ASSIST: EMPLOYING INFERENCE AND SEMANTIC TECHNOLOGIES TO FACILITATE ASSOCIATION STUDIES ON CERVICAL CANCER

P. Mitkas*, C. Maramis*, A. Delopoulos*, A. Symeonidis*, S. Diplaris*, M. Falelakis*, F. Psomopoulos*, A. Batzios*, N. Maglaveras**, I. Lekka**, V. Koutkias**, T. Agorastos**, T. Mikos** and A. Tatsis**

* Dept. Electrical and Computer Engineering, Aristotle University, Thessaloniki, Greece
** Medical School, Aristotle University, Thessaloniki, Greece

chmaramis@mug.ee.auth.gr

Abstract: Advances in biomedical engineering have lately facilitated medical data acquisition, leading to increased availability of both genetic and phenotypic patient. Particularly, in the area of cervical cancer intensive research investigates the role of specific genetic and environmental factors in determining the persistence of the HPV virus – which is the primary causal factor of cervical cancer – and the subsequent progression of the disease. To this direction, genetic association studies constitute a widely used scientific approach for medical research. However, despite the increased data availability worldwide, individual studies are often inconclusive due to the physical and conceptual isolation of the medical centers that limit the pool of data actually available to each researcher. ASSIST, an EU-funded research project, aims at facilitating medical research on cervical cancer by tackling these data isolation issues. To accomplish that, it virtually unifies multiple patient record repositories, physically located at different sites and subsequently employs inferencing techniques on the unified medical knowledge to enable the execution of cervical cancer related association studies that comprise both genotypic and phenotypic study factors, allowing medical researchers to perform more complex and reliable association studies on larger, high-quality datasets.

Introduction

During the last years, advances in the area of biomedical engineering have allowed for more accurate and detailed data acquisition in the area of health care. This has led to an increase in the availability of patient data of both phenotypic and, most importantly, genotypic nature. Such data are nowadays produced in abundance by once laborious examinations and are being used for diagnosis and successful treatment but also in medical research. However, despite the increased data availability, scientific progress is hindered by the fact that each medical center operates in relative isolation, both physical and conceptual. This means that the produced data not only reside in physically isolated repositories but are also stored in different knowledge representation forms since there is no universally accepted knowledge representation prototype for medical data acquisition, data storage and labeling.

When it comes to the area of cervical cancer (CxCa), which is the second leading cause of cancer-related deaths after breast cancer for women between 20 and 39 years old [1] and one of the leading types of cancer affecting women worldwide, it has been proven that infection by the human papillomavirus (HPV) is necessary condition for the disease [2]. However, since HPV infection is highly unlikely to be the sole cause for developing cancer, intensive ongoing research investigates the role of specific genetic and environmental factors in determining the persistence of the HPV virus and subsequent progression of the disease [3]. To this direction, genetic association studies, i.e. studies that aim at detecting associations between one or more genetic variants and a trait (e.g. a disease) [4], constitute a widely used scientific approach in medical research. If a statistical correlation is observed between genotype and phenotype, an association between the variant and the trait is inferred [5]. The quality of the association studies conclusions heavily depends on the size of the available dataset. Low numbers of patient records lead to doubtful conclusions. This is the reason why several studies are often inconclusive, since the datasets employed are small and of poor quality due to the isolation issues mentioned in the previous paragraph.

ASSIST (Association Studies aSsisted by Inference and Semantic Technologies) is an EU-funded research project that aims at facilitating medical research on CxCa by tackling these isolation issues at both physical and semantic level. ASSIST overcomes the problem of physical isolation of data sources by supporting a 3-tier architecture; Researchers conducting association studies have access to all participating patient record repositories physically located at different medical research centers and/or hospitals through the single node of ASSIST. This would be sufficient if the multiple repositories had the same internal schema, the same detail of information and the same terminology. However, this is not the case in practice. The lack of a common representation standard for CxCa related data, the detail of relevant examination result, as well as

differences in the language used in each repository are only a few of the reasons that make the representation schemes of individual repositories incompatible to each other. This calls for what in ASSIST is referred to as *semantic unification*. In order to offer virtual unification of the repositories, ASSIST introduces its own, novel CxCa ontology and maps each one of the participating heterogeneous medical repositories on this prototype. The inference techniques used, combined with the integration of data from multiple sources, allow medical researchers to perform more complex and elaborate association studies on larger, high-quality datasets and consequently to draw more reliable conclusions regarding CxCa or related precancerous stages.

