ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2021.1.023

Proposal of a Relational Database (SQL) for Zoological Research of Epigeic Synusion

Vladimír Langraf[1,✉], Kornélia Petrovičová[2], Stanislav David[3], Zuzana Krumpálová[3], Adrián Purkart[4], Janka Schlarmannová[1]

[1]Department of Zoology and Anthropology, Constantine the Philosopher University in Nitra, Slovak Republic
[2]Department of Environment and Zoology, Slovak University of Agriculture in Nitra, Slovak Republic
[3]Department of Ecology and Environmental Sciences, Constantine the Philosopher University in Nitra, Slovak Republic
[4]Department of Zoology, Comenius University, Slovak Republic
langrafvladimir@gmail.com✉

Abstract

In recent years, developments in the field of molecular biology and genetics have led to the increase in biological information stored in databases. The same increase in the volume of information occurred in the field of zoology, but the development of databases was not addressed in this area. We prepared a relational database and its diagram in the Microsoft SQL Server Management Studio (SSMS) database program. Our results represent experience with construction of a new database design for the zoology field with a focus on research of epigeic groups. The structure of the database will help with meta-analyzes with the objective to identify zoological and ecological relationships and responses to anthropic intervention.

Keywords: *Big data, SQL, SSMS, biology, zoology, epigeic groups.*

Received: 18 May 2021
Accepted: 16 June 2021
Published: 21 June 2021

1 Introduction

Already in the past, the need for analysis of biological data has led to the use of simple databases created in programmes such as dBase (1980s), FoxPro (for DOS, Windows, 1990s), Microsoft Excel, and Access. Due to the constant addition of a large amount of new data (Big data) in the field of biology (zoology, botany, anthropology, genetics, molecular biology), there is a need to create large databases. The data with which databases can be populated are usually generated from various sources (servers, sensors built into telephones, video cameras, MRI scanners, set-top boxes) [30, 45]. Bioinformatics, which includes three subdisciplines, focuses on the storage, organization and analysis of a huge amount of this data (Big data). The first subdiscipline deals with the development of new algorithms and statistics necessary to evaluate relationships among members of large data sets. The second is focused on the analysis and interpretation of data of various types and the third subdiscipline engages the development and implementation of tools enabling effective access and management of database information. To understand the structure (architecture) of biological databases, it is necessary to know the concepts of relational databases (Structured Query Language – SQL) and the concepts of obtaining information from digital libraries. Long-term development and management of biological databases is a key area of bioinformatics [44].

We currently have extensive databases focused on nucleic acids [37], DNA data-bases (GenBank) [7], RDA (RNAcentral) [14], protein databases Proteins database (PDB), Universal Protein Resource (UniProt)), Human Protein Atlas database [36], cancer-focused disease databases (Cancer Genome Atlas (TCGA), Cancer Genome Consortium (ICGC)) [43] and databases for the industrial bioeconomy (producing polymers, spider silk fermentation with recombinant Escherichia coli or yeast) [19, 21, 33]. Universal Protein Resource includes three member databases: UniProt Knowledge-base (UniProtKB), UniProt Reference Clusters (UniRef), and UniProt Archive (UniParc) [39]. The local databases for Slovakia include the Taxon and Biotope Information System (ISTB) of the ŠOP (State Nature Conservancy) [1]. The central database storing phytocenological records is the (CDF) [2]. There is a phytocenological database Pladias in the Czech Republic [3] and for taxa there is BioLib [4]. The content of databases includes tables (frequency, dimension, code), text descriptions, name of columns, attributes, entities, classifications, datatype (data format). Thus, the biological database is formed of a set of structured biological data and collected data organized so as to allow easy access to obtain, manage and update the content [11, 41].

