

# A Context Description Language for Medical Information Systems

Kurt Englmeier<sup>1</sup>, John Atkinson<sup>2</sup>, Josiane Mothe<sup>3</sup>, Fionn Murtagh<sup>4</sup>, Javier Pereira<sup>5</sup>

<sup>1</sup>Faculty of Computer Science, Schmalkalden University of Applied Science, 98574 Schmalkalden, Germany  
k.englmeier@fh-sm.de

<sup>2</sup>Department of Computer Sciences, Universidad de Concepción, Concepción, Chile  
atkinson@inf.udec.cl

<sup>3</sup>Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, 31062 Toulouse, France  
mothe@irit.fr

<sup>4</sup>Department of Computer Science, Royal Holloway, University of London, Egham, England  
fionn@cs.rhul.ac.uk

<sup>5</sup>Facultad de Ingeniería, Universidad Diego Portales, Santiago, Chile  
javier.pereira@udp.cl

**Abstract.** Contextualized delivery of information is one of the many strengths of ubiquitous computing. It makes information actionable and helps us to better understand our situations. In the realm of healthcare, contextual information provides a terse but precise picture of the patient's health situation. The patient context can have many facets, ranging from nutrition context over health heritage context to the context of symptoms, just to name a few. Setting up the correct health condition context of a patient favors better and faster recognition of the patient's actual health situation.

Context-awareness in medical monitoring mainly concentrates on gathering numerical facts depicting special aspects of a person's health condition. In this paper we want to broaden the focus on the textual dimension in context development, by considering semantic annotation in designing context-awareness. We describe an approach for a context description language (CDL) that supports the uniform presentation of textual facts in medical reports and automatic reasoning on these facts. Term clusters in medical reports represent in a unique way symptoms and findings that set up the health context reflected in this particular report. These clusters manifest potential health condition contexts where a patient can be viewed in. A reasoning engine operates on these context presentations and selects those that match best the patient's health situation. Locating the right context supports the physician in faster getting a first picture of the probable health situation of a new patient to be examined. We present experiments with a CDL applied on reports related to respiratory problems.

**Keywords:** Context-awareness, context design and development, semantic annotation, domain-specific language, information mining, natural language interaction, medical reports.

## 1 Introduction

Context can be considered as a collection of facts and their inter-relationships describing the environment of a user or an event [1]. Context awareness sharpens relevance of these facts along particular aspects. It sharpens the awareness of specific restaurants in the physical context of the user, when the user is looking for one of her favorite types of places to eat. Depending on the user's preferences specific objects of the user are drawn into the context of the users while others are pushed away. In this example, it's quite likely for the user to find a restaurant of her taste. If we draw the right objects into a given context, a terse but precise picture of a particular situation is obtained. This picture helps us to better understand that situation and to make correct decisions. Consequently, it also supports disambiguation, i.e. considering an object within its correct context and excluding irrelevant contexts. Disambiguation is important in the realm of medical information where a small number of observations and/or symptoms must be brought into the correct context.

The interpretation of a health phenomenon is formed by observations and experiences in form of patients' health records. In particular, experiences include reasoning over observations and the resulting conclusions. The situation-adequate compilation of information can be essential for the appropriate handling of any health-related situation. Contextual information is used not only to interpret available information, but also to seek additionally relevant or missing information. Contextual delivery of information makes communication more efficient, more focused on the aspects of the problem at hand. It can provide information on an interesting restaurant or building the user is probably looking for [2] or it alerts persons suffering from asthma when they enter an area where others with the same health problem previously had an asthma attack [3]. In both cases, users get focused information that takes into account their personal characteristics and qualities of their immediate environment.

The two examples show that context emerges from underlying concepts and their attributes (qualities). Correctly setting up a context depends on the correct composition of these situation-specific concepts. The context "preferred restaurant nearby" or "area reportedly causing respiratory problems" is based its own theory describing a model that merges

personal characteristics (preferred restaurants or chronic respiratory conditions, respectively) and with related objects at the physical location of people [4]. In the case of “asthma alert”, data may come from remote monitoring of people with the same respiratory problems, i.e. monitoring the location where they used their rescue inhaler.

