Semantic Medical Care in Smart Cities

Alexander Vodyaho, Nataly Zhukova, Maxim Lapaev, Andrey Koltavskiy

(Prof Alexander Vodyaho, St. Petersburg Electrotechnical University (LETI), 5, Popova str., St. Petersburg, 197376, Russia, aivodyaho@mail.ru aivodyaho@mail.ru)

(Nataly Zhukova, St Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO), 49, Kronverksky Pr.,St. Petersburg, 197101, Russia, nazhukova@mail.ru)

(Maxim Lapaev, St Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO), 49, Kronverksky Pr.,St. Petersburg, 197101, Russia, m.lapaev@corp.ifmo.ru)

(Andrey Koltavskiy, St Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO), 49, Kronverksky Pr.,St. Petersburg, 197101, Russia, andrey.koltavskiy@gmail.com)

1 ABSTRACT

Medical care is a vitally important part of successful smart cities further development. High quality medical treatment has always been a challenging task for administrative departments of cities government. The key reason is that the treatment of patients significantly depends on the skills of medical stuff that can hardly be controlled and estimated. Semantic technologies by now have showed capabilities to solve highly complicated badly formalized problems in conditions of uncertainty. It makes reasonable to apply them in medical domain. In the paper a real example of information system for semantic medical care is presented. The system is being developed for Federal Almazov North-West Medical Research Centre in St-Petersburg, Russia (http://www.almazovcentre.ru/?lang=en). The main attention is paid to the proposed solution for the problem of medical treatment estimation in administrative and managerial departments. We focus on medical treatment examinations matching, trend analysis and administrative analytical and prediction task solving making use of semantic technologies, statistical analysis and deep learning applied to huge amounts of diverse data. Semantic medical data analysis project is an attempt to proceed to semantic medicine - an interoperable approach to medical domain area.

2 INTRODUCTION

Medical data is growing dramatically every year, the volume of this knowledge doubles every two years. With large-scale digitization, several medical engines went on display, such as public searching systems of biomedical literature, specialized catalogues and indices for health workers and quality management systems. The latter of which is most significant, because it allows setting the feedback during treatment and influences the results in real time. However, while quality management systems have a significant contribution to making large medical databases accessible, their users often have to deal with the burden of browsing and filtering the numerous results of their queries in order to find and analyze the precise information they were looking for. This point is more crucial for practitioners who may need an immediate response to their queries during their work and for administrative workers who need to control and improve the process of treatment.

When we talk about the purposes of health care, the main ones are to increase the level of public health and satisfy needs for qualified medical aid. At first glance, these tasks are easily solved, but they encapsulate a set of problems. Nowadays quality of medical care is considered from different perspectives - effectiveness, sufficiency, economy and moral standards. There are also many different definitions of qualitative medical help, such as interaction between doctor and patient, physician's ability to reduce the risk of disease progression and the emergence of a new pathological process, optimal use of medical resources. So many aspects of the one process cause complexity not only for analysis, but for comprehending as well.

Formalization of medical processes meets a number of problems:

(1) Each patient is a unique living organism. Results of treatment depend on its own peculiar features, current state and whole case history. Existence of these factors makes the majority of pattern-based approaches insolvent.

(2) The state of the organisms rapidly changes in time. The changes take place in conditions of occurring planned and unexpected events. The consequences of events usually come only in some time after the event. Besides each event is supposed to be considered in relation to other events. Dealing with events requires application of fuzzy temporal logic that is highly complicated.

(3) Joint usage of medicines for patient treatment can take different effect depending on the combination of medicines, earlier prescribed medicines and conditions of their taking. It means that doctors prescribe medicines is conditions of uncertainty.

(4) Gathered data about a patient requires efforts for its understanding and consumption. The data is scrappy as the parameters that describe the patients' state are measured from time to time. Moreover, the timelines of the results of the measurements and occurred events are not coherent. Data must be preprocessed using mathematical and statistical procedures.

Unfortunately these technologies cannot be directly used for raw objective and subjective medical data processing and analyses. It is proposed to use deep learning models, methods and algorithms to extract information and knowledge from initial data and represent it in standard formats that are supported by semantic stack.

In this context, we need system able to respond to administrative worker queries fast and accurate. Responses should allow manager to track health care dynamics, identify the causes and react on them.

