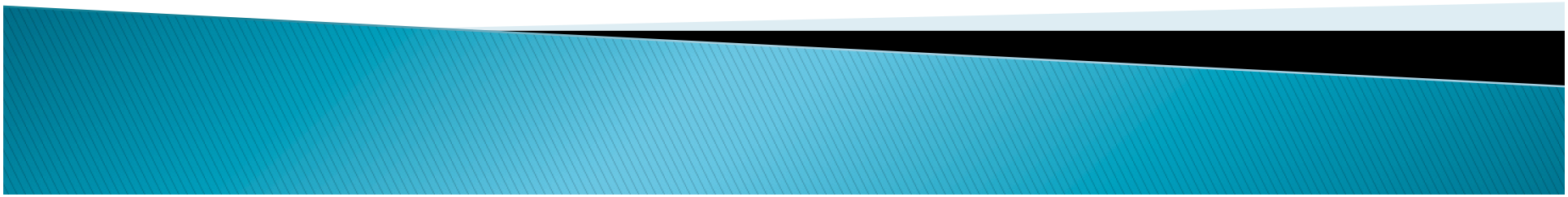
Selected innovations in Biodiversity Informatics

Presentation for the 2014 Ebbe Nielsen Prize

Tony Rees, CSIRO Australia



Submissions for the Prize

c-squares

- Rapid indexing and locator system for geographic data

OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

- Move from distributed model (v1) to "hub and spoke" model (v2, 2004), with taxonomic and spatial indexing

Taxa match

- Fuzzy matching for taxon scientific names

Interim Register of Marine and Nonmarine Genera

- A working list of "all" genera, extant + fossil, with extant/fossil and marine/nonmarine flags (plus species too)

c-squares

Rapid indexing and locator system for geographic data

Why are geographic data "special"?

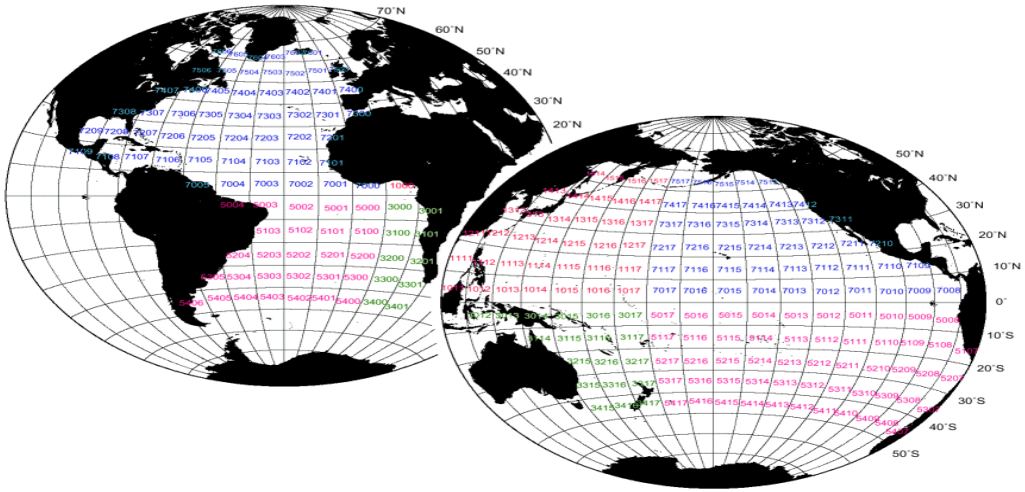
- Most familiar data items (date, time, age, name, size…) are 1-dimensional – can be indexed in a linear sequence (e.g. alphabetical term index at the back of a book)
- Book index supports 2 query types:
  - Show me all the pages containing term "X"
  - (Inverted): Show me all (significant) terms on page "Y"
- NB index construction also involves "binning" – divide e.g. continuous stream of text into numbered pages for information access.
- Geographic information is natively 2-dimensional (latitude x longitude), does not fit a 1-D index structure (unless reduced to e.g. country codes, etc.)
- Latitude and longitude are continuously variable – require binning into discrete units for rapid query.

Geographic data indexing using a grid structure

- Provides convenient “bins” for both latitude and longitude
- Familiar from town maps, etc. Index then supports:
  - Show me grid square(s) with item “X”
  - List all items within grid square “Y”

Geographic data indexing using a grid structure

- Local maps suit local scales – but run out of identifiers at map edges (also no between-map interoperability)
- Global grids tend to be large scale – e.g. WMO 10x10 degree squares, others
- C-squares (**C**oncise **S**patial **Qu**ery **a**nd **Re**presentation **S**ystem) provides a hierarchical notation for increasingly small subdivisions of a 10x10 degree global grid used as a starting point (WMO squares).

