Reengineering of Legacy Data and Knowledge for Electromyography Studies*

Mária Bieliková, Pavol Návrat, and Mária Smolárová

Slovak University of Technology
Department of Computer Science and Engineering
bielik@elf.stuba.sk, navrat@elf.stuba.sk,
smolarova@dcs.elf.stuba.sk

Abstract. We discuss different data representation formats used for collection and management of electromyography (EMG) data. Data representation is described in the context of the proposed multifunctional platform for EMG studies. The platform architecture consists of four layers: data collection, data analysis, learning and education. In this paper, we focus on the data collection layer. We describe different formats ranging from an earlier ECCO format through the designed relational data model to the proposed XML exchange format. We also survey a set of software tools we have developed for collecting the EMG data (EMMA) and determining EMG normative data (NoDaM and NoDaC).

1 Introduction

Electromyography (EMG) is a standard method for monitoring neuromuscular activity at the level of bioelectrical signals. The EMG diagnosis is based on a systematic acquisition of numeric and symbolic data. The domain of EMG is broad, covering more than a hundred existing diagnoses, exploiting knowledge belonging to various kinds of data, including about four thousand tests of nerve or muscle structures [17].

The relative complexity of EMG examination raises possible problems for physicians. Some novice practitioners propose excessive tests that are costly, time consuming and painful for patient just because they fear they might omit some relevant tests. Some physicians, on the other hand, tend to propose only a very limited set of tests. It results into insufficient EMG data, which may not allow for example to consider cases of associated pathologies. Thus experts in the area of EMG call for some kind of standardisation of the examination procedures. Moreover, to foster a more objective interpretation of test data, standardisation of values resulting from respective tests attributed to healthy population, i.e. defining so called normative data, is important, too. Yet another issue is more effective reuse of expert knowledge that is crucial in interpreting data collected during an examination and in formulating a diagnosis.

* The work reported here was partially supported by project INCO Copernicus, No. 977069, EMG-Net and by Slovak Science Grant Agency, grant No. G1/7611/20.

Computer based information processing supporting diagnostic process in neurophysiology may enhance the physician's ability to make appropriate decisions and find a correct diagnosis. Several diagnosis systems were developed to date. Most of them are systems of production rules [8, 14, 15]. Other approaches include a causal functional model [7], and bayesian probabilities model [3]. Recently, an object oriented knowledge base MYOSYS specifically used for planning neuromuscular tests was developed [16].

Such singular systems, often built with considerable but still quite limited effort, operate with a relatively small amount of knowledge (mainly due to difficulties with acquiring and representing knowledge). The knowledge has not been sufficient to create comprehensive models of concepts and entities that are dealt with in the EMG domain. As a result, most of the systems are not being upgraded and even not used any more [17].

The systems and their knowledge bases should be updated regularly according to user demands, to reflect evolution of the field, to amend errors and misinterpretations, etc. They should be updated also to cope with technological development and need to update hardware and software platforms etc. Moreover, the systems were designed without taking into account a mere existence of some other system, not to say possibility of collaboration with it.

The European project EMG-net was launched to join several efforts in the area of EMG studies. A multifunctional platform for EMG studies has been envisaged as a goal that would assist practitioners in developing standard examination procedures. This goal can be seen as a first step towards a virtual clinical electromyography laboratory that will integrate available sources of domain knowledge and data with advanced information processing techniques. Incorporating world wide web into the platform's structure has been essential from this point of view.

The project has been interdisciplinary and international, with several teams working on particular problems. In this paper, we shall not report on project's results as a whole. Quite naturally, we shall restrict ourselves to those that fall within our team's direct responsibility. We shall devote our attention to the problem of reengineering of legacy data and knowledge while keeping their previous formats as an option at the same time.

The outline of the rest of the paper is as follows. In Section 2, we briefly describe architecture of the proposed multifunctional platform for EMG studies and summarise concepts of the EMG data collection. In Section 3, EMG data representation is described. Next, we present realisation of the data collection layer by a set of software tools. We conclude the paper with our conclusions and give some proposals for the future work.

