

The Catalogue of Life: towards an integrative taxonomic backbone for biodiversity

Frank A. Bisby, Yuri R. Roskov

Abstract — The Catalogue of Life Programme is addressing the need for a comprehensive catalogue of the world's presently known animals, plants, fungi and micro-organisms. The need is for an electronic catalogue that can be used as a taxonomic back-bone. In a wide variety of programmes covering species and documenting many types of biotic materials and records. The First Phase of the programme has used an architecture based on an array of global species databases to reach coverage of about two-thirds of known species. In the Second Phase of the Programme there will be a new architecture, a new array of services, and a ring of partnerships with global programmes.

Index Terms — Catalogue of Life, global taxonomic framework, species checklist, synonymic indexing, taxonomic hierarchy.



1 INTRODUCTION

Despite 250 years of effort in the taxonomic profession, there is still, in 2010, no complete catalogue of all presently known animals, plants, fungi and micro-organisms of the world. This is a critical problem for the scientific community, and for national, regional and global organisations that organise and regulate the exchange of biotic information and materials worldwide. The set of organisms known to science is a key dimension of human knowledge concerning global biodiversity, evolution, ecology, natural resources, and biotic response to climate change. It supplies a vital set of index terms needed to access most biodiversity knowledge. There is increasing public need and expectation, focussed through the UN Convention on Biological Diversity (CBD), to complete such a catalogue of all known organisms for international uses. Many commentators are surprised that a complete catalogue does not already exist. In fact it is a non-trivial task that is too large for the individual

F. A. Bisby is with the Species 2000 Secretariat, Centre for Plant Diversity & Systematics, School of Biological Sciences, University of Reading, READING, RG6 6AS, UK.

E-mail: f.a.bisby@reading.ac.uk.

Y. R. Roskov is with the Species 2000 Secretariat, Centre for Plant Diversity & Systematics, School of Biological Sciences, University of Reading, READING, RG6 6AS, UK. E-mail: y.roskov@reading.ac.uk.

capabilities of even the largest taxonomic institutions, due to the distributed nature of the knowledge.

The Species 2000 programme, working in partnership with ITIS in N. America, has made substantial progress with resolving this problem. It has created, maintained and enlarged its Catalogue of Life to the point where it now covers 1.25 million species of plants, animals, fungi and micro-organisms, more than two-thirds of the anticipated total of 1.9 million presently known species worldwide. It has done this by employing a radical architecture of federating global sectors of taxonomic expert knowledge from a growing array of supplier databases, and integrating these into a single taxonomic hierarchy and species checklist. The distributed system harvests taxonomic knowledge provided and maintained by a community of supplier organisations in the taxonomic profession, combining work by the major taxonomic institutions with that of smaller networks and individuals. This process was brought to production scale by the EC EuroCat project funded as a scientific infrastructure under FP 5 (2003 – 2006) and further developed since then with funding from other sources.

Over the last two years the programme has concentrated on extending and improving the scientific content of the Catalogue of Life, which is now a unique and scientifically valuable resource. However, it has come as a bonus to see the rising and now substantial public usage in Europe and all over the world, including by GBIF and the Encyclopedia of Life, of what is presently an incomplete service. The 4D4Life Project provides us with a timely opportunity to develop a parallel focus on services. It will enable us to enrich the variety and technical sophistication of taxonomic services that are undoubtedly possible, exploiting the taxonomic resource that we are already building. The utility of these services will secure the sustainability of the whole programme into the future.

2 THE PRESENT CONCEPT

The Species 2000 & ITIS Catalogue of Life (henceforward 'the Catalogue') has a single purpose, to enable users throughout biological and biodiversity sciences, and across the many scientific and non-scientific disciplines that use organism information, to access data about all organisms by means of a species checklist and a taxonomic hierarchy. It is already used to access data such as organism relationships, ecology, DNA sequences, protection status, invasive properties or information in any one of a myriad of other data domains. Such a Catalogue needs to be:

- i) comprehensive:** covering all known organisms in all groups;
- ii) global:** organisms of the whole world, in terrestrial, freshwater and marine environments;
- iii) validated:** a responsible, modern and professional globalised taxonomic view of the classification, supported by and embedded in the profession's activities;
- iv) accurate:** reflecting as accurately as is practical the detail of diversity of living organisms;
- v) accessible to all:** a clear view of the taxonomy, eventually in multi-lingual

presentation;

vi) available to all: widely and freely available in a variety of forms; and

vii) dynamic: updated for taxonomic changes through time, either continuously or annually.

