# A TREE-BASED APPROACH TO MODELLING MEDICAL RESEARCH DATA. COMPARISON WITH EXISTING APPROACHES IN COMMERCIAL APPLICATIONS

**PÉTER OLÁH[1], MARIUS MĂRUŞTERI[2], CARMEN BIRIŞ[3] MARIA MIHAELA OPRIŞ[4], DANIELA DOBRU[5]**

*[1]PhD candidate University of Medicine and Pharmacy Tîrgu-Mureş, [2,3,4,5]University of Medicine and Pharmacy Tîrgu-Mureş*

***Keywords:*** *medical databases, healthcare information systems, user interface, electronic laboratory notebook*

***Abstract:*** *In our past work, we have developed a data structuring technique based on trees and multitrees to model clinical research data. In this paper, we try to assess the suitability of this model for keeping an electronic laboratory notebook by comparing its strengths and weaknesses with the products of some of the leading software providers on the market. The comparison focuses on five critical aspects: customizability, the use of tree-like structures, the use of patterns or templates, the techniques used to retrieve the data and functions that support collaboration. In the last two sections of the paper, we discuss our findings and draw some conclusions that will serve as guidelines for our future work.*

***Cuvinte cheie:*** *baze de date medicale, sisteme informatice în domeniul medical, registru electronic de laborator*

***Rezumat:*** *În lucrările noastre anterioare am dezvoltat o tehnică de structurare a datelor bazată pe arbori şi multiarbori pentru a modela date din cercetarea clinică medicală. În acest articol încercăm să evaluăm în ce măsură acest model este utilizabil pentru a dezvolta un registru electronic pentru experimente de laborator, comparând punctele tari şi punctele slabe ale acestuia cu produsele câtorva dintre ofertanţii de pe piaţa de software. Această comparaţie se concentrează pe cinci aspecte critice: adaptabilitate, utilizarea structurilor de tip arbore, folosirea de şabloane, interogarea datelor şi funcţii de colaborare în grup. În ultimele două secţiuni ale articolului vom prezenta câteva discuţii referitoare la rezultatele acestei comparaţii şi vom trage câteva concluzii care vor avea un rol în orientarea eforturilor noastre viitoare privind dezvoltarea sistemului.*

## INTRODUCTION

In our past work, (1) we have developed a data structuring technique to model clinical research data in the field of Gastroenterology using trees and multitrees.(2) our efforts to test the portability of this model to other fields of research, we have tried to adapt the application to laboratory research in the field of Bioengeneering and Regenerative Medicine. Here we have found the well established concept of Electronic Laboratory Notebook (ELN), which is intended to replace the paper notebooks most scientists still use. It is worth mentioning that a simple ELN does not cover the whole information requirements of a complex research project. The project in which we are testing our model also encompasses theoretical research, and proof of concept using an animal model.

## OBJECTIVES

In this paper, we try to assess the suitability of our model for keeping an electronic laboratory notebook and compare the strengths and weaknesses with some of the leading software providers in the market. Our comparison will be focused on the basic structures and functions that allow the researcher to model his/her own data rather than on communication and security features.

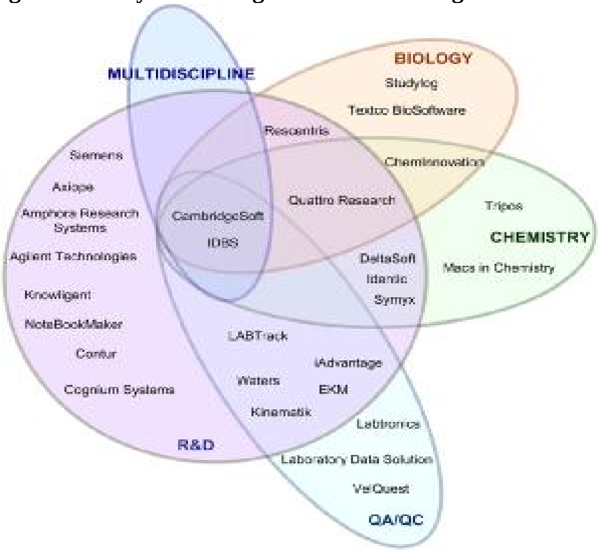
The analysis will be qualitative, focusing on two main objectives:

- Validating the concepts that we have used in our approach
- Gathering new requirements in order to improve our system to meet the industry standards

## METHODS

As a starting point for selecting the systems to be included in our comparison we have used the list and categorization compiled by Rubacha et al. (3) (see figure no. 1).

