Journal of Biomedical Informatics xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using an ECG reference ontology for semantic interoperability of ECG data

Bernardo Gonçalves *, Giancarlo Guizzardi, José G. Pereira Filho

NEMO – Ontology and Conceptual Modeling Research Group, Computer Science Department, Federal University of Espírito Santo, Vitória, Brazil

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Biomedical Ontology
ECG Ontology
Semantic interoperability
ECG standardization
Electronic Health Record (EHR)

ABSTRACT

In this paper we test the hypothesis that a domain reference ontology of the electrocardiogram (ECG) can be employed in an effective manner to achieve semantic integration between ECG data standards. Several standardization initiatives, namely AHA/MIT-BIH (Physionet), SCP-ECG and HL7 aECG, have led to heterogeneous conceptualizations of the ECG domain. We then argue that a shared anchor, the biomedical reality under scrutiny, can effectively support the semantic integration of these ECG standards into a coherent ECG representation for the sake of a unified Electronic Health Record (EHR) model. Our hypothesis is tested by means of an integration experiment that uses, on the one hand, an ECG Ontology and, on the other hand, elicited conceptual models of the ECG standards. As a conclusion, we attest the hypothesis and also provide an integration table depicting correspondence links between entities in the ECG Ontology and elements in the ECG standards.

1. Introduction

Biomedical data has been ever more available for computing as the use of information systems in Biology and Medicine becomes widespread. An example of biomedical data is the electrocardiogram (ECG), which is the most applied test for measuring heart activity in Cardiology. Since the 1970s, both the storage and transmission of ECG records have been object of standardization initiatives. Among the prominent ECG standards, one might refer to (i) AHA/MIT-BIH (Physionet) [1], an ECG data format extensively used worldwide in cardiac physiology research (cf. [2]); (ii) SCP-ECG [3], which is an ECG European standard that specifies a data format and a transmission procedure for ECG records (http://www.openecg.net); and (iii) HL7 aECG [4], which is an ECG data American standard adopted by FDA for clinical trials [5].

However, as we discuss in this text, the conceptualizations underlying these standards are heterogeneous in spite of the fact that all of them address the very same ECG domain. While this could be partially justified by their different purposes and requirements, their heterogeneity even with respect to the core ECG concepts indicates something else; namely, that their conceptualizations do not refer directly to the *biomedical reality* under scrutiny as a shared anchor (cf. [6,7]). In contrast, the focus of these ECG standards is mostly on how data and information should be represented in computer and messaging systems [8, p. 1252], [9, p. 254]. Such a concentration on the information world of *headers*, *ids* and *sections* rather than on the real world of *patients*, *ECG recording sessions*, *ECG waveform* does not favor a possible consensual model of ECG data in view of a unified Electronic Health Record (EHR).

With this in mind, in this paper we employ the *ontological approach* towards semantic interoperability of ECG data. We test the hypothesis that a reference ontology of the ECG domain (in particular, the ECG Ontology) can be employed in an effective manner to achieve semantic integration between ECG data standards. For reference ontology [10], we mean an ontology resulting from an application-independent representation of a given domain. In this case, the representation of *what the ECG is*, on both sides of the patient and of the physician. The ECG Ontology used in this article, furthermore, has been carefully developed over the years by employing ontological principles coming from the discipline of Formal Ontology in Philosophy [11]. As illustrated further on, such a principled ontological approach is of value for the semantic integration reported here.

Our program is then as follows. If we assume that the ECG Ontology does justice to *what the ECG is at the point of care* and solely this – i.e., regardless of technological issues that arise in representing it in a given information system – then it could be used to support the design of interoperable versions[1] of ECG data formats like AHA/MIT-BIH, SCP-ECG and HL7 aECG. By taking the ECG Ontology as a reference, the elements present in these data formats could be semantically mirrored to the ontology entities, instead of being object of pairwise mappings [12]. Thereby, the ECG data formats should meet Cimino's desiderata [13], namely: (i) *non-vagueness*,

* Corresponding author. Fax: +55 27 33352850.
E-mail address: bgoncalves@inf.ufes.br (B. Gonçalves).

[1] Naturally, we are not referring to an interoperation procedure as for messaging systems in the sense of computer networks, but to a structural disposition for exchanging information by sharing the same semantics.

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.08.007

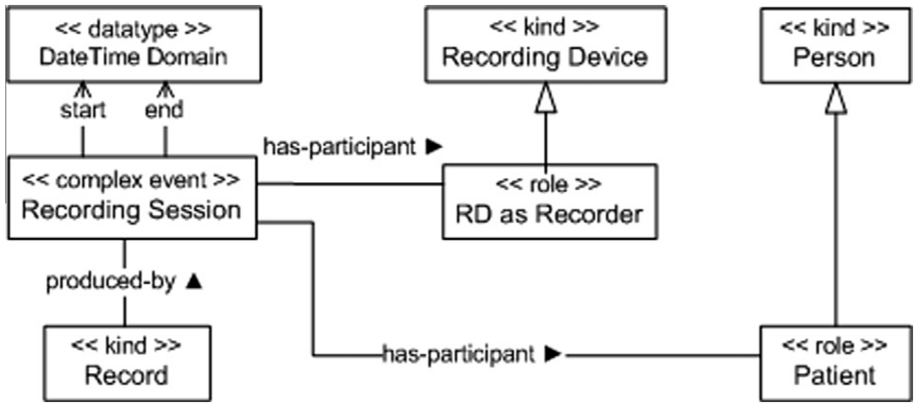

Fig. 1. Model of the ECG recording session context.

i.e., the entities which form the nodes of the data format must correspond to at least one universal in the ontology; and (ii) they must correspond to no more than one universal, i.e., *non-ambiguity*. Since the ECG Ontology axiomatization allows little freedom to both vagueness and ambiguity (cf. [14, Chapter 5]), this solution would at least force the data formats to make their assumptions explicit. Besides, the proposal is also cost-effective, since n data formats require n mappings to a reference ontology, whereas $n(n-1)/2$ pairwise mappings would be required [12, p. 311].

This methodology is applied here in an integration experiment meant either to attest or refute the hypothesis just mentioned, namely, that *an ECG reference ontology is an effective means for the semantic integration of heterogeneous ECG data formats*.

The remainder of this article is organized as follows. In Section 2, we present the fragment of the ECG Ontology which is germane to the purposes of this article. In Section 3, as an additional contribution of this article, we elicit the conceptualizations underlying the ECG standards aforementioned to then explicitly represent them as conceptual models in Section 4. Therein, we present the central contribution of this article, viz., the experiment that puts to the test whether or not the conceptual models elicited can be integrated by means of the ECG Ontology. Finally, in Section 5 we discuss (i) related work and the limitations of our own, but also the applicability of the methods proposed here; to then proceed with our conclusions in Section 6.

2. The ECG Ontology

The ECG Ontology employed in this article has been developed over the years[2] and is reported in [15]. For an in-depth discussion with a rigorous axiomatization and full documentation, the interested reader should refer to [14]. Here, in any case, we provide a brief description of this ontology, which is sufficient for the purposes of this article. This ontology uses basic relations in the biomedical ontology literature [16] and is composed of three sub-ontologies: the anatomy sub-ontology (based on the FMA [17]); the heart electrophysiology sub-ontology (based on the Ontology of Functions [18]); and finally the ECG sub-ontology. In this paper, we refer only to the latter (henceforth, just "ECG Ontology").

The ECG Ontology comprises two main models.[3] The first one is concerned with the representation of what the ECG is on the side of the patient – i.e., the ECG as it is acquired from a patient. The parts of this model which are germane to be presented here appear in Figs. 1 and 2. The second one is presented in Fig. 3, and is concerned with the ECG on the side of the physician – i.e., the waveform resulting from the process of ECG acquisition as well as the interpretation of this waveform. These models as well as the ECG standards' conceptual models are presented in this article by using UML diagrams. UML (http://www.uml.org/) is a de facto standard for creating visual conceptual models in Computer Science. For an introduction to UML, the reader should refer to [21].

