Design and Development of an Information System to Manage Clinical Data about Usher Syndrome Based on Conceptual Modeling

Verónica Burriel, M. Ángeles Pastor, Matilde Celma, J. Carlos Casamayor, Laura Mota
Centro de Investigación en Métodos de Producción de Software PROS
Universitat Politècnica de València
Valencia, Spain
<vburriel, mapastor, mcelma, jcarlos, lmota>@pros.upv.es

Abstract — **The inefficient management of clinical data in many research environments is a problem which slows down the service provided to patients. The benefits of an Information System created following the conceptual modeling rules have been proved in multiple environments with data management difficulties. The main hurdle to overcome is the large gap between the language and concepts employed by informaticians and the ones used by biologists. The work described in this paper shows how these technologies can also be applied to the clinical domain, after a long period of mutual approaching in order to understand each other. The research clinical data of an expert research group on Usher syndrome have been studied, analyzed and redesigned using conceptual modeling, helping this group to offer a better service.**

Keywords-information system; usher syndrome; database; conceptual modeling.

I. INTRODUCTION

Usher syndrome is a condition characterized by hearing loss or deafness and progressive vision loss. The loss of vision is caused by an eye disease called *retinitis pigmentosa* (RP) which affects the layer of light-sensitive tissue, at the back of the eye (the retina). Vision loss occurs as the light-sensitive cells of the retina gradually deteriorate. Usher syndrome is thought to be responsible for 3 percent to 6 percent of all childhood deafness and about 50 percent of deaf-blindness in adults. This disease is estimated to occur in at least 4 per 100,000 people.

Mutations in the CDH23, CLRN1, GPR98, MYO7A, PCDH15, USH1C, USH1G, and USH2A genes cause Usher syndrome. These genes provide instructions for proteins that play important roles in normal hearing, balance, and vision. They have influence in the development and maintenance of hair cells, which are sensitive cells in the inner ear that help to transmit sound and motion signals to the brain. In the retina, these genes are also involved in determining the structure and function of light-sensitive cells called rods and cones. In some cases, the exact role of these genes in hearing and vision is unknown. Most of the mutations responsible for Usher syndrome lead to a loss of hair cells in the inner ear and a gradual loss of rods and cones in the retina. Degeneration of these sensitive cells causes hearing loss, balance problems, and vision loss characteristic of this condition.

For research groups related to this disease, having data properly stored and classified is very important in order to access them easily. Furthermore, it is possible to detect relationships between the data if it is correctly structured and linked. This would help to improve the understanding of the disease.

However, despite the advances in information technologies, nowadays there are still research groups in health which store their data in spreadsheets or simply sheets of paper. The most advanced groups use simple databases to store information about their patients, but most of these databases are focused on the solution-space. Despite that it is not the most appropriate solution, these tools are a first approximation to a solution, but they lack a previous conceptual scheme. This problem entails poor management of the involved data and the consequent loss of quality of information and simplicity of operation. The conceptual modeling ensures the adequacy of the stored data, improving their usefulness and maintenance, aspects currently seldom taken into account in the clinical domain. The main difficulty for solving this deficiency, following our experience, is the enormous distance between the concepts and languages used by the scientists from the life areas in front of these used by the technics in informatics. This problem has been highlighted every time when our group has shared a work with biologists or doctors.

The Genoma research group of PROS (*Centro de Investigación en Métodos de Producción de Software*) is a research group with expertise in Information Systems and Bioinformatics, which is working for a while in the design of a model to represent all the knowledge acquired so far about the genomic domain using conceptual modeling technologies. This is the main step to create a Genomic Information System with a database capable of storing comprehensive genomic information [1-3]. This knowledge about genomics and Information Systems is the perfect environment to solve the problem described above.

In following sections, the information system developed will be explained in detail. In section II, similar solutions to different departments will be exposed. In section III, the current system of data storage used by the clinicians before installing this information system is detailed. The advantages of having an information system modeled by a conceptual scheme are reasoned in section IV. The conceptual scheme created to develop this information system is deeply explained in section V. In section VI and VII, the database implemented and the loading and managing processes are defined. Finally, the conclusions and future work are described in detail in section VIII of this paper.