3 BACKGROUND AND PROBLEM STATEMENT

Nowadays many quality management and medical information systems provide basic services. Information systems have penetrated and became an indispensible part of city medicine from online register offices for medical appointments to diagnostic systems based on symptoms and high-load distributed among the medical establishments medical error reporting information systems aimed at patient safety intervention (Riga M.). However, automated diagnostics and treatment systems (Edvin C.) appear to be a double-edged sword if law background is taken into account: a doctor is a legally responsible person, thus, he is punished by law in case of wrong diagnosis and treatment, or, in worst case, in case of patient's death. If obligations are imposed on the information system, there is nobody to blame, which is unacceptable in most of the establishments. Therefore, we have no intentions to overview or develop any system claiming to replace the doctor.

Most of the existing researches are focused on telecare medical information systems (TMIS), i.e. a set of various medical services for patients and practitioners (Zeeshan S. et al, Qi J. et al, Dheerenda M.), however, most of such systems are no more than a way of communication between patients and doctors and a way to store medical records and identify the patient, which is still far away from intelligent medicine managerial information system. Some of investigations are dedicated to Internet of Things among medical measuring devices to obtain telemetric indicators of patient's state (Boyi X. et al). Other researchers spot on automatization of medical processes in dedicated sections or examinations like automated optics calculation in ophthalmology (Nilanjan D., et al), which is undoubtfully a magnificent and convenient way to simplify doctors' routine tasks, but does not provide a way of comprehensive analysis and automation of all time-consuming routine processes of doctors' and administration workflow as a result of lack of structured knowledge instead of data and algorithms to process it. In accordance with the identified problems, the great solution may be adding the semantic aspect to the technology stack of the system which allows processing meaning and knowledge, but not the raw data.

Recently there has been an explosion of new data sources about diseases, drugs and treatments. Integration of these data sources and the identification of patterns that go across them are of critical interest. Through integrated and intelligent data mining, this information could provide important insights into the complex functions of the process. However, this can only be achieved when data is semantically integrated (i.e. using multiple data sources that are connected in meaningful ways) and in particular when all resources are brought together in such a framework.

There are critical problems in medicine that can only be solved through computational analysis of this kind of integrated information about every patient and its case history. For example, it is considered increasingly important to profile existing and potential new drugs for their effects across many targets, not just a single target of interest. Only by exploring the relationships of the drugs to a wide body of target information can we determine this profile.

Further, we can determine the ability of physician actions and drugs to influence at multiple points of treatment process, this will provide more robust efficacy in subsequent ones. Relationships between these pathways and potential side effects of drugs can only be determined by large-scale analysis.

Implementing such an integrated system involves creation of large networks of linked entities from multiple, heterogeneous and unstructed sources. It must give a possibility to query these data in ways that go beyond querying from a single source and allow inferencing among cross-domain information (medicine domain and complementary domains). Currently, there are significant barriers to carry out this kind of analysis. Many of the needed data sources overlap and cover similar data (we refer to them as homogenous or semi-homogenous data sources) but with slightly different foci. All data sources also tend to be published in very diverse formats (text files, journals, XML, relational databases) and may be structured or unstructured. The semantic relationship of these datasets to each other is often unclear.

Recent Semantic Web technologies provide efficient ways to integrate heterogeneous data. Various semantic languages have been established to represent and query semantic meaning of data and relationship. Our main goal and intention is to survey and design a system to support doctors and other medical establishment workers in routine task solving: matching of objective analysis data and medical notes, verification of treatment regimes including lack or excess of treatment, a set of predictive or analytical tasks for research purposes and other problems which are not contrary to law, in other words, managerial tasks and a set of doctors routine tasks.

4 SMDA PROJECT

SMDA (Semantic Medical Data Analysis) project for Federal Almazov North-West Medical Research Center (Saint-Petersburg) is an attempt to extract data in medical domain area and build an aggregated knowledge base and system following ontological approach, enabling complicated analytics, diagnosis comparison and control based on both objective numerical indicators and subjective certificates and prescriptions. Main challenges within the domain we deal with are:

(1) a multi-stage service-driven natural language processing to analyse and process medical certificates and prescriptions including concept matching to collate synonymous concepts within domain lexicon and terminology;

(2) high-performance unstructured data extraction, big data processing to gather, aggregate and analyse flows of madical data in s scalable way to provide universality;

(3) building and providing formal models for medical data, information and knowledge representation in a vivid and easy-to manage way;

(4) ontology learning, design and population and deep learning to produce a powerful knowledge base in order to meet system consistency requirements;

(5) building a linked semantic data space based on knowledge graphs and a set of ontologies related with medicine and complementary domains for continuous knowledge usage, refinement and enrichment;

(6) methods and algorithms for adaptive medical data real-time processing and analysis;

(7) semantic business logic implementation as a dynamic medical business process design, configuration and execution;

(8) design and implementation of a framework for semantical medical systems in such a way that the target establishment or a complete domain area may be changed and the system may be adopted to it using a flexible agile approach;

(9) rule-driven productive logical inference to process, analyze and verify data and make predictions;

(10) providing a functional stack of smart services for each specific task and workplace and combine them into semantic service network including coherent set of services selected from global net of services as well as internal services designed for particular specific tasks.

