Collecting, Analyzing, and Publishing Massive Data about the Hypertrophic Cardiomyopathy *

Lorenzo Montserrat[1], Jose Antonio Cotelo-Lema[2], Miguel R. Luaces[2], and Diego Seco[2]

[1] Health in Code, Ed. El Fortín, Hospital Marítimo de Oza
As Xubias, s/n, A Coruña, Spain
lorenzo.monserrat@healthincode.com
http://www.healthincode.com
[2] Databases Laboratory, University of A Coruña
Campus de Elviña, 15071 A Coruña, Spain
joseantonio@enxenio.es, {luaces,dseco}@udc.es

Abstract. We present in this paper the architecture and some implementation details of a Document Management System and Workflow to help in the diagnosis of the hypertrophic cardiomyopathy, one of the most frequent genetic cardiovascular diseases. The system allows a gradual and collaborative creation of a knowledge base about the mutations associated with this disease. The system manages both the original documents of the scientific papers and the data extracted from these papers by the experts. Furthermore, a semiautomatic report generation module exploits this knowledge base to create high quality reports about the studied mutations.

Key words: e-health, Document Management System, Hypertrophic Cardiomyopathy

1 Introduction

In the last decades, Document Management Systems (DMS) have become indispensable for many organizations. The majority of organizations need to access and consult stored information frequently. Thus, efficiency requirements must be considered by the Document Management Systems in order to provide a fast access to the information. Furthermore, documents need to pass from one person to another in many of these organizations. Therefore, DMS must define a set of rules for this process (a workflow process).

* This work has been partially supported by "Centro para el desarrollo tecnológico industrial (CDTI) del Ministerio de Industria" ref. Neotec RD 1406/1986 (IDI "20070178", title "Plataforma de diagnóstico genético de Cardiopatías") and by "Xunta de Galicia" ref. 08SIN008E. Other institutions collaborating in the support of the researchers are "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03 and "Xunta de Galicia" ref. 2006/4.

We present in this paper a Document Management System and Workflow for a medical organization (*Health In Code*). This system allows the creation, management, and exploitation of a knowledge base about genetic mutations associated with the hypertrophic cardiomyopathy.

Health In Code is a medical company dedicated to the identification of health problems that can benefit from a genetic diagnosis. Its main field of work are familial cardiovascular diseases, including cardiomyopathies and channelopaties. Cardiomyopathy means *disease of the cardiac muscle or myocardium*. Cardiomyopathies are therefore diseases characterised by the cardiac muscle having an abnormal structure and function. Although all diseases affecting the heart can damage the myocardium, cardiomyopathies do not include alterations due to ischaemic heart disease (myocardial infarction and related problems), to diseases of the cardiac valves, to disturbances produced by hypertension, or to congenital heart diseases. The term cardiomyopathy is reserved for a group of diseases in which the disturbance to the myocardium is the primary or fundamental alteration. The main cardiomyopathies are hypertrophic cardiomyopathy (affecting 1 out of every 500 adults), dilated cardiomyopathy, restrictive cardiomyopathy, arrhythmogenic right ventricle cardiomyopathy, and left ventricular noncompaction or spongiform cardiomyopathy.

Cardiomyopathies, above all hypertrophic cardiomyopathy and arrhythmogenic right ventricle dysplasia, are among the main causes of sudden death in young people and sportsmen and women, and are also an important cause in more elderly patients. Sudden death can be the first manifestation of these diseases. It is therefore fundamental to make an early diagnosis. Family check-ups and the systematic check-up of sportsmen and women play a fundamental role in this regard. The early identification of individuals affected by these diseases permits a suitable evaluation to be made of the risk of sudden death and the taking of effective prevention measures.

Cardiomyopathies are family diseases with a genetic cause. When a cardiomyopathy is diagnosed it must always be remembered that these diseases have a family background. In the case of hypertrophic cardiomyopathy and arrhythmogenic right ventricle dysplasia, the cause of the disease is practically always genetic and inheritable. In dilated cardiomyopathy, up to 50% of cases have a family background and are genetically caused, though it can also be due to infections, toxins, metabolic disturbances, or other causes.

Many different genetic alterations have been described associated with the development of one of the types of cardiomyopathy. For example, hypertrophic cardiomyopathy is caused by mutations in at least 10 genes of proteins forming part of the contraction machinery of the muscle cells. On the other hand, arrhythmogenic right ventricle dysplasia is produced by mutations in different genes related to the machinery of union and transmission of force between different cells.