The most commonly used data management model is the relational model. The language suitable for processing large amounts of data (Big data) of the relational database model is the Structural Query Language (SQL) incorporated in programs such as SQL Server Management Studio (SSMS) or MySQL. Unlike standard databases, SQL enabled the use of relational databases using a set-oriented database lan-

guage. Databases can generally be classified as primary – containing sequence or structure information (e.g. Swiss-Prot & PIR, GenBank & DDBJ)[40]. Secondary database has information derived from the primary database (napr. SCOP, CATH, PROSITE, eMOTIF) [18, 46]. More complex databases include sophisticated query facilities, bioinformatic data, analysis tools. In order to understand biological databases, it is necessary to know the concepts of relational databases (SQL) and the concepts of obtaining information from digital libraries. The main role of databases is to provide valuable knowledge in selected scientific disciplines with the help of the collected data [28].

Databases include data stored in a data format (datatype), the great heterogeneity of which gradually leads to problems in software implementation. People working with databases which transfer information to applications (software) should have good programming experience to be able to modify existing or create new scripts. This information is also needed to analyze biological data and convert the data format [9]. A recently proposed binary format for large-scale structures is the Macro Molecular Transmission Format (MMTF). It provides fast transfer of large amounts of data, visualization and analysis. The binary format enables the compaction of the data and the storage of PDB legacy archives up to 7 GB [13].

Since biological databases have been created mainly in the field of genetics, molecular biology and biomedicine so far, our goal was to design a relational database in SQL Server Management Studio (SSMS) for the storage of zoological research data focused on epigeic groups (Acarina, Aranaea, Collembolla etc.). In this paper we provide new information about the design of a relational database in zoology, which will enrich the field of bioinformatics.

2 Existing Databases

For the creation of biological databases, we must understand the information contained in the database, how to store them correctly and present them during implementation using software tools [5, 6, 7]. Functional databases which currently exist focus on the storage of protein and nucleic acid data [8, 12, 16], DNA [9], RDA [10], cancer-focused diseases [11] and bioeconomies [13, 14, 15]. Our results represent a new design of a relational database for the field of zoology with a focus on storing research data of epigeic groups. The need for such a database increases with the growing amount of information that is accumulated every day. At present, such a database is not available for the needs of zoologists focusing on the epigeic group. Establishing such a database would solve this shortcoming and at the same time contribute to the enrichment of information in the field of bioinformatics. Thus, a biological database is a collection of organized data with easy access to obtain and update data [21, 22]. The primary database consists of information on structure and sequence [23], the secondary database includes derived information from the primary database [24, 25]. The data collected in this way provide valuable information for meta-analyzes of selected disciplines [25].

Before starting the design of a biological database, it is necessary to obtain all the details for its effective solution. It is also advisable to choose a data model for a relational database, e.g. Data Description Language (DDL), Oracle SQL Developer Data Modeler, SQL Developer or E-R model (IE = Information Engineering) [10]. Our relational database data model is the E-R model, the structure of which is shown in Fig. 1. The schema was created in Microsoft SQL Server Management Studio (SSMS) programme. The Sigma link 2 application stores protein and miRNA data in a relational database in MySQL, where a schema is also created [23].

Traditional databases have data in tables stored in fields for columns. The advantage of column data storage is the faster result of aggregation queries (queries where you only need to lookup subsets of your total data) compared to row databases [20, 42]. The above accelerations will be reflected in case of Big Data with constant data filling. Our tables in the database also have data populated into fields for individual columns, to accelerate the queries. We divided the tables into frequency, dimension and code tables connected using a primary and foreign key, which also accelerates the queries. The similarity of the data type in most columns allows better compression when running compression algorithms, which also accelerates the queries [38], thus speeding up the data analysis [8, 17]. In the early days of databases, indexed ASCII text files, called "flat files", were used. The typical data type which is part of flat files contains nested records, sets, lists, and variants. Flat files form the standard in biological databases, on the basis of which data management must always be implemented [5, 12, 31, 34, 47]. In the 1970s, among the first data formats included the PDB, which was used to store and exchange data on protein structures. It was replaced and modified in 2014 to the PDBx/mmCIF format, which also stored crystallographic data [31]. Another recently proposed binary format for large-scale structures is the MMTF (Macro-Molecular Transmission Format) [26, 35]. Our data types used are supported by SQL server. We used an integer (int) to store integers, with the floating point numbers having a data type (float). The date is stored as data type date, and text strings as nvarchar.

