*i*Plants:
The World's Plants Online

Extension Grant: Final Report to Moore

Version 1.8

Author(s): Alan Paton and Bob Allkin

Location: ***i*Plants** document library: **www.iplants.intranets.com**

Directory: Group documents / Outputs Extension Grant

Filename: Extension Grant - Final Report to Moore.doc

Approval for Public Release if appropriate: Formally approved for release by Steering Committee

Date: 15th May 2006

Revision History

Date	Who	What	Version
13th March 2006	AP	First draft released to all iPlants committee	1.2
18th April 2006	BA	Second draft released to iPlants collaborators	1.3
28th April 2006	AP	Third draft released to iPlants collaborators	1.5
15th May 2006	BA	Final draft release to PI prior to publication	1.8

TABLE OF CONTENTS

1. Introduction

This report covers the work undertaken between mid June 2005 and April 2006 by the iPlants project partners (Royal Botanic Gardens, Kew, The New York Botanical Garden, and Missouri Botanical Garden) under Grant 341.01 made by the Gordon and Betty Moore Foundation. This extends the pilot work completed under Grant #341 to produce an internet accessible index of all the world's plant species together with their global distribution.

Primarily, *iPlants* is responding to the pressing need for a consistent, comprehensive and authoritative list of all scientific names for plants. It will make available electronically a single stable list of "accepted names" each linked to its alternative synonyms. Such a list is essential to enable those other than botanists to access the significant and valuable body of information that exists about plants which is highly fragmented and dispersed. Providing such a name index as a web service would significantly impact on information services about plants from many other providers operating in numerous domains across the globe.

The Global Strategy for Plant Conservation (GSPC) underlined the need for such a list by identifying this as its first target, upon which others depend. The GSPC was approved by 188 governments and comprises 16 targets for completion by 2010 which address the major challenges facing the conservation of plant diversity.

Achievement of this ambitious goal requires new ways of working. The *iPlants* pilot project represents a major advance in collaboration in the botanical community: developing methods and synergies to maximise the impact of data on key concerns in conservation and science. Our approach has generated excitement and received endorsement by key players currently providing information services and from conservation practitioners working on the ground.

The objectives of this extension grant were:
1) to complete checklists begun in the first phase of iPlants;
2) to explore and document the high level requirement for providing and maintaining checklist data; and
3) to explore and document the impact on GenBank and Consortium for the Barcode of Life end users of implementing an authoritative checklist of known plant species.

This report begins with an overview of the progress and impact of the iPlants project during the extension grant period, illustrating how work done on global checklists has been taken up by others. It then outlines the achievements under each of the objectives of the grant and concludes with an outline of future work.

2. Impact of iPlants

2.1. Accelerating checklist production

2.1.1. Compilation

The experiences and documentation produced by the iPlants project have stimulated work on an authoritative list of known plant species.

In the previous phase of the project a gap analysis of checklist coverage was carried out in collaboration with Species 2000 and GBIF. iPlants partners have since worked with GBIF, both on GBIF science committees and through e-conferences, to prioritise GBIF seed-money grants. GBIF recognized the completion of an authoritative list of accepted plant names as a priority for funding. In response to an open call, GBIF awarded two grants to help fill two of the largest gaps in coverage of the authoritative plant list.

A grant of 200,000 US dollars was made to Landcare Research , New Zealand to co-ordinate a global checklist of the Compositae- by far the largest gap. Both Missouri Botanical Garden (MO) and Kew are partners in this project, which will draw on the experiences of the iPlants partners to date .

GBIF also made a grant of 70000 $ to the University of Munich for co-ordinating an authoritative global list of the Melastomataceae. Both the New York Botanical Garden (NYBG) and MO are playing an active part in the compilation of this checklist and Kew has supplied baseline data to the project. MO has provided all available names from its Tropicos database for this project and Carmen Ulloa Ulloa, a member of the iPlants group, is one of the scientific collaborators.

The checklist compilation tool developed during the previous phase of the project is now being used by others. Irina Belyaeva, Russian Academy of Sciences, Ekaterinburg, is using the system to produce a checklist of Salicaceae and data for the Convolvulaceae is being standardised at Kew for use by George Staples, Bishop Museum, Hawaii to produce a checklist of that family. We have recently received an enquiry from the Institute of Botany, Chinese Academy of Sciences, Beijing expressing interest in providing a checklist for Ranunculaceae.