In the rest of this paper, the overall system architecture and some of the underlying design principles of ASSIST are presented.

Materials and Methods

In this section the architecture of the ASSIST system will be described briefly along with the technologies that were employed for its implementation. The ASSIST system comprises of three subsystems (Figure 1).

- The ASSIST Core
- The User Interface (UI)
- The Interfaces to Medical Archives (IMA)

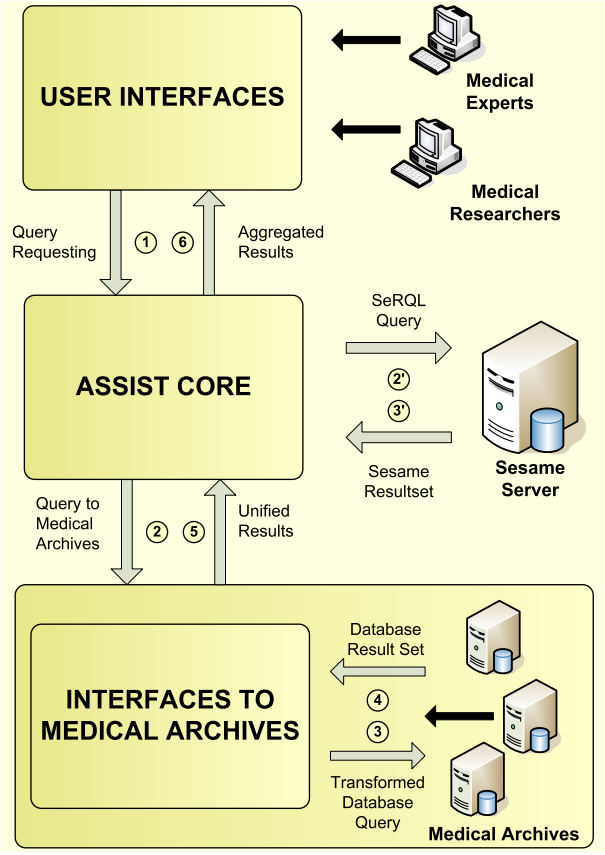

Figure 1: ASSIST framework overview and two alternative dataflow paths. Through medical archives: 1-6. Through Sesame Repository: 1,2',3',6.

The *ASSIST Core* is the most complex subsystem of ASSIST that undertakes the main semantic processing burden [6] and also the orchestration of the overall function of the system. In order to handle the heterogeneity of the data sources, a CxCa ontology (*Core Ontology*) [7], the first ever to model CxCa related concepts, has been developed. This ontology is the heart of the Core subsystem and also the common language that makes the communication between the subsystems of ASSIST possible. It is the knowledge representation form that has been chosen for the semantic integration of the participating data sources and the reason why transparency is possible at the UI side. All ASSIST subsystems are designed to be aware of this ontology. In addition and since the ontology is expected to evolve in the future, ASSIST subsystems are also designed to be adaptable to possible ontology changes. This specification ensures that the system will remain functional even in the case where the domain knowledge changes – and also in the case where other domains are included in the system.

Another component of the Core subsystem that is of great importance is the *Inference Engine*. This component is used to infer semantic medical entities from the syntactic values that are understood by the legacy medical systems of the participating medical centers. These inference functionalities are provided by Sesame [8], an of-the-shelf semantic framework Sesame was preferred over other similar tools on the basis of its efficiency in the manipulation of large amounts of semantic data and the simplicity of its accompanying query language SeRQL [9,10]. Apart from performing crisp inference on medical knowledge, the Inference Engine is also responsible for assigning validity degrees (i.e. fuzzy numbers) to the inferred entities through the use of fuzzy rules [11]. Using the inference techniques offered by ASSIST, a severity index of each patient record can be computed and the execution of association studies is made possible.

