Data & Knowledge Engineering 86 (2013) 160–178

Contents lists available at SciVerse ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Applying semantic technologies in cervical cancer research

Christos Maramis [a,b,*], Manolis Falelakis [a,b,c], Irini Lekka [d], Christos Diou [a,b], Pericles Mitkas [a,b], Anastasios Delopoulos [a,b]

[a] Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece
[b] Information Processing Laboratory, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece
[c] Department of Computing, Goldsmiths, University of London, UK
[d] Lab of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 12 January 2011
Received in revised form 18 February 2013
Accepted 18 February 2013
Available online 13 March 2013

Keywords:
Semantic unification
Heterogeneous medical data
Association studies
Cervical cancer ontology
Medical inference

ABSTRACT

In this paper we present a research system that follows a semantic approach to facilitate medical association studies in the area of cervical cancer. Our system, named ASSIST and developed as an EU research project, assists in cervical cancer research by unifying multiple patient record repositories, physically located in different medical centers or hospitals. Semantic modeling of medical data and rules for inferring domain-specific information allow the system to (i) homogenize the information contained in the isolated repositories by translating it into the terms of a unified semantic representation, (ii) extract diagnostic information not explicitly stored in the individual repositories, and (iii) automate the process of evaluating medical hypotheses by performing case–control association studies, which is the ultimate goal of the system.

1. Introduction

Cervical cancer (CxCa) is one of the most common cancers worldwide [1]. Infection by the human papillomavirus (HPV) is accepted as the central risk factor for CxCa [2]; however, it is unlikely to be the sole cause for developing cancer. Findings indicate that other factors in addition to HPV infection are likely to be important determinants in cervical carcinogenesis. Ongoing research includes investigating the role of specific genetic and environmental factors in determining HPV-persistence and subsequent progression of disease [3]. Genetic association studies [4] constitute a significant scientific approach that employs specific statistical analysis methods to investigate the association among genetic characteristics, environmental agents and virus characteristics so as to suggest pathogenetic mechanisms that will provide new markers of risk, diagnosis and prognosis, and possibly treatment. However, these association studies require large sets of patient phenotypic (e.g., virus characteristics, clinical tests, patient lifestyle) and genotypic data (e.g., polymorphisms of a gene) all provided in a structured format.

ASSIST (abbreviation for *A*ssociation *S*tudies Assisted by *I*nference and *S*emantic *T*echnologies) is a research project whose overall objective is to provide medical researchers of CxCa with an integrated environment that unifies multiple patient record repositories, physically located at different laboratories, clinics and/or hospitals. This environment enables researchers to combine phenotypic and genotypic data, utilize existing patient records from past research studies in several clinics, and, eventually, perform biomedical research in a low-cost and time-efficient way. Furthermore, ASSIST incorporates the statistical analysis tools that are required in order to automate the process of evaluating medical hypotheses of the type used in genetic case–control association studies.

---

[*] Corresponding author at: Information Processing Laboratory, Dept. Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124, Greece. Tel.: +30 2310 996 359; fax: +30 2310 996398.
E-mail address: chmaramis@mug.ee.auth.gr (C. Maramis).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.02.003

In order to provide the desired functionality, ASSIST is founded on semantic web technologies. More specifically, an ontological model is used to describe medical concepts, medical data types, relations between them and medical rules. When needed, this model can be updated by knowledge engineers – in collaboration with CxCa expert researchers – through a graphical user interface (GUI) application. Inference and Description Logics are utilized to homogenize multi-center patient data and infer diagnostic information, resulting in the construction of a unified patient record repository that adheres to the aforementioned ontological model. The system includes a GUI for its users to graphically form semantic queries and medical assumptions and also a powerful core that is able to respond to the queries and statistically validate the assumptions using data from the unified patient record repository.

