

Searching biosignal databases by content and context: Research Oriented Integration System for ECG Signals (ROISES)

Alexandra Kokkinaki, Ioanna Chouvarda, Nicos Maglaveras*

Lab. of Medical Informatics, The Medical School, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:

Received 1 February 2010

Received in revised form

28 December 2010

Accepted 12 February 2011

Keywords:

Biosignal database

Ontology

Semantic and syntactic mapping

ABSTRACT

Technological advances in textile, biosensor and electrocardiography domain induced the wide spread use of bio-signal acquisition devices leading to the generation of massive bio-signal datasets. Among the most popular bio-signals, electrocardiogram (ECG) possesses the longest tradition in bio-signal monitoring and recording, being a strong and relatively robust signal. As research resources are fostered, research community promotes the need to extract new knowledge from bio-signals towards the adoption of new medical procedures. However, integrated access, query and management of ECGs are impeded by the diversity and heterogeneity of bio-signal storage data formats. In this scope, the proposed work introduces a new methodology for the unified access to bio-signal databases and the accompanying metadata. It allows decoupling information retrieval from actual underlying datasource structures and enables transparent content and context based searching from multiple data resources. Our approach is based on the definition of an interactive global ontology which manipulates the similarities and the differences of the underlying sources to either establish similarity mappings or enrich its terminological structure. We also introduce ROISES (Research Oriented Integration System for ECG Signals), for the definition of complex content based queries against the diverse bio-signal data sources.

© 2011 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Information and communication technologies have grown rapidly during the last decades and are vastly used in every aspect of human activity. In healthcare environment like primary care and hospitals, the most visible technological developments of the last thirty years are electronic health records, tele-medicine, bio-signal processing, medical image interpretation, and knowledge processing. The wide spread use of bio-signal acquisition devices in daily life applications, generated and continues to generate large amounts of bio-

signals. Following the pervasive healthcare directions, access to bio-signals has been diversified and is no longer restricted to the hospital domain. Therefore, research resources are fostered and research community promotes the need to extract new knowledge from bio-signals towards the adoption of new medical procedures.

In the cardiology domain, the recording of the resting 12-lead ECG continues to be the most commonly used laboratory procedure for the diagnosis of cardiac disease. As a record of the electrical activity of the heart, it is a unique technology that provides information not readily obtained by other methods [1]. However, ECG bio-signals are not efficiently managed,

* Corresponding author at: Lab. of Medical Informatics, Faculty of Medicine, The Medical School, Aristotle University of Thessaloniki, P.O. Box 323, 54124 Thessaloniki, Greece. Tel.: +30 2310999281; fax: +30 2310999263.

E-mail addresses: alko@med.auth.gr (A. Kokkinaki), ioanna@med.auth.gr (I. Chouvarda), nicmag@med.auth.gr (N. Maglaveras).
0169-2607/\$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.02.008

mainly due to the diversity and heterogeneity of the various storage formats. ECG devices traditionally record ECGs in binary formats which are mainly proprietary, thus not interoperable and comprehensible. Consequently, meaningful ECG data analysis is impeded, also considering that most data formats do not provide the means to store, the equally important, context and content meta-data of the signal. Biosignal context refers to information about patient demographics, diagnosis, recording equipment, researcher/investigator, etc., while annotation and interpretation information constitute part of the bio-signal content. Retrieving biosignal metadata from other sources (if data are available and procedures are feasible) is often a prohibiting factor for important meta-analysis of the bio-signal [2].

1.1. Related work in biosignal management

As effort has been paid by the biomedical community to overcome the shortcomings of the various vendor proprietary formats, open bio-signal standards have been developed. The European standard in electrocardiography domain, SCP-ECG (Standard communication protocol for Computer-assisted electrocardiography), was developed in 1993 and specifies the interchange format and a messaging procedure for ECG cart-to-host communication and for retrieval of SCP-ECG records from the host (to the ECG cart). SCP-ECG standard, as defined by TC251 committee, specifies that information is to be structured in data sections [3]. Besides the actual time-series, and some attached annotations, various types of information regarding the patient and the medical procedure are included, either as mandatory or optional fields [4]. PhysioNet [5] is an Internet resource from the NIH (National Institutes of Health) supplying well-characterized physiological data sets and related open-source software to the biomedical research community. PhysioNet, supports WFDB (Waveform Database) format which is mostly used for ECG data, but provides data format conversions from WFDB to text and EDF plus (European Data Format plus). EDF plus is a 16 bit data format supporting multiple sampling rates and multiple scaling factors, especially suitable for multichannel recordings. It is broadly accepted in the research community for data exchange and many vendors provide import and export filters [6].

Regarding bio-signal archiving and management systems, technological advancements can be seen mainly in two directions. From one side, biosignal management systems have been developed in the context of homecare systems or integration of care and thus have to deal with patient related metadata, as well as possible temporal issues, i.e. multiple signals per patient [7], as well as multimedia and multi-level patient data integration [8]. In parallel, multicentric studies and disease specific research projects have to deal distributed databases and data heterogeneity [9–11], and therefore semantic integration and semantic web techniques [12] have been deployed. In the latter case, where integration of heterogeneous data is required, automated mapping mechanisms need to be developed [13].

In a wider scale and moving beyond specific applications and domains, PhysioNet consists a very good paradigm of a biomedical resource offering access to bio-signals as a public database, however it is not yet structured in a manner

that allows for content based search. Mechanisms that would enable the combined access to bio-signals, their features and annotations along with the medical data that describe the context of the bio-signals, would provide a valuable tool, not only for knowledge discovery by the biomedical researcher community, but also for clinical use, in terms of medical evidence and clinical decision support, or even medical error filtering.

1.2. The proposed work

In this scope, the proposed work consists a new framework for the unified access to bio-signal databases and the accompanying metadata. It allows decoupling information retrieval from actual underlying datasource structure and enables transparent content based searching from multiple data resources with context filtering. The proposed work provides a reconciled view of different ECG repositories through the use of ontologies, ECG domain standards and Unified Medical Language System (UMLS) [14].

To demonstrate the potential of the proposed methodology ROISES (Research Oriented Integration System for ECG Signals) framework was developed. Within ROISES development four interesting points were distinguished:

- The definition of a method to integrate diverse biosignal datasources.
- The introduction of a standardized global ontology based on broadly accepted ECG domain standards and UMLS huge variety of source vocabularies and terminologies to virtually aggregate the diverse resources.
- The use of ontologies to support the decoupling of ECG information from the sources.
- The specification of a terminology enhancement method, based on UMLS lexicon and data mining classification algorithms, to support the validity and extensibility of the integration scheme.

The remainder of the paper is organized as follows. Section 2 presents the proposed methodology and ROISES system. Section 2.4 specifically refers to the tools and technologies employed in ROISES. Section 3 elaborates on the results in terms of methodology application and system demonstration. Finally Sections 4 and 5 include the discussion and conclusion part of the paper respectively.

2. Materials and methods

Our approach addresses the problem of content based unified query expression on diverse and heterogeneous bio-signal sources. The approach consists of the following methodological steps:

- The generation of a conceptual model comprising the global schema the source schema and the mapping component.
- The definition of the global ontology, which provides an aggregated view of the underlying sources and allows for query definition with respect to its terminology.
- The specification of the mapping ontology to accommodate the mappings among the global ontology and the source schemas.

- The instantiation of the mapping ontology with schemas, synonyms and mappings via the semi-automatic mapping mechanisms.

2.1. Conceptual model

According to [15], the main components of a data integration system are the global schema, the sources, and the mapping. A data integration system I is formalized in terms of a triple (G, S, M) , where G is the global schema, which provides a reconciled, integrated and virtual view of the underlying sources, S is the source schema, which contains the real data, M is the mapping between G and S , constituted by a set of assertions establishing the connection between the elements of the global schema and those of the source schema. Two basic approaches have been proposed in the literature. The first approach called global-as-view (GAV), requires the global schema is expressed in terms of the data sources. The second approach, called local-as-view (LAV), requires the global schema to be specified independently from the sources. The relationships between the global schema and the sources are established by defining every source as a view over the global schema.

ROISES framework, following the LAV approach, instantiates the data integration conceptual model by defining the source schema, the global schema and the mapping as follows.

The global schema: The global schema in ROISES framework is expressed in terms of an ontology which by describing domain knowledge in electrocardiography, aims at expressing efficient combinatorial queries towards diverse ECG sources. Global ontology (GO) is based on ecgML [16], which is a markup language for supporting ECG data exchange and analysis. According to ecgML, GOs hierarchical structure conforms to existing recommendations and data format specifications such as SCP-ECG, FDA XML [17] and DICOM (Digital Imaging and Communications in Medicine) – Waveform standard [18].

