

Response to ECAT Request for Proposals 2005-2006

Pre-Proposal Draft

Title for proposed project: Global Working Checklist of Compositae

Managing Institution

Landcare Research

Address: Landcare Research, P O Box 69,
Lincoln 8152, New Zealand

Institutional Contact Person: Dr Ilse
Breitwieser

Telephone: +64 3 325 6701 extn 3796

Fax: +64 3 325 2418

Email:

breitwieseri@LandcareResearch.co.nz

Principal Investigators

Dr Ilse Breitwieser, Research Leader, Plant
Systematics; Director, Allan Herbarium

Dr Jerry Cooper, Research Leader,
Biodiversity Informatics; GBIF Node
Manager – New Zealand

Address: Landcare Research, P O Box 69,
Lincoln 8152, New Zealand

Telephone: +64 3 325 6700

Fax: +64 3 325 2418

Email:

breitwieseri@LandcareResearch.co.nz

cooperj@LandcareResearch.co.nz

Partner Institutions

Partner Institution: The International
Compositae Alliance Network

Website: www.compositae.org

Partner Institution: South African National
Biodiversity Institute (SANBI)

Institutional Contact Person: Dr Marinda
Koekemoer, Deputy Director, Curator of the
National Herbarium

Address: 2 Cussonia Avenue, Brummeria,
Pretoria, Private Bag X101, Pretoria, 0001,
South Africa

Telephone: +27 12 843 5000

Fax: +27 12 804 3211

Email: koekemoer@sanbi.org

Partner Institution: Smithsonian Institution

Institutional Contact Person: Dr Vicki A.
Funk, Senior Scientist, Curator US National
Herbarium

Address: NMNH, MRC 166 P.O. Box 37012,
Washington DC 20013-7012, United States of
America

Telephone: +1 202 633 0950

Fax: +1 202 786 2563

Email: funkv@si.edu

Partner Institution: Royal Botanic Gardens,
Kew

Institutional Contact Person: Dr Eimear
NicLughadha, Science Coordinator,
Herbarium

Address: Richmond, Surrey, TW9 3AB, United
Kingdom

Telephone: +44 20 8332 5000 (main switch)

Fax: +44 20 8332 5197

Email: e.lughadha@rbgkew.org.uk

Partner Institution: Missouri Botanical
Garden

Institutional Contact Person: Dr Peter Raven,
Director

Address: P.O. Box 199, St. Louis MO, United
States of America

Telephone: +1 314 577 5110

Fax: +1 314 577 0830

Email: peter.raven@mobot.org;

praven@nas.edu

Partner Institution: University of Tokyo

Institutional Contact Person: Dr Motomi Ito,
Associate Professor, Department of Systems
Sciences (Biology)

Address: Graduate School of Arts and
Sciences, Komaba, Meguro-ku, Tokyo 153-
8902, Japan

Telephone: +81 3 5454 6638

Fax: +81 3 5454 6638

Email: cmito@mail.ecc.u-tokyo.ac.jp

Partner Institution: Botanic Garden and
Botanical Museum Berlin-Dahlem

Institutional Contact Person: Prof. Walter Berendsohn, Director Dept. Biodiversity Informatics
Address: Freie Universität Berlin, Königin-Luise-Straße 6-8, D-14191 Berlin, Germany
Telephone: +49 30 838-50143
Fax: +49 30 841729-43
Email: w.berendsohn@bgbm.org

Partner Institution: Instituto de Botánica Darwinion
Institutional Contact Person: Dr. Fernando O. Zuloaga, Director
Address: Casilla de Correo 22, B1642HYD San Isidro, Buenos Aires, Argentina
Email: fzuloaga@darwin.edu.ar

Partner Institution: Australian National Herbarium, Centre for Plant Biodiversity Research, CSIRO

Institutional Contact Person: Prof. Judy West, Director
Address: GPO Box 1600, Canberra, ACT 2601, Australia
Telephone: +61 2 6246 5113
Email: judy.west@csiro.au

Project Summary

A Global Working Checklist of the Compositae addresses the largest identified gap in the vascular plant checklist progress. Most of the centres of high Compositae diversity are covered by existing data sources. This project will bring these together for the estimated 25,000 species of Compositae. By contributing the Compositae checklist 10% of the world flora will be completed, offering an important resource for biodiversity, biosecurity and horticultural end-users. The International Compositae Alliance (TICA) offers the network necessary to achieve this and is willing to support the generation of this checklist and function in sustaining and maintaining its data. TICA brings together experts from more than 60 countries from both the developed and developing world, making this a truly global collaborative effort. The systems being considered for this project are capable of providing access to all the data specified by the GBIF ECAT programme, and all data will be made readily available through any data provider mechanism supplied by GBIF. By the end of this project the baseline for the Global Checklist of the Compositae will be in place and 70% of the species, although not all names checked by experts, should be accessible via the Internet and linked to GBIF directly.

