MAPi DOCTORAL PROGRAM IN COMPUTER SCIENCE

2011/2012

Thesis Plan

Title:

A federated architecture for biomedical data integration

Candidate:

Luís A. Bastião Silva, bastiao@ua.pt

Supervisors:

Carlos Costa, carlos.costa@ua.pt

José Luís Oliveira, jlo@ua.pt

Introduction

During the last two decades, health centers have made significant investments in IT to integrate EHR (Electronic Health Records) and digital medical imaging systems in their daily workflow. PACS (Picture Archiving and Communication System) have been a key element in this development, proving solutions for the acquisition, distribution, storage and analysis of digital images in distributed environments. The DICOM standard architecture has also played a significant role in the exchange of structured medical imaging data and nowadays almost all medical imaging equipment manufacturers support it.

Medical researchers have also been using IT facilities to conduct cohort studies for specific population screening. Unfortunately, these studies have been gathered in isolated silos and there is no integrated view of the "whole" catalog that can take advantage of those results in a holist manner. On the other hand, a huge amount of data and scientific results are being generated by genomics-related research, and a great expectation exists about how to integrate and use these outcomes in clinical practice. The integration of different sources of knowledge promises the improvement of patient care, creates new guidelines for clinical decisions and for the development of new treatments, and gives insights towards the improvement of health services.

There are disperse databases with different information, for instance, healthcare medical records, genetic databases and molecular information. In the future, personal electronic health records (EHR) may also contain genetic information required for a fit treatment. The efficient storage and permanent availability of all clinical data are huge tasks, at least without requiring major upgrades and overhauls that significantly increase the total cost of ownership over time. The emergence of Cloud computing providers creates a great opportunity to tackle the costs of purchase hardware and software. However, despite of all the benefits of the cloud computing, it also brings new challenges, for instance, privacy and time delay to access to the data. Translational medicine is a recent research field that aims to combine genotype-to-phenotype results and put it available to create novel diseases' treatments and, generically, for the man kind benefit. Biomedical informatics is a key part of this process, by joining areas and methodologies that are crucial for transitional medicine, such as bioinformatics, medical imaging, clinical informatics, and public health informatics.

Previous work

In an attempt to create a unique access point for distributed medical information, we have been developing a set of tools using peer-to-peer, cloud storage and information retrieval technologies. One biggest problem associated to Cloud Computing is the lack of interoperability between providers, i.e. services are not interchangeable across cloud providers. Moreover, the current absence of interface normalization does not allow transparent migration of Cloud applications between providers. Our “Service Delivery Cloud Platform” (SDCP) allows applications to store information and communicate easier, using any cloud provider. With this platform developers do not care where the resources will be deployed and the applications will not be restricted to a specific cloud vendor. Over this

platform, a PACS archive over the Cloud was developed, supplying a bridge to the Cloud resources. Another tool is Dicoogle, www.dicoogle.com, an open source PACS that replaces traditional PACS databases by an indexing and retrieving engine. It has been used to support analysis of existent medical imaging repositories in some hospitals of Aveiro region retrieving data from studies that would otherwise be difficult to identify. Dicoogle communication layer is based on peer-to-peer technologies, allowing to search and retrieve disperse medical imaging data. It also supports a WAN relay service module that can be hosted in an Internet cloud computing service, namely the Google App Engine (GAE).

Objective

The objective of this project is to investigate new solutions for storing, retrieving and distributing patient-related information, for multiple medical institutions and research centers. This thesis proposal will focus the following aspects:

- Studying the field: identify the requirements for the federation of electronic patient records, cohort studies and genomic-related data, in a patient and population perspective. Special attention will be put on data integration, semantics, publishing and searching, de-identification and accreditation.
- State of art: perform a complete literature review, covering main authors and projects related to this issue.
- The question: identify better the key problem to be tackled in the following years of this doctoral program, defending its relevance and scientific impact, and describing the main strategy to conduct that research work.

Thesis Planning

Phase 1. Literature review (reading and writing)

The main goal of this task is to identify and retrieve source materials, both theoretical and empirical, which are appropriate for the topic, hypotheses, questions, and variables.

a) Some starting points:

- S. Madhavan, J. C. Zenklusen, Y. Kotliarov, H. Sahni, H. A. Fine, and K. Buetow, "Rembrandt: helping personalized medicine become a reality through integrative translational research," *Molecular Cancer Research,* vol. 7, p. 157, 2009.
- S. Szalma, V. Koka, T. Khasanova, and E. D. Perakslis, "Effective knowledge management in translational medicine," *Journal of translational medicine,* vol. 8, p. 68, 2010.
- C. A. Kulikowski and C. W. Kulikowski, "Biomedical and health informatics in translational medicine," *Methods Inf Med,* vol. 48, pp. 4-10, 2009.
- P. R. O. Payne, S. B. Johnson, J. B. Starren, H. H. Tilson, and D. Dowdy, "Breaking the translational barriers: the value of integrating biomedical informatics and translational research," *Journal of investigative medicine,* vol. 53, p. 192, 2005.

- E. D. Perakslis, J. Van Dam, and S. Szalma, "How informatics can potentiate precompetitive open-source collaboration to jump-start drug discovery and development," *Clinical Pharmacology & Therapeutics,* vol. 87, pp. 614-616, 2010.
- I. N. Sarkar, "Biomedical informatics and translational medicine," *Journal of translational medicine,* vol. 8, p. 22, 2010.

b) Review of scientific papers containing the following keywords:

- Translational research, Electronic Health Records,
- Translational medicine informatics

c) Some publications sources:

- Journal of Translational Medicine
- American Journal of Translational Research
- Journal of Biomedical Informatics
- International Journal of Medical Informatics
- Journal of the American Medical Informatics Association.
- IEEE Transactions on Information Technology in Biomedicine
- Several IEEE, ACM and Springer Conferences related to the subject Data Integration in the Life Sciences

Phase 2. Problem statement

Which issues or controversy need to be solved.

Phase 3. Plan of Attack

Plan approach to the problem in a systematic way.