A combination of solutions for mentioned challenges provides a technological stack and framework to build up the distrubuted service-based system with the following resulting functionality:

(1) use of multimodal data sources including both structured and unstructured;

(2) linked data, information and knowledge space providing both human-readable and machine-comprehensible data;

(3) complicated analytics for doctors, econimists and administration as a step to next level of solution application;

(4) adaptive model-driven business processes, solutions for complicated issues within the context;

(5) agile ontology-based architecture of SMDA systems to meet the needs of all-level end users within the domain of medicine.

4.1 Prototype overview

The platform is constituted of a number of layers including a number of components grouped by tasks: storage, internal processing and external processing. The central part of the platform is internal domain-oriented semantic processing of data including domain area (input and output of domain-oriented information), processing (based on stream of incoming data, information and knowledge, DIK) and semantic blocks (storage of knowledge and tools to process and display knowledge and produce new knowledge). The central part of the system appearing to be the server is a combination of components (Fig. 1) interacting with each other in some way and having a common source input and client output interfaces to provide the system with unstructered or semi-structured data, exucute processing stages on the server side and present the results of processing to managerial personell. The input dataflow requires data transformation from raw unstructured representation into a well-defined semantic form which includes preprocessing, processing and postprocessing followed by storing into a knowledge base formed of an integration of semantic components and services such as triple store (BlazeGraph), knowledge management (MetaPhacts) and visualization (OntoDia) sybsystems as well as a separate model of semantic service graph not discussed within the paper as a separate work is required. Note that preprocessed semi-structured data is stored as well in InterSystems DB Cache and provided with interaction interfaces to deal with numerical objective data and textual data (concepts and concept relation in described case).

