Year 2 Summary

Distributed Dynamic Diversity Databases for Life

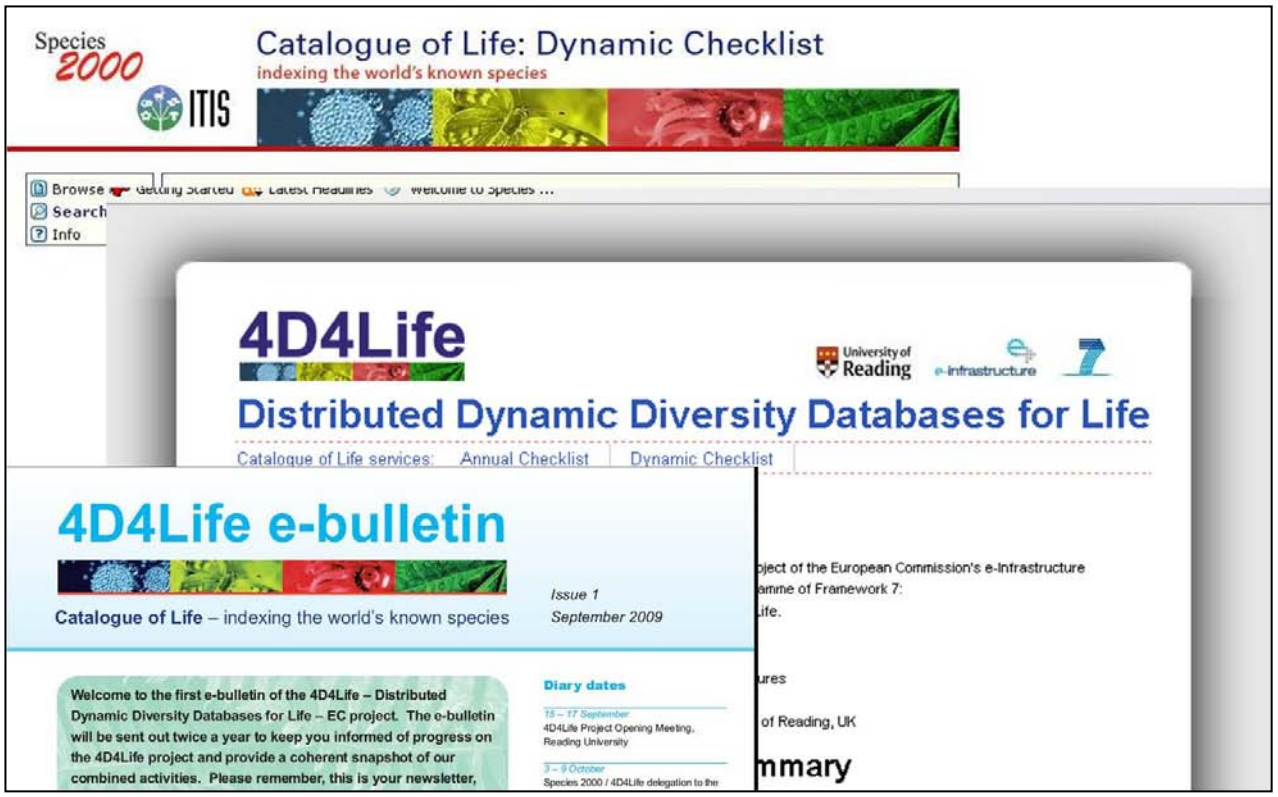
Summary

A coherent classification and species checklist of the world's plants, animals, fungi and microbes is fundamental for accessing and structuring information about biodiversity. As a scientific data infrastructure project 4D4Life builds on an existing *Catalogue of Life* programme that has started to provide the world with a unique taxonomic service. 4D4Life focuses on enhancing the electronic infrastructure within the ecosystem of databases that provide the Catalogue, and on creating a new array of services for users, as well extending the global reach of the programme, enhancing coverage of the Catalogue and seeking a sustainable future for the service.

