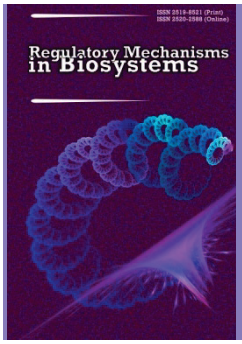

Regulatory Mechanisms in Biosystems

ISSN 2519-8521 (Print)
ISSN 2520-2588 (Online)
Regul. Mech. Biosyst.,
2022, 13(3), 207–212
doi: 10.15421/022226

Storing and structuring big data in histological research (vertebrates) using a relational database in SQL

V. Langraf*, R. Babosová*, K. Petrovičová**, J. Schlarmannová*, V. Brygadyrenko***

*Constantine the Philosopher University in Nitra, Nitra, Slovakia
**University of Agriculture in Nitra, Nitra, Slovakia
***Oles Honchar Dnipro National University, Dnipro, Ukraine
****Dnipro State Agrarian and Economic University, Dnipro, Ukraine

Article info
Received 15.06.2022
Received in revised form 17.07.2022
Accepted 18.07.2022

Constantine the Philosopher University in Nitra, Tr. A. Hlinku, 1, Nitra, 94901, Slovakia.
E-mail: langrafvladimir@gmail.com

University of Agriculture in Nitra, Tr. A. Hlinku, 2, Nitra, 94901, Slovakia.
E-mail: kornelia.petrovicova @ gmail.com

Oles Honchar Dnipro National University, Gagarin av., 72, Dnipro, 49010, Ukraine. Tel.: +38-050-93-90-788. E-mail: brigad@ua.fm

Dnipro State Agrarian and Economic University, Serhii Efremov st., 25, Dnipro, 49600, Ukraine. Tel.: +38-050-93-90-788. E-mail: brigad@ua.fm

Langraf, V., Babosová, R., Petrovičová, K., Schlarmannová, J., & Brygadyrenko, V. (2022). Storing and structuring big data in histological research (vertebrates) using a relational database in SQL. Regulatory Mechanisms in Biosystems, 13(3), 207–212. doi:10.15421/022226

Database systems store data (big data) for various areas dealing with finance (banking, insurance) and are also an essential part of corporate firms. In the field of biology, however, not much attention has been paid to database systems, with the exception of genetics (RNA, DNA) and human protein. Therefore data storage and subsequent implementation is insufficient for this field. The current situation in the field of data use for the assessment of biological relationships and trends is conditioned by constantly changing requirements, while data stored in simple databases used in the field of biology cannot respond operatively to these changes. In the recent period, developments in technology in the field of histology caused an increase in biological information stored in databases with which database technology did not deal. We proposed a new database for histology with designed data types (data format) in database program Microsoft SQL Server Management Studio. In order that the information to support identification of biological trends and regularities is relevant, the data must be provided in real time and in the required format at the strategic, tactical and operational levels. We set the data type according to the needs of our database, we used numeric (smallint,numbers, float), text string (nvarchar, varchar) and date. To select, insert, modify and delete data, we used Structured Query Language (SQL), which is currently the most widely used language in relational databases. Our results represent a new database for information about histology, focusing on histological structures in systems of animals. The structure and relational relations of the histology database will help in analysis of big data, the objective of which was to find relations between histological structures in species and the diversity of habitats in which species live. In addition to big data, the successful estimation of biological relationships and trends also requires the rapid accuracy of scientists who derive key information from the data. A properly functioning database for meta-analyses, data warehousing, and data mining includes, in addition to technological aspects, planning, design, implementation, management, and implementation.

Keywords: histology; big data; SSMS; structure database; data quality; data type.

Introduction

The development and widespread use of highly distributed systems to process big data in biology is considered as one key technological developments in bioinformatics (Silva et al., 2014). The biological sciences (zoology, botany, anthropology, genetics) are faced with storing a large amount of research data. Advances in methods and experimental research increase the amount of data, which needs to be stored and consequently analysed. In the past biological data was stored in simple databases created in Access or Microsoft excel programmes. These databases were insufficient for storing big data and therefore large-scale databases in the Structural Query Language (SQL) program began to be created (Kashyap et al., 2015).

Biological databases use the following types of databases: relational databases in SQL and object-oriented. Despite limitations of the type of database, the researchers understand the structure and output of data from the database (Baxevanis, 2011). The language used for processing big data is SQL, which is used in relational models (Shanthi et al., 2009). Data stored in biological databases have their own data format (datatype), the diversity of which leads to problems in the implementation and conversion of data formats into software (Bernstein et al., 1799).

