

Creating a Medicinal Plant Database

Carsten Kettner#, Harald Kosch*, Margit Lang*, Janine Lachner*, Doris Oborny*
and Erich Teppan*

#Beilstein-Institut
Frankfurt/Main, Germany

<http://www.beilstein-institut.de/englisch/1024.htm>

*Universität Klagenfurt
Klagenfurt, Austria

<http://www-itec.uni-klu.ac.at/~harald/>

Abstract:

This paper presents the experiences of creating the information system MEDPHYT which is built to collect data on the complete European pharmaceutical and toxicological plant world whose representatives are determined by medical and therapeutic benefit. Focus of the database content is the plant with description of their botanical characteristics, and history of discovery of therapeutic use, etymology, and synonyms. Apart the botanical characterisation there is information on both medical relevant biochemical compounds and their physicochemical characteristics, and toxicological as well as pharmaceutical facts. These data sets determine the basic system of MEDPHYT.

1. Introduction

Substances derived from plant material have been used since time in memorial for medicinal purposes. For example, plants such as garlic, sage and chamomile are well known for their positive effects of maintaining and restoring health. However, many medicinal plants (e.g. the purple foxglove, *Digitalis purpurea*) also possess toxicological properties which are largely dependent on the concentration of the active principle. In the case of the purple foxglove the biochemically active compound digitalis on one hand increases the muscle contraction ability of the heart, which is a desired (indicated) effect, but on the other hand it reduces the pulse rate significantly which is an unwelcomed side effect since this can clearly lead to death. Furthermore, it is generally not common knowledge that acetylsalicylic acid, an active ingredient of many commercially produced pain-killers was originally extracted from the cortex of the willow (*Salix spec.*). Thus, there is a plethora of native (i.e. European) plants that have a broad spectrum of both toxicological and pharmacological effects that are dependent on both dosage and application of the substances contained in the plant.

However, medicinal plants have not lost their significance as natural alternatives to synthetic drugs even in the modern world. Since the number of commercially available drugs derived from plant sources is increasing, there is a distinct need for an easy accessible data collection that provides detailed scientific information on such plants and drugs either extracted from or contained in these natural sources for both the layman and the expert. While a number of excellent texts concerning medicinal plants exist, their content must be viewed as inadequate since they represent a backward view which can only be updated by regular revision. This is also largely true of most databases with free access. In addition, these data collections are organised as flatfiles and have no capability for complex queries with Boolean operators. It should also be noted that there are no non-commercial collections available which combine text with multimedia data in an object-relational database system. Interest in pharmaceutical databases has been ongoing for more than 15 years [Horace 85]. However, our survey of free access online databases indicates that they provide only very basic information on medicinal plants found in Europe (e.g., Phytochemical DB-list¹ or Madaus DB²) and tend to focus on discovery and application of plant drugs (e.g., [Axel 97]) rather than on the comprehensive botanical data of the plants.

MEDPHYT, a new medicinal plant database, hopes to improve the current situation. This distributed web-based information system stores, organizes and provides combined text and multimedia data on

¹ <http://www.ars-grin.gov/duke/>

² <http://www.madaus.com/cd/article/homepage.jsp>

botanical, toxicological, biochemical as well as pharmacological properties of medicinal plants. In particular, from the multimedia point of view, this database will use a variety of possibilities for interactive 3D visualization of chemical compounds and, in further extensions, information about the physiological mode of action of compounds contained in these plants. From the technical point-of-view our system presents a 3-tier architecture, an Oracle DBMS was used as back-end and Tomcat was employed as Application Server.

2. The MedPhyt Application

Overview: In the scope mentioned above, we designed MEDPHYT as a distributed information system relying on object-relational database technology. Our system is not restricted to special scientific user groups and opens to experts as well as home users interested in medical plants. Therefore, the access to this database is free of charge, but registration is required to be able to gather ideas about the user's reason why to access to MEDPHYT. The content of the database is designed in such a way that scientific users and researchers can benefit, as well as the interested home users. The demands on these different applications have been realized by introducing two user main roles, one for the scientific user and one for the home user. The scientific user may be interested in botanical characteristics, pharmacology and toxicology, whereas the home user has a more general context approach. By using different weights for the queries, users can personalize MEDPHYT to ones own requirements (Table 1).