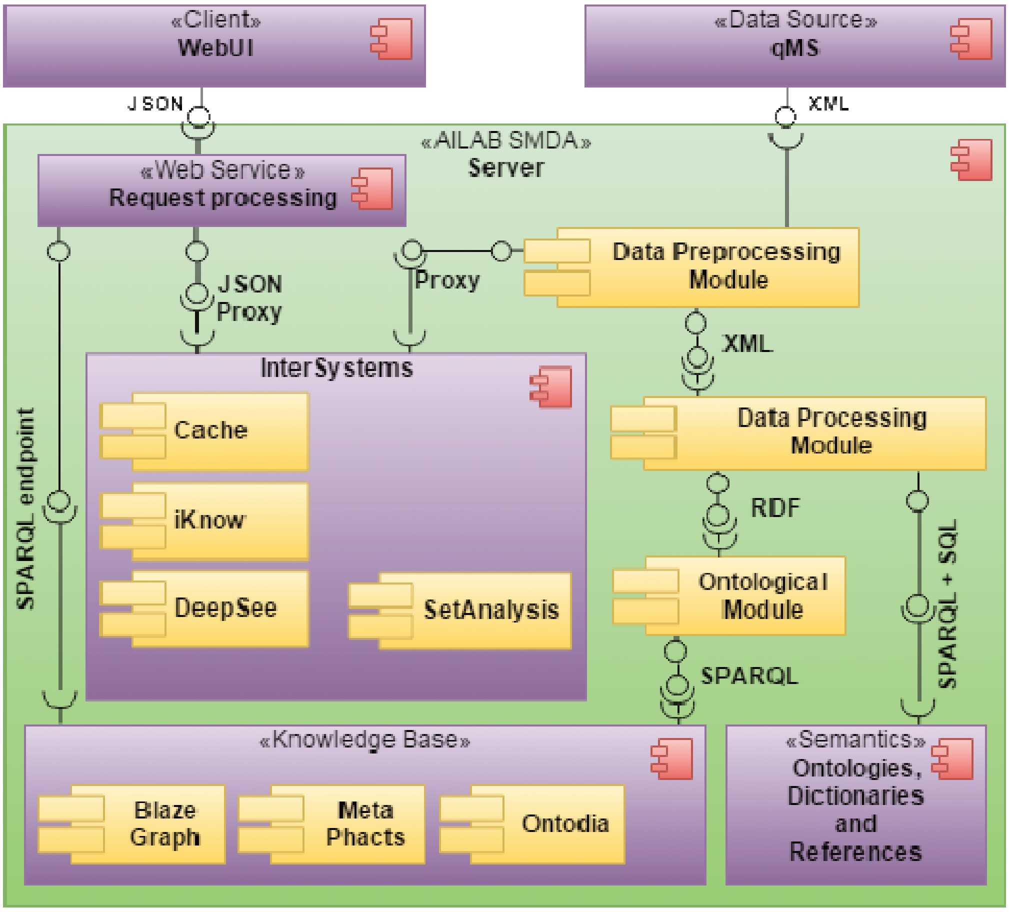

Fig. 1. SMDA simplified component diagram

All interactions between the components depending on the level of data structure are implemented via standard interfaces: XML and JSON for semi-strutured and structured data exchange, RDF upload API to import datasets into a triple store of knowledge base and SPARQL-endpoint for inference and interaction with the knowledge base while request processing.

A separate but tightly integrated component is a layer of external services and a set of complementary ontologies, dictionaries, standards and references for the task of knowledge enrichment and inference. External services are a set of single-task APIs described in the knowledge base together with internal services supported by input data format and structure requirements, output data format and structure, descriptions and deployment URLs to form a semantic service graph providing an interface for building-up a processing chain depending on input data format and desired output (requested administrative task). For better reliability and robustness all of the internal components may be either deployed at one physical server or distributed among separate machines on conditions of proper backup (Fig. 2).

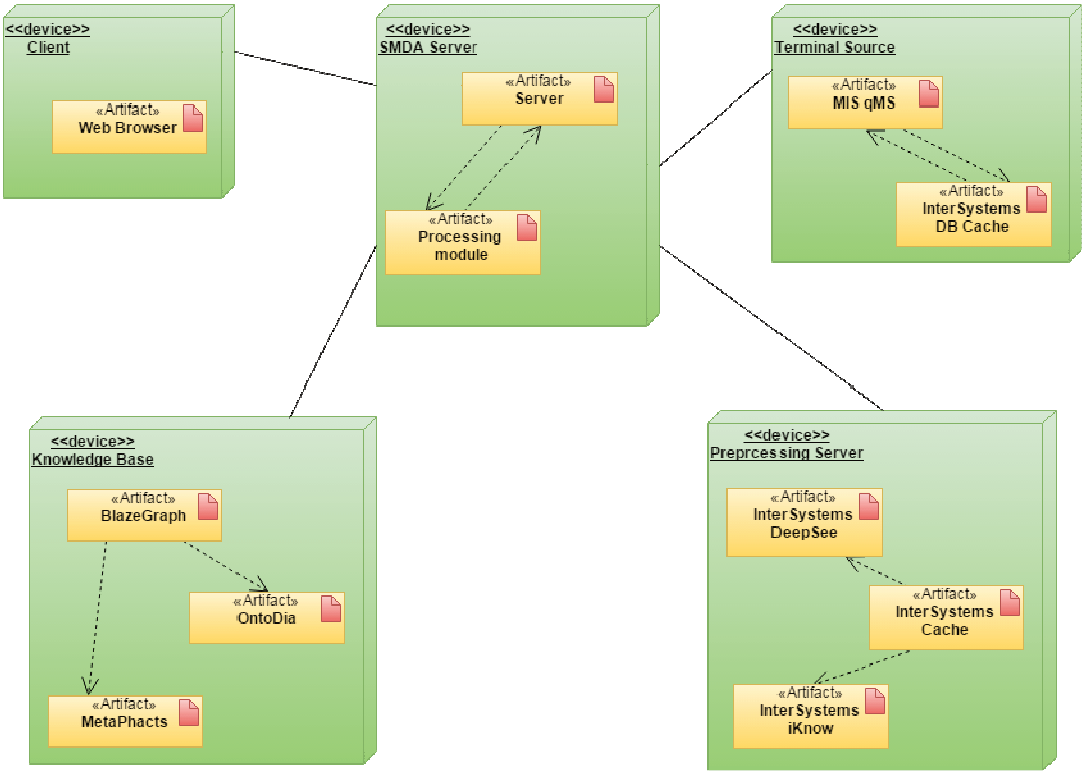

Fig. 2. Component deployment diagram

4.2 Workflow and data processing pipelines

All medical data of Almazov medical center is presented by objective numerical data (analysis and examination indicators provided with measurement units), organoleptic test indicators and textual subjective notes and image data which assumes at least three parallel data processing pipelines. We omit image data processing for the prototype as in requires special expert knowledge to develop a set of methods and focus on numerical and textual data processing pipelines (Fig. 3).