Pluralism in biological terminology occurs in the different processes and definitions used in the methodology (Leonelli, 2012). These problems can be solved using interoperable databases, for better integration of data from different sources, for subsequent use in different parts of the research (The Gene Ontology Consortium, 2019). The connection of data sets (big data) must be in the concept of real processes of nature. The connection of concepts in big data infrastructures is considered as a possibility by which we can see the biological world and the functioning of nature in an ecological context. This perception guides scientific reasoning and the direction of research in meta-analyses and accounts for new discoveries (Leonelli, 2017). Standardized tools used in big data generation and subsequent meta-analyses are designed to meet research methodologies and research designs (Bogen, 2008).

The data quality in databases on the internet may be lower, because they are not often curated by specialists in the scientific field. Many biological databases have low reliability, specification only for a certain area of research and therefore the data cannot be easily transferred to other fields of research. Based on the above mentioned facts, data quality losses occur and extensive database links are often disrupted by unreliable sources during online data collection (Illari & Floridi, 2014). Access to original datasets enhances big data implementation, which improves

replication of experiments and reuse of data. It also serves as a connection between research methodology and approaches on the internet (Leonelli & Ankeny, 2012; Dietrich et al., 2014). Recognising the original dataset and structure of big data helps us to improve research conclusions with metanalysis. This recognition is important when extrapolating data. The relational view in biological databases may be described as the removal of general approaches in the direction of context-sensitive datasets (Leonelli &Tempini, 2018).

Big data mining in biological research indicates the direction and structure of research (Nickles, 2018). Databases allow data mining to help scientists explore trends and recurring patterns in natural events that are verified by hypotheses (Pietsch, 2015; Ratti, 2015; Canali, 2019). The relational database shows that correctly interpreted results of biological research depend on approaches to original models, confrontation with scientific methods and correct contextualization of the big data (Shavit, 2009; Elliott et al., 2016). The structure of the relational database, documents of historical big data for the use of future users to judge the implementation according to their own standards, are also very important for proper data analysis. Automated data analysis and creation of algorithms increases the need for critical thinking among scientists. Collaboration between data scientists and researchers will help in the correct setting of algorithms, meta-analyses and subsequent interpretation of the results. Such cooperation will help in the correct setting of relational relations, which will capture trends in nature. It will relevantly evaluate the origin of the data in the datasets, which will ensure proper criticism in other research questions (Sterner & Franz, 2017).

We expect that our results will contribute to a new information on the structure and storing of biological databases in the field of histology. This new information will enrich the field of bioinformatics dealing with big data and databases.

Materials and methods

For the concept and design of our relational database (biological database) for histological research, we used the database program Microsoft SQL Server Management Studio 2017 (RTM-14.0.1000.169 (X64). Edition (64-bit) on Windows 10 Home 10.0 [X64]). The syntax of writing queries when uploading data uses the Microsoft platform.

Results

The relational database proposed by us contained histology research data. The histology database was made up of the following types of tables: 8 code, 3 dimension and 11 frequency, the connection between the tables is ensured by primary and foreign key (Fig. 1). Repeated writing of big data based on intervals had frequency tables which store measurements of structures in tissue and contains data of systems, classis, species. The tables were as follows: f_systemaSceleti, f_systemaSensuum, f_systemaRespiratorium, f_systemaNervosum, f_systemaMusculorum, f_systemaDigestorium, f_systemaCardiovasculare, f_organaUropoetica, f_organaGenitalia, f_integument, f_glandulaeEndocrinae. Tables that had one entry for an element, capturing its attributes, were dimensional. In our database, these were the tables: d_species, d_pageRole, d_biotops. Code tables serving as dials (class, tissue, part body) were represented by the following tables: cl_zoneBone, cl_typeSystem, cl_typeBoneTissue, cl_tissue, cl_partBody, cl_method, cl_classificationHC, cl_class. A schematic of all the tables in the program SSMS is shown in the Figure 2.

