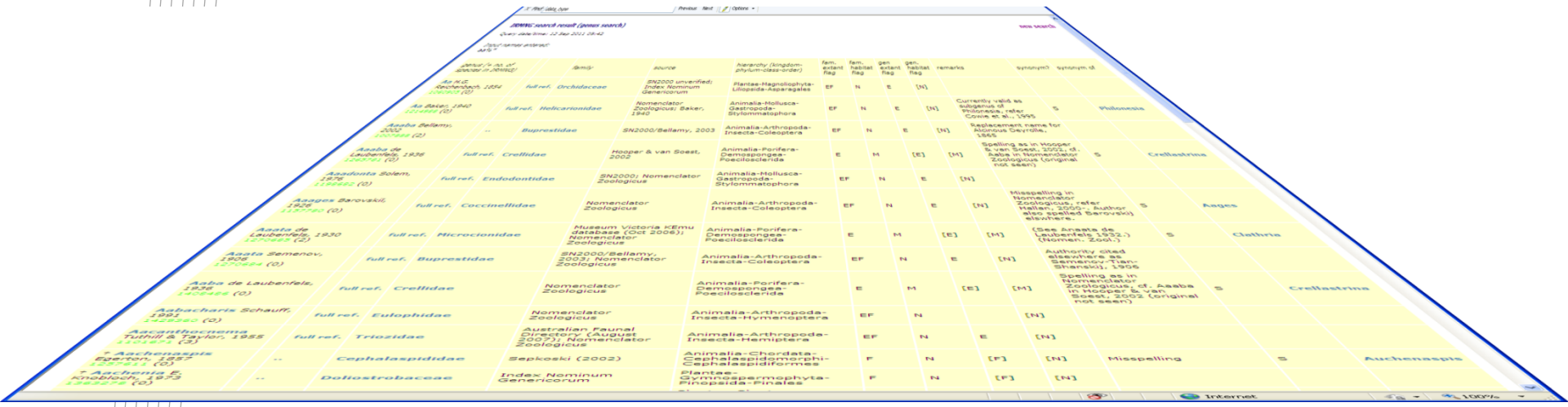


A continuously updated All Genera Index: an achievable goal for Biodiversity Informatics?

