

---

## Compiling the Global Compositae Checklist in the age of biodiversity informatics

Christina Flann<sup>1</sup> & Kevin Richards<sup>2</sup>

<sup>1</sup>Netherlands Centre for Biodiversity Naturalis (section NHN), Wageningen University, The Netherlands

<sup>2</sup>Allan Herbarium, Landcare Research, Manaaki Whenua, New Zealand

### Abstract

The days of manually compiling checklists are coming to an end as the options offered by biodiversity informatics change the systematic landscape. The Global Compositae Checklist is a project that utilises the possibilities of computer assisted checklist generation by integrating existing electronic datasets from around the world for this large plant family (approx 24,000 species). A purpose built computer program (C-INT) retains all original data and links names that are deemed the same to a 'consensus' name which reflects a summary of all contributed data. Insights from this process will be presented, discussing the advantages and pitfalls of trying to meld taxonomy and nomenclature with informatics technology. Issues that have arisen include how to obtain data and standardise the multitude of formats that data is contributed in, how to match names allowing for author variants, name orthographic variants, typographic errors and electronic artefacts as well as accommodating conflicting taxonomies.

### Keywords

Asteraceae, Biodiversity Informatics, Checklist, Compositae, Data Aggregation, Systematics, Taxonomy, TDWG Standards

### Introduction

The burgeoning field of biodiversity informatics is having a large impact on the manner of compiling taxonomic checklists. This paper is part of a symposium assessing the lessons learnt by the different approaches taken by various projects around the world to contribute to Target 1 of the Global Strategy for Plant Conservation (GSPC) - "A widely accessible working list of names of known plant species, as a step towards a complete world flora."

As part of this effort the Global Compositae Checklist (GCC) project has been running for four years, taking an electronic approach to the issue. As encapsulated in the name this project is global in scale and treats one of the largest flowering plant families, the Compositae (or Asteraceae), estimated to represent 10% of the world's vascular flora and containing approximately 24,000 species (Funk *et al.*, 2009). The GCC utilises computer assisted checklist generation by integrating existing electronic datasets from around the world to create a global list. The purpose built program (C-INT, Wilton & Richards, 2007) retains all original data and links names that are deemed the same to a 'consensus' name which reflects a summary of all contributed data. This result is then available as web-based output to allow wide access ([www.compositae.org/checklist](http://www.compositae.org/checklist)).

We have received more than seventy individual datasets (Figure 1, Table 1), ranging from local checklists to scanned flora entries to regional databases to global repositories. A number of these data providers were present at the 4<sup>th</sup> Global Botanic Gardens Conference. Fifty-three of these datasets have been integrated to date. Currently the Total number of individual Provider Name records included is 426,130 and the Total number of consensus Names across all ranks is 152,538. Species is arguably the most important rank and the current data contains 28,969 accepted species, 45,178 synonyms and 35,135 with no taxonomic concept. There is a known element of duplication in the accepted names that is likely to account for a few thousand names which are currently being dealt with.

---

## Discussion

The first issue worth noting in this process is that of expectations. There is currently a plethora of electronic based projects, all of which are collaborations between two different fields: Systematics/Taxonomy and Biodiversity Informatics. These fields have different focuses and priorities. Broadly speaking taxonomists desire definitive information reflecting the current taxonomic viewpoint (i.e. data content quality), while biodiversity informaticians are primarily interested in utilising computer tools to aggregate data in the most logical way, using the current standards for exchange of biodiversity data (i.e. the processes to deal with the data). Ultimately the lesson here is that it is very important to make sure the needs of all parties involved are covered and everyone is on the same page.

Communication is an integral issue as these two fields have different ways of communicating, using different languages eg SQL server scripts compared with the International Code of Botanical Nomenclature (McNeill *et al.*, 2006). Each side has to learn to know if we do actually want to inner join on the homonyms or not. A major lesson has been that a good relationship between the taxonomist and biodiversity informatician involved is fundamental and vital to the success of the project. The importance of the collaborative relationship built between the two authors in achieving the results we have so far cannot be stressed enough.