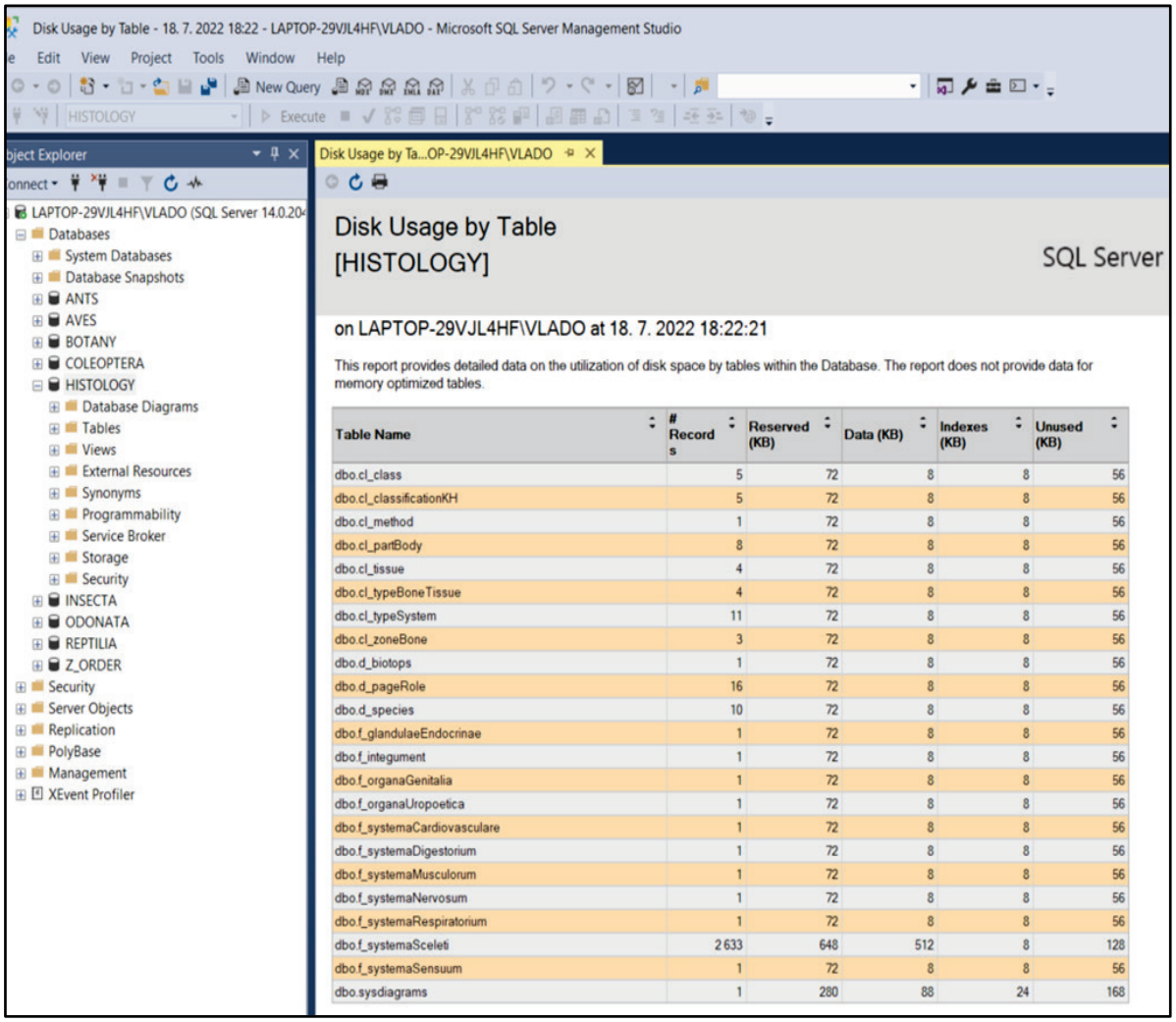

Table Name	# Records	Reserved (KB)	Data (KB)	Indexes (KB)	Unused (KB)
dbo.cl_class	5	72	8	8	56
dbo.cl_classificationKH	5	72	8	8	56
dbo.cl_method	1	72	8	8	56
dbo.cl_partBody	8	72	8	8	56
dbo.cl_tissue	4	72	8	8	56
dbo.cl_typeBoneTissue	4	72	8	8	56
dbo.cl_typeSystem	11	72	8	8	56
dbo.cl_zoneBone	3	72	8	8	56
dbo.d_biotops	1	72	8	8	56
dbo.d_pageRole	16	72	8	8	56
dbo.d_species	10	72	8	8	56
dbo.f_glandulaeEndocrinae	1	72	8	8	56
dbo.f_integument	1	72	8	8	56
dbo.f_organaGenitalia	1	72	8	8	56
dbo.f_organaUropoetica	1	72	8	8	56
dbo.f_systemaCardiovasculare	1	72	8	8	56
dbo.f_systemaDigestorium	1	72	8	8	56
dbo.f_systemaMusculorum	1	72	8	8	56
dbo.f_systemaNervosum	1	72	8	8	56
dbo.f_systemaRespiratorium	1	72	8	8	56
dbo.f_systemaSceleti	2 633	648	512	8	128
dbo.f_systemaSensuum	1	72	8	8	56
dbo.sysdiagrams	1	280	88	24	168