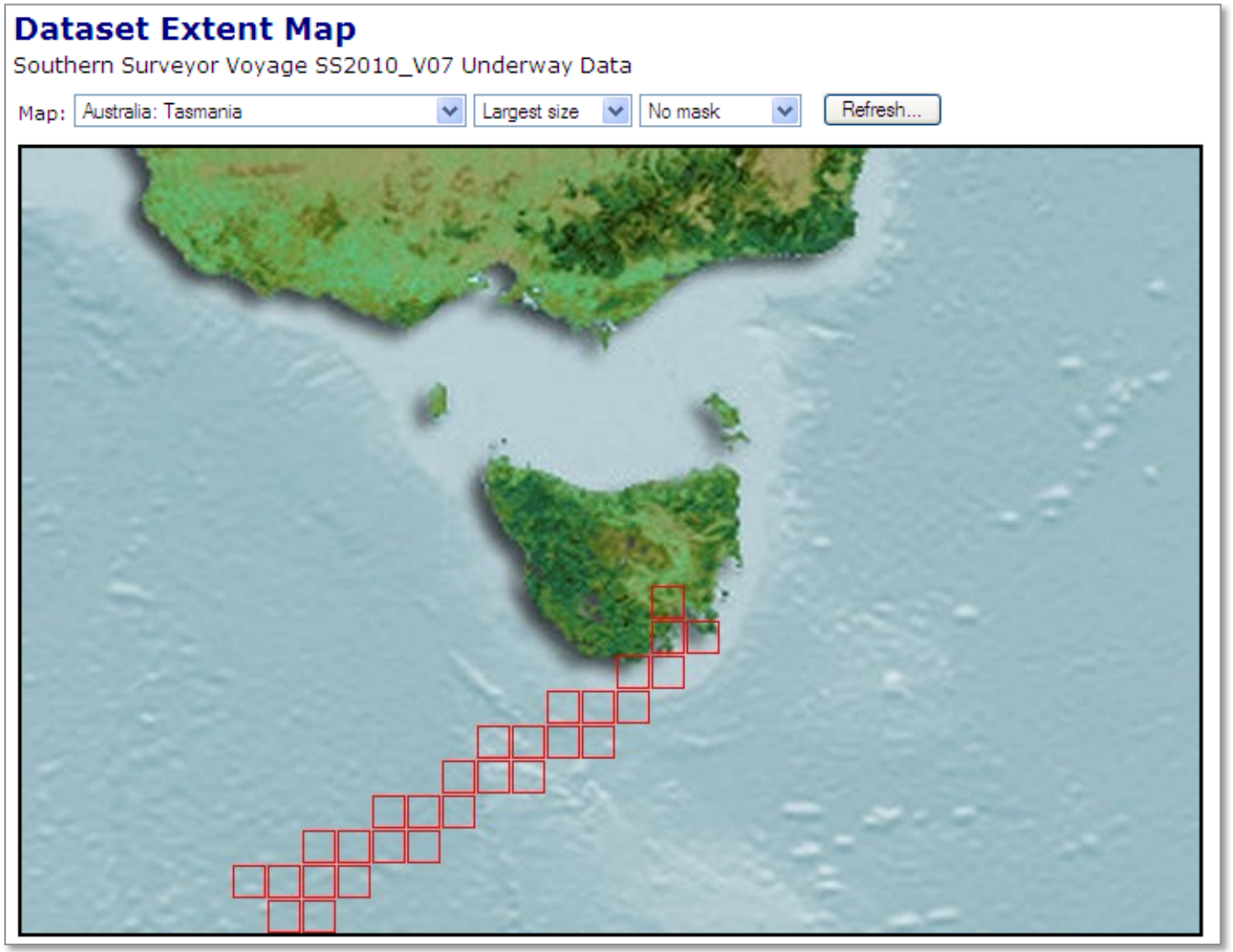


Example c-squares IDs

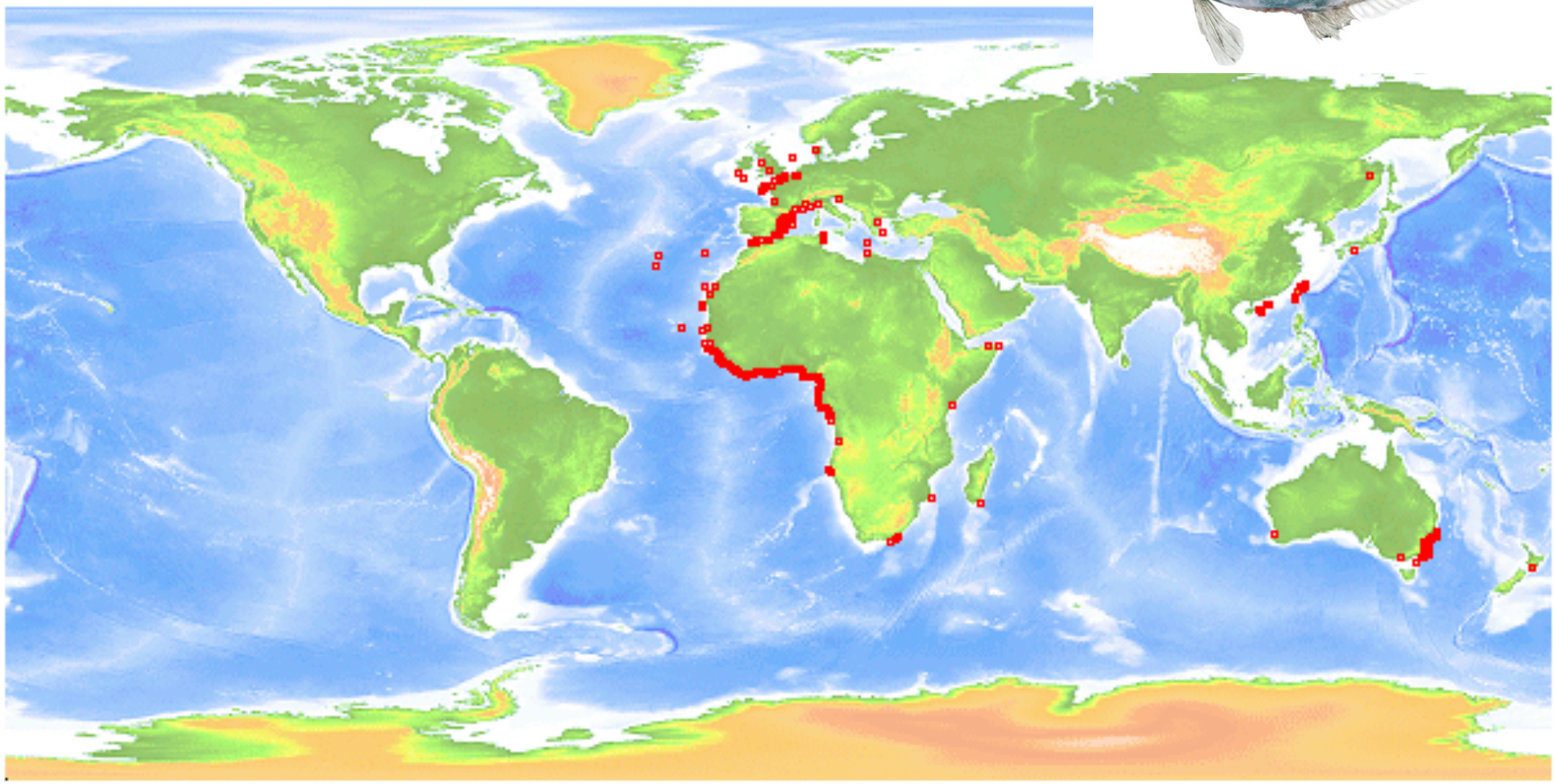
C-squares operation

- Any georeferenced data point (lat-lon) can be assigned its c-square ID at a chosen scale (example: 0.5 x 0.5 deg. square for global datasets)
- Index of data items x square IDs can be searched at that or any larger scale (e.g. 0.5, 1, 5, 10-degree squares) to retrieve relevant data
- List of squares occupied by a set of points = "dataset footprint" – easy to store, search, analyse, or send to a mapping application that understands c-squares notation
- Dataset footprint (c-squares string) need only encode each square once (discard repeat occurrences) – can give substantial data reduction for search/display purposes without losing core information
- Square IDs can be stored as simple presence/absence, or with associated density information (such as no. of records in that square, more…)

C-squares examples (1)

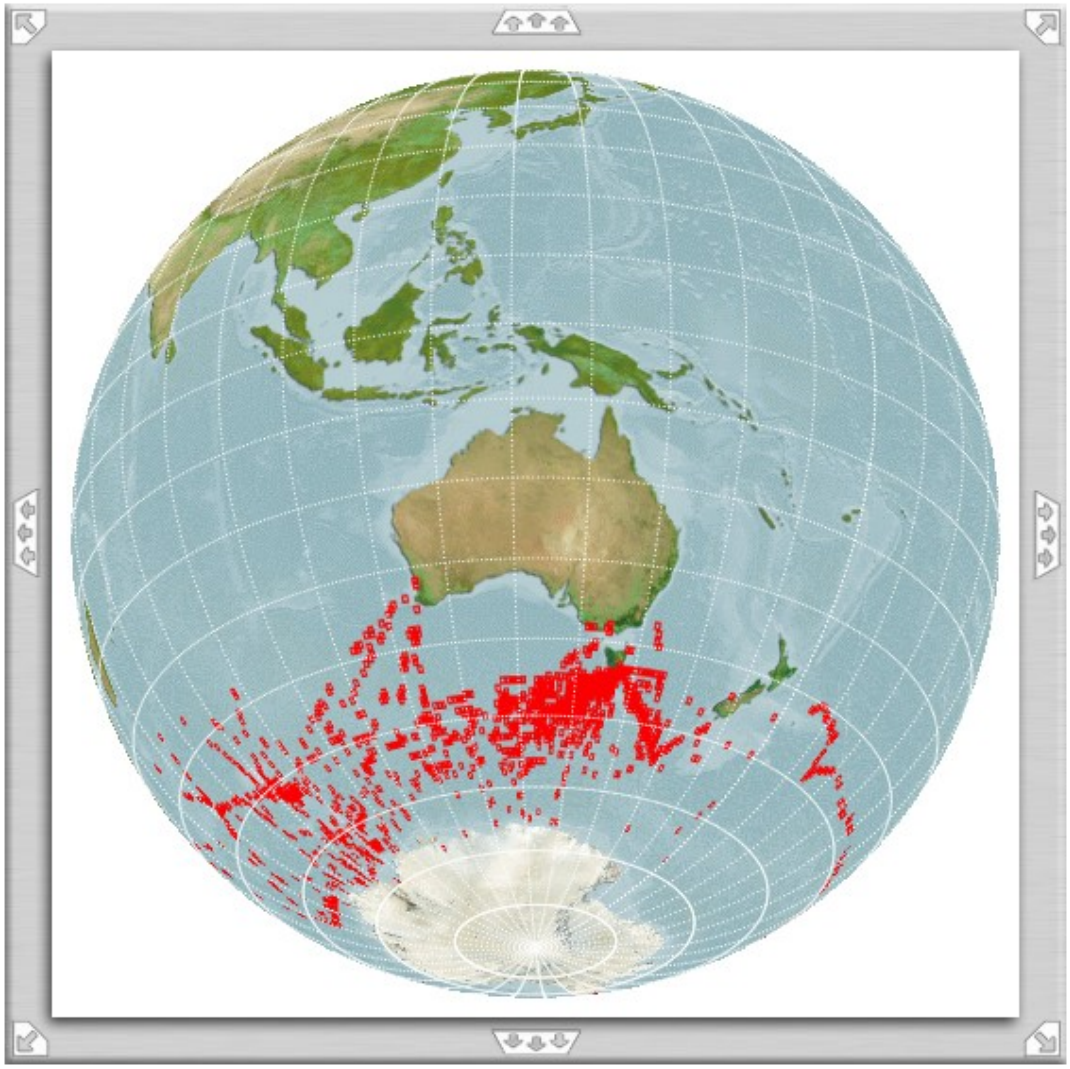


C-squares examples (2)



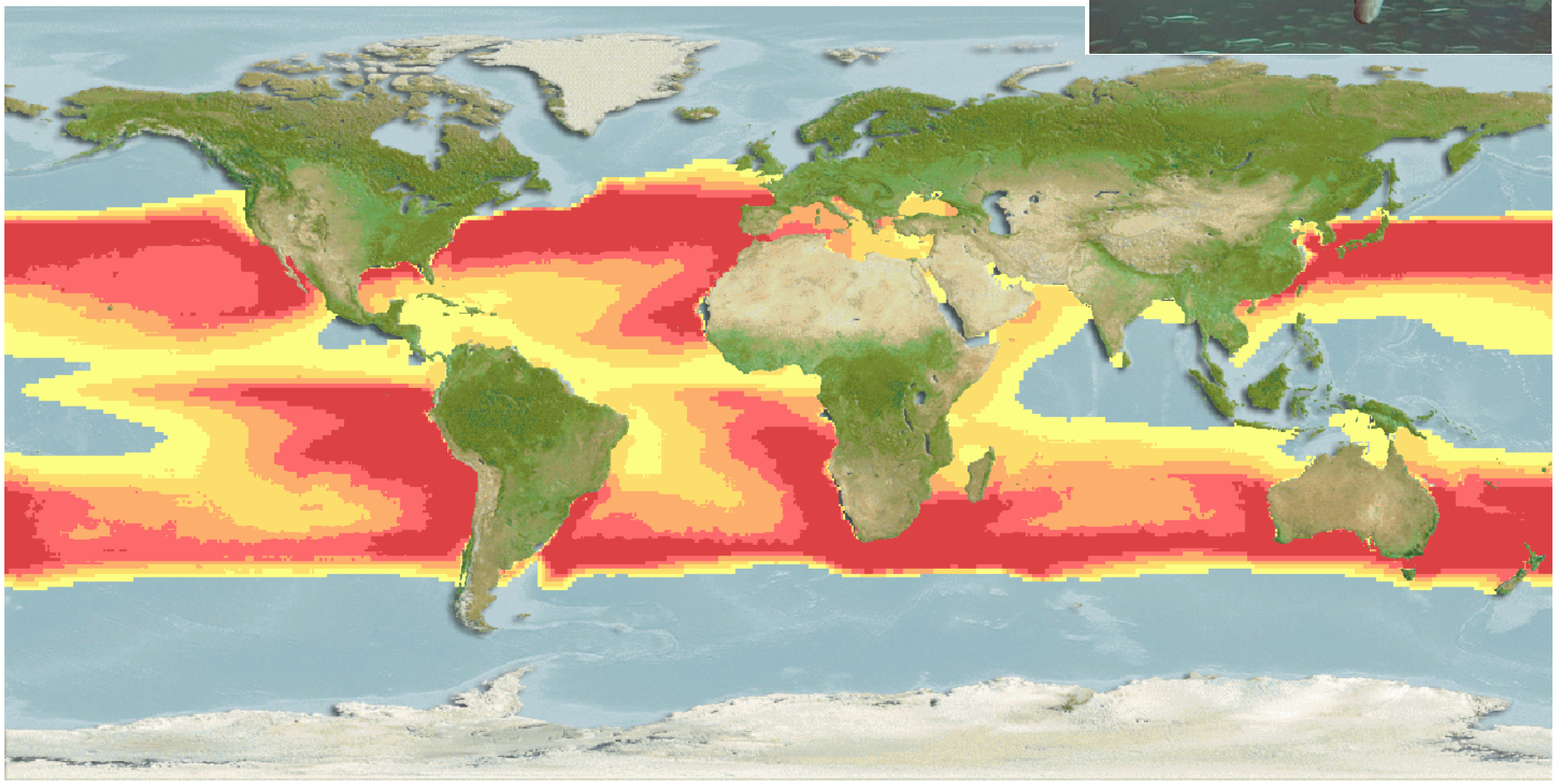
Example OBIS data (2005): species = *Zeus faber* (John Dory)

C-squares examples (3)



Rotatable/tiltable/zoomable globe view produced by the CMAR c-squares mapper (pre-dated Google Earth, still in operation…)

C-squares examples (4)



Probability-encoded squares: AquaMaps Modelled distribution for *Mola mola* (ocean sunfish) based on 0.5-degree grid + associated environmental parameters

C-squares published description

"C-Squares", a New Spatial Indexing System and its Applicability to the Description of Oceanographic Datasets

Tony Rees
CSIRO Marine Research · Hobart, Tasmania Australia

Abstract

A new method is described for representing, querying, displaying and exchanging dataset spatial extents at the metadata level. This method, entitled "c-squares" for *Concise Spatial Query And Representation System*, represents dataset spatial extents (footprints) as regular or irregular shapes built up from smaller component units (grid squares) coded according to a globally applicable system, and permits more reliable spatial queries than are presently enabled by the commonly used "bounding rectangles" method in metadata systems. The method is equally applicable to marine and terrestrial datasets but is particularly useful for marine data (which frequently have irregular, voyage-specific footprints), plus other dataset types not well represented by bounding rectangles.