During Year 2 the 4D4Life Project has enhanced the existing programme and continued the process of building a state-of-the-art *Catalogue of Life* e-infrastructure. This infrastructure is to provide a stable and sustainable electronic classification framework and species checklist for use as both taxonomic backbone and species index across global, regional and national biodiversity programmes and communities. A more robust and unified workflow, a service-oriented architecture, and a new array of electronic services are being tested in Year 2 and rolled out from Year 3.

Principal Objectives of the entire Project

- to synthesise a significantly improved global resource, the *Catalogue of Life.*
- to facilitate enhanced information exchange within the Species 2000 networks that feed into the *Catalogue*.
- to extend the reach of these networks to both inputs and public services through regional centres across the world.
- to disseminate this synthesised knowledge in a new array of modern web-services and products developed in partnership with identified user communities.
- to renew the distributed system with improved software and an enhanced service-based architecture.
- to implement a business model that will take this e-infrastructure to a sustainable future.



Year2 of the Project

Year 2 has seen the major peak of activities across all three compartments of the project – networking activities, provision of services, and research & development – although there remains a significant set of tasks for Year 3. Year 2 has seen a very large array of actions, involving all work packages, and with a high degree of parallel working – so there is much to include in this report. In general the year has worked extremely well, although there have certainly been moments when our management and communications abilities were fully stretched to keep everything co-ordinated. Many of these tasks were prepared or designed in Year 1, and needed to be completed in year 2 so that the outcomes could be brought together in Year 3 for the close of the project.

For nearly all of the Year 2 tasks, the main target during the year has been to report on and show the outcomes at the 'Prague Meeting' – the largest Project Meeting so far, that was organised through WP 4 in Prague on 28 March – 1 April. As reported below, this meeting in month 23, and attended by 57 people, was a major success for the sociology and coherence of the different networks of partners working in the project. It brought different sets of partners up to date on the huge progress in the other parallel streams, and enabled all to focus on the interlocking actions that will bring the project to a close in Year 3. The timing also worked well, as this was immediately prior to submitting several deliverables in Month 24, starting Year 3, and preparing this review for month 26.

Networking Activities (WP 2, 3, and 4)
Networking has gone ahead strongly with the Species 2000 array of 41 European global species database custodians, led by Thierry Bourgoin at MNHN, Paris in WP3. It is important to remember that this large network is the lifeblood of the Species 2000 Catalogue of Life Programme - its members provide the data from which the Catalogue is composed. The strong response, in several case carrying out more than 2 pilot-projects is a signal of real and growing confidence in the future of the Catalogue. There is also good news from Work package 4, which has broken out from the uncertainties of Year 1, and now both created a visionary plan for the Global Multi-Hub Network and made remarkable progress with the pilot-project in the hands of the CAS China partners – working jointly between Jiri Kvacek at Narodni, Prague and the Task Group installed last year. WP 2 led by Sara Oldfield at BGCI is alone in having a relatively quiet year in the gap between its proposals for new user services in Year 1, and its final work on testing the new services in Year 3. Nonetheless its influence has been felt throughout the year – from outreach meetings, such as those in Nairobi and Recife, through the ongoing e-Bulletin and promotional materials, to the implementation of the new services that WP2 is starting to test in Year 3.

Service Activities (WP 5 & 6)
There are three themes that have dominated the Service Activities throughout Year 2 and as such have also created heavy workloads. Each of these has involved close interaction and co-operation between the Software Services (WP 6) led by Peter Schalk at ETI Amsterdam, and the CoL Services (WP 5) led by Yuri Roskov at Reading. First in this list is the delivery, testing and adoption of improved production software for the Catalogue of Life. Most notable is the Unified Assembly Process tool-chain, but also the CoL Harvester, presentation and production software linked to the new Base Schema, the GSD Builder Tool being made at MNHN Paris, and the new Metadatabase – a very significant group of software tasks. Second is the implementation of the new services, now divided into three batches, and requiring both software and particularly interface alterations at ETI, but also further data acquisition at Reading and guidance from the Services Team and WP 2. And last, but most important of all, the continued creation, publication and serving of new Catalogue editions including the Annual Edition on the web and on DVD. The emphasis here has been not only to push the Catalogue to new heights in content (it now covers 1.3 million species, about 73% of known species) and to refresh more frequently for the four editions per year, but also to maintain the production and schedules in a seemingly constant work pattern despite the substitution of completely new tools and processes.

