by species, by region, or by taxonomic group, plot distributions on maps, view data, and/or export data to on-line modelling packages or to your own applications. OBIS also provides links to our data contributors and to many other sources of information about individual marine species, including a variety of educational and technical resources. **IMPORTANT NOTICE:** Use of any or all data from OBIS is subject to acceptance of the **OBIS Data Use Agreement**.

Enter a species name in the "Express Search" box below, or choose one of the other available search options:

Express Search - *search for:*

(whole word/s only; default is scientific name search)

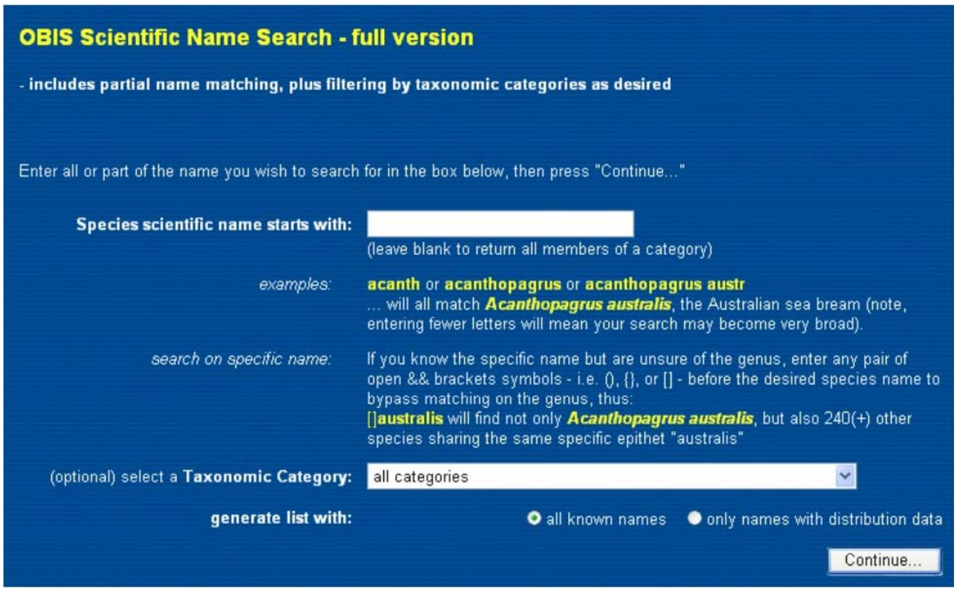
example: **Atlantic salmon**

example: **Salmo** or **Salmo salar**

other search options: ○ **Full Scientific Name/Category Search** ○ **Browse OBIS Taxonomic Categories**

○ **Advanced Search** - including search by date, depth range, custom region, etc.

Fig. 3. OBIS version 2 start page, as at November 2004 – including ‘click-on-a-map’ spatial search, and express search input text boxes for scientific and common name searches.



OBIS Scientific Name Search - full version

- includes partial name matching, plus filtering by taxonomic categories as desired

Enter all or part of the name you wish to search for in the box below, then press "Continue..."

Species scientific name starts with: (leave blank to return all members of a category)

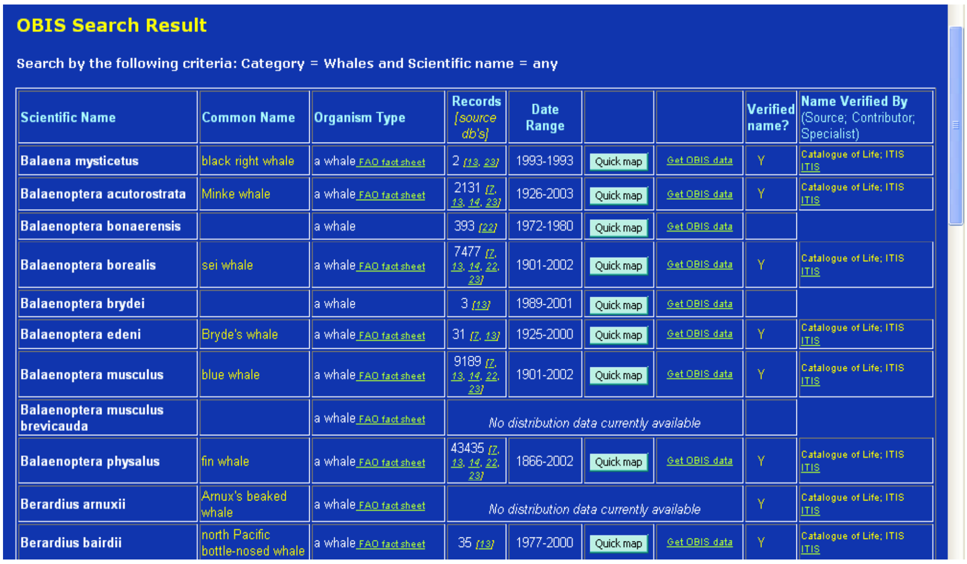
examples: **acanth** or **acanthopagrus** or **acanthopagrus austr**
... will all match ***Acanthopagrus australis***, the Australian sea bream (note, entering fewer letters will mean your search may become very broad).