Work at Kew on checklist compilation has continued with now 106 plant families available via the internet (www.kew.org/wcsp/incfamilies.do). This accounts for around one third of Flowering plants.

As a result of these activities 11 more botanical institutions and organisations are involved in global checklist compilation and many more individuals and institutions will be involved in the review of these works. In all 114 collaborators in 20 countries have been involved in the checklists available on the Kew site.

2.1.2. Standardisation

One of the lessons learned from the pilot phase of iPlants was that having a complete standardised and de-duplicated set of plant name records would greatly facilitate checklist compilation. The initial phase of the compilation procedures was a costly deduplicating and standardisation of records from IPNI, New York's name list and TROPICOS. Reducing the need for this stage will greatly increase the speed of production of checklists not just in iPlants but in the entire botanical community.

A data standardisation specialist has been employed for one year, beginning at the end of January 2006. The work will concentrate on:

- Standardising name spellings
- Standardising publication abbreviations
- Standardising authors
- Extracting the year of publication from the collation information and adding publication dates where missing or wrong
- Standardising the collation format.

Doing this work will enable faster deduplication of records in IPNI and also make the work of generating a checklist easier. First, there will be fewer duplicates to consider, and homonyms would stand out more easily. Secondly, records can be reliably linked to other data resources (particularly TROPICOS) containing concept data and synonymies. More reliable links between IPNI, TROPICOS and the New York Names list will enable those datasets to cross check and standardise themselves, and will prevent future duplication of effort as corrections to one dataset can be fed through into the others. Finally, future checklists can be more easily checked for omissions either in IPNI itself or in the checklist, allowing a virtuous cycle of reinforced standardisation to develop.

2.1.3. Measuring data quality

We need to be able to measure, objectively and automatically, the quality of plant lists used by the providers of plant information services that are NOT using an authoritative checklist and to compare them with the iPlants list.

Our thesis is that construction of an authoritative global plant checklist will result in demonstrable improvements in the quality (reliability) and coverage (comprehensiveness). To demonstrate such improvement it is necessary to establish metrics to quantify "quality" and "coverage" and tools capable of calculating these metrics on plant name lists.

iPlants has developed a collaborative relationship with the Computer Science Dept at Cardiff University (UK) who have broad research interests in bioinformatics and a particular research programme aimed at data quality issues. As part of the latter Dr. Richard White and Dr. Andrew Jones have developed a software tool "Litchi2" used by a number of projects such as Species 2000. A more recent version of this software permits lists of names and synonyms to be processed to search for "associations" between names which are similar (in textual form) or related in some way (such as by synonymy). *Litchi2* tests these associations against a number of "rules" and lists the "errors" detected in a single list and "conflicts" (differences) detected between two lists. A second set of rules is then used to infer concept relationships between names which are output as "cross-maps". *Litchi2* can be used over the web.

Working together with Drs. White and Jones and their team, we have designed a research programme evaluating the practical need to detect and measure inconsistencies within lists of names and overlaps between lists (possibly drawn up by non-systematists for specific goals). This will profile the classes of error, omission and ambiguity that can occur within plant name lists and measure the relative frequency of these error types within name lists of different origin and construction to permit us to establish procedural recommendations for those constructing checklists.

A further objective is to extend the functionality of *Litchi2* to meet the needs of *iPlants* and other projects. Planned changes include providing for the 'rule-base' to be more readily modified and controlled; including alternative formal representations of the embedded rule base to make this more intelligible to users; and to extend the rulebase in a variety of ways. An exciting product will be a set of "business" rules, expressed in a logical form and

intelligible to systematists, which encapsulate the nomenclatural quality controls that they would seek to employ when evaluating and building a checklist.

To date *iPlants* has submitted one grant proposal to NERC in the UK, in collaboration with Cardiff University, to fund this work. This was unsuccessful, but another proposal is planned.

2.2. Broadening participation: building links with users

Many organisations from the public and commercial sectors, and across a wide variety of disciplines, offer electronic information services which relate to plants. These agencies share a common need for an authoritative, synonymised and comprehensive checklist of plants in order to deliver reliable information and to ensure that their users have access to all available knowledge.

iPlants has increasingly sought to enter into dialogue with representative users of a global plant list and to document the case for providing access to such a checklist. A grant proposal prepared for the Natural Environment Research Council (UK) aimed to:
1) calculate the cost savings to individual organisations e.g. delivering plant information services
2) develop the means to measure the socioeconomic benefits arising from improvements in these services derived from the *iPlants* checklist.
3) produce business and technical designs for the means by which an electronic checklist might be made available
4) define a detailed functional specification for a machine interface to the checklist.