Introduction

Scientists have, for many years, had access to abstracting services such as "Oceanographic Abstracts", "Biological Abstracts" and others, which provide a resource discovery and description facility for articles in the scientific literature. Over the last decade, similar techniques have begun to be applied to the scientific datasets held by many research agencies, in the form of metadata records (metadata = "data about data") which are then assembled into metadata catalogs or data directories. Thus, by searching an internet-enabled metadata catalog such as NASA's "Global Change Master Directory" (gcmd.gsfc.nasa.

oceanographic data, the inquirer may wish to search for data in regions for which there is no finer scale locality name than, for example, "south-east Pacific Ocean". The second is the *bounding polygon* (supported by FGDC and some other national standards), which indicates a regular or irregular border deemed to enclose the dataset. Such polygons can potentially be uploaded to GIS (Geographic Information System) software where searching for spatial overlaps can be performed, but are generally not searchable by the simpler, numeric- or text-based query operations supported by most metadata catalogs.

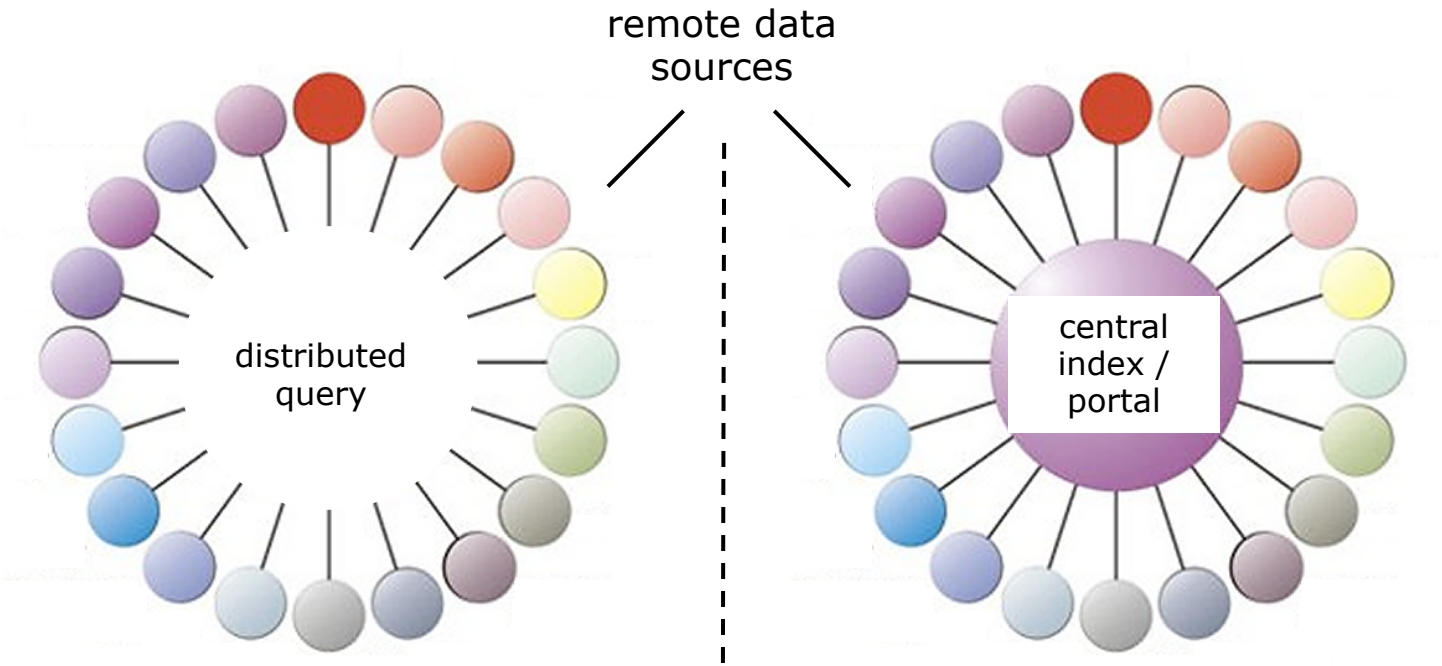
Oceanography vol. 16 (2003): 11-19

C-squares for biodiversity informatics

Spatial indexing as an enhancement for data aggregators

- Lots of point data in repositories (e.g. species x lat-lon) – can be available to remote data systems (OBIS, GBIF, etc.)
- Index (1) all the names, plus (2) the squares in which each species occurs – significant data reduction e.g. 10x – 1000x
- Store the index separately from the base data, then just query this for spatial queries and mapping
  - A lot faster than querying the raw, distributed data
  - Insulates users from potential remote provider down time
  - Pre-indexing provides valuable knowledge of system holdings which can be presented to users as picklists, etc.
- This is the model developed at CSIRO in 2002 for agency systems, offered to OBIS in 2003, implemented in 2004 as OBIS v2 (2004-2010)
- OBIS v3 has moved to full web GIS for mapping, retains c-squares as spatial index for rapid data retrieval/summaries.

OBIS architecture evolution (2002 -> 2004)



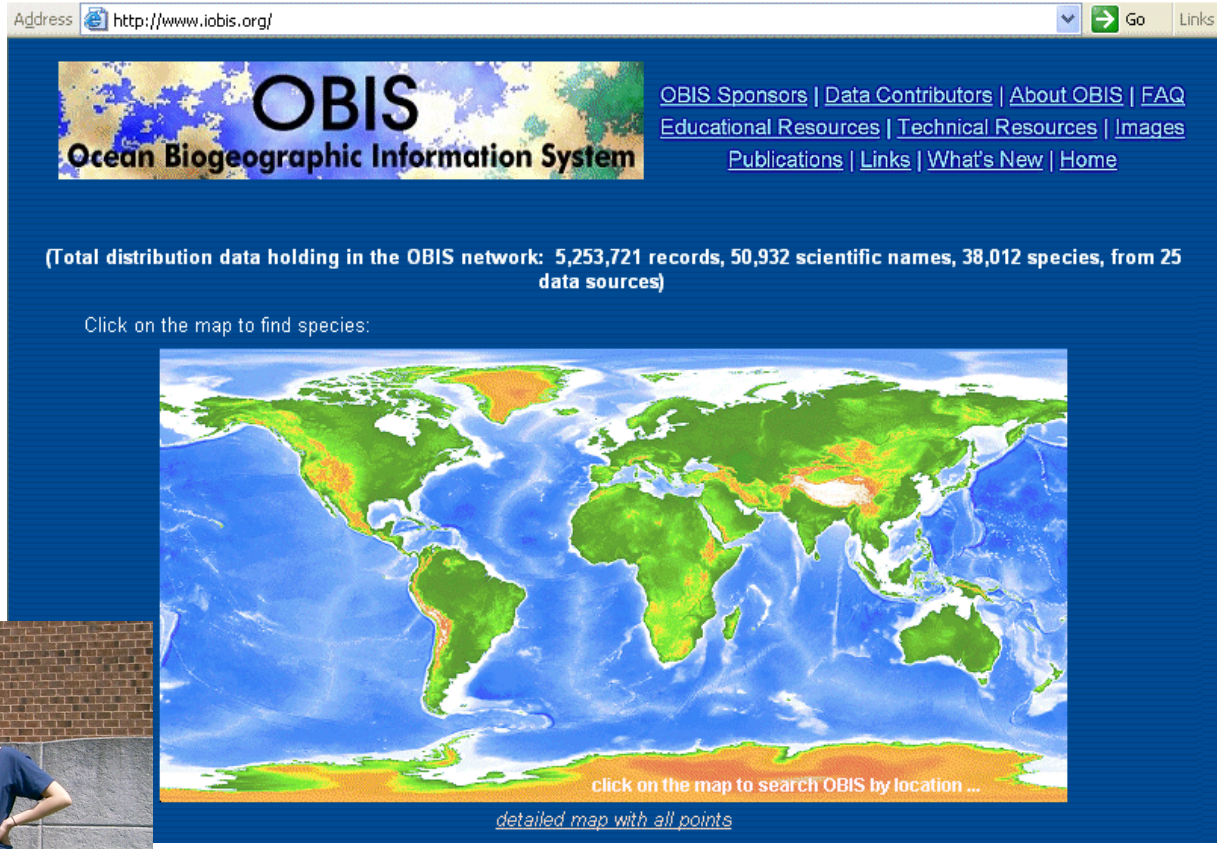
OBIS v1 (2002)

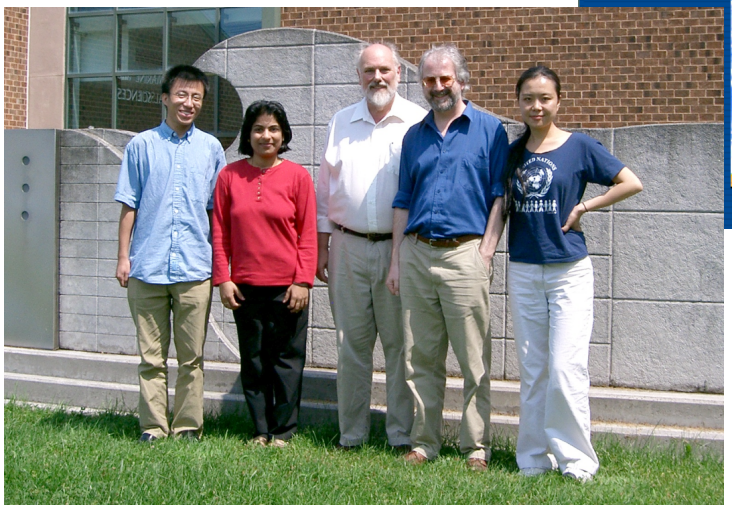
- Queries slow
- Data sources may be offline at query time
- No prior knowledge of system holdings

OBIS v2 (2004)

- Queries fast, + "quick maps" available from the index
- Offline data sources not a problem
- Taxonomic index provides key entry point for users, also shows system current coverage (& gaps)

OBIS v2 front page – highlighting spatial search (name search / browse also available)





OBIS v2 entry point – 2004
see e.g. via the Internet Archive Wayback Machine *http://archive.org/web/* (426 Billion web pages archived)

OBIS Developer team – Rutgers/CSIRO 2004

OBIS v2 published description

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.

In: Proceedings of Ocean Biodiversity Informatics, an international conference on marine biodiversity data management, Hamburg, 2004. pp. 167-176.

