

Analysis and visualization of disease courses in a semantic enabled cancer registry

Angel Esteban-Gil¹, J. T. Fernandez-Breis^{2*} and Martin Boeker³

¹ Fundación para la Formación e Investigación Sanitarias de la Región de Murcia, IMIB-Arrixaca, 30003 Murcia, Spain
angel.esteban@ffis.es

²Dpto. Informática y Sistemas, Facultad de Informática, Universidad de Murcia, IMIB-Arrixaca, 30100 Murcia, Spain
jfernand@um.es

³Center for Medical Biometry and Medical Informatics, Medical Center – University of Freiburg
martin.boeker@uniklinik-freiburg.de

Abstract. OBJECTIVE: To develop a platform that incorporates the benefits of semantic Web technologies into traditional cancer registries.

BACKGROUND: Cancer registries are important for research and the management of the cancer treatment. Many technological solutions are available to manage data for cancer registries nowadays, however, the lack of a well-defined common semantic model is a problem when customizable analysis and linking of data to external resources are required.

METHODS: OWL ontologies and an RDF repository were employed for the transformation of a cancer registry database into a semantic enabled representation. The feasibility of the approach was tested with simulated data.

RESULTS: An ontology representing the semantics of an institutional cancer registry was developed. We implemented a platform where the complex timeline of the patient's disease can be clearly represented. Moreover, the semantic structure of the representation renders it easy to analyse key figures on aggregate patient level.

CONCLUSION: The presented platform is an example of the parallel development of ontologies and applications that take advantage of semantic web technologies in the medical field.

Keywords: Biomedical Informatics, Semantic Web, Cancer Registry, Ontology

1 Introduction

Regional and epidemiological cancer registries are important for cancer research and the quality management of cancer treatment. Many technological solutions are available to collect and analyse data for cancer registries nowadays, however, the lack of a well-defined common semantic model is a problem when user defined analysis and linking of data to external resources are required.

To solve this problem, we have designed a semantic model for local cancer registries and implemented it with semantic web technology in a feasibility study. Our proposal is based on our previous results and experience working with semantic technologies. We use OWL [1] ontologies and an RDF [2] repository, for transforming a traditional cancer registry database into a triple store based on the semantic information derived from the requirements and the database structure. For semantic retrieval of data we employ SPARQL [3] queries.

Based on the requirements analysis, an OWL ontology has been developed that models the semantics of an institutional cancer registry in a pragmatic extensible way. Based on this model, we have implemented a Semantic Web platform that allows to transform and store data from cancer registries in RDF repositories. With a graphical user interface of this platform, users can also formulate incremental user-defined queries. The query results can be displayed in several customizable ways. The complex timeline of the disease of individual patients can be clearly represented. Different events, e.g. different therapies and courses of the disease, are presented according to their temporal and causal relations. Moreover, the semantic structure of the representation renders it easy to analyse key figures on aggregate patient level.

We have applied our approach using the requirements of the cancer registry of the Comprehensive Cancer Center Freiburg (CCCF)¹ and used simulated data to generate a semantic repository. A prototype web platform has been implemented demonstrating appropriate functionality and performance.

2 Background

2.1 Cancer registries

Cancer registries are complex structures for the documentation and analysis of data from patients diagnosed with cancer [4, 5]. Different types of cancer registry are organized to collect patient data from institutions (institutional), regions (regional) or countries (epidemiological). Whereas epidemiological registries provide mainly population based information on morbidity and mortality, institutional and regional registries can provide fine-grained information on treatment and conditional survival.

Information of regional cancer registries serves different requirements. Besides others, the quality control of patient care, the comparison of patient-related outcome parameters and research support are the most prominent ones. Institutional and regional registries are also the main data source for epidemiological cancer registries.

Regional cancer registries collect information about the diagnosis, therapies and course of the disease [6]. Detailed information on the histo-pathology of the primary tumor including tumor staging and grading is most important. The longterm follow-up of the patients' vital status is one of the resource intensive tasks of tumor registries providing the basis of survival analysis.

Most information is derived from primary documentation aimed at patient care. For the purpose of structured secondary documentation, tumor documentaries carefully

¹ <http://www.uniklinik-freiburg.de/cccf.html>

reprocess primary documentation and code most of the information especially of the diagnostic and treatment information with clinical coding systems (ICD-10 [7], ICD-O-3 [8], TNM [9], and others). In many countries, a standardized common dataset has been developed to better support exhaustive data exchange with the epidemiological cancer registries.

