Medical Archetypes and Information Extraction Templates in Automatic Processing of Clinical Narratives

Ivelina Nikolova[1], Galia Angelova[1], Dimitar Tcharaktchiev[2], and Svetla Boytcheva[3]

[1] Institute of Information and Communication Technology,
Bulgarian Academy of Sciences,
Sofia, Bulgaria,
iva@lml.bas.bg galia@lml.bas.bg
[2] University Specialised Hospital for Active Treatment of Endocrinology,
Medical University Sofia, Bulgaria,
dimitardt@gmail.com
[3] American University in Bulgaria, Blagoevgrad, Bulgaria,
svetla.boytcheva@gmail.com

Abstract. This paper discusses the notion of medical archetype and the manner how the archetype elements are documented in hospital patient records. This is done by interpreting the archetypes as information extraction templates in automatic text analysis of clinical narratives. The extensive extraction experiments performed over thousands of anonymous discharge letters show the actual instantiation of the required and expected items in the narrative clinical documentation; in fact much tacit and implicit medical knowledge is implicitly presented in the real clinical texts. This fact suggests that the archetype approach to defaults and inheritance might need certain development.

Key words: Clinical knowledge, Medical archetypes, Interoperability of clinical software systems, NLP of clinical narratives, Information extraction, Template filling

1 Introduction

Archetypes are chunks of declarative medical knowledge that are designed to capture maximally expressive and internationally reusable clinical information units. They encode knowledge about clinical observations, evaluations, actions and instructions in a coherent and holistic manner with the intension to provide language-independent specifications. Archetypes are based on conceptual structures of medical knowledge and provide standardised clinical content. Medical ontologies conceptualise domain objects, actions and relationships among them; the archetypes, representing the blueprints of defined medical domains, are focused on capturing clinical information about the patient. Archetypes are not linked a priory to any medical terminology but they can refer to multiple external medical classifications (e.g. SNOMED) from where controlled vocabularies are incorporated as labels of archetype elements. The openEHR project aims at the acquisition of a representative set of freely available archetypes thus enabling information sharing between clinical systems. Hundreds of archetypes in ADL (Archetype Definition Language) are publicly available via the openEHR Clinical Knowledge Manager. openEHR expresses health information systems and interoperability mechanisms in UML (Unified Modelling Language).

Automatic processing of free clinical texts, however, might reveal whether medical experts keep the requirements to document clinical units in a manner which ensures their unambiguous export to other clinical systems. Analysing the free text of 6,204 anonymous discharge letters of diabetic patients, we present empirical observations whether the slots of the diabetic-relevant archetypes, published by openEHR, are filled in by the necessary information of classification codes or free text. In a sense we discuss how the theoretical models of clinical knowledge are applied in practical settings when the medical case is documented.

The article is structured as follows. Section 2 overviews the notion of archetypes. Section 3 discusses the archetypes as Information Extraction templates applied in automatic text processing. Section 4 presents the experiments performed on a large corpus of discharge letters. Section 5 contains some discussion and the conclusion.

2 Archetypes as Conceptual Structures

Archetypes are designed during the last decade to make health information systems properly and safety interoperable [1]. They are based on the notion of recording in medicine. The health record content is likely to be a small, selective choice of notes about real events, situations etc. intended for interpretation by other professionals rather than some more general notion of comprehensive fact representation. Analysing the important types of information in the health care process, the authors propose the Clinical Investigator Record Ontology where the observations (evidences) and opinions (inferences) are different categories (see Figure 1). This taxonomy provides the categories in the Entry classes of the openEHR reference model. For our purposes we shall be interested in the archetypes capturing the observations (findings of examinations, measurement, questioning, or testing of the patient or related substance like blood, tissue etc.), because automatic information extraction from clinical narratives is most successful for declarative statements. An archetype is a computable expression of a domain content model in the form of structured constraint statements, based on a reference (information) model [2]. Archetypes define conceptual items and relationships among them as well as constraints on the values of their instances: e.g. allowed types, ordering, cardinality, (referent) values etc. We are interested mostly in the conceptual background of the clinical archetype model which is defined together with the openEHR software development requirements.