Communication between computers is also not as easy as it may first appear. At the beginning of this project there were many groups working on aggregation software and the prospects of avoiding reinventing the wheel seemed positive. However, the reality of reusing code designed for a slightly different purpose is that there are many computer languages, systems, programs, and setups, and making things truly interoperable is seldom practical. Despite this, we have successfully used an author thesaurus from IPNI (The International Plant Name Index, 2008) and a name de-duplication tool by Julius Welby (EPU, 2009). The lesson is that everyone wants to work together and it is possible, but not quite to the degree that would be ideal.

Now a series of related issues regarding the data itself, the names and taxonomic concepts, will be dealt with. First, obtaining the data to aggregate has been slow and patchy. Some data has been promised and not delivered, some data has arrived years later than expected, and data for some regions has not been easy to find. At the time of the Gardens Conference, the GCC owed a download of data to another project presenting in the symposium. This has since been delivered, but these deadlines have a tendency to get moved. Part of the issue is that there is little leverage available when asking people to contribute their data voluntarily and time needs to be allowed for delays.

The most problematic issue encountered is that of data format. Every dataset is different as they were all made for slightly different purposes. Fundamentally they are all files full of plant names, but some have all name elements parsed into separate fields, some have everything in one field, some have authors abbreviated, some have year of publication separate to the citation and so on. The step of standardising data in order to be able to integrate it has been a huge bottleneck. The lesson here is that there is a need for standards.

The Taxonomic Databases Working Group (TDWG) would agree with this, but there are also issues with standards. They have a tendency to be more fluid than standard. There is continually new best practice. In 2006 the Taxon Concept Schema was the standard so we adopted it, now it is not strictly considered a current standard by TDWG themselves, but a "Current (2005) Standard" (TDWG). Biodiversity informatics is an innovating field by definition, which continually improves the tools it works with. Much of the data used in this project was developed before TDWG and it is unlikely that for example, a taxonomist's private Cuban endemic checklist is ever going to be set up according to any standard. The lesson here is that standards only help when widely used and stable.

Details about the approach to matching taxonomic name data can be found in Richards *et al.* (in press). The problems encountered include dealing with author variants, orthographic variants and typographic errors as well as accommodating conflicting taxonomies. There are tricky cases such as 'ex' authors and misapplied names which data sources have dealt with in varying manners. Regardless of how good the matching algorithms used are, the approach needs to account for straight-out mistakes and electronic artefacts. To successfully encompass all of this makes matching algorithms a non-trivial matter.

After negotiating all of the above mentioned issues and having successfully achieved a standardised dataset that has been integrated into the database through matching with the data previously received then you come to the data content vetting. At this stage in assessing the aggregated summarised consensus data that should cover the global group and has been sorted using logical rules we have found unexpected inconsistencies. For example, accepted genera with no accepted species; or the opposite, genera coming up as synonyms while species in the genus are accepted. These cases are often the result of a difference of scale in datasets, one dealing only in genera, another only in species. Other issues have also become apparent, such as multiple accepted homonyms and more author abbreviation variants than covered by the included thesaurus leading to duplication of names in the database. The lesson here is to expect the unexpected and that data aggregation may not be the panacea we wish it to be.

Is aggregation enough? The simple answer is no. Regardless of how smart your algorithms are, having got all of the data together, it still has to be checked by experts. This is always a voluntary vetting process. The GCC website has been internally released to The International Compositae Alliance (TICA) community for a few months now and Google Analytics show 278 unique visitors who, on average, spend around 10 minutes on the site and look at about 8 separate pages. These visits have come from 46 different countries covering all continents, which is very positive. However, the majority of concrete content feedback has come from two workshops where experts were asked to look at the website and check the content for their groups. After this around 100 individual feedback emails were received from seven experts. We are very pleased by this response. However, for a group with 24,000 species, this is a drop in the ocean in terms of producing an expert validated list. The lesson is that checking the quality of data content is a huge task that is necessary and we are expecting taxonomic experts to do this on a voluntary basis.