To this end we have made contact with and entered into negotiation with potential partners in a user network. We continue to seek to expand this network of partners such as a recent presentation at the Royal Pharmaceutical Society of Great Britain hosting an international conference on Pharmacovigilance of herbal medicines which led to subsequent proposals for collaboration from the European Medicines Evaluation Agency – the European regulatory body for pharmacovigilance based in London and the UK Government's own regulatory body MHRA (Medicines and Healthcare products Regulatory Agency)..

Partners in discussion include:

- European Medicines Evaluation Agency (London, European Union)
- Global Biodiversity Information Facility (Copenhagen, Denmark)
- IUCN Red List of Threatened Species Programme (Gland, Switzerland)
- National Centre for Biotechnology Information – GenBank (New York, USA)
- MHRA (Medicines and Healthcare Products Regulatory Authority) London, UK)
- UNEP World Conservation Monitoring Centre (Cambridge, UK)
- World Agroforestry Centre (Nairobi, Kenya)
- World Health Organisation –Monitoring Centre (Uppsala, Sweden)

3. Output 1- Compilation and review of Checklists

3.1. Compilation of Checklists begun during the first phase of iPlants

A checklist of the Lecythidaceae (300 accepted species, 1200 names) was compiled by Melissa Tulig and reviewed by Scott Mori, both of the New York Botanic Garden (NYBG).

A checklist of the Bignoniaceae (900 Species 4087 names) was compiled by Carmen Ulloa Ulloa, Missouri Botanical Garden (MO) and reviewed by Lúcia Lohmann (Universidade de São Paulo, Brazil).

A checklist of Iridaceae (1500 species, 8000 names) is being compiled by Christine Barker at Kew. This checklist is c 80% complete. The delay in completion of the Iridaceae list was due to an underestimation of the size and complexity. The complexity arose from compilation of the central Asian species, a major centre of diversity in Iridaceae for which access to reference sources and resolution of inconsistencies in the literature proved particularly difficult. Lessons from this exercise include 1) new compilers, unfamiliar with the family in question, take longer to compile lists and 2) there are significant benefits to be derived from compilers contacting potential reviewers early for their input. There is a working version of the checklist of Iridaceae on-line, and Kew will ensure its completion in the near future.

The Checklists for Bignoniaceae, Lecythidaceae, Madagascan Endemic Families, Schhegeliaceae and the partially compiled Iridaceae are all now available via the prototype iPlants website- www.iplants.org.

3.2. Review of Checklists

Various checklists were reviewed during the grant period using a variety of review processes. The Lecythidaceae was reviewed by a member of staff of the same institution as the compiler (NYBG), while the Bignoniaceae was reviewed at MO by a visiting staff member from another institution (Universidade de São Paulo). The review of the Bignoniaceae involved bringing the reviewer to MO, with expenses covered, to conduct the whole review process during a fixed time frame. The reviewer used ca. 25 days, an average of 8+ hours a day, to go through all names, i.e. 163 names per day.

As the Iridaceae checklist is still under compilation, Kew's experiences in reviewing the large families Myrtaceae (3500 species) and Rubiaceae (13000 species) were used to help refine and document the requirements and lessons learned from the review process. Both the Myrtaceae and Rubiaceae were compiled in a similar way to the Lecythidaceae and Bignoniaceae. However, the review of the Myrtaceae involved holding a two week workshop at Kew, attended by 13 delegates from around the world, followed by a six month review period when comments were fed back to Kew from the reviewers. The review of the Rubiaceae involved employing a staff member for one year, using a GBIF grant. The staff member co-ordinated review comments from a wide circle of reviewers and added distribution and taxonomic information using the herbarium collection and library.

Additional information was also gathered from Kew staff members involved in the production and review of an authoritative list of the plants of North East Brazil. A workshop was held at Kew in late February 2006 for the co-ordinators of the various review processes and the issues surrounding the review processes and lessons learned were documented.

The main findings can be summarized under the following headings: Accreditation, Composition of Review Panel, Value of Review, Guidance to Reviewers, and Sustainability and Resources.