Project Description and Importance

The outcome of this two year project will be a Working Checklist of the Compositae on a worldwide scale accessible via the Internet and linked to GBIF. It will mobilize and integrate the content of many existing regional species databases, in addition to other digital sources. Commencement of work on a global checklist of this family is imperative in order to fulfill the Global Strategy for Plant Conservation (GSPC) Target 1 by 2010, “A widely accessibly working list of names of known plant species, as a step towards a complete world flora”. This project offers progress towards a final global synonymised checklist of the largest angiosperm family, directly addressing the largest identified gap in the vascular plant checklist progress and meeting the priorities stated in section 1.4.2.2 of the Request for Proposals. It is recognized that a complete checklist for all Compositae is beyond the resources provided by this project, but the output will form an essential and substantial resource towards that goal.

This project directly addresses Target 1 of the GSPC which underpins the other 15 Targets. This project fits with the goals of GBIF, ECAT, Species 2000/ITIS COL, IOPI and Species Plantarum. Implementation of several Articles of the Convention on Biological Diversity (CBD) relies on access to correct name information. Additionally this project works towards implementing the Global Taxonomy Initiative (GTI).

As stated in the Request for Proposals, GBIF is in a unique position to initiate international collaborative projects such as this collaboration to compile a Working Checklist of the Compositae. Without GBIF funding to act as a catalyst for the co-ordination and synthesis of existing data, the likelihood of compiling a working list of Compositae species and thus achieving the GSPC targets by 2010 is negligible. The funds sought from GBIF will cover the salary of a project co-coordinator for 2 years and will contribute to the development of the TICA hosted Compositae portal and the data network making the global checklist data available as it becomes available.

Most of the centres of high Compositae diversity are covered by existing databases, published checklists or floras (- the largest missing areas are parts of Asia, Brazil, and Colombia). These range from regional to national and offer a rich source of data for populating a global checklist. The aims of this project are to: 1) facilitate the collation of these distributed digital resources, 2) when possible resolve and complete nomenclatural content (including homotypic synonyms), 3) examine, report and resolve, as much as possible, differences in taxon concepts, 4) feed relevant information back to data providers, 5) provide a framework for facilitating information flow among data contributors. These aims will be achieved by creating an initial data warehouse with data integrity/quality managed by a sophisticated data management suite that can be used over the Internet. The project will result in a working checklist of the estimated 25,000 species of Compositae. We will create a web portal to deliver these, and ancillary data to end users, at the earliest opportunity. The work will result in a period of consolidation of data content and quality throughout our network of data providers and from input from the TICA community. There will be discussion on appropriate workflows and protocols with the aligned IPlant project. We will encourage and assist regional data providers to become independent ECAT data providers. Consequently, we propose at that stage that in the longer term the data will be managed consistently by the range of data providers best equipped to manage their regional content – thus ensuring longevity of the content. The ECAT network will itself then provide the taxonomic data for the Compositae portal which will concentrate on providing other data services relevant to the end-user community. The central warehouse would be available to TICA at the completion of the project to help manage the content of the checklist if desired. However, an alternative model

for the long-term sustainability needs to be discussed and evaluated as part of this project. In this scenario the family will be taxonomically delimited and managed by a number of global expert networks agreeing on a common data provision mechanism. This project is therefore about mobilizing existing data, facilitating cooperation, capacity building, filling digital gaps, improving data quality, and identifying sustainable data management models for ECAT. The 'clearing house' operating through a 'people-network', will enable contributors to mobilize their data by providing both the necessary technical and cooperative structures. Differing levels of information are present in different datasets and some of the discrepancies will be dealt with during this project. However the huge nature of this undertaking will mean that many names will remain to be resolved at the completion of this project hence the working nature of the checklist that will be produced. Despite this, by the end of this project the checklist will be complete for around 70% of the Compositae species and all available names will be included in the database. An outcome of this project will be identifying remaining gaps in the coverage of the checklist and outlining a strategy to complete those areas. The International Compositae Alliance will assist with fundraising for future work.