Fig. 1. A schema of the histology database

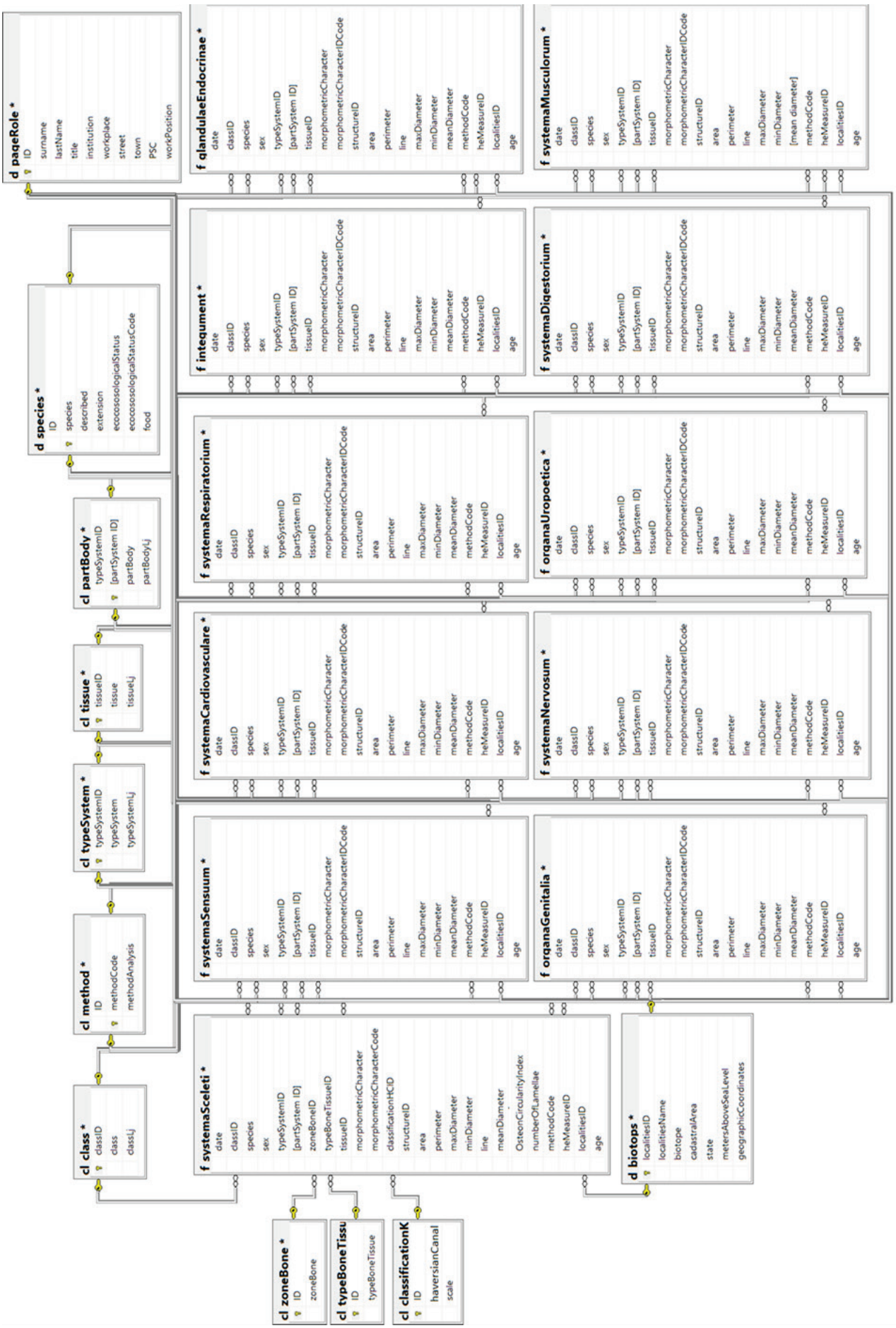

Fig. 2. Biological database proposal for histology research

The frequency tables contained data with histology attributes assigned to individual systems. The tables with biotopes was filled with the characteristics of biotopes with the names of locations. Data about scientific workplaces and the scientists themselves were stored in a dimension table d pag-eRole. The IDs assigned to individual entities of the given element were in code lists. The connection between individual tables was made using the primary key (PK; yellow key) and the foreign key (FK; infinity symbol).