## Conclusion

In conclusion, this is progress towards a list of known species for this significant plant family, which has a decent looking output to contribute to Target 1. This data is contributed to the Species2000 Catalogue of Life, the Encyclopedia of Life and has been added to the Target 1 project also discussed in this symposium by Chuck Miller and Bob Allkin. The current funding for the GCC is coming to an end and while this is an impressive base, it is only ultimately worthwhile if the project is continued and vetted.

## Acknowledgements

We gratefully acknowledge the assistance of all of the generous data providers (see Table 1); Ilse Breitwieser, Aaron Wilton, Jerry Cooper, Adam Thomas (Landcare Research, NZ), Julius Welby, Nicholas Hind, Christine Barker (RBG Kew UK), Vicki Funk, Erika Gonzalez (Smithsonian Institution, USA); The International Compositae Alliance; Wageningen University, NL. Funding: GBIF, NWO, The Systematics Association, EDIT, Smithsonian Institute, 4D4Life.

---

**References**

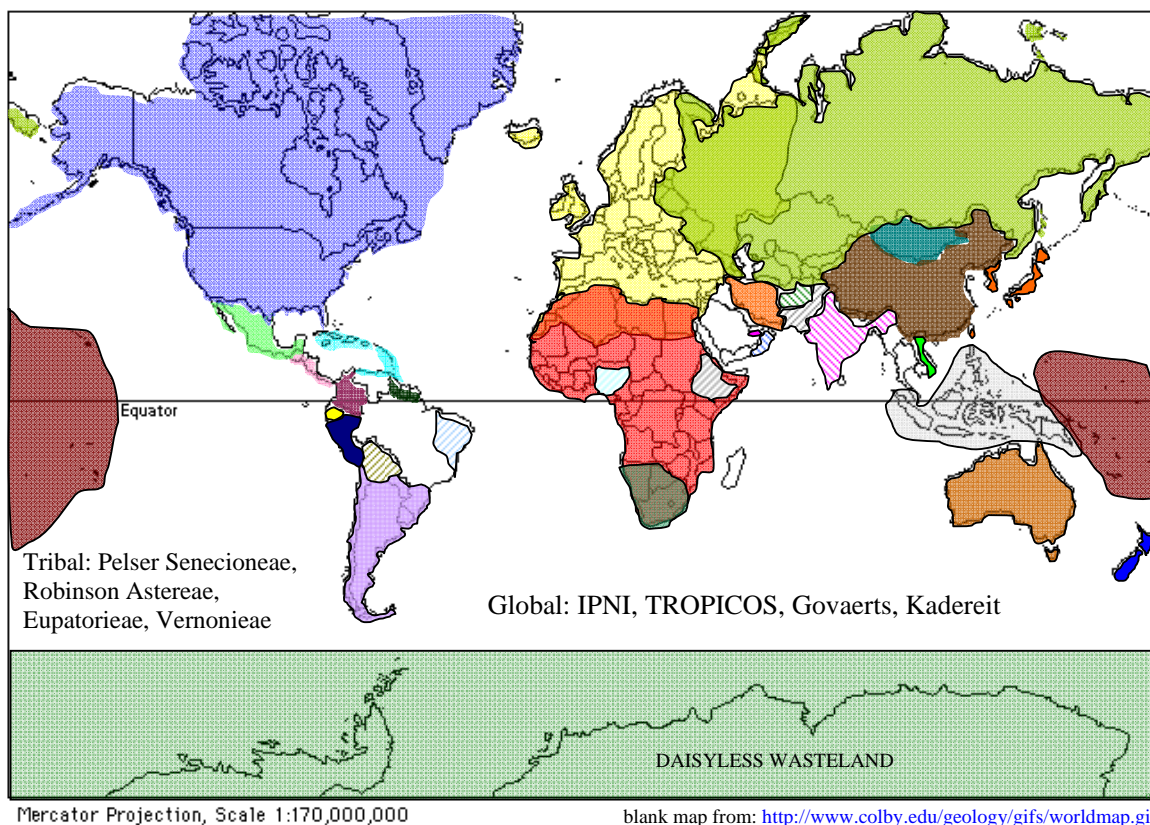
- African Plants Database, <http://www.ville-ge.ch/musinfo/bd/cjb/africa/index.php?langue=an>
- Baikov, K.S. (ed.), 2005. *Conspectus florum Sibiriae: plantae vasculares*. Science Publishers, Novosibirsk. 362 pp. (In Russian).
- Bostock, P.D. & Holland, A.E. (eds), 2007. *Census of the Queensland Flora*. Queensland Herbarium, Environmental Protection Agency, Brisbane.
- Brown, G. & Sakkir, S., 2004. *The Vascular Plants Of Abu Dhabi Emirate*. Terrestrial Environment Research Centre Environmental Research & Wildlife Development Agency [Internal Report].
- Castelo, E., Ricalde, O. & Panero, J.L., 2005. *Actualización del catálogo de autoridades de las Asteraceae, Tribu Heliantheae y Eupatorieae*. Herbarium, The University of Texas. Base de datos SNIBConabio, proyecto CS011. [http://siit.conabio.gob.mx/pls/itisca/taxaget?p\\_ifx=itismx&p\\_lang=es](http://siit.conabio.gob.mx/pls/itisca/taxaget?p_ifx=itismx&p_lang=es)
- Czerepanov, S.K., 1995. *Vascular Plants of Russia and Adjacent States (the Former USSR)*. Cambridge University Press.
- Electronic Flora of South Australia, <http://www.flora.sa.gov.au/>