While the use of the rescue inhaler (location and time) can be monitored quite easily, it is a bit more complicated to represent chronic respiratory conditions. Medical reports typically provide facts representing health conditions in mainly textual form. Of course, these reports refer to a variety of symptoms that furthermore differ gradually. They represent thus variations of health condition concepts (e.g. “severe bronchial asthma” or “light bronchial asthma”).

In this paper, we present a context description language (CDL) for the medical domain. With our approach we want to draw attention to semantic integration of textual facts into situation-specific contexts. We demonstrate how the CDL forms a meta-language that supports (1) the development of context-related controlled vocabularies and (2) concept-based reasoning, which, in turn, controls the integration mechanism. Our experiment with reports on respiratory health problems exemplifies how context will be described and how a context is selected in accordance to an individual patient’s health situation.

## 2 Natural Language Representation of Context

In the recent past, mobile computing brought us a large number of location-based services thanks to the popularity of GPS [5]. Many context engines today are built for location-based services using inference models that operate on geographical data and related mass data [6]. In automotive systems, the context engine takes the location of the car, projects the car’s trajectory, and compares this path with the position or trajectory of objects in the car’s way. It infers a potential collision from these data and alerts the driver correspondingly. In our case, we have to “locate” a person within the most appropriate health condition context. This helps the physician to compare the suggested contexts with the patient’s health condition in order to get a solid first glance on the patient’s situation and to find clues for further examination steps.

A context picture can be composed of purely numerical data or a combination of numerical and textual data. In health condition contexts, prevail textual facts interspersed with a few numerical facts. Reasoning is thus based more on methods for the semantic analysis of textual information. Our approach roots thus in Semantic Annotation that deals with analysis of unstructured content. By adding annotation tags we describe the meaning behind terms. If this description takes standardized form, machines can easier recognize facts than be solely applying text retrieval algorithms. Semantic Web knows numerous approaches for the description of content elements and relationships among them. RDF [7] is a popular standard to add structured annotations to text. It helps to bring text terms in certain semantic relationships, which supports understanding of text terms, to a certain extent, but does not solve the problem of understanding the tags. This analysis, in turn, benefits from a uniform presentation of the facts in a text. RDF does little to sufficiently address the understanding of meaning [8]. A solution to this problem provides the agreement on fixed terms to be used in text annotation. The Dublin Core Metadata Initiative<sup>1</sup> developed a number of concept definitions, i.e. definition of terms with attributes and properties. However, Dublin Core does not cover medical concepts. Nevertheless, a practical approach to solve the problem of understanding of meaning (in texts as well as context descriptions) emerges from the agreement on a relatively simple and precisely specified language. If this language is, in addition, machine-processable, we can develop engines that reason on context descriptions and the semantic closeness of contexts. Usually, this is the point where ontologies come into play. They define logic-based knowledge-representation formalism and are thus a powerful candidate when it comes to define context and reasoning processes to support context-awareness. OWL [9] is currently the most prominent candidate among ontology languages. However, ontologies and their design languages are not only powerful but also complex. The application of ontologies, even promising, is too labor-intensive and expensive in many cases.

We therefore opt for domain-specific languages (DSL) [10] that are domain-related markup languages fostering the uniform presentation of semantics related to textual facts [11-12]. They manifest an agreement on language elements for semantic tagging and a couple of reasoning rules. Description logics expressed in DSL are not as powerful and scalable like ontologies. However, in a thematically not too complex environment, they are just as useful. Our CDL is a DSL that semantically depicts a patient’s health context. The development of our CDL is guided by the assumption that humans apply a specific language in their working environment, and this becomes essentially a formal working jargon. Within this domain and community framework, natural language statements represent literal meaning [13]. Literal meaning can be interpreted correctly in the absence of any explicit and implicit context. When describing objects or processes of their working environment, users apply a language in their statements that aims at unambiguous understanding of those descriptions. The same holds for medical personnel: they want to express literal meaning in order to assure that their statements are interpreted correctly by their colleagues.

