Digital Patient: Personalized and Translational Data Management through the MyHealthAvatar EU Project*

Haridimos Kondylakis, Emmanouil G. Spanakis, Stelios Sfakianakis, Vangelis Sakkalis, Manolis Tsiknakis, Kostas Marias, Xia Zhao, Hong Qing Yu, Feng Dong

Abstract— **The advancements in healthcare practice have brought to the fore the need for flexible access to health-related information and created an ever-growing demand for the design and the development of data management infrastructures for translational and personalized medicine. In this paper, we present the data management solution implemented for the MyHealthAvatar EU research project, a project that attempts to create a digital representation of a patient's health status. The platform is capable of aggregating several knowledge sources relevant for the provision of individualized personal services. To this end, state of the art technologies are exploited, such as *ontologies* to model all available information, *semantic integration* to enable data and query translation and a variety of *linking services* to allow connecting to external sources. All original information is stored in a *NoSQL database* for reasons of efficiency and fault tolerance. Then it is semantically uplifted through a *semantic warehouse* which enables efficient access to it. All different technologies are combined to create a novel web-based platform allowing seamless user interaction through *APIs* that support personalized, granular and secure access to the relevant information.**

I. Introduction

A recent report by the eHealth Task Force entitled "Redesigning health in Europe for 2020" [1] focuses on how to achieve a vision of affordable, less intrusive and more personalized care, ultimately, increasing the quality of life as well as lowering mortality. Such a vision depends on the application of ICT and the use of data and requires a radical redesign of health to meet these challenges. A main driver for change is currently taking place under the term "*liberate the data*". The secondary use of care data for research, quality assurance and patient safety is still rarely supported and the main barriers to this are the lack of interoperability, common standards and terminologies [2]. Large amounts of data currently sit in silos within health and social care systems. If these data are integrated and used effectively they could transform the way that care is provided.

The MyHealthAvatar (MHA) EU project [3] is an attempt for the digital representation of patient health status. The goal is to create a *"digital avatar"*, i.e. a graphical representation/manifestation of the user, acting as a mediator between the end-users and health related data collections. It is designed as a lifetime companion for individual citizens that will facilitate the collection, the access and the sustainability of health status information over the long-term. Among others, key questions that should be answered in this context is how to develop optimal frameworks for large-scale data-sharing, how to exploit and curate data from various Electronic and Patient Health Records, assembling them into ontological descriptions relevant to the practice of systems medicine and how to manage the problems of large scale medical data.

In this paper, we attempt to provide answers to the aforementioned questions by presenting a novel data management infrastructure. This infrastructure is capable of combining large-scale and multidimensional data that are semantically enriched and intrgrated to be further used for a variety of diverse use-cases. More specifically our contributions are the following:

- A modular ontology named *MHA Semantic Core Ontology* capable of modeling all health-related information.
- A scalable *NoSQL Data Repository* for storing all original information received from external sources.
- A novel *Data Translation Module* that uses mappings to semantically uplift and translate the data to be stored in a central *Semantic Warehouse*. These data can come either from the NoSQL data repository or from other external sources that are linked through mappings.
- A variety of *Linking Services* to external sources to enable interaction with them. Although these sources do not support direct linking through mappings, they provide standard interfaces for exporting data.
- A wide range of programmatic interfaces (*APIs*) allowing the granular and secure access to all relevant information in the platform.

The remaining of this paper is structured as follows: In Section 2 we present shortly the use-cases that the platform should cover and the different types of data that should be used. Then in Section 3 we demonstrate the building blocks of our data management solution. Section 4 reviews other related projects with similar goals and Section 5 summarizes and presents an outlook for further work.

II. Use-Case Requirements

In MHA two general categories of scenarios are investigated: a) system use cases, describing the

*This work was partially support by the iManageCancer (H2020-643529) and the MyHealthAvatar (FP7-600929) EU projects.

Haridimos Kondylakis, Emmanouil G. Spanakis, Stelios Sfakianakis, Vangelis Sakkalis, Kostas Marias are with Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH), (e-mail: {kondylak, spanakis, ssfak, sakkalis, kmarias} @ics.forth.gr).

Manolis Tsiknakis is with the Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH) and the Technical University of Crete (e-mail: tsiknaki@ics.forth.gr).