Numerical data is a subset of objective data obtained from results of analyses, examinations and measurements of patient's state (blood test, blood pressure, body temperature and others) as well as demographics (current age, date of birth, sex and others). This kind of data, in general, is presented by numbers, number pairs or number intervals accompanied by measuring unit and timestamp in various formats which are to be unified for further processing and analyses. Objective data is the main source for managerial statistical and research tasks which requires a powerful math unit provided by semantic service graph. Numeric data processing includes measuring unit refinement as far as units are textual data pieces and lack in standardization within the domain; formatting of floating-point numbers as they are initially provided as text strings; storing preprocessed data into a database with respect to domain object model. In the course of further postprocessing tasks on administration request all required numerical data undergoes specific calculations and comparison with norms depending on patients' age and sex to present statistical overview and trends of needed data section. However, not only is objective data required for most of the administrative tasks, information distributed over medical certificates and notes matters as well.

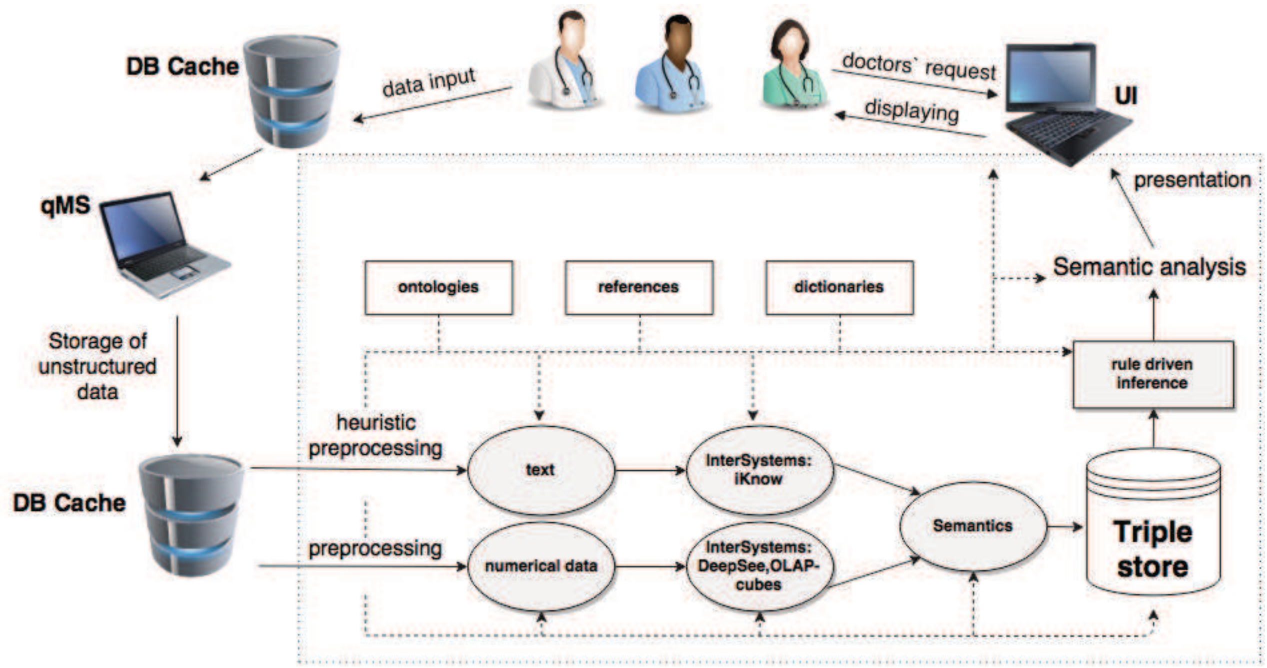

Fig. 3. Workflow and data processing pipelines

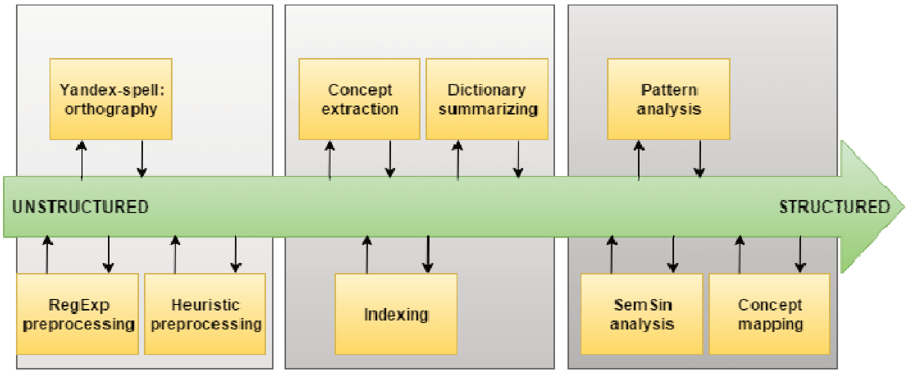

Fig. 4. Textual data processing pipeline

Manual thorough textual data analysis has shown that medical texts are specific domain-area related text pieces possessing a number of traits which leads to lack of possibility to treat them as usual textual data by widely used classification tools. Our tool needs deep analysis of medical documents in order to extract relevant information. At the first level of this information come the medical entities (e.g. diseases, drugs, symptoms). At the second, more complicated level, comes the extraction of semantic relationships between these entities. Information extraction is a complex task which is necessary to develop high-precision information retrieval tools. Our approach is based on the use of linguistic patterns. For every couple of medical entities, we collect the possible relations between their semantic types (e.g. between the semantic types Therapeutic or Preventive Procedure and Disease or Syndrome there are a few relations: treats, prevents, complicates, etc.). We construct patterns for each relation type and match them with the sentences in order to identify the correct relation. The relation extraction process relies on two criteria: (i) a degree of specialization associated to each pattern and (ii) an empirically-fixed order associated to each relation type which allows ordering the patterns to be matched. We target six relation types: treats, prevents, causes, complicates, diagnoses and sign or symptom.