The rest of this paper is structured as follows. Section 2 outlines the system specifications of ASSIST and then presents the conceptual architecture of the system emphasizing the semantic web technologies that have been used and their interactions. Section 3 presents two – real – usage scenarios of ASSIST and evaluates the usability and response speed of the system. The extensibility of ASSIST in multiple directions is discussed in Section 4. The state-of-the-art on related semantics-enabled biomedical systems is reviewed in Section 5. Finally, Section 6 draws the conclusion of this work and provides insights for the future of semantic web technologies in medical research.

2. System architecture

In this section we present the overall architecture of ASSIST from a knowledge engineering perspective, i.e., emphasizing the system's *semantic modules*, their role in the system, and their interactions; the term "semantic" refers to the ability of a module to process information defined through an ontology with RDF or OWL annotations. ASSIST consists of three main communicating subsystems, namely the *System Core*, the *User Interface* (*UI*), and the *Medical Repositories & Associated Interfaces*, which are complemented by two standalone applications, namely the *Medical Knowledge Base* (*MKB*) *Editor* and the *UI View Extractor*. An outline of ASSIST's architectural components is provided in Fig. 1.

Before proceeding with the detailed description of system modules, we present in the following section the "big picture" of ASSIST: The motivation behind ASSIST project, the project's objectives and associated system specifications, the issues that had to be addressed and how all these are mapped to the modules of the final system.

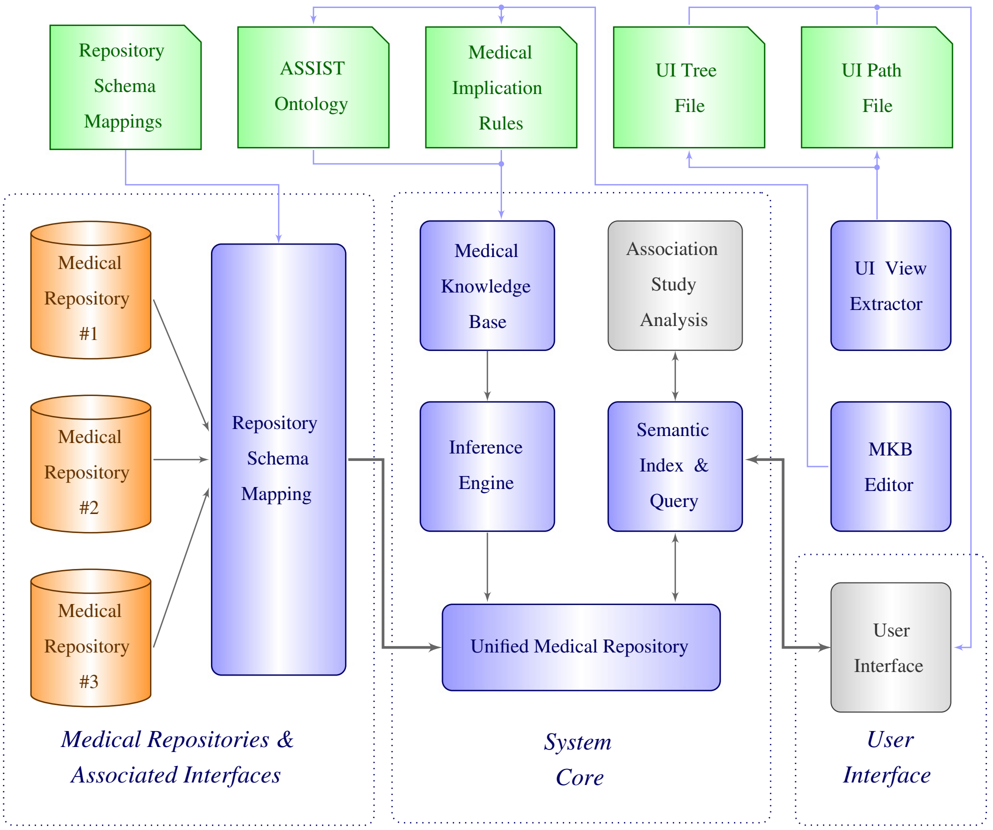

Fig. 1. Outline of ASSIST's architectural components (system modules and configuration files). Semantic modules are indicated in blue color. System configuration files are colored green.