Xia.Zhao, Hong Qing Yu and Feng Dong are with the Department of Computer Science and Technology, University of Bedfordshire, Luton, UK (e-mail: {xia.zhao, hongqing.yu, feng.dong}@beds.ac.uk).

functionalities of the MHA system from the perspectives of both clinicians and citizens/patients and b) clinical use cases describing how to use the data from the MHA system in real clinical scenarios. After performing an extended requirement analysis, four distinct and diverse clinical use cases were selected for further implementation, demonstration and evaluation: Diabetes, Nephroblastoma Simulation Model and Clinical Trial, Personalized Congestive Heart Failure (CHF) risk analysis and Osteoarthritis. From the above we present important functional requirement of the last two use-cases:

Personalized CHF risk analysis:

- *Assist individualized* self-monitoring of patient's own health-status through a "*CHF Real-time patient monitoring*" and a "*CHF Risk Assessment*" service.
- Provide *risk analysis for personal risk monitoring* for developing a cardiovascular related episode in the future.
- *Provide comorbidities and drug interaction information* in both the treating physicians, but also the patient him/herself regarding negative drug interactions.
- *Create a monitoring tool* for Personalized CHF risk assessment using medical sensors together with mobile application and MHA's schematics layer.
- Link, through MHA, *with external clinical information systems* to acquire specific EHR patient related data.
- *Incorporate verified risk assessment* models for CHF.
- Create *individualized mobile apps* for easy access to the service and MHA platform.

Osteoarthritis:

- *Visual analytics* should be used to display aggregated lifestyle data aiming to easy interpretation by both citizens (patients and healthy) and medical professionals.
- *Data collection methods* to easily upload health data.
- *Personal Diary* managing patients/citizens' health status and behaviors, including diet, movement, environment, mood, smoking, symptoms etc.
- *Guided interventions for patients/citizens.*
- Provide the means to be able to also incorporate *genomic predisposition evaluation* for estimating the risk of developing osteoarthritis.

To support all aforementioned requirements an advanced data management infrastructure is required to enable real-time analysis of big data and the interconnection of all heterogeneous available information

III. Conceptual and Technical Architecture

The conceptual architecture of the data management platform is shown in Figure 1. In the bottom layer, external sources are pushing data to the original data repository by using a variety of linking services. In addition, there are sources that allow access to the available information (such as the Linked Life Data[1] or the DrugBank[2]) directly from the semantic integration module. The data are semantically linked and integrated using the aforementioned module and stored as triples at the Semantic Data Warehouse to be served. On top of these repositories various APIs allow the granular and secure access to the available data either directly from the original data repository or from the semantically integrated data warehouse. In the following sections we present in detail each one of the aforementioned components.

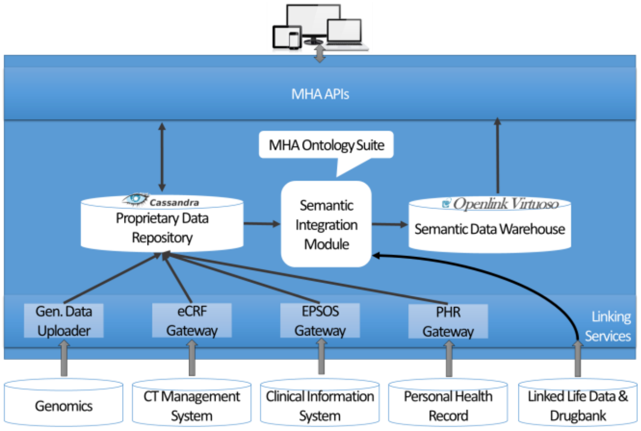

Figure 1. The architecture of the Data Management Approach

A. *MHA Semantic Core Ontology*

The MHA Semantic Core Ontology [4] is used as the virtual schema of all data stored within MHA. It is able to semantically describe the different types of data required and processed by the platform.