2 Basic Concepts of Computer Supported Collection of EMG data

We have proposed a multifunctional platform for EMG studies. The platform consists of four layers: data collection layer, data analysis layer, learning layer and educational layer (see Figure 1). Individual layers operate on EMG data and knowledge stored in the EMG case base and the EMG knowledge base. Architecture of the multifunctional platform for EMG studies evolved from the proposal introduced in [2].

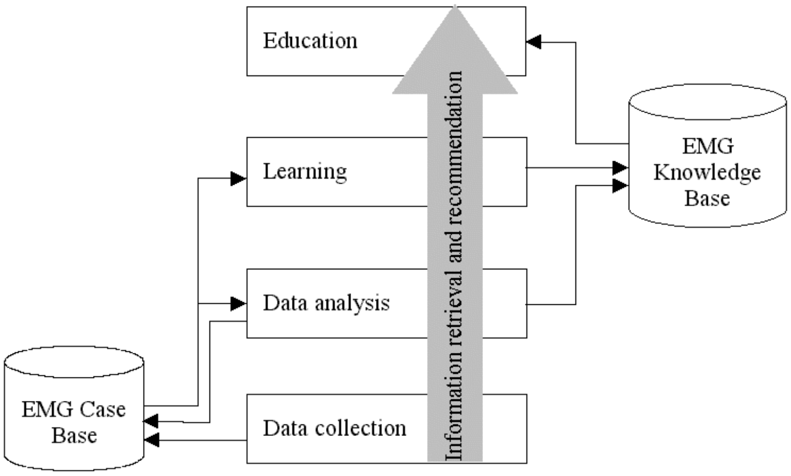

Fig. 1. Multifunctional EMG platform.

Collection of data related to EMG examination cases forms the first and basic layer of the proposed architecture. Collected EMG data are stored in the EMG case base. The EMG case base contains roughly three groups of data:

- *general data* (patient, examination),
- *primitive measurement data* (mainly numerical values for conduction velocities, amplitudes, etc.),
- *symbolic interpretations* of these measurements at various levels of detail (symbolic parameter values, pathophysiological test conclusions, pathophysiological structure conclusions, EMG diagnoses and clinical diagnoses).

EMG case base is actively used by data analysis and learning layers. These layers consist of a set of data analysis, data mining and machine learning tools. The objective of these tools is to analyse collected EMG data for consistency, completeness, performing useful statistics and "mine" new EMG knowledge from data acquired at the data collection layer.

EMG knowledge is stored in the EMG knowledge base, which is realised as a web server containing available consensual knowledge on the EMG domain [17]. The server, practically, reflects the current state of expertise in the EMG domain and provides a possibility to access this expertise via world wide web.

Each layer can use the EMG data and knowledge through information retrieval and recommendation facilities. This part of the architecture is inevitable, mainly due to extensiveness of the EMG domain. One promising approach is sharing knowledge about quality of information resources using combination of content-based and collaborative filtering [13].

Collecting the EMG data is carried on in distinct regions. In our approach, we consider the communication at several levels: the local level, the national level, and the European level. The local level is represented by a particular EMG workplace. The national and the European levels are intended mainly for exchange of the EMG expertise accumulated during years of practice. This would improve the quality of an early diagnosis and prevention of neuromuscular diseases in each particular country, or a whole region. The national level becomes also important when the aim of the EMG data collection is development of normative EMG data.

Figure 2 illustrates two sides of the EMG data collection. Local side (left part) corresponds to the computer support of local EMG data collection through a combination of an automated support of data acquisition from the EMG machine and a manual input (or modification) of the EMG data by a physician. Manual input is inevitable at least in case of inferred data (physician's conclusions). Proposing conclusions at different levels of diagnosis can be supported by a decision support tool (analysis or learning layer of the architecture).