In order to improve the readability of the remainder of this text, we adopt the following typesetting convention: (i) names of the entity types pertaining to the ECG Ontology are written in Sans serif typeface with Upper Case initials; (ii) relations coming from the same ontology are written in *italics*. In addition, the elements coming from this ontology which are most relevant for the purposes of this paper are listed in Table 1 alongside the corresponding definitions. The "category" column in Table 1 refers to corresponding types of the ECG entities in an underlying foundational system of categories. Most of the categories used in the ECG Ontology are well-known in the Ontology literature (see e.g., [22] for an initial overview). We use them here as defined in [23] and related publications. Those categories also appear as stereotypes that label the UML classes in the models that follow. These stereotypes are syntactical markings making explicit which (ontological) category in the metamodel predicates over the marked element. For instance, when stereotyping the entity Person with $\ll$ kind $\gg$, we are making clear in the model that Person is an example of a Kind, thus, possessing all its meta-properties (e.g., being an ultimate substance sortal which supplies a unique principle of identity for its instances). This stereotyping mechanism is not necessary to understand the models, but provide a tracing mechanism to the ontological analysis that supported our modeling decisions, thus, allowing for an in-depth ontological reading.

2.1. The ECG on the side of the patient

An ECG Record[4] is *produced by* a Recording session that has a start and end date/time (see Fig. 1). They are two properties of the session which are projected into the Date Time datatype. A session has as participants an RD as recorder (subtype of Recording device) and a Patient. Notice that the ECG recording session is an example of complex event – i.e., an event composed of two or more events. Indeed, (many) observations (or measurements, loosely speaking) are made

[2] Cf. the project website at http://nemo.inf.ufes.br/biomedicine/ecg.html.

[3] These models account for a descriptive commonsensical view of reality, focused on structural (as opposed to dynamic) aspects of the ECG. For an introduction to the domain of electrocardiography, one should refer to [19,20].

[4] Note that although the term "record" is very general even within Biomedical Informatics in particular, we refrain from assigning a rubric like "ECG record" to avoid verbosity. This is because, if that were the case, many entities in our domain would require for the "ECG" prefix as well. Nevertheless, in the ECG ontology each entity is accompanied by a unique ID with the "ecg" signature before it (cf. Table 1), and this will suffice for us.

Please cite this article in press as: Gonçalves B et al. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform (2010), doi:10.1016/j.jbi.2010.08.007

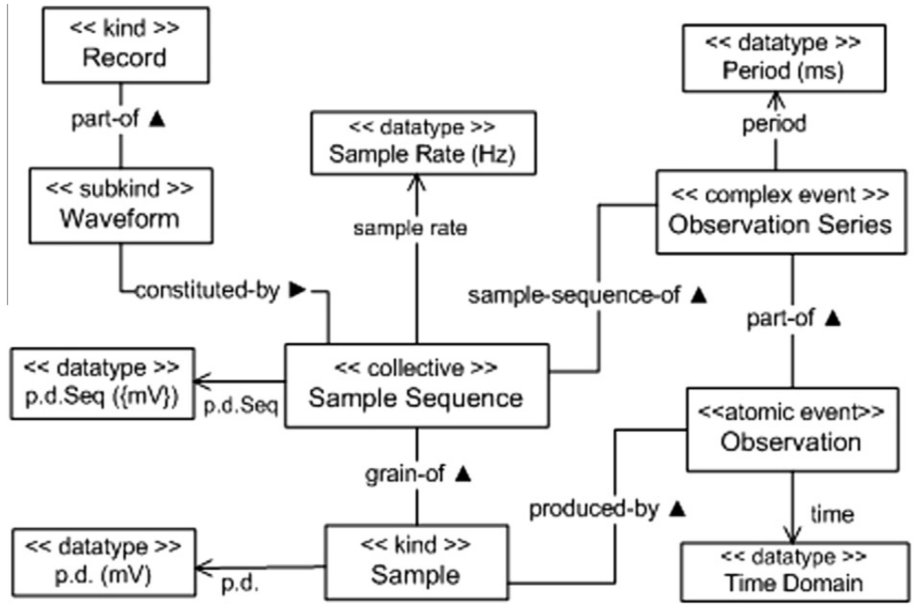

Fig. 2. Model of the ECG acquisition mechanism.

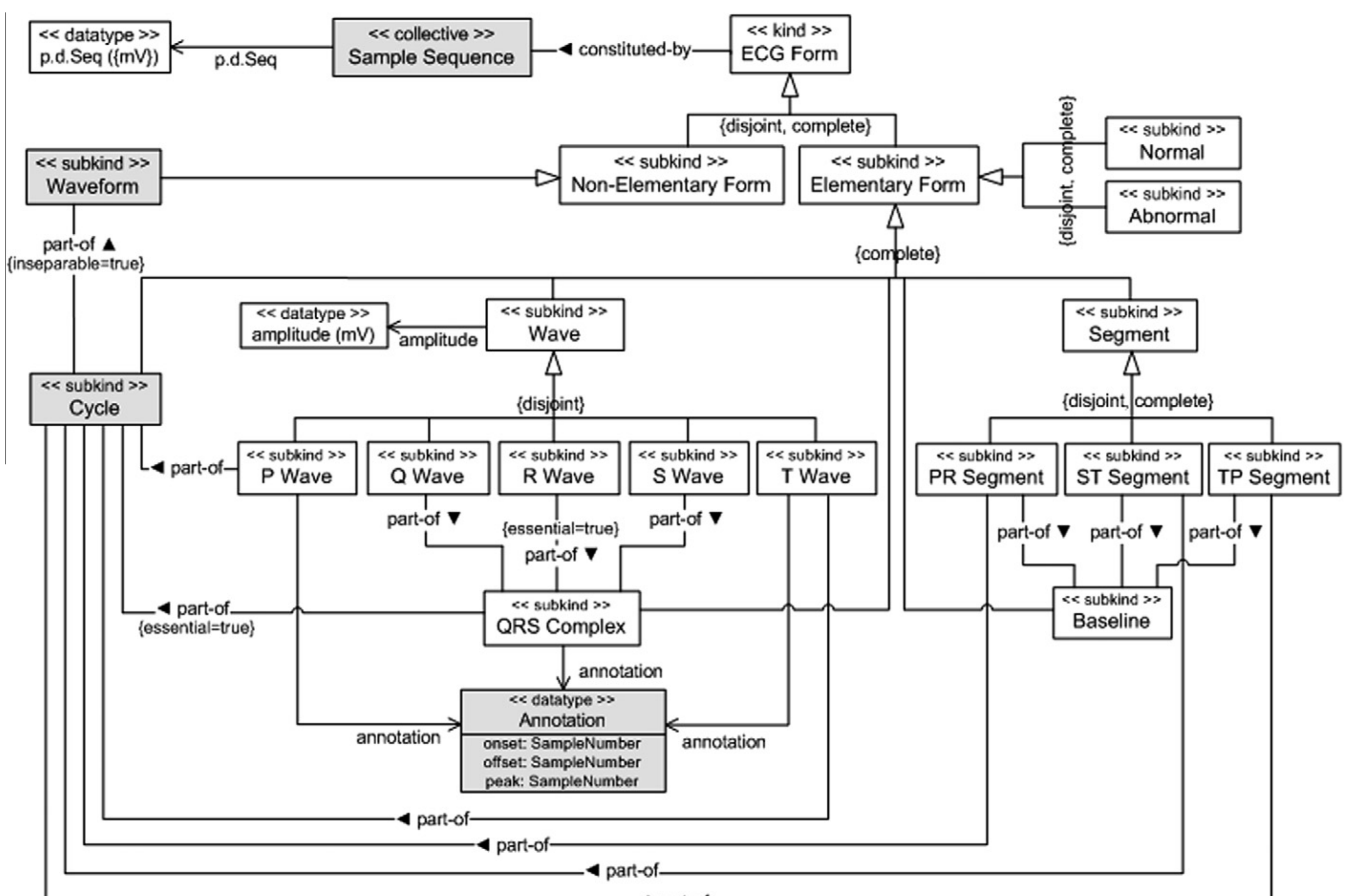

Fig. 3. Model of the ECG on the side of the physician. The entities highlighted in gray are those which are more important for the semantic integration of the ECG standards. Nonetheless, in order to make feasible their understanding, we present them here in their full context.

by the recording device between the session's temporal boundaries (see Fig. 2). These observations are actually what allows the ECG to gradually take form. Observations are atomic events, i.e., events which do not have other events as parts. They are evenly spaced in time, forming then an Observation series that lasts a non-zero period of time. The sample rate of the ECG Waveform accounts for the in-

Table 1
Entities of the ECG Ontology that are relevant for the semantic integration of ECG standards.

