Tropical Diversity (2023) 3(1): 1-6.
ISSN: 2596-2388
DOI: 10.5281/zenodo.10911800

RESEARCH NOTE

ACACIA: a generic conceptual schema for taxonomic databases

ACACIA: um esquema conceitual genérico para bancos de dados taxonômicos

Mauro José Cavalcanti[1*] https://orcid.org/0000-0003-2389-1902

Ecoinformatics Studio, PO Box 18123, CEP 20720-970, Rio de Janeiro, RJ, Brazil.

*Email: maurobio@gmail.com

Received: 3, July 2023 / Accepted: 1, November 2023 / Published: 3, November 2023

Resumo ACACIA é um esquema de dados genérico para bancos de dados taxonômicos relacionais e um pacote de programas que o implementa. Tal esquema permite a representação eficiente de todas as classes de dados requeridas para o armazenamento e recuperação de dados de biodiversidade, do nível taxonômico aos níveis ecológico e genômico. O pacote de programas ACACIA para o gerenciamento de bancos de dados de biodiversidade, desenvolvido na linguagem PHP, está disponível em https://github.com/maurobio/acacia sob uma licença livre GPL v3.

Palavras-Chave: Bancos de dados de biodiversidade, cibertaxonomia, modelo relacional, BAOBAB, DELTA.

Abstract ACACIA is both a generic data schema for relational taxonomic databases and a software package which implements it. Such schema allows for the efficient representation of all data classes required for the storage and retrieval of biodiversity data, from taxonomic to ecological and genomic levels. The ACACIA software package for the management of biodiversity databases, written in the PHP language, is available from https://github.com/maurobio/acacia under the GPL v3 free license.

Keywords: Biodiversity databases, cybertaxonomy, relational model, BAOBAB, DELTA.

Research supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [grant #350389/2011-0] and Fundação da Amparo à Pesquisa do Estado do Amazonas (FAPEAM) [grant #2275/11].

Introduction

In the last three decades, a large number of taxonomic databases have been developed to address curatorial management processes, taxonomic revisions and applied biology needs (Pankhurst, 1991) as well as the growing demand for large-scale global biodiversity information systems accessible over the World Wide Web (Bisby, 2000; Curry & Humphries, 2007).

A basic step towards developing a taxonomic database system is to build a data model to describe the entities involved and relationships among them. This has lead to the development of several models for the design of taxonomic databases (Allkin & Bisby, 1988; Allkin & White, 1982, 1988; Berendsohn, 1997; Berendsohn et al., 1999). Although no doubt useful in clarifying relationships among taxonomic entities and their attributes, these models are, however, invariably of such complexity as to make them rather difficult to implement and manage (Morris, 2005).

More recently, this approach has been expanded towards the development of comprehensive solutions for taxonomic computing, as the Scratchpad (Smith et al., 2009) and the EDIT Platform for Cybertaxonomy (Berendsohn, 2010; Berendsohn et al., 2011). The former is built on top of the generic content management system Drupal, whereas the latter comprises both a common data model and a set of specialized software tools for interacting with it. These solutions, however, still do not offer the average user an independent environment, at the same time simple to use and to maintain, for dealing with the complexity of taxonomic data, as nomenclatural information, the taxonomic hierarchy, structured and unstructured descriptive data, geographic information, literature citations, and ecological and genomic data.

The ACACIA design is an attempt to overcome this difficulty, offering a simple but flexible and extensible data model that can be used as a framework for taxonomic information systems. ACACIA intends to be a practical implementation of the ideal of a "universal biological database structure" (White & Allkin, 1993).

The name "Acacia" is an allusion to "Baobab", a comprehensive database design for biologists developed by Allkin & White (1982, 1988) and partially implemented in the Alice species diversity database management system (White & Allkin, 1993; White et al., 1993). The ACACIA model can be conceived of as a simpler version of the full BAOBAB design. The genus *Acacia* has also been frequently used as an example in taxonomic database studies (Allkin et al., 1992).

Materials and Methods

As a database scheme, ACACIA is a set of 15 entities or tables based on the relational database model (Codd, 1970), designed to convey all basic classes of information required for taxonomic databases and facilitate the recording of taxonomic data from literature and other sources (biological collections, field surveys, and other databases). Four tables in the set are mandatory in any ACACIA database: species (storing valid, *ie.*,

currently accepted, species names), synonyms (storing any other names by which a species may also have been previously described), higher taxon (storing basic taxonomic categories above species-level, *i.e.*, kingdom, division or phylum, class, order, family, as well as intermediary levels if required) and bibliography (storing bibliographic references to each valid name and any synonyms). These four tables constitute the "Taxonomic Core" of the schema and are mandatory in any ACACIA database, plus a metadata table storing data about the database itself (Figure 1). This generic schema can be used in building taxon-oriented databases, using any desired combination of computer platform, operating system, database engine, and application programming language. Once implemented, the ACACIA schema can be used to create a wide variety of taxonomic databases of diverse content including monographic databases, species inventories, annotated checklists or identification keys (when used in combination with the DELTA system; Dallwitz, 1980).