Tony Rees – CSIRO Marine and Atmospheric Research, Australia

TDWG Conference, October 2011

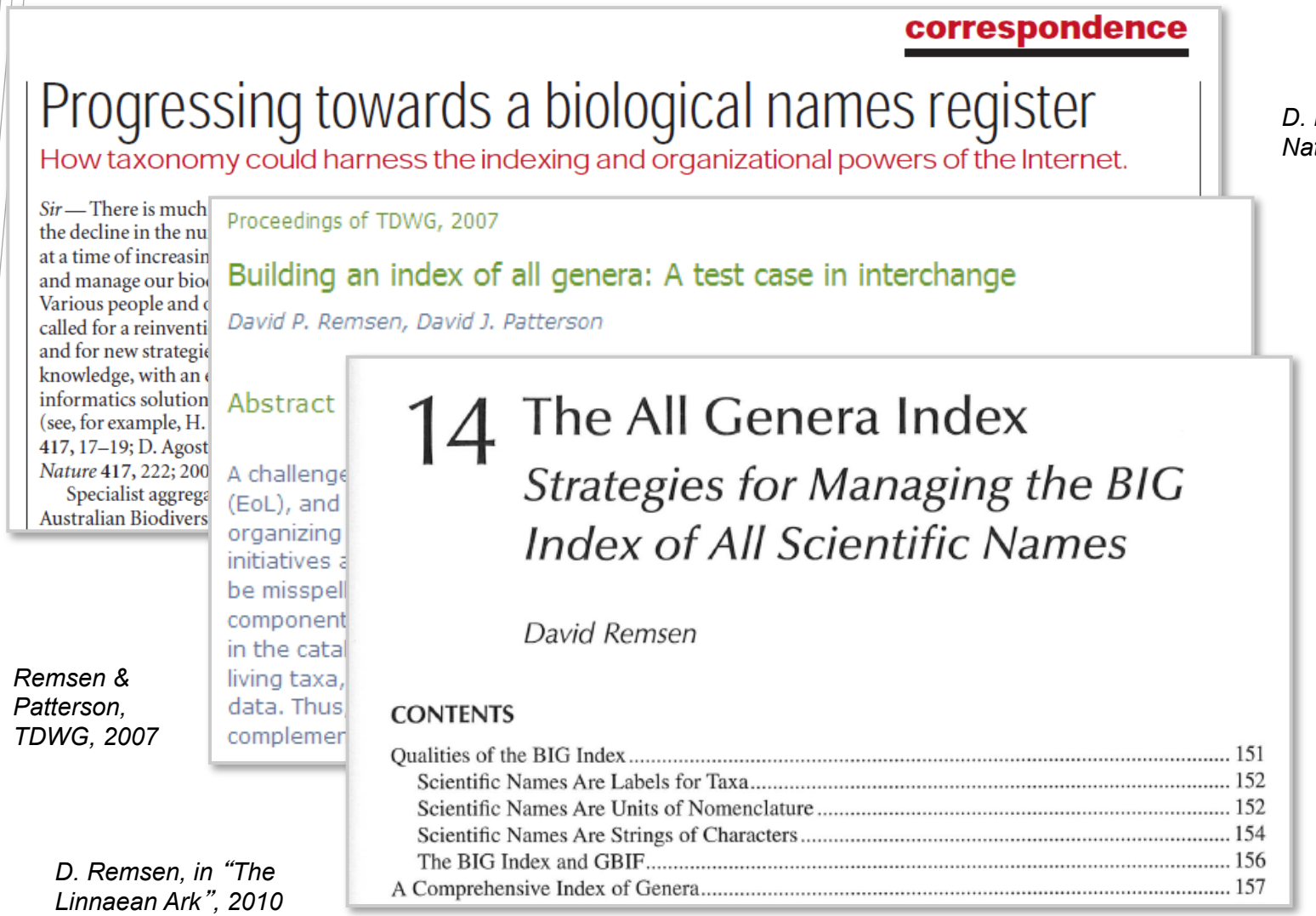
www.csiro.au

CSIRO

Why an All Genera Index?

- All-species index(es) will take time to complete, all-genera potentially more tractable:
  - ~10x smaller task (~2m valid species, maybe 250k genera)
  - leverage off existing genus-level compilations e.g. ING for plant names, Nomenclator Zoologicus for legacy animals, maybe ZooBank for future animal names, IPNI/others for plants
  - prokaryote, virus names also well curated and accessible
- Aim for horizontal coverage first (no missing tax. sectors, also include both extant + fossil names), vertical completeness e.g. to species level can be secondary consideration
- Can carry the burden of tax. assignments – then species merely need to be attached to the correct genus instance
- Genera can have significant nomenclatural and taxonomic interest i.e. valid vs. invalid names, author / year and place of publication (i.e. original work), genus-level synonyms and homonyms
- Can carry other attributes / assertions e.g. all species have trait “x”, occur in habitat “y”, within geological range “z”

Continuing a distinguished tradition…



correspondence

Progressing towards a biological names register

How taxonomy could harness the indexing and organizational powers of the Internet.

Sir — There is much
the decline in the nu
at a time of increasin
and manage our bio
Various people and
called for a reinventi
and for new strategie
knowledge, with an
informatics solution
(see, for example, H.
417, 17–19; D. Agost
Nature 417, 222; 200
Specialist aggrega
Australian Biodivers

Proceedings of TDWG, 2007

Building an index of all genera: A test case in interchange

David P. Remsen, David J. Patterson

Abstract

A challenge
(EoL), and
organizing
initiatives
be misspel
component
in the cata
living taxa,
data. Thus
complemen

