Integrative Zoology 2006; 1: 18-21 doi: 10.1111/j.1749-4877.2006.00007.x

ESSAY

Towards an index of all known species: the Catalogue of Life, its rationale, design and use

Monalisa CACHUELA-PALACIO

Species 2000 Project, WorldFish Center - Philippine Office, Laguna, Philippines

Abstract

The Catalogue of Life, created by Species 2000 and the Integrated Taxonomic Information System, has the ambitious aim of creating a uniform and validated index of the world's known species for use as a practical tool in inventorying and monitoring biodiversity worldwide. This is achieved by accessing an array of taxonomic databases created and made available by individual, project and institutional custodians through taxonomic and biodiversity initiatives. A standard data set is delivered for every known species, drawn from Species 2000 contributing databases. Data are provided to the Species 2000 Philippine office, compiled annually and published in CD-ROM format. Species 2000 also includes the Species Locator Service, which gives a responsible opinion as to the actual species currently recognized by taxonomists and their accepted names, and the Name Service that assists users in checking the spelling and original publication details for a wider range of names, including species not yet incorporated in the Species Locator Service. The Species Locator Service is available as the Dynamic Checklist, which is used to search for the scientific name of an organism online, and the Annual Checklist, which is a stable index produced once a year and is available on CD-ROM and online. Catalogue of Life currently covers more than 272 000 species and 32 000 infraspecies, with 181 000 synonyms. It has 217 000 common names, 87 000 references and 20 contributing databases that encompass all known animals, plants, fungi and microorganisms.

Key words: Annual Checklist, biodiversity, global species database, species, Species 2000, synonyms.

INTRODUCTION

More than 1.75 million organisms have become known to science since Linnaeus began classifying all forms of life 250 years ago (United Nations Environment Programme 2002). These names of organisms are the key to understanding biodiversity and providing access to accumulated knowledge for all life on Earth. An inventory of species that exist in an ecosystem is also used as the starting point for assessment of the world's biological resources. But is there such a list? A catalogue is indeed of vast importance, especially given that no comprehensive indexing system yet exists for all these millions of animals, plants, fungi, protists and bacteria. This lack of a widely accessible index, with inbuilt mechanisms for maintenance and updating, is a significant constraint on all nations wishing to fulfill their obligations worldwide under the Convention on Biological Diversity.

This challenge is the driving force that led a scientific union of different international organizations in 1996 towards the formation of Species 2000: a "federation" of database organizations working closely with users, taxonomists and sponsoring agencies. This group agreed in 2001 to work jointly with the Integrated Taxonomic Information System (ITIS) to create a common "Catalogue of Life." The ITIS database is an automated reference of scientific and common names of biota of interest to North America. The goal of Catalogue of Life is to provide a uniform and validated quality index of all known species for use worldwide

Correspondence: Monalisa Cachuela-Palacio, 17 Marshall Laing Ave. Mt. Roskill Auckland 1004 New Zealand. Email: monalisapalacio@yahoo.com

as an electronic baseline species list in inventorying projects, an index for an internet gateway to species databases, a reference for comparisons between inventories and a comprehensive catalogue for checking the status, classification and naming of species (Species 2000 2004). To cite an application of this goal, Catalogue of Life is used as a core species index by the Global Biodiversity Information Facility.

The thrust of the Species 2000 plan is to create an array of participant global species databases covering each of the major groups of organisms. Each database will contain all known species using a consistent taxonomic system. Organizations with databases covering more than 85 major groups have joined the program. It is projected that existing database projects may provide approximately 55% of known species (Bisby 2000). The taxonomic database organizations starting the program already provided indexes for viruses, bacteria, corals, mollusks, crustaceans, flies, ichneumon wasps, geometrid moths, weevils, fishes, birds, mammals, fungi, algae, mosses and seed plants.

The Catalogue of Life specifically aims to:

1. Operate a dynamic Common Access System on the Internet through which users can locate a species by name across an array of online taxonomic databases.

2. Produce a stable annual edition of the species index, the Species 2000 Annual Checklist, which is available on the Internet and on CD-ROM, updated annually.

3. Stimulate completion of the array of taxonomic databases by seeking resources both for the completion of existing databases, and to help establish new databases to cover identified taxonomic gaps.