Taxa match

Fuzzy matching for taxonomic names

Fuzzy matching for taxonomic names

- Taxon scientific names are often not simple to spell, errors creep in…
  - ***Coelorinchus <> Caelorinchus <> Coelorynchus <> Coelorhynchis***
  - ***antactica <> antarctica , flaveolata <> faveolata…***
- Either input name or target name can be misspelled (or both)
- Lists formed by aggregation can contain same name in multiple variants – need detecting and reconciling (problem for GBIF, OBIS, etc.)
- Errors can be phonetic (soundalike) or non-phonetic (keying errors, OCR errors, more)
- Existing fuzzy match tools not ideal for taxonomic names:
  - Return too many false hits (poor precision)
  - Miss some true hits (poor recall)
  - Phonetic algorithms are fast but miss non-phonetic errors (<50% of total)
  - Non-phonetic algorithms e.g. edit distance catch more errors but are SLOW – especially against millions of target names, also still not optimal for precision + recall

The Taxamatch solution (T. Rees, 2007 onwards)

- **Recall (find all true hits):** look for all possible matches, i.e.:
  1. Test for both phonetic + non-phonetic matches (combination of algorithms)
  2. Close + distant (allow up to 3 character errors in genera, 4 in binomials)
  3. Look for omitted, inserted, substituted and transposed characters/syllables as well as phonetic errors
  4. Look for gender mismatches in species epithets
  5. Handle errors in genus alone, species epithet alone, or both

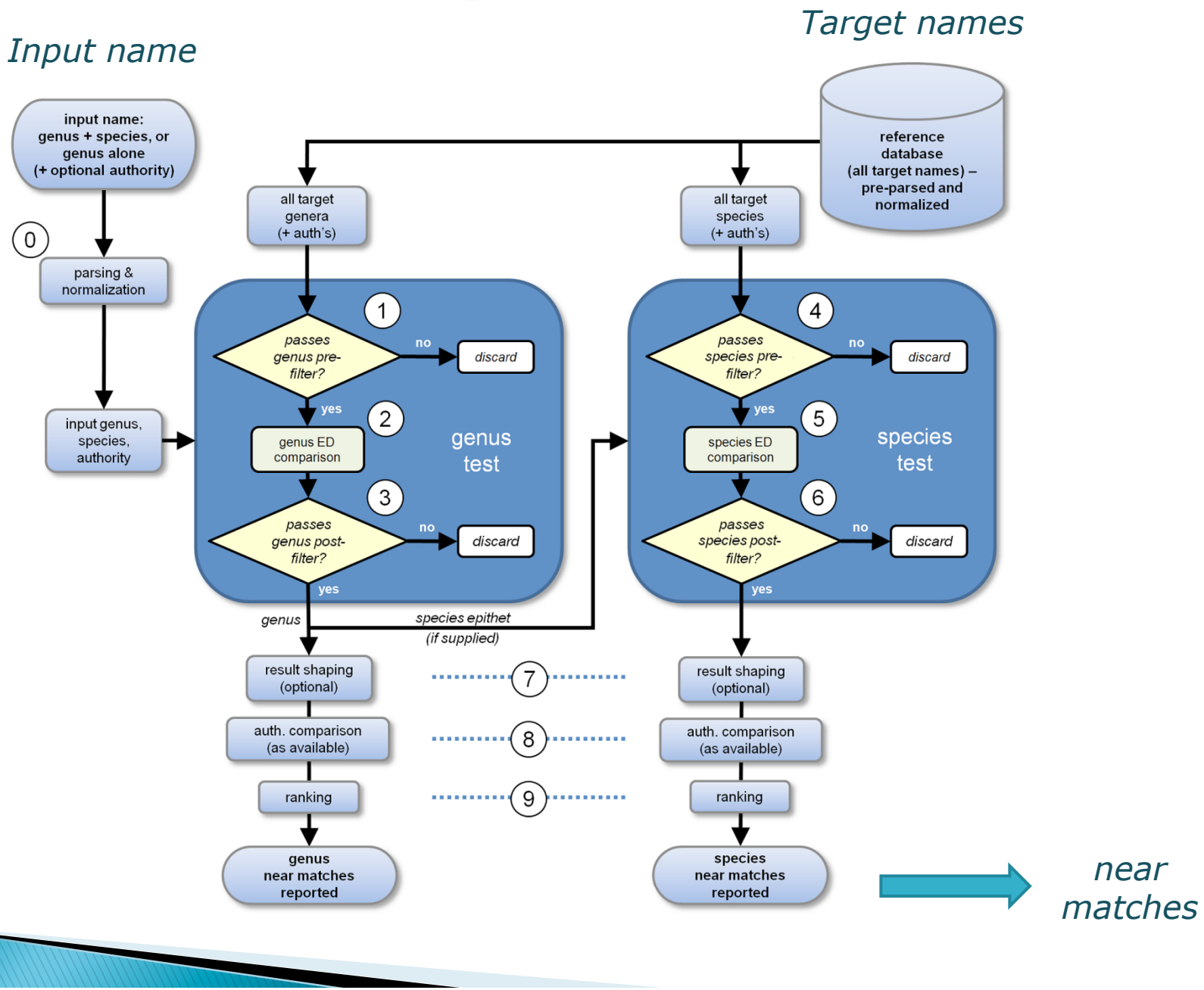
The Taxamatch solution (T. Rees, 2007 onwards)

- **Precision (rejection of false hits):**
  1. Use pattern matching (rule-based approach) to discard candidate matches which do not match patterns seen in real-world misspellings
  2. Have sliding scale for maximum edit distance according to word length (tighter threshold for short words)
  3. Dynamic thresholding ("result shaping"): only return more distant matches in absence of close+phonetic ones (can switch off in special circumstances)
  4. Test only genus names against genera (not higher ranks, some of which may have similar names)
  5. Switchable taxonomic filtering – restrict matches to only a desired taxonomic group – not specific to Taxamatch, but a useful feature to provide in user interface (and is in author's reference implementation)

The Taxamatch solution (T. Rees, 2007 onwards)

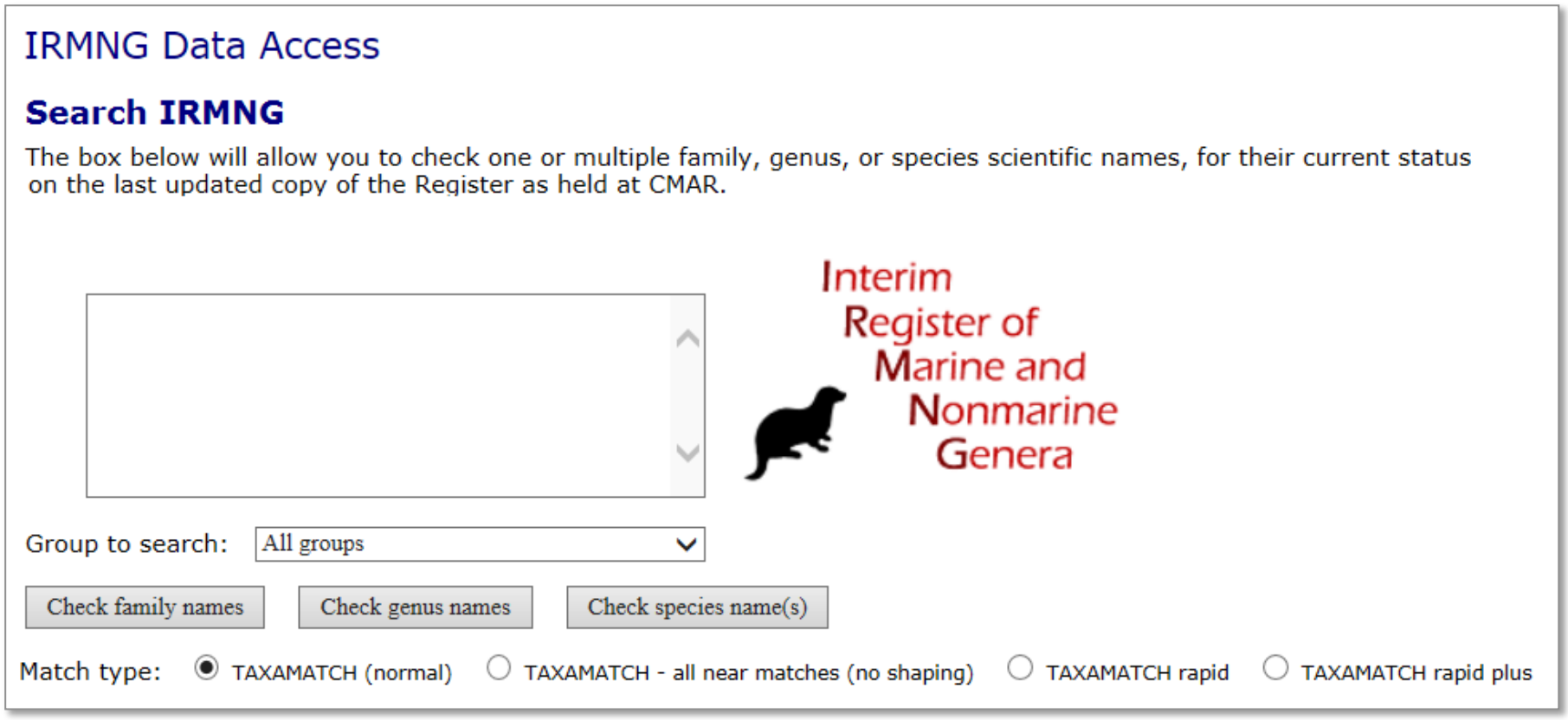
- **Efficiency (rapid execution time):**
  1. 2-stage approach for species: first test genus portion only, then just species of near-matching genera
     - a lot less genera than species to match against, plus many species can be eliminated without testing
  2. Many names need not be tested – if they do not match predicted error patterns, or are too long/short cf. the input name
     - can avoid testing ~99% of all names, reducing overall test time to 1-2 secs or better against reference DB of 2m names
  3. For even faster execution time (0.1 sec/input name or better): presume that *either* genus or species epithet is at least a phonetic match (= "Taxamatch rapid")
     - avoids testing more names, but can fail in rare cases; also does not suit genus names tested alone.