3.2.1. Accreditation

Reviewers are generally happy to assist with the review process. However, it is important that their work is acknowledged. It is desirable to be able to accredit reviewers for their contribution to individual records or any subset of records. A barrier to participation is that contributions to checklists are not considered highly by the broad academic community whose normal measures of value include such devices as citation indices. Enabling accreditation for individual records and subsets of checklist data may help reviewers demonstrate the impact of their work, particularly where checklists are linked to other data sources.

3.2.2. Composition of Review Panel

Experience suggests that multiple reviewers give a more thorough review as single reviewers may not feel comfortable commenting outside the area of their immediate expertise. This is especially true of larger families where reviewers concentrate on Old or New World taxa. Disagreements between reviewers are few and can be resolved through discussion, for example of the 5,800 accepted species of Myrtaceae the reviewers only disagreed on the circumscription of ten species. If reviewers cannot be found to cover the whole scope of the checklist it should still be made available via the internet. Good compiled, but unreviewed, data is more useful than inaccessible data.

3.2.3. Value of Review

The review process led to an improvement in the data, particularly in terms of additional distributional data for accepted taxa and modifications to the taxonomy. Review has a particularly high impact in areas where existing baseline data is incomplete or highly fragmented. These problems are exemplified by taxa occurring in Brazil and Mexico. The draft for review produced by compilers is of high quality, large changes requested by review being very few.

3.2.4. Guidance to Reviewers

Clearer guidance is required for those carrying out the review process. A draft guide for reviewers was produced, lesson's documented and suggestions for improvement made. Issues included the need to store reviewer and compiler comments so that investigations into specific data would not have to be redone at a later date; and the need to be able to produce reports of the data in a more flexible way.

3.2.5. Sustainability and Resources.

Review should be seen as a process of building a lasting relationship between a group of reviewers, compilers and editors who feel responsible for the checklist product. It is likely that review of any individual family checklist will be periodic rather than continuous and mechanisms need to exist for keeping and acknowledging review and other comments on the checklist so that they can be fed into a periodic review process. The review process should be an open system to facilitate as broad a range of input as possible. Statistics on the use of the checklist and website encourage participation (see accreditation above).

The greatest barriers to the review process are:

i) lack of time – because the poor recognition and rewards (from employers and research funding agencies) for this work to be prioritised over other activities;
ii) lack of access to herbaria, libraries and other taxonomic resources – particularly for those working in smaller institutions with fewer historical resources.

Small grants can be an effective way of leveraging reviewer input and workshops which have been found to be particularly effective for facilitating reviewer input. See also http://www.plants2010.org/targets/target1.html.

However, full time co-ordination, institutional support and commitment are required for long term maintenance. Kew is involved in an NERC funded project which will further explore the issue of facilitating collaborative taxonomic work to produce and maintain taxonomic revisions.

4. Output 2 - High level requirement for providing and maintaining checklist data

4.1. GenBank and the Consortium of the BarCode of Life (CBOL)

The US National Center for Biotechnology Information (NCBI) was established in 1988 in the USA as a national resource for molecular biology information. It creates public databases and disseminates biomedical information for the better understanding of molecular processes affecting human health and disease. NCBI hosts numerous information resources containing molecular, genomic, protein data and the associated literature (e.g. the PubMed database).Among the online resources offered by NCBI is "GenBank": a comprehensive database of publicly available DNA sequences for more than 165,000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. GenBank is accessible through NCBI's retrieval system, "Entrez", which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed.

The Consortium of the Barcode of Life (CBOL) is an international initiative devoted to developing DNA barcoding as an accurate and reliable tool for scientific research on the taxonomy of plant and animal species. CBOL has reached an agreement with GenBank to create an open archive of standardized DNA sequences derived from voucher specimens held in reference collections.

The NCBI site presents a disclaimer as to the coverage and reliability of its taxonomic index:

"The taxonomy database within NCBI is *not a primary source* for taxonomic or phylogenetic information" and the database does not follow a single taxonomic treatise. Consequently, the NCBI taxonomy database is not a phylogenetic or taxonomic authority and should not be cited as such."

This approach is probably the only option for NCBI given their business goal, expertise and resource levels. Nevertheless the self-avowed limitations of the taxonomy underlying their data resources have serious implications for users of their services – some of which may not at first be apparent.