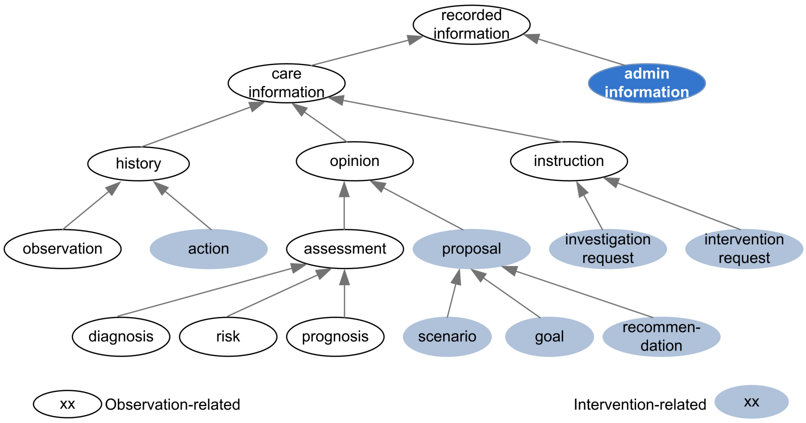

Fig. 1. The Clinical Investigator Record Ontology [1].

The Clinical Knowledge Manager supports two major kinds of archetypes: the Electronic Health Record (EHR) Archetypes where patient-centered data is kept and the Demographic Model Archetypes. The EHR Archetypes are Clusters, Compositions, Elements, Entries, Sections and Structures. Table 1 shows the Items of the Examination of thyroid Cluster (the Header of the Cluster is skipped as it contains metadata related to the creation, author, date etc.) The indexing Keywords, included in the Header of this Cluster, are examination, physical, thyroid. They are included manually in order to facilitate the advanced search within the archetype collection. Table 1 illustrates the hierarchy of embedded (included) sub-clusters which are referred to by citation of the archetype names e.g. an instance of openEHR-EHR-CLUSTER.inspection.v1 is included in the description of the findings concerning the Left lateral lobe. We note that a significant number of the descriptions are assumed to be typed in as free or coded text, therefore the archetype is a kind of template where text fragments might be entered in narrative form. Optional elements might be omitted in the instances in case there is no abnormality observed during the examination. Without entering in details we remind that declarative specifications are hard to define and standardize for broader use; in addition the support and maintenance of the archetype collection requires significant efforts. But despite these shortcomings it is clear that the Archetype model responds to the needs of establishing standards in the EHR content (and the clinical documentation practice in general) in order to ensure semantic interoperability between the healthcare systems.

In 2008 the archetype approach to structuring patient-related records was accepted as ISO standard 13606-2:2008. It specifies the information architecture required for interoperable communications between systems and services dealing with EHR data [3]. In this way ISO 13606-2:2008 defines how to organise hierarchically the EHR content, how to define the individual data items and their aggregations, what types of values or measurement units are appropriate and so on. Archetypes are viewed as a serialised representation, an exchange format for communicating individual archetypes between archetype libraries. Current efforts of the openEHR-related community are dedicated to the definition of further archetypes at the optimal level of granularity and specificity in order to ensure their wide adoption. In this way more medical experts could be involved in the creation of archetype repositories. Best practices are sought to achieve multi-professional clinical consensus. Having in mind all the recent developments, we think that Natural Language Processing (NLP) of clinical narratives can help much in the tests whether archetypes are properly defined. The automatic text analysis might reveal the actual status of clinical event documentation and suggest potential drawbacks in the archetype definition. This paper presents such tests for some essential archetypes, related to diabetic patients.

Authoring and review of archetypes is viewed as a knowledge acquisition task with highest priority. An Archetype Editorial Group has been established as an expert clinical team to lead the authoring of archetypes within the openEHR community. The national eHealth programs in several countries (Australia, Denmark, Singapore, Sweden, and UK) include archetype-related initiatives in order to involve medical professionals, agencies and educational institutions into development activities. International agreements should be sought by international authorities (like the World Health Organisation and relevant standardisation bodies). Actually the unification of clinical description content is a long process which is still in its infantry. Nevertheless it is important that this process has started and an ISO standard has been adopted.