The source schema: As mentioned previously, the sources represent real data, comprising ECG biosignals stored in diverse and heterogeneous data formats. Source schemas are expressed either in terms of ontologies or as relational databases.

The mapping: In the proposed methodology the LAV approach is engaged to define the mappings between the global schema and the sources. The mapping consists of the mapping ontology and the mapping mechanism. The mapping ontology, which is analysed in Section 2.1.2, accommodates the mappings between the source schemas and the global schema, thus specifying how data stored at the sources are linked to the concepts and roles in the global schema. The mapping mechanism automates the instantiation of the mapping ontology and is elaborated in Section 2.1.3.

2.1.1. The global ontology

Global ontology (GO) provides the common shared and standardized description of the appropriate concepts attributes and relationships to query ECGs and the accompanying metadata. Instance storing is beyond the scope of this ontology, since actual data are retrieved from their remote repositories. Instances are solely generated during the Synonym Instantiation procedure elaborated in Section 2.1.3. Every concept in GO essentially corresponds to one or more UMLS synonyms. Every

synonym, in turn, is specified by three fundamental attributes namely CUI, CName and SemType representing Code Unique Identifier (CUI), Concept Name (CN) and Semantic Type (ST) UMLS concepts.

In global ontology taxonomy, ecgML elements are modeled as classes while ecgML attributes correspond to either classes or properties. Element attributes which constrain the accepted value range to a predefined set of values, like Sex (Female, Male, Unknown, Undefined) or Race (Caucasian, Oriental, Black, Unknown), are modeled as classes subsuming the attributes' values. Free text attributes like first name, last name, acquisition date and time, etc. also preserve their datatype in GO. A partial view of GOs top class hierarchy is illustrated in Fig. 1.

Apart from ecgML hierarchy and terminology, GO is also enhanced with taxonomies originating from either standardized terminological systems or widely accepted medical data formats. ECG Lead Definition class subsumes a number of classes defining ECG Lead Identification Codes originating from the updated version 2.2 of the SCP-ECG standard. Patient Medical History concept subsumes a set of hierarchically defined classes organized under Diseases, imported in the ontology through UMLS plugin, as shown in Fig. 2. Drugs are organized in categories like Antiarrhythmic, Diuretics, etc. which in turn include drug descriptions like Amiodarone, Dilantin and Potassium Chloride, Thiazide respectively. Annotation subsumes PhysionetAnnotation and WaveNotation both describing annotation naming conventions and codings.

2.1.2. The mapping ontology

A mapping may be expressed as associations, constraints, rules, templates, with parameters that must be assigned during the mapping or other forms yet to be determined. More specifically, in the current application, mappings are established among ontologies and between ontologies and database models.

Considering the conclusions induced by [19] concerning RDB to RDF mapping approaches and those yielded by [20,21,26,27] pertaining ontology to ontology mapping and carefully examining popular mapping tools such as PROMPT [22], MapOnto, OntoMerge, AUTOMS [23], Chimaera [24], ONION [25], we eventuated that building our own mapping mechanism would be more efficient and less time consuming, since it would serve our own specific needs. In this scope, we developed the mapping ontology and the mapping mechanism to automatically determine linguistic similarity among concepts of the global ontology and the candidate ECG repositories. To accomplish this task, we employed UMLS Metathesaurus to retrieve the synonyms of the concepts located in the candidate ECG repositories.

A version of a mapping ontology (MO) is represented as a triple $MO = (G, T, S)$, where G describes GOs schema and metadata, T describes the target structure and S is the set of synonyms corresponding to G and T concepts. g , t and s represent instances regarding G , T and S respectively.

Mappings are stored in the mapping ontology in the form of instances and support the query formulation process. The instantiation procedure is performed in three successive steps elaborated in Section 2.1.3.

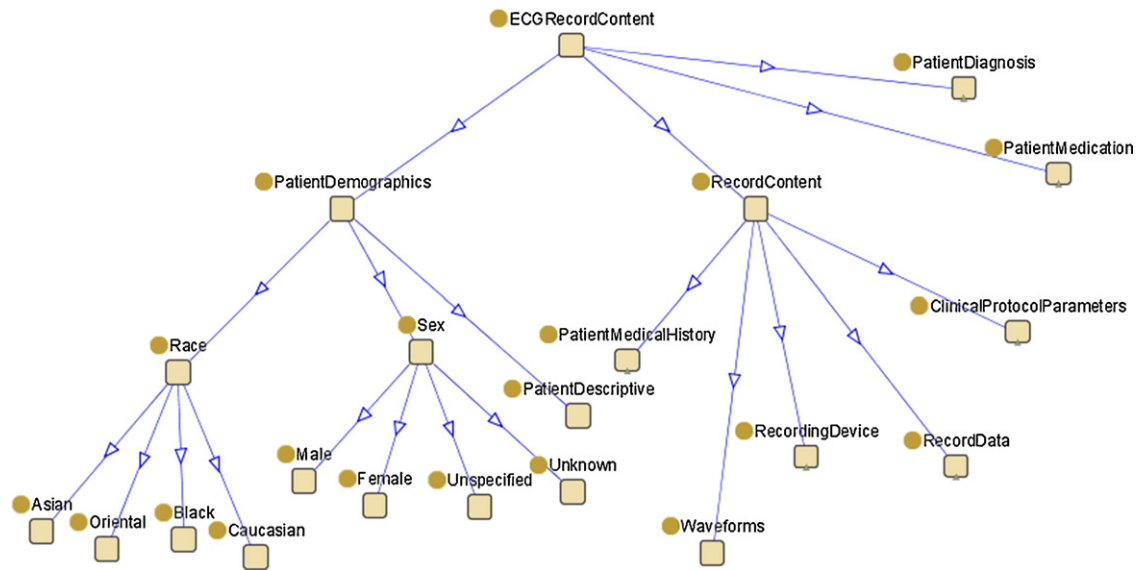


Fig. 1 – The upper class hierarchy of global ontology consisting of the expanded classes PatientDemographics and RecordContent and the collapsed – due to size – classes of PatientDiagnosis and PatientMedication.

The mapping ontology is expressed in OWL and incorporates an hierarchy of seven classes. ClassAttributes and ClassNames accommodate GOs schema and relevant metadata. TargetClassAttributes, TargetClassNames and TargetClassInstances are populated with target ontology schema information. Similarly, classes TableNames, TableProperties and TableRecords store database schema and metadata. The

classes are associated with each other via mapsWith object property which relates the instances of GO to the instances of the target schemes. Class CUI contains synonyms that correspond to mapping ontology instances. Each synonym instance is essentially described by a CUI code, a CUI description and a Semantic Type, all retrieved by UMLS Knowledge Server.

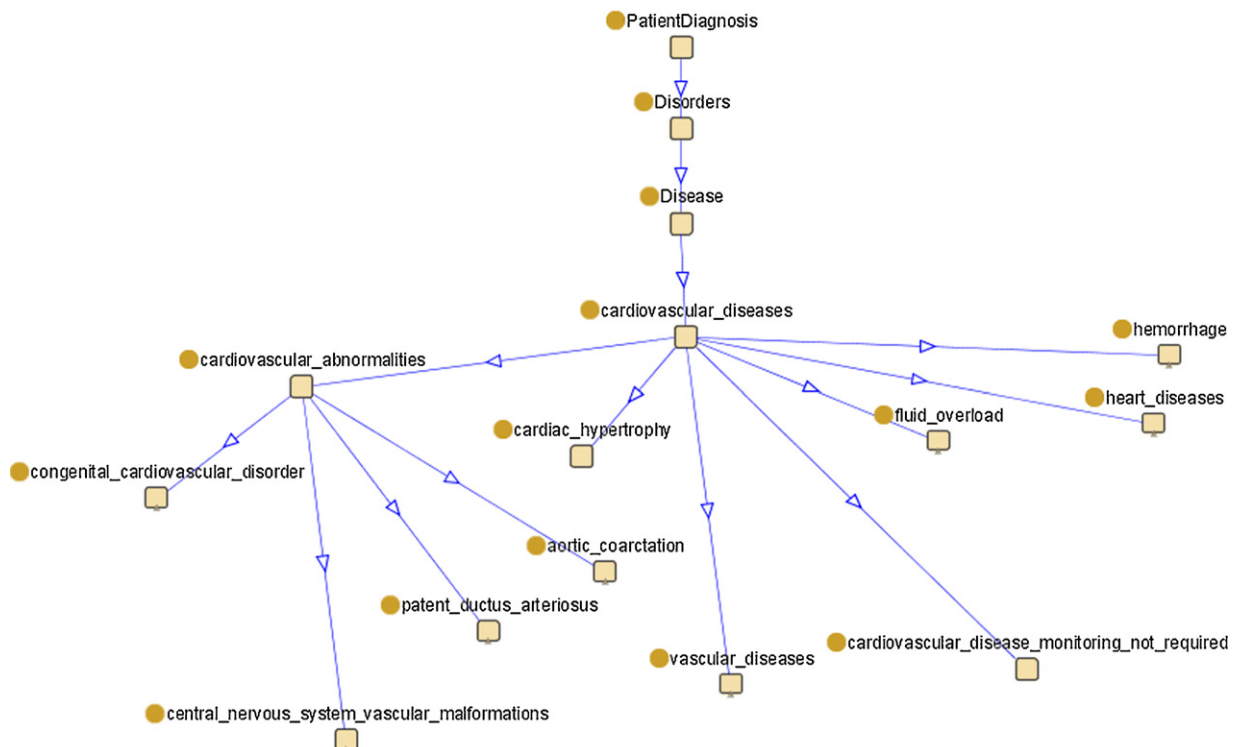


Fig. 2 – PatientDiagnosis taxonomy subsuming disorders and disease. The rest of the hierarchy is imported directly from UMLS.