Sally Hinchcliffe and Alan Paton (Kew) attended a meeting organized by the Database Working Group of the Consortium for the Bar Code of Life in April 2005 which discussed species names in Barcode Records in Genbank. CBOL concluded that the highest standard of sources for taxonomic names would: "have been reviewed for their adherence to taxonomic standards, objective and subjective synonymy, and reflect expert opinions. [These] Gold standard sources have the added value of being well-maintained; that is, as a name is revised in subsequent releases of an index, new information on the status of that name will be retrievable through the GenBank record" (Report of meeting). iPlants records aim to meet these criteria, and clearly a comprehensive index of plant names would greatly facilitate the work of CBOL.

This meeting was followed by a visit of Mark Jackson and Alan Paton to visit GenBank to discuss these issues.

At present a Genbank view on taxonomy is presented through the efforts of staff taxonomists, who base their decisions on checklist resources available directly to them or online. For the near future this is how GenBank will continue to operate. iPlants was welcomed as a useful additional 'behind-the-scenes' resource for their taxonomic editors. It should be noted that

GenBank is an archival database and its policy is to only make corrections to data records if and when the submitter requests a change.

The important data for GenBank provided by checklists are the taxon names, synonomies and references, with distributions not greatly valued. The ability to differentiate between homotypic and heterotypic synonyms is important, as the former would be automatically linked in the Genbank taxonomy though this is not always the case currently. It is likely that GenBank would want to acquire updates either of all names, or of names changed since the last date the list was uploaded.

The current requirement therefore is to send the iPlants list as files which they will incorporate into the GenBank system. A copy of Kew's Orchidaceae checklist was provided to GenBank and checklist information from the Kew Monocot Checklist made available through the linkout mechanism provided by GenBank. A similar link will be made from the iPlants checklists once the new names and combinations they contain are validly published by the authors. Kew's electronic plant information center (ePIC) and the International Plant Names Index (IPNI) have also been linked. Missouri Botanical Garden's Tropicos database has been linked to GenBank for some time. These links will be found as part of the GenBank taxonomy database and through their Entrez system. This means that users browsing the NCBI system may, if they choose, consult Kew or Missouri checklists seeking clarification of synonymy but will not need to do so and will mostly not be aware of the advantages.

These links require the creation of standard XML documentation which is then passed to GenBank for inclusion in their system via the use of automated queries on their database and the provision of html links to Kew, Missouri and iPlants web pages. The XML documents are then transferred manually to GenBank via FTP and are read automatically by their systems during their regular update period. This means that updates usually appear on the GenBank website within a day or two of their provision by the provider.

In conclusion, GenBank uses authoritative checklists within its own internal processes and integrates them into its services to end users. The more complete the authoritative index, the greater value it will have to GenBank and CBOL.

5. Output 3. Impact on Information Retrieval.

5.1. Introduction

In the previous phase of the project we outlined how the construction of a global plant checklist could result in demonstrable improvements in the quality (reliability) and coverage (comprehensiveness) of information services about plants that employed such a checklist. That study was largely theoretical focused on the type of issues which undermine the quality and comprehensiveness of information retrieval.
These potential errors resulting from lacking a comprehensive authoritative list of species include:

- failure to find all relevant information about a given species – where data has been recorded under two or more different synonyms;
- obtaining false information where a name exists as two or more homonyms (same genus and species epithet, but differing author and publication);
- obtaining false information where a name has been misapplied to another species.

In this extension project the aim was to gain a quantifiable measure of the impact an authoritative checklist would have on users of GenBank and the Consortium of Barcode of Life.

We initially hoped to examine usage logs from GenBank to see how an authoritative checklist might improve data retrieval given real queries to the system. Unfortunately the required level of detailed usage log was unavailable to us. However, Kew's World Checklist of Selected Plant Families (WCSP) covers 106 plant families and around one third of flowering plants. So we measured the frequency of errors and omissions found by users of GenBank for that third of the world's plants by comparing the WCSP with the name lists stored in NCBI for those 106 families.