II. State of the Art

Despite being well-known that Health Information Systems have a great potential to improve quality of clinical services and reduce costs, only about 17 percent of doctors and 8 percent to 10 percent of U.S.A. hospitals use electronic medical records. Most medical institutions do not want to risk installing these systems due to lack of formal evaluations and evidence regarding its successful implementation [4] .

However, there are some groups that have implemented Information Systems for a clinical environment with satisfactory results. In University Medical Centre Ljubljana (Slovenia) a clinical information system is used to support medication process (prescribing, ordering, dispensing, administration and monitoring) and offer participating medical teams real time warnings and key information regarding medications and patient status, thus reducing medication errors [5].

Other example of the success of Health Information Systems is installed in the University Cancer Centre in Frankfurt (Germany). The hospital information system gives physicians the possibility to access to all patient information in a hospital. Furthermore, a special query and reporting tool has been integrated in the health information system to recognize patients with a specific disease and with basic inclusion and exclusion criteria for a specific clinical trial [6].

Instead, this progress of health information systems has not been so relevant in Information Systems with genomic information. Predefined online databases have been used to store information about the patients in these situations, instead of personalized health information systems. One example of this situation is the locus-specific database for mutations in GDAP1 gene created in the INSERM of Angers (France). It helps in the analysis of genotype-phenotype correlations in Charcot-Marie-Tooth diseases type 4A and 2K [7]. A complete health information system can offer much more services than a simple database and could be more useful in this purpose.

Due to the different ways of working in every health and research centre, a specific information system adapted to every center or centers with similar needs will be the easier way to introduce information systems in these domains without changing too much their current working methods.

III. Current System of Data Storage

As an example of this problem, the storage of data from a center for research in neurosensory diseases has been analyzed [8-10]. This center studies patients with Usher Syndrome coming from every hospital in Spain, gathering together a lot of data from this disease. To obtain all these data, they use several methods: questionnaires, which are given to the patients and relatives to be fulfilled to obtain certain data about the severity of the disease, and genetic analysis in order to obtain some objective data from the blood sample of a patient.

When a patient arrives to the hospital with signs and symptoms of suffering Usher Syndrome, a first diagnosis is done by the doctor, who creates a diagnosis report with these signs and symptoms. During the same visit, a questionnaire is given to him or, in case he is not able to fill it, to his parents. Next, a sample of blood is extracted from the patient in order to be analyzed genetically. Sometimes, some genetic analyses are also performed to the relatives, looking for the disease genetic heritage. These samples are analyzed using a sequencer which directly produces the results in electronic support. All the information obtained from the diagnosis report, questionnaires and genetic analyses are used by the clinicians to make the definitive diagnosis of the disease suffered by the patient.

The information extracted from these processes is stored using different methods. The questionnaires and diagnosis reports are filed in physical folders sorted by families and following an alphabetical order. On the other hand, the genomic data is saved in Excel files and one Access table without any kind of structure behind it.

Once the questionnaires have been filled by the patients and their relatives on a set of sheets of paper, introducing all this information into an information system is a hard work that requires a lot of time, and the clinicians are too busy to do it. If these questionnaires had been done by computer directly, the information would have been introduced instantly into a database. The same problem appears with the diagnosis reports. This solution would help a lot to the clinicians and would save a lot of time and space.

Sorting the information by families is a peculiar classification which is very common when information about a genetic disease is stored. It is very important and relevant to know the genetic profile of the members of the family of the patient, because they have a high probability of suffering the same disease or being carriers of it.

Another problem found among the data is the lack of formalism. This means that information is repeated in multiple places creating redundancy. Often some data are missed or represented using cryptic codes without clear semantics.

This case study is not an isolated case. There are many research groups or clinical departments that do not store their data in an information system based on conceptual models.

IV. Advantages of Conceptual Modeling

Nowadays, the most advanced approximations of Information Systems development, which are oriented towards producing quality systems, propose the use of conceptual model-based methodologies[11]. Conceptual modeling is widely used in the Information Systems field because it helps developers in the understanding and description of the problem domain before implementation. In this way, conceptual modeling improves the developed system helping to manage its evolution in the future [11, 12]. For a long time, conceptual modeling techniques have been used successfully to build IS in many different domains [13].