Hundreds of different mutations have been described associated with these diseases and in fact each one of the mutations could be considered as a different disease, hence the interest in finding a genetic diagnosis that permits a better

individualising not just of the diagnosis but also the prognosis and treatment of these diseases.

The Document Management System that we present in this paper makes possible a comprehensive genetic diagnosis of all the mutations and genetic variants associated with the hypertrophic cardiomyopathy (from now on HCM), where the differential diagnosis can be confuse. Furthermore, this system makes possible to identify many genetic interactions. Thousands of papers related to the HCM are published each year, and many of these papers are contradictory. Therefore, most of this information is completely ignored by people in charge of the diagnosis of the disease. The system that we present can help clinicians to take into account this information, and thus, it can help to improve the clinical diagnosis. There are a lot of people that can be benefit from this improvement:

- Patients with a clinical diagnosis about HCM (80000 people in Spain, more than 500000 people in the EU, etc.).
- Relatives of these patients (a mean of 4 per patient).
- Patients with left ventricular hypertrophy with doubts about the possibility of HCM. We consider in this group hypertensive patients with moderate to severe hypertrophy (5% of the hypertensives), differential diagnosis with athlete's heart, obese patients with hypertrophy, etc.
- Patients with an abnormal electrocardiogram (ECG) without apparent cause.
- Sudden death patients.

The HCM is probably the most studied cardiomyopathy and thus it was selected as the main goal of the organization. However, the system, and the overall operative process in general, was designed considering other cardiomyopathies, and even channelopaties. Hence, the resulting system has a robust design that can be easily extended to collect data about other possible diseases.

The rest of the paper is organized as follows. First, the operative process of the organization is described in Section 2. Second, in Section 3 the architecture of the system and some technical details are described. Then, in Section 4, we present the user interface of the collaborative web-based application to create the knowledge base. After that, the interest of its use is emphasized in Section 5. Finally, Section 6 presents the conclusions and some ideas for future lines of work.

2 Operative Process

Figure 1 shows the operative process supported by the developed system. First, interesting scientific papers are collected by a group of documentalists from different databases, conferences, journals, etc. After that, the scientific committee evaluates these papers. Selected papers that pass the quality controls are assigned to the experts who are in charge of reviewing them. A critical reading of the paper is the first task that must be done by these experts. After that, they have to analyze the data about patients described in the paper and they have to

type these data in the application. An evaluation of the quality of the information is done by the scientific committee at this stage of the process. When the process passes these controls, the information is marked as valid and it becomes part of the knowledge base.

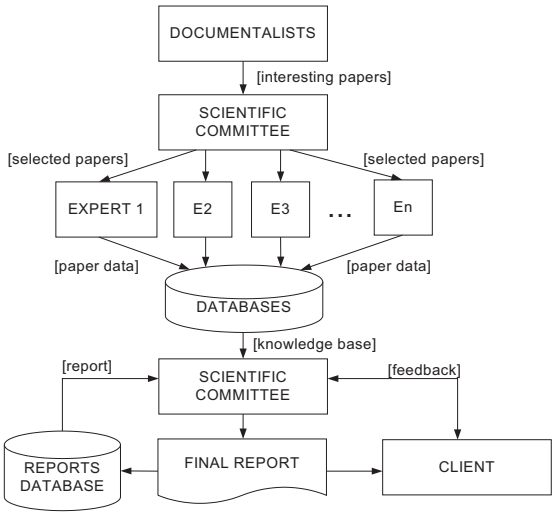

Fig. 1. Operative Process

This operative process ends with an e-commerce stage where clients provide samples about patients and the company analyzes these samples and provides all the relevant clinical information available about the identified mutations. A semi-automatic subsystem to generate reports about mutations is the base of the business model of the company. In this context, a report is a set of pre-processed information from the knowledge base or a meta-analysis of the information about each mutation. These reports must be reviewed and edited to elaborate the final version of them. These final reports are elaborated for all the mutations identified in the samples provided by the clients. The company does not replace clinicians, but it supplies them with the best tools to take the best decisions based on the available knowledge of each mutation. Furthermore, the operative process contemplates a *feedback offer* where clients can provide clinical information about new mutation carriers and their relatives. This information is evaluated and introduced in the knowledge base if it passes the quality controls. The company offers the possibility of reinterpreting the mutation implications when clients provide clinical details about these new mutation carriers.