To be effective in the many applications in which it is used, the classification and the naming of species and higher taxa must be as close to 'agreed and correct' as is possible in taxonomy. This means for each taxon either using a consensus system, or selecting by peer review and using consistently one of the competing classifications where alternatives are in wide use. Because alternative classifications have been used both today and in the past, users must be able to locate species known by other names (or concepts) in the Catalogue, and discover alternative names under which to access data on the internet or in other resources. Consequently synonymy and common names must be included for each species. As much as possible should be 'concept-based', a precision provided by some of the supplier databases.

The dream is simple - to create a Catalogue that contains an accurately maintained synonymic species checklist covering all known groups, connected in a validated taxonomic hierarchy.

3 THE EXISTING (PHASE 1) PROGRAMME

The present Catalogue of Life Programme, led by the global Species 2000 organisation based in Reading, and working with the N. American organisation ITIS, was set up as an international programme at a UK-funded (BBSRC) workshop in 2001. Bringing the programme up to production scale was funded by the EC as one of its scientific infrastructures (EuroCat), with further funding by the Japanese Government, the US Government (through ITIS) and GBIF. Output is via the *Catalogue of Life Annual Checklists* on the web [1], and on free DVD [2], and the *Dynamic Checklist* on the web, both also available as web-services for electronic use.

In March 2007 an EC-funded 'Million Species Day' symposium was held to celebrate reaching one million species. The *2010 Annual Checklist* now provides a quality species checklist of 1,257,735 species with unique identifiers and a hierarchy for all organisms (animals, plants, fungi, chromista, protozoa, bacteria, archaea, viruses). The estimate for the number of known extant species is currently 1.9 million [3].

The present Catalogue benefits from simplicity of structure incorporating minimal but standardised data for each species. These contribute to its success in providing a universal baseline needed by all biologists, and in making the project practicable. It consists of two knowledge structures, and software that enables the user to search or traverse them, and to toggle between them. i) The Species Checklist is a series of Species pages that are located by name searches, with automated synonymic indexing. Each page gives the Standard data for a Species, including common names, the higher taxa it belongs to in the hierarchy, and geographical distribution. ii) The Taxonomic Hierarchy is an expandable tree that can be followed down through the classification to the 1.25 million individual species. Or it can be used to navigate upwards to the higher

taxa containing the one that is viewed. By clicking on a higher taxon listed on a species page, the user can transfer to the tree for that taxon, and see all its daughters. Conversely, by clicking on a species at a twig in the tree, the user can visit the relevant Species page in the Checklist.

A comprehensive checklist cannot be made simply by adding together regional or single-country lists. Different classification and naming schemes mean that a simple additive list would be massively duplicative and of little use. The current system is a successful development of the original BBSRC SPICE project. It federates the taxonomic sector checklists provided by a distributed array of global species databases (GSDs), which are globalised checklists of a whole taxon, harvested across the Internet, and fitted together 'end-to-end' within a single overall classification. When enough sectors are fitted this process can eventually create a complete list. The number of GSDs contributing one or more taxonomic sectors to the Catalogue reached 77 for the 2010 edition, including 47 based in Europe, 18 in the USA, 5 in Brazil, and 7 in New Zealand, Russia, Japan, Taiwan, Australia and the Philippines. The model ensures that sectors are enhanced taxonomically by the supplier databases, and ca. 3,000 experts globally contribute to these databases. The whole programme depends on the integration and aggregation of expert knowledge from these key suppliers.