At the end of this section we present the data fields included in two other archetypes:
(i) Blood pressure $openEHR - HER - OBSERVATION.blood\_pressure.v1$ and
(ii) Body weight $openEHR - HER - OBSERVATION.body\_weight.v1$.
Extracting automatically these items from the discharge letters of diabetic patients we can check their availability and actual use in the clinical documentation.

3 Information Extraction Templates

Information Extraction (IE) is a popular technique for Natural Language Processing (NLP) which aims at partial text understanding in order to provide fast and efficient analysis of texts in specialised domains. The IE systems identify specific events or topics, searching for relevant information only and disregarding the remaining text fragments. IE typically extracts named entities and words referring to objects or events in order to identify their roles in event descriptions. The identification is supported by the so called templates feature-value structures that capture the entities recognised by the text analysers. Most generally, the IE success is measured by the accuracy of filling in the template slots by proper words encountered in the text.

Structure: Cluster Occurrences: 1..1 (mandatory) Cardinality: 1..* (*mandatory, repeating, unordered*)		
Normal statements Cluster Occurrences: 0..1 (optional) Cardinality: 1..* (*mandatory, repeating, unordered*)	A group of statements about the normality of the examination.	
Normal statement Text, Occurrences: 0..* (optional, repeating)	A specific statement of normality.	Free or coded text
Clinical description Text, Occurrences: 0..1 (optional)	Textural description of the part examined.	Free or coded text
Findings Cluster Occurrences: 0..1 (optional) Cardinality: 1..* (*mandatory, repeating, unordered*)	Clinical findings.	
Visible abnormality Boolean Occurrences: 0..1 (optional)	There is a visible thyroid abnormality.	
Mobility on swallowing liquid Text, Occurrences: 0..1 (optional)	Description of thyroid mobility on swallowing liquid.	Free or coded text
Left lateral lobe Cluster Occurrences: 0..1 (optional) Cardinality: 1..* (*mandatory, repeating, unordered*)	Findings of left lobe of thyroid.	
Description Text, Occurrences: 0..1 (optional)	Text description of clinical findings.	Free or coded text
Left lateral lobe SLOT (Cluster) Occurrences: 0..1 (optional)	Detailed findings of left lobe of thyroid.	**Include:** openEHR-EHR-CLUSTER.inspection.v1 and specialisations *Or* openEHR-EHR-CLUSTER.exam-generic.v1 *Or* openEHR-EHR-CLUSTER.palpation.v1

Fig. 2. Items of the CLUSTER Examination of thyroid .

Right lateral lobe Cluster Occurrences: 0..1 (optional) Cardinality: 1..* (*mandatory, repeating, unordered*)	Findings of right lobe of thyroid.	
Description Text, Occurrences: 0..1 (optional)	Text description of clinical findings.	Free or coded text
Right lateral lobe SLOT (Cluster) Occurrences: 0..1 (optional)	Detailed findings of right lobe of thyroid.	**Include:** openEHR-EHR-CLUSTER.inspection.v1 and specializations ***Or*** openEHR-EHR-CLUSTER.exam-generic.v1 ***Or*** openEHR-EHR-CLUSTER.palpation.v1
Isthmus Cluster Occurrences: 0..1 (optional) Cardinality: 1..* (*mandatory, repeating, unordered*)	Findings of isthmus of thyroid.	
Description Text, Occurrences: 0..1 (optional)	Text description of clinical findings.	Free or coded text
Isthmus SLOT (Cluster) Occurrences: 0..1 (optional)	Findings of isthmus of thyroid.	**Include:** openEHR-EHR-CLUSTER.inspection.v1 and specialisations ***Or*** openEHR-EHR-CLUSTER.exam-generic.v1 ***Or*** openEHR-EHR-CLUSTER.palpation.v1
Detail SLOT (Cluster) Occurrences: 0..* (optional, repeating)	More focused examination findings	**Include:** openEHR-EHR-CLUSTER.exam-generic.v1 and specialisations ***Or*** openEHR-EHR-CLUSTER.auscultation.v1 ***Or*** openEHR-EHR-CLUSTER.inspection.v1 ***Or*** openEHR-EHR-CLUSTER.palpation.v1 ***Or*** openEHR-EHR-CLUSTER.percussion.v1 ***Or*** openEHR-EHR-CLUSTER.physical_properties.v1
Image Multimedia Occurrences: 0..* (optional, repeating)	Drawing or image of the area examined.	image/gif, image/png, image/jpeg

Fig. 3. Items of the CLUSTER Examination of thyroid (continued).