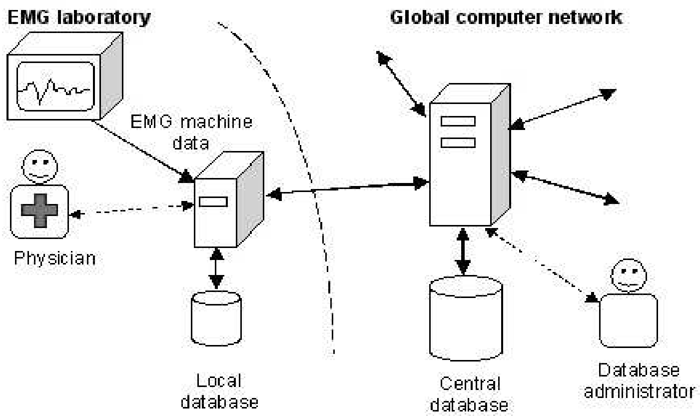

Fig. 2. Distributed EMG data collection.

The right part of the Figure 2 depicts a computer support of the EMG data distribution through global computer network or Internet. EMG data are stored in a global database, which should be synchronised with local databases in order to contain correct data from all EMG laboratories. In fact, several global databases can exist: global national databases and global regional databases. This hierarchy allows the use of a software tool at all three mentioned levels: local, national and European. Separating particular levels enables use of the collection software tool in off-line mode, i.e. physician is allowed to enter data without the Internet connection.

The described hierarchy is employed also for collecting measured EMG data suitable for determination of normative EMG data. Normative data enable EMG practitioners to compare clinical data of patients to those of healthy persons. To find values that are as representative as possible (to become a true standard), collaboration of several laboratories distributed throughout the territory is inevitable. The measurements are coordinated by the EMG laboratory that should issue requests for EMG measurements to participating laboratories, accumulate, store and monitor the measured EMG values provided by all partners. However, all participating laboratories should closely collaborate on formulating the requests. Well-defined requests strongly influence results of a process of the EMG data collecting, because each laboratory can have specific possibilities according the measured values (also in respect with healthy volunteers).

The physician in an EMG laboratory receives a request containing information about required EMG examinations, requirements on a subject being examined (sex, age, temperature, etc.) together with additional information, which allows standardisation of the examination process. The measured EMG values (response) are passed to the coordinating laboratory.

3 EMG Data Representation

During the previous EMG project (ESTEEM - European Standardised Telemetric tool to Evaluate knowledge-based EMG systems and Methods), the software tool for EMG data collection (called CASETOOL) was developed. For storage and exchange of EMG data, the binary ECCO format was designed. Several existing tools use the ECCO format (e.g., KANDID [8]). Moreover, EMG laboratories, which form the EMG-net consortium, have collected alltogether more than 1 000 EMG examination cases in the last years, all stored in the ECCO format. There exists also so called *golden EMG data collection*, which stores more than 200 examination cases resulted from the consensus of partners from about a dozen laboratories (each from different country in Europe). The golden part of the EMG case base is intended for data analysis and mining.

Main difficulty with the ECCO format lies in its inflexibility with respect to extensions. The problem arose during the EMG data collection and evaluation. Binary format is also not very suitable for exchange using the Internet for reliability reasons.

In order to preserve continuity in the EMG-net project as well as to support future developments, we use several data representation formats, which range from binary format (ECCO version 3.2) through relational format (represented by a relational database) to text format (represented in the XML - eXtensible Markup Language).

The data collection layer of multifunctional platform for EMG studies should support all three formats. ECCO is retained in the EMG examination cases collection and management tool in the form of export/import capabilities. The relational format is advantageous because of its flexibility, extensibility and portability. It is used for storing EMG examination cases and serves as a base for data mining. It is refined on the higher layers according to specific requirements for data analysis and knowledge discovery [9, 12].

Textual XML format is suitable for distribution of EMG data between physicians or collaborating software tools. Exchange of EMG examination cases is important in the process of consensus exercises, where a group of physicians discusses particular EMG examination case in order to reach consensus regarding an EMG diagnose as well as a clinical diagnose. Distribution of requests and collecting responses serves for collecting data suitable for normative EMG data determination.

Detailed description of the ECCO format can be found in [10]. The following sections provide more information about relational and textual formats of the EMG data.