The ACACIA design is based on and fully compatible with the International Legume Database and Information Service (ILDIS) Type One Data Fields (Bisby, 1989, 1993), as well as with the Species 2000 (Bisby, 2000) and the Catalogue of Life (Bisby & Roskov, 2010) Standard Dataset (which is itself loosely derived from the ILDIS standard).

The ACACIA scheme is completely neutral and can be implemented in any relational database management system, from Sqlite and MySQL to PostgreSQL and Oracle. So far, all implementations have been based on MySQL, because of the widespread availability, ease of installation, and low footprint of that particular DBMS.

An integrated software package, written in the PHP language, has been implemented to manage databases designed with the ACACIA scheme. This tool is an interactive data entry, querying, and editing system for taxonomic databases based on the generic ACACIA conceptual schema. It combines the automated use of scientific names and synonyms in a species checklist with online access to geographical data, morphological descriptors, genomics, ecology, vernacular names, economic uses, structured notes and conservation status about each species, with all these data being cross-indexed to a citation list. Interactive keys may be easily created and published online from morphological data stored in database and automatically translated into DELTA format (Dallwitz, 1980), using the NaviKey Java applet (Neubacher & Rambold, 2005). For DNA sequence data, several manipulation routines from the BioPHP project (http://www.biophp.org) are provided (for computation of reversals, complementarity, G+C content, nucleotide composition, and conversion to RNA).

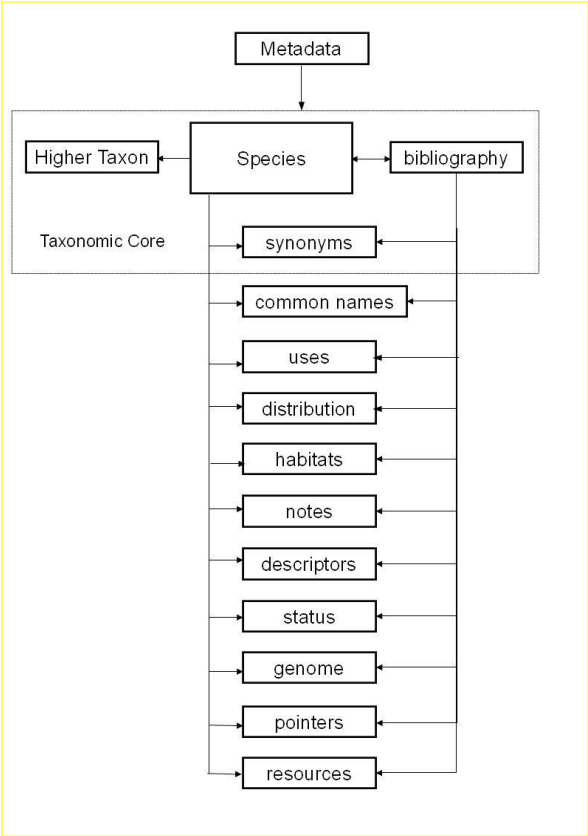

Figure 1 – The ACACIA database schema.

Results and Discussion

The software design allows for rapid customization to suit its application to any taxonomic group (plant, animal, or microorganism). Since only the data tables comprising the taxonomic core are mandatory in a given database based on the ACACIA scheme, every other data table can be individually activated or deactivated when configuring the tool, therefore allowing the database to be tailored to better suit a specific field of application, as taxonomy, conservation, or ethnobiology. This also allows reducing the database server overload, by avoiding the need to keep in the system empty unused tables (for example, a genome table in a database which does not store genomic data).

Three production databases have already been implemented with the ACACIA schema and are currently available on the web: (1) the Neotropical Copepoda (NEOCOP) database (http://neocop.biotupe.org); (2) the Marine-derived Amazonian biota Research (MAR) database (http://mar.biotupe.org); and (3) the Brazilian Coral Reef Fish (http://coralfish.scienceontheweb.net) database. The development of a third database, the Neotropical Cladocera (NEOCLAD) database (http://neoclad.biotupe.org), has been started and is currently in prototype stage.

Conclusions

The current version of the ACACIA software package works in tandem with a program called Feronia (in allusion to a goddess associated with wildlife and fertility in Etruscan and Roman mythology), written in the Python programming language; this program allows the fast collation and building of databases by using several available web services in order to harvest and/or check data on nomenclature (from the Catalog of Life, http://www.catalogueoflife.org), distribution (from the Global Biodiversity Information Facility, http://www.gbif.org), genomics (from GenBank, http://www.ncbi.nlm.nih.gov/genbank) conservation status and habitats (from the IUCN Red List (http://www.iucnredlist.org), and literature (from http://www.itis.gov). Future versions are expected to improve on this, both by offering scripts for harvesting data from more biological web services.