14 The All Genera Index

Strategies for Managing the BIG Index of All Scientific Names

David Remsen

CONTENTS

D. Patterson, Nature, 2003

Remsen & Patterson, TDWG, 2007

D. Remsen, in "The Linnaean Ark", 2010

CSIRO

Different use cases, different approaches

- Remsen / Patterson / uBio approach (if correctly understood)
  - Assemble largest possible list of taxonomic names from multiple sources / provenance, reconciliation / deduplication / assignment to tax. hierarchy is subsequent activity
  - Main initial use case is for information retrieval / query expansion (multiple variants of name authorship are seen as valuable)
- Author / OBIS interest and approach
  - Starting point is a tax. hierarchy (kingdom through family), all names must live in this structure
  - Names from "trusted sources" given precedence, others used sparingly and subject to additional verification, multiple variants of name authorship are rationalized to single preferred version
  - Important focus (after tax. assignment) for OBIS is on attributes, in particular marine vs. nonmarine, extant vs. fossil – i.e. use the power of the list for non-tax. as well as taxonomic purposes
- Linkages to primary taxonomic literature also of potential value (allows harvesting of attributes, expanded understanding of original tax. concepts, more…)

Leverage existing genus-level compilations

Index Nominum Genericorum (ING)

[C] Aa H. G. Reichenbach, Xenia Orchid. 1: 18. 1 Apr 1854.
LT.: *A. paleacea* (Kunth) Schlechter (*Ophrys paleacea* Kunth) (vide Schlechter, Repert. Spec. N
309. 1892)
PHAN.-ORCHIDACEAE (111/104) 15 Mar 2010
Aachenia E. Knobloch, Neues Jahrb. Geol. Paläontol., Monatsh. 1972: 401. 1972 (post 18 Mar).
T.: *A. debeyi* E. Knobloch
Cone scales; Cretaceous (Senonian); Aachen, Germany.
FOSSIL-GYMNOSPERMAE-CONIFEROPHYTA (104) 19 Feb 1999
Aachenipollis W. Krutzsch, Paläontol. Abh., Abt. B, Paläobot. 3: 408. 1970.
T.: *A. aachenensis* Krutzsch
Cretaceous.
FOSSIL-SPORAE DISPERSAE (107) 9 Feb 1996
Aachenosaurus G. Smets, Aachenosaurus Multidens Reptile Foss. 20. 1888.

Tony Rees: Continuously Updated All Genera Index

CSIRO

Leverage existing genus-level compilations

Index Nominum Genericorum (ING)

[C] Aa H. G. Reichenbach, Xenia Orchid. 1: 18. 1 Apr 1854.
LT.: *A. paleacea* (Kunth) Schlechter (*Ophrys paleacea* Kunth) (vide Schlechter, Repert. Spec. N
309. 1892)
PHAN.-ORCHIDACEAE (111/104) 15 Mar 2010
Aachenia E. Knobloch, Neues Jahrb. Geol. Paläontol., Monatsh. 1972: 401. 1972 (post 18 Mar).
T.: *A. debeyi* E. Knobloch
Cone scales; Cretaceous (Senonian); Aachen, Germany

Ababa Casey 1897, Ann. N. York Acad., **19,** 653.—Col.
Ababactus Sharp 1885, Biol. Centr.-Amer., Zool., Col., **1** (2), 533.—Col.
Ababes Gray 1842, Syn. Cont. Brit. Mus., ed. 44, 150 [*n.n.*].—Pisces.
Abacella Stechow 1920, S.B. Ges. Morph. Phys., Münich, **31,** 37.—Coel.
Abacena Walker 1865, List Specimens Lep. Ins. Br. Mus., **34,** 1270.—Lep.
Abacetus Dejean 1828, Spec. gén. Coléopt., **3,** 195.—Col.
Abacidus Leconte 1863, List Coleopt. N. Amer., 9; 1873, Proc. Acad. nat. Sci. Philad., **1873,** 305.—Col.
Abacion Rafinesque 1820, Ann. of Nature, 9.—Pisces.
Abaciscus Butler 1889, Ill. Lep. Heteroc. Brit. Mus., **7,** 102.—Lep.
Abacistis Meyrick 1913, Ann. Transvaal Mus., **3,** 318.—Lep.
Abacobia Dietz 1905, Trans. Amer. ent. Soc., **31,** 29.—Lep. (See *Dietzia* Busck 1906.)
Abacobius Lacordaire 1866, Hist. nat. Ins., Gen. Col., **7,** 285.—Col.