A well-designed relational database and efficient communication between tables are needed for analysis [24, 29]. It has been used in global prediction of the response of terrestrial biodiversity to anthropogenic activity in various parts of the planet [35, 41]. Meta-analyzes have also contributed to the study of medicinal plants for pharmaceutical use and the potential for the discovery of new species [22, 27]. Metabolite analysis of Taxus baccata L., the world's leading source of anticancer drugs, has also been performed on the basis of a large dataset [6, 15, 16, 48]. Our

relational database for epigeic groups is designed to analyze metadata of zoological-ecological relationships and responses to anthropogenic activity.

3 Proposal of a Database

3.1 Database Structure

In the Microsoft SQL Server Management Studio (SSMS) [32] program, we designed a relational database for storing data from epigeic groups research. The zoological database consists of frequency tables (18), dimension tables (18) and code tables (5), connected using a primary and foreign key. The connection of individual tables is illustrated in the diagram (Figure1). Frequency tables are filled with data in repeated intervals based on the frequency of collection and contain data of the types of the given orders:

- f_speciesAcarina
- f_speciesAraneae
- f_speciesColeoptera
- f_speciesCollembola
- f_speciesDermaptera
- f_speciesDiplopoda
- f_speciesGlomerida
- f_speciesHaplotaxida
- f_speciesHymenoptera
- f_speciesIsopoda
- f_speciesJulida
- f_speciesLithobiomorpha
- f_speciesOpilionidea
- f_speciesPolydesmida
- f_speciesPseudoscorpion
- f_speciesStylomatophora

The frequency table f_order is a summary table for all series and f_environmental. Variables is populated with environmental variables (e.g. temperature, humidity, pH). Dimension tables with one record for an entity (e.g. species, localities, biotope) are represented by following tables:

- d_speciesAcarina
- d_speciesAraneae
- d_speciesColeoptera
- d_speciesCollembola
- d_speciesDermaptera
- d_speciesDiplopoda
- d_speciesGlomerida
- d_speciesHaplotaxida
- d_speciesHymenoptera
- d_speciesIsopoda
- d_speciesJulida
- d_speciesLithobiomorpha
- d_speciesOpilionidea
- d_speciesPolydesmida
- d_speciesPseudoscorpion
- d_speciesStylomatophora

The above mentioned tables store data with species characteristics belonging to epigeic groups. The d_biotope table stores data of localities with biotope characteristics. Non-duplicate data are also included in the dimension table d_pageRole with data from the workplaces of scientists collecting and determining epigeic groups. Code tables are code lists with an assigned unique ID and one record and include the following tables cl_order (orders), cl_familia (families), cl_genus (genera), cl_method (collection method), cl_fertility (soil fertility). The connection of tables via the primary key (PK) and the foreign key (FK) is executed as follows. Dimension tables of species of individual epigeic groups (e.g. d_speciesColeoptera) are linked via PK (ID, yellow key) to speciesID (FK, infinity symbol) of frequency tables of species of epigeic groups (e.g. f_speciesColeoptera). The dimension table d_pageRole has a PK (ID) linked to the frequency tables of species and genera via the FK (heDetermined, heCollected). The d_biotope PK (localitiesID) table is linked via the FK (localitiesID) for the frequency tables of species, genera and environmental variables. In the same way, frequency tables of species and orders (f_order) are linked, using the FK (orderID, familiaID, genusID, methodCode) to code tables cl_order, cl_familia, cl_genus via the PK (ID) and cl_method PK (methodCode). The f_environmentalVariables table is linked by the FK (fertilityID) to the code table cl_fertility PK (ID). The connection link of dimension and code tables with frequency tables is 1: N. abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

