Developing a Data Infrastructure for Enabling Breast Cancer Women to BOUNCE Back

Haridimos Kondylakis
Institute of Computer Science, FORTH
Heraklion, Greece
kondylak@ics.forth.gr

Lefteris Koumakis
Institute of Computer Science, FORTH
Heraklion, Greece
koumakis@ics.forth.gr

Dimitrios G. Katehakis
Institute of Computer Science, FORTH
Heraklion, Greece
katehaki@ics.forth.gr

Angelina Kouroubali
Institute of Computer Science, FORTH
Heraklion, Greece
kouroub@ics.forth.gr

Kostas Marias
Institute of Computer Science, FORTH
Heraklion, Greece
kmarias@ics.forth.gr

Manolis Tsiknakis
Institute of Computer Science, FORTH
Heraklion, Greece
tsiknaki@ics.forth.gr

Panagiotis G. Simos
Medical School, University of Crete
Heraklion, Greece
akis.simos@gmail.com

Evangelos Karademas
Department of Psychology, University of Crete
Rethymno, Greece
karademas@uoc.gr

***Abstract*— Breast cancer is the most common cancer disease in women and is rapidly becoming a chronic illness due recent advances in treatment methods. As such, coping with cancer has become a major socio-economic challenge leading to an increasing need for predicting resilience of women to the variety of stressful experiences and practical challenges they face. In this paper, we present the data infrastructure developed for this purpose, demonstrating the various components that will contribute to the developing the resilience trajectory predictor. Special emphasis is given to the semantic tier, presenting the project solution already implemented for effectively collecting, ingesting, cleaning, modelling and processing data that will be used throughout the lifetime of the project.**

Keywords—Breast Cancer, Resilience, Semantic Infrastructure

I. Introduction

Breast cancer is the most common cancer in women world-wide accounting for 28% of the total cancer cases in the WHO European Region. In Europe, screening programmes with good compliance have achieved a reduced breast cancer mortality of at least 20% in women aged over 50 [1]. As survival rates increase, coping with breast cancer becomes a socio-economic challenge. Novel strategies are needed for understanding, predicting and increasing the resilience of women with breast cancer to all the stressful challenges and experiences and ensuring better and faster recovery through personalized interventions [2].

Coping with breast cancer increasingly becomes a major socio-economic challenge not least due to its constantly increasing incidence in the developing world. There is a growing need for novel strategies to improve understanding and capacity to predict resilience of women to the variety of stressful experiences and practical challenges with breast cancer. This is a necessary step toward efficient recovery through personalized interventions.

The Horizon 2020 project BOUNCE [3] uses techniques and expertise from modelling, medical, and social sciences to advance current knowledge on the dynamic nature of resilience as it relates to efficient recovery from breast cancer. BOUNCE takes into consideration clinical, cancer-related biological, lifestyle, and psychosocial parameters in order to predict individual resilience trajectories throughout the cancer continuum. Eventually the target is to increase resilience in breast cancer survivors and help them remain in the workforce and enjoy a better quality of life. BOUNCE develops and deploys advanced computational tools to validate indices of patients' capacity to bounce back during the highly stressful treatment and recovery period following diagnosis of breast cancer. Elements of a dynamic, predictive model of patient outcomes are incorporated in building a decision-support system to be used in routine clinical practice to provide physicians and other health professionals with concrete, personalized recommendations regarding optimal psychosocial support strategies. The process of successful adaptation to breast cancer and the various accompanying stressors can be conceptually defined as the person's resilience, marking the will to 'fight for life' and bounce back. When faced with such potentially life-threatening events each person engages coping strategies that can vary widely on their capacity to provide adaptive solutions and ensure optimal recovery with respect to the disease itself as well as to the patient's overall quality of life. BOUNCE engages multi-scale, patient-specific modeling approaches for delivering a personalised resilience model for breast cancer patients. The model is built by considering a number of biological, social, environmental, lifestyle, occupational, and socio-economic factors. The resilience trajectory predictor is a dynamic model aiming to anticipate the patient's ability to cope. As such, the model needs to be accompanied by other in silico technologies related to disease prognosis and therapeutic outcome for the individual patient. This will allow a more holistic approach in predicting both the patient's disease and resilience trajectories (since they interact with each other).

2372-9198/19/$31.00 ©2019 IEEE
DOI 10.1109/CBMS.2019.00134

In this paper, we focus on the technological infrastructure that is currently under development describing the various layers of the platform architecture. Then we focus on the data management layer presenting the decisions made by the consortium for the effective and efficient usage of the data that will be collected through the lifetime of the project.