Each datum in the database was assigned a data type (data format). Numeric data with datatype "smallint" had ID attributes: ID, classID, typeSystemID, partSystemID, tissueID, typeSys-temID, localitiesID. Numbers had an "int" data type for variables PSC, age, numberOfLamellae. Data type float had area, perimeter, maxDiameter, minDiameter, line, meanDiameter, eccentricity, OsteonCircularityIndex. The date was written in the standard format YYYY-MM-DD. Data type nvarchar with variable length had the data stored in a text string where the columns belonged: class, classLj, HaversianCanal, scale, methodAnalysis, methodCode, partBody, partBodyLj, tissue, tissueLj, typeBoneTissue, typeSystem, typeSys-temLj, street, zoneBone, localitiesName, title, biotope, cadastralArea, state, geographicCoordi-nates, surname, lastname, workPosition, street, town, workplace, workPosition, species, sex, described, extension, ecozoologicalStatus, ecozoologicalStatusCode, institution, food,morphometricCharacter, morphometricCharacter-Code,town, structureID. The mentioned code below presents the entry of data types for individual columns in the f_systemaSceleti table.

```
CREATE TABLE [dbo].[f_systemaSceleti]
([year] [int] NULL,
[classID] [smallint] NOT NULL,
[species] [nvarchar](50) NULL,
[sex] [nvarchar](5) NULL,
[typeSystemID] [smallint] NOT NULL,
[partSystem ID] [smallint] NOT NULL,
[zoneBoneID] [smallint] NULL,
[typeBoneTissueID] [smallint] NULL,
[tissueID] [smallint] NULL,
[morphometricCharacter] [nvarchar](50) NULL,
[morphometricCharacterID] [nvarchar](5) NULL,
[classificationHCID] [smallint] NULL,
[structureID] [nvarchar](10) NULL,
[area] [float] NULL,
[perimeter] [float] NULL,
[maxDiameter] [float] NULL,
[minDiameter] [float] NULL,
[line] [float] NULL,
[meanDiameter] [float] NULL,
[OsteonCircularityIndex] [float] NULL,
[numberOfLamellae] [int] NULL,
[methodCode] [nvarchar](5) NULL,
[heMeasureID] [smallint] NULL,
[age] [int] NULL);
```

Research data was written into tables in the database using the INSERT INTO function. Based on the USE [HISTOLOGY] function, we queried a database and then after the INSERT INTO function, we queried a table into which we will insert the data. When entering data, we must have queries attributes (columns) and define the inserted values with the VALUES function. An example of the syntax of inserting a single element was given in the code below.

```
USE [HISTOLOGY]
GO
INSERT INTO [dbo].[f_systemaSceleti]
([date],[classID],[species],[sex],[typeSystemID],[partSystem],[ID],[zoneBoneID],[typeBoneTissueID],[tissueID],[morphometricCharacter],[morphometricCharacterCode],[classificationHKID],[structureID],[area],[perimeter],[maxDiameter],[minDiameter],[line],[meanDiameter],[eccentricity],[OsteonCircularityIndex],[numberOfLamellae],[methodCode],[heMeasureID],[age])
VALUES (2022,90,'Martes foina', 'F',1,8,9,1,2,'Haversian canals','HK',4,'E1',188.9,49.3,8.6,7, 15,6,-0.228571428571428, 0.97617517455328, 20, 'PM',2,10)
GO
```

Data stored in Excel, text documents or csv must be imported using the tasks and import data option, which is included in the Microsoft SQL Server Management Studio program itself (Fig. 3). During the successive steps by which we import external data, we also set data tips using the Edit Mappings function (Fig. 4). After importing data from external sources, the program warns with a report about the correct loading of data and the number of records (elements, Fig. 5).