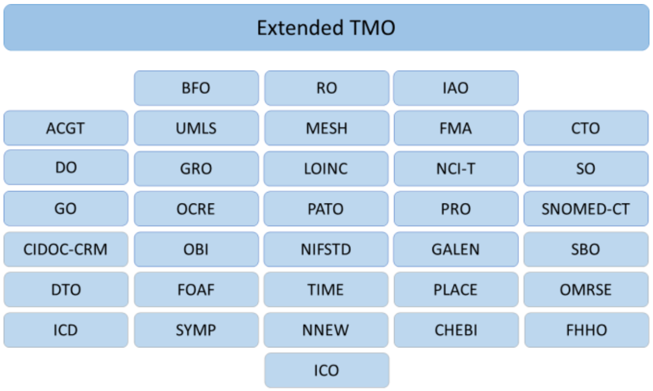

Figure 2. The modules of MHA Semantic Core Ontology[3]

[1] http://linkedlifedata.com/

[2] http://www.drugbank.ca/

[3] ACGT: ACGT Master Ontology, BFO: Basic Formal Ontology, CHEBI: Chemical Entities of Biological Interest, CIDOC-CRM: CIDOC Conceptual Reference Model, CTO: Clinical Trial Ontology, DO: Human Disease Ontology, DTO: Disease Treatment Ontology, FHHO: Family Health History Ontology, FMA: Foundation Model of Anatomy, FOAF: Friend of a Friend Ontology, GALEN: Galen Ontology, GO: Gene Ontology, GRO: Gene Regulation Ontology, IAO: Information Artifact Ontology, ICD: International Classification of Diseases, ICO: Informed Consent Ontology, LOINC: Logical Observation Identifier Names and Codes, MESH: Medical Subject Headings, NCI-T: NCI theraurus, NIFSTD: Neuroscience Information Framework Standardized ontology, NNEW: New Weather Ontology, OBI: Ontology for Biomedical Investigation, OCRE: Ontology for Clinical Research, OMRSE: Ontology of Medically Related

The development of the MHA Semantic Core Ontology was based on the following principles: a) *Reuse*: Exploit already established high quality ontologies; b) *Granularity*: A single ontological resource is not adequate to model the multi-faceted ecosystem of eHealth so multiple ontologies should be used; c) *Modularity*: Create a framework where different ontologies would be able to integrate many modules through mappings and equivalences between ontology terms.

In our case, after an initial evaluation, 34 sub-ontologies were selected and integrated through an extension of the Translational Medicine Ontology [5] (eTMO). The result is shown in Figure 2. The integration is achieved by introducing terms from these sub-ontologies to the eTMO ontology and via relations of equivalence (using owl:equivalentClass) and subsumption (rdfs:subClassof) from eTMO to the various ontology modules. These relations (~300) were manually identified and verified using the NCBO BioPortal[4] .

B. *NoSQL Data Repository*

The lifelong patients' data to be stored is complex, with hundreds of attributes per patient record that will continually evolve as new types of calculations and analysis/assessment results are added to the record over time. Apache Cassandra[5] is used to store these large-scale original data. Cassandra is an open-source, peer-to-peer and key value based store, where data are stored in key spaces. Cassandra has also built-in support for the Hadoop implementation of MapReduce [6], considered currently state of the art for real-time data analysis and has advanced replication functions. As such, Cassandra is used to store all original data and to provide input to applications that require big-data real-time analysis. In the MHA platform the Cassandra repository is an instantiation of a "data lake[6]" concept that stores the raw data supplied by the different information sources.

C. *Semantic Integration & Data Warehouse*

Although Cassandra is an excellent choice for storing and processing large amounts of data, the restrictions imposed on querying (e.g. lack of joins) prohibit the interconnection and the real integration of the data. However, the integrated information of the whole or parts of the patient profile is required for the provision of specific health care services. To achieve the semantic integration of the available data we use an extension of the *exelixis* [7] [8] platform, a novel data integration engine that has two main functionalities: a) It achieves query answering by accepting SPARQL queries that are rewritten to the data sources; b) it allows the transformation of data from original models to RDF/S data according to the MHA Ontology Suite. The platform allows the integration of a variety of data sources such as relational and NoSQL databases, XML, RDF/S and CSV documents, web services etc. Using the aforementioned platform we select which of the available data should be semantically linked and integrated by establishing the appropriate mappings. Then these data are queried, transformed into triples and loaded to the Semantic Warehouse where they are available for further reasoning and querying. A benefit of the approach is that we can recreate from scratch the resulting triples at any time. However for reasons of efficiency the *exelixis* transforms periodically only the newly inserted information by checking the timestamps of the data.