4. Establish a system of onward links connecting each species entry in the checklist with a wide range of other species databases with information on the species being searched; for example to include germplasm, museum/herbarium, ecosystem and other data systems (Species 2000 2004).

METHODOLOGY

The following methodology was implemented to achieve the goals of the Catalogue of Life:

1. Formation of a "federation" of existing taxonomic databases: Species 2000.

2. Stimulation of the establishment of global species databases for all groups of organisms, by accelerating the completion of existing databases and developing new ones.

3. Work towards an ultimate goal of providing a computer-based index of all known species.

4. Development of procedures to maintain the databases and to update the taxonomy.

5. Cooperation with international nomenclatural authorities in stabilizing nomenclature.

Compiling the Catalogue of Life (Annual Checklist)

Data provided by database custodians are compiled using Microsoft Access. Their global species databases!/standard datasets are contributed to Species 2000 through its Philippine office at the WorldFish Center, where all these files are incorporated in the Catalogue of Life database in Microsoft Access format. Data available are then crossmapped and appended to the eight major tables of the Catalogue of Life database. This is followed by a sequence of testing and checking to ensure data validity and consistency. Checking routines include sending copies of the synthetic database for scrutiny to global species databases custodians, catalogue editors, the Species 2000 Secretariat Office at the University of Reading in the United Kingdom, taxonomists and database experts, and members of the Species 2000 team. All comments and reports are then gathered and revised on the database. These are the protocols followed before publishing the data as a CD-ROM, thus ensuring that the Catalogue of Life will provide a uniform and validated quality index of all known species for use worldwide in different aspects.

THE CATALOGUE OF LIFE: OVERVIEW

The Catalogue of Life was first produced in 2001 with 14 databases and 222 135 species. The Catalogue of Life Year 2003 Annual Checklist included 304 710 species with 181 774 synonyms. It had 217 589 common names, 87 409 references with 656 507 bibliographic records. All these data were provided by 20 global species databases that encompass all five kingdoms.

Catalogue of Life is essentially a synonymic species checklist of animals, plants, fungi, protists and bacteria. It is completed by accessing different taxonomic sectors provided by an array of taxonomic databases or global species databases (GSDs). GSDs are unique because they contain worldwide coverage of all the species within one taxon. They treat species as taxa, and contain synonymy and taxonomic opinion. GSDs have an explicit mechanism for seeking at least one responsible or consensus taxonomy and for applying it consistently. And GSDs cross-index significant alternative taxonomies via synonymy and have mechanisms to enhance the taxonomy over time (Froese *et al*. 2003). GSD can be put together end-to-end because they do not overlap, and they contain a consistent taxo-

nomic treatment for the whole of one higher taxon. These databases are made available by individual, project and institutional custodians through taxonomic and biodiversity initiatives.

Species 2000 delivers a standard dataset for every species included in GSD, with eight field groups defined for each species (Bisby & Roskov 2003), these are:

1. Accepted scientific name with references: valid or correct name that is currently accepted for the species as a taxon (one per species).

2. Synonyms with references: a list that can include zero to many species or infraspecific names. There are three synonymic possibilities Species 2000 considers to create a uniform, accurate but broad set of synonymic links: unambiguous synonyms that point unambiguously at one species, ambiguous synonyms that are ambiguous because they point at the current species and one or more others, and misapplied names that have been wrongly applied to the current species and may also be correctly applied to another species.

3. Common names with references: may include zero to many names.

4. Latest taxonomic scrutiny: details of when and by whom the species was last reviewed in the source database, thus an indicator of quality and a form of credit.

5. Source database information: attached to each species record.

6. Comment field: includes information from one or several data fields from the source database as decided by its custodian.

7. Family name: contains only one valid Latin name to which the species belongs.

8. Distribution: a list of geographic records from zero to many areas.

Additional information is made available either within the appropriate source database or through the hyperlinks to other databases. The user can communicate directly with one or more species databases, without intervention through Species 2000.