Term	Entity ID	Category	Textual definition
Record	ecgOnto:085	kind	ECG data record (in any medium) resulting from a Recording session and essentially composed of an ECG Waveform.
Recording session	ecgOnto:086	complex event	Medical service in which the Patient is subject of ECG recording by some Recording device. The Recording session (event) can be said to temporally coincide, albeit in a different level of abstraction, with the Observation series (event). In other words, these two events have the same time boundaries.
Recording device	ecgOnto:087	kind	Device used to acquire (to record) an ECG from a given Patient by means of electrodes. Also called electrocardiograph.
RD as recorder	ecgOnto:088	role	Recording device as it plays the role of an ECG recorder.
Patient	ecgOnto:090	role	Person as he/she plays the role of being subject of care, i.e., scheduled to receive, receiving, or having received a healthcare service (based on ISO/TC 18308:2003).
Waveform	ecgOnto:091	subkind	Non-elementary Geometric form *constituted by* the Sample sequence resulting from the Observation series which makes up an ECG Recording session.
Observation	ecgOnto:092	atomic event	Measurement of the p.d. between two regions of the patient's body. It is carried out by an ECG Recording device by means of two Electrode placements on those regions. The placements are defined according to an ECG Lead.
Observation series	ecgOnto:093	complex event	Series of observations evenly spaced in time carried out in an ECG Recording session.
Sample sequence	ecgOnto:095	collective	Ordered sequence of samples resulting from an Observation series.
Lead	ecgOnto:096	kind	Viewpoint of the heart activity that emerges from an Observation series of the p.d. between two electrode placements on specific regions of the surface of the patient's body.
Cycle	ecgOnto:105	subkind	Elementary form periodically repeated in the ECG Waveform that indirectly indicates a heart beat. The QRS complex is an *essential part of* it, as the peak of the R wave is considered a reference point to define it.

verse of that period. Sample rate values are projections into the quality structure[5] of Hertz (Hz, samples/s) by considering the period in seconds.

The observations are meant to measure electrical potential differences (p.d.) over the patient's body surface with the result of producing samples. Every observation produces an electrical potential Sample (in the scale of millivolts). The sample values are projections into the voltage (p.d.) quality structure within the set of the real numbers. Every sample is, in turn, a *grain of* a Sample sequence – the relation *grain of* [24] is adopted here since the waveform is dense, with typical sample rates of 256 or 300 Hz. A sample sequence is an ordered collective of samples, and thus, a projection into the quality structure of p.d. values as an ordered sequence (in the mathematical sense) within the set of the real numbers.

Since measuring the p.d. between two points provides only a partial point of view of the heart activity, usually multiple observations are made at the same time to capture multiple views of the heart activity. While a series of observations denotes an Observation series, multiple series that share the same structure in the time axis (i.e., the same beginning, end and period) denote a correlated observation series. Each of those viewpoints that emerge from a single Observation series defines an ECG Lead, cf. Fig. 2.

2.2. The ECG on the side of the physician

By shifting to the physician's perspective, we put in focus the objects of ECG analysis. Heart beats are mirrored to cardiac cycles that compose the ECG Waveform (see Fig. 3). A given ECG Form can be either elementary or not. Every ECG Elementary form is, as the name suggests, elementary for ECG analysis. These are the forms appearing within every cardiac cycle that either directly map a relevant and cohesive electrophysiological event in the heart behavior or connect two forms which do it. For instance, the P wave maps the depolarization of atria, while the QRS complex maps the depolarization of ventricles; the PR segment, rather, connects the former to the latter. A Non-Elementary form is, in turn, any arbitrary ECG form which is not an elementary one. An example of such ECG form can be given by a proper part of the T wave, or by the right half of the P wave and the left half of the PR segment.

The ECG Waveform is the ECG form *constituted by* the Sample sequence resulting from the full Observation series carried out in the ECG session's time boundaries, or, in other words, the ECG form resulting from an ECG Recording session. The Waveform *is a* Non-elementary form, as it as a whole is not (an elementary) object of the physician's analysis. Elementary forms are of different natures. Namely, Wave, complex (only the QRS complex), Segment and Cycle. This partition is complete, and is made under the following differentiae. While a Wave must have a peak (higher *y*-coordinate value in module), a complex, viz., the QRS complex, can be composed of one or more waves. A Segment, in turn, connects two waves and does not have a peak.

Finally, a Cycle is composed by waves and segments that connect them – the QRS complex *is an essential part of* any Cycle. In a canonical view, the combination of the P wave, PR segment, QRS complex, ST segment, T wave and TP segment composes the cardiac cycle. In practice, however, there are cycles with missing ECG elementary forms (e.g., a missing Q wave) for several reasons, e.g., because it is not well visible from the respective ECG lead.

Once more, we emphasize that the presentation of the ECG Ontology put forth here has the sole purpose of a brief overview. All the modeling decisions previously mentioned are carefully elaborated, ontologically justified and formally characterized in [14]. In what follows we concentrate on eliciting and extracting, or "excavating", the conceptualizations underlying the ECG data standards aforementioned.

3. The ECG Standards' Conceptualizations

Every design or implementation artifact commits to an underlying conceptualization. However, frequently this commitment is made in an implicit and ad hoc manner. We use here the term "Conceptual Model Excavation" to refer to the activity of, firstly, making explicit these implicit conceptualizations, followed by their representations via concrete engineering artifacts, i.e., conceptual models.

The ECG data formats are presented here in the same spirit and by employing the same jargon used by their maintainers in the corresponding specifications. We strive as much as possible to maxi-

[5] A quality structure comprises a space of values, possibly embedded with topological or geometrical features. This space of values delimits the possible information tokens that can be associated with a (quality) property. Consider the example of the "sample rate" property characterizing a "Sample Sequence" universal. Associated with this property, we have a quality structure of Hertz (Hz). This quality structure comprise an underlying (totally-ordered) quality dimension which is isomorphic to the positive half-line of real numbers. A specific qualia (member of that quality structure) is, for instance, 256 Hz. For an in-depth discussion about quality structures, please refer to [23, Chapter 6].

Please cite this article in press as: Gonçalves B et al. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform (2010), doi:10.1016/j.jbi.2010.08.007

mize fidelity to the original terminology, in order to be able to proper address a number of subtleties present in their underlying conceptualizations. For readability, the elements of the ECG standards are introduced here in sans serif typeface with lower case initials.

The ECG conceptualizations underlying the standards are extracted from their specifications based on the textual descriptions of their composing elements and relations. The conceptualizations are then represented using standard UML diagrams (Section 4). The refined process of obtaining these conceptual models by extracting the ECG standards' data models is described in depth in [14, Chapter 7].

3.1. AHA/MIT-BIH (Physionet)

In what follows, we briefly describe the AHA/MIT-BIH data format following the WFDB programming guide (version 10.4.19) [25]. A more detailed analysis of this ECG format can be found in [14].