In terms of this project's value to science and society, the preliminary database of the Compositae within a global index of organism names is integral to the success of larger programmes with regard to vascular plants. When the Compositae checklist is finished it will contribute about 10% of the world flora. The product offers considerable scientific contribution to the GSPC (see attached letter of support). There are a host of questions about biodiversity that could be answered using such a list and there are also issues of biosecurity. The Compositae have contributed many invasive aliens to the floras of the world and the existence of a Checklist of the Compositae would be an essential resource for biosecurity end-users (see attached letter of support). In addition, this family probably contributes the largest number of species of cultivated plants for gardens. Having such a list would facilitate communication in the commercial horticulture community. It would also target plants for consideration for cultivation.

The International Compositae Alliance

This project would not be possible without the existence of The International Compositae Alliance (TICA) which brings together experts from around the world on this large and important group. TICA has approximately 200 people on its email list from about 35 countries. The *Compositae Newsletter* is mailed to around 600 addresses (about 100 are libraries). The newsletter coverage includes an additional 27 countries. In total more than 60 countries from both the developed and developing world are represented by the network, making this a truly global collaborative effort. The mailing list is continuing to expand and we will be working to increase the number of email contacts available for input to this project. When the email list is compared to the printed list we find that many of those from developing nations have joined the email list as it provides easy contact with the community.

TICA offers the network necessary to attempt a checklist of a family of this size and is willing to support the generation of this checklist and function in sustaining and maintaining its data. Barriers to the completion of checklists of large taxonomic groups such as the Compositae exist due to the sheer size of the task, time necessary to complete it and lack of funding. Many of the barriers identified in the June 2004 Workshop "Towards a Working List of Known Plant Species: Coverage, Gaps and Metadata" are overcome by the TICA network. The taxonomic knowledge is available, review and validation can be organized through TICA as can synthesis and co-ordination. TICA allows access to a global pool of expertise on the group and is

therefore integral to the success of this project. However, such a project as proposed here cannot be started without Seed Money.

Although TICA has been in existence and growing for several years, the existing website requires further development. Having the Compositae portal operate out of TICA will increase the usage of the website and its development. TICA members maintain their own lists and will send updates to the relevant data providers as they become available. Once the Compositae portal is established, TICA will help seeking other funding to maintain the content and to continue to build on a global resource for information on Compositae extending beyond the initial taxonomic resource of the checklist.

TICA has regular international meetings providing an existing forum for interaction of collaborators in this project. The next full meeting is scheduled for Barcelona in July of 2006. During the meeting TICA will hold a workshop on the checklist project.

Data Integration, Taxonomic Data Management, and Data Provision

An existing, related GBIF funded project “Implementation of an extensible register of the European and Mediterranean Compositae” has used the IOPI-based Berlin Model software for taxonomic data management. In addition Landcare Research has developed a sophisticated taxonomic data management system that is currently used to manage 80,000 taxonomic names and supporting literature relating to the New Zealand national collections of many organism groups. Missouri Botanical Garden’s TROPICOS database contains a significant taxonomic contribution to this effort (874,000 vascular plant names) including checklists for many parts of the world. IPNI, Australian Plant Name Index (APNI) and TROPICOS have information on nomenclatural content although we will be careful to resolve nomenclatural differences. As part of the project, discussions will take place among these institutions, and other proposed data providers, to decide on an appropriate framework for data integration and managing the resulting taxonomic content. One potential framework is to use this project to explore the possibility of data sharing between separate systems. Consequently, this proposal contains a significant technical component relating to data interoperability, data integration, data cleansing, resolution of overlapping taxon concepts, data repatriation, and a subsequent ECAT data provider network. A long-term strategy of maintenance of the checklist will also be established as part of the project. We will welcome feedback from reviewers on alternative frameworks with reduced implementation costs. A detailed budget is provided to enable reviewers to understand the cost allocations.

Existing systems are capable of providing access to all the data specified by the GBIF ECAT programme, and all data will be made available through data exchange using a schema compatible to TCS of TDWG.