The *User Interface* subsystem enables transparent and advanced access to the data repositories incorporated in ASSIST. The UI is implemented as a user-friendly and simple yet powerful web-based interface that offers the end-user the ability to express queries and also visualizes the resulting patient and statistical data. The two main functionalities provided by the UI is mere patient data retrieval and complete association study execution.

The *Interfaces to Medical Archives* (IMA) subsystem is responsible to map the information contained in each legacy data repository to corresponding entities that are defined in the knowledge model of ASSIST (i.e. classes of the CxCa ontology). Currently, three pilot sites are participating in the project by offering an anonymized view of their CxCa related data. This mapping between ontology and database concepts is used to transform the data of the individual archives to the representation scheme centrally supported by ASSIST and essentially store these data in a central repository. The latter is in essence

the ABox of ASSIST Core ontology maintained as a Sesame repository.

The subsystems of ASSIST are separate software systems. The communication between them is carried out with the exchange of XML documents that are aware of the semantics of the CxCa ontology. Figure 2 below illustrates the overall architecture of ASSIST, detailed at the modules level, as well as the data exchange means between the subsystems.

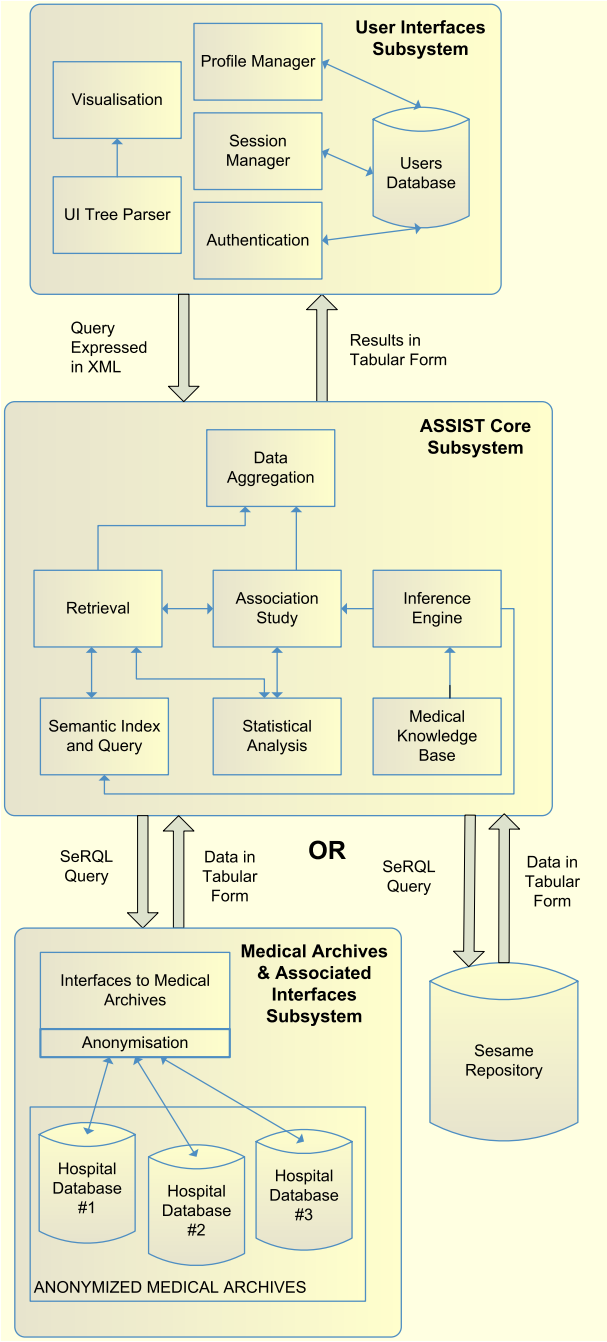

Figure 2: ASSIST overall system architecture (detailed at modules level) and data workflow.