Therefore, the operative process requires the ordered execution of a set of activities, with several people participating in each of them. In such a complex process, the lack of control on the workflow can result in dead times, errors in the obtained results, and loss of data. In general, an unsatisfactory coordination

of the people increases the overall cost and decreases the quality of the results. Because this process requires significant effort, the more automated tools that can be built and used, the better the use of human resources will be. The control of the workflow inside this work team is a key factor in the success of the process. This control can be achieved by the use of a workflow management tool specially designed for this process. That is, a system which allows to coordinate and control all the involved people, monitor and manage factors as the current state of each scientific paper, store intermediate results, control the average time to process each document, record all the people who have worked in each scientific paper, etc.

3 System Architecture and Technology

3.1 System Architecture

According to [1], workflow is concerned with the automation of procedures where documents, information, or tasks are passed between participants following a defined set of rules to achieve or contribute to an overall business goal; the computerized facilitation or automation of a business process, in whole or part. Workflow management systems can be classified in several types depending on the nature and characteristics of the process [2][3]. Collaborative workflow systems automate business processes where a group of people participate to achieve a common goal. This type of business processes involves a chain of activities where the documents, which hold the information, are processed and transformed until that goal is achieved. As the problematic of building a repository about mutations associated with the HCM fits perfectly in this model we based the architecture of the system in this model.

In general, we can differentiate three user profiles involved in the repository building:

- Administrator. Administrators are the people responsible of the process as a whole. They are the responsible of managing (add/delete/update) users and controlling the state of the application.
- Supervisor users. The supervisor users are the people in charge of carrying out critical activities such as the metadata storage, assigning tasks to different workers (*inspectors*), or supervise their work. They are members of the scientific committee and they are in charge of performing the quality controls over the scientific papers and the inspection process.
- Inspector users. The inspector users are the workers who carry out tasks such as inspecting scientific papers. This task involves a critical reading of the paper and typing relevant data in the system. This role is played by users with some knowledge in the mutations associated with the HCM but without any responsibility on the management of the system.

Supervisor and inspector users are the two user profiles involved in the repository building process. Therefore, a communication protocol between these user

profiles has been implemented. *Supervisor* users can publish news, send messages to the *inspector* users, and answer their questions.

Figure 2 shows the overall system architecture. When we defined it, we followed the recommendations of the Workflow Reference Model [2], a commonly accepted framework for the design and development of workflow management systems, intended to accommodate the variety of implementation techniques and operational environments which characterize this technology. Thus, although we used this architecture for the implementation of a specific system, it can be used in other environments and situations. Furthermore, design patterns [4][5][6] were used in order to obtain a modular, robust, and easy to extend architecture.

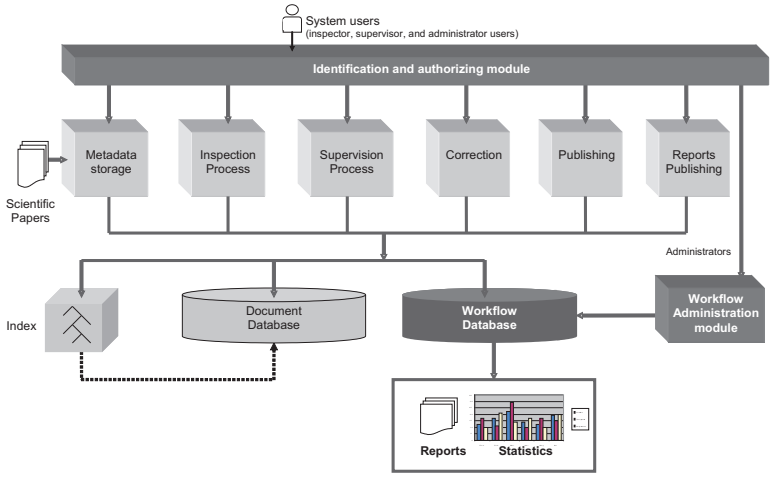

Fig. 2. System architecture

As we can see in Figure 2, the authentication and authorizing module is in charge of the authentication of the users who want to access to the system. Each user has a system role depending on the tasks he/she is going to work on. According to this system role, the authorizing module only provides the user with access to the needed features. The system architecture is composed of a module for each activity carried out during the operative process.