Data Sources & Approach

The recently completed list of accepted genera for the Compositae (Kubitzki, ed. *Vascular Plant Families and Genera* in press, expected in 2006) will be used as a basis for the project. The larger regional datasets will be the first port of call to minimise duplication of entries and cleaning of the data to work towards a global list with a consensus synonymy. These sources of Compositae names will include those listed below and complement the existing register of the European and Mediterranean Compositae previously funded by GBIF. Selected letters of support are attached.

Global: Species2000/ITIS COL consortium, GBIF, IOPI, *World Checklist of Seed Plants* (A to G), w3TROPICOS, IPNI. **The New World:** Flora North America, Mexico (CONABIO), Flora Mesoamericana, checklist for Northeastern South America plus other national level data such as a recently published checklist of Panama, the Missouri Botanical Garden checklists (with partners) for Ecuador, Peru and Bolivia, the new Checklist of the plants of Venezuela and a checklist for the "southern cone" of South America including Argentina, Chile, Paraguay, Uruguay. **Europe and the Mediterranean:** Flora Europaea (the ESFEDS database), the Flora of Macaronesia database, and data from a number of European and Mediterranean country Floras have been integrated into the original Euro+Med database. The ongoing GBIF-supported database project on European and Mediterranean Compositae builds on this, updates the information, adds full coverage of the S and E Mediterranean countries (from the unpublished Med-Checklist data) plus Caucasia, and introduces a coherent and modern classification at all levels. **Eurasia:** Checklist for Russia and adjacent countries and the Flora of the USSR. **Africa:** African Plant Checklist (sub-Saharan Africa), data for all the islands off the coast of Africa. **Pacific/Asia:** Flora of China, draft checklist of China, Flora of Taiwan, Flora of Japan, Checklist for Flora of Korea, Flora of Thailand, Flora Malesiana, APNI, Species 2000 New Zealand, and The Pacific islands checklist.

In a group this large there is the potential for differences between species concepts and name usage between datasets. Expertise in the TICA membership, with various checklists from around the world, and the existing experience of the Euro+Med Compositae project, will be used to develop a consensus taxonomic opinion, however the proposed data management tools will also allow multiple taxonomic concepts to be captured.

At the end of this project it will be possible to use the checklist data warehouse to report on how many Compositae names and concepts are present in which provider databases, what data coverage is provided, what percentage need additional information added, from which regions, and for which taxa. This will encourage national funding to improve data quality and to fill gaps.

Milestones

Month three (April 2006)

Agreed on coordination and collaboration with Berlin Euro+Med Compositae project, MBG TROPICOS and others, on approaches to data management and sharing. Established data contributor network, negotiated data access, and established prioritization.

Month six (July 2006)

Data exchanged from contributing regional databases and integrated into data warehouse. Where possible, distribution information will be captured but reconciling of data will be dealt with in a second phase of the project. A workshop at the TICA meeting in Barcelona will focus on the way forward.

Month 9 (October 2006)

Coordinator using data management tools – work started on quality/completeness/synonymy.

Month 12 (January 2007)

Data exchanged from priority national databases and added. Initial working list composed from collaborating datasets and accessible via Compositae web portal.

Year 2 (January 2008)

TICA lists used to notify experts in various taxonomic groups of the availability of the checklist and asked for feedback. Investigation of extension of data management to distributed network of experts.

Prioritized information found only in word processing files will be incorporated.

ECAT data provider operational on checklist warehouse.

A detailed list of missing coverage will be compiled and an action plan prepared.

Updated checklist information made available via data exchange standard to originating data providers.

Long-term strategy of maintenance established.

It is difficult to estimate an accurate percentage of entries that will be editorially checked and complete by the end of two years. For initiating the checklist there are the extensive database resources listed above. For example, within the Compositae, TROPICOS has 73,000 names, Berlin (incorporating Euro+Med) has 23,175 names representing 8,523 accepted names and 1,332 misapplied names, ITIS currently provides data on 9,500 scientific names and 4,500 common names to COL/Species2000 representing 4,650 accepted names (but ITIS provides no publication data for these names). Within New Zealand's database there are 2,300 names with 1,000 editorially checked, representing around 500 taxa. An online review also revealed that many national parks, even in small countries that have few resources, have a preliminary species list available. It seems reasonable that if we negotiate access to all that available data, we should be able to complete approximately 70% of the taxonomic data. It should be noted however, that even with the large sum requested, we are unlikely to be able to process complete data for both publication details and common names in multiple languages and alphabets (these data do not exist in many existing digital resources). We will need to seek guidance to establish prioritization of these data components according to available resources.