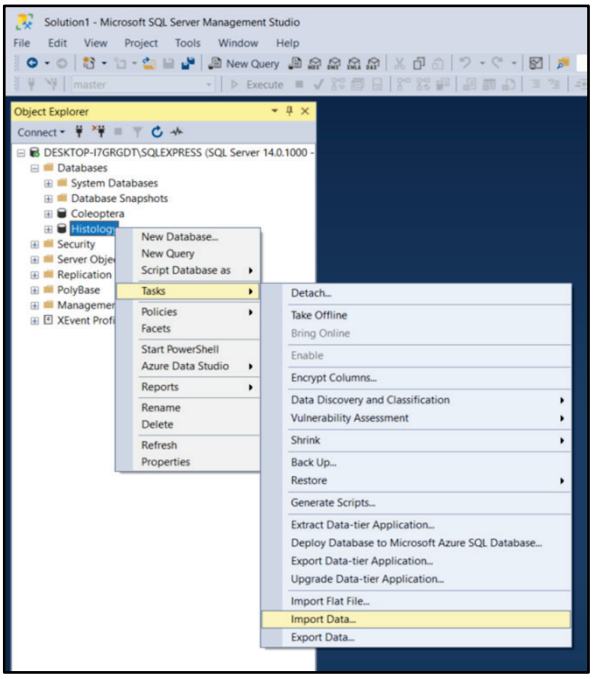

Fig. 3. Importing data using the Tasks option

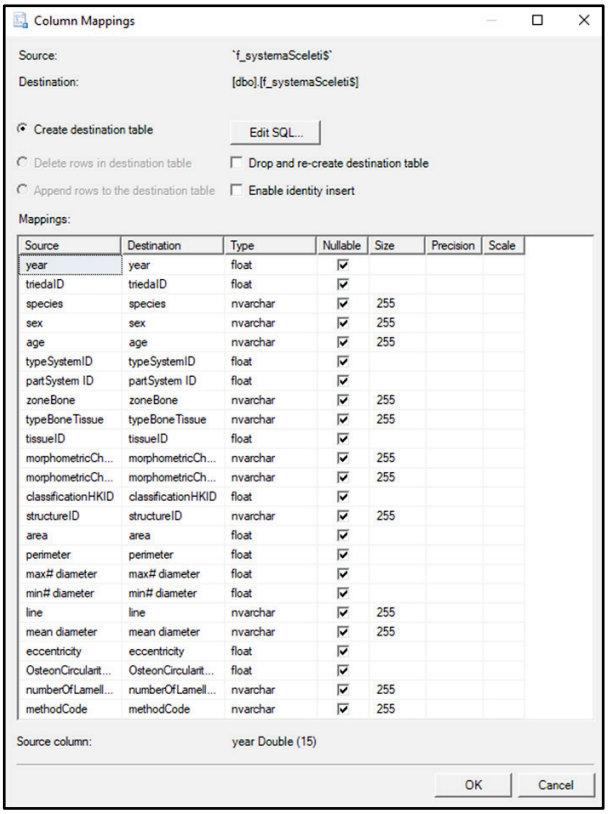

Fig. 4. Editing data types in Edit Mappings

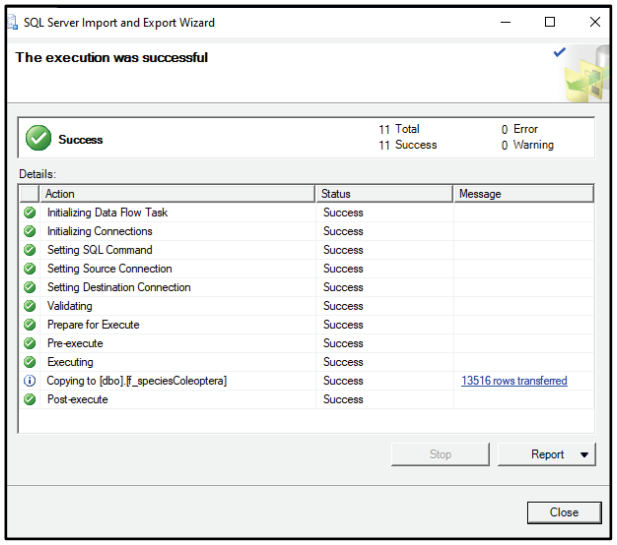

Fig. 5. Report on completed data import

Discussion

Understanding the information stored in databases is a key fact for choosing the right database structure and its subsequent implementation. It is also important for data mining and interpretation of results (Leonelli, 2020). Currently in existence there are functional databases in which the datasets are for DNA, RDA (Birney, 2004), cancer-focused diseases (Bourne, 2005), structure protein (Benson et al., 2014; Dalmaris, 2020) and and bio economies (Bradley, 2017; Burge, 2013). In our work, we presented a new structure of the relational database and its filling with data from the field of histology. The necessity of such a database grows with the information obtained by research, and at present such a type of database does not exist. Establishing such a database, this shortcoming would be solved and it would contribute to new information in the field of bioinformatics.