2.1.3. The mapping mechanism

The mapping module comprises three interdependent components:

- (a) Schema instantiation component (SCIC) is developed to automatically extract the structures of the global ontology and the sources and transform them to instances in the mapping ontology. SCIC accepts ontologies or databases as input and automatically generates instances of their logical schema. Possible essential initialization records and instances are elucidated in a subsequent step, where the user annotates the appropriate mapping ontology instances in order to be included in the procedure.
- (b) Synonym instantiation component (SYIC) retrieves the relevant CUI codes for mapping ontology instances. SYIC connects to the UMLS Knowledge Source Server (UMLSKS), carefully scanning the instances of the mapping ontology to locate concept unique identifier (CUI)/concept name pairs. The module retrieves additional details about the concepts like concept name, synonyms, semantic type, semantic group and definition. The extracted CUIs instantiate class CUI while relevant details fill the corresponding attributes. In order to validate the resulting synonyms SYIC requires user feedback.
- (c) Mapping instantiation component (MIC) establishes the actual mappings between the instances of the participating structures. To successfully establish the mappings, MIC performs multiple comparisons per term. Not only it compares CUIs with each other but it also retrieves class names, labels and comments to be included in the mapping procedure. CUI comparisons need no further processing to be accomplished opposed to string comparisons where similarity measures are employed. Levenshtein algorithm (also called Edit-Distance) [28] was engaged as a metric for lexical similarity between strings. MIC calculates the edit distance of the source string compared to the target strings, and allows the user to adjust the required threshold. A mapping is registered if two CUIs are found identical, or a mapping suggested by MIC, is manually selected by the user. The same procedure is repeated for class ClassAttributes in order to finalize with the mapping process.

2.2. Dynamic Terminology Enhancement Method (DTEM)

During the mapping procedure elements originating from the sources are not mapped to any of GOs concepts. Therefore, possibly significant terms are excluded from the query criteria and thus from the induced results. We propose the encapsulation of unmapped terms in the terminology of the global ontology while diversifying its application domain. In this sense we developed Dynamic Terminology Enhancement Method (DTEM) to manipulate unmapped terms by dynamically incorporating them in GO. Dynamic Terminology Enhancement Method (DTEM) is performed in two sequential steps namely concept suitability investigation and hierarchy suitability investigation, to primarily determine the suitability of a medical term as a potential class in the global ontology

hierarchy and secondarily, given the suitability of the medical term, elucidate its location in the hierarchy.

2.2.1. Concept suitability investigation

DTEM, in the first step of concept suitability investigation, manipulates UMLSs huge variety of source vocabularies, terminologies, and hierarchies to compare the required medical term to its sources. A medical term is compared to UMLS Metathesaurus to locate CUI/CName pairs, considering it does not correspond to a respective one in GO. Additionally, DTEM identifies a number of features to discriminate between suitable or inapt synonym results in terms of semantic and syntactic similarity. Support Vector Machines (SVMs) method is adopted for the classification of the relevant and the irrelevant synonym results.

For a medical term i , the set of features takes the value of vector $V_{Fi} = (f_1, f_2, f_j, \dots, f_N)$, where f_j is the value of the j th feature, and N is the number of collected features, with $N > 0$. To classify the relevant from the irrelevant terms Support Vector Machines (SVMs) are employed, selecting kernels to range from Linear to Radial Basis Function (RBF).

As far as datasets are concerned the classification task engages training and testing datasets. A training set is a set of $D = \{(V_{Fi}^1, y^1), \dots, (V_{Fi}^M, y^M)\}$, where M is the number of training data instances $f \in \mathbb{R}$, $y \in \{-1, 1\}$ and determines if a concept is a potential class (1) or not (-1). Given a testing dataset, $T_s = (V_{F_k}, k = 1, 2, \dots, L)$ where L is the number of data, an SVM classifier predicts the y_i values.

The elements in the feature vectors, which are numerical indicators of the semantic and syntactic distance between two medical terms, are defined as follows:

- $f_1 \in \mathbb{R}^+$ is the Edit Distance between a medical term and its UMLS synonym,
- $f_2 \in \{0, 1, 2\}$, where f_2 equals 0 when the Semantic Type/Group of the medical term and the UMLS synonym are irrelevant, 1 when they belong to the same Semantic Group, 2 when they share the same semantic type, and
- $f_3 \in \mathbb{R}^+$ is the number of UMLS results returned for a medical term.

To evaluate the accuracy of a classifier, three measures are typically employed, namely the sensitivity (sens.), the specificity (spec.) and the accuracy.

2.2.2. Hierarchy suitability investigation

Finalizing with concept suitability investigation, leads to DTEMs second step pertaining the insertion of the previously selected suitable classes into GOs hierarchy. To accurately determine the appropriate insertion points, UMLS Knowledge Source Server (UMLS KSS) is adopted to retrieve the list of indented hierarchies of the ancestors of a concept's context. The results represent a list of ancestries for the required concept/CUI, originating from different vocabulary sources specified in UMLS Source Abbreviation field (SAB). In this work, SAB is set to "MSH", "NCI", "SNOMEDCT" corresponding to Medical Subject Headings,¹ National Cancer Institute

¹ <http://www.nlm.nih.gov/mesh/>.

Thesaurus² and Systematized Nomenclature of Medicine Clinical Terms³ respectively, since they most efficiently resemble GOs hierarchy. To initially compare GO to the resulted hierarchies, Required Starting Class (RSC) is also introduced, to be the upper GO class under which the candidate hierarchies will be subsumed. In this sense DTEM constructs K feature vectors per medical term where K is the number of retrieved ancestor tree results per concept. The feature vector of concept i , $V_{pi} = (p_1, p_2, p_j, \dots, p_S)$, represents the features of a candidate ancestor hierarchy where p_j is the value of the j th feature, and S is the number of features, with $S > 0$. The classification, which depends on the suitability of ancestries to be encapsulated in GOs hierarchy, is performed via decision trees. The elements in the feature vectors are numerical indicators of the similarity between the retrieved taxonomy and GOs taxonomy, promoting the selection of the minimum complete ancestry. The features are defined as follows:

- $p_1 \in \mathbb{R}^+$ is the number of UMLS returned results per dictionary,
- $p_2 \in \mathbb{R}^+$ is the number of retrieved UMLS hierarchical nodes,
- $p_3 \in \mathbb{R}^+$ is the number of common nodes among UMLS resulted hierarchy and GO,
- $p_4 \in \mathbb{R}^+$ is the depth of the first occurrence of the Required Starting Concept,
- $p_5 \in \mathbb{R}^+$ is the last common node in terms of depth in UMLS result hierarchy,
- $p_6 \in \mathbb{R}^+$ is the last common node in terms of depth in GO hierarchy,
- $p_7 \in \{0, 1, 2\}$, where 0, 1 and 2 correspond to results originating from “MSH”, “NCI” and “SNOMEDCT” respectively.

The classification process produces a list of “suitable” hierarchies per concept. Each hierarchy begins with the concept under which it will be subsumed (e.g. Disease) and continues with the ancestors of the candidate class. The final decision concerning the encapsulation of the hierarchy is obtained by the user, who selects which of the suggested hierarchies, best fits in GOs taxonomy. Hierarchy addition, triggers the execution of hierarchy suitability investigation procedure, to recalculate the values of the feature vectors. The procedure ends, when all the candidate hierarchies are inserted into GO.