„Home-User“	Doctors/Pharmacists	Research: Pharmacologists, Biologists, Chemists, Medics	Teaching (university and school): Students and Teachers
<ul style="list-style-type: none"> • botanical interest • botanical guide • additional to the instruction leaflet to phytopharmaca • toxicological relevance: diagnosis, first aid, • historic data 	<ul style="list-style-type: none"> • additional to the instruction leaflet to phytopharmaca • documentation for e.g. teas, oils,... • effects • contraindications • side effects • personal training 	<ul style="list-style-type: none"> • drug-targeting • target research • drug-simulation • simulation of synthetic components in metabolic paths • parameterisation of targets and drugs • cell modelling • (new-) development of plant drugs 	<ul style="list-style-type: none"> • application in: phytochemistry, physiology, biochemistry, molecular biology and biotechnology • representation of e.g. cellular toxication /detoxication • botanical guide

Table 1: Potential User groups

Content Organization: Fig. 1 gives an overview of the content organization with the main sections: Botany, History, Bio/-Chemistry, Toxicology, and Pharmacology. The Botany section contains data about the plant systematic (the position of a species within the plant world), descriptions and geographic distribution as well as brief comments on their pharmaceutical and toxicological meaning for therapeutic use. Images and coloured drawings help here to identify the text-described plant species. The Bio/-Chemistry section provides information about the chemical properties of the main compounds, representations of chemical structures in two- and three-dimensional ways as well as descriptions and schematic drawings of the branch of the plant metabolism in which these compounds are synthesised. The section Pharmacology contains tables with data about the molecular and cellular mechanism of drug effect, data about the pharmacokinetic properties of the plant drug and data about the use of this drug in commercial products.

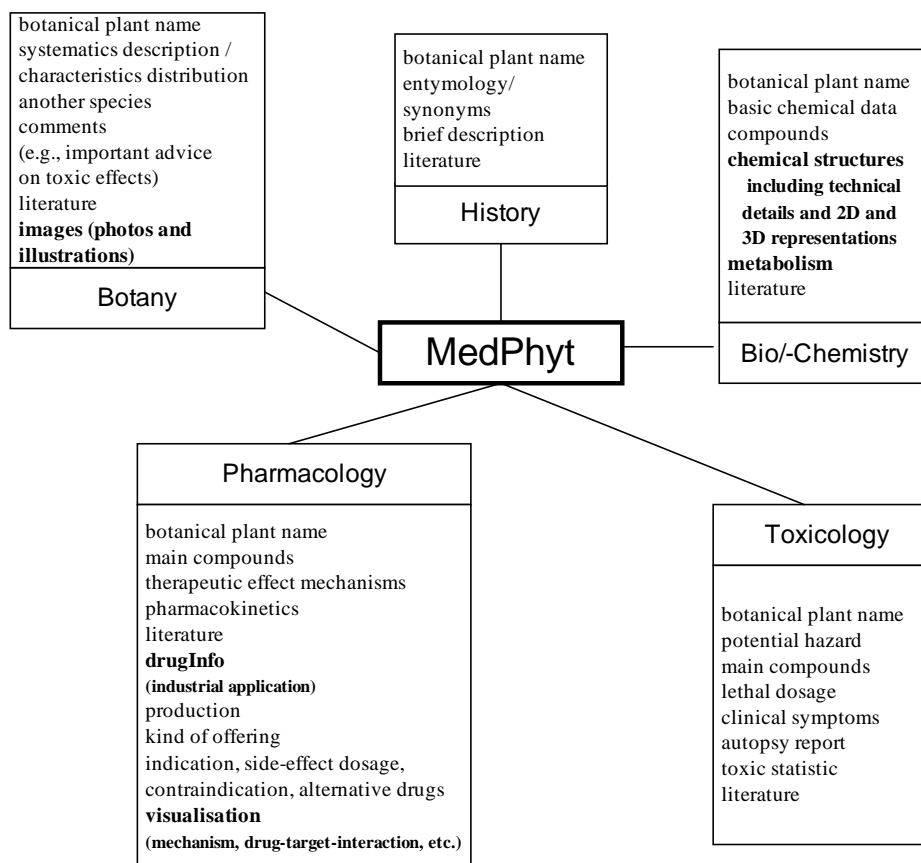


Figure 1: Simplified Organisational diagram of the database

Query Mechanisms: MEDPHYT offers three different access and search mechanisms: simple query, complex query and browsing.

The *Simple Search* allows to search the whole database for one or more terms (separated e.g., by the word "and") in a google-like manner. In case of hits, the results are presented as a summary (see Figure 3 below for a sample of the result size).

The *Expert Search* is designed for users who want to specialise their search by defining what parts of the database should be searched and how this should happen. It offers different search sections that are divided into subsections.