The rest of the paper is structured as follows: Section 2 elaborates on the BOUNCE reference architecture comprising three tiers, and focuses in more detail on the semantic tier. Section 3 outlines architectures developed in related projects for cancer care, whereas Section 4 summarizes the present work and outlines future steps in developing and validating the BOUNCE architecture.

II. BOUNCE Reference Architecture & Methodology

BOUNCE aims to build an open architecture to maximize the benefits of combining technologies and data from different partners and organizations. The architecture is constructed based on an iterative incremental process of software development. Short iterations are important to keep quality under control by driving to a releasable state frequently, which prevents the project from collecting a large backlog of defect correction work. Refinements of the architecture will continue to take place during the whole lifetime of the project driven by the iterative feedbacks from all stakeholders. The general BOUNCE architecture is envisioned as a framework, which integrates several building blocks oriented to support/predict the resilience of women with breast cancer. The building blocks are organized in three tiers:

- **the applications tier** which contains components such for biological and medical modelling, rule based decision support system and psycho-emotional models;
- **the security tier**, a privacy framework able to handle user privacy and data security;
- **the semantic tier** based on hybrid architecture which contains data extraction, transformation and serving the applications tier;

Every component/service designed within BOUNCE can be mapped to one of these tiers (or span over multiple ones). Modules and components designed and built within the project seamlessly operate through well-defined interfaces on different levels (i.e. interoperability on the level of IT-protocol, data format, information content, etc.).

Fig. 1 shows the main components of the BOUNCE architecture and their interconnections. As shown, patient data are collected via a mobile application called 'Noona' [4] and stored within the individual hospitals' premises. Data that are collected include answers to several questionnaires (psychological instruments, medication and symptom rating scales). Healthcare professionals, such as physicians, research nurses and psychologists, are able to access the Noona tool and provide clinical assessment data for the patients. The data anonymizer module anonymizes the collected data and pushes it into the BOUNCE data lake (via the Data API). Then the semantic tier modules clean and harmonize the data that is accessible with the Data API.

Health professionals have access to the decision support tool and in the first stages of the development to the temporary research-supporting tool. The Temporary Research Supporting Tool is the short-term internal project tool to facilitate data exploration and visualization of data and patient-completed self-report scales. The Decision Support System is the final online tool that will produce (a) an overall "resilience predictor" score, and (b) scores for specific psychological variables that could be used by mental health professionals in adapting psychological interventions to individual patient needs. The Decision Support System will replace the Temporary Research Tool and will be able to retrieve data, apply prediction model(s) and view/ store the results of the analysis. Furthermore, model developers will be able to upload models to the models repository.

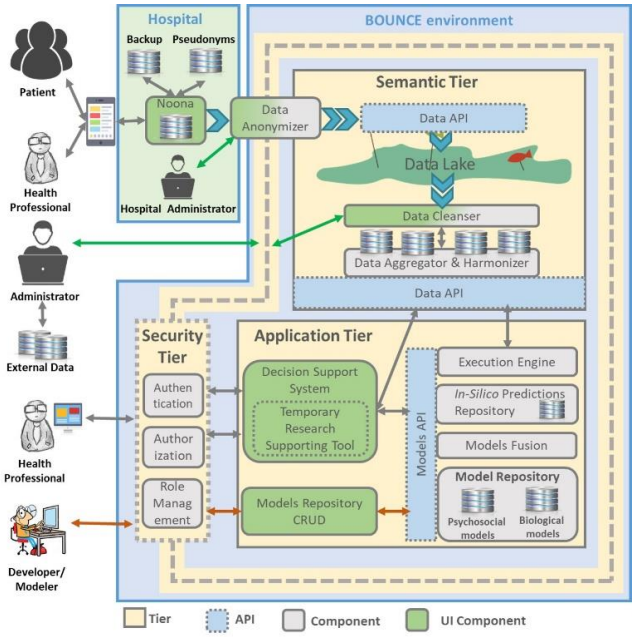

Fig. 1. BOUNCE reference architecture.

The main challenge of the BOUNCE architecture is the interoperability of systems, tools and services that are available to the users of the environment with the ultimate goal of secure, transparent, and unobtrusive sharing of data and functionality. Most of the currently identified scenarios in the project focus on data access and processing of data but there are also tasks involving computational jobs and visualization. In order to fulfil the requirements imposed by these scenarios a scalable and flexible environment is needed and the following technologies, which have gained momentum in the recent years, will be adopted: (i) Web/ REST Services technologies and (ii) Semantic Web technologies.