(Nomenclator Zoologicus extract)

Characteristics of nomenclator-style compilations

- Emphasis is on nomenclatural information i.e. facts (name X was established by Y in publication Z on date D) and nomenclatural synonyms / rationale, subsequent tax. treatment ("opinions") may or may not be included
- Literature citations seen as critical component (excellent!), often verified from the original – i.e. a nomenclator can be considered a proxy for the primary literature
- Recent / on-line nomenclators often have full citation information / reference modules (e.g. Catalog of Fishes, Index Fungorum, Systema Dipterorum, more…)
  - ING and Nomenclator Zoologicus use the more terse "nomenclator style" or microcitation (no article title, full authorship or page range included) – less obvious for verifying/sourcing relevant attributes, or cross-linking to bibliographic lists
- Non-taxonomic attributes may also be included in some compilations, but not all.

Assembling the “desired” data set

- In practice, for the full set of desired information it may be necessary to supplement information from nomenclators with that from other sources i.e. subsequent tax. treatments and opinions, bibliographies / literature indexes, sources for attributes such as eco- and geo- characteristics
- Additional effort may be needed to massage supplied fragmentary / inconsistent taxonomies into a coherent whole at higher levels
- Higher tax. itself is a moving target too – e.g. for Angiosperms (APG, APG II, APG III…), protists, viruses and prokaryotes
- Information varies from readily available / well curated / comprehensive / current (for “examplar” groups) to fragmentary / out-of-date / hard-to-access / no recent overviews for others
- Desired level of detail is not available at genus level from current Cat. of Life, need to go to contributing GSDs, checklists, primary literature and elsewhere at this time (also to relevant sources for fossil taxa).

Author's experience to date

- First "cut" in 2003-4 as names indexing operation for OBIS, ramped up in 2006 as **IRMNG**, the **Interim Register of Marine and Nonmarine Genera**
  - Concept name follows ERMS, the European Register of Marine Species (now WoRMS), also including "Interim" for incomplete / provisional, but hopefully useable in its present state
  - Initial guesstimate to complete was 3-6 months (slight underestimate!)
  - All names sourcing and ingestion based on manual data loading at this time, would like to move to automated data feeds / updates as available in future versions
  - Uploading initial batches of data straightforward, problems come with subsequent ones required for gap filling, i.e.:
    - Duplicate and near-duplicate detection
    - Genus-level homonyms are a significant issue
    - Dealing with data conflicts – same name, different tax. opinions or orthographies for supplied information.

A portion of the IRMNG master genus table (as at Oct 2011)

MASTER_GENLIST

GENUS_ID	GENUS	AUTHORITY	FAMILY_	SOURCE	!	IS_SYN_O	IS_SYN_OF_NAME	GEN_TAX_REMARKS	C	I
1142224	A-Thienemannia	Viets, 1920	106933	Nomenclator Zoologicus	S	1035099	Athienemannia	Original incorrect spelling, see Athiener	E	N
1185070	A-omidimeroceros	Sobolev, 1914	103222	Nomenclator Zoologicus	S	1189249	Omadimeroceras	Variant spelling / representation	F	
1452697	AHJD-like viruses		106549	ICTVdb (Jul 2011)					E	
1060905	Aa	H.G. Reichenbach, 1854	114451	SN2000 unverified; Index Nominum Genei					E	N
1214988	Aa	Baker, 1940	111173	Nomenclator Zoologicus; Baker, 1940	S	1037448	Philonesia	Currently valid as subgenus of Philonesi	E	N
1265781	Aaaba	de Laubenfels, 1936	113273	Hooper & van Soest, 2002	S	1025271	Crellastrina	Spelling as in Hooper & van Soest, 2002,	E	M
1007888	Aaaba	Bellamy, 2002	100103	SN2000/Bellamy, 2003				Replacement name for Alcinous Deyroll	E	N