As already described, in order to select the information that is integrated, the proper mappings are established between parts of the source schemata and the MHA Semantic Core Ontology. However, the definition of those mappings is a time-consuming, labor-intensive and error-prone activity. To assist human in this difficult task, we created an innovative mapping workflow that manages the core processes needed to create, maintain and manage mapping relationships between different data sources over the long term, with high level of quality control. This novel workflow is named X3ML [4] and is composed of two main steps, shown in Figure 3: a) *Schema Matching*: The domain experts define a matching between the individual schemata and the ontology with the help of a graphical tool which is documented in a *schema matching definition file*. This file is human and machine readable and is the ultimate communication mean on the semantic correctness of the mapping; b) *Mapping definition*: In this step the actual mappings are generated based on the input of the previous step. In this step only the IT experts are involved and domain experts have no interest or knowledge about it.

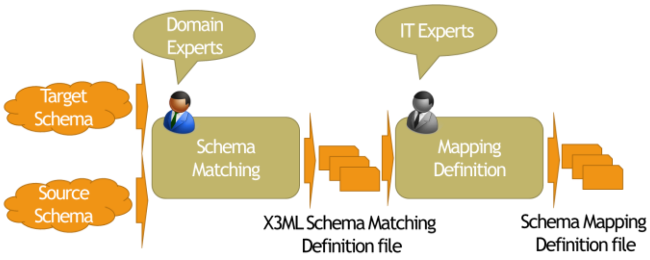

Figure 3. X3ML Mapping workflow

D. *Linking Services to External Sources*

Besides sources allowing the direct integration through mappings, the data management infrastructure supports the incorporation of parts of the patient's clinical and social history that are already stored and managed by third party systems.

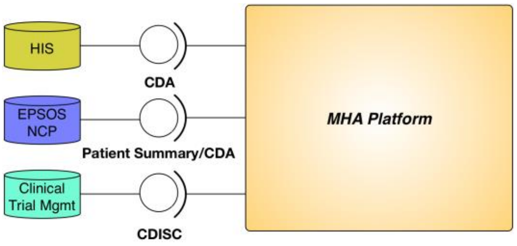

Figure 4. Linking MHA with external systems through well-defined interfaces

For this reason the proper mechanisms are in place for retrieving relevant user information from these external data sources. This "linking" mechanism is based on well-known

Social Entities, PATO: Phenotypic Quality Ontology, PLACE: Place Ontology, PRO: Protein Ontology, RO: Relation Ontology, SBO: Systems Biology Ontology, SNOMED-CT: SNOMED clinical terms, SO: Sequence Ontology, SYMP: Symptom Ontology, TIME: Time Ontology, UMLS: Unified Modeling Language System.

[4] http://bioportal.bioontology.org/

[5] http://cassandra.apache.org/

[6] http://martinfowler.com/bliki/DataLake.html

and established standard interfaces since they allow the building of generic ports and interfaces and the reuse of existing code bases. Figure 4 shows some notable examples for the realization of these links to external resources: Clinical data are retrieved from Hospital Information Systems (HIS) through the Clinical Document Architecture (CDA[7]) guidelines and set of specifications, clinical trial specific patient data are acquired using the Operational Data Model (ODM) of the Clinical Data Interchange Standards Consortium (CDISC[8]), whereas cross-border healthcare provisioning is supported by the adoption of epSOS[9] Patient Summary (again, based on CDA) interfaces. All these interfaces enable pushing data that are stored within the NoSQL data repository.

E. *APIs for Third-Party Access*

The APIs focus on providing functions to potential applications for accessing and managing data stored in the original data repository and the semantic data warehouse. The APIs are implemented in RESTful style and JSON is used as communication format. In addition oAuth[10] provides secure delegated access to the available resources.

IV. Related Work

Projects with similar goals for collecting, storing and accessing eHealth data were the eHealthMonitor[11] [9] and the INTEGRATE[12] projects whereas currently running projects on the area include the p-Medicine[13] and the EURECA[14] projects. In all those projects, the need for integrating disparate data sources has led to the adoption of multiple ontological resources as well. However, it is the first time that a separation is achieved between the original data collected from a variety of sources and a semantic repository to support the data needed integration and interconnection.

Besides European projects, there also exists a set of initiatives and personal health record systems concerned with the management of patient data. Some of them are Microsoft HealthVault, PatientsLikeMe, Indivo-X, Tolven, Dossia etc. (see [10] for a comparison). However, opposed to the work presented here they act as a static knowledge spaces – providing only storage and role based access to the knowledge resources.

