Report on the First International Workshop on Semantic Web Technologies for Health Data Management (SWH 2018)

Haridimos Kondylakis
ICS-FORTH
Heraklion, Greece
kondylak@ics.forth.gr

Kostas Stefanidis
Tampere University
Tampere, Finland
kostas.stefanidis@uta.fi

Praveen Rao
University of Missouri-Kansas
Kansas City, USA
raopr@umkc.edu

ABSTRACT

Better information management is the key to a more intelligent health and social system. To this direction, many challenges must be first overcome, enabling seamless, effective and efficient access to various health data sets and novel methods for exploiting the available information. The First International Workshop on Semantic Web Technologies for Health Data Management aimed at bringing together an interdisciplinary audience interested in the fields of semantic web, data management and health informatics to discuss the challenges in health-care data management and to propose novel and practical solutions for the next generation data-driven health-care systems. In this paper, we summarize the outcomes of the first instance of the workshop, and we present interesting conclusions and key messages.

1. INTRODUCTION

Key in achieving the vision of affordable, less intrusive and more personalized care, is the efficient and effective exploitation of health data. Ultimately this has the potential to increase the quality of life as well as to lower mortality. However, the lifelong patient data to be exploited for this purpose are complex, with hundreds of attributes per patient record, that will continually evolve as new types of calculations and analysis/assessment results are added to these records over time. In addition, data exist in many different formats, from textual documents and web tables to well-defined relational data and APIs. Furthermore, they pertain to ambiguous semantics and quality standards resulted from different collection processes across sites. The vast amount of data generated and collected comes in so many different streams and forms from physician notes, personal health records, images from patient scans, health conversations in social media, to continuous streaming information collected from wearables and other monitoring devices.

To this direction, semantic technologies can provide effective solutions for enabling interoperability and common language among health systems, and can lead to the disambiguation of the information through the adoption of various terminologies and ontologies available. On the other hand, cloud-based technologies and micro-services are the key for large-scale health systems deployment, data storage and analysis. The goal of this workshop is to bring together researchers cross-cutting the fields of semantic web, data management and health informatics to discuss the challenges in health-care data management and to propose novel and practical solutions for the next generation of data-driven health-care systems. Developing optimal frameworks for integrating, curating and sharing large volumes of Health Record data has the potential for a tremendous impact on health-care, enabling better outcomes at a lower cost.

Next, we summarize the outcomes of the first workshop instance held in conjunction with ISWC 2018 in Monterey, USA.

2. INVITED TALK

2.1 Benchmarking Big Linked Data: The case of the HOBBIT Project

Irini Fundulaki, in her keynote talk, discussed the results of the European H2020 HOBBIT (Holistic Benchmarking for Big Linked Data) Project. More specifically, she talked about the benchmarks developed in the context of the project, that target all the steps of the Big Data Value Chain. Special focus was given on the following two benchmarks: the link discovery benchmarks that aim at testing link discovery systems which address one of the most important problems of data integration, and the versioning benchmark that can be used to check the performance of systems that manage versions of linked datasets.

The HOBBIT project worked towards providing innovative benchmarks with the following characteristics:

- Realistic benchmarks: Benchmarks are commonly generated with synthetic data that reflect a single and specific domain. HOBBIT created mimicking algorithms to generate synthetic data from different domains.
- Universal benchmarking platform: HOBBIT developed a generic platform that is able to execute large scale benchmarks across the Linked Data lifecycle. The platform provides reference implementations, as well as dereferenceable results and automatic feedback to tools developers.
- Industry relevant Key Performance Indicators (KPIs): In addition to the classical KPIs developed over the last decades, HOBBIT collected relevant KPIs from industry to assess technologies based on the industrial needs.

Finally, Irini Fundulaki discussed what are the benchmarks that are of interest for health-care systems and applications, and provided some vision as to what are the lines of research that we should pursue in order to be able to have good quality benchmarks for such critical systems.

3. PAPER PRESENTATIONS

3.1 An Ontology-Driven Elderly People Home Mobilization Approach

Proposing exercise games to elderly is a challenging area of research. Age-related changes in balance, gait, strength, visual and hearing senses, memory and attention, and their deterioration over time make it difficult to assess individual status and adapt appropriately the corresponding recommendations. The authors in [5] propose an intelligent agent, that automatically and continuously adapts to the user profile, and provides corresponding incentives for mobilization at home. In order to construct the user profile, the agent incrementally builds a knowledge base capturing behavioral characteristics and movement sequences. The tracking of the users is realized by 3D sensors, which capture individual tracks throughout the day. Processing those tracks, information regarding *active time in room*, *active time of day*, *average gait velocity*, *average stand up time* and *average walking time* is calculated and stored in the knowledge base. The information in the knowledge base is modeled using an ontology. Instances of this ontology match the condition of the available rules for providing personalized recommendations. Those rules are in essence SPARQL queries, which propose personalized recommendations regarding games which include walking exercises or mind games. REST APIs implemented on top provide CRUD functionality on the available data and expose the movement sequences processed by other components of the agent. The evaluation performed confirm the effectiveness and efficiency of the approach.