1345036	Aagyrus		100453	CoL2006/UCD
1232551	Aaka	Dworakowska, 1972	117103	Nomenclator Zoologicus
1058352	Aalatettix	Zheng & Mao, 2002	104521	SN2000 unverified/Stang, 2004-present; N
1364492	Aaleniella	Plumhoff, 1963	100808	Sepkoski (2002); Knitter, 1983
1364514	Aaleniella	Conti & Fischer, 1981	115598	Sepkoski (2002); Nomenclator Zoologicus
1387375	Aalenilla	Phemhoff, 1963	110737	Nomenclator Zoologicus
1057863	Aalenirhynchia	Shi & Grant, 1993	117179	Sepkoski (2002); Nomenclator Zoologicus
1309319	Aalius	Rumphius ex O. Kuntze, 18!	114198	CoL2006/RBG Kew Checklist; Index Nomini
1088737	Aalolana		109534	Museum Victoria KEmu database (Oct 200
1006518	Aancistroger	Bei-Bienko, 1957	101558	SN2000 unverified/Stang, 2004-present; N
1091104	Aaosphaeria	A. Aptroot, 1995	105753	SN2000/O.E. Eriksson, 2006; Index Nominu

1232551	Aaka	Dworakowska, 1972	117103	Nomenclator Zoologicus						N
1058352	Aalatettix	Zheng & Mao, 2002	104521	SN2000 unverified/Stang, 2004-present; N					E	N
1364492	Aaleniella	Plumhoff, 1963	100808	Sepkoski (2002); Knitter, 1983					F	M
1364514	Aaleniella	Conti & Fischer, 1981	115598	Sepkoski (2002); Nomenclator Zoologicus					F	M
1387375	Aalenilla	Phemhoff, 1963	110737	Nomenclator Zoologicus				Unconfirmed elsewhere; likely misspell	F	
1057863	Aalenirhynchia	Shi & Grant, 1993	117179	Sepkoski (2002); Nomenclator Zoologicus					F	M
1309319	Aalius	Rumphius ex O. Kuntze, 18!	114198	CoL2006/RBG Kew Checklist; Index Nomini					E	N
1088737	Aalolana		109534	Museum Victoria KEmu database (Oct 200	S	1379323	Aatolana	Presumed misspelling	E	M
1006518	Aancistroger	Bei-Bienko, 1957	101558	SN2000 unverified/Stang, 2004-present; N					E	N
1091104	Aaosphaeria	A. Aptroot, 1995	105753	SN2000/O.E. Eriksson, 2006; Index Nominu					E	

Services / views this currently supports

- High-level overview + relevant statistics for “all life” (currently possible for names, in future for valid taxa)
- Navigate the tax. hierarchy in any direction
- Generate hierarchical lists
- Generate alphabetic lists
- Sort / filter by any desired criteria, both taxonomic and non-taxonomic
- Generate lists of homonyms, within or across Codes
- Indicate current tax. hierarchy, nomenclatural / taxonomic status, and attributes (to varying degrees) for any input name
- Holds partial species lists for selected genus names e.g. from Cat.of Life (with permission) and elsewhere (could be developed further as desired)
- Indicate near match targets to any input name (“did you mean…”) – using TAXAMATCH fuzzy matching (latter also adopted by iPlant, PESI, GNI, more…)

IRMNG-generated statistics for "all life" (web query 6 Oct 2011)

Kingdom (+ no. of phyla/ classes/ orders/ families/ genera/ species in IRMNG)

Animalia (47/212/1454/14650/362475/1134339)

Archaea (3/11/17/31/111/315)

Bacteria (32/56/122/418/2837/12964)

Fungi (9/45/181/852/16217/87935)

Plantae (10/45/275/1626/52078/188833)

Protista (30/105/369/1913/18508/33512)

Viruses (2/3/10/100/424/3317)

Unaccepted (1/1/1/1/10/0)