Below we briefly describe the application tier and the security tier and we present additional details about the semantic tier that is already available.

A. Applications Tier

The product of BOUNCE will be a set of clinically validated algorithms designed to predict each of the three main resilience

outcomes assessed in the prospective, multicentre clinical pilot (clinical relapse, physical/psychological well-being, functionality). Using statistical and machine learning methodologies to mine clinical [5], molecular [6] and psychological data [7] in conjunction with mechanistic models [8], BOUNCE aims to provide prediction tools for the targeted trajectories. These tools will deliver reliable indications to health professionals, enable the translation of the conclusions of the resilience trajectory models and ultimately enhance the capacity of individual breast cancer patients to efficiently adapt and resume a full life.

An innovative feature of the BOUNCE project relies on a clinically validated fusion strategy, which is expected to optimize overall prediction model accuracy. In addition, since the output of the models concerns multiple types of trajectories, weighted average schemes (e.g. linear combination, Bayesian combination, Bayesian Model Averaging (BMA), fuzzy rules, evidential reasoning technique) as well as supervised learning methods will be applied. The fusion approach will also consider the different reliability factors of model outcomes. Model fusion will take place at the decision level using supervised learning methods. Furthermore, the interaction between the different trajectories will take place through specific parameters. The value of such a parameter generated by one trajectory will modulate the course of the other trajectory.

B. *Security by Design*

BOUNCE aims to strictly abide by the EU General Data Protection Directive [9], as well as any subsequent possible improvement EU defines, fully adopting the "privacy by design" concept. Following the INTEGRATE [10] and the iManageCancer paradigm [11], security protection of personal and sensitive information in BOUNCE has been addressed through a two-fold procedure, which includes Data Access Control, and Security of data across their whole lifecycle, from storage, to transit and to use. It should be noted however that the holistic security approach is not a standalone component, but rather a set of technologies and tools that are utilised within the components of the BOUNCE platform in order to enable cross-platform security.

C. *Semantic Tier*

The semantic tier of BOUNCE provides a unified, homogenized view over the underlying data sources. A detailed view of the Semantic tier is shown in Fig. 2 comprising the following components: the data lake, the data cleaning tools, the cleaned databases, the semantic integration tool, the BOUNCE ontology and the data access APIs.

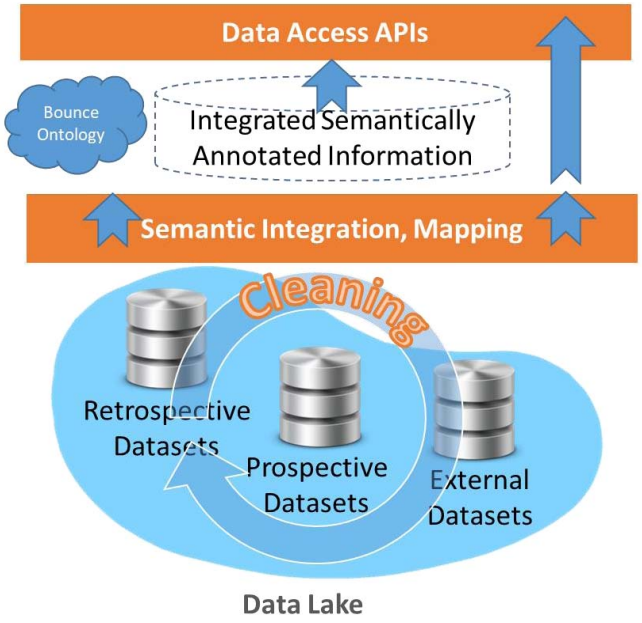

Fig. 2. The Semantic Tier.

Data Lake: At the bottom layer, all inserted data are staged in a repository with multiple databases forming a data lake. Data available already include the following anonymized retrospective datasets:

- *Data from European Institute of Oncology (IEO), Milan*: Comprises multiple datasets from psycho-oncology studies including biomedical, psychosocial, functional, demographic and psychological data.
- *Data from Rabin Medical Center and Shaare Zedek Medical Center – under the coordination of The Hebrew University of Jerusalem (HUJI), Israel*: The dataset includes a sample of 198 women after breast cancer including sociodemographic data, physiological, psychosocial self-reported questionnaires and life style data.
- *Data from Helsinki University Hospital (HUS), Finland*: This dataset includes clinical, breast and treatment data, sociodemographics, physical performance and activity, survival data and psychosocial self-report questionnaires.
- *Data from Champalimaud Foundation (CHAMP), Portugal*: This dataset includes information from an existing cohort of 110 breast cancer patients focusing on biological, socio-demographic, functional and psychological variables that could influence resilience processes. Biological variables include cancer type, treatment characteristics, and medical outcomes. Socio-demographic data includes sociological (e.g. marital status) and demographic information (e.g. age). Psychological variables refer to emotional, cognitive and relational aspects of an individual.