- *Scientific papers selection.* As we noted in the previous section, thousands of scientific papers related to the HCM are published each year. Therefore, the scientific committee has to select the most prominent papers. This subsystem provides integration with software for publishing and managing bibliographies such as EndNote.
- *Metadata storage.* This subsystem is in charge of the introduction and storage of the metadata for each scientific paper (title, author, year, source, described mutations, etc.). Furthermore, each paper must be assigned to the inspector

user who will be in charge of analyzing it. This task is performed by the supervisor users of the system, therefore only they have access to this module.
- *Inspection process.* This module allows inspector users to access the scientific papers previously assigned to them. They analyze these papers and type relevant data in the application.
- *Supervision process.* It provides access to the scientific papers and analyzed data introduced using the previous module. This task is performed by the supervisor users who are in charge of performing quality controls about the previous stage.
- *Correction.* If supervisor users detect some mistakes in the *inspection process* they can correct them using this module. Moreover, supervisor users can delegate this task to the inspector.
- *Publishing.* Once the analysis process is accepted, this module is in charge of committing its contents.
- *Reports publishing.* Reports about the mutations associated with the HCM can be generated using the data analyzed in previous stages of the workflow. Supervisor users must review and edit these reports in order to commercialize them. This subsystem provides functionalities to generate, edit, and store the reports in a database.
- *Workflow administration module.* This subsystem is in charge of managing the workflow between all these activities. It also provides reporting tools for monitoring purposes.

3.2 Data Model

Figure 3 shows our proposal for the data model that supports the architecture. This data model is organized around the entities *Paper* and *Report*. Both of them constitute the core of this architecture because the system is feeding with information from the *papers*, and the *reports* are generated by the system to be commercialized.

As we noted before, both scientific papers analyzed by the experts and the resultant information of this analysis are stored together by the Document Management System (DMS). Therefore, the entity *Paper* (and the entities associated with it such as *Mutation*, *Gene*, and *Patient*) represents both the original scientific paper (scanned PDF file, title, authors, and other metadata) and the data obtained from the analysis of this paper (mutations associated with each paper, number and type of control cases, information about described patients, etc.).

Furthermore, the workflow process described in the previous section is contemplated, and therefore, there are several entities in the data model to support it. *Admin*, *Supervisor*, and *Inspector* represent the three *User* profiles involved in the workflow process. All the entities, and modules that manage these entities, are associated with a set of user profiles. Therefore, only these user profiles have access to the information stored in such entities.

The most important characteristic of a DMS is the information that can be managed with it. As we noted, *Paper* is one of the most important entities of the data model. Metadata about original scientific papers are directly stored in

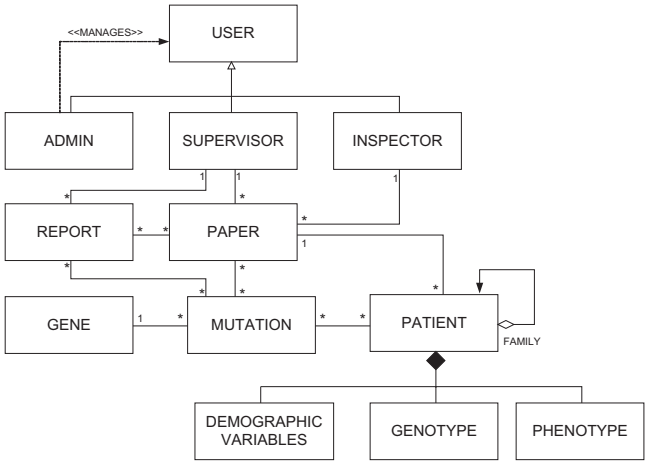

Fig. 3. Data Model

this entity. Data obtained from the analysis of the papers are organized centred around the entity *Patient*. Therefore, there is a relationship between the entities *Paper* and *Patient* (a paper describes N patients and a patient is described in 1 paper). Thousands of scientific papers about the mutations associated with the HCM have been published in journals, conferences, etc. Moreover there are many internal documents in cardiology departments of the hospitals that describe patients with these mutations. Both types of data sources describe patients and information about their relatives to a greater or lesser extent. Information about these families can be entered in the system in a very intuitive, progressive, and simple way (the carefully designed user interfaces can be seen in Section 4). Information about patients and their relatives can be categorized in the following types:

- *Identification data*. Both application internal identifiers (patient identifier, family identifier, etc.) and external domain identifiers (position in the *pedigree*) are included in this category.
- *Demographic variables*. Data about the sex, ethnicity, age at the diagnosis of the disease, etc. are included here.
- *Genotype*. This category includes the results of the genotype study. These results present the relationship between the patients and the mutations associated with the paper where the patient is described. *Obligate carrier*, *homozygous carrier*, *normal carrier*, and *not carrier* are the possible values for this relationship. However, not all the scientific papers describe this relationship for all the mutations and patients. Therefore, an *unknown* value is available for the relationship. This philosophy is applicable for most of the variables managed by the system.
- *Phenotype*. This category includes the results of the different clinical tests that can be done in order to determine the appearance of a patient resulting

from the interaction of the genotype and the environment. There are several subcategories in accordance with its nature. First, results about the *clinical diagnosis* are collected. These results determine whether the patient is affected or not by some phenotypes, which phenotypes, etc. The second group includes *environmental factors or triggers* (alcohol, hypertension, tobacco, obesity, etc.). There are many variables that can be determined in a *echocardiography, MRI, or autopsy*. These variables constitute the third group. Hypertrophy, dilatation, systolic and diastolic dysfunctions are some examples of these variables. The fourth group includes *symptoms and risk factors* (dyspnea, chest pain, abnormal blood pressure response, etc.). Variables of the *ECG* (rhythm, pre-excitation, abnormal voltage or repolarization) constitute the fifth group. The sixth group includes data about the *electrophysiological study* (inducibility of malignant arrhythmias, conduction disturbance, etc.). Finally, the last two groups include data about the *treatment* (medical treatment, surgery, etc.) and the *events* (death, cerebrovascular accident, etc.).

In brief, more than 200 variables are currently collected about each patient. However, new variables of interest can be easily introduced in the system.

3.3 Technology

This section briefly describes the most important technologies used in the development of the system. First, *Java 2 Platform, Enterprise Edition* (J2EE) [7][8] was the selected development platform. J2EE is a widely-used platform for server programming. This platform allows developers to create portable and scalable applications. J2EE provides a set of technologies that make the development process easier. JDBC (an API to access relational databases), JavaServer Pages (JSP, a technology to dynamically generate HTML), or JavaServer Pages Standard Tag Library (JSTL, a tag library for JSP) are several examples of such technologies provided by J2EE and used in this project. Furthermore, other technologies can be easily integrated with this platform. For example, *Jakarta Struts* [9], a framework that allows software engineers to develop applications following the architectural patterns *Model-View-Controller* and *Layers*, has been used. CSS [10] is the technology used to enhance the user interface. Finally, we have widely used JavaScript [11] to improve the dynamism and interaction of the user interface.

Technologies employed in the development of the *reports generation module* deserve special mention. *eXtensible Stylesheet Language Formatting Objects* (XSL-FO) [12] is the most important technology used in this module. XSL-FO is a mark-up language for XML document formatting which is most often used to generate reports. An XSL-FO document is an XML document where the format of a dataset is defined. This format defines the presentation of these data in a paper, screen, or other media. The XSL-FO document does not describe the layout of the text on various pages. Instead, it describes what the pages look like and where the various contents go. However, the developed system does not write XSL-FO documents. An XSLT transformation is used to convert the semantic

XML, generated from data in the knowledge base of the system, into XSL-FO documents. This issue is very important because it provides independence between data and their output format. Finally, Apache FOP [13] is used to render the XSL-FO document to a specified output format. Output formats currently supported by Apache FOP include PDF, PS, PCL, AFP, XML, Print, AWT and PNG, RTF and TXT. The primary output target is PDF. However, our system generates RTF documents because the reports must be editable by the experts.

4 User Interfaces

The usability of the system is a key factor to guarantee its acceptance. In this context, this term denotes the ease with which users employ the application [14]. Therefore, in application domains where users do not have the required expertise level about computers, it is very important to design the user interfaces in accordance with the user preferences. For example, this system is used by experts in HCM but they do not have to know the way typical web applications work. The main issues of the user interface design are presented in this section.

Typing in the application the data resultant from the analysis of the scientific papers is the most expensive task. An average time of 3 hours has been estimated by the experts in charge of this task. Therefore, the user interface has to allow doing this task in several steps. Furthermore, scientific papers about the HCM usually present a common format. Patients and their relatives are described to a greater or lesser extent. However, sometimes these descriptions are organized by family, other times they are organized by type of data, etc. Therefore, the user interface has to take this issue into account. Figure 4 shows the developed user interface for typing data resultant from this process in the system.