Literal meaning, in turn, is almost immediately processable by machines without language transformation that is usually required for the interpretation of (natural language) statements representing non-literal meaning that often refers to implicit knowledge of the broader statement context. In sentences representing literal meaning, there is a close corre-

---

<sup>1</sup> [dublincore.org](http://dublincore.org)

spondence between the term patterns and their inherent semantics. Furthermore, these patterns also reflect computational representations supporting taxonomic, temporal, and spatial reasoning. Text patterns have thus a clear relationship to their inherent meaning. The typical problem of ambiguity that usually comes with natural language analysis is by far less severe in statements representing literal meaning. Consequently, the logic inherent in Named Entity Recognition and some more text pattern recognition algorithms combined with a sparse semantic annotation suffice to provide enough potential to represent meaning in medical reports.

Our reasoning algorithms constituting text pattern analysis can therefore operate immediately on the users' statements. However, instead of traditional text retrieval methods [14] we propose grammar-free, machine-processable description logic [15] that assigns specific concepts to term (text) patterns within or across sentences. These concepts, in turn, form a controlled vocabulary that semantically covers the language elements for the context (model) descriptions. In order to find a proper health condition context for a particular person, a new patient for instance, we need a basic set of context descriptions (expressed in terms of the controlled vocabulary) that act as candidates for the context of this patient. From this perspective, context matching resembles information mining in unstructured data. We compare statements about an individual patient with the context descriptions in our database.

### 3 Representation of Context

Our CDL provides, at first, abstract concepts for both, the semantic annotation of medical reports and the abstract representation of condition contexts (Figure 1). A health condition or problem usually comprises one or more symptoms. These generic concepts represent a symptom described by (one or more) body parts (organs) and (one or more) phenomena occurring in the context of diseases (congestion, dilation, etc.). Each of these elements usually has qualities like "normal" or "calcified". In particular, phenomena have as quality also their specific location ("anterior rib arc"). The objective of text analysis is to classify all terms in the text according to these abstract concepts and to group them in accordance with their appearance in the individual reports. It is important to note, that each symptom is related to one body part and each report addresses one individual health condition.

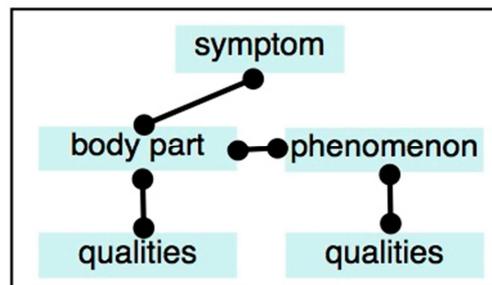


Fig. 1. Generic concepts describing health conditions. All text terms are classified along these abstract concepts.

Classification and grouping are the next features of the CDL. They are achieved by a stepwise approach (bottom-up) starting with the most simple language constructs (step 1) to more complex ones (step 5). For determining the more simple language elements we apply the well-known approach of "Parsing and Standardization" that uses stereotypical (standardized) text patterns to identify (simple and more complex) named entities.

#### 1. Identification of elementary expressions

For the identification of basic elements we apply Named Entity Recognition (NER) tasks by using a set of generic patterns for Named Entities. These entities are expressions like word, number, date, or ID number. Any identified Regular Expression ("13 mm", "03-JAN-1996") here represents an instance of a basic element. Instances of these elements stand for basic concepts like word, date, number, distance, name, etc.

#### 2. Identification of minimal semantics

Minimal semantics focus on generic text elements (or basic elements) and terms adjacent to these elements. The concept "age", for instance, is represented by a generic element consisting of a pattern reflecting a 3-digit Regular Expression ("1?[0-9]{1,2}") combined with terms like "age", "old", "years" or "hours". An instance of "price" may have a leading symbol ("€" or "\$"); the expression "hrs." may follow an expression of digitals indicating that this element is instance of "hour".