In this study, we investigate three classification techniques: two decision tree algorithms and Naïve Bayes [29]. Among the plethora of algorithms for constructing decision trees, the alternating decision tree (ADTree) [30] and the C4.5 [31] algorithms are selected.

Alike SVMs, the performance metrics of accuracy, precision and recall are employed to compare the results of the above mentioned techniques.

2.3. ROISES Architecture

ROISES is organized in three inter-operating layers namely the UI Layer, the Mediation Layer and the Repository Layer. The overall architecture of ROISES is depicted in Fig. 3.

UI Layer: The UI Layer comprises the Interrogator UI, the Results UI and the Conversion and Management UI. Users can manage, query and convert the underlying sources through the software applications that reside in this Layer. The Conversion and Management UI allows end-users to open and edit binary ECG files and store them as ontology instances in the relevant ontology. Interrogator enables end users to express integrated queries against the participating sources. The query criteria are dynamically created by dragging and dropping the nodes of GO to the interrogator. The query mechanism and the conversion mechanisms adhere to the Mediation Layer which is elaborated below.

Mediation Layer: Mediation Layer represents ROISES' core mechanism. It supports two basic functionalities: repository population functionality and query management functionality.

Repository population functionality: ECG binary files stored in open source and structured bio-signal data formats are converted to OWL instances, which are then stored to the appropriate ontology scheme. ECG2OWL conversion mechanism automatically extracts the textual content and context from ECG binary files and populates the ontologies.

Query management functionality: Query management functionality encapsulates query transformation modules. Every signal source is essentially associated to a query transformation module. This module interacts with the criteria defined by the end user, stemming from GO, the instances originating from the mapping ontology and an XML file enclosing paths among different concepts of the target ontology, to construct efficient dynamic combinatorial queries. The same rational applies for database sources where information is obtained from the logical schema.

Repository layer: The Repository layer consists of the ECG bio-signal sources and the ontologies. The resources are collections of ECG files and ECG relational databases, which can potentially derive from domain-specific views of Electronic Medical Record databases or other patient related databases. The ontologies reflect the conceptual models of the participating data formats.

2.4. Hardware and software specifications

The ontologies incorporated in ROISES were written in OWL [32] Language, since it is currently the most widely adopted standard in ontology languages Protégé 3.3.1 [33] was selected as ontology editor, and Racer⁴ as Description Logic Reasoner, to check the consistency of the ontology and automatically compute the ontology class hierarchy. All the components of the Mapping Mechanism, and the Conversion and Management mechanisms are Java based software programs supported by OWL API open source Java library for accessing and manipulating OWL data models. SPARQL query engine included in Jena Semantic Web Framework⁵ was employed for the Query transformation mechanism while SPARQL⁶ was selected as RDF query language. Access to UMLS Knowledge

² <http://ncit.nci.nih.gov/>.

³ <http://www.nlm.nih.gov/research/umls/Snomed/snomed-main.html>.

⁴ <http://www.sts.tu-harburg.de/~r.f.moeller/racer/>.

⁵ <http://jena.sourceforge.net/>.

⁶ <http://www.w3.org/TR/rdf-sparql-query/>.

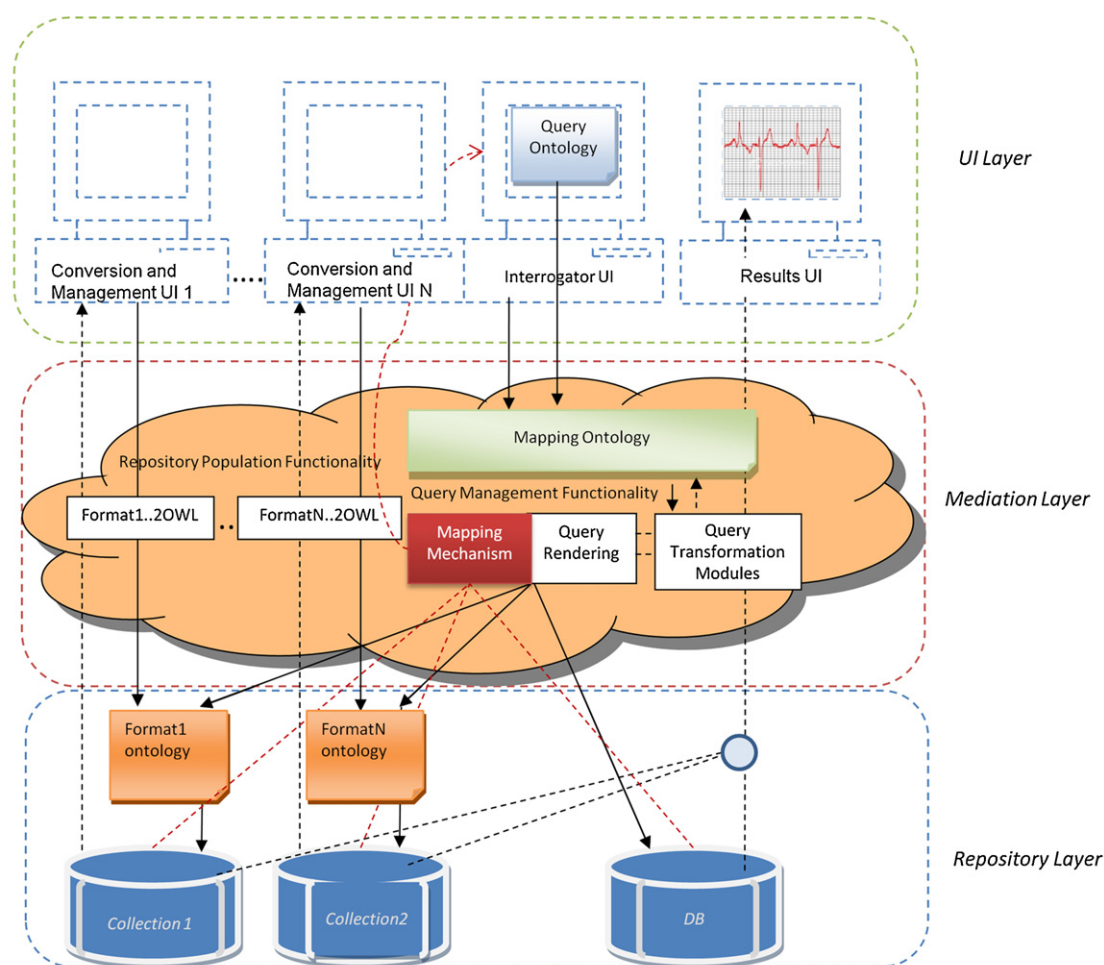


Fig. 3 – ROISES Architecture consisting of three layers, namely the UI Layer the Mediation Layer and the Repository Layer.

Server and Metathesaurus was established via the Application Programmer Interface (API) provided by UMLS. Jambalaya⁷ Protégé plugin was used to visualize the global ontology. To efficiently handle free texts during the DTEM procedure, the SPECIALIST text tools⁸ were also employed from UMLS. Access to Support Vector Machine Software was established through LibSVM library [34]. Weka toolkit was engaged to support the experiments with the three data mining algorithms [29].

3. Results

The application of the proposed methodologies/mechanisms to three diverse biosignal sources induced the results elaborated in the following subsections.

3.1. Method application

Source description: The ECG sources comprise two repositories of binary formatted files encoded in SCP-ECG and EDF plus specifications and a relational Database.

The Database constitutes part of the INTERLIFE patient and home monitoring system, which was used for the management of patients with diabetes and chronic heart disease and developed during INTERLIFE EU 1ST e-TEN project [35]. The database model reflects the structure of an Electronic Health Record specialized for diabetes and chronic heart disease. The database stores patient related information like demographics, diagnosis, medication, measurements and 1-lead ASCII formatted ECGs. The relational database was implemented in MS-SQL DBMS.

SCP-ECG files originate from a testing data set coming from real implementations and provided by openECG Consortium. Given the dearth of freely available, contextually enhanced, real SCP-ECG data, we generated several artificially constructed datasets to efficiently test the query process, by collating existing SCP data with annotations and medical information available in Physionet's PTB diagnostic database [36].

The third resource includes EDF files originating from PhysioBank. The collection of ECGs was acquired from the MIT-BIH Arrhythmia Database [37] and transformed to EDF format via Physionet's ATM conversion functionality. Since Physionet does not support writing EDF+ annotation streams, signal annotations were overwritten during WFDB to EDF conversion

⁷ <http://www.thechiselgroup.org/jambalaya>.

⁸ <http://SPECIALIST.nlm.nih.gov/textTools>.

process. In order to incorporate WFDB annotations in EDF+ formatted files, data pre-processing⁹ was adopted.