Aside from the main ASSIST platform but within the context of the project, three auxiliary standalone software applications have been implemented. *CAT* (Custodix Anonymization Tool) is a powerful anonymization tool that is being used to provide anonymized views of each medical data repository before inclusion in the ASSIST system, in full compliance to the legal national and EU medical research requirements and code of practice. The *ASSIST M.K.B. Editor* is a java-based application implemented on top of the semantic Jena API [12] that is used as a lightweight ontology editor. As medical knowledge in the area of CxCa is about to change with the introduction of the HPV vaccine, such a tool is most certainly useful in updating the Core Ontology. Finally, the *UI-Ontology mapping tool* is a java-based application that maps entities of the Core Ontology to the medical entities to be displayed to the end-user by the UI. This way, a display tree is constructed that contains only the medical entities that are being used for query construction. The nodes of the tree are labeled with familiar medical terms instead of the more technical terms in the ontology.

Discussion

The ASSIST system will soon be available to medical researchers to perform association studies. Statistically significant CxCa related association studies will be possible by using the virtually combined dataset of three major European clinics in this domain. These three pilots participating in the ASSIST project will provide feedback before releasing the ASSIST framework to the wider community.

ASSIST has been designed with scalability and extendibility in mind. With regards to scalability, the system is by design capable of incorporating new medical repositories - when they become available - without any impacts on its functionality and performance. Adding new data sources means that the association studies are performed on larger, high-quality -owing to the semantic integration- datasets and consequently the drawn conclusions become more reliable and statistically significant. Moreover, the semantic technologies employed in the project allow for the extension of the system to include other medical domains apart from the domain of CxCa by either adding new ontologies representing these new domains or updating the existing Core Ontology.

Conclusions

Owing to the constantly increasing research interest in genetic association studies, several projects are currently under development, aiming to provide researchers with more powerful tools for investigating associations among various types of clinical and genetic data [13]. ASSIST is one of the most ambitious of these projects and this is partly due to its scalability and extensibility virtues. Upon successful completion, the ASSIST platform aspires to function as an IT tool enabling association studies linked at present with CxCa research by establishing a collaborative environment

and allowing any medical group active in this area to use its facilities and/or contribute their own data/results. However, in the future, the ASSIST system may be expanded in terms of its underlying knowledge model in order to facilitate genetic association studies for other diseases, e.g. colon cancer and cardiovascular diseases.

Acknowledgements

This work is supported in part by the IST-2004-027510 project, entitled "Association Studies Assisted by Inference and Semantic Technologies (ASSIST)", funded by the Commission of the European Community (CEC).

References

1. SH. Landis, T Murray, S Bolden, and PA. Wingo (1999) Cancer Statistics, CA: A Cancer Journal for Clinicians, (49)1:8-31
2. J.M.M. Walboomers et al. (1999) Human papillomavirus is a necessary cause of invasive cervical cancer worldwide, The Journal of Pathology, 189(1):12-19
3. T. Agorastos et al. (2005) Human papillomavirus testing for primary screening in women at low risk of developing cervical cancer. The Greek experience, Gynecologic Oncology, 96(3): 714-720
4. H. Cordell and D. Clayton (2005) Genetic association studies, The Lancet, 366(9491):1121-1131
5. J.N. Hirschhorn and M.J. Daly (2005) Genome-wide association studies for common diseases and complex traits, Nature reviews. Genetics, 6(2):95-108
6. G. Antoniou and F. van Harmelen (2004) A Semantic Web Primer, MIT Press
7. X.H. Wang, D.Q. Zhang, T. Gu and H.K. Pung (2004) Ontology based context modeling and reasoning using OWL, In Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 18-22
8. J. Broekstra, A. Kampman and F. van Harmelen (2002) Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema, In Proceedings of the First Internation Semantic Web Conference, 58-64
9. J. Broekstra and A. Kampman (2003) SeRQL: A Second Generation RDF Query Language, In Proceedings of SWAD-Europe Workshop on Semantic Web Storage and Retrieval
10. T. Furche et al. (2006) RDF Querying: Language Constructs and Evaluation Methods Compared, Reasoning Web, 4126:1-52
11. GJ Klir and B Yuan (1994) Fuzzy sets and fuzzy logic: theory and applications, Prentice-Hall
12. B. McBride (2002) Jena: A Semantic Web Toolkit, IEEE Internet Computing, 6(6):55-59
13. M. Dugas (2002) Impact of integrating clinical and genetic information, In Silico Biology, 2(3):383-391