In this standard, the ECG data is available in *PhysioNet* (http://www.physionet.org) as part of one of the databases at *PhysioBank*. The databases comprehend ECG records, each of which containing a continuous recording from a single subject. Signals are commonly understood to be functions of time, obtained by observation of physical variables. In AHA/MIT-BIH, a signal is defined more restrictively as a finite sequence of integer samples. These are usually obtained by digitalizing a continuous observed function of time at a fixed sampling frequency (expressed in Hz, i.e., samples per second). The time interval between any pair of adjacent samples in a given signal is a sample interval. All sample intervals for a given signal are equal. MIT DB records are each 30 min in duration, and are annotated throughout. This means that each beat (marked by its QRS complex) is described by a label called an annotation. The annotations describe a feature of one or more signals at a given time instant in the record. The main feature to be annotated is the type of the beat (normal, ventricular ectopic, etc.). Other types of annotations include indications of changes in the predominant cardiac rhythm, or in the signal quality. The conceptual model resulting from our excavation of this standard is depicted in the left-hand of Fig. 4.

3.2. SCP-ECG

The SCP-ECG data format has been consolidated as a standard by CEN/TC-251 and is thoroughly described in the SCP document (version N02-15) [3]. A detailed analysis of this ECG data format can also be found in [14, Chapter 7]. For brevity, we refer here only to some of the most important SCP elements.

An ECG record in the SCP format is composed of several sections. The sections containing relevant elements are listed as follows. One or more leads are identified in Section 3; the data corresponding to each of these leads is laid in the rhythm data element of Section 6. This data comprises samples ordered as they are acquired according to a sample interval. Section 1 lies in a header which contains the record acquisition date/time and an identification of the acquiring cardiograph and the patientID. Sections 4 and 5 are devoted to encode reference beats, in case they are automatically recognized by the acquiring device. Finally, Section 8 bears the record interpretation made further by a physician and incorporated in the record. The element definitions are listed in Table 2. The SCP-ECG conceptual model we have obtained is presented in the left-hand of Fig. 5.

3.3. HL7 aECG

The HL7 aECG conceptualization is extracted from the implementation guide released in 2005 [4]. In HL7 aECG, sets of leads collected simultaneously are packaged together as a complete ECG session dataset. One can have a hierarchy of derived (filtered or otherwise transformed) representations based explicitly on some defined subset of the original data, termed regions of interest (ROI) [4].

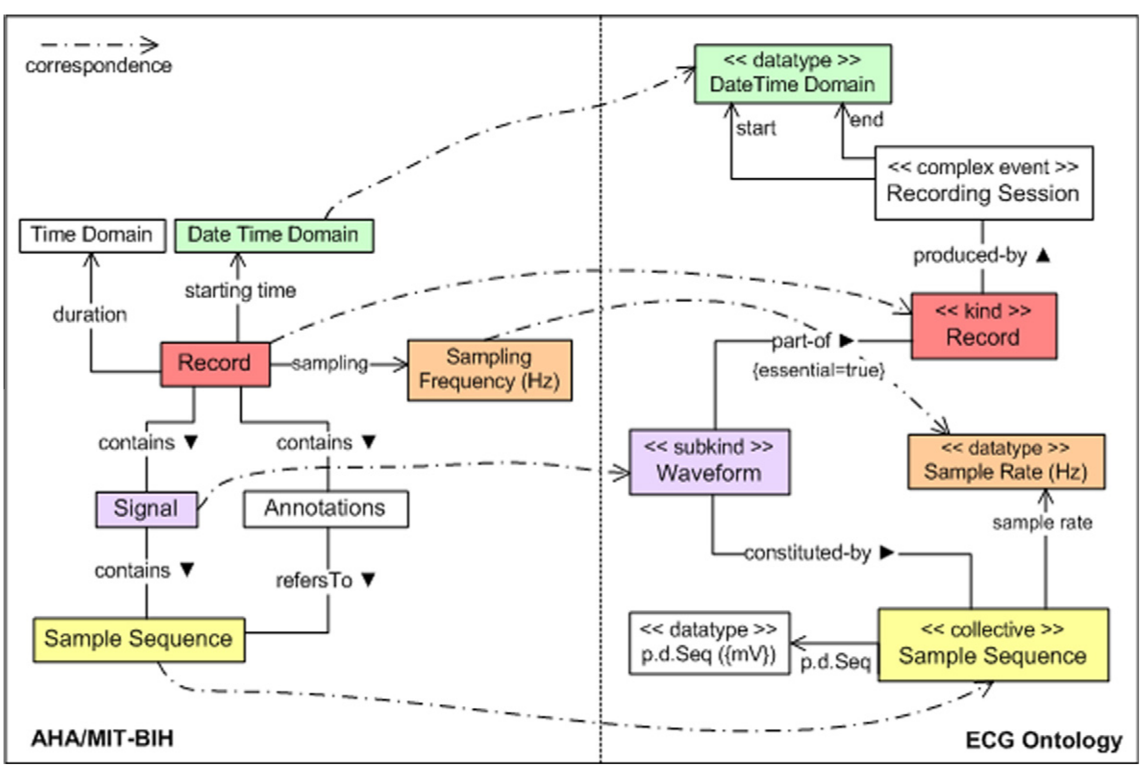

Fig. 4. Integration between the AHA/MIT-BIH conceptual model and the ECG Ontology.

Please cite this article in press as: Gonçalves B et al. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform (2010), doi:10.1016/j.jbi.2010.08.007

Table 2
Definitions of SCP-ECG elements (source: [3]).

Element	Definition
section	Aggregate of data elements related to one aspect of the electrocardiographic recording, measurement or interpretation.
acquiring cardiograph	Cardiograph that records the original ECG signal.
record	Entire data file, including the ECG data and associated information, such as patient identification, demographic and other clinical data.
reference beat	Reference/representative ECG cycle, computed by any algorithm, comprising the P, QRS and the ST-T waves.
rhythm data	Full original ECG data, or the decompressed and reconstructed ECG data at reduced resolution.