- Enomoto, T., 1992. *Weed and Alien Species in Chichi-jima and Haha-jima, Ogasawara Islands*. Ogasawara Kenkyu Nenpo 16: 3-17. (In Japanese)
- Euro+Med PlantBase, <http://www.emplantbase.org/home.html>
- Flora Mesoamericana, <http://www.tropicos.org/Project/FM>
- Flora of Japan Database, <http://foj.c.u-tokyo.ac.jp/gbif/>
- Flora of Tasmania Online: an eFlora for the State of Tasmania, <http://demo1.tmag.tas.gov.au/index.html>
- Funk, V., Hollowell, T., Berry, P., Kelloff, C. & Alexander, S.N., 2007. *Checklist of the Plants of the Guiana Shield (Venezuela: Amazonas, Bolivar, Delta Amacuro; Guyana, Surinam, French Guiana) Contributions*. <http://botany.si.edu/bdg/index.html>
- Funk, V., Susanna, A., Stuessy, T.F. & Robinson, H., 2009. Chapter 11: Classification of Compositae. In: Funk, V., Susanna, A., Stuessy, T.F. & Bayer, R.J. *Systematics, Evolution, and Biogeography of Compositae*. IAPT, Vienna. pp. 171-176.
- Gubanov, I.A., 1996. *Conspectus of the Flora of Outer Mongolia (Vascular Plants)*. Moscow. (In Russian).
- Hind, D.J.N. & Jeffrey, C., 2001. A checklist of the Compositae of Vol. IV of 'H.B.K.'s' Nova Genera et Species Plantarum. *Compositae Newsletter* 37: 1-84.
- Hind, D.J.N. & Miranda, E.B., 2008. *Lista preliminar da familia Compositae na regio Nordeste do Brasil/Preliminary list of the Compositae in Northeastern Brazil* (Repatriation of Kew Herbarium data for the Flora of Northeastern Brazil Series, vol. 4). Royal Botanic Gardens, Kew. pp. i-xxv, 1-104.
- Hind, D.J.N., 2009. *An annotated preliminary checklist of the Compositae of Bolivia*. The Herbarium, Library, Art & Archives, Royal Botanic Gardens, Kew. [www.kew.org/science/tropamerica/boliviacompositae/checklist.pdf](http://www.kew.org/science/tropamerica/boliviacompositae/checklist.pdf).
- Kadereit, J.W. & Jeffrey, C. (eds), 2006. *The families and genera of vascular plants. VIII, Flowering plants. Eudicots: Asterales*. Springer, Berlin.
- Kerrigan, R.A. & Albrecht, D.E., 2007. *Checklist of Northern Territory Vascular Plant Species*.
- Le Kim Bien., 2005. *Danh lục các loài thực vật Việt Nam/ Checklist of plant species of Vietnam. Vol. 3*. Nhà xuất bản Nông Nghiệp