The following external datasets were also considered as potential sources of clinical information and incorporated into the BOUNCE data lake in their original format.

- *The Breast dataset* [1] is a comprehensive dataset that contains nearly all the PLCO study data available for breast cancer incidence and mortality analyses. For many women the trial documents multiple breast cancers, however, this file only has data on the earliest breast cancer diagnosed in the trial. The dataset contains one record for each of the approximately 78,000 women in the PLCO trial.
- This dataset contains the clinical and MRI data from *the ISPY* [2] *clinical trial* of patients with breast cancer. It includes various clinical and outcome data for around 180 women with breast cancer.

Besides having those datasets in their original form, there are also cleaned copies of the data in the data lake. Cleaning included missing value handling, identification of erroneous records etc.

Besides those data, prospective longitudinal data will be collected in several measurement waves from the four hospitals participating in the pilots. The data will be collected using the Noona tool and include the following information:

- Socio-demographic and lifestyle data such as age, education level, employment status, diet, exercise etc.;
- Clinical variables such as medical, treatment, patient care data and laboratory tests;
- Psychosocial measures captured from standardized psychological instruments such as the Ten item Personality measure, PTSD Checklist, Connor Davidson Resilience Scale, Family Resilience Questionnaire, NCCN Distress Thermometer etc.

All prospective datasets will be collected by the individual hospitals, who are the data owners and then they will be anonymized locally at the hospitals. Then, the data will be pushed into the data lake where they will again be cleaned and staged. Detailed descriptions for all aforementioned datasets can be found online [12] [13].

Semantic Integration & Mapping: To fully exploit the information potential of these datasets, they should be homogenized, semantically uplifted and transparently accessed [14].

To achieve this integration, a semantic model, i.e. the BOUNCE Ontology [13], is currently under development with the purpose of effectively representing and modeling all data that will be collected and analyzed within the BOUNCE project. The BOUNCE ontology is actually extending the iManageCancer Semantic Core Ontology [11] with a novel module modelling psychological constructs. The iManageCancer Semantic Core Ontology contains 36 sub-ontologies integrated equivalence and subsumption mappings, whereas *the ontology of psychological constructs* is describing and interrelating psychological measures. An overview of the specific module can also be found online [13] whereas a small part is shown in Fig. 3, depicting factors based on which the quality of life of a patient can be assessed.

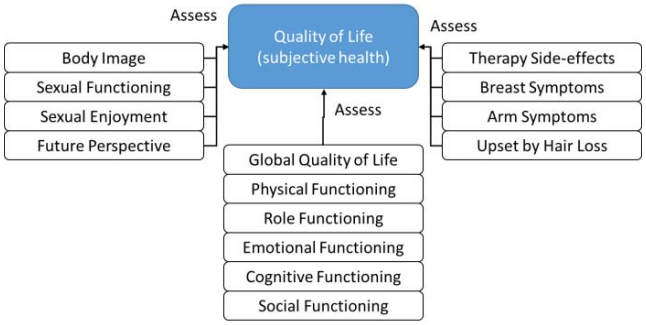

Fig. 3. Data Integration & Homogenization through an ontology

Having a semantic model available, the next step in order to homogenize and integrate the various datasets is to define the necessary mappings between the individual schemata and the semantic model [15] [16]. The process is a time-consuming and iterative process as trying to produce the mappings the semantic model might have to be updated, which might also lead to updating individual mappings. Those mappings provide declarative correspondences between the data columns and specific ontology terms. This iterative process is shown in Fig. 4.