Quality clinical practice must necessarily be based on a data set of maximum reliability and accurate interpretation. To make this possible, it is necessary to develop conceptual schemas that characterize the information to be stored, used and modified to ensure its continued updating.

To create this conceptual schema, a strong knowledge of the domain is strictly necessary. To deal with this problem we were helped by the clinicians from the Center for Research in Neurosensory Disease, who explained us the domain in great detail, enabling us to understand every piece of stored information and validating our interpretation of the domain. After the understanding of the domain, all the information has been correctly represented in a conceptual schema by our group..

From the conceptual schema, the corresponding database schema will be generated. This schema, then, will be used to create a database to correctly store the data ensuring their quality. This will provide a system to store and access data in a much easier and faster way, allowing much more complex queries on these data than before. Having all the information linked into this information system, the possibility of discovering the relation between data and improving the treatments for the patient is more feasible.

The new Information System will improve the quality of information stored by the clinicians and consequently, the service provided to the patient, as well as the time saving for doctors in the management and recovery of the information.

V. Conceptual Scheme Description

The conceptual schema shown in Figure 1 represents all the information handled by the Center for Research in Neurosensory Group. To understand it easily, this section explains in detail its contents. The information about the persons studied in its laboratory is represented in this schema by *Person* class. There is a relevant attribute of *Person* to take into account, *family_id*, which connects the relatives. This attribute is very useful to analyze samples of new members of the same family, who can have the same variations. Other relevant attribute is *origin_code*, where the code from the department that sends the sample is stored. Basic data for a person registration as *name, surname* and *date* of registration in included in this class too. A person can be related with two persons through the associations *father* and *mother* which represent the parents. Being one of the few hospitals in Spain experts in this disease, samples of every Spanish hospital are received in the laboratory. The provenance of these samples is conveniently stored in the *Provenance* class.

In addition, a person can be a patient or not, depending if this person suffers this disease or not. This is represented by the *Patient* class. It is also important to take into account if studies about *segregation* and *consanguinity* have been done to the patient. Sometimes, the diagnosis of the patient is not totally clear and this lack of sureness is represented in *diagnosis_reliability* attribute.