3.2 Integrating clinical data from hospital databases

Research in various fields of medicine often requires the process and analysis of large amounts of possibly heterogeneous data that appear in different sources, like hospitals or scientific laboratories. By integrating such data, researchers extract new knowledge related to their field of study, that are not able to obtain when working with each data source independently. In general, the goal of data integration is to provide a uniform access over a set of data sources that have been created and stored autonomously [4, 3].

To fully exploit the integrated clinical data, it is important to be able to reveal the semantic relationships among them. For example, as stated in [6], to translate patients data describing dementia symptoms into effective Alzheimer's disease diagnosis, it is important that these data are related to additional patients information, such as genetic data, as well as to biological markers, such as proteins and electroencephalography. Specifically, the authors are interested in integrating clinical data related to the human brain. This work has been developed to meet the needs of the Medical Informatics Platform (MIP) of the Human Brain Project (HBP) that aims to develop technologies that enhance the scientific research related to human brain. Towards this effort, MIP provides a data integration mechanism to collect clinical data, such as Electronic Health Records (EHR) and imaging features stored in hospitals local databases.

3.3 Knowledge Engineering Framework to Quantify Dependencies between Epidemiological and Biomolecular Factors in Breast Cancer

The relationship between social determinants of health and chronic disease risks is crucial for the prevention of chronic diseases. Such associations are relatively easier to uncover for simple diseases, like obesity. But for complex diagnoses like cancer, a large number of factors contribute to the onset of the disease. Cancer Registries as the source of health data are used widely in epidemiological

research. Being collected by health professionals, they reduce research costs and embrace the whole population. However, the primary purpose of those sources is not being used for research, e.g., many times the structure of records is not appropriate in order to build an epidemiological model. Therefore, a data adjusting issue arises. To fit data from Cancer Registries to the epidemiological model, the authors create a knowledge engineering framework utilizing controlled vocabularies, using Bayesian Networks to quantify and predict factors that influence hormonal patterns of breast cancer, which can lead to better patient care.

3.4 The FairGRecs Dataset

FairGRecs is a synthetic dataset that can be used for evaluating and benchmarking methods [8] that produce recommendations related to health documents based on individual health records. Specifically, FairGRecs can create, via a fully parametrized API, synthetic patients profiles, containing the same characteristics that exist in a real medical database, including both information about health problems and also relevant documents. More specifically, [10] relies on the EMRBots dataset[1], which contains synthetic patients profiles, containing the same characteristics that exist in a real medical database, such as patients admission details, demographics, socioeconomic details, labs and medications, extending it with a document corpus and a rating dataset. By exploiting the FairGRecs dataset, interested users can create patients that have provided rankings for health documents. To link document contents with patients, the authors use the ICD10[2] ontology, namely the International Statistical Classification of Diseases and Related Health Problems, which is a standard medical classification list maintained by the World Health Organization. FairGRecs is fully parametrized, is offered via an API, and has been used already in [11].

3.5 Towards the Development of a National eHealth Interoperability Framework to Address Public Health Challenges in Greece

Large amounts of health data are daily generated and stored in regional health systems across Europe. Opening and reusing these data can be the key for improving healthcare efficiency and effectiveness. As such, the development of national interoperability frameworks (NIF) is essential and towards this direction the EU has announced guidelines for a European interoperability framework (EIF) [1], including 47 concrete recommendations for legal, organizational, semantic and technical interoperability. In Greece, healthcare is provided by the national health system with multiple of services already available such as ePrescription, eReferral for primary care, eConfirmation for insurance status verification, eReimbursement, eAppointments for doctors in the primary care etc. [7]. The Greek national health system has recognized the importance of implementing a NIF. The prerequisites for enabling data reuse include a well-defined process model, available and agreed terminology and reliable clinical content.

3.6 The Case for Designing Data-Intensive Cloud-Based Healthcare Applications

Cloud computing is a major source of revenue for companies like Amazon, Google, and Microsoft. Today, it is attracting a lot of interest among the healthcare community due to its benefits such a lower cost and ease of deployment. [2] argues for a microservices-based architecture for designing data-intensive cloud-based healthcare applications. A healthcare vendor may consider moving an existing software application by deploying them through virtual machines (VMs). However, VMs are heavyweight and are not suited for rapid deployment and recovery. Rethinking the design of the software application using microservices will provide the ability to quickly scale, be fault-tolerant, and provide high levels of security and availability. Microservices allow an application to be composed of loosely coupled services. Kubernetes[3] is a popular framework for orchestrating microservices. The authors propose a generic architecture for designing healthcare applications using microservices by decoupling storage, compute, and non-functional requirements such as availability, security, scalability, capacity planning, and others. Services for data analytics and machine learning can be incorporated using the software-as-a-service model.