Each GSD sector is attached at its 'top point' (its highest ranking taxon) in the hierarchy, and in addition to harvesting the checklist, the system also harvests branches of the tree beneath this top point for the hierarchy leading down to the species in that sector of checklist. The checklist and hierarchy created from a growing array of GSDs in this present-day architecture ('Architecture 1') referred to as the 'Global Hub'.

Despite the evident success of Architecture 1 in permitting the rapid build up of the Catalogue to its present point, its limitations have been evident for some time. The difficulty is simply that no-one anywhere in the world is creating global species databases for some of the least known taxonomic groups, so by this model these would be destined always to remain as gaps in the Catalogue. In the EC EuroCat project (2003 – 2006), we additionally experimented with making a Regional or 'Euro-hub' with a further set of European regional databases, and versions of SPICE that could handle multiple hubs, and the first steps towards integrating their contents using the LITCHI 2 taxonomically intelligent integrity tracking. We then started to plan an 'Architecture 2', in which an array of Regional Hubs might be connected to the Global Hub, this providing linkage to regional databases from many parts of the world, but also the potential for the Global Hub to harvest data or checklist sectors from Regional treatments for the species groups that were missing from the Global Hub. Good progress is being made with initiating these Regional hubs now, and plans in the 4D4Life Project are to develop a unified concept and specification for this Multi-Hub Network working with the designated centres for China, New Zealand, Brazil, and Australia.

4 THE PHASE 2 PROGRAMME

In June 2009 Species 2000 and ITIS launched the Phase 2 programme of the Catalogue of Life with a fresh funding initiative and extended partnerships around the world planned for the 5-year period 2009 – 2014. In outline Phase 2 involves:

1. A new array of electronic and other services
2. A new service-based cyber-infrastructure: an ecosystem of services
3. A strategy for completing taxonomic coverage of the Catalogue
4. A world-wide multi-hub network of regional hubs
5. A 2nd Edition Catalogue of Life Management Hierarchy
6. A ring of partnerships with global biodiversity programmes

5 4D4LIFE PROJECT

The 4D4Life Project in the EC e-Infrastructure programme has now taken responsibility for the array of new services, the new cyber-infrastructure, and designing the world-wide multi-hub network. Sara Oldfield at Botanic Gardens Conservation International is co-ordinating the Services Team, and Alex Hardisty at Cardiff University is co-ordinating the System Design Team.

6 I4LIFE PROJECT

The i4Life Project in the EC e-Infrastructures Programme will shortly take responsibility for the ring of partnerships with global biodiversity programmes intended to harmonise and integrate between the taxonomic catalogues.

7 CONCLUSION

Substantial progress has been made with developing a comprehensive Catalogue of Life. However, there remains much to be done in the ambitious Species 2000 & ITIS Catalogue of Life Programme. The Catalogue is still far from complete in terms of taxonomic groups and known species; there is much to be done in improving both quality and fill of the Standard data set across all taxa; the new public services need to be fully tested and rolled out, and the programme needs to make progress with becoming sustainable as a scientific infrastructure for use around the world.

ACKNOWLEDGEMENT

This work was supported in part by the EC DG INFSO FP7 e-Infrastructures Programme under the 4D4Life Project (Grant 238988).

REFERENCES

- [1] F. A. Bisby, Y. R. Roskov, T. M. Orrell, D. Nicolson, L. E. Paglinawan, N. Bailly, P. M. Kirk, T. Bourgoïn and G. Baillargeon, *Species 2000 & ITIS Catalogue of Life: 2010 Annual Checklist*, www.catalogueoflife/annual-checklist/2010, Species 2000, Reading, 2010.
- [2] F. A. Bisby, Y. R. Roskov, T. M. Orrell, D. Nicolson, L. E. Paglinawan, N. Bailly, P. M. Kirk, T. Bourgoïn and G. Baillargeon, *Species 2000 & ITIS Catalogue of Life: 2010 Annual Checklist*, DVD, Species 2000, Reading, 2010.
- [3] A. D. Chapman, *Numbers of Living Species in Australia and the World*, 2nd Edition. Australian Biological Resources Study, Australian Government, Canberra, 2009.