The common use of markers to identify which genes

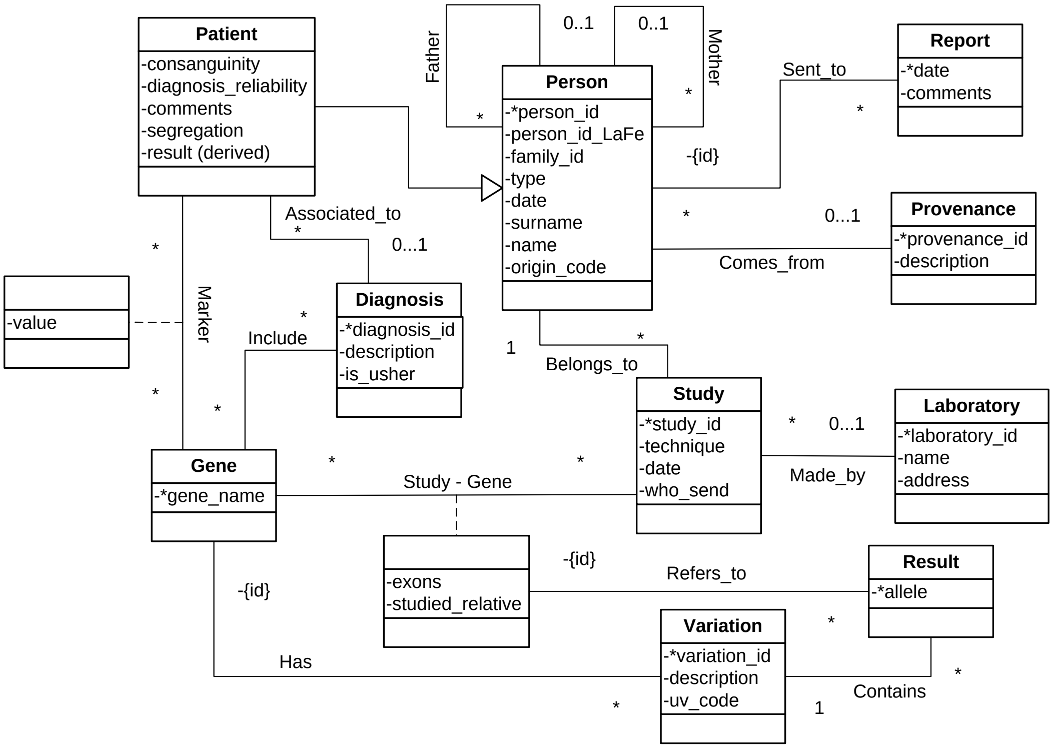


Figure 1. Information System Conceptual Schema

(represented by *Gene* class and its *gene_name* attribute) are inherited from the patient's mother and father has promoted the inclusion of a *Marker* class in this conceptual scheme. These markers help to exclude the non-mutated genes from the study, knowing previously the non-mutated genes of the parents. The value obtained from the marker is stored in *value* attribute, which is referred only to one patient and one gene.

As in any clinical service, a *Diagnosis* is associated to a patient and one or more genes. In this case, the *diagnosis_id* content represents one of the three cataloged types of the disease (Usher I, Usher II or Usher III), a combination of them if it is not clear, to the Usher Syndrome if it is not possible be more specific, or any other disease if it is necessary.

The process to analyze a sample of a person to define if he suffers the disease or if he is a carrier of it, is defined by the *Study* class. Knowing the person *who_send* the sample to an external laboratory to analyze it and the *date* of the sending is very interesting to value if the study is precise enough. To ensure the reliability of the study, storing the *technique* used to analyze the genes is very important too.

The studies are done gene by gene (represented by the *Study-Gene* class), so if the variation is not found in that gene, a new study has to be done. It is also important to store the studied *exons* (transcriptable DNA within a gene), defining which exons have been analyzed when the gene has not been completely analyzed, and to notify if a previous study of that gene in a relative has been done, which is expressed by the *studied_relative* attribute.

As mentioned before, every study can be done in a different laboratory so, it is important to store basic information about the laboratory, such as *name* and *address*, where the study has been done using the *Laboratory* class.

Obviously, after a study has been done, some results, which are represented by *Result* class, have been found. These results are related only to one *allele* and they can contain some variations which may have been found during the study.

The variations found into the DNA are represented by the *Variation* class and are related to a gene. The HGVS code of the variation is included in the *variation_id* attribute, complementing this information with a *description* attribute. Additionally, the standard *uv_code* is used to express the pathogenicity of the variation.

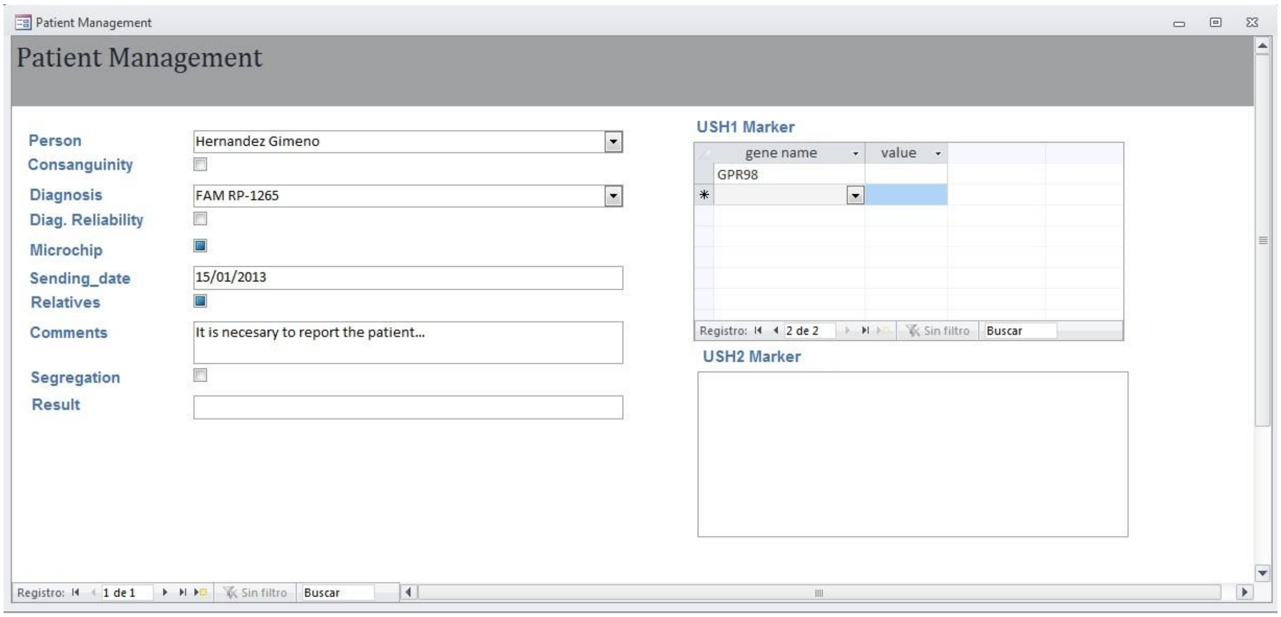
Finally, after making a complete genetic study of the person, a report has to be presented to inform the person about the results. The *date* and the *comments* related to this report are represented by the *Report* class.