5.2. Impact of Checklist

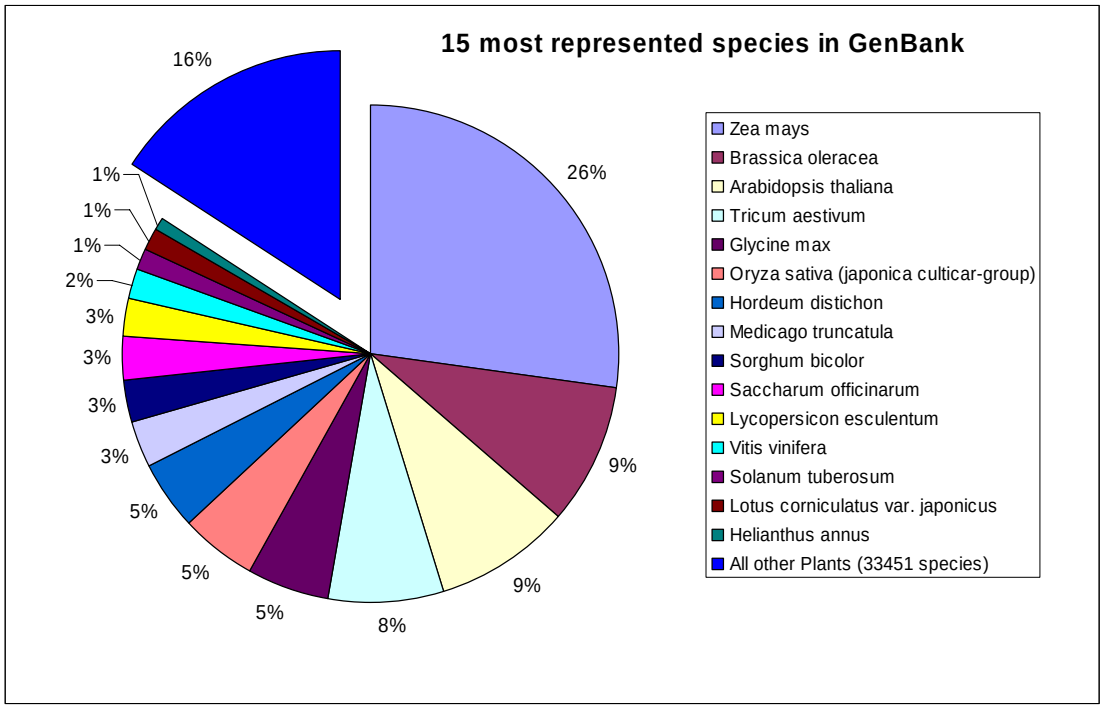
5.2.1. Coverage of names

An authoritative checklist (such as WCSP) would be expected to contain many more plant names than a list (such as NCBI) produced from contributed records. To illustrate this, the NCBI system contains less than 8% of the plant names covered in the WCSP.

This large discrepancy is problematic to those looking for information across the entire plant kingdom. GenBank contains data records for just over 33,000 plant species, less than 10% of the Plant Kingdom. Furthermore, the 15 most frequently represented plant species (which are all crop plants) together account for 84% of all plant sequences. Thus the majority of plant sequence records are associated with a very few well known plants which present limited nomenclatural confusion. 16% of the sequence records in GenBank refer to around 33,000 other plant species (where there will be more confusion).