3.2 Data Type of Database

Data type (data format) were selected based on the nature of the entered data. Numbers had a data type numeric. The above data type is for the following attributes ID, numberTraps, orderID, familiaID, genusID, speciesID, heDetermined, heCollected, rolaID, localitiesID, biotopID, fertilityID, PSC, numberIndividuals, ageGrowth, occurrenceMounth, pHSoil, moistureSoil, luminosity, fertilityCoef, codes, temperature, precipitation, canopyOpenness, flooding, groundWater, temperatureSoil, metersAboveSeaLevel, RakeThickness, moosFloor, herbalFloor, shrubFloor, treeFloor, nitrogenDown, nitrogenUp, phosphorusDown, phosphorusUp, potassiumDown, potassiumUp, length-body, heightBody, widthBody, Bv, Ev, lengthProthorax, lengthElytry, lengthCaput, lengthEyea, lengthPronotum, lengthAntennas, lengthFemur, lengthTrochanter, lengthtibia, lengthTarsus, widthProthorax, widthElytry, widthCaput, widthPronotum, widthFemur, heightPronotum, shapePronotum. The YYYY-MM-DD format had a date attribute with a date data type. Variable length text strings are

Figure 1: Relational database design scheme for epigeic groups (E–R model).

stored as data type nvarchar (maximum size 536 870 912 characters). This data type was assigned to the following attributes methodCode, methodName, species, sex, shapePronotum, order, familia, genus, described, ecologicalCharacteristics, occurrenceBiotops, extension, categoryIndexIKS, categoryIndexEV, categoryEndangermentIUCN, surname, last-Name, title, institution, workplace, street, town, workPosition, localitiesCode, locali-tiesName, biotope, cadastralArea, studies, geographicCoordinates, soilType, fertility. The assigned data type for the attributes in the f_speciesColeoptera table is listed on the left side (Object Explorer area) (Fig. 2).

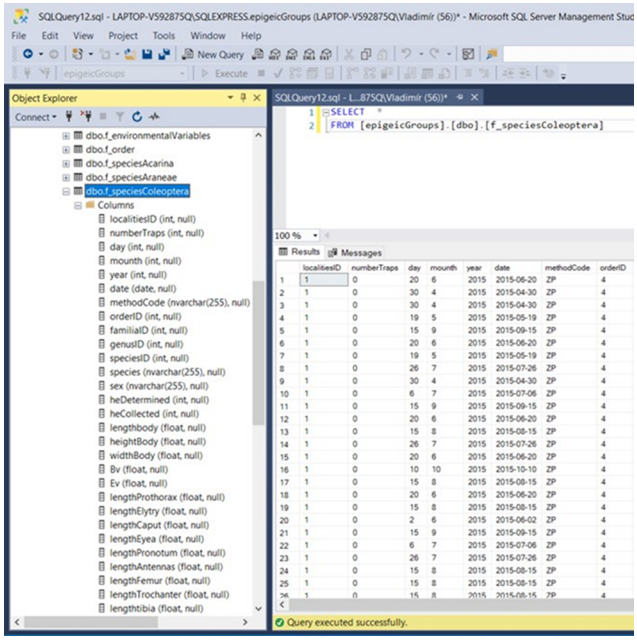

Figure 2: Sample of data type for attributes in the species Coleoptera table.

4 Conclusion

Zoology is increasingly becoming a data-rich science, and therefore, the need to store and communicate large files has grown tremendously. Databases constitute an important tool to help scientists understand and explain biological phenomena. Until now, biological databases have focused mainly on molecular biology and genetics. Our results bring the design of a new relational database for the needs of zoological research focused on epigeic groups. The new structure of the zoological database helps facilitate the analysis of metadata, in obtaining current information on zoological-ecological relationships and the response of animals to anthropogenic interventions in the environment.

Acknowledgement: This research was supported by the grants VEGA 1/0604/20 Environmental assessment of specific habitats in the Danube Plain. KEGA No. 019UKF-4/2021 Creation and innovation of education - Zoology for Ecologists, part — Invertebrates.

References

[1] http://www.sopsr.sk/web/ [online, accessed 15 June 2021].

[2] https://ibot.sav.sk/cdf/ [online, accessed 15 June 2021].

[3] https://pladias.cz/en [online, accessed 15 June 2021].

[4] https://www.biolib.cz/ [online, accessed 15 June 2021].