VI. DATABASE AND LOADING PROCESS

From the conceptual scheme explained below, a database, which is able to store data from the sources mentioned below, has been created. The conceptual scheme, on which this database is based on, ensures the correct structure of data and the efficiency of this database. It has been produced following the Conceptual Modeling rules, ensuring the quality of information and the efficiency of storage of genomic information.

Furthermore, this database has been designed and implemented using Microsoft Access technologies. This tool used to develop the database, has been chosen taking into account that it is a well-known environment used by the clinicians. Another aspect that was taken into account is that the quantity of data handled in this research centre is not too large, so it is not necessary to use a more sophisticated System Management Database. This choice will make the information system easy to be used by clinicians.

Once the database has been developed, loading the information into the database is the next step. This process



Patient Management

Person	Hernandez Gimeno
Consanguinity	☐
Diagnosis	FAM RP-1265
Diag. Reliability	☐
Microchip	☑
Sending_date	15/01/2013
Relatives	☑
Comments	It is necesary to report the patient...
Segregation	☐
Result	

USH1 Marker

gene name	value
GPR98	
*	

USH2 Marker

Figure 2. Patient Management Form

can't be done directly due to the peculiar form of the information in the original sources. As explained in section II, the data is saved in sheets of paper, Excel files and some Access tables without any kind of structure behind it. Storing different data into the same cell of the spreadsheet or expressing the same information in different formats, are some of the encountered problems. Facing this situation together with the clinicians trying to clarify the information found in the sources is the essential step to start the loading process.

Some loading modules have been developed in order to introduce the information. These modules have been implemented analyzing the information contained in the sources, extracting the information, transforming it to the new format and loading it into the database. The "transforming" step is essential to introduce the information with the appropriate format and structure into the new Information System, avoiding the mentioned above format mistakes.

VII. Managing the data

To achieve an efficient use and avoid inaccurate management of the created Information System, a user interface has been implemented. Specific Microsoft Access forms have been developed to introduce new data and manage the database. There are five different forms that can be accessed by a *Main Menu* form.

These five forms help users correctly enter data into the information system and make queries to extract the information previously introduced. Information about *Diagnosis, Gene, Study, Patient* and *Person* is correctly stored thank to the corresponding forms.