Taxamatch block diagram



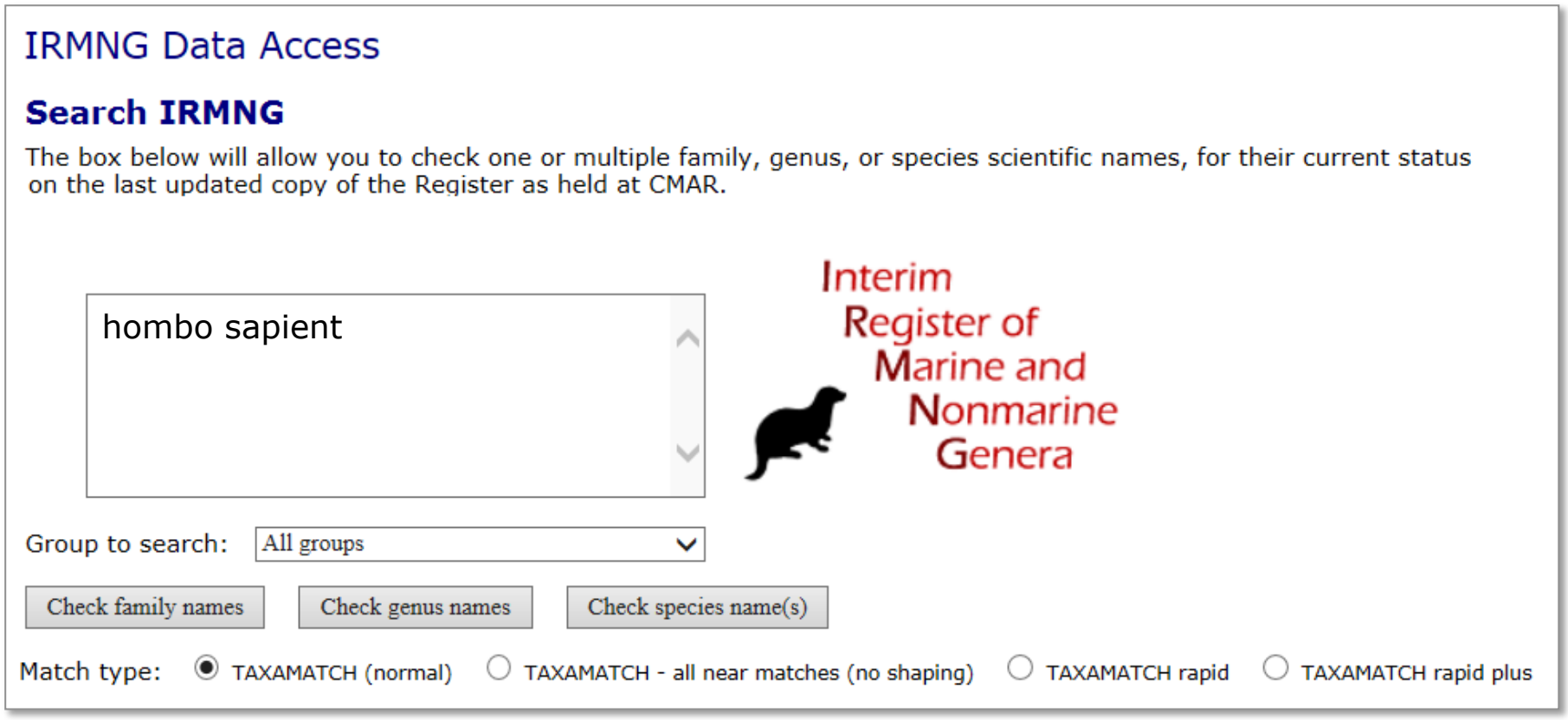
Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)



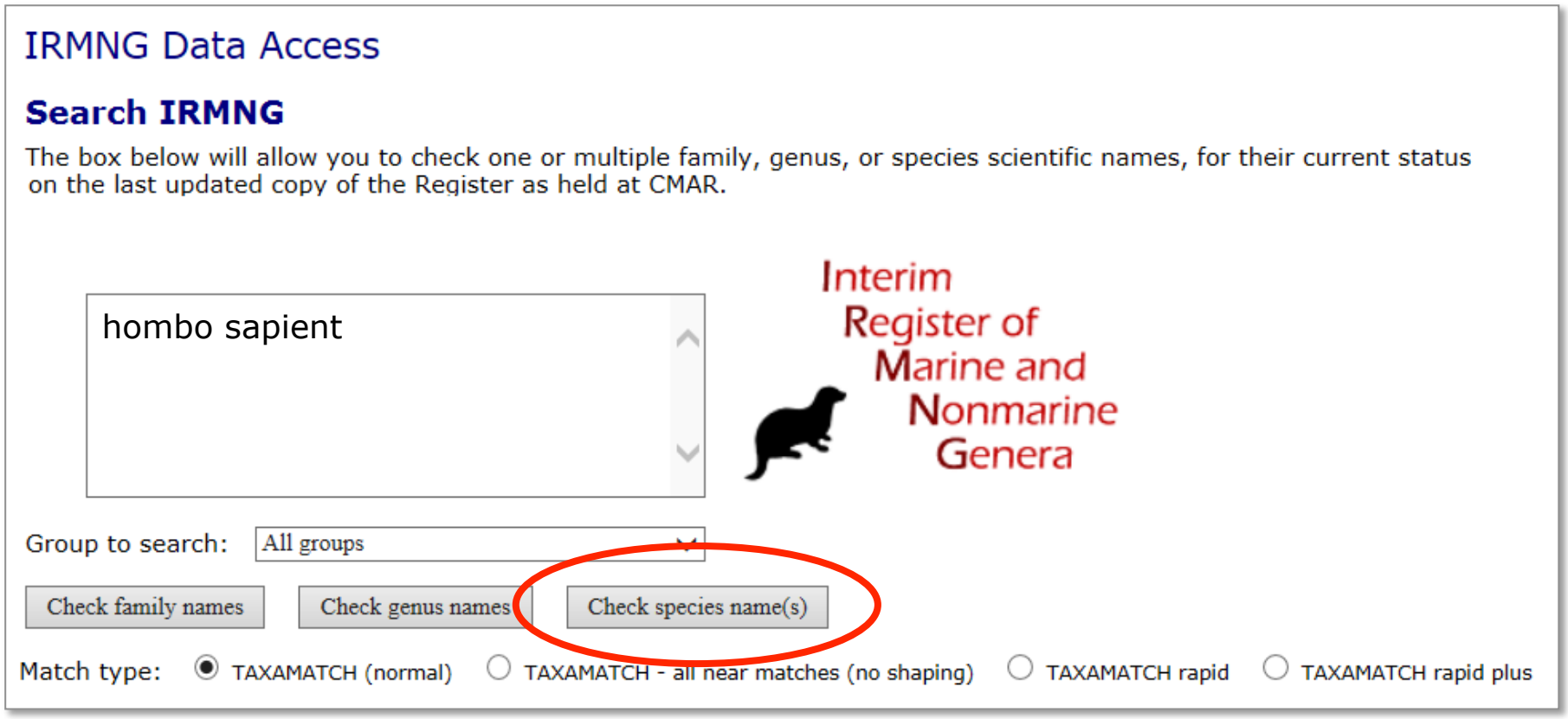
IRMNG web access point (www.cmar.csiro.au/datacentre/irmng)

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)



IRMNG web access point (www.cmar.csiro.au/datacentre/irmng)

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)



IRMNG web access point (www.cmar.csiro.au/datacentre/irmng)

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)

IRMNG search result (species search)

Query date/time: 10 Sep 2014 19:32

Species name entered: *Hombo sapient*

No exact genus matches found.

Genus nearest matches: **Homo** Linnaeus, 1758 (Animalia-Primates)

Other genus near matches: **Hamba** Distant, 1907 (Animalia-Hemiptera) , **Homaïd** Adanson, 1763 (Plantae-Alismatales) , **Hombak** Adanson, 1763 (Plantae-Brassicales) , **Homea** Fleming, 1822 (Animalia-Myxiniformes) , **Homia** Blackwelder, 1952 (Animalia-Coleoptera)

Species, authority	*authority match?*		*family*	*source*	*hierarchy (kingdom-phylum-class-order)*
Species near matches:					
Homo sapiens Linnaeus, 1758 [ED: 1,1] 10857762			**Hominidae**	Integrated Taxonomic Information System (ITIS) via **Catalogue of Life** (2006 version)	Animalia-Chordata-Mammalia-Primates

The following IRMNG taxa have either the same, or a very similar species epithet:
(none)

Query time: 0.701 seconds
548 of 469307 IRMNG genera tested
50 of 1900255 IRMNG species tested

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)

IRMNG search result (species search)

Query date/time: 10 Sep 2014 19:32

Species name entered: *Hombo sapient*

Genus test results

No exact genus matches found.

Genus nearest matches: **Homo** Linnaeus, 1758 (Animalia-Primates)

Other genus near matches: **Hamba** Distant, 1907 (Animalia-Hemiptera) , **Homaïd** Adanson, 1763 (Plantae-Alismatales) , **Hombak** Adanson, 1763 (Plantae-Brassicales) , **Homea** Fleming, 1822 (Animalia-Myxiniformes) , **Homia** Blackwelder, 1952 (Animalia-Coleoptera)

Species, authority	*authority match?*		*family*	*source*	*hierarchy (kingdom-phylum-class-order)*
Species near matches:					
Homo sapiens Linnaeus, 1758 [ED: 1,1] 10857762			**Hominidae**	Integrated Taxonomic Information System (ITIS) via **Catalogue of Life** (2006 version)	Animalia-Chordata-Mammalia-Primates

The following IRMNG taxa have either the same, or a very similar species epithet:
(none)

Query time: 0.701 seconds
548 of 469307 IRMNG genera tested
50 of 1900255 IRMNG species tested

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)

IRMNG search result (species search)

Query date/time: 10 Sep 2014 19:32

Species name entered: *Hombo sapient*

No exact genus matches found.

Genus nearest matches: **Homo** Linnaeus, 1758 (Animalia-Primates)

Other genus near matches: **Hamba** Distant, 1907 (Animalia-Hemiptera) , **Homaïd** Adanson, 1763 (Plantae-Alismatales) , **Hombak** Adanson, 1763 (Plantae-Brassicales) , **Homea** Fleming, 1822 (Animalia-Myxiniformes) , **Homia** Blackwelder, 1952 (Animalia-Coleoptera)

Species near matches:

Species, authority	*authority match?*		*family*	*source*	*hierarchy (kingdom-phylum-class-order)*
Homo sapiens Linnaeus, 1758 [ED: 1,1] 10857762			**Hominidae**	Integrated Taxonomic Information System (ITIS) via **Catalogue of Life** (2006 version)	Animalia-Chordata-Mammalia-Primates

The following IRMNG taxa have either the same, or a very similar species epithet:
(none)

Query time: 0.701 seconds
548 of 469307 IRMNG genera tested
50 of 1900255 IRMNG species tested

Species test results

Taxamatch in practice – e.g. against IRMNG reference database (>460k genera, 1.9m species)

IRMNG search result (species search)

Query date/time: 10 Sep 2014 19:32

Species name entered: *Hombo sapient*

No exact genus matches found.