- Lepschi, B.J., Mallinson, D.J. & Cargill, D.C. (eds), 2008. *Census of the Vascular Plants, Hornworts and Liverworts of the Australian Capital Territory. Version 2.0.* <http://www.anbg.gov.au/cpbr/ACT-census/index.html>
- Luteyn, J.L. (ed.), 1999. Páramos, a Checklist of Plant Diversity, Geographical Distribution and Botanical Literature. *Memoirs of the New York Botanical Garden*, Vol. 84.
- McNeill, J., Barrie, F.R., Burdet, H.M., Demoulin, V., Hawksworth, D.L., Marhold, K., Nicolson, D.H., Prado, J., Silva, P.C., Skog, J.E., Wiersema, J.H., & Turland, N.J. (eds), 2006. *International Code of Botanical Nomenclature (Vienna Code) adopted by the Seventeenth International Botanical Congress Vienna, Austria, July 2005.* A.R.G. Gantner Verlag, Ruggell, Liechtenstein. [Regnum Veg. 146]
- Mota, J.F., Medina-Cazorla, J.M., Navarro, F.B., Pérezía, F.J., Pérez-Latorre, A., Sánchez-Gómez, P., Torres, J.A., Benavente, A., Gabriel Blanca, G., Gil, C., Lorite, J. & Merloa, M.E., 2008. Dolomite flora of the Baetic Ranges glades (South Spain). *Flora* 203: 359–375.
- New South Wales Flora Online, <http://plantnet.rbgsyd.nsw.gov.au/floraonline.htm>
- New Zealand Plant Names Database, <http://nzflora.landcareresearch.co.nz/default.aspx?NavControl=home>
- Plants of Southern Africa: an online checklist, <http://posa.sanbi.org/intro.php>
- Rechinger, K.H. (ed), 1906 – 1998. *Flora Iranica*. Akademische Druck und Verlagsanstalt, Austria.
- Richards, K, Wilton, A., Flann, C. & Cooper, J., In press. A Consensus Method for Checklist Integration. *Conference Proceedings, 14th Annual KES Conference*, Cardiff, Wales, UK, 8-10 September 2010.
- Tadesse, Mesfin, 2004. Asteraceae (Compositae). In: Hedberg, I., Friis, I. & Edwards, S., (eds), *Flora of Ethiopia and Eritrea*. Vol. 4 (2), National Herbarium, Addis Ababa University, Ethiopia, and Department of Systematic Botany, Uppsala University, Sweden.
- Taxon Concept Transfer Schema (TCS) <http://www.tdwg.org/standards/117/>
- Taxonomic Databases Working Group (TDWG) <http://www.tdwg.org/>
- The International Plant Names Index (2008). Published on the Internet <http://www.ipni.org> [accessed 14 February 2008].
- The Western Australian Census of Plant Names (WACensus) database
- Toyoda, T., 2003. *Flora of Bonin Islands*. Aboc-sha Co., Ltd. Kamakura pp.522 (In Japanese)
- Tropicos® – Missouri Botanical Garden, [www.tropicos.org](http://www.tropicos.org)
- Ulloa Ulloa, C. & Neill, D.A., 2005. *Cinco años de adiciones a la flora del Ecuador: 1999-2004*. Editorial Universidad Técnica Particular de Loja, Loja. 75 Pp.
- Ulloa Ulloa, C., J. L. Zarucchi & León, B., 2004. *Diez años de adiciones a la flora del Perú: 1993-2003*. Arnaldoa Edición Especial Nov. 2004: 1-242.
- Veldkamp, JeF., 2006. Draft Compositae for Flora Malesiana.
- Walsh, N.G. & Stajsic, V., 2007. *A Census of the Vascular Plants of Victoria, Eighth Edition*. Royal Botanic Gardens, Melbourne
- Welby, J., 2009. EPU, initially developed at Royal Botanic Gardens (Kew), United Kingdom.
- Wilton, A. & Richards, K., 2007. C-INT Checklist Integration Software. Landcare Research, New Zealand.







































Zuloaga, F., Morrone, O. & Belgrano, M. (eds), 2008. Catálogo de las plantas vasculares del cono sur (Argentina, southern Brazil, Chile, Paraguay y Uruguay). *Monographs in Systematic Botany from the Missouri Botanical Garden* 107. <http://www2.darwin.edu.ar/Proyectos/FloraArgentina/FA.asp>









**Figure 1** – Map of Data Provider coverage for Global Compositae Checklist, September 2010  
 Key to Colours provided in Table 1.



