Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241061

Organizing an Interdisciplinary Platform for Knowledge Sharing on a Class of Compounds of Natural Origin

Ylenia MURGIA[a,1], Valeria IOBBI[b], Angela BISIO[b], Nunziatina DE TOMMASI[c] and Mauro GIACOMINI[a]

[a] *Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy*

[b] *Department of Pharmacy, University of Genova, Viale Cembrano 4, 16148 Genova, Italy; Sea Study Centre, University of Genova, Via Balbi 5, 16126, Genova, Italy.*

[c] *Department of Pharmacy, Via Giovanni Paolo II, 132, 84084 Fisciano, Salerno, Italy.*

ORCiD ID: Ylenia Murgia https://orcid.org/0009-0009-3303-3160, Valeria Iobbi https://orcid.org/0000-0003-4883-4840, Angela Bisio https://orcid.org/0000-0002-7559-7732, Nunziatina De Tommasi https://orcid.org/0000-0003-1707-4156, Mauro Giacomini https://orcid.org/0000-0001-5646-2034

Abstract. Sesterterpenoids, a subset of the terpene family, exhibit notable biological activities. These natural compounds are present in a variety of organisms such as plants, fungi, bacteria, insects and marine life. The therapeutic potential and structural diversity of sesterterpenoids have attracted considerable interest in pharmacological and chemical research. This study illustrates the development of a database to structure and manage data on these compounds. The design process involves the collection of user requirements, creation of a conceptual model with and Entity-Relationship Diagram (ERD), development of a logical model, and implementation in Microsoft SQL Server 2022. Data collection began with an extensive literature review and organization in an Excel spreadsheet. The resulting database improves data acquisition, organization, and accessibility. Future work will include building a website to facilitate data entry, editing, reading and extraction, and automation of data updates via external web services.

Keywords. Sesterterpenoids, Relational Database, Interdisciplinary Platform

1. Introduction

Sesterterpenoids are a relatively small class of natural compounds that belong to the large family of terpenes. Specifically, sesterterpenoids are pentaprenyl terpenoids that originate from the linear precursor geranylfarnesyl diphosphate. Theses compounds have been identified in a variety of organism, including plants, fungi, bacteria, insects, and marine organisms, and are known for their biological properties, such as antimicrobial, anti-inflammatory, and antitumor activities. The growing interest in sesterterpenoids is

[1] Corresponding Author: Ylenia Murgia; Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS) - University of Genoa, Via all'Opera Pia 13, 16145 Genoa, Italy; E-mail: ylenia.murgia@edu.unige.it.

due to their therapeutic potential and structural diversity, making them valuable for pharmacological and chemical research [1,2]. As we can observe from various previous studies [3,4], to effectively organize and share knowledge, it is essential to arrange data in a database, to facilitate data collection, data access and collaboration among researchers. Database design is a crucial part of the informatics process, and, usually, the design phase includes four main steps [5]: User Requirements, i.e. gathering data and outline design specifications; Conceptual Design, i.e. translating requirements into a graphical data model, often using an Entity-Relationship Diagram (ERD); Logical Design, i.e. converting the ERD into tables and optimizing them through normalization to avoid redundancies and data anomalies; Physical Design, i.e. defining database parameters and implementing the design in a specific database management system.
The aim of this work is to describe the database design process, from collecting requirements and creating the conceptual model to developing the physical model, in order to structure and organize data related to sesterterpenoids. This comprehensive approach ensures that data are efficiently acquired, accurately organized, and easily accessible, thus facilitating research and analysis in the study of these natural compounds.

2. Materials and Methods

2.1. Description of data collection

This study began with an extensive literature review focusing on articles describing the extraction of sesterterpenoids from a wide range of organisms, including sponges, mollusks, bacteria, fungi, plants, and insects. At first, a team of experts collaborated to compile and enter data on these natural compounds into an Excel spreadsheet. The dataset was organized into multiple sheets with different columns depending on the type of organism from which each compound was extracted. Despite these divergences, some columns were common to all types of organisms. These shared columns, which provide a standardized framework for the data set, are shown in Table 1.