For collection, storage, retrieval and analysis several electronic solutions are available (e.g. METRIQ², OncoLog Registry³ or CNEXT⁴). Proprietary systems have been developed for large institution. Software for cancer registries has to fulfil a large set of rapidly changing requirements. Scientific progress and changing legal conditions complicate the development process and the standardization of cancer registry software.

2.2 Standards and Classification Systems

The most important clinical classification system applied in cancer registries is the ICD-10 (International Statistical Classification of Diseases and Related Health Problems). This classification system is divided in chapters, with blocks of diseases. For example, chapter II includes the classification for neoplasms between the blocks C00 and D48. These blocks are subdivided in hierarchies that further specify the diagnosis. The ICD-O (International Classification of Diseases of Oncology) is a domain-specific extension of ICD for cancer diseases. The classification system is dual allowing the coding of topography (tumor site) and morphology of the tumor. This coding system is of particular interest because SNOMED-CT [10] (Systematized Nomenclature in Medicine – Clinical Terms) has adopted the ICD-O classification of morphology.

Several staging systems for cancer have evolved over time and continue to change with scientific progress. The most important classification system is the TNM classification for the description of the anatomical extent of the disease. This system is under constant development of the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC). The TNM staging is based on the size and/or extent (reach) of the primary tumor (T), the amount of spread to nearby lymph nodes (N), and the presence of metastasis (M) or secondary tumors formed by the spread of cancer cells to other parts of the body.

Clinical procedures are coded, e.g., with the ICD10-PCS (Procedure Coding System) [11]. In this system, each code has seven alphanumeric characters. The character position indicates the clinical classification of the procedure. For example, the first position is used to define the surgical section; the second position is used to define the body system.

² <http://www.elekta.com/healthcare-professionals/products/elekta-software/cancer-registry.html>
³ <http://www.oncolog.com/?cid=7>
⁴ <http://www.askcnet.org/>

2.3 Semantic web technologies

The Semantic Web can be seen as the next-generation web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [12]. Ontologies [13] constitute the standard knowledge representation mechanism for the Semantic Web. Technologies like OWL for the ontology construction enable a formal representation of the domain.

Important international initiatives [14, 15] strive to ensure that the Semantic Web becomes a fundamental system to achieve consistent and meaningful representation, access, interpretation and exchange of clinical data. The focus of our work lies in the intersection between the research domains of knowledge management and health applications, in which there is an increasing use of the semantic web technologies for managing the knowledge of health information systems. There are a lot of studies that use semantic web technologies like OWL to represent cancer diseases. To mention some examples, in [16] an ontology for modeling clinic-genomic trials on cancer has been developed. In others works [17, 18], ontologies that represent a concrete type of cancer disease have been developed and used in the construction of technical solutions.

In our approach, we try to take advantage of the best features of the combination of semantic technologies like OWL and RDF. This means that part of the processing will be performed using OWL (domain level) and the rest using RDF (data level). For querying the information we will use SPARQL.

3 Methods

3.1 Ontology construction

Best practices in ontology engineering recommend to reuse existing and to create modular ontologies [19]. These recommendations are implemented reusing concepts from different ontologies so that the resulting ontology infrastructure is likely to be a networked ontology. The OBO Foundry has also developed a series of principles for ontology construction which propose principles for modularity, orthogonality and reusability [20].

Our method for constructing the domain ontology used in this work consisted on identifying the main entities that should be represented, searching for existing ontologies containing classes representing these entities, selecting the most appropriate ones (by our subjective criteria), and extending them when necessary. The final ontology has been implemented using Protégé⁵ in OWL-DL, which is the OWL subset based on Description Logics. By proceeding in this way, the domain knowledge is made explicit in a set of OWL axioms and therefore ready to be exploited by means of automated reasoning.