The most important feature of this user interface is that it can manage a lot of information in each screen. As we noted before, each article can describe several families and each of them could have tens of studied cases. Moreover, more than 200 variables about each studied case are collected in the system. The designed interface provides the users with a centralized access point where they can introduce, consult and update all the data about the studied cases described in an article. Furthermore, a requirement of flexibility has been considered in the design of this interface. The experts that have to enter the data in the application can do it following the same organization presented in the original scientific paper. For example, they can introduce the data that identify all the patients and their relatives and, after that, they can complete other information such as genotype, phenotype, etc. But, they could type all the data about a patient before introducing other one. Some graphical icons help the user to know the state of each data category.

This design of the user interface implies that there is a lot of information in the same screen. All this information is organized and categorized in a natural way for the experts that have to use the system. First, data about each patient are organized in categories (demographic variables, genotype, phenotype, etc.). Moreover, these patients are grouped in families. A dynamic technique has been

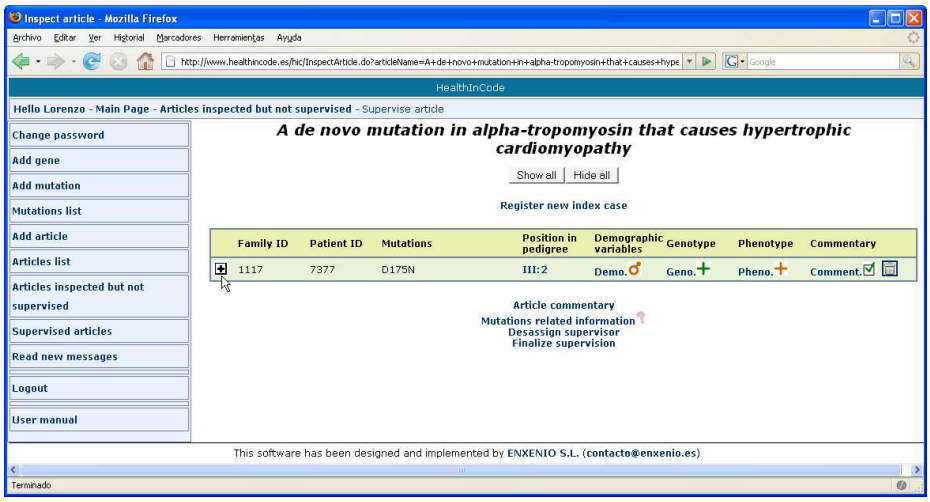

Fig. 4. Screenshot of the user interface (I)

used in the implementation of the interface that allows the experts to fold and unfold the families. Figure 5 presents the same family of the previous screenshot unfolded.

These screenshots belong to a small example, with just a few cases, to improve the quality of the figures. However, there are scientific papers where more than one hundred cases are described. These papers would be unapproachable without the techniques presented in this work.

5 Use Interest

The main goal of this section is to emphasize the usefulness of the developed system. As we noted in previous sections, thousands of scientific papers about the mutations related to the HCM are published each year. Therefore, the task of generating reports that summarize each of these mutations would be unapproachable without a system like the one presented in this paper. There are two key factors in the design of the system. First, our system defines a workflow process. The scientific document repository building requires the ordered execution of a set of activities on the documents with several people participating in each of them. In such a complex process, the lack of control on the workflow can result in dead times, errors in the obtained results and loss of data. Second, our system is web-based. This allows the users to collaborate all over the world in order to keep the information constantly up to date.

Nowadays, there are 57 users registered in the system. One of this users has the role of *administrator*, 24 users have the role of *supervisor*, and 32 users have the role of *inspector*. These users have collaborated in the creation of a database with more than 1 GB of information. More than 3280 mutations, categorized