search on specific name: If you know the specific name but are unsure of the genus, enter any pair of open && brackets symbols - i.e. (), {}, or [] - before the desired species name to bypass matching on the genus, thus:
[]**australis** will find not only ***Acanthopagrus australis***, but also 240(+) other species sharing the same specific epithet "australis"

(optional) select a **Taxonomic Category:** all categories

generate list with: ● all known names ○ only names with distribution data

Fig. 4. OBIS version 2 full scientific name search page, as at November 2004 – including 'partial name matching, and filter by taxonomic category.

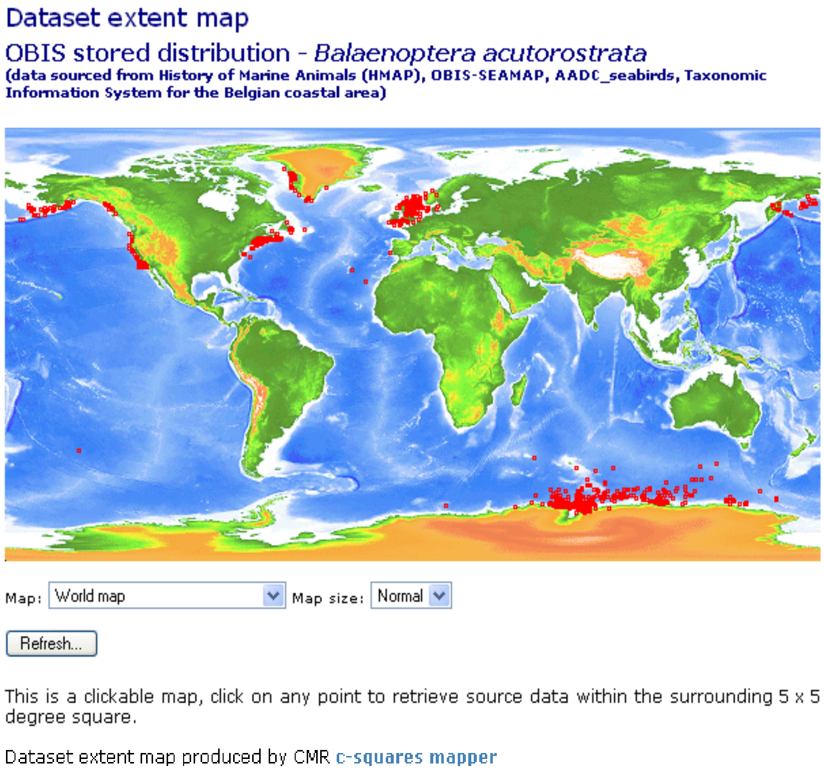


OBIS Search Result

Search by the following criteria: Category = Whales and Scientific name = any

Scientific Name	Common Name	Organism Type	Records *[source db's]*	Date Range			Verified name?	Name Verified By (Source; Contributor; Specialist)
Balaena mysticetus	black right whale	a whale FAO fact sheet	2 [13, 23]	1993-1993	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera acutorostrata	Minke whale	a whale FAO fact sheet	2131 [7, 13, 14, 23]	1926-2003	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera bonaerensis		a whale	393 [22]	1972-1980	Quick map	Get OBIS data		
Balaenoptera borealis	sei whale	a whale FAO fact sheet	7477 [7, 13, 14, 22, 23]	1901-2002	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera brydei		a whale	3 [13]	1989-2001	Quick map	Get OBIS data		
Balaenoptera edeni	Bryde's whale	a whale FAO fact sheet	31 [7, 13]	1925-2000	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera musculus	blue whale	a whale FAO fact sheet	9189 [7, 13, 14, 22, 23]	1901-2002	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera musculus brevicauda		a whale FAO fact sheet	*No distribution data currently available*					
Balaenoptera physalus	fin whale	a whale FAO fact sheet	43435 [7, 13, 14, 22, 23]	1866-2002	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Berardius arnuxii	Arnux's beaked whale	a whale FAO fact sheet	*No distribution data currently available*				Y	Catalogue of Life; ITIS ITIS
Berardius bairdii	north Pacific bottle-nosed whale	a whale FAO fact sheet	35 [13]	1977-2000	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS

Fig. 5. Example 'stage 1' search result, OBIS version 2, comprising a list of matching species names with associated metadata, plus links to 'quick maps' and 'get OBIS data' (=stage 2) searches.



Dataset extent map

OBIS stored distribution - *Balaenoptera acutorostrata*