⁵ <http://protege.stanford.edu/>

3.2 Data transformation and exploitation

Clinical data are usually stored in relational databases. Different methods and tools are available for the transformation of relational data into semantic formats like D2RQ⁶ or Triplify⁷. Most approaches perform a syntactical transformation, that is, they are based on the mappings between the relational and semantic primitives of the corresponding models resp. languages. Consequently, the meaning of the content is not really exploited in such transformation processes. In this work, we use a transformation approach which more likely preserves the meaning of the contents based on the specification of mappings between the entities of the source relational schema and the entities of the target domain ontology.

For this purpose, we use SWIT [21], which is our semantic transformation engine capable of generating RDF and OWL repositories from both relational and XML-based databases. Besides transforming the data, SWIT prevents the generation of inconsistent semantic data sets with the support of DL reasoners by not transforming inconsistent source content. The transformation method has three main steps: (1) definition of the mapping rules between the fields of the database and the ontology; (2) generation of the OWL data; and (3) importing the OWL data into the semantic data store. We use Virtuoso⁸ as data store. Virtuoso has been used in other medical works as [22]. Our semantic data store uses OWL files to save the ontology axioms and an RDF repository to persist the individuals. This permits separating the semantic model from the semantic data.

SPARQL is used as retrieval language used for expressing and issuing queries over the data store. We use our ontology-guided input text subsystem called ODS [23] to facilitate clinicians the exploitation of the data store,. This tool is an editor for SPARQL queries supported by OWL models. The tool uses the underlying domain ontologies to show the necessary information to visually define SPARQL queries.

3.3 Generation of simulated data

In this work, we have generated a set of simulated data of a cancer registry by using random functions to assign a value of each property of a patient with a diagnosed cancer. We have ensured that some patients have more than one cancer diagnosis. Each diagnosis was assigned more than one treatment episode with a maximum treatment period of twelve months. Plausibility of simulated data was ensured by rules.

3.4 Semantic profiles

We define a semantic profile as a set of relations and properties that some individuals have. Semantic profiles permit to identify groups of patients that share some

⁶ <http://d2rq.org/>

⁷ <http://triplify.org>

⁸ <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/>

properties and are therefore useful for comparing and studying such groups. Ontologies are of special interest for creating profiles because they allow for aggregation and selection of individuals from a conceptual perspective.

In this work, the semantic profiles are built by using the entities defined in a domain ontology. We have generated two major semantic profiles, namely, timeline representation of a patient with cancer and aggregated disease timeline representation of a patient group with some common properties.

Disease timeline of a cancer patient

In the disease timeline of a patient various health related events occur (e.g., diagnosis, patient conditions, therapies and the disease courses). Retrieving these events of a patient required us to perform some data normalization for representing therapies by months. Figure 1 shows that every diagnosis has an associated timeline which includes therapies and the disease course, both ordered by month. E.g., for a breast cancer patient we can show the timeline with the applied therapies (surgical treatment, chemotherapy, etc.) for every period. Furthermore, we can show the course of the disease and its relation with changes in therapy. It also includes the date of the diagnosis and the date of the last encounter with the patient. At last, the profile contains the list of all patient diagnoses and a list with the patient conditions, including all its properties.

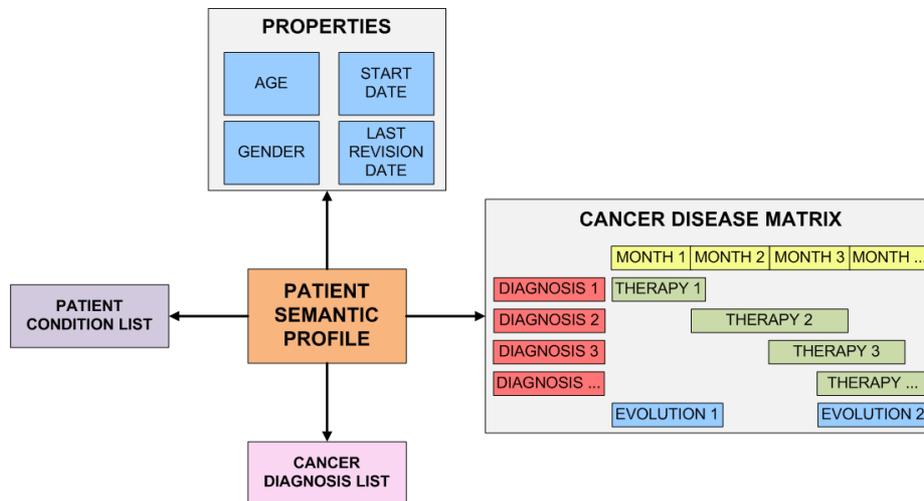


Fig. 1. Schema of semantic profile of a cancer patient

Aggregated disease timelines of a group of patients

The aggregated timelines of a patient group (see Figure 2) include all events of the selected patients that have the selection criteria in common, for a given period, and