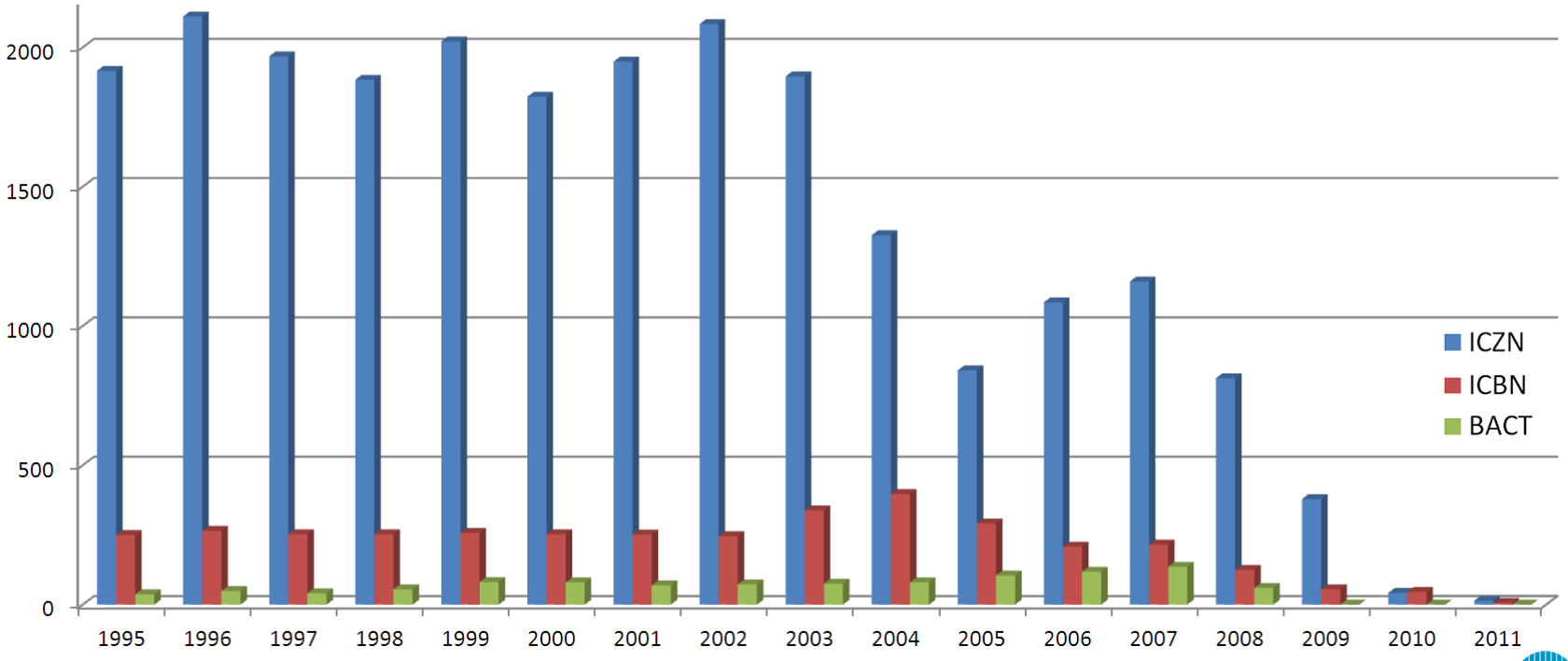
Unallocated (1/1/1/1/35/40)

IRMNG complete lists: **kingdoms** | **phyla** | **classes** | **orders** | **families**

- (NB, can also generate these lists as required via the web, by navigating the hierarchy, or enter the hierarchy at any level)

Current IRMNG status

- >450k genus names, in 17k+ families as at October 2011 (however significant subset, ~30%, still await family-level allocation)
- Start made on resolving genus-level synonyms on group-by-group basis, but much more to do
- Genus coverage considered >95% complete 1753-2003, less so for more recent data:



CSIRO

Some questions for this meeting

- Is this a worthwhile effort more generally i.e. as a community resource, cf. ongoing equivalent activities e.g. Catalogue of Life, GSDs, ITIS, PaleoDB, more…
- If so, where should it reside, who should manage/curate for the future
- To what extent can it leverage or synergise with emerging GN* activities and infrastructure
- To what degree can existing manual data upload / infill processes be automated
- How best to achieve continuing population and currency, e.g. as new names appear (~2k genera, 25k new species / yr if relevant).