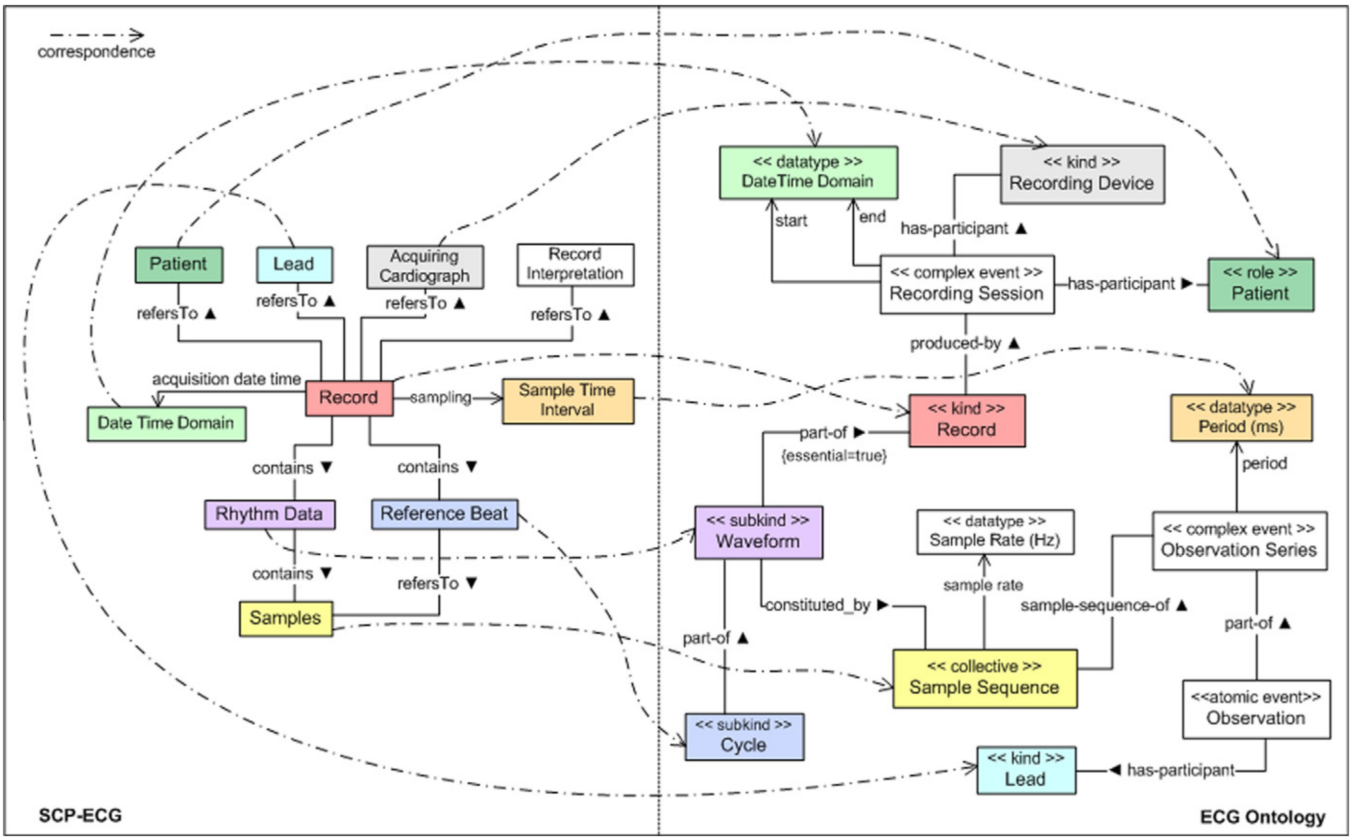

Fig. 5. Integration between the SCP-ECG conceptual model and the ECG Ontology.

A series contains all the sequences, regions of interest, and annotations sharing a common frame of reference for a single ECG. If multiple ECGs are contained within a single aECG file, a different series is used for each of them. A series can be derived from another series. For example, a series containing representative beat waveforms can be algorithmically derived from a rhythm series. Alternatively, a series containing waveforms with special filtering applied can be algorithmically derived from the "raw" rhythm waveforms [4, p. 26]. If two sequence sets were collected from different leads then they cannot be part of the same series, cf. [4]. A series can be further classified into two different types, rhythm and representative beat [4]. The definitions of the most relevant HL7 aECG elements are collected in Table 3. The HL7 aECG conceptual model we have extracted is shown in the left-hand of Fig. 6.

3.4. Discussion

By following the descriptions in the preceding sections, one can notice that the elements present in these ECG standards refer mostly not to real instances from clinical reality, but to their counterparts in the information world in which computer programs operate. This is evidenced by the use of terms like "component", "value", "code", "header", "section", etc. Indeed, to interpret a conceptualization that mixes-up *what exists* with *the information entities that refer to what exists* is not trivial. We have been beset by this issue in the capture of the ECG conceptual models presented here. However, our main point here is not to propose ultimate conceptual models of those ECG standards, but rather to develop our ideas with respect to the use of domain reference ontologies to foster semantic interoperability in Biomedical Informatics.

As examples of the emphasis on the information world in the discussed standards, we can cite the following: AHA/MIT-BIH represents the ECG domain by focusing mainly on storage issues; SCP-ECG, in turn, addresses mostly communication issues; finally, HL7 aECG is primarily concerned with the use of XML technology for favoring data interchange flexibility. On the one hand, as their focus on information requirements vary strongly, some heterogeneity in a "data-level" has in fact been expected. On the other hand, since they deal with the very same ECG domain, it would be intuitive to expect them to share at least a set of core concepts. However, as we have seen, we have found heterogeneity at the "conceptual-level" too – viz., different terms for the same entities, and different entities for the same term.

Please cite this article in press as: Gonçalves B et al. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform (2010), doi:10.1016/j.jbi.2010.08.007

Table 3
Definitions of HL7 aECG elements (source: [4]).

Element	Page	Definition
aECG	5	Standing for Annotated ECG. The annotated ECG is the name given to a file or message conforming to HL7's "Annotated ECG" standard. It contains one or more series of ECG waveforms.
annotation	5	An observation made on or associated with a series. E.g. a P-wave onset, a period of atrial fibrillation, etc.
electrocardiograph	5	A device that records the electrical activity of the patient's heart by tracing voltage-vs-time waveforms.
lead	5	A vector along which the heart's electrical activity is recorded as a waveform.
ROI	5	Standing for Region Of Interest. The region of interest is used to define a region within an ECG series so an annotation can be associated with it – e.g., the ROI between the onset and offset of the P-wave can be associated with a P-wave annotation.
series	5	Contains one or more sequence sets sharing a common frame of reference.
sequence Set	5	A set of sequences all having the same length and containing related values.
sequence	5	An ordered list of values sharing a common code – e.g., sequence of voltage values with code "LEAD II".
value	35	The list of values in the sequence.
subject	14	Identifies the subject from which the ECG waveforms were obtained.
series effective time	26	Physiologically relevant time range assigned to the ECG waveforms contained within the series. This is typically referred to as the "acquisition time" which is determined by the device that collected the waveforms.
aECG component	26	The component parts of the aECG, viz., the waveforms and annotations.
series author	28	This describes the device that "authored" (recorded) the series waveforms. This would typically describe an electrocardiograph or Holter recorder
manufactured series device	28	The unique identifier of the device, independent of its role (i.e. independent of trial). This would typically be the serial number assigned by the device manufacturer.
annotation set component	39	The annotation set made up of one or more annotations (the components of the set).
annotation	39	An observation made on the series by the annotation set's author – e.g., if the electrocardiograph has algorithms to find the beginning of every QRS, an annotation set authored by the electrocardiograph could be made with component annotations for every QRS it finds. If the algorithm can also suggest a disease diagnoses (i.e. "interpretation"), the annotation set could include interpretation statements.
rhythm	26	The series contains rhythm waveforms. These are the waveforms collected by the device. The voltage samples are related to each other in real time (wall time).
representative beat	26	The series contains the waveforms of a representative beat derived from a series of rhythm waveforms.