Source transformation: Two new ontologies were developed for the decoupling of ECG information from SCP-ECG and EDF encoded files, namely SCP-ECG OWL ontology [38] and EDF ontology. Both of them adhere to the specifications of the respective formats.

Additionally, two software programs were implemented to facilitate the SCP-ECG to OWL and EDF to OWL conversion processes. SCP-ECG V/W/C (Viewer/Writer/Converter) had already been implemented in the scope of openECG contest¹⁰ participation, to read SCP-ECG files with relevant content and convert them to OWL instances in SCP-ECG ontology. EDF files were managed via open source Java EDF Reader,¹¹ which was actually mutated to incorporate ECG annotations and serve the EDF to OWL conversion process.

Source schema instantiation: Three mapping ontologies were instantiated to express the correspondences among GO and SCP-ECG, EDF and Interlife database. Schema instantiation Component (SCIC) registered 244 instances subsumed under ClassNames and 34 instances under ClassAttributes for GO. 585, 140 and 40 were the instances of the records, attributes and tables corresponding to Interlife database. SCP-ECG ontology resulted in 213 and 131 instances for classes and attributes respectively. Finally EDF ontology contributed 54 class and 17 attribute instances in the related mapping ontology.

Synonym instantiation: The previously defined mapping ontologies and GO were instantiated with synonym concepts and relevant details by SYIC. Since the instantiation process was fully automated, the resulting synonyms were manually validated by Medical Doctors.

Mapping instantiation: The mapping procedure was executed three times, once for each mapping ontology. As far as the Interlife database is concerned, the edit distance value was initially set below 1, but was eventually adjusted to 2 since it matched terms with spelling errors. Some exemplary matches are Gender to Sex, DBP to DiastolicBP, SBP to SystolicBP, Amoidarone to Amiodarone and Drug to Medication concepts. Table 1 illustrates some of the DB-GO matches, where the first column lists the database element names, the second column shows GO's corresponding concepts and the third column displays the edit distance. Edit distance refers to the distance among either the original concept or one of its synonyms. As an example the result "MEDICATION Drugs 0" refers to the comparison between "MEDICATION" or its synonym "drug" to "Drugs" or its synonym "medication" resulting in edit distance value 0. Instances of the class TableRecords attained the majority of matches, as it was initially expected, since database records were typically registered with real world names. On the contrary table and attribute names are usually incomprehensible since they adhere to programming naming conventions. Therefore, the UMLS synonym retrieval procedure returned no matches for naming like DEV_DESCR, which actually represents the utilized device description.

Table 1 – The results of the mapping procedure applied to Interlife database. The first column lists the words, acquired from Interlife database, mapped to GO concepts – listed in the second column – via their edit distance listed in the third column.

Database names	Global ontology names	Edit Distance
Diabetes	Diabetes	0
Verapamil	Verapamil	0
Aspirin	Aspirin	0
Pindolol	Pindolol	0
Hydralazine	Hydralazine	0
Atenolol	Atenolol	0
Propranolol	Propranolol	0
Clonidine	Clonidine	0
Metoprolol	Metoprolol	0
Simvastatin	Simvastatin	0
Nifedipine	Nifedipine	0
Captopril	Captopril	0
Nitrates	Nitrates	0
Amiodarone	Amoidarone	2
Prasozin	Prazosin	2
Sex	GENDER	0
patientId	PATIENT_ID	1
Beta_Blockers	Beta_Blockers	0
DIAGNOSES	Diagnosis	0
MEDICATION	Drugs	0
DiastolicBP	DBP to	0
SystolicBP	SBP	0

MIC discovered a satisfying number of mappings among SCP-ECG and GO, since most of the former concepts were adopted by the latter as mentioned in Section 2.1.1. EDF to GO mapping procedure generated the least mappings as initially expected, considering that EDF ontology contains the least number of classes (standardized texts and textual information) compared to the other sources.

3.2. Dynamic Terminology Enhancement Method evaluation

In this section we present the results of the classification processes engaged for the evaluation of DTEM. The concept suitability investigation method employs Support Vector Machines to classify suitable from inapt concepts while hierarchy suitability investigation method engages decision trees to evaluate its accuracy. The datasets for both methods were partitioned into training and test sets, to build and test the model respectively.

3.2.1. Concept suitability investigation evaluation

Dataset: The dataset constitutes a combined set of medical terms stemming from unmapped table records related to medication, medication categories and free text entries of patient medical history, all originating from the SCP-ECG ontology. The testing set contains 85 medical terms retrieved from UMLS Metathesaurus. 76 were positive (suitable classes) and 9 were negative (inapt classes) with each medical term being represented by 3 features (see Section 2.2.1). The training set consists of 28 medical terms 9 of which were labelled as negative and 19 as positive ones.

SVM classification: To classify the suitable from the inapt classes, we applied SVMs with linear and Radial Basis

⁹ <http://www.neurotraces.com>.

¹⁰ <http://openecg.ific.cnr.it/alko.html>.

¹¹ <http://www.loria.fr/~kerkeni/jEDF.php>.

Function (RBF) kernels. Concerning the RBF kernel, grid search was applied to the training dataset, to select the best-fitting C and γ parameters within a range of $[-5, 5]$ and $[-4, 0]$ respectively, with step 1 in the log scale. Parameters C and γ were set to 2 and 0.125 respectively during the training procedure. To fairly compare the different kernels, the same training set was also employed to train the linear classifier. The generated models were then adopted to predict the target value of the data instances in the testing set. For these models, sensitivity, specificity and accuracy were employed to evaluate the method's accuracy. RBF classifier succeeded much higher sensitivity, specificity and accuracy (0.986842, 0.777778, 0.964706 (82/85), than the linear SVM classifier (0.973684, 0.666667, 0.941176 (80/85)) applied to the same training and test datasets.

After being applied, the concept suitability investigation procedure validated the following medical terms as being suitable to form a class in GO: Hyperlipoproteinemias (C0020476), Obesity (C0028754), Diabetes MellitusClass (C0011849), Hyperuricemia (C0740394).

On the other hand, the following concepts were characterized as unsuitable for class creation, since they did not match to any UMLS CUI: Struma (euthyroid), Scoliosis, Hypercholesterinemia, peripheral atherosclerosis, Hypercholesterinemia, Adiposis hepatis, Hyperlipoproteinemia Type II.

3.2.2. Hierarchy suitability investigation evaluation

Dataset: The datasets, engaged in the evaluation of hierarchy suitability investigation, were generated by UMLS Metathesaurus upon queried for the ancestry of a series of medical terms. The terms originated from patient diagnosis and medication domains and were extracted by SCP-ECG ontology and Interlife database. The training set comprises 55 instances, 49 of which being negative (−1) and 6 being positive (1). The testing set contains 61 instances partitioned into 51 negative and 10 positive ones.

Classification techniques: Two decision tree algorithms were employed for the evaluation of the hierarchy suitability investigation method, namely C4.5 and Alternating Decision

Trees. Additionally, Naïve Bayes was engaged as a representative from a different classification category. In order to have a fair measure of the classifiers' performance; all classifiers were trained with the previously defined training set. Similarly, the same test set was applied to the classifiers concerning the prediction task. The values of sensitivity, specificity and accuracy, concerning Alternating Decision Tree algorithm's performance, were 0.8, 0.727 and 91.803% (56/61) respectively. The relevant scores for both C4.5 and Naïve Bayes algorithm were 1, 0.8333 and 96.721 (59/61).

Naïve Bayes and C4.5, although representing diverse classification categories, share the same performance in terms of accuracy, sensitivity and specificity. On the other hand, ADTree, does not achieve a satisfying performance compared to the prior two classifiers. Consequently C4.5 and Naïve Bayes methods entail the efficacy of our method.

A significant conclusion induced from the experimental results is the high sensitivity value achieved by the classifiers, considering that the positive values were very few in both the training and the testing datasets. Consequently, the evaluation procedure's results, foster our method in terms of accuracy.

The application of hierarchy suitability investigation procedure produced the results depicted in Table 2. The procedure was executed again without significant change in the results. In the next two steps, the hierarchies of concept hyperuricemia (third row) and Hyperlipoproteinemia (first row) were selected. The final configuration of GOs taxonomy is illustrated in Fig. 4.

3.3. ROISES application

ROISES Interrogator UI applies to two different target groups in terms of functionality: the medical researcher, who studies the evolution of ECG parameter values, constrained by specific criteria and the medical student who seeks exemplary annotated ECGs in order to self practice in the identification of the structural and pathophysiologic abnormalities recognized in ECG patterns.

Table 2 – The results of applying hierarchy suitability investigation procedure. The second and third columns of the table list the CUIs and names of the investigating concepts. The fourth column reveals the lexical source that produced the ancestry introduced in the fifth column.