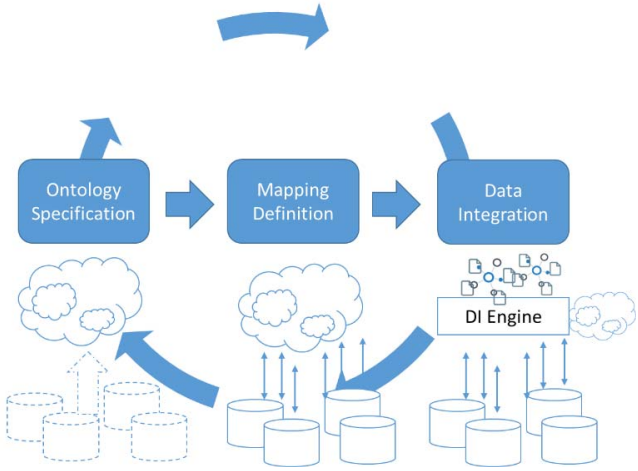

Fig. 4. Data Integration & Homogenization through an ontology

When those mappings are available, the data integration engine can either transform, semantically uplift and store them in a central data warehouse or enable federated access to the data where they reside. The former approach has benefits in terms of efficiency but has to deal with outdated data, whereas the latter enables access always to the latest snapshot of the data but suffers from efficiency. In our case, for the underlying data integration engine, we adopt the Ontop system [17]. This system accepts R2RML mapping rules and is able to perform both real time integration of the various sources and also to export the data as triples to be further served by a triple store.

Currently, we use the real time integration as the limited size of the current data allow efficient federated access. However we also design and implement a central warehouse where all data

[1] https://biometry.nci.nih.gov/cdas/datasets/plco/19/

[2] https://data.world/julio/ispy-1-trial

transformed into triples, will be served through the APIs via LAWA, a Spark based query execution engine [18].

Data Access APIs: All data, both integrated and those residing in the data lake are accessible through RESTfull APIs. As data access needs to be regulated this is ensured by an authentication and authorization server which is contacted prior to answering specific data requests, ensuring compliance with all necessary security and confidentiality requirements established by the security layer of the BOUNCE project.

III. Related Data Architectures

Similar architectures have been designed by other recent EU funded projects like iManageCancer, DESIREE and eSMART on the cancer domain.

Within iManageCancer [19], for example again a data lake is constructed, storing all external and internal datasets and a semantic model, i.e. the IMC Semantic Core ontology is used to effectively model all required terms. Then the *exelixis* data integration system [20] [21] [22], uses established mappings to integrate, in real-time, selected information out of the available sources and to publish offline a semantic repository. Although the BOUNCE data infrastructure uses a similar architecture, it allows both real time and offline data integration and progresses in modelling, storing and processing psychological terms as well.

ESMART, the "Electronic Symptom Management System Remote Technology" [23] demonstrates the effects of a real-time, mobile phone based, remote patient monitoring intervention on key patient outcomes, and delivery of care provided to people with cancer during and after chemotherapy. However, the data scheme is predetermined, and the information sources used are limited.

Desiree [24] is a web-based software ecosystem for personalized, collaborative and multidisciplinary management of primary breast cancer by specialized breast units, from diagnosis to therapy and follow-up. For managing the available data the project develops a novel complex Digital Breast Cancer Patient (DBPC) model, which incorporates information on clinical history and therapeutic procedure. However, beyond this model no further integration mechanisms are offered. We believe that the capability to integrate in real time various sources of information is essential in such a complex disease as cancer where multiple sources of information are required for optimal planning.

IV. Conclusions

Work presented focuses on estimating and improving resilience of women to the variety of stressful experiences and practical challenges related to breast cancer, which is a necessary step towards efficient recovery through personalized interventions. The overreaching goal of BOUNCE is to incorporate elements of a dynamic, predictive model of patient outcomes in building a decision-support system, to be used in routine clinical practice and provide health professionals with concrete, personalized recommendations [25] regarding optimal psychosocial support strategies for breast cancer women. Work takes into account clinical, cancer-related biological, lifestyle, and psychosocial parameters in order to predict individual resilience trajectories throughout the cancer continuum, and eventually increase resilience in breast cancer survivors while helping them to remain in the workforce and enjoy a better quality of life.

To achieve the aforementioned goal effective and efficient access to homogenized and integrated data is of utmost importance. So far the project has successfully achieved the implementation of the semantic tier, including a first version of a novel ontological module for modeling psychological constructs, which has been subsequently used to generate mappings to the data sources and enable effective and efficient data integration. APIs implemented on top allow regulated access to the available, homogenized information.

The next steps of the project include the deployment of application tier components and their integration with the semantic tier in order to be able to query the available data. Patient enrollment to the prospective clinical pilot has already begun by the four participating hospitals and data will be received in batches, one every three months according to the project protocol. The expected number of patients to be recruited is 660.