Every textual note undergoes a number of preprocessing and processing stages before being stored into a knowledge base which are out of the scope of this work and are to be proposed in a dedicated paper. Nevertheless, we mention the chain in brief (Fig. 4):

(1) text refinement and purification by means of regular expressions heuristics to eliminate domain-specific noise as a result of human factor and restricted time reception at the doctor causing contractions and abbreviations;

(2) further text preprocessing making use of internal and external services mostly aimed at spell-checking and mistake correction;

(3) text conceptualization (concept and relation extraction) by means of InterSystems iKnow;

(4) semi-automated thesauri organization with a help of involved experts within the domain area or from the establishment, mainly to match establishment-specific synonymous lexemes and collations;

(5) structural analysis of the text by means of SemSin tool to extract sentence structure in a machine-processible xml form;

(6) pattern analysis based on predefined templates (experts are involved) and mentioned above thesauri to form the triples to be stored into a knowledge graph.

Now, when all raw data, both numerical and textual, is transformed into semantics (a data-set of triples "subject" - "predicate" - "object") and stored into the triple store, it becomes available for semantic analyses on requested user tasks involving external linked data sources, ontological and productive inference.

5 SEMANTIC IN SMDA PROJECT

Semantic layer of semantic medical data analysis project is represented by a number of third-party platforms, specific tools and services, triple store, medicine ontology and complementary ontology, inference engine and s set of thesauri, references and dictionaries to form a semantic knowledge and service graph for flexible task solving. The third-party tools we use for particular storage and analysis tasks include:

(1) BlazeGraph used as a triple store and SPARQL-endpoint for graph search and knowledge inference and a high-performance knowledge base graph platform, supporting RDF in a scalable and flexible way;

(2) MetaPhacts as a platform providing solutions and a number of services to describe, query and interchang graph-based data, as well as a way for visual analysis and interaction with knowledge graphs;

(3) OntoDia as free and user-friendly online OWL and RDF diagramming tool to present a knowledge base;

(4) Apache Jena as an open source fre of charge Java framework for Semantic Web and Linked Data applications based on onologies and elements of enference.