Genus nearest matches: **Homo** Linnaeus, 1758 (Animalia-Primates)

Other genus near matches: **Hamba** Distant, 1907 (Animalia-Hemiptera) , **Homaïd** Adanson, 1763 (Plantae-Alismatales) , **Hombak** Adanson, 1763 (Plantae-Brassicales) , **Homea** Fleming, 1822 (Animalia-Myxiniformes) , **Homia** Blackwelder, 1952 (Animalia-Coleoptera)

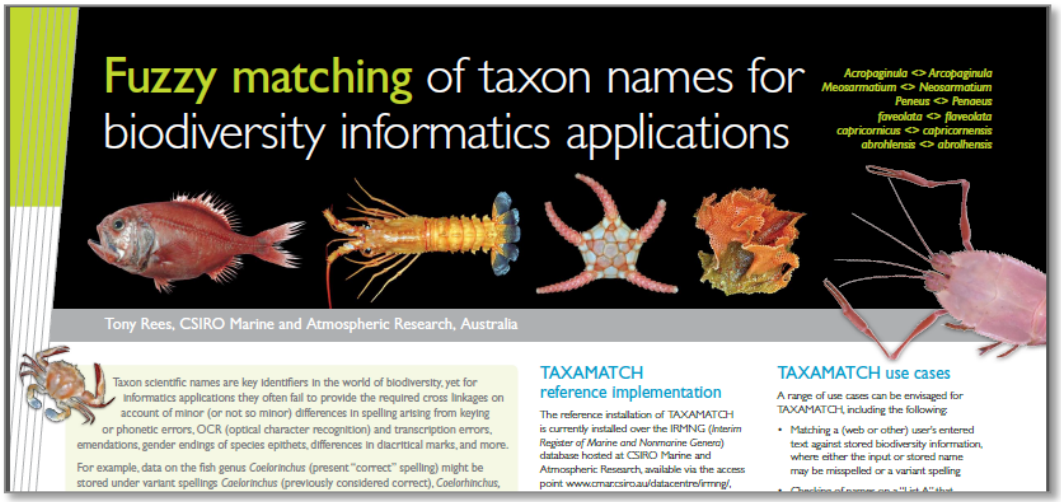
Species, authority	*authority match?*	*family*	*source*	*hierarchy (kingdom-phylum-class-order)*
Species near matches:				
Homo sapiens Linnaeus, 1758 [ED: 1,1] 10857762		**Hominidae**	Integrated Taxonomic Information System (ITIS) via **Catalogue of Life** (2006 version)	Animalia-Chordata-Mammalia-Primates

The following IRMNG taxa have either the same, or a very similar species epithet:
(none)

Query time: 0.701 seconds
548 of 469307 IRMNG genera tested
50 of 1900255 IRMNG species tested

Algorithm performance results

Taxamatch published description



1: Poster, "e-Biosphere" international meeting, London, 2009

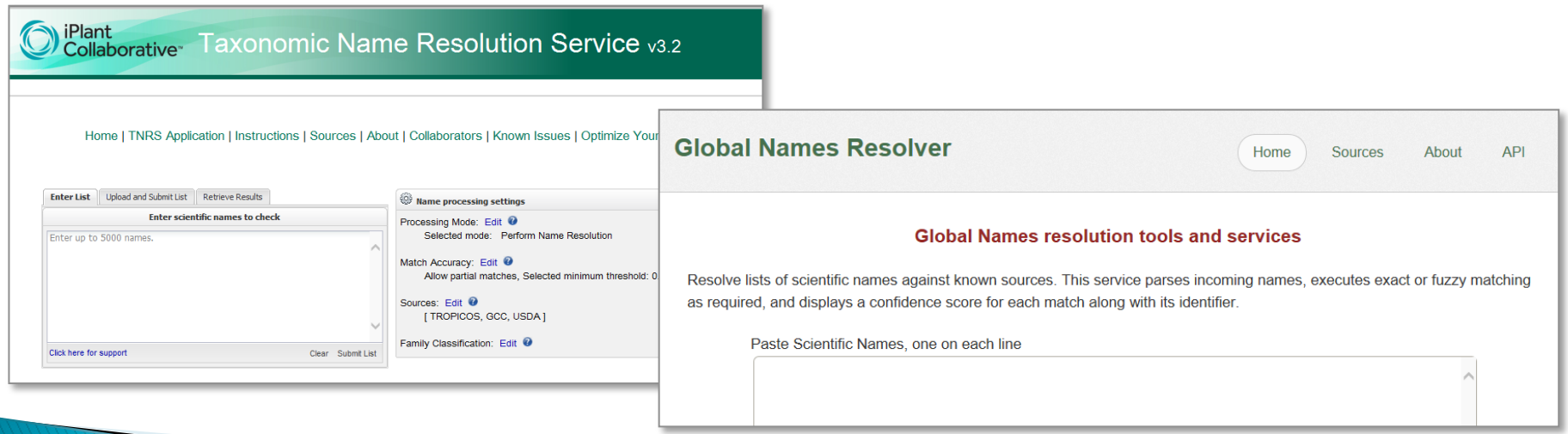
2: Manuscript in press (accepted by PLOS One), 2014

1	**Title**
2	Taxamatch, an Algorithm For Near ('Fuzzy') Matching of Scientific Names in Taxonomic Databases
3	**Author**
4	Tony Rees
5	**Affiliation**
6	CSIRO Oceans and Atmosphere Flagship, GPO Box 1538, Hobart, Tasmania, Australia

3: Taxamatch website at CSIRO Marine & Atmospheric Research

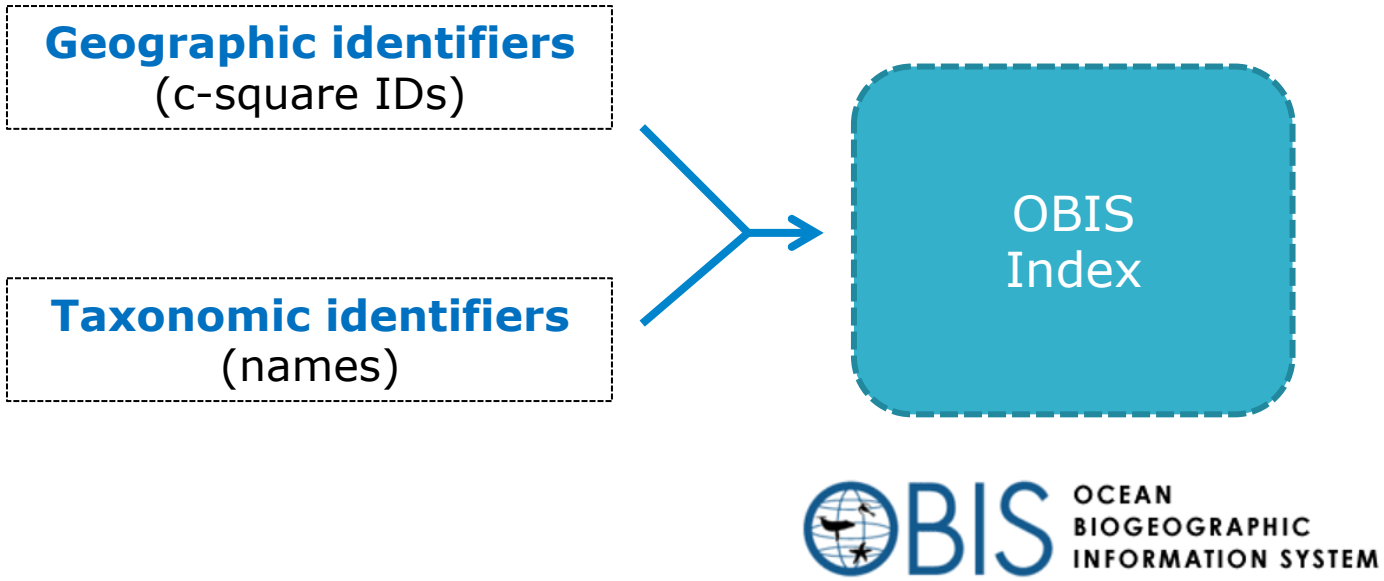
Taxamatch uptake

- Initially: in author's own systems at CSIRO (particularly IRMNG), also accessible by external web users
- Subsequently in WoRMS (World Register of Marine Species) + 30 co-hosted databases; Euro+Med PlantBase; Global Names Resolver; iPlant Taxonomic Names Resolution Service (TNRS) (USA); University of Vienna Herbarium; Pan-European Species Directories Infrastructure (PESI); more…
- GBIF and ALA have working code, not sure if integrated into production systems at this time.



IRMNG – a taxonomic name index for OBIS (initially)

IRMNG – a taxonomic name index for OBIS (initially)



A taxonomic name index for OBIS (1)

- Need to index the names as well as the occurrences (preferably in a tax. hierarchy)
- Need to reconcile synonyms, spelling variants to a standard form
- Need to distinguish marine from nonmarine, extant from fossil taxa (in OBIS, show only marine, extant taxa at this time)
- Ideally: need all relevant names, not just those in today's providers
  - Show current data gaps – i.e. name A: 100 records / name B: 0 records
  - Anticipate future providers coming on-line (know in advance where names belong in hierarchy)
  - Provide a "clean" list against which to test potential misspellings.