Species 2000 presents two types of services: the Species Locator Service and the Name Service. In the Species Locator service, responsible opinion as to the actual species currently recognized by taxonomists and their accepted names are given. In the Names Service, users can check the spelling and original publication details for a range of names, including many groups of organisms not yet incorporated in the Species Locator Service. The Species Locator Service connects to taxonomic databases in two ways: through the Dynamic Checklist, where online taxonomic databases are connected over the Internet to the Species 2000 Common Access System. Thus, a real-time connection is achieved where the catalogue seen by users is dynamic in the sense that sectors may be updated as new editions come online at any time. Another way of connecting to the taxonomic databases is through the Annual Checklist, which is a stable index for reference and comparison, produced annually using data from the Species 2000 array of taxonomic databases and available on CD-ROM and online.

Species 2000 has two access architectures. Architecture One is the virtual array of GSDs or digital library of biodiversity databases connected by onward links, of which the majority are regional databases. Architecture Two is the combination of global and regional hubs, such as Species 2000 Europa EC Project in the United Kingdom, Species 2000 Asia Oceania in Japan and ITIS for North America in the United States. Catalogue of Life uses these architectures as a strategy to locate and work with as many GSDs as there exist with metadatabase as the coverage, establish linkages between GSDs and regional databases and encourage integration for the remaining groups (Species 2000 2004).

The Annual Checklist is presently created by the contributing GSD, who provide an annual download of the standard dataset for their taxonomic sectors. Data are provided to the Species 2000 Philippine Office, compiled using Microsoft Access and made available online and on CD-ROM. The ITIS database is used as the baseline for the Annual Checklist because it contains at least some species records for nearly all of the major taxa. For other data, each available GSD for a particular taxonomic sector is used, and the corresponding taxa in ITIS are deleted to avoid duplication of data. A summary of the data in the Catalogue of Life 2003 CD-ROM is shown in Table 1, with the 20 source databases and number of data records contributed to the Catalogue of Life database. As a result of standard procedures for encoding GSDs to the Annual Checklist, some parts are GSD sectors, either of ITIS or Species 2000 origin, and the remaining sectors are filled on an interim basis from ITIS (Bisby *et al*. 2003).

There are three ways of performing a search using the Annual Checklist. The user can either ask for species scientific or common name, or browse by hierarchy from kingdom level to species epithet. Standard information on the species will be returned to the user. This includes valid name, synonyms, complete hierarchy, common names, distribution, references, name and contact information of the taxonomist who checked the data, details on contributing database, online links to the species information in the Internet version of the Catalogue of Life, and links to the species in the website of the contributing database. All

these data can easily be printed as a report.

Catalogue of Life envisions a total of up to 200 component GSD for all species to be listed. At least 150 global species databases, each initially covering 10 000 to 25 000 species, will be needed for all species to be included. It will seek resources for the completion of the existing databases and help establish new databases to cover the 45% of species that are not yet catalogued in a database format (Bisby 2000). After Catalogue of Life achieved a listing of 300 000 species in 2003 and 500 000 species in 2005, it endeavored to attain 800 000 species in 2006 and to completely enumerate all of the 1.75 million species by 2011 (United Nations Environment Programme 2002). This massive scope of a species record can only come to reality if a consortium approach is implemented. This challenge is still ahead of us. Catalogue of Life calls for taxonomists in each discipline to collaborate and build the global species database on the taxa of their expertise. It is high time to fight the threats to biological diversity. Whatever conservation effort we employ, it always rests on a very basic need, to have at least a complete index of these species that we are trying to save.

REFERENCES

Bisby FA (2000). The quiet revolution: biodiversity informatics and the internet. *Science* **289**, 2309-12.

Bisby FA, Roskov YR (2003). Species 2000 Standard Dataset, version 2. Paper presented at the Species 2000 Annual Meeting, University of Reading, 9-11 May 2003, Reading, United Kingdom.

Bisby FA, White RJ, Roskov YR (2003). Species 2000 Database Connection Protocols, version 1. Paper presented at Species 2000 Annual Meeting, University of Reading, 9-11 May 2003, Reading, United Kingdom.

Froese R, Bisby FA, Wilson KL, eds. (2003). *Species 2000 and ITIS Catalogue of Life 2003: Indexing the World's Known Species; Year 2003 Annual Checklist.* Species 2000, Los Baños, Philippines.

Species 2000 (2004). *Species 2000*. [Cited 24 February 2004.] Available from URL: http://www.sp2000.org

United Nations Environment Programme (2002). *Global Environment Outlook 3: Past, Present and Future Perspectives*. EarthScan Publications, London.