VIII. Conclusions and Future Work

The implementation of developed information system has allowed the clinicians to overcome previous drawbacks as the waste of time looking for physical papers, the manual work, the use of different sources of information, the personal use of identification symbols to describe the data, etc. as we mentioned in section III. With this information systems-based policy, the quality of the stored data has significantly improved. The ambiguity of their data has disappeared and they can access to all the information faster and easier. It has also increased the usefulness of their data because the data can be more precisely searched and multiple parameters can be used, and even the information that previously was stored in folders can be searched through. This information system has helped this clinical research group to properly manage their data, but the problem is not an isolated case and an information system based on conceptual models can resolve problems like this in many other research centers or medical departments. Conceptual modeling technique has been used in many areas to ensure a correct management of the data, but in the clinical area, most information systems lack a conceptual model base. This work has shown that conceptual modeling is also effective in the clinical field. This work has reaffirmed us in the idea that the main difficulty for achieving this improvement in quality is the long time consuming meetings that are needed in order to experts from the informatics area understand the experts from the life sciences area.

As future work, connecting all the information about this disease would be very useful to the research about Usher Syndrome. The information system explained in this paper can help to achieve this future goal. If this information system was installed into every centre of research in Usher Syndrome, connecting these information systems to share the information between them would be possible. This big information system about Usher Syndrome would join all the information about the disease, making easier and faster the research about it.

Acknowledgment

This work has been developed under support of the director of PROS, Dr. Oscar Pastor, and jointly with Dr. José María Millán, Dr. Elena Aller and Dr. Teresa Jaijo, who belong to the *Centro de Investigación en Enfermedades Neurosensoriales* of the *Instituto de Investigación Sanitaria La Fe* in Valencia (Spain).

In addition, it is also important to highlight that this work has been done under the framework of the *Cátedra Tecnologías para la Salud* of the *Universitat Politècnica de València* financed by INDRA Systems.

References

[1] O. Pastor, "Conceptual Modeling Meets the Human Genome," Conceptual Modeling-ER 2008, pp. 1-11, 2008.

[2] O. Pastor, A. Levin, J. Casamayor, M. Celma, L. Eraso, et al., "Enforcing Conceptual Modeling to improve the understanding of human genome," 2010, pp. 85-92.

[3] M. Pastor, V. Burriel, and O. Pastor, "Conceptual Modeling of Human Genome Mutations: A Dichotomy Between What we Have and What we Should Have," in BIOSTEC Bioinformatics, Valencia, 2010, pp. 160-166.

[4] J. C. Goodman, "Health Information Technology: Benefits and Problems."

[5] A. Cufar, A. Droljc, and A. Orel, "Electronic medication ordering with integrated drug database and clinical decision support system" Studies in health technology and informatics, vol. 180, p. 693-697, 2012.

[6] M. Koca, G. Husmann, J. Jesgarz, M. Overath, C. Brandts, et al., "A special query tool in the hospital information system to recognize patients and to increase patient numbers for clinical trials" Studies in health technology and informatics, vol. 180, p. 1180-1181, 2012.

[7] J. Cassereau, A. Chevrollier, D. Bonneau, C. Verny, V. Procaccio, et al., "A locus-specific database for mutations in GDAP1 allows analysis of genotype-phenotype correlations in Charcot-Marie-Tooth diseases type 4A and 2K" Orphanet Journal of Rare Diseases, vol. 6, p. 87-94, 2011.

[8] F. P. M. Cremers, W. J. Kimberling, M. Külm, A. P. de Brouwer, E. van Wijk, et al., "Development of a genotyping microarray for Usher syndrome" Journal of medical genetics, vol. 44, pp. 153-160, 2007.

[9] J. M. Millán, E. Aller, T. Jaijo, F. Blanco-Kelly, A. Gimenez-Pardo, et al., "An update on the genetics of usher syndrome" Journal of ophthalmology, 2010.

[10] C. Nájera, M. Beneyto, J. Blanca, E. Aller, A. Fontcuberta, et al., "Mutations in myosin VIIA (MYO7A) and usherin (USH2A) in Spanish patients with Usher syndrome types I and II, respectively" Human Mutation, vol. 20, pp. 76-77, 2002.

[11] A. Olivé, Conceptual modeling of information systems, Springer, 2007.

[12] O. Pastor and J. C. Molina, Model-driven architecture in practice: a software production environment based on conceptual modeling, Springer, 2007.

[13] E. D. Falkenberg, W. Hesse, P. Lindgreen, B. E. Nilsson, J. L. H. Oei, et al., "A framework of information systems concepts" in IFIP WG, 1998.