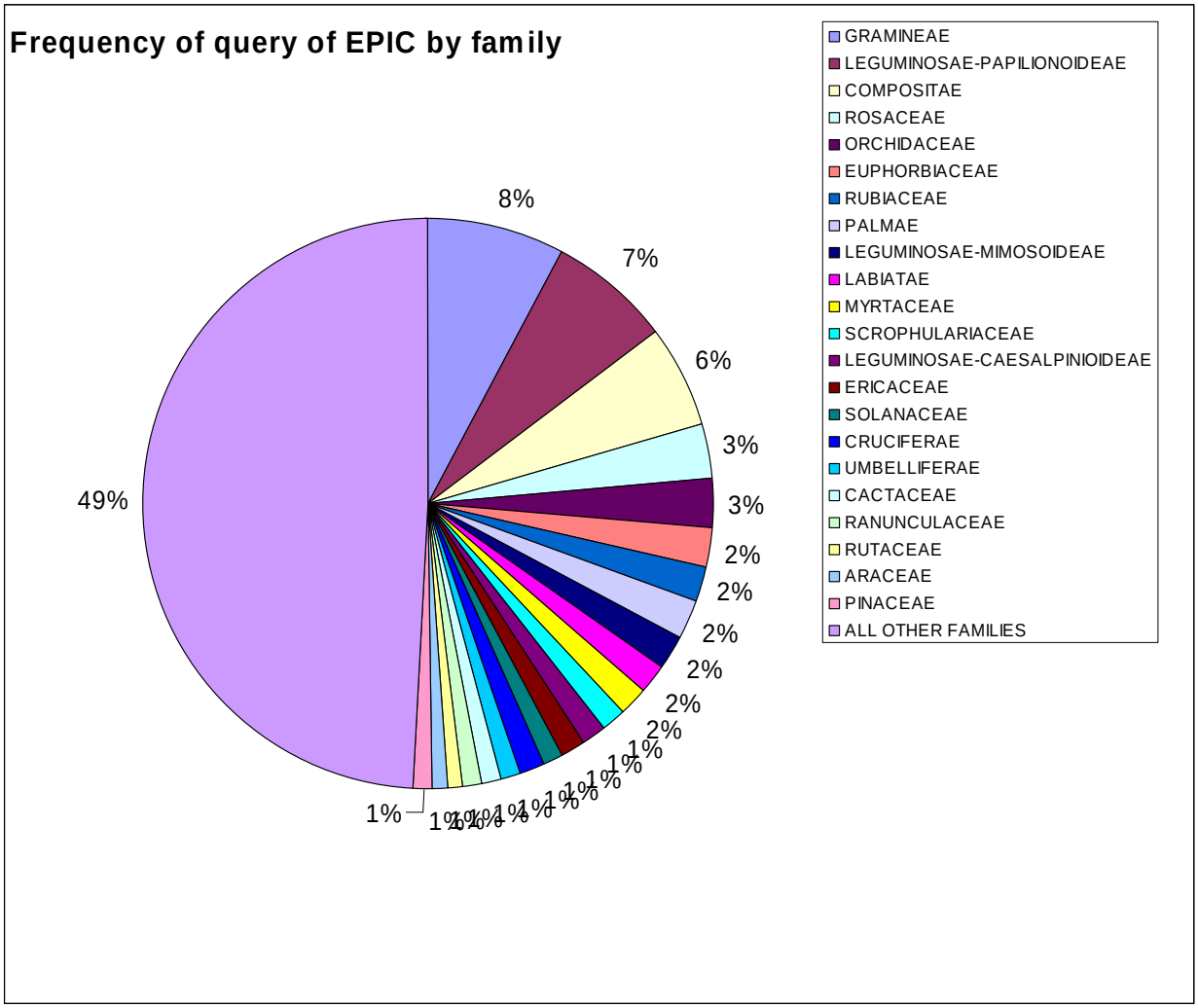
We were unable to access detailed usage logs on which species users are most interested in, because GenBank do not archive this level of information. Nevertheless it is a fair assumption that the current GenBank users are mostly interested in the small number of species with a large amount of sequence information.

Distribution of sequence information in GenBank.



However, as work progresses to establish a barcode of life it will be necessary for NCBI to hold at least one sequence for every plant species. Also as DNA sequencing gets cheaper, faster and more useful, so inevitably will more and more demands develop in other areas of science. In this context, the lack of authoritative name data for the great majority of plants will become a significant problem, seriously hampering the value and usability of data held in the NCBI system.

Interest in information on plant species is, at least in Kew's experience, very wide indeed. Of 671,000 queries made of EPIC last year, 51% were associated with 20 plant families (7,578 genera).The remaining 49% were queries relating to 490 different families of vascular plants and bryophytes (6921 genera). This demonstrates a far wider curiosity and need for information about all plant diversity which differs significantly from the research being undertaken and recorded by GenBank – where a few species have most data associated with them.



5.2.2. Quality of name data.

The NCBI database contains some names that have not been formally published and as such cannot be traced to a species description or type specimen (around 1.5 % of names in the sample of 106 families from the NCBI database fell into this category). These may represent misspellings or transliteration errors and cast doubt on the identity of the organism referred to. An authoritative checklist could be used as means to check names and prevent such errors occurring.

Another factor adversely affecting data quality in GenBank is homonymy. The same binomial may have been validly published by different authors at different times. Scientific names which share the same genus name and the same species epithet but were published by different authors in different publications are called "Homonyms". Homonyms occur when an author, unaware of the earlier use of that genus and species epithet combination, uses it to refer to another plant. These two alternative names (homonyms) will normally refer to different plants.