for a concrete diagnosis. An ontology-driven search is performed for the selection of the patients. This permits to create groups of patients with the same diagnosis, staging, grading and age range. Second, the semantic profiles of each group member are obtained. Third, the semantic profiles are globally analyzed to retrieve all data on the included events. The result is a matrix that contains the disease courses of the included patients for every month of the disease.

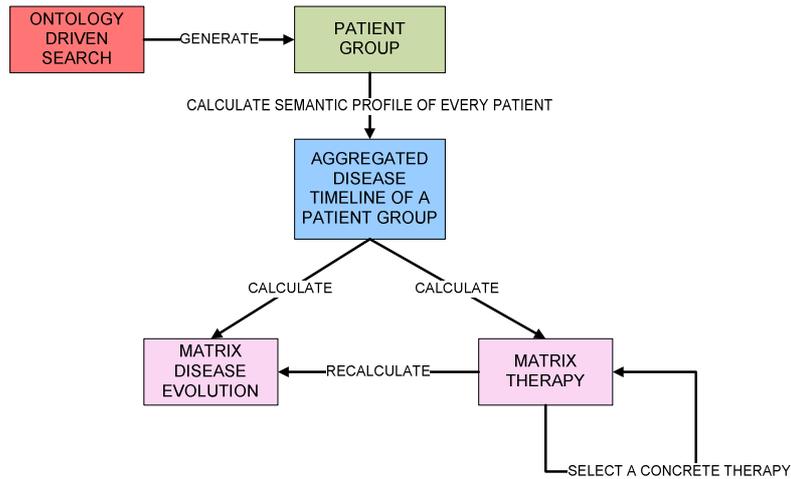


Fig. 2. Overview of the generation of aggregated disease timeline of a patient group

4 Results

The approach described in the previous section has been applied in an institutional cancer registry scenario. Based on the requirement analysis [24, 25], an ontology was developed that models the semantics of an institutional cancer registry. We have used this model and SWIT for transforming and storing simulated data from cancer registries in a semantic data store. We have implemented a Semantic Web platform that permits users to formulate incremental user defined queries with a graphical user interface based in ODS. The query results can be displayed in several customizable ways, allowing the generation of dashboards on demand. The complex timelines of the disease of individual and aggregated patients can be clearly represented.

4.1 The ontology

In the last few years, several cancer ontologies have been developed. For the purposes of this work, we have built a preliminary cancer registry ontology⁹ based on the existing ontologies and fulfilling the requirements of a local cancer registry. This first draft ontology represents some aspects of cancer diseases and their treatment prag-

⁹ <http://sele.inf.um.es/ontologies/cancer-registry.owl>

matically. The ontology has been defined in OWL-DL. The ontology contains a total of 335 classes, 18 properties and 29 object properties, with 2.581 logical axioms. The ontology covers the following classes:

- *Patient* represents a person with any type of cancer disease. Properties: gender, birth date, diagnosis, therapies and disease courses.
- *Patient condition* represents the health condition of a patient at a given time. Properties: reference date, age, weight, height, Karnofsky index [26], ASA index [27] and the menopause status.
- *Diagnosis* represents the patient diagnosis at a given time. Properties: ICD10 code, grading, staging, therapies, date, pathological structure, anatomical structure and tumor type.
- *Therapy* represents the patient therapies of a diagnosis at a given time. Different kinds of therapy as *Chemotherapy*, *Surgical Treatment*, *Nuclear Medicine* and others have been modeled in the ontology as subclasses of *Therapy* class. Properties: medication, start date and end date.
- *Disease course* represents the patient disease course of a diagnosis at a given time. Different kinds of course as *Complete remission*, *Progression*, *Recurrence* and others have been modeled in the ontology as subclasses of *Disease course* class. Properties: patient conditions and date.
- The ontology also includes some classes to represent the TNM classification system of malignant tumors. They include anatomical entities for cancer grading and staging, e.g. *Primary tumor*, *Regional Lymph Nodes* and *Distant Metastasis* hierarchies.
- *Health Classification System* is the superclass of all classes representing coding artifacts of health related classification systems. To build the taxonomies of classifications for a cancer registry, we tried to reuse other ontologies. For the ICD10 code we use the ontology built in [28]. For ICD-O and ICD10-PCS¹⁰, we have transformed spreadsheet versions into OWL.