To the best of our knowledge the proposed data management solution is the only one allowing the separation of the original and the semantically enhanced information, combining Ontologies, Semantics and NoSQL databases allowing a wide range of methods for pushing and retrieving information. The added value of our approach is that real-time analysis can directly be performed on the original data whereas semantic queries on integrated data can be efficiently answered using the Semantic Data Warehouse allowing a clear *separation of concerns* between the Cassandra and the Semantic Repository.

V. Discussion & Conclusion

Our architecture adopts a variation of the *command-query responsibility segregation* principle[15] where one uses a different model to update information than the model one is using to read. Although the mainstream approach people use for interacting with an information system is to treat it as a create, read, update and delete data-store, as the needs become more sophisticated state of the art approaches steadily move away from that model. In our case we rely on NoSQL technologies to store the original data due to their ability to handle enormous data sets and the “schema-less” nature, which makes, to a large extent, the import of new information to be frictionless. But their limitations in the flexibility of query mechanisms are a real barrier for any application that has not predetermined access use cases. The Semantic Warehouse component in the MHA platform fills these gaps by effectively providing a semantically enriched and search optimized index to the unstructured contents of the Cassandra repository. Therefore, our approach tries to offer best of both worlds: efficient persistence and availability of heterogeneous data, and semantic integration and searching of the “essence” of the ingested information.

A key next step is to evaluate the whole platform in a real-world context in the four clinical scenarios in three European countries (United Kingdom, Greece and Germany). Preliminary evaluation performed, provided initial evidences about the added value and the usability of our approach which will be extensively reported in a follow-up paper. Without a doubt data managements is an important area for healthcare that will only become more critical as healthcare delivery continues to grapple with current challenges.

References

[1] eHealth Task Force Report, “Redesigning health in Europe for 2020”, 2012.

[2] E.G. Spanakis, P. Lelis, F. Chiarugi, C. Chronaki, “R&D challenges in developing an ambient intelligence eHealth platform”, *EMBEC*, pp. 1727-1983, 2006.

[3] E.G. Spanakis, D. Kafetzopoulos, P. Yang, K. Marias, Z. Deng, M. Tsiknakis, V. Sakkalis, F. Dong, "MyHealthAvatar: Personalized and empowerment health services through Internet of Things technologies", *Mobihealth*, pp. 331-334, 2014.

[4] MyHealthAvatar Consortium: D4.2 Extension of the Semantic Core Ontology, February 2015.

[5] J.S. Luciano, S. Joanne S et al., “The Translational Medicine Ontology and Knowledge Base: Driving Personalized Medicine by Bridging the Gap between Bench and Bedside.” *Journal of Biomedical Semantics* 2.Suppl 2 (2011): S1. 2015.

[6] J. Dean, S. Ghemawat, “Map Reduce: simplified data processing on large clusters”, *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[7] H. Kondylakis, D. Plexousakis, “Exelixis: Evolving Ontology-Based Data Integration System”, *ACM SIGMOD*, pp. 1283-1286, 2011.

[8] H. Kondylakis, D. Plexousakis, “Ontology Evolution without Tears”, *Journal of Web Semantics*, 19, pp. 42-58, 2013.

[9] H. Kondylakis, D. Plexousakis, V. Hrgovcic, R. Woitsch, M. Premm, M. Schuele, “Agents, Models and Semantic Integration in support of Personal eHealth Knowledge Spaces”, *WISE*, pp. 496-511, 2014.

[10] I. Genitsaridi, H. Kondylakis, L. Koumakis, K. Marias, M. Tsiknakis, “Towards Intelligent Personal Health Record Systems: Review, Criteria and Extensions”, *Procedia Computer Science*, vol. 21, pp. 327-334, 2013.

[7] http://www.hl7.org/Special/committees/structure/index.cfm

[8] http://www.cdisc.org/

[9] http://www.epsos.eu/

[10] http://oauth.net/

[11] http://ehealthmonitor.eu/

[12] http://www.fp7-integrate.eu/

[13] http://www.p-medicine.eu/

[14] http://eurecaproject.eu/

[15] http://en.wikipedia.org/wiki/Command%E2%80%93query_separation