In our approach, stop words are not eliminated, as they can be significant for the correct interpretation of entities. For identifying minimal semantics, we used n-grams that may be interlaced by stop words. To identify composed entities only the n-grams are used without their interlaced terms. These are used only for the purpose of further specification. For instance, "... valid until 30-NOV-2011" is an expression of date, represented by a 2-gram ("valid" and date) ignoring the term "until" in the first place. This term is used later when the concept "validity period" is expanded towards

“end of validity period”. An instance of the distance element (“21 mm”) becomes a “diameter” element when followed by an expression like “in diameter”. Otherwise it just represents the size of an object.

### 3. Transformation into standardized forms

Natural language is ambiguous by nature even if we consider a quasi-formal language like the one medical professionals apply in their reports. Humans tend to apply different words to the same concept. We handle this synonym problem by standardization, i.e. mapping different terms representing one concept onto one representative term for this concept (“caliber” to “size” or “anterior” to “front”). It also includes the transformation of nouns into their corresponding adjective if used in order to qualify another noun like in “Aortic wall with calcification”. This is transformed into “Aortic wall calcified”.

Negations require a special treatment in any automatic text analysis. We treat negation as a special quality of an object. A negation just states if a quality of an object or the object itself is present or absent. In medical reports, it is important to explicitly mention the absence, for instance, of a certain phenomenon (“No pleural effusion”) in order to depict a correct picture of the health condition or distinguish two otherwise similarly described conditions.

### 4. Identification of key nouns

Named entities reflect basic language elements. These expressions are input to further pattern analysis. We repeat the identification of terms frequently appearing in close proximity. This process leads to language elements that stand for concepts representing represent parts of the body (“lungs”, “bronchi”, etc.) or phenomena like “dilation”. In the case of medical reports, nouns usually reflect these concepts.

For the development of our controlled vocabulary we extracted instances of the generic concepts mentioned above. Instances of body parts are then clustered into concepts representing larger parts of the body (“lung” comprises “right lobe”, “trachea”, etc.). Here, classification systems like the International Classification of Diseases (ICD) of the WHO [16] provide useful sources to set up a concept structure for body parts.

### 5. Conceptualization

The final step identifies key patterns as a combination of basic language elements with key nouns. This process adds qualities to concepts represented by key nouns by relating key nouns to simple language elements. Both parts usually appear within the same phrase. A concept represented by key nouns is thus considered in a broader context including terms that reflect qualities. By using n-grams we ensure that only terms in close proximity are considered in a single conceptualization.

For example, in the phrase “Trachea and bronchi permeable, both normal caliber” we identify “Trachea” and “bronchi” as key nouns representing body parts with the qualities “permeable” and “size:normal”. As already mentioned, we apply also a normalization process that transforms, for instance, all size-related expressions into one standardized form “caliber” to “size”.

It is important to deal with pronouns and similar relationships, because they may address qualities that need to be assigned to more than one noun. In the example phrase here, the qualities mentioned address the two body parts. The keyword “both” explicitly indicates the pronoun relationship between the qualities and the two terms reflecting the body parts. It is the simple sequence of terms that determines the word pattern and thus the role of the terms (key noun - “and” - key noun - indicator of a pronoun relationship - quality-related terms) appearing in the phrase. Similar indicators include “all”, “that” or “which”.