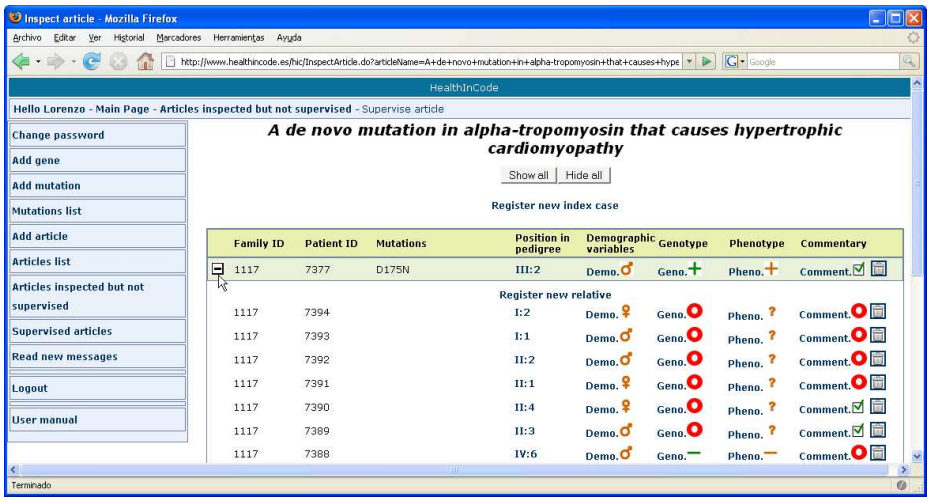

Fig. 5. Screenshot of the user interface (II)

in 85 genes, and more than 3405 scientific papers are registered in the system database. Furthermore, data about more than 14000 patients (or relatives) described in these scientific papers are available to generate high quality reports.

On the other hand, carefully designed user interfaces have a significant reduction in the time that experts need to analyze a scientific paper and to type relevant data in the application. This means a reduction of cost for the organization. But also, and much important, information can be up to date easily without expending a lot of money to hire more experts.

6 Conclusions and Future Work

We have presented in this paper the architecture and some implementation details of a Document Management System (DMS) to help at the diagnosis of the hypertrophic cardiomyopathy. The creation of the DMS repository is not a simple process. It requires the coordination of people and tools to carry out every activity that is part of the process. For all these process to be correctly and efficiently made, it is necessary the use of support tools that facilitate the work of each participant and ensure the quality of the obtained results.

The proposed workflow strategies and system architecture support the control and coordination of people and tasks involved in the whole process. The use of this architecture automates the completion of prone to error activities and optimizes the performance of the process and the quality of the obtained results. This architecture was defined following the recommendations of the Workflow Reference Model. Furthermore, several architectural and design patterns were used in order to obtain a modular, robust, and easy to extend system. This

system was built as a web application which provides an integrated environment for the execution of all the tasks.

As lines of future work, the developed system is going to be applied to new projects of similar characteristics involving other Cardiomyopathies (Dilated, Restrictive, etc.) and Channelopathies (Brugrada syndrome, Long QT syndrome, Short QT syndrome, etc.). In addition, we are working on different implementations of the activities considered to optimize the performance of the overall process. Another future development could be the extension of the report generation module in order to support reports about several mutations. It could be very useful to analyze some mutations that occur in nearby areas of the protein. Finally, a statistical module could be developed to provide research capabilities inside the system. For example, a study about the correlation between variables could improve the quality of the final reports.

References

1. Hollingsworth, D.: Workflow management coalition - the workflow reference model. Technical report, Workflow Management Coalition (1995)
2. van der Aalst, W., van Hee, K.: Workflow management: Models, methods, and systems. (2002)
3. Fischer, L.: Workflow handbook 2003. Future Strategies Inc., USA (2003)
4. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elments of Reusable Object-oriented Sofware. Addison-Wesley (1996)
5. Grand, M.: Patterns in Java. Volume 1. John Wiley & Sons (1998)
6. Alur, D., Crupi, J., Malks, D.: Core J2EE Patterns. Prentice Hall (2003)
7. Perrone, P.J., Chaganti, K.: J2EE Developer's Handbook. Sam's Publishing (2003)
8. Bodoff, S.: The J2EE Tutorial. Addison-Wesley (2004)
9. Holmes, J.: Struts: The Complete Reference, 2nd Edition. McGraw-Hill Osborne Media (2006)
10. Shafer, D.: HTML Utopia: Designing Without Tables Using CSS. Sitepoint Pty Ltd. (2003)
11. Flanagan, D.: JavaScript: The Definitive Guide, 5th Edition. O'Reilly (2006)
12. W3C Recommendation: Extensible stylesheet language (XSL) version 1.1. Technical report, W3C (2006)
13. Apache FOP: Retrieved from http://xmlgraphics.apache.org/fop/ in october 2007. (2007)
14. Shneiderman, B.: Designing The user interface, Strategies for effective Human-computer interaction. Addison-wesley (1998)