4.2 The semantic cancer registry system

We implemented a prototype system¹¹ based on the methods described in previous sections. Figure 3 shows the three main parts of this system. The upper part of the figure shows the data transformation module, which makes use of SWIT to transform the original data in semantic information stored in the Semantic data store.

The cancer registry ontology is the core of the system that allows for computational management of the information related to the cancer patients. All the services offered by the prototype are implemented on top of this core. The data transformation requires mapping the source data schema to the cancer registry ontology.

The lower part of the figure shows the other two modules of the system. The right one shows the module for the analysis of individual patients, that is, extraction of profile and timeline analysis. The left one shows the module for the analysis of groups of patients, which also includes the functionality for graphical access to the disease

¹⁰ http://sele.inf.um.es/ontologies/ICD10_PCSv1.owl

¹¹ <http://sele.inf.um.es/SECARE/>

courses of patient groups. The Ontology Driven Searcher (ODS) permits to create group of patients with semantic properties in common. The ODS generates charts and tables with accumulated data of the semantic repository. In this case, the ODS has an option for adding the concept or property by which to group, so that it can be considered as a customizable dashboard designer. With this dashboard, the user can select and aggregate the information on every class of the semantic model. This tool is used as base for the construction of other services like the graphical representation of the aggregated timelines of a group of patients or the customizable dashboard.

The dashboard visualizes the concepts of the model as charted and grouped form. It is based in the ODS and can generate multiple incremental dashboards on demand. E.g., the user can generate a pie chart selecting patients by their first therapy. The user can save any dashboard for querying the results without having to generate it again.

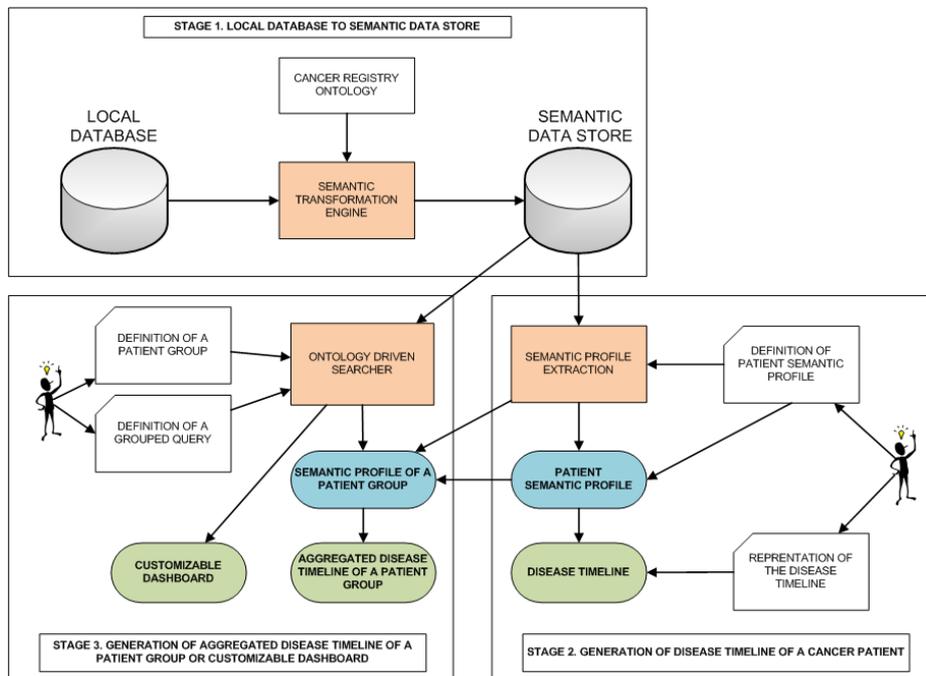


Fig. 3. Overview of the system

4.3 Simulated use case

We have performed an initial evaluation of the system. We have generated a simulated database with 207.190 patients. The time for the transformation since relational database to semantic data store of the simulated data has been thirty-two minutes.