2.2. Database design

After a meticulous requirements analysis, to create a structured and efficient database, we first "translated" the columns in the Excel file into a conceptual model, using an ERD. The ERD serves as a model, illustrating the entities, their attributes, and the relationships between them, which are essential to understand the structure of the data and ensuring that they meet the user requirements. Following the ERD, we developed the logical model for the relational database. This involved creating tables for each entity defined in the ERD and implementing primary and foreign keys to establish relationships between the tables. The database was created using a relational Database Management System (DBMS). In particular, the Microsoft SQL Server 2022 DBMS was selected and installed on a server operating with Microsoft Windows Server 2022.

2.3. Nomenclature preservation

To preserve the graphical peculiarities of the nomenclature rules for both compounds and organisms, such as the use of italics, in the database, we applied a macro function to specific columns in the Excel spreadsheet. This macro can recognize the graphic peculiarity and apply appropriate HTML tag.

Table 1. Main columns of the Excel spreadsheet, shared by all organisms.

Column Name	Description	Example
Code	Code associated with the natural compound.	TeC-158
Compound name	Name of the natura compound.	19-deoxyscalarin
Notes to the name of the compound	Such notes are useful when the name of the compound is not specified in the chosen reference	Reported as compound 1
Chemical class	Chemical class the compounds belongs to.	Tetracarbocyclic sesterterpenoids (TeC)
Taxon name as reported in the publication	Taxonomic name of the organism from which the natural compound was extracted as given in the article. It may differ from the accepted species name.	*Glossodoris pallida*
Reference	Reference related to the article. It includes authors' last name and year of publication.	Rogers et al., 1991
Taxonomy-related columns: Kingdom, Phylum, Class, Order, Family, Genus, Species	These columns specify the classification of the organism according to the current taxonomic system	Kingdom: Animalia Phylum: Mollusca Class: Gastropoda Order: Nudibranchia Family: Chromodorididae Genus: *Glossodoris* Species: *Glossodoris pallida*
Collection site and Native range	The collection site column refers to the place where the organism was found, while the native range column refers to its native habitat	Collection site: Guam Native Range: Pacific

3. Results

3.1. Conceptual model

The results of this study outline a comprehensive database design process aimed at structuring and organizing data related to sesterterpenoids and the associated organisms from which these natural compounds are extracted. Starting with a deep analysis of the requirements, the study identified the need for 26 entities and 37 relationships to form the ERD structure. These entities are essential to capture the full range of information about sesterterpenoids, organisms, and publication references. Some entities are interconnected through specific relationships to highlight the interactions and dependencies that exist within the dataset. In Figure 1 we can find a simplified version of the ERD, where only the most relevant entities and relationships with their attributes are present. Below is a simple description regarding the relationship between the shared columns (Table 1) in the initial dataset and the major entities in the ERD:

- Compound: it includes attributes related to the "Code" and "Compound name" columns. The "Notes to the name of the compound" column is not directly included in the Compound entity because it is also partially dependent on the bibliographic reference. In addition, the "Chemical class" column is present in the form of a foreign key and is related to an entity of the same name given the 1:N relationship;

- Organism: it includes items belonging to the “Taxon name as reported in the publication” column;
- Bibliographic Resource: it is the entity related to the “Reference” column, but it also contains foreign keys related to “Compound” and “Organism” entities, since the same organism can produce more than one compound and the same compound can be extracted from more than one organism (N:N relationship). This entity contains the elements that are part of the “Notes to the name of the compound” column;
- Taxonomy: this macro entity actually consists of multiple entities that serve to represent the hierarchical rules of the taxonomy. Therefore, this macro entity refers to “Kingdom”, “Phylum”, “Class”, “Order”, “Family”, “Genus”, and “Species” columns;
- Geography: both elements that are part of the “Collection site” column and the “Native range” column are included in this entry.