Systematics	Searchterm:	<input type="text"/>	included in:	Gender <input type="text"/>
Botany	Searchterm:	fox	included in:	Plantname <input type="text"/>
Biochemistry	Searchterm:	digitalis	included in:	Contents <input type="text"/>
History	Searchterm:	<input type="text"/>	included in:	Entymology <input type="text"/>
Toxication	Searchterm:	<input type="text"/>	included in:	Symptom <input type="text"/>
Lethal	Searchterm:	<input type="text"/>	included in:	Statistics <input type="text"/>

QueryExpression:

Figure 2: Expert Search

Search fields can be combined by a Boolean expression. By default all the search fields are combined by a logical "AND". In the example above, we are looking for plants which name contains "fox" and

which main substance of content is "digitalis". More powerful expressions are entered below by specifying the sections which have to be combined. In case of a successful search, a summary of the plants found and an explanation for the hit is given. The name of the plant links to the detailed information site.

[Purple Fox Glove \(Digitalis Purpurea\)](#)

Notice:

very harmful, are all parts of the plant are toxic, dry loading or cooking does not inactivate the toxins, all fox gloves are very toxic, very bitter taste

Dangerousness:

very toxic

Explanation [\[hide\]](#)

- Plantname contains 'fox'
- Clinic Symptom contains 'diarrhoea'
- chemical contents contains 'digitalis'

Figure 3: Search Result

In addition, the MEDPHYT query functionality provides an alphabetical browsing mechanism as the easiest way to gain first insight in the content of the database. Either by scrolling through the entire plant list or by clicking the demanded letter in the alphabetical strip on top of the browsing site, the user will see the corresponding plant list and short information on each plant species.

3. Application and Technical Aspects

3.1. Biology and Chemistry Aspects

Apart from the inappropriate situation that free accessible on-line medical plant databases are mainly flat file text collections called monographs, these data presentations usually do not consider the technical means and advantages of object-relational databases, as well as of Web technology. Information is mainly presented in huge text fields with only little structure. Search is provided only as full-text search. Therefore a top-down approach has been chosen for the development of this new database, from the general anticipated presentation of data down to the content itself.

Thus, from the application point-of-view, the challenge was the realisation of a *scientific* object-relational database. We started from the view of the specialist and designed the database from the application aspect. The implementation of the web-based prototype also was driven by the intuition of the application aspects and not as a typical database project affected by IT requirements. The specification put the main emphasis on usability, relationality and kind of content. The simple usability is intended to be supported by a plain user interface which is comparable to that of Google. The relationality of the database is proposed to provide complex queries through all tables and fields, whereas the content is planned to differ significantly from common on-line medicinal plant databases regarding the multimedia data presentation. The first draft considering these aspects has been implemented by the use of a FileMaker prototype to demonstrate the basic concept of data representation. On the basis of this prototype, a complex ER schema has been generated which resulted in creation of two main tables, botany (Species) and biochemistry (ChemicalCompound). By linked interdependencies between these main tables and further related tables the focus of the database has been realized to be on physiology, bio- and phytochemistry, molecular biology, analytics as well as the visualization of the mode(s) of action of the active principle. Slight adjustments of the final database structure were necessary for at least two reasons: on one hand the varieties of table relations were very complex and has initially been underestimated and on the other hand the development of the insert application required changes to the structure to ensure data integrity.

Data on medicinal plants derive from primary and secondary literature which is excerpted by scientists. Originally, during the first step so called monographs, flat file text documents, are created

from which appropriate data are extracted and entered in the insert masks. Due to the time spent for the extraction of data from the monographs which has to be done even by a further scientist, this workflow will be changed to an input of excerpted data in a flat file document. This file corresponds to the structure of the insert masks and is provided for correction by the input control before the data can be entered in the insert masks by any person. Ongoing developmental work is concerned with the use and implementation of bio- and cheminformatic tools to visualize the drug binding and drug-targeting on the molecular level. The integration and application of such tools is known to be of particular interest of both biologists and pharmacologists.

One of the proposed features in MEDPHYT is the 3D presentations of chemical structures. In order to display the chemical information in an appropriate way, two important questions had to be answered:

1. Which freely available Editor is best for our issues?
2. Which format of chemical structures (in terms of presentation and implementation) can/should be used in connection with the database (Object-relational database)?