No.	CUI	CName	Lexical source	Ancestry
1	C0020476	Hyperlipoproteinemias	SNOMED CT	Disease → metabolic disease → disorder of lipoprotein and/or lipid metabolism → disorder of lipoprotein storage and metabolism
2	C0020476	Hyperlipoproteinemias	MeSH	Diseases (MeSH category) → nutritional and metabolic diseases → metabolic diseases → lipid metabolism disorders → dyslipidemias → hyperlipidemias
3	C0740394	Hyperuricemia	SNOMED CT	Disease → metabolic disease → inborn error of metabolism → disorder of purine and pyrimidine metabolism → disorder of purine metabolism → increased uric acid level
4	C0740394	Hyperuricemia	SNOMED CT	Disease → congenital disease → inborn error of metabolism → disorder of purine and pyrimidine metabolism → disorder of purine metabolism → increased uric acid level
5	C0740394	Hyperuricemia	SNOMED CT	Disease → hereditary disease → inborn error of metabolism → disorder of purine and pyrimidine metabolism → disorder of purine metabolism → increased uric acid level
6	C0028754	Obesity	MeSH	Diseases (MeSH category) → nutritional and metabolic diseases → nutrition disorders → overnutrition

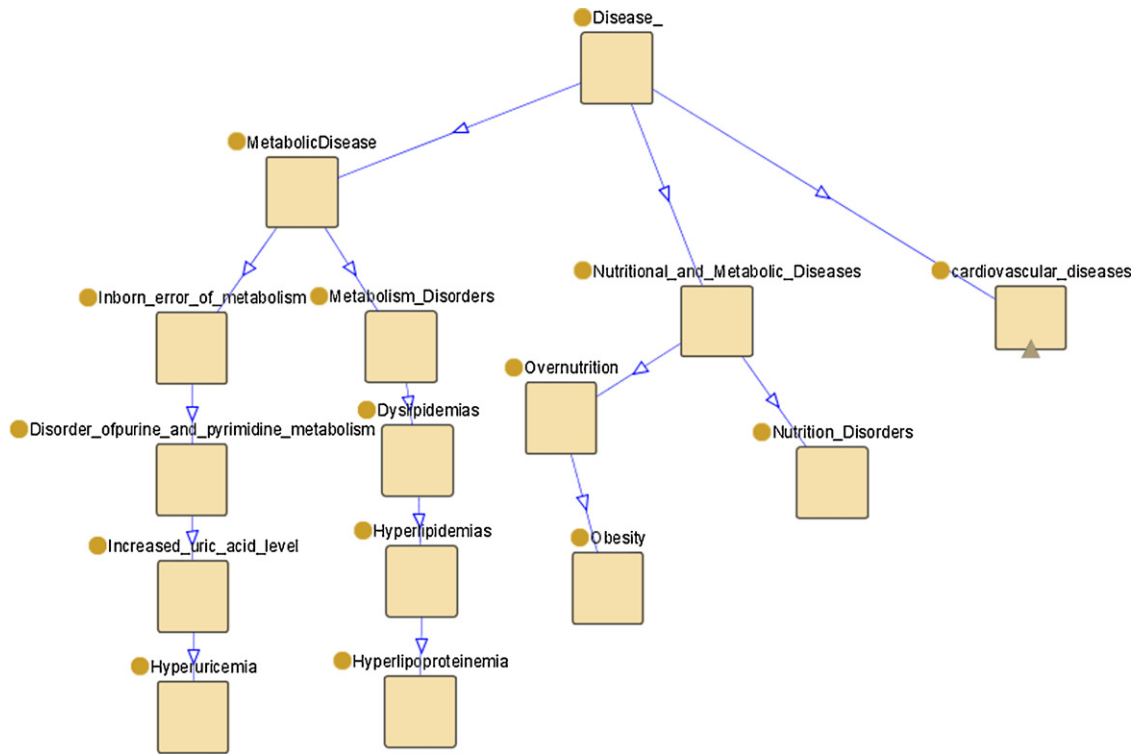


Fig. 4 – The modified Disease taxonomy, after the addition of classes MetabolicDiseases and Nutritional and Metabolic Diseases.

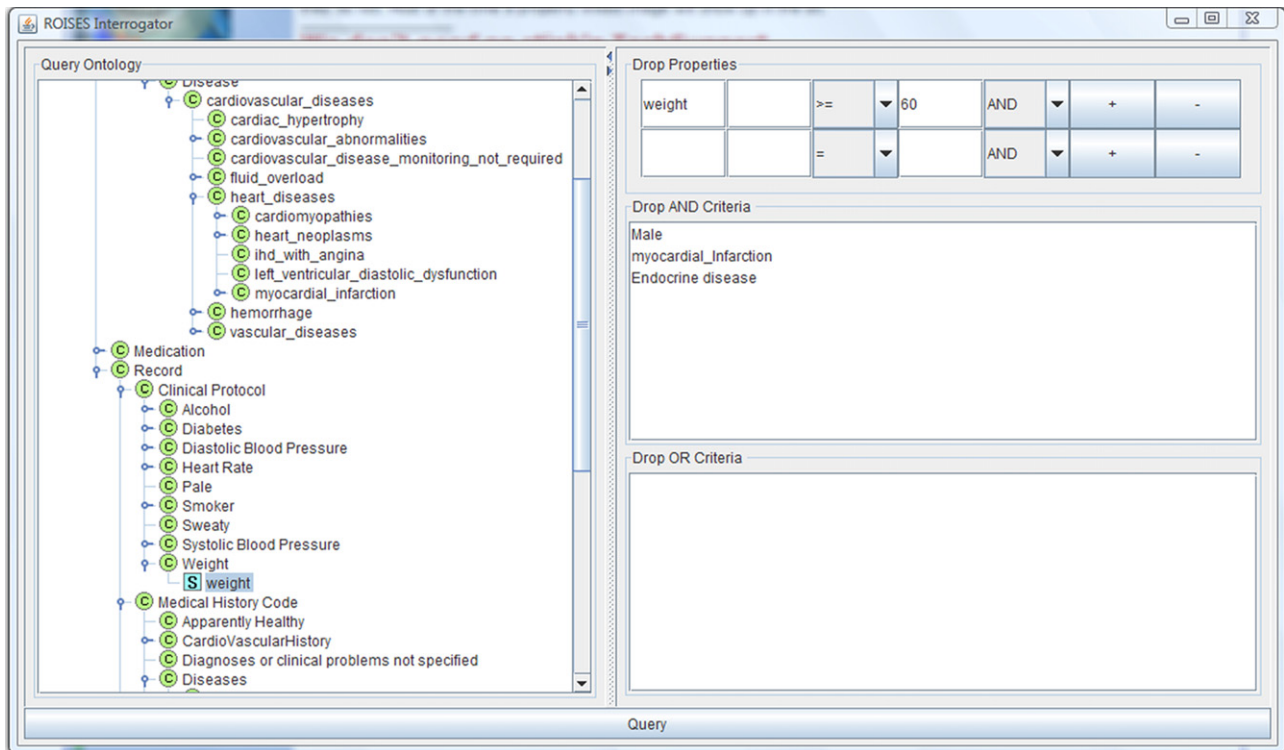


Fig. 5 – The Interrogator UI, demonstrating query a. Classes MyocardialInfarction, Endocrine.disease and Male were “dropped” in AND criteria panel. Weight property was dropped in the properties panel limited to 60 kg.

ROISES Interrogator window is divided into two panes, as depicted in Fig. 5. The left hand panel depicts GO hierarchical tree which actually introduces the search criteria for query formulation. Query criteria can be activated by simply dragging them from left to right pane. The right pane is divided in three regions: The top pane accepts blue square nodes (datatype properties) and defines criteria constrained to user specific values; the middle pane accepts green circled nodes (classes) and introduces criteria connected together with AND Boolean operator. The middle panel's functionality is also issued for the lower pane, where OR class criteria are rendered. Finally, a search box facilitates the user navigation to the complex criteria hierarchy.

The following queries illustrate Interrogator's functionalities:

- "Show ECGs with Diagnosis: Myocardial Infarction, Medical History: Endocrine.disease, Weight ≥ 60 and Sex: Male".
- "Show ECGs with number of Leads ≥ 2 , Including Lead Names: V4, V2 and Sampling Frequency = 250 Hz".
- "Show ECGs with Low Pass Filter = 30 Hz, average RRInterval (ms) > 800 ".

In order to formulate the first query, which is depicted in Fig. 5, the user drags class Myocardial Infarction which is subsumed under Diagnosis. EndocrineDisease node is also dragged from its superclass Medical History, while class Male is dropped in the middle pane. The previously defined criteria are linked with the AND operator as specified by the middle panel. Finally the user drops property weight, from class visit, in the upper panel and constrains its value to 60 kg.