**Figure no. 1. Systems categorization according to Rubacha**




[1]Corresponding author: Péter Oláh, Str. Ghe. Marinescu, Nr. 50, Cod 540136, Târgu Mureş, România, E-mail: olah_peter@yahoo.com, Tel: +40265 212111

Article received on 09.11.2012 and accepted for publication on 17.12.2013
ACTA MEDICA TRANSILVANICA March 2013;2(1):172-174

The final list was constructed based on the following criteria:

- Only systems that are considered to be tailor made for R&D by the authors of the original categorization (3) were included;
- The solutions which have no working websites at the time of our study were eliminated;
- The list was further narrowed by selecting only the applications that provide screenshots or a demo version directly on their websites.

Thus, the systems that we have evaluated for our comparison are:

- Agilent Technologies' OpenLAB ELN
- Axiope's e-CAT
- Cognium Systems' iPad ELN

The features that we have looked for are:

- Customizability – can the user define new structures without the intervention of the software developer?
- Use of trees – are tree-like structures used in structuring the data?
- Patterns - are there any predefined structures and/or can patterns be constructed by the user?
- Retrieving data – is it done by structured queries or general purpose search engines?
- Collaboration – is there collaborative functionality implemented in the systems?

Furthermore, we have analyzed in detail the structuring of data from two different points of view:

- Administrative data - what structures are used to organize the experiments that the scientists are conducting over time?
- Scientific data - how is the data structured for each experiment?

The same analysis was conducted on the prototype of our own system in order to compare it's functionality to those on the market.

## RESULTS

Using the above mentioned criteria we have found some similarities but also a number of differences between the analyzed applications. In this section we will present our findings.

**Agilent Technologies' OpenLAB ELN**

The central concepts of the system's build-up are **templates** and **forms**.

The *customizability* of the system is supported by the use of „dynamic templates" and „dynamic forms" to provide on-demand structuring for the patterns and the scientific data stored by the user.

*Trees* are used to structure the administrative data of the system, such as the opened windows, the existing templates or forms. The scientific data is defined within the forms themselves, thus not having a tree-like structure.

*Patterns* are represented by the templates used to model the workflows of the researchers. They can be configured by associating forms to them and re-used as needed.

*Data retrieval* is done using a search engine, augmented by a structured search instrument to find experiments in various stages of completion or review.

*Collaboration* is achieved via the WEB by configuring virtual teams as needed.

**Axiope's e-CAT**

The system's design revolves around the concept of creating and storing **records**.

The system is *customizable* by the use *of* a series of templates to structure both the administrative data and the scientific data. The template list is one and the same though for both categories, making the distinction between them somewhat blurry.

The administrative part of the system is structured using *trees*. A structure called the Record Tree organizes Projects, Users and Other data (ex. Supplies). Lists and fields are preferred to structure the scientific data.

The system uses *patterns* as it ships with over 20 templates, with the possibility to define additional ones.

*Data retrieval* is unstructured, but the system features a powerful search engine with the possibility to refine searches by field type, project or field name.

*Collaboration* is WEB enabled. Sharing can be set up so that the Principal Investigator has access to all group members' non-private records.

**Cognium Systems' iPad**

This application is a Content Management System, which allows the creation and management of XML documents, that has been adapted to the domain of experimental research. Being XML based, the system relies heavily on the use of tags.

Revolving around the „document" and „tag" concepts, the system can categorize virtually anything by tagging the data in various ways, thus being completely *customizable*.

The application uses *trees* to organize the administrative data in a Document Browser, but also to navigate in the structure of the experiments. The tags used constitute a hierarchy which can be navigated using trees. There is also a functionality to create links between tags, suggesting the building of a graph type structure.

Searching for *patterns*, we have determined that three pre-defined templates are used: research notes, project notes, and generic document. The generic document template allows the user to create new patterns as needed.

*Data retrieval* is supported by a complex search engine which converts the search criteria to the corresponding tags, in order to search in the database.

*Collaboration* of the system's users is achieved by accessing external repositories.

**The Tree/Multitree based approach**

The central concepts in our approach are trees and **multitrees**.