A biological database has an organized structure for proper data storage, easy retrieval and updating of datasets (Duigou et al., 2019). The primary database contains sequences and structure, which are further derived into the secondary database and used in meta-analyses (Fazekas et al., 2013; Gharajeh, 2018). Each database must have an appropriately chosen data model, which can be, for example, Oracle SQL, SQL, E-R (IE = Information Engineering), DDL. In our relational database, we used the E-R model, which was designed in the database program SSMS.

The aggregation queries have the data of individual attributes stored in columns, and the search works on the basis of subsets, which speeds up the processing of the result (Duggirala, 2018; Srinivasa & Hiriyannaiah, 2018). Individual types of tables in our histology database (code, frequency, dimension) have data of attributes also stored in columns, and relationships are linked using primary and foreign keys, which also speeds up the result of queries. The speed of aggregation queries is also affected by data compression in the algorithms used in the procedures. The above mentioned fact also accelerates the availability of the metanalysis result (Benson et al., 2014; Raj, 2018). In the early days of using databases, "flat files" (ASCII) were used. It is a type of text file containing subset, nested indexes and datasets. Nowadays, this type is still used on the basis of better data compatibility during the implementation of datasets. In the 1970s, the PDB was among the first data formats, which was used in a protein database. The flat files format was modified to PDBx/mmCIF (Kinjo et al., 2017) during 2014 and replaced by the binary MMTF (Macro-Molecular Transmission Format) format (Gharajeh, 2017; Pejić et al., 2020). Our suggested data types are "int" for integers, "smallint" for small-valued numbers, "float" for floating-point numbers, "date" for storing dates, "nvarchar" for storing text. All mentioned data types are supported by the Microsoft and Oracle platforms together with MYSQL. A properly designed database structure that interprets biological relationships and trends is very important for fast and reliable meta-analysis of biological data. It is also very important for implementation and communication between datasets for functional use of trend predictions of biological data (Feld et al., 2010; Sarita et al., 2010; Canali, 2019). Our histology database is designed for metadata analysis with the objective of identifying responses of histological structures in species in relation to the diversity of habitats in which species live.

Conclusions

Biological databases are often filled with histological data that need to be stored and properly implemented when using meta-analyses. Well-structured histology databases help scientists understand biological processes. Until now, attention has been focused mainly on data storage in the field of molecular biology and genetics. In our contribution, we created a new database structure for storing histological research data and set the correct data types for fast communication during data mining. The structure of the histological database created by us will ensure the integrity of the data during meta-analyses, which will contribute information on responses of histological structures in species in relation to diversity of habitats in which species live.

This research was supported by the grants VEGA 1/0604/20 Environmental assessment of specific habitats in the Danube Plain. KEGA No. 002UKF-4/2022 Metaanalyzes in Biology and Ecology (Databases and Statistical Data Analysis).

References

Baxevanis, A. D. (2011). The importance of biological databases in biological discovery. Current Protocols in Bioinformatics, 34(1), 111–116.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2014). GenBank. Nucleic Acids Research, 42, 32–37.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. Journal of Molecular Biology, 112(3), 535–542.

Birney, E. (2004). Biological database design and implementation. Briefings in Bioinformatics, 5(1), 31–38.

Bogen, J. (2008). Experiment and observation. In: Machamer, P., & Silberstein, M. (Eds.). The Blackwell guide to the philosophy of science. Blackwell Publishers Ltd. Pp. 128–148.

Bourne, P. (2005). Will a biological database be different from a biological journal? PLoS Computational Biology, 1(3), e34.

Bradley, A. R., Rose, A. S., Pavelka, A., Valasatava, Y., Duarte, J. M., Prlić, A., & Rose, P. W. (2017). MMTF – an efficient file format for the transmission, visualization, and analysis of macromolecular structures. PLOS Computational Biology, 13(6), e1005575.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., & Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. Nucleic Acids Research, 41, 226–232.

Canali, S. (2019). Evaluating evidential pluralism in epidemiology: Mechanistic evidence in exposome research. History and Philosophy of the Life Sciences, 41, 4.

Dalmaris, E., Avramidou, E. V., Xanthopoulou, A., & Aravanopoulos, F. A. (2020). Dataset of targeted metabolite analysis for five taxanes of hellenic *Taxus baccata* L. populations. Data, 5(1), 22.

Dietrich, M. R., Ankeny, R. A., & Chen, P. M. (2014). Publication trends in model organism research. Genetics, 198(3), 787–794.

Duggirala, S. (2018). NewSQL databases and scalable in-memory analytics. Advances in Computers, 109, 49–76.