(data sourced from History of Marine Animals (HMAP), OBIS-SEAMAP, AADC_seabirds, Taxonomic Information System for the Belgian coastal area)

Map: World map Map size: Normal

This is a clickable map, click on any point to retrieve source data within the surrounding 5 x 5 degree square.

Dataset extent map produced by CMR c-squares mapper

Fig. 6. Example 'quick map' generated from the spatial index holdings for a species of whale (Balaenoptera acutorostrata, 2,131 records) in a few seconds using the c-squares mapper located at CSIRO Marine Research.

'Stage 2' searches and the OBIS data cache

As described in the previous section, with the initial redesign of OBIS incorporating the new indexing functions, the requirement for 'stage 2' (= 'get data') queries is deferred in many instances until the user has familiarised his or herself with relevant system content using a 'stage 1' or index level search, leading to much faster and more satisfactory performance in the initial stages, and a reduction in the load on the system since many initial queries can be answered from the index alone (such as whether or not data exist for a species of interest, what species occur in a given area, and even the production and browsing of 'quick maps'). Nevertheless, it is essential to provide access for 'stage 2' searches from this point onwards, and in the hybrid 'index plus distributed search' architecture such queries are themselves still subject to a number of the disadvantages of a fully distributed system as described above, even with the introduction of upgraded 'wrapper' technology, introduced when OBIS moved to the DiGIR data retrieval protocol (Blum *et al.*, 2001) in place of the initial custom database wrappers, concurrently with the present upgrade.

These residual negative aspects have been addressed by the introduction of a data cache on the Portal, holding a subset of the full record (as a copy) for every OBIS data item accessible via the remote data providers, updated on a rolling basis. The purpose of this cache is to insulate the user from any individual provider being off line or unresponsive at time of querying, and also to provide a faster and more uniform response to user queries. (As a by-product, it also facilitates creation of the Index, which otherwise would require numerous and possibly slow queries to the remote providers on a species-by-species basis). Together with the Index, this cache is shown in the revised architecture as implemented for OBIS version 2, below (Fig. 7).