The *customizability* in this approach derives from the design of the database itself. The structure of the scientific data is not modelled intentionally, but extensionally, thus being changeable by simply adding or modifying records in the metadata part of the database.

*Trees* and *multitrees* are used to model the structure of scientific data rather than the structure of the administrative part of the system. Every type of experiment can be structured as a tree. Different variations of the same experiment will become sub-trees of a multitree. If necessary, nodes from different trees can be linked, resulting in a more complex structure, known as a Directed Acyclic Graph (DAG).

*Patterns* are an inherent part of the system's design. Every time a high level user defines a new structure for a new type of experiment, a reusable pattern is created. Patterns can also be transferred from one structure to another by copy/paste. It is a special type of copy and paste that will replicate the structure of a sub-tree starting from a specific node.

*Data retrieval* is implemented primarily by structured queries. These are complex interrogations of the underlying database that make use of the tree structures that the user has already defined.

*Collaboration* is for the time being limited to the users of the same server, as the system is not WEB enabled. In its current form the collaboration is possible by granting specific access rights to users or user groups.

## DISCUSSIONS

In this section we will focus the discussions on the suitability of the tree based model for implementing an ELN for academic research. It should be mentioned that the requirements of academic users differ from those of the chemical and pharmaceutical industry.(4)

In our model flexibility is considered paramount, so the system is highly *customizable* by the end-user. The analysis of the other three applications from above supports the idea that customizability is a key factor for a successful ELN.

*Tree*-like structures are ever-more present in Graphical User Interfaces (GUI) in a variety of systems. In the analyzed ELNs this type of structure was used primarily to organize the various types of records (experiments, patterns, data types etc.). None of these approaches used trees to model the structure of the scientific data. The GUIs that offer data entry functionality to the researcher generally use some sort of object aggregation. These objects can be viewed as highly specialized building-blocks that can be put together to represent the structure or flow of an experiment. These building-blocks can be combined in limitless ways offering a high degree of flexibility. However the data collected in this manner does not appear to be structured according to the rules of a relational database. By comparison, the tree based model offers similar flexibility, but keeps the data highly structured, supported by a normalized relational database at the back-end of the system.

Research often implies repeating the same steps, so the need for *patterns* is inherent. All of the above described systems offer some sort of functionality for defining patterns or templates. While this function is optional in these applications, in our approach it is embedded in the design of the system. Thus a researcher cannot record an experiment before defining a pattern for it. This can constitute a drawback in some situations, but on the long term it will assure the proper structure of all data entered in the system.

In terms of *data retrieval*, the search engine seems to be the overall norm in the commercial applications. Our system does not feature one yet, but it can be developed by implementing full text searches on the database.

Some of the search engines that we have analyzed are very complex, using several layers of search refinement. However, the complexity of the search algorithms cannot substitute the accuracy of queries run against structured data. This type of data retrieval offers the user the possibility to target specific elements of information and to expect precise results. This type of querying is in our opinion the major strength of the tree based approach.

The analysis of the commercial ELNs has also revealed that *collaboration* is a key factor in conducting successful scientific experiments. Furthermore, collaboration has to occur regardless of the distance between the members, so web enabled systems clearly set the trend in this regard.

## CONCLUSIONS

The analysis of three commercially available ELN systems has revealed that customizability, the use of patterns/templates and collaboration are the most sought after functions in this type of application.

The use of trees to model the scientific data meets these requirements and, in our view, presents some clear advantages related to data retrieval. However, this approach has not been fully proven yet.

**Aknowledgement:**

*This paper is partly supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU 64331 and POSDRU 60782.*

## REFERENCES

1. Olah P, et al. A database design pattern for structuring hierarchical medical data. Acta Medica Marisiensis, 2013. to be published in vol. 58 no. 6.
2. Furnas GW, Zacks J. Multitrees: Enriching and Reusing Hierarchical Structure, in ACM CHI 94 Human Factors in Computing Systems; 1994. p. 330-336.
3. Rubacha M, Rattan AK, Hosselet SC. A review of electronic laboratory notebooks available in the market today. J Lab Autom; 2011;16(1):90-8.
4. Rudolphi F, Goossen LJ. Electronic laboratory notebook: the academic point of view. J Chem Inf Model. 2012;52(2):293-301.