3.1 EMG Data Relational Model

Designed relational data model is based on the analytical model of the ESTEEM Communication Protocol (ECCO version 3.2). The proposed model fully comprises the attributes and their corresponding values from the ECCO format. Moreover, we proposed additions, which are inevitable for efficient data management and extensions proposed in the User-5 specification [11]. Beside this, the relational data model is capable to store normative EMG data.

The basic part of our model is the interconnection between the *Examination*, *Structure* and *Test* entities. It is based on the current situation in the EMG examination process, where a single EMG examination includes examination of several structures and each structure can be tested by several tests.

The central entity of our model is the *Examination* entity. It contains general information acquired from one examination. The *Patient* entity contains only those attributes, which are permanent for all examinations of a patient. At the present it contains only attributes obtained from the ECCO format and two new attributes - *Status* and *Consensus* (both for marking purposes in developed software tool). However, we assume some other attributes will emerge. We also expect the EMG examiners would like to add some attributes for their internal purposes (e.g., address, comment, phone number). The relation between the *Patient* and the *Examination* is 1:N.

There exist several entities with the *Type* attribute, which affect contents and meaning of attributes in their corresponding entities. This approach is used to simulate generalisation relationship between particular entities.

We identified some parts of the data model, which were designed with the likely future changes of requirements in mind. The identified entities contain their symbolic values in separate tables. Each of such tables contains the $CodeTableVer$ attribute, which determines version of the given code table (at the present the values are equivalent with the ECCO, v. 3.2 values). This allows possible changes and support of different symbolic value assignment versions by a collection tool without a change of the software. As an example we can mention flexible representation of clinical diagnosis names and their numeric representations; EMG conclusion names and their numeric representations, or anatomical structures and their codes. Described approach complicates the data model, so we had to carefully decide between simplicity of the data model and its flexibility. A complete report of the proposed relational model is available at `http://www.dcs.elf.stuba.sk/emg/`.

3.2 XML and EMG data

Since our aim has been to develop a set of software tools, which collaborate in collecting the EMG data at several levels (local, national and European), availability of an appropriate exchange format was cruical. For the purposes of exchange, we proposed the textual representation of the EMG data in the eXtensible Markup Language (XML). XML was chosen because its facility to describe the content of (semi)structured data flexibly and across multiple platforms. The readability of XML documents as well as the availability of software tools supporting XML manipulating was equally important.

The proposed markup language for the EMG data exchange (named EMG-ML) is intended to create XML data-centric documents [6]. EMG-ML reflects contents of the EMG case base. It defines markup for data about the patient's examinations (EMG examination document), together with data related to request and response documents used in the process of normative EMG data determination.

The structure of request and response documents is almost the same, the only difference is that request does not contain values of particular parameters. These will be measured and thanks to our NoDaC software tool automatically included in the response document.

The EMG examination document consists of data about a patient and data related to (possibly several) examinations for this patient. Bellow is depicted part of the Document Type Definition for the EMG examination:

```
<?xml version="1.0" encoding="windows-1250"?>
<!DOCTYPE emd [
<!ELEMENT EMD (PATIENT, EXAMINATION+)>
    <!ELEMENT PATIENT ((NAME | PATIENTID), BIRTHDATE,
                                          RACE, SEX)>
    ....
    <!ELEMENT EXAMINATION (HEAD, NONEMG, EMGDIAG?,
                         FINALDIAG?, STRUCTURES, TESTS)>
    ....   ]>
```

Using XML files as the exchange format has several advantages: the content of the document is dynamic, depending on the actual anatomical structures that are examined and used examination techniques; the content of the document is self-describing because of the text-based nature of XML documents; the EMG data transmission is highly error-free because its non-binary character and the transmitted data are "open" for third party software processing. Main disadvantage of such data format is the length of the document and inefficient manipulation of the data. However, its combination with a relational database solves these problems.

It should be noted that data about some patient' examination are to be treated with an extreme care with respect to possible enfringement of individual privacy. All data sent in this format for the purpose of collaboration on such themes as consensus exercises are strictly de-personalised.

4 Realisation of Data Collection Layer

At this moment the data collection layer consists of a set of software tools aimed at collecting data about EMG examination cases [5] and normative EMG data determination [4].