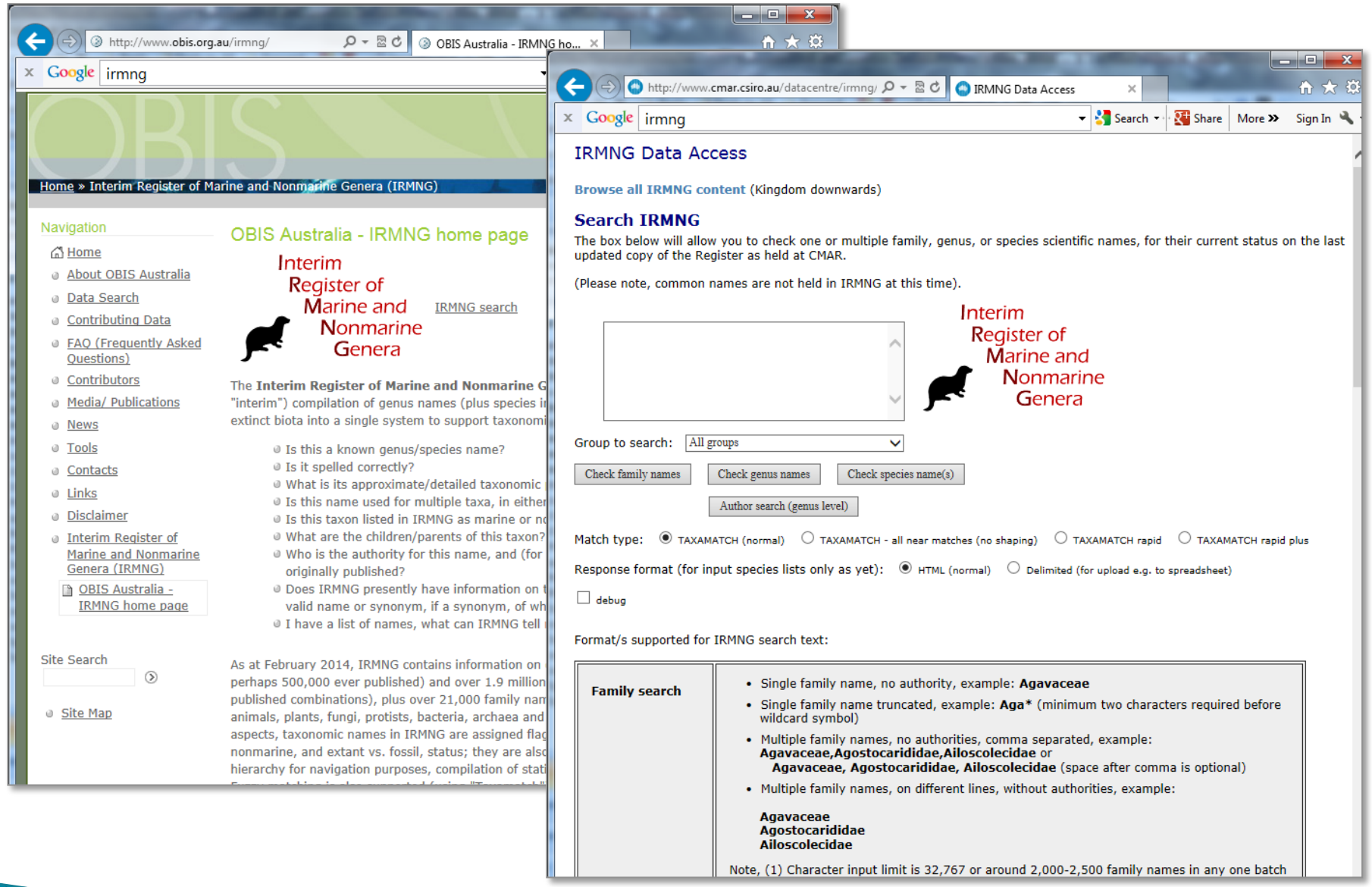
A taxonomic name index for OBIS (2)

- Listing all species is *hard* - Catalogue of Life task since 2000, still only 70% complete (actually: less), no fossil names included
- Listing all genera should be easier: 10x fewer names to compile, major compilations already exist (animals, plants, prokaryotes, viruses) – however level of taxonomic resolution varies
- 2006: Decision made to start a new compilation "IRMNG", based on genus names from Nomenclator Zoologicus (animals), Index Nominum Genericorum (plants) plus other compendia

IRMNG data compilation

- Family allocations, synonymy, marine/nonmarine flags added from multipe online + print sources (also extant/fossil flags) – ongoing activity
- Species also added to correct genus instances where available, e.g. from Cat. of Life, museum lists, regional lists, more…
- Target seems to be around 500k genus names for "all life", extant+fossil; 469k currently held (>92%), more in pipeline
- Also held: ~1.9m species names, from Cat. of Life + elsewhere (of maybe 5m ever published)
- Entire database is searchable via the web, also supplied as data dump to "power users" e.g. OBIS, WoRMS, GBIF, ALA, more…
- Master copy presently at CSIRO, plan to relocate it to VLIZ (Belgium) later this year – as per WoRMS, OBIS, other taxon DB's

Present IRMNG web access point(s)



Example IRMNG web query (1) – check species name/s

IRMNG search result (species search)

Query date/time: 09 Sep 2014 16:32

Input names entered:
Aricidea (Acmira) catherinae
Aricidea (Acmira) horikoshii
Aricidea (Acmira) lopezi

Species, authority	*authority match?*	*family*	*source*	*hierarchy (kingdom-phylum-class-order)*	*sp. extant flag*	*sp. habitat flag*	*remarks*	*synonym?*	*synonym of*
Species exact matches:									
Aricidea catherinae Laubier, 1967 10236747		**Paraonidae**	CoL2006/ITS; WoRMS (Mar 2013)	Animalia-Annelida-Polychaeta-Cirratulida	[E]	[M]		S	***Acmira catherinae***
(no near match species held in IRMNG at this time.)									
Aricidea lopezi Berkeley & Berkeley, 1956 11270211		**Paraonidae**	CoL2006/ITS; WoRMS (Mar 2013)	Animalia-Annelida-Polychaeta-Cirratulida	[E]	[M]		S	***Acmira lopezi***
(no near match species held in IRMNG at this time.)									

Species names not found:

Aricidea horikoshii **web links**
Genus exact match/es: ***Aricidea*** Webster, 1879 (Annelida, Cirratulida)

Species nearest matches: ***Aricidea horikoshi*** Imajima, 1973 (Annelida, Cirratulida)

Example IRMNG web query (1) – check species name/s

IRMNG search result (species search)

Query date/time: 09 Sep 2014 16:32

Input names entered:
Aricidea (Acmira) catherinae
Aricidea (Acmira) horikoshii
Aricidea (Acmira) lopezi

Species, authority	*authority match?*	*family*	*source*	*hierarchy (kingdom-phylum-class-order)*	*sp. extant flag*	*sp. habitat flag*	*remarks*	*synonym?*	*synonym of*
Species exact matches:									
Aricidea catherinae Laubier, 1967 10236747		**Paraonidae**	CoL2006/ITS; WoRMS (Mar 2013)	Animalia-Annelida-Polychaeta-Cirratulida	[E]	[M]		S	***Acmira catherinae***
(no near match species held in IRMNG at this time.)									
Aricidea lopezi Berkeley & Berkeley, 1956 11270211		**Paraonidae**	CoL2006/ITS; WoRMS (Mar 2013)	Animalia-Annelida-Polychaeta-Cirratulida	[E]	[M]		S	***Acmira lopezi***
(no near match species held in IRMNG at this time.)									

Species names not found:

Aricidea horikoshii **web links**
Genus exact match/es: ***Aricidea*** Webster, 1879 (Annelida, Cirratulida)

Species nearest matches: ***Aricidea horikoshi*** Imajima, 1973 (Annelida, Cirratulida)

Near match result (using Taxamatch)

Example IRMNG web query (2) – list genera (e.g. for all groups, starting with "Ma"…)

IRMNG search result (genus search)

Query date/time: 09 Sep 2014 16:41

Input names entered:
Ma*

genus (+ no. of species in IRMNG)		*family*	*source*	*hierarchy (kingdom-phylum-class-order)*	*gen. extant flag*	*gen. habitat flag*	*remarks*
Maabella Hastriter & Bush, 2006 1274667 (0)	full ref.	**Streblidae**	Hastriter & Bush, 2006	Animalia-Arthropoda-Insecta-Diptera	[E]	[N]	
† **Maaccoyella** 1059863 (1)	..	**Oxytomidae**	Museum Victoria KEmu database (Oct 2006)	Animalia-Mollusca-Bivalvia-Pterioida	[F]	[M]	Misspelling.
Maackia Clessin, 1880 1353560 (0)	full ref.	**Hydrobiidae**	Nomenclator Zoologicus	Animalia-Mollusca-Gastropoda-Littorinimorpha			
Maackia Ruprecht, 1856 1353590 (13)	full ref.	**Fabaceae**	SN2000/Takhtajan, 1997; Index Nominum Genericorum	Plantae-Magnoliophyta-Magnoliopsida-Fabales	E	N	
Maackiana Krivosheina, 1973 1437379 (1)	full ref.	**Stratiomyidae**	SN2000 unverified/Stang, 2004-present; Nomenclator Zoologicus	Animalia-Arthropoda-Insecta-Diptera	E	[N]	
Maackiechinogammarus Dybowski, 1926 1161647 (0)	full ref.	**Gammaridae**	Nomenclator Zoologicus	Animalia-Arthropoda-Malacostraca-Amphipoda	E	N	Unavailable r suppressed w Opinion 105
Maacoccus Tao, Wong & Chang, 1983 1018090 (5)	full ref.	**Coccidae**	SN2000 unverified/Stang, 2004-present; Nomenclator Zoologicus	Animalia-Arthropoda-Insecta-Hemiptera	E	[N]	
Maacynips Yoshimoto, 1963 1254385 (0)	full ref.	**Cynipidae**	Nomenclator Zoologicus	Animalia-Arthropoda-Insecta-Hymenoptera		[N]	
Maafu Solem, 1983 1448789 (0)	full ref.	**Stylommatophora (awaiting allocation)**	Nomenclator Zoologicus	Animalia-Mollusca-Gastropoda-Stylommatophora			
Maaia Gressitt, 1951 1354555 (1)	full ref.	**Cerambycidae**	CoL2006/TITAN; Nomenclator Zoologicus	Animalia-Arthropoda-Insecta-Coleoptera	E	[N]	

Example IRMNG web query (3) – search by genus author (includes "Page")

genus (+ no. of species in IRMNG)		*family*	*source*	*hierarchy (kingdom-phylum-class-order)*	*gen. extant flag*	*gen. habitat flag*	*remarks*
† **Kampecaris** Page, 1856 1174003 (0)	full ref.	**Myriapoda (awaiting allocation)**	Nomenclator Zoologicus	Animalia-Arthropoda-Myriapoda (awaiting allocation)-Myriapoda (awaiting allocation)	F	N	
† **Slimonia** Page, 1856 1353936 (1)	full ref.	**Hughmilleriidae**	Museum Victoria KEmu database (Oct 2006); Nomenclator Zoologicus	Animalia-Arthropoda-Merostomata-Eurypterida	[F]	[M]	
† **Stylonarus** Page, ? 1856 1317971 (0)	full ref.	**Eurypterida (awaiting allocation)**	Nomenclator Zoologicus	Animalia-Arthropoda-Merostomata-Eurypterida	[F]	[M]	
† **Stylonurus** Page, 1856 1355702 (0)	full ref.	**Eurypterida (awaiting allocation)**	Sepkoski (2002); Nomenclator Zoologicus	Animalia-Arthropoda-Merostomata-Eurypterida	[F]	[M]	err. pro -narus Page 1856 (Nomen. Zool.).
† **Ictinocephalus** Page, 1859 1165895 (0)	full ref.	**Pisces (awaiting allocation)**	Nomenclator Zoologicus	Animalia-Chordata-Pisces (awaiting allocation)-Pisces (awaiting allocation)	F		Unavailable name (nomen nudum).
† **Lyonothamnoxylon** Page, 1964 1111601 (0)	full ref.	**Rosaceae**	Index Nominum Genericorum	Plantae-Magnoliophyta-Magnoliopsida-Rosales	F	[N]	
Filamoeba Page, 1967 1045075 (0)	full ref.	**Filamoebidae**	SN2000/Cavalier-Smith, 2004; Nomenclator Zoologicus	Protista-Amoebozoa-Variosea-Varipodida	[E]		
Platyamoeba Page, 1969 1400559 (6)	full ref.	**Vannellidae**	SN2000/T. Cavalier-Smith et al., 2004; Nomenclator Zoologicus	Protista-Amoebozoa-Flabellinea-Vannellida	[E]	[M]	
Mesodixa Belkin, Heinemann & Page, 1970 1404111 (0)	full ref.	**Dixidae**	Nomenclator Zoologicus	Animalia-Arthropoda-Insecta-Diptera		[N]	
† **Mulleroxylon** V.M. Page, 1970	full ref.	**Magnoliopsida (awaiting**	Index Nominum Genericorum	Plantae-Magnoliophyta-Magnoliopsida-Magnoliopsida (awaiting	F	[N]	