We tested some chemical editors, and decided to work with Marvin³ (kindly provided by Chemaxon⁴ for free of licence fees for academic use). With this java-based tool, editing and viewing chemical structures in the web-application is as intuitive as with a stand-alone application. Furthermore, it smoothly integrates with the other applications. Generally, Marvin supports several file formats like SMILE, Pdb, Mol and CML (and many more), but none format was suitable for the integration in MEDPHYT.

The selection of the right file format required to examine the existing description formats for molecules. The decision was dependent on both the capabilities of the description language and the way of administration and storage of the description file in the database. Marvin Sketch (the chemical editor) can export the molecule definition as a text files, thus we decided to insert these structures in the database as a CLOB (Character Large Object). However, not all of the above mentioned file formats can be used. The formats Mol, CML and PdB, had to be post-processed after retrieval as CLOB objects from the database before viewed in MarvinView (the chemical viewer). Only SMILE provided complete functionality for direct import and export and was therefore chosen as format.

Another aspect we had to take care of was the appropriate way to display and arrange all desired information on the website. The content spans over several sections from biology to chemistry. Moreover, our application is intended to be used by home users as well as not only experts. So we decided to split the information into different sections: Overview, Biochemistry, History, Toxicology and so on. Some of the content cannot be exactly matched to one of those categories and it was decided to display these parts redundantly in different sections.

3.2 Technical Aspects

Based on previous experiences, we used an Oracle database as DBMS backend (to be obtained from <http://technet.oracle.com>). We decided to employ the Oracle 9i second release. It is the latest reliable version of Oracle's database system suite. Recent attempts to migrate to Oracle 10g failed due to important memory consumption problems of this new release.

The insertion process of the medicinal plants was session-based. The user enters the different elements into a session first. After the successful control of the dataset, the complete plant data is stored permanently in the database. Update functionality is provided by a separate tool, available to the administrator only, which displays the different tables and content for editing.

A main focus was put on multimedia elements. We introduced interfaces to upload images and videos for the plant and chemical compounds, .e.g. for the molecular effects. The initial idea was to hold these data in the database and to retrieve them with the help of the Oracle *interMedia* tools⁵. However, it was technically impossible to retrieve and then show jpeg and gif images from the database without

³ <http://www.chemaxon.com/marvin/>.

⁴ <http://www.chemaxon.com/>

⁵ <http://www.oracle.com/technology/products/intermedia/index.html>

storing them intermediately in the file system. The provided JDBC functionality together with the library was not fully functional. Therefore, we decided to maintain the application images in special directories on the application server.

Another important issue was the choice of the appropriate server technology. The whole project is implemented in java and the Apache Tomcat servlet container was used as application server. In the beginning it seemed to be a good idea to use the Apache Webserver in coexistence with Tomcat, as many other web-based information system's projects do. However, since a closer look revealed, that the Tomcat server includes all the capabilities of the Apache Webserver (in terms of security configuration, serving static content, etc.) and only a few static pages were used (for which the Apache Webserver is optimized due to performance), there was no further reason to keep running the Apache Webserver.

Another important question was the installation of security techniques. We installed a SSL/TLS connection using RSA encoding with a 2048 bit key. The concept for user authentication is based on form-based authentication in addition to a mysql-database-realm. The developed small log-in system includes besides the authentication a user registration and an administration application

4. Conclusion and Outlook

Although with MEDPHYT it is intended to improve the current situation of the electronic presentation of pharmacologic and toxicologic data by introduction of 3D visualisation capabilities, the primary aim of this database is not to substitute currently present medicinal plant databases but to be a supplement since the focus of MEDPHYT differs significantly from such databases available on-line. However, the here presented medicinal plant database MEDPHYT did not reach its final developmental endpoint. By contrast, this system is called a basis system which provides all necessary interfaces for further extensions such as comprehensive 3D visualisation techniques, implementation of the system for mobile systems and expansion of the content. The first and second aspects require some additional work on both new scientific approaches to the action of mode of the drugs on molecular level and the database system MEDPHYT. The third aspect only requires the addition of further tables for animals, bacteria and fungi. However, even now it is easily possible to insert plant data from other parts of the world without changing the data structure. This indicates that there is a great potential for the MEDPHYT database system to play a paramount role in the world of natural compound databases with free access.

References

- [Horace85] Horace D. Brown: A drug is born: its formation facets in pharmaceutical research and development. *Journal of Chemical Information and Computer Sciences* 25(3): 218-224 (1985)
- [Axel97] Melinda G. Axel and Il-Yeol Song: Data Warehouse Design for Pharmaceutical Drug Discovery Research. DEXA Conference Workshop 1997: 644-650. IEEE CS Press.