Duigou, T., du Lac, M., Carbonell, P., & Faulon, J. L. (2019). RetroRules: A database of reaction rules for engineering biology. Nucleic Acids Research, 47, 1229–1235.

Elliott, K. C., Cheruvelil, K. S., Montgomery, G. M., & Soranno, P. A. (2016). Conceptions of good science in our data-rich world. BioScience, 66(10), 880–889.

Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekö, M., Lenti, K., Farkas, I. J., Vellai, T., Csermely, P., & Korcsmáros, T. (2013). SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. BMC Systems Biology, 7(1), 7.

Feld, C. K., Sousa, J. P., da Silva, P. M., & Dawson, T. P. (2010). Indicators for biodiversity and ecosystem services: Towards an improved framework for ecosystems assessment. Biodiversity and Conservation, 19(10), 2895–2919.

Gharajeh, M. S. (2017). A learning analytics approach for job scheduling on cloud servers. In: Peña-Ayala, A. (Ed.). Learning analytics: Fundaments, applications, and trends. Springer, Cham. Vol. 94. Pp. 269–302.

Gharajeh, M. S. (2018). Biological big data analytics. Advances in Computers, 109, 321–355.

Illari, P., & Floridi, L. (2014). Information quality, data and philosophy. In: Floridi, L., & Illari, P. (Eds.). The philosophy of information quality. Berlin, Springer. Pp. 5–23.
Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K. (2015). Big data analytics in bioinformatics: A machine learning perspective. Journal of Latex Class Files, 13(9), 1–20.
Kinjo, A. R., Bekker, G. J., Suzuki, H., Tsuchiya, Y., Kawabata, T., Ikegawa, Y., & Nakamura, H. (2017). Protein Data Bank Japan (PDBJ): Updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Research, 45(1), 282–288.
Leonelli, S. (2012). When humans are the exception: Cross-species databases at the interface of biological and clinical research. Social Studies of Science, 42(2), 214–236.
Leonelli, S. (2017). Global data quality assessment and the situated nature of "best" research practices in biology. Data Science Journal, 16, 32.
Leonelli, S. (2020). Scientific research and big data. In: Edward, N. Z. (Ed.). The Stanford encyclopedia of philosophy. Stanford University, Stanford.
Leonelli, S., & Ankeny, R. A. (2012). Re-thinking organisms: the impact of databases on model organism biology. Studies in History and Philosophy of Science, 43(1), 29–36.
Leonelli, S., & Tempini, N. (2018). Where health and environment meet: The use of invariant parameters in big data analysis. Synthese, 198(10), 2485–2504.
Nickles, T. (2018). Alien reasoning: Is a major change in scientific research underway? Topoi, 39(4), 901–914.
Pejić Bach, M., Bertoncel, T., Meško, M., Suša Vugec, D., & Ivančić, L. (2020). Big data usage in European countries: Cluster analysis approach. Data, 5(1), 25.
Pietsch, W. (2015). The causal nature of modeling with big data. Philosophy and Technology, 29(2), 137–171.
Raj, P. (2018). A detailed analysis of NoSQL and NewSQL databases for big data analytics and distributed computing. Advances in Computers, 109, 1–48.
Ratti, E. (2015). Big data biology: Between eliminative inferences and exploratory experiments. Philosophy of Science, 82(2), 198–218.
Sarita, S., Kumar, G. S., Anuradaha, N., Sanjay, K., Rajendra, N., Kishore, S. P., & Kumar, P. K. (2010). Comparative modeling study of the 3-D structure of small delta antigen protein of hepatitis delta virus. Journal of Computer Science and Systems Biology, 3(1), 47.
Shanthi, V., Ramanathan, K., & Sethumadhavan, R. (2009). Role of the cation-π interaction in therapeutic proteins: A comparative study with conventional stabilizing forces. Journal of Computer Science and Systems Biology, 2(1), 51–68.
Shavit, A., & Griesemer, J. (2009). There and back again, or the problem of locality in biodiversity surveys. Philosophy of Science, 76(3), 273–294.
Silva, Y. N., Dietrich, S. W., Reed, J. M., & Tsosie, L. M. (2014). Integrating big data into the computing curricula. In: SIGCSE '14: Proceedings of the 45th ACM technical symposium on computer science education. Machinery, Ney York. Pp. 139–144.
Sterner, B., & Franz, N. M. (2017). Taxonomy for humans or computers? Cognitive pragmatics for big data. Biological Theory, 12(2), 99–111.
The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still going strong. Nucleic Acids Research, 47(1), 330–338.