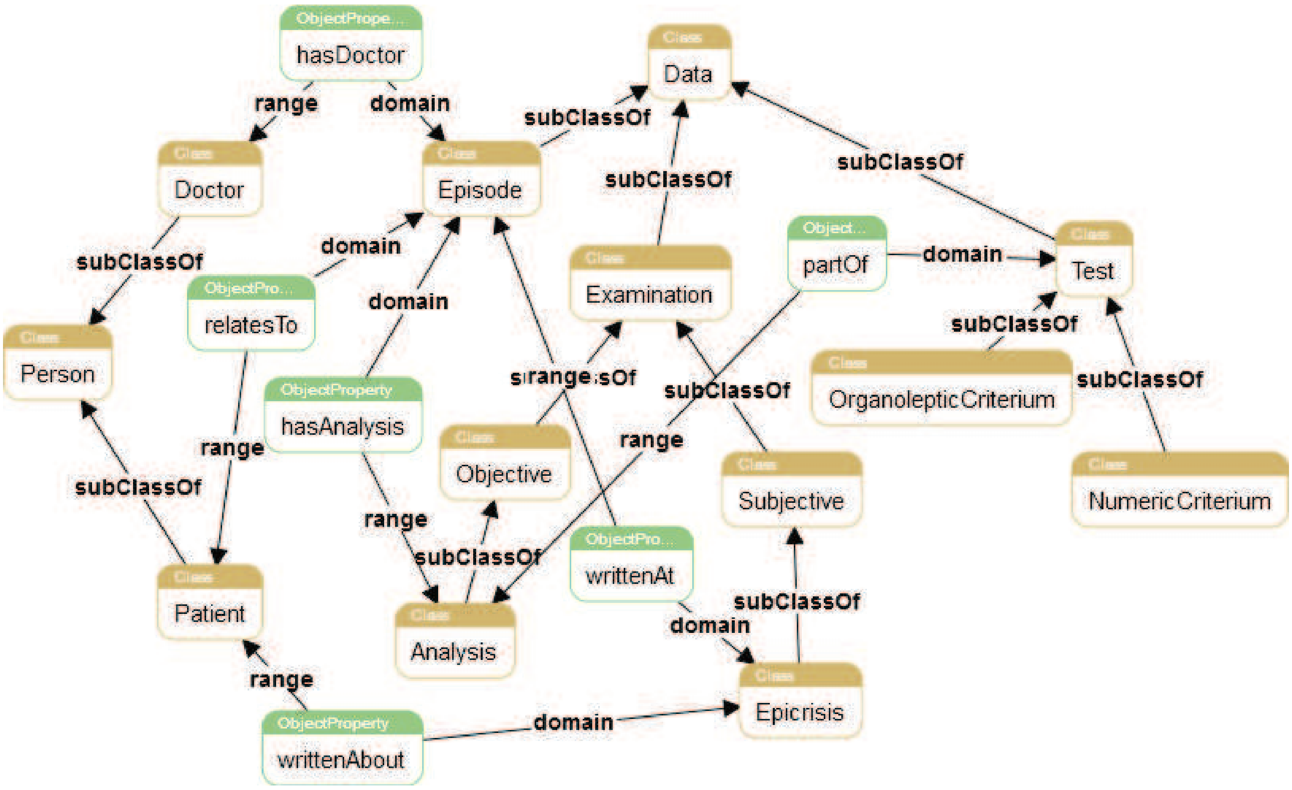

Fig. 5. A prototype subontology of SMDA-ontology

The central part of ouur system semantic core is a knowledge base, which includes semantic service knowledge graph, medical ontologies, references and thesauri as well as context-aware well-structured semantic establishment-specific medical data covered by SMDA-ontology (Fig. 5). The processes, technologies and tools for performing preprocessing tasks to populate the knowledge base are outside the scope of this paper and describes in dedeicated works, the brief process is presented in previous sections.

6 CASE STUDY

6.1 Objective and subjective data matching

One of the best ways to evaluate the effectiveness of physician is to match objective and subjective data of the treatment process. It means that we compare current situation with a standard sequence. Objective data includes precise results of analysis and examinations. It is presented by objective numerical data (indicators provided with measurement units). Subjective data are physician notes and records, which are usually textual, unstructed and hard to distinguish and process automatically. This data is often based on objective parameters. Comparison of these data types allows administrative officer to determine physicians' mistakes. The most common of them are misdiagnosis, prescribing unnecessary test and procedures, wrong medication reception, choosing wrong treatment regimen and incompetence in the matter of a particular field.

At first glance, this method of comparison allows us to consider the error only after the fact, when the patient's health is already irretrievably damaged. But firstly, it will prevent such incidents in the future, by excluding negative influencing factors. Secondly, it will provide us with basic knowledge for further identification of such patterns of behavior or qualification and solve the problems in the early stages.