**Table 1.** Original Data Provider Sources by Country or Region for Global Compositae Checklist, September 2010  
Key to colours in Figure 1; \*online; ^received but not yet integrated; AU - Australia

Key	Country/Region	Data Provider/Source	No. Names
	Abu Dhabi Emirate	Brown & Sakkir 2004	36
	Afghanistan	Checklist of Afghan Plants <sup>^</sup>	687
	Africa	African Plants Database*	14862
	Australia	Naturalised Australian Asteraceae <sup>^</sup>	434
	Australian Capital Territory, AU	Lepschi et al. 2008*	182
	Bolivia	Hind 2009 <sup>^</sup>	1044
	Bonin Islands	Toyoda 2003, Enomoto 1992	40
	Caribbean	Caribbean Checklist	1181
	China	Flora of China Checklist	2630
	Colombia	Colombia Database	2078
	Cuba	Database of Cuban Endemic Plants	220
-	Dolomite, Spain	Mota et al. 2008	18
	Ecuador	Ulloa Ulloa & Neill 2005	73
	Ethiopia & Eritrea	Tadesse, Mesfin 2004 <sup>^</sup>	473
	Europe, Mediterranean	Euro+Med PlantBase*	18881
	Guiana Shield	Funk et al. 2007*	1043
	India	Draft Checklist <sup>^</sup>	1061
	Iran	Rechinger 1906 – 1998	1154
	Japan	Flora of Japan Database*	1704
	Korea	Flora of Korea	403
	Malesiana	Veldkamp 2006	2981
	Mesoamericana	Flora Mesoamericana*	6733
	Mexico	Castelo et al. 2005*	7920
	Mongolia	Gubanov 1996	595
	New South Wales, AU	New South Wales Flora Online*	1049
	New Zealand	New Zealand Plant Name Database*	2967
	Nigeria	Asteraceae in Nigeria <sup>^</sup>	315
	Northeastern Brazil	Hind & Miranda 2008 <sup>^</sup>	486
	Northern America	Preliminary Checklist of North American Compositae	10878
	Northern Territory, AU	Kerrigan & Albrecht 2007	284
	Oman	Flora of Oman <sup>^</sup>	119
	Pacific Islands	14 individual data sources	483
	Pakistan	Flora of Pakistan Compositae (partial) <sup>^</sup>	194
	Panama	Flora of Panama	400
-	Paramo	Luteyn 1999	1299
	Peru	Ulloa Ulloa et al. 2004	359
	Queensland, AU	Bostock & Holland 2007	512
	Russia	Czerepanov 1995	7248
-	Siberia	Baikov 2005 <sup>^</sup>	600+
	South Africa	Plants of Southern Africa: an online checklist*	3384
	South Australia	Electronic Flora of South Australia*	1121



Key	Country/Region	Data Provider/Source	No. Names
	Southern Cone	Zuloaga et al. 2008*	8360
	Taiwan	Flora of Taiwan	342
	Tasmania, AU	Flora of Tasmania Online*	474
	Victoria, AU	Walsh & Stajsic 2007	715
	Vietnam	Le Kim Bien 2005	860
	Western Australia	Western Australia Census	1058
-	World	Hind & Jeffrey 2001^	
-	World	Govaerts World Compositae Checklist A-G	61035
-	World	International Plant Name Index*	159046
-	World	Kadereit & Jeffrey 2006	1620
-	World	Pieter Pelser Senecioneae Database	10020
-	World	Robinson Astereae^	9000+
-	World	Robinson Eupatorieae^	10650
-	World	Robinson Vernonieae^	7000
-	World	Tropicos*	74405