Research and Development Activities (WP 7)
Finally, there has also been a peak of activity in the one RTD work package, led by Richard White at Cardiff. The Cardiff Team has not only put the final touches to its Service Oriented e-2 Architecture and made up time in its Recovery Plan, but also, because of a decision of the Design Team, brought forward several implementation items in the e-2 Architecture to be included now in the Unified Assembly Process tool-chain.

Cross-co-ordination
This has been a demanding year for the co-ordination of the project by the Co-ordinator and work package leaders supported by Alex Hardisty, convenor of the Design Team, and Sara Oldfield, convenor of the Services Team. Nonetheless, it is a compliment to all involved that the main tasks were delivered and reported at the Prague Meeting, and, with just one possible exception, fully ready for the tasks of Year 3.

Results during Year 2

We give here just the highlights from Year 2. Many more detailed results are given in the work package reports below.

Result 1. Upgrade to uniformity of GSD Supplier Data Sets
WP 3 has succeeded in organising a significant upgrade to the standards compliance and data fill of the 4D4Life GSDs. This is expected to yield a significant improvement to the data uniformity across the whole of the CoL when these data sets are refreshed in Year 3.

Result 2. Conceptualised a new public service for future implementation from the Multi-Hub Network.
WP 4 has generated the concept for an exciting new public service based on the Global Multi-Hub Network – presently referred to as '*Taxa of the World/Biota of the World'*.

Result 3. Created a working prototype of the Multi-Hub Engine
Within WP 4 CAS China has created a working prototype for the Multi-Hub Engine as a pilot-project.

Result 4. A large batch of CoL production software delivered and installed
In WP 6 & 5 a large batch of replacement core production software has been delivered and installed successfully – CAS Harvester, Interface linked to New Base Schema, Converter into New Schema, New Metadatabase.

Result 5. A second batch of software (Unified Assembly Process tool chain) delivered but needing continuing development.
A second very major software ensemble – the Unified Assembly Process tool-chain – is being delivered and tested. The development of this software is associated with a significant change to the working practices and automation of the production process. While it represents a great step forward, it requires both further development of rule-sets for each provider GSD and further development of the software, including hardening and improvement of the interface, before it can be 'set to work' as the mainstream production tool.

Result 6. Software and Content for Batches 1 and 2 of New Services implemented for test
The main software alterations needed to implement the new services in Batches 1 and 2 are implemented by WP 6 and now under test by WP 5 & 2. They will be reviewed and commented on by the User Panels in WP 2 early in year 3, before fine tuning, and launching later in the year. Examples already attracting user interest include:

i) Species numbers statistics at all levels of the hierarchy.
ii) Taxonomic sector quality indicators in the checklist
iii) Web-service to expand a species name with synonyms for searches.
iv) Web-service to return common names for a species.

Result 7. Continued on-time production and extension of Catalogue of Life editions, 4 per year, despite significant changes to the process and tools in use.
A significant achievement is the full production schedule of 4 editions of the Catalogue of Life during Year 2: (1 Jul, 2 Oct, 3 Jan, and 1 April-Annual Edition). Not only has the Executive Editor continued to locate and incorporate significant taxonomic sectors from new suppliers (now reaching 99 suppliers and 1.3 million species), but also to upgrade substantial existing data sets including

some of the new data from WP 3, and to operate the whole process against the background of testing and replacing software related to the Base Schema, the Interface and the Metadatabase.