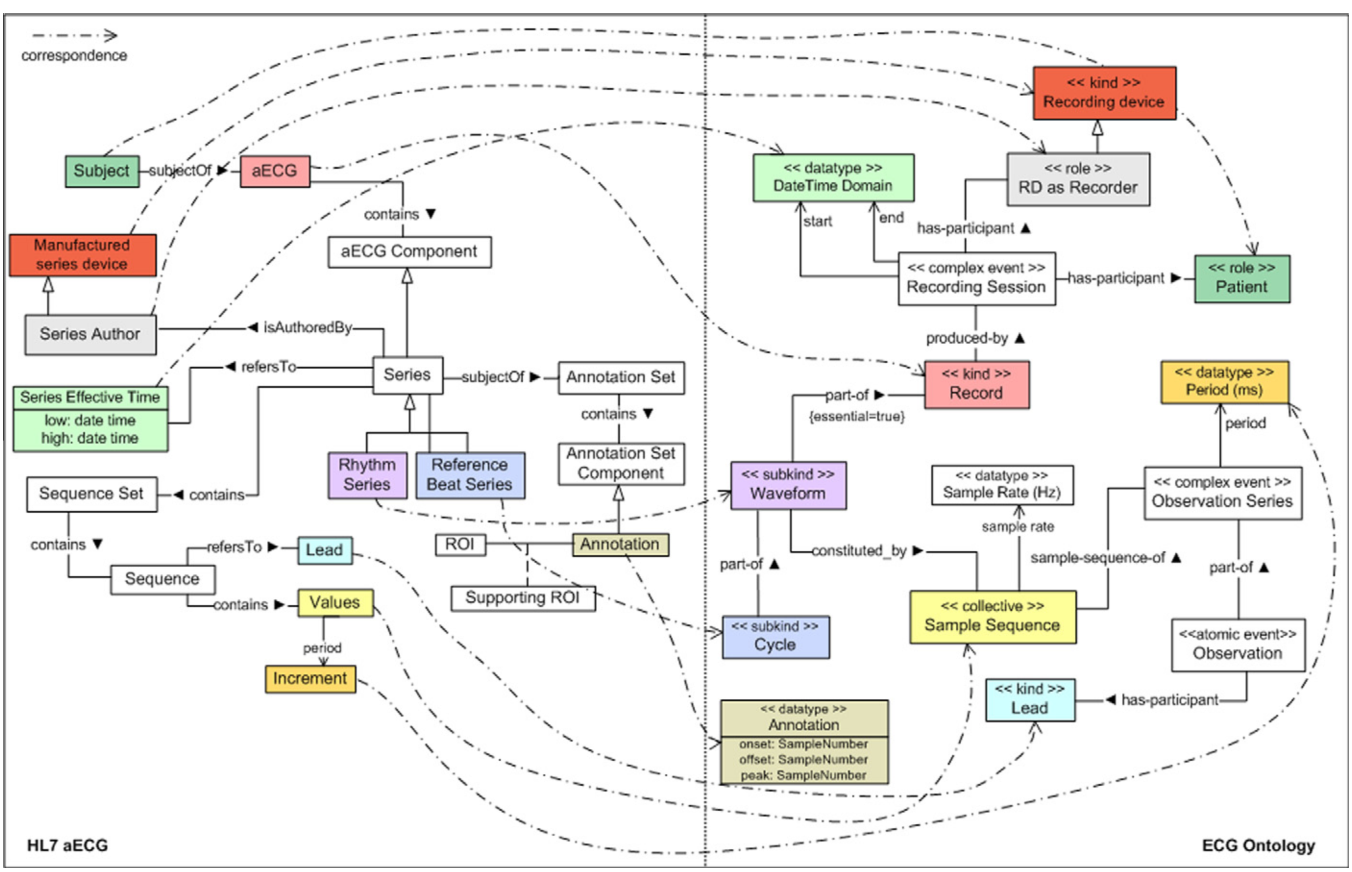

Fig. 6. Integration between the HL7 aECG conceptual model and the ECG Ontology.

In face of this, if we consider the three different ECG domain conceptualizations that underlie the data formats just mentioned, a number of question begging issues arise, namely: (i) what is the "correct" conceptualization for the ECG domain (if any)? (ii) As each of them addresses different purposes and requirements, how to evaluate which purposes and requirements are more fair?

In this case, one would come up to the conclusion that all these different conceptualizations reflect genuine purposes in their own right and, hence, that all of them reflect different aspects or perspectives of the ECG reality. In this case, the most important question is: how to relate these conceptualizations in a sound manner?

Naturally, this is only possible by relying on some shareable anchor. Something that, no matter what community α is in need – likely different to the needs of community β – could draw the attention of both to the fact that there exist acquiring devices, periodic observations, samples resulting from these observations, geometric patterns of interest for diagnostic that emerge from those samples, and so on. This anchor, as discussed by Johansson [6], has been over the centuries the object of interest of Philosophy and Science. Tracking and systematizing a system of referents in reality for the terms used in scientific theories is in line with the very business of Formal Ontology. As advocated by Smith [7], a discipline grounded in formal ontology provides a step forward for initiatives contributing to a unified EHR model. Along these lines, and to echo Johansson [6], we can look *through* terms appearing in medical terminologies or in information systems in order to look *at* the entities instantiated in the health environment where the patient is subject of care by the physician. In next section, we report an integration experiment based on this assumption.

4. The integration experiment

Once we have excavated the conceptual models underlying the ECG data formats, we are able to perform their semantic integration by using the ECG Ontology. This has been conducted in the following manner. Firstly, each element of each ECG data format has been object of analysis in order to match a correspondent entity of the ECG Ontology. We use *correspondence* here as a relationship that gives to the ECG standards' symbolic elements a real-world semantics according to the ECG Ontology. Ergo, this relation is neither one of *equivalence* nor *identity*, i.e., two entities holding a correspondence relation are not the same.

By starting with the AHA/MIT-BIH, the result of the mapping is presented in Fig. 4. The starting time and duration properties refer to record in AHA/MIT-BIH. At a first sight, in an analysis grounded in reality, this association can seem odd. This is because these properties are typically associated with occurrents (perdurants, processes), while a record in AHA/MIT-BIH represents a continuant (endurant, object). However, as we know that the data format is oriented to information systems, we understand that these properties actually correspond to the Recording session's start and end time in the ECG Ontology. The annotations element misses a correspondence in the ECG Ontology, since it does not match the datatype property Annotation in the ECG Ontology but rather to a multitude of annotations only justifiable as a programming resort.

We then turn to the SCP-ECG conceptual model, in order to find correspondence between its elements and the ECG Ontology entities (see Fig. 5). In SCP-ECG, the time associated with a record (acquisition date time) has an ambiguous meaning; it could intuitively be either the start or the end date time. However, in a reasonable reading, we set its corresponding relation to the start and end date time of the Recording session in the ECG Ontology. Instead of referring to the record's sample rate (or sample frequency, in AHA/MIT-BIH), SCP-ECG refers to sample time interval. The latter happens to correspond to the period (duration of each observation) of the Observation series, which results in the samples, or Sample sequence in the ECG Ontology. The element record interpretation comprises an interpretation of the record as a whole, and has no correspondent entity in the ECG Ontology. Apparently, such a general interpretation may be derived from the interpretation of particular elementary forms (viz., normal/abnormal), but this is not clear in the SCP documentation.

Finally, we have performed the mapping from the elements in the HL7 aECG conceptual model to the ECG Ontology, as depicted in Fig. 6. The element aECG component is meant in aECG only to distinguish the series element from other aECG components related to organizational issues (the latter which are not relevant for our purposes here). The element series, in turn, despite being the entity that encompasses all the ECG data in an aECG message (or Record in the ECG Ontology), does not correspond to an entity in the ECG Ontology. In fact, this notion is only relevant in HL7 aECG in case the ECG waveform(s) is (are) fragmented into parts for the purpose of addressing computational needs of ECG processing or viewer programs.

A manufactured series device corresponds to the Recording device in the ECG Ontology. That element comprehends two specific types of device, viz., electrocardiograph and holter. Such a series device can perform the role of series author, just like the RD as recorder. Elements such as sequence, sequence set, annotation set and annotation component also miss a corresponding entity in the ontology. Once again, this is due to the fact that these elements merely represent auxiliary information elements, hence, devoid of real-world semantics.

There is additionally an important element in aECG that miss a correspondence relation to an entity of the ontology, viz., ROI (region of interest). The element ROI refers to any arbitrarily defined interval in the waveform that can hold an annotation. It is somewhat similar to the entity Elementary Form (a ROI is usually an elementary form), but still not an exact match. From its textual definition given in the aECG document, a ROI could be any waveform interval defined for any particular purpose, i.e., its definition is kept too vague preventing an exact mapping to be ascribed to it. For an analogous reason, the element supporting ROI cannot find a direct association in the ontology either.

As result of the mapping from each of the ECG standards to the ECG Ontology, we have achieved, as suggested by Burgun [12, p. 308], indirect mappings between all of them. These mappings are summarized in Table 4. As a conclusion, we advocate that our integration experiment provides evidence for the following statement: *a reference ontology of the ECG domain can be effectively used to foster semantic interoperability between the ECG standards.*

Besides the heterogeneity in the adoption of terms, the heterogeneity involving the conceptualizations underlying these ECG standards is chiefly caused by their computational biases. For instance, the important element annotations in AHA/MIT-BIH misses a correspondent one in the ECG Ontology. This is because technological motivations such as efficient data access have led their maintainers to use a single file with few distinguished fields to keep several annotations of very different sorts. As consequence, the element annotations happens to have no more than a weak connection to the ECG record, of which annotations are intuitively supposed to be about. The element sequence in HL7 aECG, in turn, exemplifies how an integration endeavor guided only by term alignment could lead to mistaken associations. In this case, the element values is actually the one which corresponds to the entity Sample sequence.[6] Currently, nonetheless, the technique known as "ontology alignment" which relies mostly in lexical similarity is, as reported by Yu [9], one of the most used methods for data integration in Biomedical Informatics. As demonstrated here, an integration based uniquely on term alignment cannot generally be said to guarantee semantically sound results (see next section).