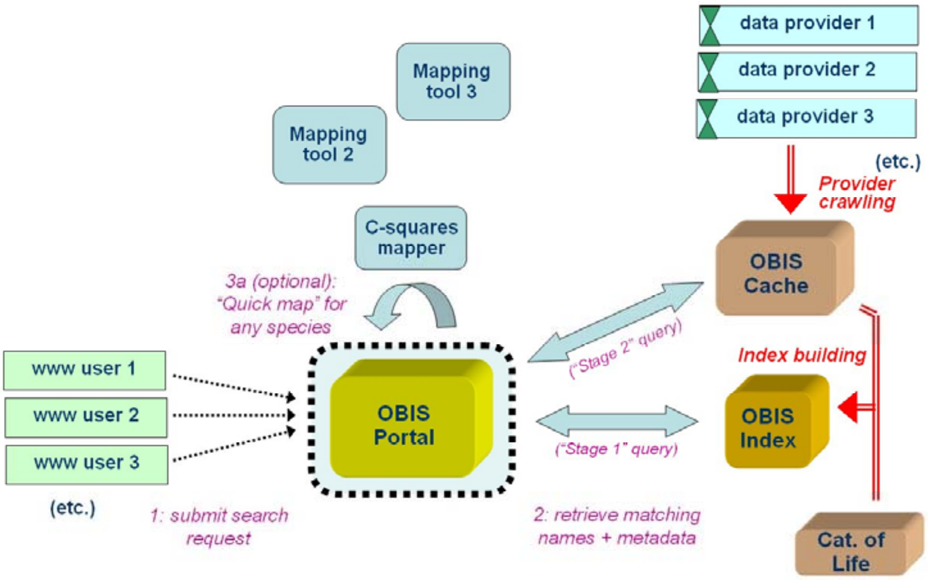


Fig. 7. Schematic of overall architecture for OBIS version 2.

The practical implementation of this architecture requires initial provider crawling to populate the cache, then creation of the index (name and spatial components) by parsing the cache content and also incorporating relevant information from the Catalogue of Life. As described above, 'stage 1' queries can then be issued against the index (relatively small in size, *e.g.* 100,000 rows at this time for the main 'obis_species' table) and used for the generation of 'get OBIS data' links and 'quick maps', while 'stage 2' queries operate against the cache (5 million + rows) but are still substantially faster (and potentially more complete) than querying the remote data sources in real time. The main disadvantages of this new architecture are its increased complexity and requirement for resources of data storage and maintenance, and the necessity to keep both cache and index content up-to-date by continual re-crawling of the remote data providers as new data are added to the system, or individual records are altered or deleted at the provider end.

Conclusion

The new version of OBIS released in 2004 has achieved a quantum leap in usability, and addressed all of the weaknesses described above for a fully distributed system, while at the same time incorporating additional innovative approaches to spatial indexing, searching and mapping, search by a custom taxonomic hierarchy, 'fuzzy' matching of species scientific names, and more. With all such systems, a degree of continuous improvement and evolution to reflect changing user demands or system possibilities will be inevitable over time, however it is felt that the current 'OBIS version 2' offers a satisfactory balance of user-weighted features as against the increased complexity and requirement for technical resources from the portal and system design points of view. Further development of OBIS will incorporate our experiences with the current system over the next 12-18 month time frame as well as the potential to exchange experiences with the developers of GBIF, the Global Biological Information Facility (www.gbif.org) and others working in similar areas of distributed biological information retrieval.

Acknowledgements

The authors are grateful for the contributions of members of the OBIS technical subcommittee (Rainer Froese, Daphne Fautin), other members of the 'OBIS web development team' (Karen Stocks, James Wood), and the OBIS International Committee (in particular Fred Grassle and Mark Costello) in reviewing progress and suggesting particular aspects for implementation. The help of Pamela Brodie and Miroslaw Ryba (CSIRO) and Lissa Jerry and Wei Zhuang (Rutgers) in assisting with practical aspects of OBIS development and deployment is also gratefully acknowledged.

References

Bisby F.A., R. Froese, M.A. Ruggiero and K.L. Wilson. 2004. Species 2000 and ITIS Catalogue of Life, Annual Checklist 2004: Indexing the world's known species. CD-ROM. Species 2000: Los Baños, Philippines.

Blum, S., D. Vieglais and P.J. Schwartz. 2001. DiGIR – Distributed Generic Information Retrieval. Powerpoint presentation, available at http://digir.sourceforge.net/events/20011106/DiGIR.ppt.

Grassle, J.F. and K.I. Stocks. 1999. A Global Ocean Biogeographic Information System (OBIS) for the Census of Marine Life. Oceanography 12(3):12-14.

Rees, T. 2003. "C-squares", a new spatial indexing system and its applicability to the description of oceanographic datasets. Oceanography 16(1):11-19.

Rees, T. 2004. Use of c-squares spatial indexing and mapping in the 2004 release of OBIS, the Ocean Biogeographic Information System. Abstract and Presentation, EOGEO 2004, London, UK. Available via the EOGEO website at http://www.eogeo.org/Workshops/EOGEO2004/.

Zhang, Y. and J.F. Grassle. 2003. A portal for the Ocean Biogeographic Information System. Oceanologica Acta 25:193-197.