Example: Percentage of diagnosis "hypertension" increased in the abstract cardiology department. After a couple of complaints on the treatment the comparison according to the above scenario was done. That revealed the incompetence of the doctor which having objective indicators of low pressure of the patient made the opposite diagnosis. The reasons for this behavior have been clarified on the spot, measures have been taken, and most importantly - primary knowledge about the incident (exact numeric statistics) have been received. Soon it will identify similar chain of events.

The above method will work even if the problem will not come from the doctor, and lies, for example, under the faulty equipment, which initially gave the wrong allegedly objective data. Any deviation from the standard will be noticed.

6.2 Statistical data analysis for managerial and research tasks

In addition to improving efficiency in specific areas, it is important to observe the general dynamics of the processes that are directly related to medicine. Basically it is simple and well-structured global data. The most common statistic - a mortality distribution by diagnosis and detected diseases, trends in the market of medicines and medical equipment. Even these raw data bring benefits, suffice to visualize them in the right way and to provide a convenient interface for manipulating them.

6.3 Productive and analytical functionality for trend analysis

Having a knowledge base with sufficient precision, but different data, you can make the process of interlinking these data with other data pieces. On the basis of these relationships we can make inferences. This will reveal previously unnoticed dependencies and trends which significantly optimizes the processes taken separately.

As in the previous method, the information we get will be used as the initial data for further machine learning, in order to avoid repetition of implicit errors .

7 CONCLUSION

In this study we showed that introduced techniques and technological stack are applicable for the issue of medical routine tasks automatization and managerial analysis efficiency and trends. Along with the techniques we introduced a tool-chain that potentially dramatically improves the quality medical service and treatment as a result of continuous effectiveness control an preventive measures after early identification of mistakes or poor qualification. However, the system we propose is not a way to undermine the authority of physicians. We propose a way to make city medicine more smart on the way to semantic medicine as a new paradigm of healthcare. We just assume that human factor is possible in any domain.

Despite the prospects we spot, the prototype is to be improved dramatically to provide a powerful, stable and reliable solution. Investigation and analyses of further cases are desired as well as smart context-sensitive user interface so that managerial tasks are implemented in a completely semantic way.

8 REFERENCES

Riga M. et al. "MERIS (Medical Error Reporting Information System) as an innovative patient safety intervention: A health policy perspective." Health Policy 119.4 (2015): 539-548.

Iliff, Edwin C. "Computerized medical diagnostic and treatment advice system including network access." U.S. Patent No. 9,005,119. 14 Apr. 2015.

Zeeshan S. et al. "Smart environment as a service: Three factor cloud based user authentication for telecare medical information system." Journal of medical systems 38.1 (2014): 1-14.

Qi J. et al. "A privacy enhanced authentication scheme for telecare medical information systems." Journal of medical systems 37.1 (2013): 1-8.

Dheerendra M. "A study on id-based authentication schemes for telecare medical information system." arXiv preprint arXiv:1311.0151 (2013).

Boyi X. et al. "Ubiquitous data accessing method in IoT-based information system for emergency medical services." Industrial Informatics, IEEE Transactions on 10.2 (2014): 1578-1586.

Nilanjan D. et al. "Firefly algorithm for optimization of scaling factors during embedding of manifold medical information: an application in ophthalmology imaging." Journal of Medical Imaging and Health Informatics 4.3 (2014): 384-394.

Boyarsky, K., Kanevsky, E.: "The semantic-and-syntactic parser SEMSIN". In: International Conference on Computational Linguistics Dialog-2012 (2012).

Mouromtsev D., Kovriguina L., Emelyanov Y., Pavlov D., Shipilo A., "From spoken language to Ontology-Driven Dialogue Management", Lecture Notes in Computer Science, 2015, vol. 9302, pp. 542-550

Brands, Michael Rik Frans, and Dirk Medard Helena Van Hyfte. "Data integration and knowledge management solution." U.S. Patent No. 7,428,517. 23 Sep. 2008.

Brands, Michael Rik Frans, and Dirk Medard Helena Van Hyfte. "Data processing based on data linking elements." U.S. Patent No. 7,912,841. 22 Mar. 2011.

Brands, Michael Rik Frans, and Dirk Medard Helena Van Hyfte. "Data analysis based on data linking elements." U.S. Patent No. 9,053,145. 9 Jun. 2015.