Software tool for collection of EMG examination cases is named EMMA. It stores and enables manipulation of all the above mentioned data. Moreover, we also incorporated EMG normative data into the design, which serve for accurate examination interpretation. Input and modification of normative data in all the required formats (e.g., in table format, functional format) together with calculating symbolic parameter values according these data in an actual test are supported. The EMMA tool is designed in such way that a patient is considered a central entity in the system. A physician can work with patient data (general and clinical information) and create several EMG examinations for the patient. This approach enables observing changes in subsequent examination of a particular patient (see Figure 3).

Collection of normative data is a complex process, which requires EMG examination of many healthy persons. Based on the proposed data representation, we have developed two software tools: NoDaM (Normative Data Manager) and NoDaC (Normative Data Client).

NoDaM is directed for the coordinating EMG laboratory. Its primarily functions are to produce the correct XML request document according to physician's requirements, process the data from the XML response document and transform them into the EMG case base.

NoDaC software tool allows the examiner to record the values of EMG parameters. It lets him/her to review the requirements from the coordinating laboratory and sent measured data back in the form of XML response document. Our main goal was to keep this tool as simple as possible by providing the minimal functionality actually needed. We therefore designed NoDaC as a dynamic form configured by the XML request document. Only those parameters specified in the obtained request document are displayed in the form. The other ones, non-relevant or not required, have not to be considered.

5 Conclusions

Presented work is aimed towards EMG standardisation by development of multifunctional platform for EMG studies. We concentrated in this paper on data

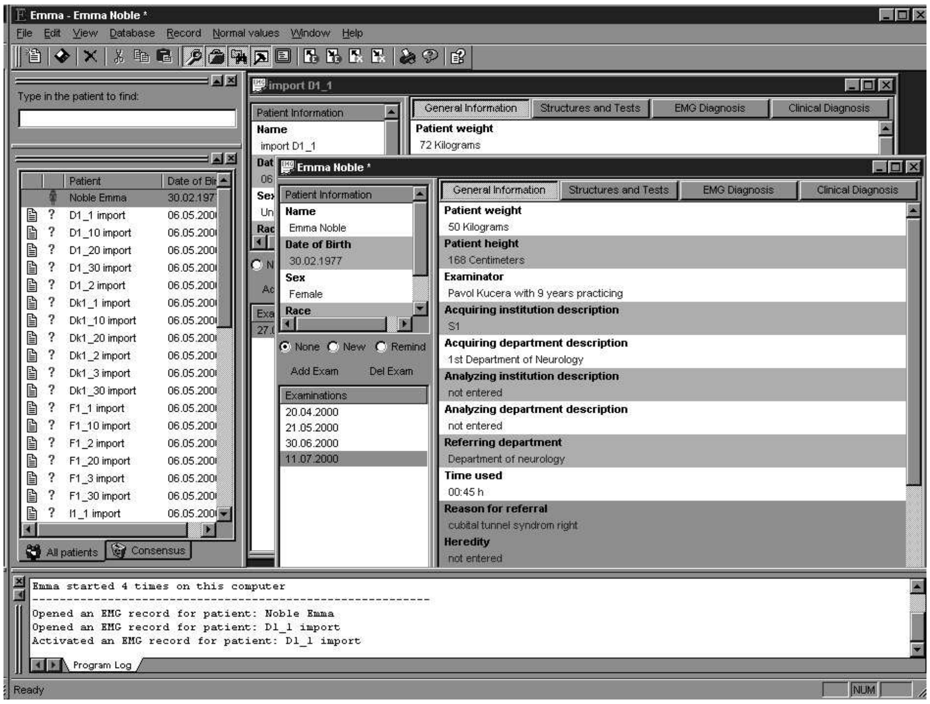

Fig. 3. Example of EMMA interface.

representation, which is crucial task in such data intensive area as is the EMG examination. Proposed data representation serves for all layers in the EMG studies platform. It allows different EMG data management features:

- support of tests and techniques selection during EMG examination;
- support of conclusion statements determination (at all levels, i.e. structure conclusion, EMG diagnose conclusion and clinical diagnose conclusion);
- support of retrieving cases based on their similarity;
- support of monitoring patient's course of illness.