The presence of homonyms could lead users of GenBank to make significant errors and false conclusions by treating data records from more than one species as if it were all from the same species. 3% of the binomials in GenBank were found to have two or more homonyms in WCSP. If we assume that queries are distributed in the same ratio as data held, then 3 in every 100 queries would hit a homonym and potentially return unreliable data. Examination

of the WCSP data suggests that 3% is a typical rate of honomymy at species level. An authoritative checklist can resolve homomyms and could eliminate this cause of unreliable data return.

An authoritative checklist could greatly assist in improving the quality of name data in information sources such as GenBank. Use of such a checklist could prevent misspellings or incorrect transliterations being added to the database, eliminate the problems of homonymy and help ensure that data could be linked to an accepted name and thus allow all available information for that species to be made available.

5.2.3. Taxonomy and data retrieval.

Although GenBank does not aim to follow a single taxonomic treatment, there are significant implications for information retrieval in not linking accepted names with their synonyms. Following the consensus taxonomy offered by WCSP for the 106 plant families covered, 10.1 % of the names within GenBank have data stored under a synonym. In these cases queries using an accepted name will return null responses if the synonym was not linked to the accepted name. A total of 6.1% of the species within GenBank have data stored under more than one name (preventing users from finding all data associated with a particular plant) and 5.9 % of the species which have data under their accepted name also have data stored under one of their synonyms within GenBank. Queries using these names will only return a portion of the data available for the species as a whole if the accepted names and synonymies were not linked. A comprehensive checklist would minimise the number of species for which data is stored under unlinked names or under names which are not current or are incorrect.

6. Future work

6.1. Gap Analysis.

A revised gap analysis of global checklist coverage indicates that the proportion of a working list of known plant species available on the Internet is now 45% complete and will be close to 60 % complete by the end of 2007. It was only 15% complete at the beginning of the iPlants pilot phase project in 2004. The iPlants project partners will continue to facilitate and foster the necessary collaborations to complete this important goal.

6.2. Maintenance

Kew, MO and NY are excited about the significant advance that iPlants would offer. iPlants will be a fundamentally important, outward-looking service provided by botanists to the wider community of those who need access to information about plants. It will be fundamental to resolving many of the conservation and biodiversity problems faced by the modern world. The three institutions commit to underwriting the maintenance of the iPlants system and data in the long term using core resources and will also seek to attract additional funding and collaborators.

The resources necessary for long term maintenance of the iPlants system and data include:

1) The assimilation of 10,000 changes to name records every year as a result of ongoing scientific research and the description of new species. This is estimated to require two full time posts.
2) A database editor will be required to assimilate and respond to feedback and to assist with the administration of the review process.
3) Maintenance of the IT infrastructure. The system will need, at every site running the system, the following:
    a) Ongoing third-party maintenance costs for the infrastructure (hardware, operating systems, software, Internet connection, etc).
    b) Replacement cycle for the above components.
    c) Ongoing technical support and management for the infrastructure (support people to notice and diagnose problems, do backups, manage storage, servers and databases, and generally keep things working). It may be possible to provide some of this remotely.
    d) Continuing development of the system as new opportunities come along and new technologies emerge. If the system is successful then we should anticipate having to respond to new ways of merging it into the wider bioinformatics network and of presenting information just to keep it relevant and useful.

7. Conclusions

Much progress has been made in producing a working list of known plant species since the inception of the iPlants project. The work done in this extension grant has confirmed that checklist compilation can be done by the iPlants partnership leveraging their resources. We also believe that an authoritative checklist for all vascular plants and mosses can be completed within a reasonable time scale and at a reasonable cost.

This extension phase of the project demonstrates the enormous value that such a checklist could supply to other information providers in ensuring the quality of plant name data, comprehensiveness of coverage, and the linkage of disparate information. Work with GenBank, CBOL, IUCN, WHO and European Medical Evaluation Agency confirms the demand for checklist data and that much of this demand is for datasets and webservices to be built into other systems. On-line checklists will be a necessary long-term part of the online biodiversity infrastructure.

The participating institutions are committed to maintain the checklist data and more institutions have been involved recently in the production and review of authoritative lists. The complete index is likely to be around 60% complete by the end of 2007. However, resources for the compilation of the remaining portion of the index for all plants, estimated at 150,000 species, is required in order to complete the index in a timely manner. Funding is also required to develop an adequate service to the users of this information.

We would like to discuss the possiblity of funding this important endeavour with GBMF.