3.7 A De-centralized Framework for Data Sharing, Ontology Matching and Distributed Analytics

The HarmonicSS platform [9] is a decentralized platform with the target to address all the aforementioned needs, tackling appropriately all ethical, legal and privacy issues for data sharing. The data sharing framework includes the data assessment and the data sharing management modules, ensuring that the framework respects all General Data Pro-

[1]http://www.emrbots.org
[2]http://www.icd10data.com/
[3]https://kubernetes.io

tection Regulation requirements for both the data providers and data processors. The clinical data are stored on a private cloud, specifically designed for each cohort, whereas a data curation module performs data cleaning. For data harmonization, an ontology has been defined and ontology matching is used for mapping the schema of each individual cohort to the ontology.

The proposed architecture is used to integrate and harmonize 7500 records out of 22 cohorts, including a variety of patients with primary Sjogren syndrome. A data processor who wishes to process cohort data has first to request access. Then the corresponding data providers can allow or deny data access. Data analytic services on top are then executed locally on the private cloud and the results are combined in a distributed learning fashion ensuring that the data never leave the clinical center.

4. WORKSHOP CONCLUSIONS

One important message was made clear by the workshop presentations and the participants: given the proliferation of health data and applications, there is a need to view and manage health data from different perspectives. A number of key observations and research directions emerged that we summarize below.

- Semantic technology can leverage behavioral characteristics into personalized recommendations. In this direction, user context and interactivity need to be emphasized.
- Personal health data can be leveraged for exploring the past and personalizing the user experience. Personal data exploration even in the health domain can take into account psychological and behavioral patterns to build novel exploration paradigms.
- System performance, in particular response time experienced by the user, remains a major challenge in the domain of health data management.
- Given the growing interest among the healthcare community to adopt cloud services for large-scale health data management, microservices come to the foreground, holding the promise for designing the next generation of data-intensive health-care applications.
- Researchers can employ different clinical terminologies (e.g., ICD10, SNOMED CT) for building knowledge/data management solutions for healthcare data based on the country of deployment.

This first instance of the Semantic Web Technologies for Health Data Management Workshop made clear that a lot of research work still needs to be done in the general area of semantic health data management. Given the growing interest in industry and academia, we are looking forward to the next instance of this workshop.

5. REFERENCES

[1] The new european interoperability framework. `https://ec.europa.eu/isa2/eif_en`. Accessed: 2018-10-30.

[2] S. Bhagavan, K. Alsultan, and P. Rao. The case for designing data-intensive cloud-based healthcare applications. In *SWH*, 2018.

[3] V. Christophides, V. Efthymiou, and K. Stefanidis. *Entity Resolution in the Web of Data*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, 2015.

[4] A. Doan, A. Y. Halevy, and Z. G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.

[5] S. Karagiorgou, D. Ntalaperas, G. Vafeiadis, D. Alexandrou, K. Perakis, D. Baltas, C. Amza, A. Wanka, H. Freitag, M. Blok, M. Kampel, V. de Rond, T. Münzer, and R. Planinc. An ontology-driven elderly people home mobilization approach. In *SWH*, 2018.

[6] K. Karozos, I. Spartalis, A. Tsikiridis, D. Trivela, and V. Vassalos. Integrating clinical data from hospital databases. In *SWH*, 2018.

[7] D. G. Katehakis, A. Kouroubali, and I. Fundulaki. Towards the development of a national ehealth interoperability framework to address public health challenges in greece. In *SWH*, 2018.

[8] H. Kondylakis, L. Koumakis, E. Kazantzaki, M. Chatzimina, M. Psaraki, K. Marias, and M. Tsiknakis. Patient empowerment through personal medical recommendations. In *MEDINFO 2015*, page 1117, 2015.

[9] V. C. Pezoulas, K. D. Kourou, T. P. Exarchos, V. Andronikou, T. A. Varvarigou, A. G. Tzioufas, S. de Vita, and D. I. Fotiadis. A de-centralized framework for data sharing, ontology matching and distributed analytics. In *SWH*, 2018.

[10] M. Stratigi, H. Kondylakis, and K. Stefanidis. The fairgrecs dataset: A dataset for producing health-related recommendations. In *SWH*, 2018.

[11] M. Stratigi, H. Kondylakis, and K. Stefanidis. Fairgrecs: Fair group recommendations by exploiting personal health information. In *DEXA*, 2018.