Fig. 6 depicts ROISES results panel. The left panel illustrates the list of ECG files that satisfy a previously defined query while the right panel graphically displays the selected signal. Users can access additional information encompassed in the file, through the viewer's functionalities. Since ECG files are diversely formatted, visualization is performed by the invocation of the appropriate viewer (SCP-ECG, jEDF).

4. Discussion

ECG has certainly the longest tradition in bio-signal monitoring and recording, because it is a strong (amplitude near 1 mV) and relatively robust signal [39]. Electrocardiography can be monitored with various lead setups for diagnostic or treatment related purposes (e.g. signs of ischemia or heart attack, response to medication) in the clinic, in a general physician's office, and even at home or during the activities of daily life, with relatively simple and inexpensive devices. In order to draw conclusions based on the signals, patient specific characteristics, i.e. the context of the signal measurement, as well as information characterizing signal's content, i.e. detection of specific temporal events or signal features, would be required.

In general, virtually unified content based querying of different sources of bio-signals can be regarded as a step towards further signal processing and analysis. Additionally, studying a plethora of carefully annotated bio-signals can support human users or computer algorithms to extract new knowledge from specific signal patterns. In this scope, we presented a methodology and a framework for the aggregation of different types of bio-signal data sources through the definition of a

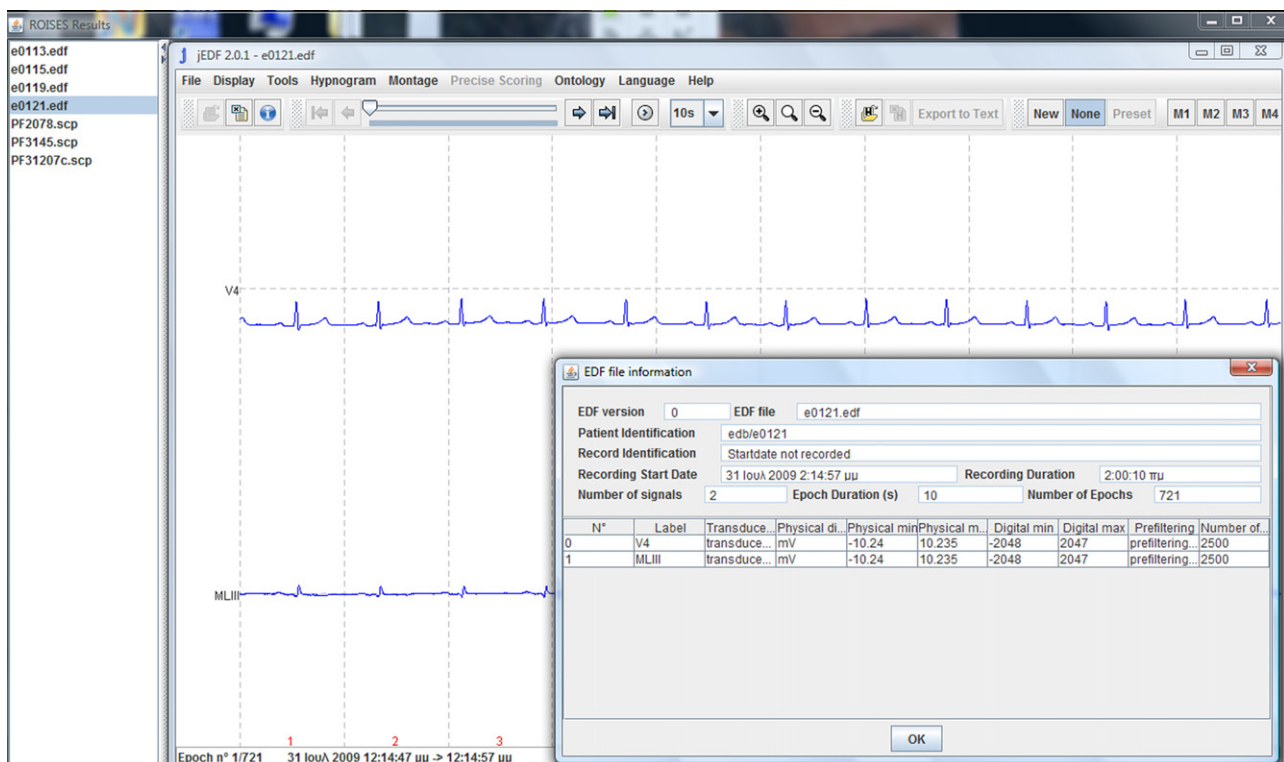


Fig. 6 – ROISES results panel. The left panel lists the ECGs which satisfy the query criteria. The central panel graphically displays the ECG with its content and context.

standardized, robust, interactive and extensible global schema employed for mapping and querying the underlying sources.

The proposed methodology currently focuses on ECG, and currently two bio-signal data formats (SCP, EDF) and two data structures (relational database, ontology) have been consolidated in ROISES framework. More specifically, if for example EEG signals were encompassed in ROISES, the global ontology would be the element to be affected the most. Specifically, some upper classes containing the word ECG would be renamed e.g. ECGRecordContent while some others, like patientDiagnosis, annotations, Disease, waveforms, patientMedication would be leveraged with EEG domain specific terminology. Although the transformations on GO are not so dramatic, they have to be accomplished by the EEG domain expert, who would select the appropriate concepts to be used as query parameters by the researcher. The alterations specified in GO do not induce any modifications in the mapping procedure and the querying mechanism, thus the newly introduced signal needs no further adjustments to be encapsulated in ROISES. It has to be stated though, that the global ontology is ECG oriented since its structure is based on ecgML. Nevertheless the previously referenced alterations are feasible, revealing that a robust and interoperable system can be easily extended horizontally or vertically.

Biosignal context constitutes a challenging issue, as it might involve a very wide medical domain along with its terminology, hierarchy and associations. In the work presented, current state of the art methods for modeling and mapping patient information have been applied. Nevertheless, there is space for improvement and extension.

As far as biosignal content is addressed, up to now, not all formats allow the embodiment of signal interpretation information in the binary file. When such information is already available, it can be queried. In [38] we proposed an approach towards invoking methods for automatic generation of signal annotations and interpretation content. Even if event annotations are incorporated in a signal format (HL7 aECG, WFDB, EDF+, etc.), the insufficiency of a standardized structure to capture events impedes integration. A prominent paradigm towards this direction is PhysioNet's annotation scheme which is well designed, easily comprehensible and provides encoded mnemonics for event types. Nevertheless, as this is a whole new domain, the enhancement of a bio-signal framework with methods for automated calculation of annotations and signal related features, towards an integrated biomedical data and function framework, is seen as one of the interesting future perspectives of this work.

The query results previously illustrated, induced some interesting considerations. The aggregation of differently formatted and structured sources of information significantly confines the expression of common query criteria. As each resource captures the content and context of the signal in various levels of detail, the definition of one or more criteria induces the exclusion of a particular resource. We therefore plan to provide the minimum and maximum requirements for the structure of the global ontology, guided by the statistical results of the query criteria selections. ROISES aims to provide the researcher with a flexible and consistent environment to define criteria not only for the queries but also for the kind of parameters to appear in the results.

5. Conclusion

Technological advances applied in the medical domain have generated huge amounts of information, which if efficiently exploited by the research community, could usher the adoption of new medical procedures in healthcare domain. In this sense, we presented a new framework for the unified access to bio-signal sources and the accompanying metadata. It allows decoupling information retrieval from actual underlying data source structure and enables transparent content based searching from multiple data resources. Three different types of resources were virtually aggregated under a global ontology, SCP-ECG and EDF+ formatted files and a relational ECG database. ROISES exploits the global ontology to support the seamless definition of unified queries towards the heterogeneous sources. Although there are still several challenges to be met, as discussed above, the framework presented is considered as an approach addressing the major direction of modeling, mapping, querying in a novel and efficient manner, raising also the concepts of content and context in the biosignal querying domain.

It is believed that the methodology and the framework presented consists an approach that contributes towards the generation of an advanced biosignal resource space, in terms of both data and services. Such an integrated resource could be of great value to the biomedical community, in terms of research, education and practice.

Authors contributions

Kokkinaki Alexandra drafted the manuscript in the scope of her PHD study on the development of an integrated interoperable framework for the unified access to multimedia medical data. Chouvarda Ioanna participated in the design and coordination of the study and the manuscript. Maglaveras Nicos conceived and coordinated the study, and helped to draft the manuscript. All authors read and approved the final manuscript.

Statements on conflicts of interest

None declared.