Entity name	Content	Value
Systolic	Peak systemic arterial BP	Units: mm[Hg]
Diastolic	Minimum systemic arterial BP	Units: mm[Hg]
Mean arterial pressure MAP	Average arterial pressure	Units: mm[Hg]
Pulse pressure	Difference between the systolic and diastolic pressure	Units: mm[Hg]
Comment	Comment about the measurement	Free or coded text
		Standing; Sitting; Reclining; Lying; Lying with tilt to left
Position	Description	
Confounding factors	Free or coded text: factors that may impact the measurement	For instance: level of anxiety; pain or fever
	Details about physical activity undertaken at the time of measurement	Includes openEHR-EHR-CLUSTER.level_of_exertion.v1 and specialisations
Exertion		
Sleep status	Supports interpretation of 24-hours BP measurement	Alert & Awake; Sleeping
Tilt	Surface craniocaudal tilt	Angle, plane, degrees
Cuff size	The size of the cuff used for the measurement	Adult thigh; Large adult; Adult; Small adult; Paediatric/Child; Infant; Neonatal
Location /cluster		
		Right arm; Left arm; Left thigh; Right wrist; Left wrist; Right ankle; Left ankle; Finger; Toe; Intra-arterial
Location of	Body site where BP is recorded	
measurement		
Specific	Specific details about the site where the BP is recorded	Free or coded text
location		
Method	Method of measurement	Auscultation; Palpation; Machine; Invasive
Mean arterial pressure formula	Formula used to calculate MAP	Free or coded text
Diastolic endpoint	Which Korotkoff sound is used	Phase IV; Phase V
Device	Details about the device used to measure BP	Includes openEHR-HER-CLUSTER.device.v1 and specialisations
Event	Description	Any relevant event
24 hour average	Estimate of the average BP	Math function Mean

Table 1. Entities included in the Blood Pressure (BP) archetype

Entity name	Content	Value
Weight, quantity	Weight mass	Units: kg, lb
Comment	Comment about the measurement of weight	Free or coded text
		Lightly clothed/Underwear; Naked; Fully clothed including shoes; Nappy/diaper
State of dress	Description	
Confounding factors	Free or coded text: factors that may impact the measurement	For instance: timing of menstrual cycle, timing of recent bowel motion, noting of amputation
Confounding factors	Free or coded text: factors that may impact the measurement	For instance: timing of menstrual cycle, timing of recent bowel motion, noting of amputation
Device	Details about the weighing device	Includes openEHR-EHR-CLUSTER.device.v1 and specialisations
Event	Description	Any relevant event

Table 2. Entities included in the Body Weight archetype, which is indexed by the keywords *weight*, *gain*, *loss*, *increase*, *decrease*, *mass*, *estimate*, *actual*

Early IE papers consider the template design as an essential step in the IE system development. Templates are flat or object-oriented [4] and their design should satisfy a number of requirements:

- descriptive adequacy the template should represent all the information necessary for the task at hand, having in mind that adding features often requires to add further features;
- clarity the ability to represent all the information in the template unambiguously;
- determinacy there should be only one way of representing a given item or a complex of items;
- perspicuity the degree to which the design is conceptually clear to the human analyst who will input or edit information in the template or work with the results;
- monotonicity the template should reflect the data content monotonically or incrementally (adding a new value should not cause update, restructuring or removal of the values in other template slots);
- application considerations the particular task might impose constraints e.g. evaluation metrics and further limitations; reusability the template objects should be potentially reusable in other domains and applications.