We have proposed the communication method between the EMG laboratories at different levels (local, national, or European). The communication is based on using the EMG-ML format, which we have developed according the XML standard.

An unsolved problem in current state of data representation is a fairly large amount of non-structured information, which is entered into the database as free text. The structured clinical information from MYOSYS [16] can be used to enhance possibility of data mining over created databases.

Developed software tools for collection of EMG examination cases and for EMG normative data determination belong to the first layer of the proposed architecture of multifunctional platform for EMG studies. The tools we have developed ease not only collection of EMG examination cases, but more importantly reviewing the examination cases during consensus exercises and making decisions about a diagnosis. We find our results to be the firm basis for future works when several other tools may be built on the top of it.

References

1. Agre, G. and Ziébelin, D. (Eds.): Proceedings of Workshop on Application of Advanced Information Technologies to Medicine. Varna, 2000.
2. Agre, G.: A Multifunctional Platform for EMG Studies. In: [1], 2000, pp. 117-124.
3. Andreassen, S. et al.: MUNIN - an expert EMG assistant. In: Computer-Aided Electromyography and Expert Systems, Desmedt, J.E. (Ed.), Elsevier Sciences Publishers B.V. (Biomedical Division), 1989.
4. Bieliková, M., Kučera, P., Návrat, P., Smolárová, M., Traubner, P. and Záhon, P.: Computer Support for Normative EMG Data Determination. In: [1], 2000, pp. 39-47.
5. Bieliková, M., Návrat, P., Smolárová, M.: Collection and Management of EMG Examination Cases. ERCIM News. No. 45, April 2001, 43-44.
6. Bourret, R.: XML and Databases. Technical Report. Technical University of Darmstadt, September 1999.
7. Cruz, J. and Barahona, P.: A causal functional model for applied to EMG diagnosis. In: Proc. of 6th Int. Conf. on Artificial Intelligence in Medicine, Grenoble, Springer, LNCS 1211, March 1997, pp.249-260.
8. Fuglsang-Frederiksen, A., Ronager, J. and Vingtoft, S.: PC-KANDID: an expert system for electromyography. Artificial Intelligence Medicine, 1989, 1: 117-124.
9. Garbay, C.: Data Mining in EMG: An Agent-Centered Perspective. In: [1], 2000, pp. 72-83.
10. Jakobsen, L.S., Fog, B. and Talbot, A. ESTEEM communication protocol, ver. 3.2. ESTEEM report Advanced Informatics in Medicine (A2010), 1993,1-87.
11. Johnsen, B. and Fuglsang-Frederiksen, A.: USER-5: EMG dataset specification version 3. ESTEEM report Advanced Informatics in Medicine (A2010), 1994, 1-44.
12. Pilka, M. and Bieliková, M.: Mining the Gold EMG Case Base. Tech. report, Department of Computer Science and Engineering, Slovak University of Technology in Bratislava, 2001.
13. Polčicová, G., Slovák, R. and Návrat, P.: Combining Content-Based and Collaborative Filtering. In: Proc. of ADBIS-DASFAA'2000 - Symposium on Advances in Databases and Information Systems, Y. Masunaga, J. Pokorný, J. Štuller and B. Thalheim (Eds.), MATFYZ press, Praha, Sept. 2000, pp.118-127.
14. Rialle, V., Vila, A. and Besnard, Y.: Heterogeneous knowledge representation using a finite automation and first order logic: a case study in Electromyography. Artificial Intelligence in Medicine, Vol. 3, 1991.
15. Schofield, I.: HINT High Level Inferencing Tool: An Expert System for the Interpretation of Neurophysiologic Studies. In: Proc. of MIE'94, 12th Int. Congress, Lisbon, Portugal, May 1994.
16. Ziébelin, D. and Vila, A.: Medical planning environment. In: Proc. of 6th Int. Conf. on AI in Medicine, Grenoble, Springer, LNCS 1211, March 1997, pp. 343-348.
17. Ziébelin, D. and Villa, A.: Building knowledge base in electromyography through the WWW. In: [1], 2000, pp. 117-124.