Acknowledgements

The authors wish to acknowledge ASSIST¹² project for its contribution to the query formulation process.

REFERENCES

- [1] A.H. Kadish, A.E. Buxton, H.L. Kennedy, B.P. Knight, J.W. Mason, C.D. Schuger, C.M. Tracy, A.W. Boone, M. Elnicki, J.W. Hirshfeld Jr., B.H. Lorell, G.P. Rodgers, H.H. Weitz, ACC/AHA clinical competence statement on electrocardiography and

¹² <http://assist.iti.gr/>.

- ambulatory electrocardiography. A report of the ACC/AHA/ACP-ASIM Task Force on Clinical Competence (ACC/AHA Committee to Develop a Clinical Competence Statement on Electrocardiography and Ambulatory Electrocardiography), *J. Am. Coll. Cardiol.* 38 (2001) 2091–2100.
- [2] A. Schlögl, An overview on data formats for biomedical signals, *Image Process. Biosignal Process. Model. Simul. Biomech.* (2009) 1557–1560.
 - [3] C. Zywiec, R. Fischer, Integrated content and format checking for processing of SCP ECG records, *Comput. Cardiol.* 31 (2004) 37–40.
 - [4] OpenECG Consortium, SCP-ECG Standard Last Work Document, 2005, Available at <http://www.openecg.net/member/EN1064.lastworkdocument.pdf> (accessed 20.1.10).
 - [5] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* 101 (2000) 215–220.
 - [6] B. Kemp, J. Olivan, European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data, *Clin. Neurophysiol.* 114 (2003) 1755–1761.
 - [7] S. Yoo, D. Rho, G. Cheon, J. Choi, A central repository for bio-signal data, in: *IEEE International Symposium on Bio-Informatics and Biomedical Engineering Proceedings*, 2008, pp. 275–277.
 - [8] P.C.Y. Sheu, B. Cummings, C. Cotman, C. Chubb, L. Hu, T. Wang, J. Johnson, S. Mobley, T. Stitch, Y. Inagaki, An object relational approach to biomedical database, in: *IEEE International Symposium on Bio-Informatics and Biomedical Engineering Proceedings*, 2000, pp. 91–98.
 - [9] T. Agorastos, V. Koutkias, M. Falelakis, I. Lekka, T. Mikos, A. Delopoulos, P.A. Mitkas, A. Tantsis, S. Weyers, P. Coorevits, A.M. Kaufmann, R. Kurzeja, N. Maglaveras, Semantic integration of cervical cancer data repositories to facilitate multicenter association studies: the ASSIST Approach, *Cancer Inform.* 8 (2009) 31–44.
 - [10] Y. Vadim, R.C. Bichutskiy, K.B. Rainer, H. Richard, Lathrop heterogeneous biomedical database integration using a hybrid strategy: a p53 cancer research database, *Cancer Inform.* 2 (2006) 277–287.
 - [11] L. Marenco, T.Y. Wang, G. Shepherd, P.L. Miller, P. Nadkarni, QIS: a framework for biomedical database federation, *J. Am. Med. Inform. Assoc.* 11 (2004) 523–534.
 - [12] M. O'Connor, R. Shankar, S. Tu, Using Semantic Web technologies for knowledge-driven querying of biomedical data, *Lect. Notes Artif. Intel.* (2007) 267–276.
 - [13] L. Marenco, R.X. Wang, P. Nadkarni, Automated database mediation using ontological metadata mappings, *J. Am. Med. Inform. Assoc.* 16 (2009) 723–737.
 - [14] D.A.B. Lindberg, B.L. Humphreys, A.T. McCray, The unified medical language system, *Methods Inform. Med.* 32 (1993) 281–291.
 - [15] M. Lenzerini, Data integration: a theoretical perspective, in: *21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS02)*, Proceedings, 2002, pp. 233–246.
 - [16] H. Wang, F. Azuaje, B. Jung, N. Black, A markup language for electrocardiogram data acquisition and analysis (ecgML), *BMC Med. Inform. Decis. Mak.* 3 (2003) 3–4.
 - [17] FDA XML Data Format Design Specification, 2010, Available at http://www.cdsc.org/discussions/EGC/FDA_XML_Data_Format_Design_Specification_DRAFT_B.pdf (accessed 20.1.10).
 - [18] DICOM Suppl. 30, Waveform Interchange, Nat. Elect. Manufacturers Assoc.: ARC-NEMA, Digital Imaging and Communications, NEMA, Washington DC, 1999.
 - [19] S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau, S. Auer, J.F. Sequeda, A. Ezzat, A Survey of Current Approaches for Mapping of Relational Databases to RDF, W3C RDB2RDF XG Report, 2009, Available at http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_Survey_Report.01082009.pdf (accessed 20.1.10).
 - [20] N.F. Noy, M.A. Musen, Evaluating ontology-mapping tools: requirements and experience, in: *Workshop on Evaluation of Ontology Tools at EKAW'02 (EON2002)*, Proceedings, Sigüenza, Spain, 2002, pp. 1–14.
 - [21] S.M. Falconer, N.F. Noy, M.A. Storey, Ontology mapping – a user survey, in: *The Second International Workshop on Ontology Matching at ISWC 07 + ASWC 07*, Busan, Korea, 2007.
 - [22] N.F. Noy, M.A. Musen, PROMPT: algorithm and tool for automated ontology merging and alignment, in: *Seventeenth National Conference on Artificial Intelligence (Aaai-2001)*, *Twelfth Innovative Applications of Artificial Intelligence Conference (Iaai-2000)*, 2000, pp. 450–455.
 - [23] K. Kotis, A. Valarakos, G. Vouros, AUTOMS: Automating Ontology Mapping through Synthesis of Methods, in: *OAEI (Ontology Alignment Evaluation Initiative) 2006 Contest*, *Ontology Matching International Workshop*, Atlanta, USA, 2006.
 - [24] D.L. McGuinness, R. Fikes, J. Rice, S. Wilder, An environment for merging and testing large ontologies, in: *Proc. 17th Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'2000)*, Colorado, USA, April 2000, pp. 483–493.
 - [25] P. Mitra, G. Wiederhold, M. Kersten, A graph-oriented model for articulation of ontology interdependencies, in: *Advances in Database Technology-EDBT 2000*, Proceedings 1777 (2000) 86–100.
 - [26] Y. Kalfoglou, M. Schorlemmer, Ontology mapping: the state of the art, *Knowl. Eng. Rev.* 18 (2003) 1–31.
 - [27] N. Konstantinou, D.E. Spanos, N. Mitrou, Ontology and database mapping: a survey of current implementations and future directions, *J. Web Eng.* 7 (2008) 1–24.
 - [28] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, *Sov. Phys. Dokl.* 6 (1996) 707–710.
 - [29] H.I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition, Morgan Kaufmann, San Francisco, 2005.
 - [30] Y. Freund, L. Mason, The alternating decision tree learning algorithm, in: *Machine Learning*, Proceedings, 1999, pp. 124–133.
 - [31] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, 1993.
 - [32] OWL Web Ontology Language Guide, 2010, Available at <http://www.w3.org/TR/owl-guide> (Accessed 20.1.10).
 - [33] J.H. Gennari, M.A. Musen, R.W. Fergerson, W.E. Grosso, M. Crubezy, H. Eriksson, N.F. Noy, S.W. Tu, The evolution of Protege: an environment for knowledge-based systems development, *Int. J. Hum.-Comput. Stud.* 58 (2003) 89–123.
 - [34] C.C. Chang, C.J. Liu, LIBSVM: A Library for Support Vector Machines, 2001, Available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (Accessed 20.1.10).
 - [35] N. Maglaveras, S. Maglaveras, I. Lekka, I. Chouvarda, V. Kaimakamis, V. Kilintzis, A. Prentza, Quality home telemedicine services for chronic cardiac disease patients through the INTERLIFE platform, *Comput. Cardiol.* (2006) 245–248.
 - [36] R. Bousseljot, D. Kreisler, A. Schnabel, A. Nutzung der, EKG-Signaldatenbank CARDIODAT der PTB über das

- Internet, Biomedizinische Technik, Band 40, Ergänzungsband 1 (1995) 317.
- [37] A. Taddei, G. Distanti, M. Emdin, P. Pisani, G.B. Moody, C. Zeelenberg, C. Marchesi, The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography, *Eur. Heart J.* 13 (1992) 1164–1172.
- [38] A. Kokkinaki, I. Chouvarda, N. Maglaveras, An ontology-based approach facilitating unified querying of biosignals and patient records, in: *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2008, pp. 2861–2864.
- [39] W.J. Tompkins (Ed.), *Biomedical Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1993.