Acknowledgements

Thanks to Haroldo Cavalcante de Lima (Jardim Botânico do Rio de Janeiro) and Robert Allkin (Royal Botanic Gardens, Kew) for providing information on the ILDIS data standard and the Alice database system, to Eduardo Dalcin (Jardim Botânico do Rio de Janeiro) for earlier discussions of the ACACIA model and useful suggestions, and to Gustavo Martinelli (Jardim Botânico do Rio de Janeiro) and Edinaldo Nelson dos Santos Silva (Instituto Nacional de Pesquisas da Amazônia) for support and interest. Veridiana Scudeller (Universidade Federal do Amazonas) and Valéria Gallo (Universidade do Estado do Rio de Janeiro) provided constructive reviews and valuable suggestions which much contributed to the improvement of the original manuscript. This paper is dedicated to the memory of Frank A. Bisby (1945-2011), for his pioneering work on taxonomic and biodiversity databases.

References

Allkin, R. & Bisby, F. A. 1988. The structure of monographic databases. Taxon 37**:** 756-763.

Allkin, R. & White, R. J. 1982. Design criteria for a computer program to facilitate the acquisition, storage, retrieval and reformatting of biological descriptions. Southampton: University Research Fund Report.

Allkin, R. & White, R. J. 1988. Data management models for biological classification. In: Bock, H.H. (ed.). Classification and Related Methods of Data Analysis. Amsterdam: Elsevier, pp. 653-660.

Allkin, R., White, R. J. & Winfield, P. J. 1992. Handling the taxonomic structure of biological data. Mathematical and Computer Modelling 16: 1-9.

Berendsohn, W. G. 1997. A taxonomic information model for botanical databases: the IOPI model. Taxon 46**:** 283-309.

Berendsohn, W. G. 2010. Devising the EDIT Platform for Cybertaxonomy. In: Nimis, P. L & Vignes Lebbe, R. (eds.), Tools for Identifying Biodiversity: Progress and Problems. Trieste: Edizioni Università di Trieste, pp. 1-6.

Berendsohn, W. G., Anagnostopoulos, A., Hagedorn, G., Jakupovic, J., Nimis, P. L, Valdés, B., Güntsch, A., Pankhurst, R. J. & White, R. J. 1999. A comprehensive reference model for biological collections and surveys. Taxon 48**:** 511-562.

Berendsohn, W. G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K., & Müller, A. 2011. Biodiversity information platforms: From standards to interoperability. ZooKeys 150: 71-87.

Bisby, F. A. 1989. Databases, information systems, and legume research. Monographs in Systematic Botany from the Missouri Botanical Garden 29: 811-825.

Bisby, F. A. 1993. Species diversity knowledge systems. The ILDIS prototype for legumes. Annals of the New York Academy of Sciences 700: 159-164.

Bisby, F. A. 2000. The quiet revolution: biodiversity informatics and the internet. Science 289: 2309-2312.

Bisby, F. A. & Roskov, Y. R. 2010. The Catalogue of Life: towards an integrative taxonomic backbone for biodiversity. In: Nimis, P. L & Vignes Lebbe, R. (eds.). Tools for Identifying Biodiversity: Progress and Problems. Trieste: Edizioni Università di Trieste, pp. 37-42.

Codd, E. F. 1970. A relational model of data for large shared data banks. Communications of the ACM 13**:** 377-387.

Curry, G. B. & Humphries, C. J. 2007. Biodiversity Databases: Techniques, Politics, and Applications. Boca Raton: CRC Press.

Dallwitz, M. J. 1980. A general system for coding taxonomic descriptions. Taxon 29**:** 41-46.

Morris, P. J. 2005. Relational database design and implementation for biodiversity informatics. PhyloInformatics 7*:* 1-66.

Neubacher, D. & Rambold, G. 2005. NaviKey – a Java applet and application for accessing descriptive data coded in DELTA format. http://www.navikey.net.

Pankhurst, R. J. 1991. Practical Taxonomic Computing. Cambridge: Cambridge University Press.

Smith, V. S., Rycroft, S. D., Harman, K. T., Scott, B. & Roberts, D. 2009. Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. BMC Bioinformatics 10 (Suppl 14): S6.

White, R. J. & Allkin, R. 1993. A strategy for the evolution of database designs. In: Bisby, F. A., Russell, G. F. & Pankhurst, R. J. (eds.). Designs for a Global Plant Species Information System. Oxford: Oxford University Press, pp. 284-303.

White, R. J., Allkin, R. & Winfield, P. J. 1993. Systematic databases: the BAOBAB design and the ALICE system. In: Fortuner, R. (ed.). Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision. Baltimore: Johns Hopkins University Press, pp. 297-3