[6] A more suitable XML design for it would let the sequence to hold the samples themselves (for term appropriateness) with an attribute, rather than a child element, indicating the sample rate.

Table 4
Correspondences between entities in the ECG Ontology and corresponding elements in each of the ECG standards.

ECG Ontology	AHA/MIT-BIH	SCP-ECG	aECG
Record	record	record	aECG
Sample rate	sampling frequency	–	–
Period	–	sample time interval	increment
Sample sequence	sample sequence	samples	values
Waveform	signal	rhythm data	rhythm series
Cycle	–	reference beat	reference beat series
(Date time domain): start and end time	(date time domain): starting time	(date time domain): acquisition date time	(series effective time): low and high
Recording device	–	acquiring cardiograph	manufactured series device
RD as recorder	–	–	series author
Patient	–	patient	subject
Lead	–	lead	lead
Annotation	–	–	annotation

5. Discussion

5.1. Related Work

We are aware of some preliminary initiatives in terms of related work on ontologies for the ECG domain. There are two of them which fall roughly in heart electrophysiology. Rubin et al. [26] present a symbolic and ontologically-guided methodology for representing a physiological model of the circulation, as an alternative to the mathematical models commonly employed. Cook et al. [27] in turn are putting effort in an extension of the FMA to cover physiology. However, the most correlated work is the NCBO project entitled "Electrocardiography Ontology".[7] This project, which started in early 2009 seems to be in a preliminary stage from a Formal Ontology point of view. The results so far (which can be visualized on the project's website) seem to have a significant overlap with HL7 aECG. Moreover, the ontology elements presented there still lack proper formal definitions and characterization. We conclude that, to the best of our knowledge, there is still no consolidated related work on ECG Ontology.

Regarding initiatives for data integration in the Biomedical Informatics literature, we have a different scenario. Currently, we can find several initiatives dedicated to (automated) ontology alignment (cf. [9, p. 261]). An illustrative example can be given by the work of Zhang and Bodenreider [28] as cited by Burgun in [12, p. 309]. In that paper, they investigate the possibilities of indirectly aligning two anatomical ontologies by using the FMA as a domain reference ontology. They showed that 91% of the direct matches between the Adult Mouse Anatomical Dictionary and NCI Thesaurus were discovered by the indirect alignment. In their work, Zhang and Bodenreider use structural similarity in addition to lexical similarity. We argue that structural similarity analysis can indeed provide a valuable support to avoid a number of semantic mismatches that recurrently take place when lexical similarity is the only employed technique. However, as one would expect, the quality of structural similarity analysis is strongly dependent on the quality of the structures at hand. Now, as reported on in the biomedical ontology literature, the available ontologies are still beset by several structural problems of ontological nature. This is particularly true for the FMA itself. In [29], Donnelly et al. demonstrate that the FMA collapses at least three different distinctions of parthood which are in general relevant in anatomy.

To echo Bittner and Donnelly's discussion on the automated alignment of relations in biomedical ontologies [30], "*at least one major obstacle to such integration is that many existing bio-medical terminology systems and ontologies handle foundational relations such as parthood ambiguously and inconsistently*". For this reason, we believe that, for now, qualitative results in terms of data integration can hardly be achieved only with automated alignment. The challenge of semantic integration still requires, rather, careful investigations by using the state-of-the-art tools of Formal Ontology and Conceptual Modeling.

5.2. Applicability

We discuss in the sequel two possible classes of applications of the work reported here, reflecting different modes of support for the semantic interoperability of ECG data.

- *EHR-driven computational applications*. By being able to establish clearly the semantic relations between data elements in two different information sources, one could, for instance: (i) create a wrapper for query translation from one source to the other (supposing these sources are implemented as databases of ECG data following each of the standards); (ii) create a composed web service which relies on individual services that make each of these standards with reliable translation of parameters. For example, a query to collect the Sample sequence which constitutes the Waveform that composes a Record will be translated in SCP-ECG as a query to collect the samples contained in the rhythm data element contained in turn in a record (see Fig. 5). In contrast, in HL7 aECG, the same query would be translated as one to collect the values contained in the sequence element, which is contained in a sequence set contained in turn in a series contained finally in aECG (see Fig. 6). Such a computer application could be used in an EHR system to retrieve patient data from heterogeneous data sources.
- *Support for redesign and possible unification of existing standards*. A domain reference ontology can be used to support the design of interoperable versions of (otherwise) heterogeneous data formats. As a reference model whose primary goal is adherence to reality, an ontology can serve as an anchor to both conceptual and terminological negotiation processes. This marks an striking contrast to the existing standards in this area, each of which biased to a different set of information systems' concerns. From the results presented in Table 4, one could (i) propose conceptual and terminological changes in the corresponding standards in order to facilitate the unification of these standards; or even (ii) have a starting point for an EHR unification program that would be grounded in reality since its very beginning (as proposed by Smith [7, p. 297]), while still being sensitive to the existing conceptualizations and terminologies proposed by standardization initiatives.

5.3. Limitations and future work

In the sequel, we discuss the main limitations of the ECG Ontology employed in this paper, which shall be addressed in future

[7] Cf. http://bioportal.bioontology.org/ontologies/39893.

Please cite this article in press as: Gonçalves B et al. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform (2010), doi:10.1016/j.jbi.2010.08.007

work. Firstly, the current version of our ECG Ontology is limited to a canonical representation of heart anatomy and physiology. We believe that a strong research effort is required to extend the ECG ontological theory presented here to cover pathological aspects of heart electrophysiology recognizable in the ECG waveform. Secondly, in the present ontological theory, we have covered only a single-lead ECG scope. However, contrary to the first limitation, the extension of the ontology to cover multiple-lead ECGs (at least the standardized 12-lead ECG) is expected to be straightforward.

These limitations in the ECG Ontology delimit a scope for which it can be used as a reference model for interoperability. In particular, the integration we have carried out in this article has a scope restricted to the core aspects of the ECG, viz., to what the ECG is in essence. Therefore, the two aspects aforementioned – for which the ECG Ontology has to be extended to cope with – shall enable a broader applicability of this ontological theory in terms of a definitive integration of the ECG standards. To put it baldly, an endeavor committed to such a full integration of the ECG standards (which has not been the objective of the work reported in this article) cannot be accomplished only with the tools we have in hand at this point.

However, this by no means impinges a limitation on the methodology proposed here. For that the hypothesis put to test in this article has been fully proved. That is, we have successfully used an ECG reference ontology to semantically integrate ECG data standards. An additional contribution of the article is to provide Table 4 depicting correspondence links between entities in the ECG reference ontology and elements in the ECG standards. This particular integration of the ECG standards AHA/MIT-BIH (Physionet), SCP-ECG and HL7 aECG, however, can be extended to a full scope if we extend the ECG Ontology. Firstly, to cover multiple-lead ECGs, and secondly, to cover pathological patterns in the ECG waveform. It is worthwhile to highlight that reference ontologies are, just like any other scientific theory [31], evolving artifacts. Therefore, they shall eventually become effective means for the full integration of data formats in their respective domains.

Along these lines, the methodology used and proposed here could be replicated in other domains in the field of Biomedical Informatics. Finally, we must remark that one of the main difficulties faced in this work has been the intricateness of excavating conceptualizations from specifications that are imprecise from an ontological point of view and biased from a computational one.