These tools delivered by the BOUNCE project offer a unique opportunity to advance clinical research and shed light on pathways to efficient recovery and optimal quality of life through validated assessment of risk factors and preventive strategies which can be readily incorporated into routine practice of clinical oncologists and mental health professionals involved in supporting breast cancer patients.

Acknowledgment

Work presented in the paper is part of the BOUNCE project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme. Any opinions, results, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of BOUNCE or the European Commission.

References

[1] WHO, Cancer Key Statistics, Available Online: https://www.who.int/cancer/resources/keyfacts/en/ (visited January 2019).

[2] H. Kondylakis, et al. "Digital patient: Personalized and translational data management through the MyHealthAvatar EU project." Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. IEEE, 2015.

[3] BOUNCE project website, Available Online: https://www.bounce-project.eu/ (May 2019).

[4] Noona website, Available Online: http://www.noona.com/ (May 2019).

[5] G. Potamias, L. Koumakis, and V. Moustakis, "Mining XML Clinical Data: the HealthObs System.," Ingénierie des systèmes d'information,. 10(1), pp. 59–79, 2005.

[6] L. Koumakis, V. Moustakis, M. Zervakis, D. Kafetzopoulos, and G. Potamias, "Coupling regulatory networks and microarays: Revealing molecular regulations of breast cancer treatment responses," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7297, pp. 239–246, 2012.

[7] E. Kazantzaki et al., "Psycho-emotional tools for better treatment adherence and therapeutic outcomes for cancer patients," in Studies in Health Technology and Informatics, 224, pp. 129–134, 2016.

[8] A. Bucur et al., “Workflow-driven clinical decision support for personalized oncology,” BMC Med. Inform. Decis. Mak., 2016.

[9] The EU General Data Protection Regulation (GDPR), Available online: https://eugdpr.org/ (May 2019).

[10] H. Kondylakis, et al. "Donor’s support tool: Enabling informed secondary use of patient’s biomaterial and personal data." International journal of medical informatics 97, pp. 282-292, 2017.

[11] H. Kondylakis et al., “iManageCancer: Developing a platform for Empowering patients and strengthening self-management in cancer diseases,” in , 30th IEEE International Symposium on Computer-Based Medi-cal Systems - IEEE CBMS, 2017.

[12] The BOUNCE Consortium, “D3.1 Identification of Internal and External Data Sources and Registries”, July 2018.

[13] The BOUNCE Consortium, “D3.2 Initial Semantic Model”, November 2018.

[14] L. Martín, et al. "Ontology Based Integration of Distributed and Heterogeneous Data Sources in ACGT." HEALTHINF (1). 2008.

[15] Y. Marketakis, et al. "X3ML mapping framework for information integration in cultural heritage and beyond." International Journal on Digital Libraries, 18(4), pp. 301-319, 2017.

[16] N. Minadakis, et al. “X3ML Framework: An Effective Suite for Supporting Data Mappings”. In EMF-CRM@ TPDL, pp. 1-12, 2015.

[17] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, “Ontop: Answering SPARQL queries over relational databases”. Semantic Web, 8(3), pp. 471-487, 2017.

[18] I. , LAWA: Locality Aware Partitioning for Efficient Query Answering over RDF Data, Master Thesis, University of Crete, 2019.

[19] H. Kondylakis, L. Koumakis, M. Tsiknakis, K. Marias, “Implementing a data management infrastructure for big healthcare data”. BHI, pp.361-364, 2018

[20] H. Kondylakis, D. Plexousakis. "Ontology evolution: assisting query migration." In International Conference on Conceptual Modeling, pp. 331-344. Springer, Berlin, Heidelberg, 2012.

[21] H. Kondylakis, D. Plexousakis. "Ontology evolution without tears." Web semantics: science, services and agents on the world wide web, 19, 42-58, 2013.

[22] H. Kondylakis, D. Plexousakis. "Ontology evolution: assisting query migration." In International Conference on Conceptual Modeling, pp. 331-344. Springer, Berlin, Heidelberg, 2012.

[23] F.A. Patricia, et al. "The eSMART Project: real time symptom management in the oncology setting." Cancer Professional, pp. 22-24, 2016.

[24] L. Nekane, et al. "DESIREE-a web-based software ecosystem for the personalized, collaborative and multidisciplinary management of primary breast cancer." 20th International Conference on e-Health Networking, Applications and Services (Healthcom). IEEE, 2018.

[25] H. Kondylakis, et al.. "Patient Empowerment through Personal Medical Recommendations." MedInfo 216, pp. 1117, 2015.