It is easy to see the similarities between the definition of template (a chuck of declarative knowledge automatically extracted from text) and archetype (an ultimate, universal chunk of clinical knowledge, to be declared manually and used as standard aggregation of atomic elements). Without loss of generality we can consider the attributes, listed at Tables 2 and 3, as prototypical elements of flat templates to be used in IE from clinical texts. It is obvious that simple conceptual graphs [5] can capture the semantics of the feature-value pairs in Tables 1, 2 and 3. In the next section we shall present the results of IE experiments using the archetypes listed above.

Finally we note that the notion of template evolves in the NLP field; recent papers suggest learning template structure automatically from raw text without using predefined template schemes [6].

4 Extracting Archetype Items from Clinical Texts

Here we report the results of experiments with 6,204 anonymised patient records (PRs) of diabetic patient and assessment whether the archetype elements are explicitly documented or not. Our attention is focused on the three archetypes that have been previously discussed: examination of thyroid, measurement of blood pressure and measurement of patient body weight. The experiments are performed using the IE environment that has been recently developed by the authors [7, 8].

4.1 Examination of Thyroid

More than 97% of the PRs in our corpus contain explicit descriptions of thyroid examination. Many PRs contain more than one discussion of thyroid because they include basic description in the Status section and more detailed tests (like echography) in the Lab values and/or Consultations sections. Due to this reason some 11,606 instances of the archetype are found in 6,058 PRs.

Total PRs	**6 204**
PRs with no explicit data for thyroid	146
PRs containing description of thyroid	6 058
Total extracted records for thyroid	11 606

Table 3. Availability of thyroid descriptions in 6,204 discharge letters

Our IE components identified text fragments describing certain abnormalities, the left/right thyroid lobe, the mobility of the swallowing liquids and the isthmus (see Table 5). More than 82% of the PRs (5,144 out of 6,204) contain a statement about normality which can be positive or negative. Comparing the available descriptions to the map view of the archetype in Figure 2 we see that almost all data items are regularly filled in.

Items/Findings	Visible abnormality	1 556
	Mobility of swallowing liquid	1 892
	Left lateral lobe	1 836
	Right lateral lobe	2 304
	Isthmus	1 846
Items/Normal statements	Normal statement	5 144

Table 4. Availability of thyroid descriptions in 6,204 discharge letters

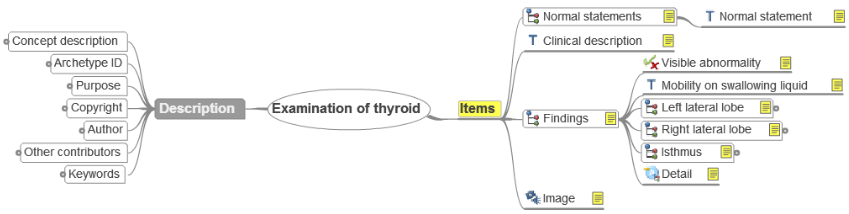

Fig. 4. Map View of the Examination of thyroid archetype.

4.2 Measurement of Blood Pressure

About 78% of PRs in our corpus contain explicit BP values. Table 6 illustrates the findings. In the 2,111 PRs without explicit values, there could be phrases referring to normal and default values like: Blood pressure in the norm, No data/signals for Arterial Hypertonic illness and so on. Some PRs contain more than one occurrence of BP values and this explains the fact that 4,841 items were extracted from 4,093 PRs.

Total PRs	**6 204**
Total PRs	6 204
PRs with no explicit data about BP	2 111
PRs containing data about BP	4 093
Total extracted records about BP	4 841

Table 5. Availability of thyroid descriptions in 6,204 discharge letters

Further details about available descriptions are given in Table 7. Only 47 PRs discuss the position when the BP measurement is performed (less than 0,01% of all PRs). About 12,6% of the PRs discuss confounding factors. Both systolic and diastolic values are given in the 4,841 particular measurements cited in the corpus. Some 8% of the PRs discuss the mean arterial BP. Pulse pressure occurs in 57% of the analysed discharge letters. The abbreviation (RR) in Protocol/Method denotes BP measurements taken with the technique of the sphygmomanometer invented by Scipione Riva-Rocci. It occurs in 26,6% of all PRs.