6. Conclusions

In this paper we have made the case that an ECG reference ontology can be effectively used to accomplish semantic interoperability of ECG data. We have reflected on this domain-specific case study to shed light on how the ontological approach can be applied to the problem of semantic integration in Biomedical Informatics. The key points developed throughout the article which are worthwhile to recall are:

- This case study shows that a syntactic approach to model integration cannot work in general because heterogeneity is not confined to the syntactic (term) level. In contrast, heterogeneity can also be found in the conceptualizations underlying these data models. Hence, for semantic interoperability to be conducted in an effective way, the heterogeneity between conceptualizations has to be resolved before handling the term-level syntactic heterogeneity. The methodology proposed here then complements the methods commonly employed in the literature (which rely on term and structural comparison of concepts) by reaching the semantic level.
- As conceptual models are usually produced to meet specific purposes and requirements, a shareable anchor is required for effective integration to be achieved. We have demonstrated in this article that a domain reference ontology (thus, an axiomatized conceptual model grounded in reality) can be used as a cost-effective means to foster semantic interoperability of heterogeneous conceptual models.
- We have conducted an integration experiment which provides evidence that the ECG Ontology can be effectively used to support the design of interoperable versions of the ECG standards AHA/MIT-BIH, SCP-ECG and HL7 aECG. The methodology proposed here can be applied in other domains.
- Eventhough the excavated conceptual models we convey here are not meant to be final representations of those standards, to have them expressed in an explicit manner puts forth concrete media for discussion and elaboration in the community of Biomedical Informatics.

As argued by Smith [8], "*the value of any kind of data is greatly enhanced when it exists in a form that allows it to be integrated with other data*". Time has come for us to start integrating data not only at the syntactic level, but also at the semantic level. However, as demonstrated here, this semantic integration can hardly be conducted effectively without the proper formal tools of ontological analysis to address the common reality underneath these data. This is specially the case when the reality at hand is one of sheer complexity, as it is the case of the domains of interest for Biomedical Informatics.

Acknowledgement

This research has received financial support from the Brazilian Research Funding Agencies FAPES (Grant No. 45444080/09) and CNPq (Grant No. 481906/2009-6).

References

[1] Goldberger A et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–20.
[2] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng Med Biol Magaz 2001;20(3):45–50.
[3] CEN/TC-251. SCP Document CEN/TC-251 N02-15, 2002. Retrieved from: http://www.centc251.org/. August, 2006.
[4] Brown B, Badilini F. HL7 aECG implementation guide, 2005. Available at: http://www.amps-llc.com/UsefulDocs/aECG_Implementation_Guide.pdf. Accessed on March 15, 2009.
[5] Stockbridge N, Brown B. Annotated ecg waveform data at fda. J Electrocardiol 2004;37(Suppl 1):63–4. doi:10.1016/j.jelectrocard.2004.08.018.
[6] Johansson I. Bioinformatics and biological reality. J Biomed Inf 2006;39(3):274–87. doi:10.1016/j.jbi.2005.08.005.
[7] Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. J Biomed Inf 2006;39(3):288–98. doi:10.1016/j.jbi.2005.09.005.
[8] Smith B et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251–5. doi:10.1038/nbt1346.
[9] Yu A. Methods in Biomedical Ontology. J Biomed Inf 2006;39(3):252–66. doi:10.1016/j.jbi.2005.11.006.
[10] Guarino N. Formal Ontology and information systems. In: Guarino N, editor. Proceedings of the 1st Formal Ontology and information systems. Amsterdam: IOS Press; 1998. p. 3–15. Trento, Italy.
[11] Smith B, Welty C. Ontology: towards a new synthesis. In: Smith B, Welty C, editors. Proceedings of the 2nd international conference on Formal Ontology in information systems (FOIS). New York: ACM Press; 2001. p. 3–9.
[12] Burgun A. A desiderata for domain reference ontologies in Biomedicine. J Biomed Inf 2006;39(3):307–13. doi:10.1016/j.jbi.2005.09.002.
[13] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4–5):394–403.
[14] Gonçalves B. An ontological theory of the electrocardiogram with applications. Master's thesis. Computer Science Department, Federal University of Espírito Santo, 2009. Available at: http://www.inf.ufes.br/bgoncalves/contents/Bernardo_MScThesis.pdf.
[15] Gonçalves B, Zamborlini V, Guizzardi G. An ontological analysis of the electrocardiogram. Elect J Commun Inf Innov Health (Suppl on Ontologies,

Semantic Web and Health) 2009;3(1):907–28. doi:10.3395/reciis.v3i1.242en. Rio de Janeiro, Brazil.
[16] Smith B et al. Relations in biomedical ontologies. Genome Biol 2005;6(5). doi:10.1186/gb-2005-6-5-r46.
[17] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inf 2003;36(2003):478–500. doi:10.1016/j.jbi.2003.11.007.
[18] Burek P et al. A top-level ontology of functions and its application in the Open Biomedical Ontologies. Bioinformatics (Oxford) 2006;22(14):e66–73. doi:10.1093/bioinformatics/btl266.
[19] Dupre A, Vincent S, Iaizzo PA. Basic ECG theory, recordings, and interpretation. In: Iaizzo P, editor. Handbook of cardiac anatomy, physiology, and devices, Chapter 15. Totowa, New Jersey: Humana Press; 2005.
[20] Guyton A, Hall J. Textbook of medical physiology. 11th ed. Philadelphia: Elsevier Saunders; 2006.
[21] Weilkiens T. Uml – Unified Modeling Language. Systems Engineering with SysML/UML 2007:143–221. doi:10.1016/B978-0-12-374274-2.00003-1.
[22] Welty C, Guarino N. Supporting ontological analysis of taxonomic relationships. J Data Knowledge Eng 2001;39(1):51–74.
[23] Guizzardi G. Ontological foundations for structural conceptual models. Ph.D. thesis. The Netherlands: University of Twente; 2005. Also published as the homonymous book in the Telematica Instituut Fundamental Research Series; No. 15; Universal Press, The Netherlands. ISBN 90-75176-81-3. Available at: http://purl.org/utwente/50826.
[24] Rector A, Rogers J, Bittner T. Granularity, scale and collectivity: when size does and does not matter. J Biomed Inf 2006;39(3):333–49. doi:10.1016/j.jbi.2005.08.010.
[25] Moody GB. WFDB programmer's guide, 2009. Version 10.4.19. Available at: http://www.physionet.org/physiotools/wpg/. Accessed on March 2009.
[26] Rubin DL et al. Ontology-based representation of simulation models of physiology. In: AMIA Annual Symposium Proceedings 2006, Washington, USA, 2006, p. 664–8.
[27] Cook DL, Jose LV, Mejino CR. Evolution of a Foundational Model of Physiology: symbolic representation for functional bioinformatics. In: Proceedings of the 11th world congress on Medical Informatics (MEDINFO'04). Vol 107 Stud Health Technol Inform. Amsterdam: IOS Press; 2004, p. 336–40.
[28] Zhang S, Bodenreider O. Alignment of multiple ontologies of anatomy: deriving indirect mappings from direct mappings to a reference. In: AMIA Annual Symposium Proceedings 2005, Washington, USA, 2005. p. 864–8.
[29] Donnelly M, Bittner T, Rosse C. A formal theory for spatial representation and reasoning in biomedical ontologies. Artif Intell Med 2005;36(1):1–27. doi:10.1016/j.artmed.2005.07.004.
[30] Bittner T, Donnelly M. Logical properties of foundational relations in bio-ontologies. Artif Intell Med 2007;39(3):197–216. doi:10.1016/j.artmed.2006.12.005.
[31] Smith B. Ontology (science). In: Proceedings of the 5th international conference on Formal Ontology in information systems (FOIS). Amsterdam: IOS Press; 2008. p. 21–35.

Please cite this article in press as: Gonçalves B et al. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform (2010), doi:10.1016/j.jbi.2010.08.007