Although many projects have existed for the Compositae of regional or national scope this project stands apart in the global nature of the undertaking, drawing on the existing work but not duplicating it and identifying remaining gaps for future work. By the end of this project the baseline for the Global Checklist of Compositae will be in place and many regions will be covered completely.

Preliminary project budget

ECAT Seed Money

	NZ/US \$ 0.7	conversion rate (10/08/2005)
Year 2006	NZ\$	US\$
Breitwieser Ilse (PI – Systematics)		
Project Management	7400	5180
Compositae portal development	4625	3237.5
Wilton Aaron (Systematics, Informatics)		
Compositae portal development	1300	910
Cooper Jerry (PI – Informatics)		
Project Management	3300	2310
Compositae portal development	825	577.5
ECAT Data Provider installation	1650	1155
Data harvesting, integration & interop	24750	17325
Scott Karen (Web designer)		
Compositae portal development	1900	1330
Richards Kevin (Programmer)		
ECAT Data Provider installation	3900	2730
Data harvesting, integration & interop	6500	4550
Portal Maintenance	5200	3640
Labour Amounts Total	61350	42945
Operating		
Travel – International Budget	4000	2800
Supplies – Science Budget	1000	700
Subcontract – Christina Flann salary	75000	52500
Operating Total	82000	57400
Total 2006	143350	100345
Year 2007		
Breitwieser Ilse (PI – Systematics)		
Project Management	7600	5320
Compositae portal development	4750	3325
Wilton Aaron (Systematics, Informatics)		
Compositae portal development	1350	945
Cooper Jerry (PI – Informatics)		
Project Management	3400	2380
Compositae portal development	850	595
ECAT Data Provider installation	1700	1190
Data harvesting, integration & interop	25500	17850
Scott Karen (Web designer)		
Compositae portal development	2000	1400
Richards Kevin (Programmer)		
ECAT Data Provider installation	4050	2835
Data harvesting, integration & interop	6750	4725
Portal Maintenance	5400	3780
Labour Amounts Total	63350	44345

Compositae portal development

This will be the TICA hosted web portal providing community access to the integrated data warehouse. It also hosts the web-service mediated remote data editing interface and will host the ECAT TCS data provider wrapper (see below). It is intended in the longer term to find support for additional services operating on this portal, which will eventually derive the bulk of its data from the ECAT network.

ECAT Data Provider installation

This will involve setting up an installation of the proposed TCS wrapper to provide ECAT access to the warehouse. This is likely to be one of a number of preliminary providers testing this new GBIF capability.

Data harvesting, integration & interoperability

This component involves: 1) agreeing and implementing data exchange mechanisms between the contributing partners, 2) any subsequent necessary transformation into TCS, 3) import of TCS data into at least the NZ-based warehouse and the Berlin & MOBOT data structures, 4) subsequent data driven integration/merge/purge of disparate data representing multiple overlapping nominal and taxonomic concepts of various completeness – including structured literature elements (not structured within TCS), 5) implementation of subsequent repatriation of edited data to originating data providers in agreed formats, 6) assistance for data providers in becoming independent ECAT providers (as resources allow).

Subcontract - Christina Flann salary

Christina Flann, who recently received her PhD, will be engaged as a self-employed subcontractor. Christina's role is to coordinate between the various data providers. The main function of her role will be to use taxonomic editing tools to complete fields, make appropriate synonymy linkages associated with references (consensus concepts), identify incorrect data, identify gaps and fill in as many as possible, and locate and enter as many common names as possible. This activity will be prioritized and worked through until resources are exhausted.

International Budget

This is to facilitate collaboration between the technical staff at Landcare Research, the project coordinator/editor, and the key data providers.

Supplies

To provide minimal funds for access to copies/loan of necessary literature, data media costs.

Portal Maintenance

The portal will be based on Landcare Research's existing portal investment. This receives on-going development and updates which will require rolling out to the TICA portal. Landcare Research's DBAs will also maintain the data warehouse and delivery platform, and programmers will maintain the editorial software suite.

No contingency is included in this budget. Landcare Research is a non-profit organization without core funding, and salary costs/overheads are expected to be covered.