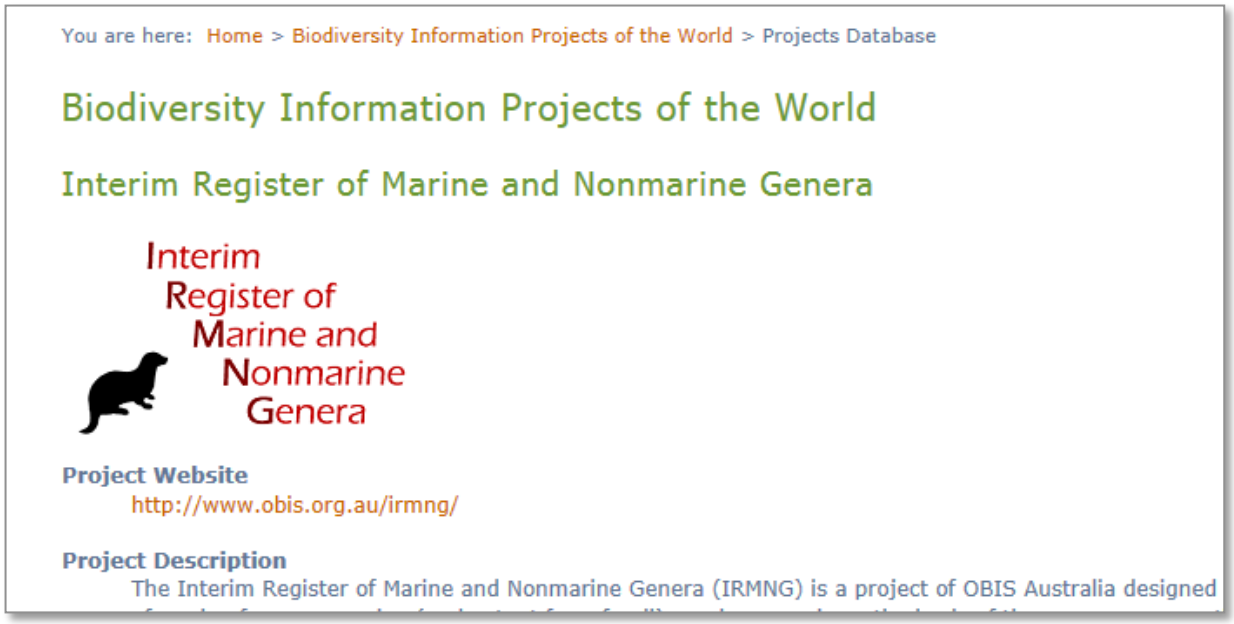
Other functions

- Browse entire hierarchy from top, or any entry point
- Filter query results by major taxonomic group e.g. molluscs, mammals, fungi, higher plants, more…
- Generate list of homonyms (only) as presently held at family, genus or species level
- Group genus names by publication year
- Other custom queries as desired at database level (limited only by level of information held for each name)

IRMNG "to do" list...

- Look for and add "missing" genus names (e.g. 30k+ , many 2005 onwards, mostly animals)
- Get detailed family placement for ~100k animal names presently resolved only to class or order
- Upgrade "orphan" names not in major nomenclators
  - some will be misspellings (detectable using Taxamatch), others genuine names that the nomenclators have missed!
- Add cross links to ION IDs, expand original publication info using R. Page dump of "Bionames" content (2012 version)
- Add more species from CoL 2014, The Plant List, etc.
- GBIF Ebbe Nielsen award will assist with the above!

IRMNG published description



1: Entry on TDWG "Biodiversity Information Projects of the World" site

2: Current IRMNG home page (on OBIS Australia web site) and data access/search page

3: Future IRMNG home page at VLIZ (to be constructed late 2014)

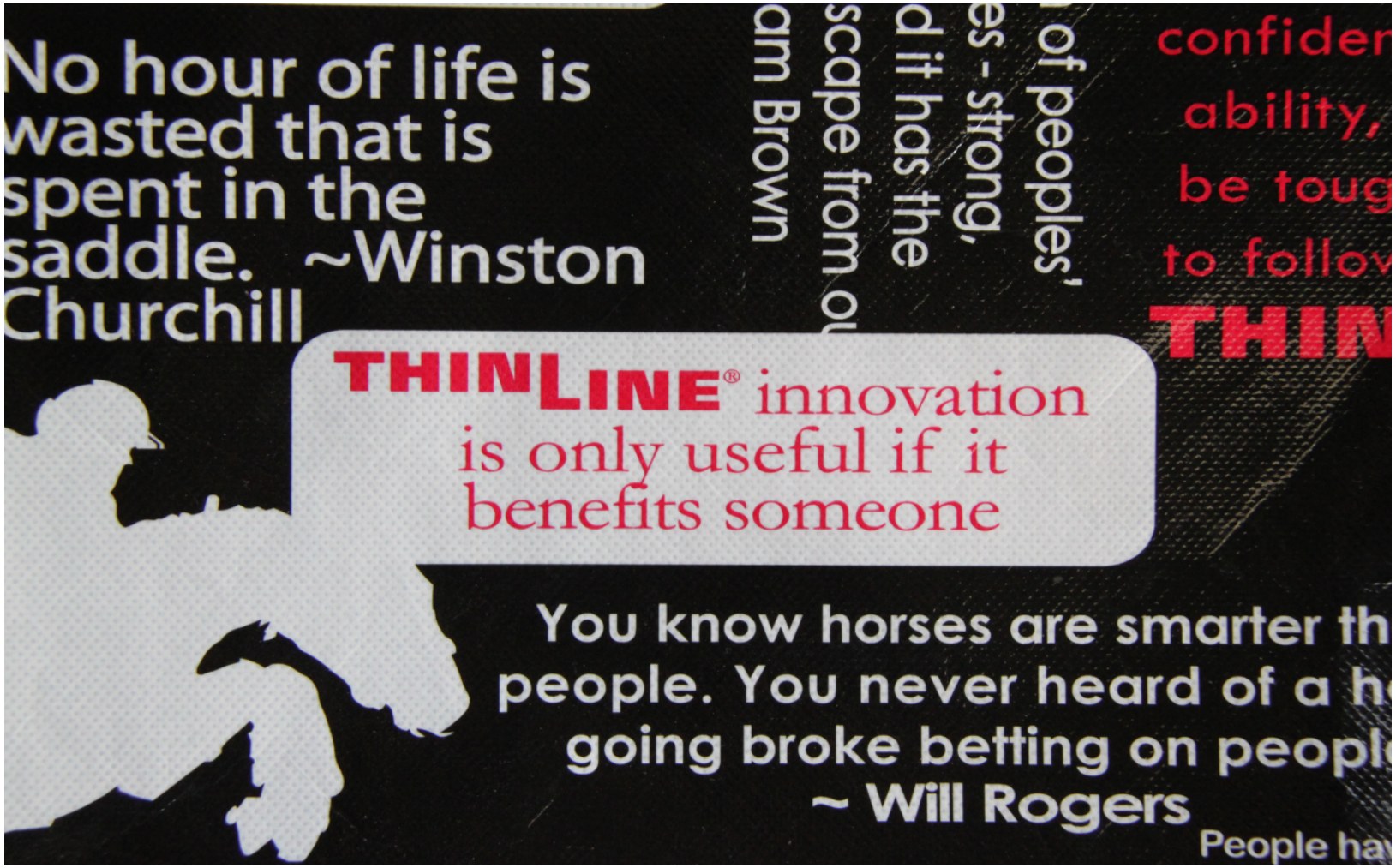
IRMNG uptake

- Standalone/master copy is used for ongoing data cleaning / sorting at author's agency and public access via external web queries
- Copy supplied to WoRMS and OBIS is used for OBIS data validation, WoRMS extension and data cleaning
- Copy supplied to ALA is used for marine/nonmarine flags (traits), some taxonomic allocation
- Copies supplied to GBIF, Open Tree of Life (OTOL), and EOL are used as input to "master taxonomic backbone" creation / providing alternative trees
- Copy supplied to Global Names is accessible via Global Names Resolver (look up names against multiple data sources)
- More in future??

Concluding remarks

- These tools developed (mostly) independently, but can work well together – e.g. OBIS and c-squares, IRMNG and Taxamatch
- IRMNG content is currently used as a contributor to GBIF taxonomic backbone, could possibly play a larger role (further discussion/thought needed)
- "Global Names" project is supposed to play an integrating role between many such compilations – e.g. support automated data flows etc. (though not really happening yet)
- Further integration with e.g. taxonomic literature, GBIF data points would be valuable areas to progress – while not losing sight of the initial "names problem" (shared by many data aggregators e.g. GBIF).

Seen on a shopping bag:



Acknowledgements

- Funding: CSIRO, OBIS, GBIF, ALA
- IT assistance: Miroslaw Ryba, Pamela Brodie, Philip Bohm
- IRMNG bulk data sources: Sheila Brands, Ely Wallis, Ellen Farr, Edward Vanden Bergh, Joel Hallan, Dennis Gordon, David Remsen, Markus Döring, Rod Page/BioNames
- IRMNG content entry/markup assistance: Deng Palomares, Helen Morgan, Anna Povey, Robin Wilson, Gary Poore, Stevie Davenport
- Travel/meetings support: OBIS, AquaMaps, Global Names, ALA
- Colleagues and others for helpful discussions, manuscript reviews etc.
- GBIF Science Committee for 2014 ENP award and associated travel!

Useful links

c-squares

- www.cmar.csiro.au/csquares

OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

- www.iobis.org

Taxa match

- www.cmar.csiro.au/datacentre/taxamatch.htm

Interim Register of Marine and Nonmarine Genera

- www.obis.org.au/irmng ; www.cmar.csiro.au/datacentre/irmng

- Author: Tony.Rees@csiro.au (to end Oct), then tonyrees49@gmail.com

- Thank You!