A pattern recognition engine takes the CDL, analyzes the medical reports and transforms them into an XML representation. The behavior of the engine is determined by the role descriptions of the concepts listed in the CDL. When applied to medical reports the CDL, the result of the analysis looks like in Figure 2.

```

2② <observations>
3②   <trachea quality="permeable">
4     Trachea and bronchi permeable, normal caliber. Secretions in the trachea.
5     <size>normal</size>
6     <secretion/>
7   </trachea>
8②   <bronchi quality="permeable">
9     Trachea and bronchi permeable, normal caliber.
10    <size>normal</size>
11  </bronchi>
12②  <lungs>
13④    <consolidation quality="none">□
16④    <granuloma quality="calified">□
20②    <dilatation>
21      Cylindrical bronchial dilatation in anterior segments of both upper lobes,
22      middle lobe and posterior segments of lower lobes.
23④    <location>□
29    </dilatation>
30②    <surface quality="normal">
31      Normal pleural surfaces.
32    </surface>
33②    <aorta quality="calified">
34      Aortic wall calcification.
35    </aorta>
36  </lungs>
37②  <heart>
38    <size>normal</size>
39  </heart>
40②  <mediastinum>
41    <lymph_nodes quality="unchanged" />
42  </mediastinum>
43②  <lump quality="benign">
44    In the anterior arch of the third left rib nodular lesion shows a well defined
45    contours of 13 mm diameter, nonaggressive aspect.
46②    <size>
47      13 mm
48    </size>
49②    <location>
50      <rib quality="left,third, front" />
51    </location>
52  </lump>
53 </observations>
54② <conclusions>

```

Fig. 2. Medical report expressed in terms of the CDL.

The example shows the observation part of the report listed in the XML structure. This part is usually followed by a conclusion part (shown in Figure 3). The section shows minimal semantics (“size”, “location” and “quality”), and keywords for body parts (“lungs”, “aorta”, “heart”, etc.) and phenomena (“lump”, “dilation”, etc.). It also shows some results of the normalization process that standardizes “nonaggressive” to “benign” (lines 43 and 44) when appearing in proximity to the concept “lump” (lines 43 though 52) and “anterior” to “front” (lines 44 and 50).

The first phrase in line 4 implicitly contains a pronoun relationship that semantically links the qualities “size” and “permeable” to the two body parts. The representation of the two related concepts (“trachea” and “bronchi”) is thus replicated.

```

54② <conclusions>
55②   <bronchiectasis>
56     Bilateral bronchiectasis
57     <location>bilateral</location>
58   </bronchiectasis>
59②   <atherosclerosis>
60     Aortic atherosclerosis
61     <location>aortic</location>
62   </atherosclerosis>
63②   <lump>
64     Nodular lesion of the anterior arch of the left third rib.
65②   <location>
66     <rib quality="left,third, front" />
67   </location>
68   </lump>
69 </conclusions>

```

Fig. 3. Conclusion section of the medical report, expressed in CDL.



```

1 ⊖ <observations>
2 ⊖   <lungs>
3 ⊖     <dilatation>
4     <location>left, bottom</location>
5     </dilatation>
6   </lungs>
7 </observations>

```

**Fig. 5.** A patient context description (“bronchial dilation in the lower lobes of the left lung”) expressed in CDL.

We acknowledge that we get such a favorable result because of the relatively small set of candidate contexts. Nevertheless, our database comprises reports on respiratory health problems recorded over a period of about six months by the radiology department. Even if we take into account the possibility that this collection grows to a size 10-fold of the one of our example base, we will see the number of recommendations growing just to 40 cases. A pre-requisite, however, to keep the size of the recommendation list actionable, would be the classification of each patient description along the main health problems before the pattern recognition engine starts its work.

As an initial proof of concept, our work was validated by the user group of physicians involved. It showed us that our stepwise analysis of text patterns yields a controlled vocabulary that, classified along the generic concepts mentioned above, can provide machine-processable descriptions of health contexts. In a next step, we are going to expand this experiment to further health problems including more medical professionals in order to cover also a broader variety of first-time statements on patients’ health conditions.

## 5 Conclusions

Mobile Computing sees many initiatives, projects, and product developments under way that monitor key indicators of a person’s health situation. They all address portable devices that act as sensors gathering health data such as pulse, temperature and so on besides the coordinates of the person’s physical location. Connected to a remote data mining system, the user of such a wearable device can expect to get alerts warning her that her health situation sparks concerns. Or she may be warned when she is entering an area where other persons experienced respiratory problems.