Table 1¹² shows that the time performance of the semantic datastore is slower than the relational one for basic queries, without joins. However, when we try to join the data, the semantic datastore is better than the relational model, even with indexes. The results obtained when we filter by a property of the class or the table, the semantic datastore also is faster than the relational one.

Table 1. Results of the migration of the relational database to the semantic data store

Query	SQL count result	SQL time	SPARQL count result	SPARQL time
Recovery all Patients	207.190	0.060s	207.190	0.189s
Recovery all Therapies	400.290	0.132s	400.290	0.317s
Recovery all Diagnosis	240.088	0.070s	240.088	0.220s
Recovery all Courses	108.297	0.030s	108.297	0.155s
Recovery patients with diagnosis, therapies and courses	207.190	1.048s	207.190	0.204s
Recovery all female Patients	105.714	0.231s	105.714	0.189s
Recovery all female Patients with more of 60 years old	62.603	0.245s	62.603	0.192s

Graphical representation of the disease timeline of a patient

This service permits users to observe the main properties of the timeline of a patient with a cancer disease. In this view, users can see the details of the diagnosis and of every applied therapy in each period. Besides, users are provided with two evolution charts, which are based on the patient course and the Karnofsky index. Figure 4 shows an excerpt of the therapy and course timeline of a patient with pharynx cancer.



Fig. 4. Excerpt of the timeline representation

Graphical representation of the aggregated disease timeline of a patient group

¹² The test has been realized in a local machine with MySQL 5 as relational database and Virtuoso 7 as RDF repository.

Figure 5 shows patient selection and aggregation by the following criteria: male patients with an age between 50 and 70 diagnosed with colorectal cancer who have received Chemotherapy.

After selection and aggregation, the system generates charts that contain the therapies and the disease courses of the patients. This service can be employed as an exploratory therapy simulator. Optionally, the entire time matrix can be recalculated by the selecting a certain therapy. This can help the user to estimate which therapy is likely to be the most appropriate. Figure 6 shows an excerpt of the first two months of the therapies analysis panel of a group of 60 patients.

I want to recovery where:

Fig. 5. Ontology-driven searcher view

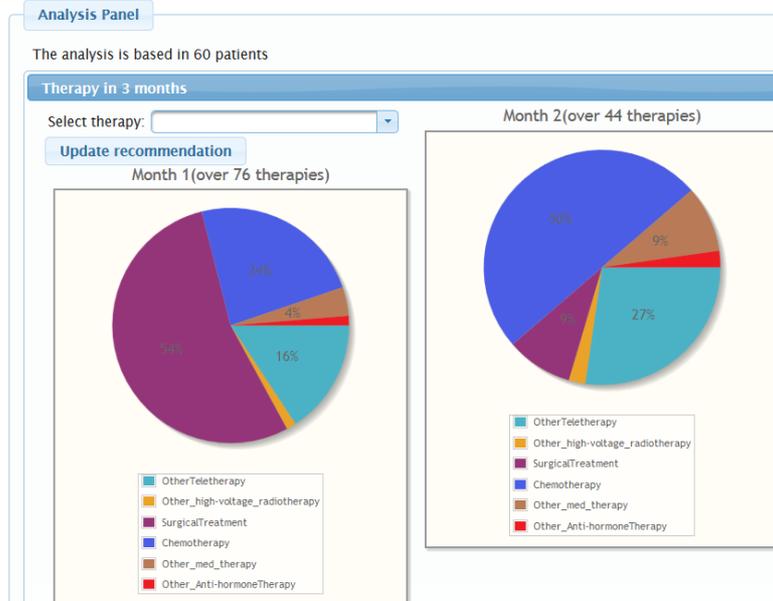


Fig. 6. Excerpt of the aggregated disease timeline of a patient group

5 Discussion

The main result of this study is the development of a Semantic Web platform that facilitates the analysis and visualization of data from cancer registries including (1) the representation of the disease course of a patient, (2) the representation of the aggregated disease courses of a group of patients, and (3) the definition of customizable dashboards for patient selection and visualization of the data.