Result 8. e-2 Architecture completed and first e-2 Tool fast-tracked into production.
As well as completing the WP 7 proposals for the new e-2 Architecture, WP 7 has also created a much-needed tool – the Hierarchical Data Editor – that has been fast tracked for use in the Unified Assembly Process tool-chain.

Expected final results of the entire project, and current status

i) New array of Catalogue of Life public services,
In test, On track for Year 3
ii) Modern e-infrastructure with strengthened supplier base and system infrastructure,
Many components in place, but not yet complete.
iii) Roll out of a state of the art service-based architecture,
Ready to roll (Year 3), and implementation continuing into i4Life project.
iv) Extended community participation and taxonomic coverage through the Global Multi-Hub Network, around the world and in Europe,
Already successful, and on track.
v) Mixed mode business model that takes the Catalogue of Life and its residual legal body, Species 2000, towards sustainable operation,
Plan ready (for Year 3), key components in place, effectively started
vi) Substantial progress towards completing the Catalogue of Life.
Strongly on track, but still some way to go (- a huge & challenging task!)
(including with i4Life & OpenBio)

Expected scientific and socio-economic impact, and status

We are already starting to see the first steps in the scientific and socio-economic impact envisaged for this project.

As stated in the First Periodic Report, the Catalogue is a true knowledge infrastructure used to *organise* and to *index* other biodiversity information, in addition to its use as an information source itself. It is in this sense that we refer to it as a taxonomic backbone – on which others can hang and organise biodiversity knowledge or records of biotic resources.

A number of user occurrences that occurred in Year 2 are given here to illustrate the key areas of impact:

i) By individual scientists

Prof. Nigel Stork from Melbourne Univ. visited the Secretariat to enquire about using the Catalogue and its new statistics services as comparators in his research into the true numbers of species in the major groups of Invertebrates. There is a pattern of similar requests, such as another from Copenhagen Museum. Another work in this topic (Chapman, A, 2009: *Numbers of Living Species...2nd Edition*, Australian Gov't, Canberra*)* cites the Catalogue extensively.

ii) By Public Portals and Biodiversity Science infrastructure programmes

In addition to well-established usages by GBIF and EoL, *and* the developing usage by partners in the i4Life project (EBI/GenBank, IUCN Red List, iBOL/CBOL/ECBOL Barcode programmes) *and* the usage by regional centres such as IABIN in S. America and ACB in ASEAN countries, two new linkages have been established during Year 2. First – it has become clear that the Catalogue will be needed and used by two if not three of the new ESFRI Infrastructures now going ahead – by ELIXIR based at EBI/EMBL Cambridge, and by LifeWatch based in Seville & Amsterdam. Formal partnerships are being established. Second, two FP7 EC projects (OpenUp! and BHL-E) wish to use the Catalogue as synonymic indexing framework for species-linked materials in their portals to the large digital media networks, Europeana and Biodiversity Heritage Library, with special interest in our new web-services.

A number of further new usages come from distant parts of the world – e.g. a 'mirror site' established by the Peruvian Government Ministry of the Environment, and permission to embed the Catalogue within a biodiversity inventory system of the Philippines Government.

iii) By Publishers and the Information Industry

A second commercial publisher has in Year 2 launched an electronic information product that makes use of the Catalogue for organising content.

iv) As a Global Biotic Resources Documentation System by public and governmental utilities and commercial companies across the world.

There is a growing perception of usage as a global biotic resources documentation system. A number of trial contracts and licences have been set up during the year relating to national import regulations requiring scientific naming of wild materials, both with trading organisations wishing to make importations, and with government agencies setting up regulations and monitoring. Particular interactions have been with a multinational trading company and the Biosecurity Services Group of the Australian Government. Similarly, part of the remit of our New Zealand partner (partner 38, Landcare Research (NZ) and its NZOR system www.nzor.org.nz/about-nzor) also comes from biosecurity concerns.

v) In the broader domain of biodiversity science and its contributions to the well-being of society.