Visit IRMNG at www.obis.org.au/irmng/

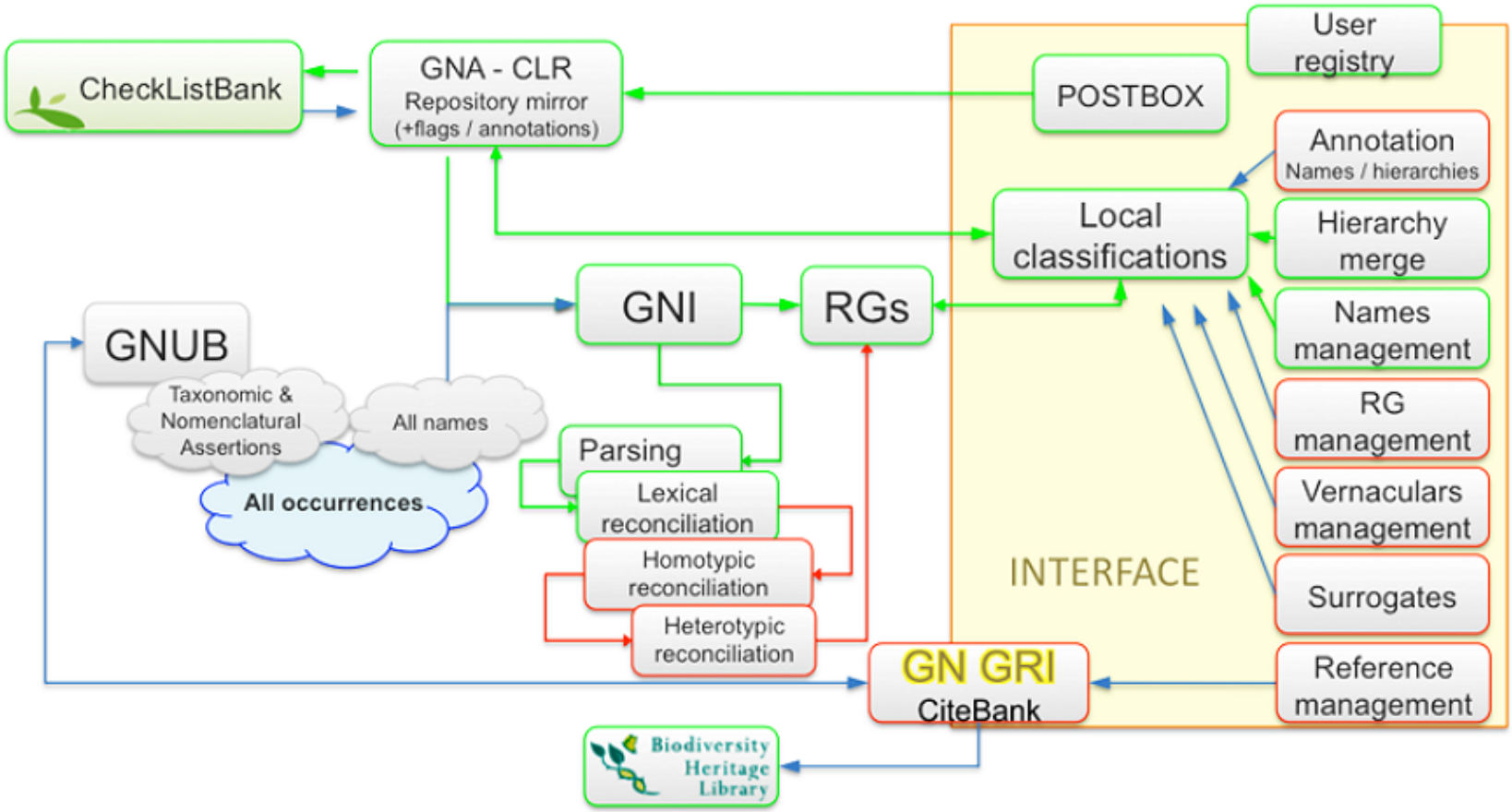
Thanks to data sources and funders who have contributed to development of IRMNG to date!

Thank you

Contact Us
Phone: 1300 363 400 or +61 3 9545 2176
Email: Tony.Rees@csiro.au Web: www.cmar.csiro.au/datacentre/

Supplementary slide

The emerging GN* world… – which elements relevant to this task?



CSIRO