The context constructed in these location-based services is exclusively built on numerical information and the corresponding mathematical models. This approach is beneficial for a huge number of mobile and embedded applications that operate on sensory data.

With our paper we want to broaden the focus on the benefits of mining textual information for the design of context-awareness. Many medical facts are expressed in natural language statements. Fortunately, the language used in reports describing health conditions often lacks the ambiguity that usually can be found in natural language texts. Due to this quality medical reports lend themselves for the development of a meta-language that identifies the meaning of terms in the light of depicting the context of a patient’s health situation. This meta-language defines a domain-dependent Context Description Language that controls text pattern recognition engines and context engines.

By broadening the focus on textual facts, i.e. facts extracted from unstructured texts, we can contribute to the informative performance of context-aware systems. Our context- and ontology-free approach is not restricted to health-related topics. However, it thematically scales only when we consider domains where facts are represented in the same clear manner. The ambiguity rises in textual information the more we need ontologies for a clear identification of meaning in text.

## References

1. Mehra, P.: Context-Aware Computing. Beyond Search and Location-Based Services. *IEEE Internet Computing* 16 (2), 12-16 (2012).
2. Fröhlich, P., Oulasvirta, A., Baldauf, Nurminen, A.: On the Move, Wirelessly Connected to the World. *Communications of the ACM* 54 (1), 132-138 (2011).
3. McKenna, M.: The New Age of Medical Monitoring. *Scientific American* 308 (3), 16-17 (2013).
4. Goodchild, M.: Citizens as sensors: The world of volunteered geography. *GeoJournal* 69, 211–221 (2007).
5. Mehra, P.: Context-Aware Computing. Beyond Search and Location-Based Services. *IEEE Internet Computing* 16 (2), 12-16 (2012).
6. Lee, J.-H., Kim, J.-T., Lee, H.-K., Paik, E.-H.: Design and implementation of the Geo-Context Engine for semantic social media service. *Conference for Internet Technology and Secured Transactions (ICITST)*, 383-387 (2011).
7. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. [www.w3.org/TR/2004/ REC-rdf-concepts-20040210/](http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/) (2004)
8. Horrocks, I.: Ontologies and the Semantic Web. *Communications of the ACM* 51(12), p. 67 (2008).
9. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language. [www.w3c.org/TR/owl-features/](http://www.w3c.org/TR/owl-features/) (2004).

10. Mernik, M., Heering, J., Sloane, A.M.: When and How Develop Domain-Specific Languages. *ACM Computing Surveys* 37(4), 316-344 (2005).
11. Stahl, T.; Voelter M.: *Model-Driven Software Development*, Wiley & Sons (2006)
12. Cao, L., Ramesh, B., Rossi, M.: Are Domain-Specific Models Easier to Maintain Than UML Models? *IEEE Software* 26(4), 19-21 (2009).
13. Iwanska, L.M.: Natural Language Is a Powerful Knowledge Representation System: The UNO Model. In: L.M. Iwanska and S.C. Shapiro (eds.), *Natural Language Processing and Knowledge Representation*, AAAI Press, Menlo Park, USA, 7-64 (2000).
14. Agosti, M., Gradegnio, G., Marchetti, P.: A hypertext environment for interacting with large databases. *Information processing and management* 28, 371-387 (1992).
15. Englmeier, K., Koinig, R.: Domain-Specific Deployment and Configuration Language for Composition and Adaptation of Coarse-Grained Services. *Proceedings of IEEE-SCC 2009 conference*, 490-493 (2009).
16. WHO: International Classification of Diseases (ICD); retrieved on February 20, 2013 on <http://www.who.int/classifications/icd/en/#> (2013)
17. Mahmud, U., Mohammed, Y.J.: Context Inference Engine (CiE): Inferring Context. *International Journal of Advanced Pervasive and Ubiquitous Computing* 4 (3), 13-41 (2012).
18. Garcia-Molina, H., Koutrika, G., Parameswaran, A.: Information seeking: convergence of search, recommendations, and advertising. *Communications of the ACM* , 54 (11), 121-130 (2011).