One of the goals of the biodiversity informatics community is to provide a global-scale biodiversity virtual laboratory or cyber-infrastructure, as a platform that can that can bring together both global species-based, ecological and environmental data sets, and the electronic tools needed for aggregative, analytical and modelling studies. The result will be to open up the evidence-based analysis of the global biota and the first steps towards analysis, modelling and management – scientific areas with impact on the livelihood and quality of life of societies and individual citizens all over the world. Centre stage to the implementation of such virtual laboratories, for instance in the FP7 D4Science II, OpenBio and BioVel projects will be the key role of an electronic taxonomic backbone such as the Catalogue produced in this 4D4Life project.

vi) And an impact on taxonomists themselves.

Both the 4D4Life project and the growing community confidence in the Catalogue of Life product are having a socio-economic impact on the taxonomists and taxonomic institutions themselves. These are not easy times for taxonomists, with many in the molecular community questioning their role. So the shared vision of a clear product, of opening

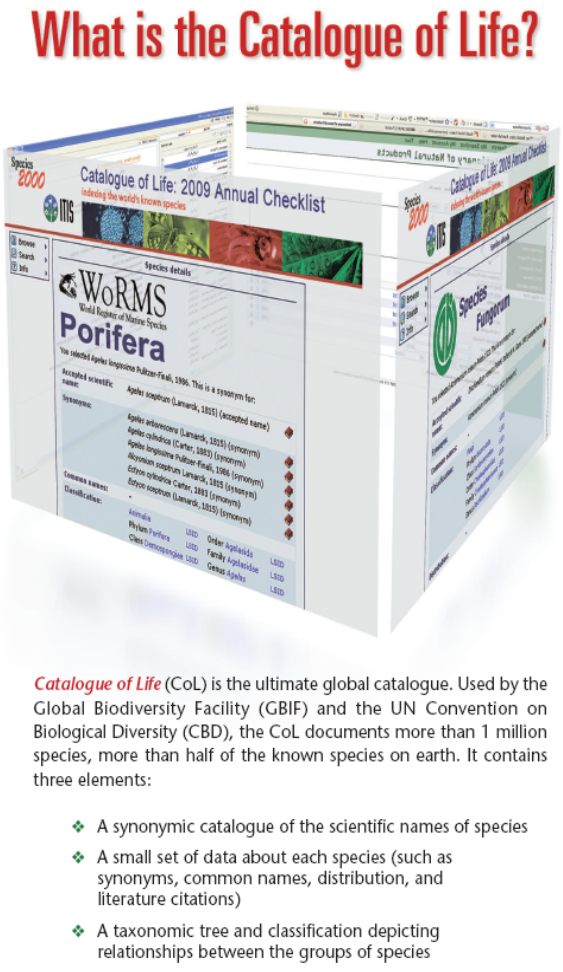
expertise to outside users, and of very substantial world-wide usage are proving strong stimuli across the profession. A single feature of the CoL usage stops everyone short. During many months of 2009 and 2010 the single largest user of the CoL online services was not Google or EoL or GBIF – but GenBank, the service for molecular biologists!

So to draw together these practical and scientific threads – our impact is developing in two key areas:

- in science: as the first comprehensive catalogue and taxonomic backbone for all organisms and drawn from an ecosystem of database services. A significant new layer of knowledge provision from across the taxonomic profession, and primarily for use by citizens and professionals across the broad sweep of biodiversity and biological disciplines.

- in e-science – as an electronic knowledge infrastructure for documenting the world's biotic records and resources, utilised by scientific biodiversity informatics developments and as a framework in biodiversity, commercial and regulatory organisations.

plus a real but diffuse contribution to the greater wellbeing and quality of life that citizens may benefit from developments in the science and management of the global biota.



The Catalogue of Life as shown in the advertising materials of one of our commercial users.