State/ Position	Standing	25
	Sitting	3
	Reclining	0
	Lying	19
	Lying with the tilt on the left	0
State/Confounding factors	Under therapy	350
	Without Orthostatic Symptoms	428
	With Orthostatic Symptoms	6
Data/ Systolic - Diastolic		4 841 - 4 841
Data/ Mean Arterial Pressure	Usually/Average	501
	Max	456
	Min	150
Data/ Pulse Pressure		3 566
Protocol/ Method	RR	1 834

Table 6. Recording measurements of BP values in 6,204 discharge letters

Comparing the extracted values to the map view in Figure 3, we see the elements that are rarely instantiated: most items in State section (position, exertion, sleeping status, tilt) and in Protocol section (cuff size, location of measurement, method, mean arterial pressure formula, diastolic endpoint).

4.3 Measurement of Body Weight

The absolute value of body weight is a factor when diagnosing with diabetes but even more important is the deviation from the patients ordinary body weight. For the professional it is necessary to know whether the patient has experienced any significant change in the weight during the recent months or year(s). Along with the thyroid gland and limbs and skin description, one of the most often met descriptions in the patient records is body weight change. Table 8 summarizes the number of events extracted from the patient records.

Total PRs	**6 204**
PRs containing data about exact weight	3 815
Total extracted occurrences of exact weight	3884
PRs containing data about increase of weight	2 613
Total extracted occurrences of increase of weight	5533
PRs containing data about decrease of weight	1 083
Total extracted occurrences of decrease of weight	1273
PRs containing data about weight change	3 197
Total extracted occurrences of weight change	6 806

Table 7. Available values of weight and weight change in 6,204 discharge letters

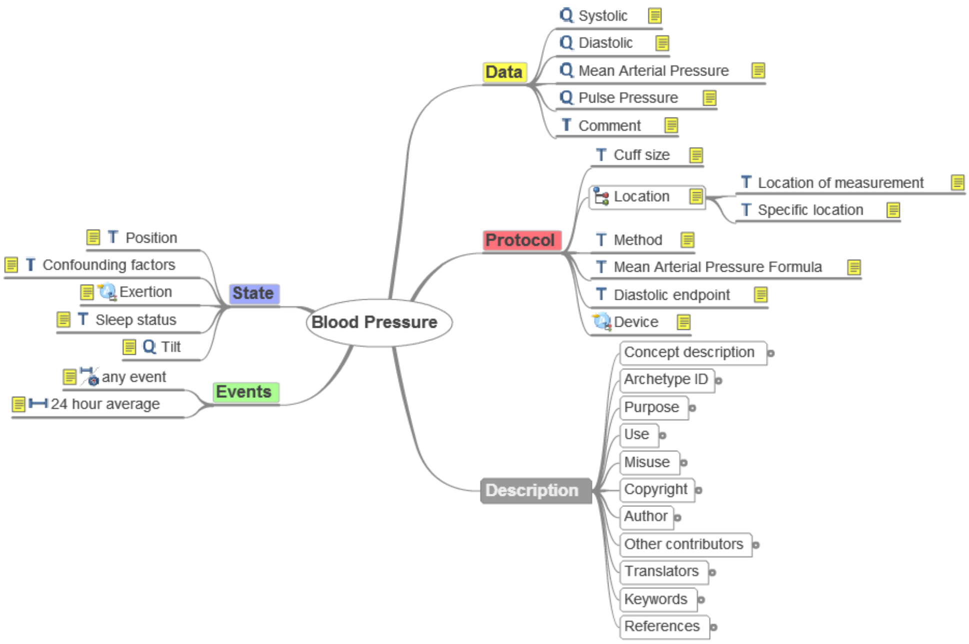

Fig. 5. Map View of the Blood Pressure measurement archetype.