[5] Altschul, S., Gish, W., Miller, W., et al. Basic Local Alignment Search Tool. *Journal of Molecular Biology 215* (1990), 403–410.

[6] Benham, S., et al. *Taxus baccata in Europe: Distribution, habitat, usage and threats*. Publications Office of the EU: Luxembourg, 2016.

[7] Benson, D., Karsch-Mizrachi, I., Lipman, D., et al. GenBank. *Nucleic Acids Res 28* (2000), 15–18.

[8] Benson, D., Karsch-Mizrachi, I., Lipman, D., et al. GenBank. *Nucleic Acids Res 42* (2014), 7–32.

[9] Bernstein, F., Koetzle, T., Williams, G., et al. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology 112*, 3 (1977), 535–542.

[10] Birney, E., and Clamp, M. Biological database design and implementation. *Briefings in Bioinformatics 5*, 1 (2004), 31–38.

[11] Bourne, P. Will a biological database be different from a biological journal? *PLOS Computational Biology 1*, 3 (2005).

[12] Bourne, P. E., et al. Macromolecular crystallographic information file. In *Macromolecular Crystallography Part B*, vol. 277 of *Methods in Enzymology*. Academic Press, 1997, pp. 571–590.

[13] Bradley, A. R., Rose, A. S., Pavelka, A., et al. Mmtf—an efficient file format for the transmission, visualization, and analysis of macromolecular structures. *PLOS Computational Biology 13* (2017), 1–16.

[14] Burge, S. W., et al. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research 41*, D1 (2012), D226–D232.

[15] Clarkson, C., et al. In vitro antiplasmodial activity of medicinal plants native to or naturalised in south africa. *Journal of ethnopharmacology 92* (2004), 177–91.

[16] Dalmaris, E., et al. Dataset of targeted metabolite analysis for five taxanes of hellenic taxus baccata l. populations. *Data 5*, 1 (2020).

[17] Davidson, S., Overton, C., Tannen, V., and Wong, L. Biokleisli: A Digital Library for Biomedical Researchers. *International Journal on Digital Libraries 1* (1997), 36–53.

[18] Dawson, W., and Kawai, G. Modeling the chain entropy of biopolymers: Unifying two different random walk models under one framework. *Journal of Computer Science & Systems Biology 2* (2009), 1–23.

[19] de Lorenzo, V., et al. The power of synthetic biology for bioproduction, remediation and pollution control. *EMBO reports 19*, 4 (2018), e45658.

[20] Duggirala, S. Newsql databases and scalable in-memory analytics. In *A Deep Dive into NoSQL Databases: The Use Cases and Applications*, P. Raj and G. C. Deka, Eds., vol. 109 of *Advances in Computers*. Elsevier, 2018, pp. 49–76.

[21] Duigou, T., du Lac, M., Carbonell, P., and Faulon, J.-L. RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Research 47*, D1 (2018), D1229–D1235.

[22] Fabricant, D., and Farnsworth, N. The value of plants used in traditional medicine for drug discovery. *Environmental Health Perspectives 109* (2001), 69–75.

[23] Fazekas, D., et al. SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol 7* (2013), 7.

[24] Feld, C., et al. Indicators for biodiversity and ecosystem services: towards an improved framework for ecosystems assessment. *Biodivers Conserv 19* (2010), 2895–2919.

[25] Gharajeh, M. Biological big data analytics. In *A Deep Dive into NoSQL Databases: The Use Cases and Applications*, P. Raj and G. C. Deka, Eds., vol. 109 of *Advances in Computers*. Elsevier, 2018, pp. 1–48.

[26] Gharajeh, M. S. *A Learning Analytics Approach for Job Scheduling on Cloud Servers*. Springer International Publishing, Cham, 2017, pp. 269–302.

[27] Heink, U., and Kowarik, I. What criteria should be used to select biodiversity indicators? *Biodivers Conserv 19* (2010), 3769–3797.

[28] Hoskeri, J., Krishna, V., and Amruthavalli, C. Functional annotation of conserved hypothetical proteins in rickettsia massiliae mtu5. *Journal of Computer Science & Systems Biology 3* (2010), 50–52.