Our approach allows to provide powerful and precise search capabilities assisted by a customizable dashboard adaptable to the requirements of each user. The query editor has been developed guided by OWL, providing the user the possibility of define statements in a more intuitive way than if we used a traditional relational model. Furthermore, the use of a NoSQL database like an RDF repository allows building a robust and scalable architecture for big clinical data warehouses [20]. Other important advantage of semantic knowledge modeling is the possibility of sharing information and comparison of clinical cases and processes.

The use of simulated data has demonstrated the viability of incorporating a local cancer registry to this model. A comparative performance analysis of relational databases and semantic repositories demonstrated excellent performance measures for the semantic repository.

Rule-based systems and logic-based models have been semantic approaches applied to cancer registries, like analysis of cancer registry processes [29], quality assurance [30] and decision support [31]. Our approach innovates by combining traditional technologies like relational databases and semantic web technologies like OWL and RDF. We have created an OWL ontology for representing some aspects of an institutional, local cancer registry. We have developed an RDF repository whose structure is driven by the OWL ontology and permits to work by exploiting the semantics of the content, so avoiding the need to care about the relational data structures, which facilitates the sharing of the content. Our technological infrastructure has permitted us to develop a semantic searcher for navigating through the complete cancer registry, to extract semantic profiles of the patients, and to analyze the structure of disease courses.

One limitation of this work has been to work with a preliminary version of an ontology of epidemiological cancer data. This ontology needs to be reviewed and extended, although it has been enough to demonstrate that the semantic exploitation of this type of clinical data is possible in a robust and scalable way.

Another limitation of this work is the lack of real data to test all the services and the quality of our domain ontology. The use of real data would enable to (1) test the performance of the system with a realistic volume of data and (2) evaluate the impact of missing data in the performance of the system, because cancer registry quality assurance studies have revealed that they often lack some data [30]. We plan to perform a real study with data from a large local cancer registry, which might also include a clinical validation.

Furthermore, we would like to perform term enrichment analyses [32] to estimate the use, for instance, of the ICD-10 codes. Finally, we would like to use this model to generate rules that serve to generate patient groups automatically or for quality assurance of the data.

6 Conclusion

This work demonstrates that the ontologies and the RDF repositories can be effectively combined for exploiting a local cancer registry. On the one hand, we constructed an ontology that models the knowledge of local cancer registry. On the other hand, we have used semantic web technologies for building a platform to analysis the complex timeline of a patient with cancer. Besides, our semantic structure has allowed representing the aggregated disease timeline of a patient group.

The semantic infrastructure has also permitted the generation of graphical representations of the stored knowledge in the cancer registry with the generation of customizable dashboards.

The presented platform is an example of the parallel development of ontologies and applications that take advantage of semantic web technologies in the medical field.

Acknowledgments

This project has been possible thanks to the Spanish Ministry of Science and Innovation and the FEDER program through grant TIN2010-21388-C02-02 and the Fundación Séneca through grant 15295/PI/10.

References

1. W3C, OWL2 web ontology language. <http://www.w3.org/TR/owl2-overview/> (accessed September 2014)
2. W3C, RDF Resource Description Framework. <http://www.w3.org/RDF/> (accessed: September 2014)
3. W3C, SPARQL Query Language for RDF. <http://www.w3.org/TR/rdf-sparql-query/> (accessed: September 2014)
4. Muir CS, Nectoux J. Role of the cancer registry. *Natl Cancer Inst Monogr.* 47: 3-6 (1977).
5. Jensen OM, Whelan S. Planning a cancer registry. In Jensen OM, Parkin DM, Maclennan R et al. eds: *Cancer Registration: Principles and Methods*. Lyon, France: IARC; 22-28 (1991).
6. Altmann, U., Frank RK., Ali GT., Volker H., Joachim D. GTDS—a tool for tumor registries to support shared patient care. In *Proceedings of the AMIA Annual Fall Symposium*, 512. American Medical Informatics Association. (1996).
7. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). <http://apps.who.int/classifications/icd10/browse> (accessed: September 2014).

8. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). <http://www.who.int/classifications/icd/adaptations/oncology/en/> (accessed: July 2014).
9. Sobin, L., Gospodarowicz, M., Wittekind, C.: TNM Classification of Malignant Tumors. ISBN: 978-1- 4443-3241-4; (2009).
10. SNOMED Clinical Terms (SNOMED CT). <http://www.ihtsdo.org/snomed-ct/> (accessed: September 2014).
11. The 2014 ICD-10-Procedure Coding System (ICD-10-PCS). <http://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-PCS.html> (accessed: July 2014).
12. Berners-Lee, T., Hendler, J., & Lassila, O.: The semantic web. *Scientific American*, 284, 34-43 (2001).
13. Studer, R., Benjamins, V. R., & Fensel, D.: Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25, 161-197 (1998).
14. European Commission. Semantic interoperability for better health and safer healthcare. Deployment and research roadmap for europe. ISBN-13: 978-92- 79-11139-6; (2009).
15. SemanticHealthNet. <http://www.semantichealthnet.eu/> (accessed: September 2014)
16. Brochhausen, M., Spear, AD., Cocos, C., et al.: The ACGT Master Ontology and its application – Towards an ontology-driven cancer research and management system. *Journal of Biomedical Informatics*, 44, 8-25 (2011).
17. Min, H., Manion, FJ., Goralczyk, E., Wong, Y., Ross, E., Beck, JR.: Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics*, 42, 1035-1045 (2009).
18. Abidi S.: Ontology-Based modeling of Breast Cancer follow-up Clinical Practice Guideline for providing Clinical Decision Support. *Computer-Based medical systems, CBMS'07*, 542-547 (2007).
19. Rector AL, Brandt S, Drummond N, Horridge M, Puleston C, Stevens R. Engineering use cases for modular development of ontologies in OWL. *Applied Ontology* 2012, 7(2):113-132 (2002).
20. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Lewis, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11), 1251-1255 (2007).
21. Miñarro-Giménez. JA., Fernández-Breis JT.: Ontology-driven Method for Integrating Biomedical Repositories, *Lecture Notes in Computer Science*. 7023, 473-482 (2011).
22. Rea, S., Pathak, J., Savova, G., Oniki, T., Westberg, L., Beebe, C., Tao, C., Parker, C., Haug, P., Huff, S., Chute, C.: Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPN project, *Journal of Biomedical Informatics*. 45, 736-771 (2012).
23. Esteban-Gil, A., Garcia-Sanchez, F., Valencia-Garcia, R., Fernandez-Breis, JT.: SocialBROKER: A collaborative social space for gathering semantically-enhanced financial information, *Expert Systems with Applications*. 39, 9715-9722 (2012).
24. Dorn, HF. The Use of Statistics in Cancer Control Programs. *American Journal of Public Health and the Nations Health* 39 (5_Pt_1): 602-606 (1949).
25. Neitlich, HW., Stephen L., Vincent J.O.. "Development of a regional computerized cancer registry and its impact on medical care". *Journal of Medical Systems* 7 (3): 251-255. (1983).
26. Yates JW, Chalmer B, McKegney FP. Evaluation of patients with advanced cancer using the Karnofsky performance status. *Cancer*. 45, 2220-2224 (1980).

27. ASA Physical Status Classification System. <https://www.asahq.org/clinical/physicalstatus.htm> (accessed: July 2014)
28. Cardillo, E., Tamilin, A., Eccher, C., Serafini, L.: ICD-10 Ontology, <https://dkm.fbk.eu/technologies/icd-10-ontology> (accessed: July 2014)
29. Shiki, Naomi, Yuko Ohno, Ayumi Fujii, Taizo Murata und Yasushi Matsumura. 2008. Unified Modeling Language (UML) for hospital-based cancer registration processes. *Asian Pacific Journal of Cancer Prevention* 9:789-796 (2008).
30. Caldarella, A., Gianni A., Catia A., Crocetti, E., Di Costanzo, F., et al. Feasibility of evaluating quality cancer care using registry data and electronic health records: a population-based study". *International Journal for Quality in Health Care*: mzs020 (2012).
31. Tafazzoli, A. G., Altmann U., Wächter, W., Katz, FR., Hölzer, S., Dudeck, J. Integrated knowledge-based functions in a hospital cancer registry-specific requirements for routine applicability. *Proceedings of the AMIA Symposium*: 410-414. (1999).
32. Boyle, EI., et al.: GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18), 3710-3715 (2004).