It is visible from Table 8 that mentions of increase or decrease of body weight are almost double the mentions of exact weight in our document collection. Often when the weight is discussed in a file it is mentioned more than once describing the changes during the development of the disease and this also explains why the percentage of files containing mentions is rather similar (52% contain weight change and 61contain exact weight). Mentions of increased body weight are almost 3 times more often than mentions of decreased weight. The expressions related to weight in the patient records are all related to quantity: (i) body weight change which can be found in the anamnesis or patient status section and is expressed as an interval value, exact value or by an expression showing the direction of the change:

- ◇ increased her body weight with about 10-12 kg in the last 6 months or
- ◇ reduction of body weight /15 kg for 2 years
- ◇ overweight

It is often followed or preceded by the period for which the given change is observed. (ii) *exact weight values* which can be found in the laboratory results section:

- ◇ weight - 89 kg
- ◇ 170/86kg

(iii) unclassified expressions of the following type:

◇ succeeded to go back to his regular weight

which value is rather relative to a previous condition in the history of the disease and it is hard to select a slot in the archetype to fill them in.

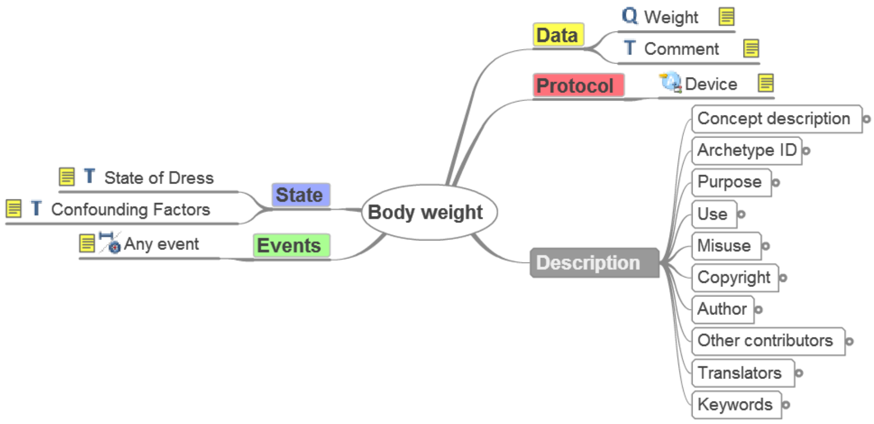

Fig. 6. Map View of the Body weight archetype.

There are no other expressions related to weight, which are informative for fillfing in the archetype slots (state of dress, confounding factors, device, event). Obviously these are not a subject of interest in the endocrinology.

4.4 Evaluation Results

For IE is used rule-based approach. Due to variety of paraphrases and keywords used to represent blood pressure the precision is relatively low, in contrast for "Thyroid" we have only few keywords and their abbreviations used in our Training and Test banks "щитовидна жлеза" (thyroid gland) and "тиреодид" (Thyroidism) disorder of Thyroid gland. Many paraphrases and abbreviations are used to represent "Blood pressure" like "артериално налягане", "", or disorders like "артериална хипертония" (arterial hypertention), on many places are used only data and method "rr" or the data are listed without explicit notation as examinations of the superclass "аритмична сърдечна дејност" (arrhythmic heartbeat).

5 Discussion

Our IE components work in the following manner:

- The English terms, available in the Tables 1, 2 and 3 are translated to Bulgarian;

	Precision	**Recall**	**F-measure**
Thyroid	96.25%	93.42%	94.81%
Blood Pressure	71.37%	90.63%	79.86%
Body Weight			

Table 8. Available values of weight and weight change in 6,204 discharge letters

- Their synonyms (terms or paraphrases) are found in the dictionaries that we have developed previously;
- Then the target terms for the selected archetypes are searched in the texts of the corpus PRs.

In this way we identify availability and type of the recognised descriptions. There might be other items, expressed by different words that remain unidentified; however, the observations centered on the terms mentioned in Tables 1-3, deliver a relevant generalised view about text content.