[29] Hudson, L. N., et al. The predicts database: a global database of how local terrestrial biodiversity responds to human impacts. *Ecology and Evolution 4*, 24 (2014), 4701–4735.

[30] Kashyap, H., et al. Big data analytics in bioinformatics: A machine learning perspective. *arXiv 1506.05101* (2015).

[31] Kinjo, A., et al. Protein data bank japan (pdbj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Research 45* (2017), D282–D288.

[32] Microsoft. Microsoft SQL Server. 2017: (RTM) - 14.0.1000.169 (X64) Aug 22 2017 17:04:49 Copyright (C) 2017 Microsoft Corporation Express Edition (64-bit) on Windows 10 Home 10.0 [X64] (Build 18362:).

[33] Nielsen, J., and Keasling, J. D. Engineering cellular metabolism. *Cell 164* (2016), 1185–1197.

[34] Pearson, W., and Lipman, D. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Science USA 85* (1988), 2444–2448.

[35] Pejić Bach, M., Bertoncel, T., Meško, M., Suša Vugec, D., and Ivančić, L. Big data usage in european countries: Cluster analysis approach. *Data 5*, 1 (2020).

[36] Pontén, F., Schwenk, J. M., Asplund, A., and Edqvist, P.-H. D. The human protein atlas as a proteomic resource for biomarker discovery. *Journal of Internal Medicine 270*, 5 (2011), 428–446.

[37] Ragunath, P. K., Venkatesan, P., and Ravimohan, R. New curriculum design model for bioinformatics postgraduate program using systems biology approach. *Journal of Computer Science & Systems Biology 2* (2009), 300–305.

[38] Raj, P. A detailed analysis of nosql and newsql databases for bigdata analytics and distributed computing. In *A Deep Dive into NoSQL Databases: The Use Cases and Applications*, P. Raj and G. C. Deka, Eds., vol. 109 of *Advances in Computers*. Elsevier, 2018, pp. 1–48.

[39] Rose, P. W., et al. The rcsb protein data bank: redesigned web site and web services. *Nucleic Acids Research 39* (2011), D392–D401.

[40] Shanthi, V., Ramanathan, K., and Sethumadhavan, R. Role of the cation-π interaction in therapeutic proteins: A comparative study with conventional stabilizing forces. *Journal of Computer Science & Systems Biology 2* (2009), 51–68.

[41] Singh, S., et al. Comparative modeling study of the 3-d structure of small delta anti-gen protein of hepatitis delta virus. *Journal of Computer Science & Systems Biology 3* (2010), 1–4.

[42] Srinivasa, K., and Hiriyannaiah, S. Comparative study of different in-memory (no/new) sql databases. In *A Deep Dive into NoSQL Databases: The Use Cases and Applications*, P. Raj and G. C. Deka, Eds., vol. 109 of *Advances in Computers*. Elsevier, 2018, pp. 133–156.

[43] Stratton, M., Campbell, P., and Futreal, P. The cancer genome. *Nature 458* (2009), 719–724.

[44] Toomula, N., Kumar, A., Kumar, D. S., and Bheemidi, V. S. Biological databases - integration of life science data. *Journal of Computer Science & Systems Biology 4* (2012), 87–92.

[45] Turner, V., Gantz, J., and Minton, S. The digital universe of opportunities: Rich data and the increasing value of the internet of things. Tech. rep., 2014.

[46] Vaseeharan, B., and Sivakamavalli, J. In silico homology modeling of prophenoloxidase activating factor serine proteinase gene from the haemocytes of fenneropenaeus indicus. *Journal of Proteomics & Bioinformatics 4* (2011), 53–57.

[47] Velankar, S., et al. Pdbe: improved accessibility of macromolecular structure data from pdb and emdb. *Nucleic Acids Research 44* (2016), D385–D395.

[48] Wheeler, N., et al. Effects of genetic, epigenetic, and environmental factors on taxol content in taxus brevifolia and related species. *Journal of natural products 55* (1992), 432–440.