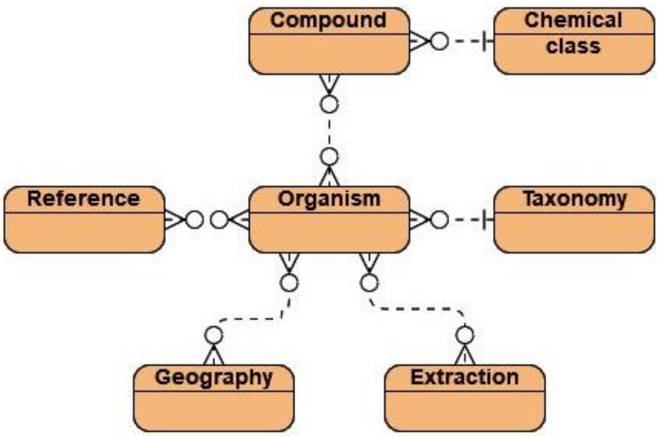

Figure 1. Simplified version of the ERD: only the major entities and relationships are shown.

3.2. Logical and physical model

The conceptual ERD was used to implement a specific schema for SQL Server 2022 DBMS. This step involved converting entities and their relationships into 36 tables, converting attributes into columns, defining data types for each column, and defining primary and foreign keys to ensure data integrity and efficient querying. Primary keys uniquely identify each record within a table, while foreign keys establish relationships between tables to maintain referential integrity. We also applied normalization techniques to eliminate redundancies and minimize data anomalies, optimizing the structure for reliability and performance.

4. Discussion and Conclusions

Through this project, we significantly improved the structure and accessibility of sesterterpenoids and related organisms by developing a comprehensive database. By converting the initial data from an Excel spreadsheet into a well-defined relational database, we enhanced the organization, retrieval, and analysis of this information. This structured approach ensures that researchers can access and use the data efficiently, thus facilitating advanced research and analysis in the field. The next step will be to develop a dedicated website to further improve data accessibility. This website will provide an intuitive interface to make it easier for authorized people to enter, edit, extract, and view data, making the information even more user-friendly and accessible to researchers and other interested parties. Preserving the graphic peculiarities of the nomenclature has already been implemented in this project, with a view to future use via website. This ensure that when the data will be displayed online, the names of organisms and compounds will be formatted accurately and correctly. Organizing and storing data from bibliographic references in a relational database offers the advantage of obtaining well-structured data. However, when it comes to organisms, this work can quickly become obsolete, as the taxonomy of some organisms can change rapidly. Therefore, it is imperative to find an automatic way to indicate whether the current status of a name is still in use or has become obsolete. In the future, consideration will be given to link the platform to existing external web services, such as the World Register of Marine Species (WoRMS) [6]. This type of linkage would help maintain the accuracy of organism names and synonyms, which change frequently. By incorporating these automated and integrative approaches, the future platform can remain a valuable and reliable resource for researchers.

References

[1] Li K, Gustafson KR. Sesterterpenoids: Chemistry, biology, and biosynthesis. *Natural Product Reports* 2021; 38: 1251–1281. doi: 10.1039/D0NP00070A

[2] Kinghorn AD, Falk H, Gibbons S, et al. *Progress in the Chemistry of Organic Natural Products*. Available from http://www.springer.com/series/10169.

[3] Giacomini M, Pastorino L, Soumetz FC, et al. Data Modeling for Tools and Technologies for the Analysis and Synthesis of NANOstructures (TASNANO) Project. *Journal of Information Technology Research* 2009; 2: 49–70. doi: 10.4018/jitr.2009070104

[4] Giacomini M, Bisio A, Giacomelli E, et al. Data collection and advanced statistical analysis in phytotoxic activity of aerial parts exudates of Salvia spp. *Revista Brasileira de Farmacognosia Brazilian Journal of Pharmacognosy*; 21: 856–863. doi: 10.1590/S0102

[5] Thompson CB, Sward K. Modeling and teaching techniques for conceptual and logical relational database design. *J Med Syst* 2005; 29: 513–525. doi: 10.1007/s10916-005-6108-3

[6] WoRMS [webservice]. Available from *https://www.marinespecies.org/aphia.php?p=webservice*.