Here are some examples how we capture thyroid gland descriptions in our data starting from the archetype description. The recognition modules are built on the base of archetype keywords and slot descriptions. We know from previous experiments that the description of the status of an anatomical part is normally present in a single sentence. The rules are constructed to capture expressions starting from one mention of the anatomical part of interest (thyroid gland in this case) and try to find subsequent descriptions of the archetype slots. Below are given examples which include description of the left lobe and thyroid gland properties, which are listed one after another and separated from the anatomical part by hyphen:

◇ щитовидна жлеза – увеличена ... левия лоб, с еластична консистенция (thyroid gland enlarged... left lobe with elastic consistency)
◇ ехо на щитов. жлеза – уголемени размери, хипоехогенна (echography of the thyroid gland enlarged size, hypoechogenic)
◇ щитовидна жлеза – увеличена, плътна консистенция, чувствителна при палпация (thyroid gland enlarged, solid consistency, sensitive when palpated)

This way we easily grep expressions which contain terms used in the archetype description but it is difficult to capture narratives such as comments. They are free text fields in the archetype and their nature does not allow to suggest any keyterms to search for. For exhaustiveness we rely on the characteristics of our data to contain one body part description within a single sentence.

The results of this study show that the doctors hardly explicate:

- hospital-dependent implicit knowledge in clinical narratives when reporting about patient cases, for instance type of devices (e.g. cuffs for blood pressure measurements);
- values that are irrelevant for the particular disease (e.g. exact weight of diabetic patients and conditions when it was measured, or location where the blood pressure is measured). Instead, they document relevant features like weight change for given period which are not included in the archetype.

There is also tacit knowledge which holds in the respective domain and it is regularly omitted in the particular texts. These observations show the difference between theoretical information models in medicine and their practical application. The standards of writing clinical documentation do not affect quickly the established tradition in writing domain-specific texts.

6 Conclusion

In this paper we present evidences about availability of standard elements in clinical descriptions. It is evident that the conceptual structures, designed to capture patient- related clinical information in order to ensure its systematic representation, need a long period of development, standardisation and wide adoption in order to provide interoperable resources of clinical knowledge. Perhaps thinking in terms of archetypes and conceptual structures needs to be incorporated in the medical training as well. The Topic Maps, as illustrated in Figures 2 and 3, are a suitable visualisation tool that might help to advertise the archetype methodology.

We propose that the archetype design process should integrate language technologies for information extraction which enable immediate checks whether the theoretical conceptual model is aligned to the clinical practice of reporting events and observations. For instance, if the checks show that the medical experts regularly omit device descriptions, then this element might be included in the specific archetype instance by default for the particular clinical units. In this way some information in the instantiated archetype might be imported from a separate hospital unit description without burdening the doctors with too much documentation. Another possibility is to offer specific predefined menus for item selection that are contextualised for the hospital unit. Simplifying the documentation process will facilitate the wide archetypes adoption.

Acknowledgments. The research work presented in this paper is supported by grant DO 02-292 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2012.

References

1. Beale, T. and S. Heard. An Ontology-based Model of Clinical Information. In K. Kuhn et al. (Eds), Proceedings MedInfo 2007, IOS Publishing 2007, pp. 760764.
2. Beale, T. and S. Heard (Eds.) Archetype Definitions and Principles. openEHR Report, March 2007.
3. ISO 13606-2:2008 Health informatics - Electronic health record communication - Part 2: Archetype interchange specification, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50119
4. Onyshkevych, B. Template Design for Information Extraction. In Proc. of the TIPSTER Text Program: Phase I, September 1993, Virginia, USA, pp. 141145, available in the ACL Anthology http://www.aclweb.org/anthology/X93-1015

5. Sowa, J. Conceptual Information Processing in Mind and Machines. Reading, MA (1984)
6. Chambers, N. and D. Jurafsky. Template-Based Information Extraction without the Templates. Proc. of the 49th ACL Ann. Meeting, Oregon June 2011, pp. 976986.
7. Boytcheva, S. Structured Information Extraction from Medical Texts in Bulgarian. In Proc. of the SINUS Workshop Semantic Technologies in the Humanities, 7-8 June 2012, Sozopol, Bulgaria, to appear in a Special Issue of the Journal Cybernetics and Information Technologies, 2012.
8. Nikolova, I. Unified Extraction of Health Condition Descriptions. In Proc. of the NAACL HLT 2012 Student Research Workshop, Montreal, Canada, June 3-8, 2012, pp. 2328, see http://www.aclweb.org/anthology-new/N/N12/N12-2005.pdf
9. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
10. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
11. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
12. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
13. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
14. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov