Cloud-Based Data and Knowledge Management for Multi-Centre Biomedical Studies

Amalia Tsafara Christos Tryfonopoulos Spiros Skiadopoulos* Lefteris Zervakis*

Department of Informatics & Telecommunications
University of Peloponnese, GR 22100 Tripoli, Greece

{amtsafara,trifon,spiros,zervakis}@uop.gr

ABSTRACT

Among the basic research tools for (bio)medical science are epidemiological studies that typically involve a number of hospitals, clinics, and research centres scattered around the world, and are often referred to as multi-centre studies. Clearly, the effectiveness and importance of a multi-centre study increases with the number of participating centres and enrolled patients, but at the same time this natural distribution in the production of research data requires sophisticated data/knowledge management infrastructures to support the participating units. This kind of infrastructure is not only expensive to build and maintain, but also cannot be reused as it is often tailored to a specific study. In this work, we present a cloud-based system, that allows users without any computer science background to design, deploy, and administer platforms aimed for managing, sharing, and analysing clinical data from multi-centre studies. The proposed system provides a zero-administration, zero-cost online data/knowledge management tool that *(i)* enhances re-usability by introducing study templates, *(ii)* supports (bio)medical needs through specialised data types able to capture specialised knowledge like repeated therapies or treatments, and *(iii)* emphasises data filtering/export through an expressive yet simple graphical query engine.

1. INTRODUCTION

Large-scale epidemiological studies typically involve a number of different stakeholders, including hospitals, clinics, and research centres, physically distributed around the world, and are often referred to as *multi-centre studies*. These studies are invaluable as they collect large amounts of data from different regions, correlate them, and draw useful conclusions on important research questions. However, the physical distribution of the participants and the asynchronous nature of data acquisition pose a number of issues including the collection, organisation, and processing of data.

*Supported by EU and Greek funds through the EICOS project under the National Strategic Reference Framework Research Funding Program: Thales.

K-CAP'15, October 7–10, 2015, Palisades, NY, USA.
 ISBN 978-1-4503-3849-3/15/10
DOI: http://dx.doi.org/10.1145/2815833.2816949 ...$15.00.

To tackle data/knowledge fragmentation and address the issues arising from the coordination of geographically distributed participants, a number of platforms, that focus on the storage and management of (bio)medical data and knowledge, have been proposed [2, 4, 5, 6]. However, all these platforms are either designed for a *specific task or study* [2, 4, 6] (and are thus unusable in any other study), or require significant *computing infrastructure* and an *expert* in Information Technology (IT) for setup and tuning [5]. Typically, reconfiguring an existing platform for another study or setting one up from scratch results in *(i)* time-consuming meetings between scientists of different principles trying to understand each other's needs and *(ii)* resource-consuming IT infrastructure, that requires outsourcing to IT specialists and regular maintenance/upgrades to keep up with technological requirements. Due to these issues, a great number of multi-centre studies that lack the resources are still performed by resorting to *manual procedures*, such as collecting data on paper, exchanging data by post (either as hard copies, or as electronic copies stored in removable media), or emailing enormous (often outdated) spreadsheet files with (often sensitive) patient data among participants. Therefore, concerns like data freshness/integrity, participant coordination and degree of involvement, and timeliness of results are lost between versioning in exchanged spreadsheets, hard copies of patient data, and unanswered email requests.

In this work, we present a *cloud-based* service for *managing, sharing, and organising* clinical and patient data from *multi-centre studies*. The proposed system covers all the functional requirements posed by multi-centre studies, and enables researchers to easily organise and share data and knowledge generated by the research activity. We propose an innovative, integrated framework for creating platforms for multi-centre studies that enables users with *no prior IT knowledge* to *(i) design and launch*, in an easy and transparent way, platforms tailored to the specific needs of their studies, *(ii)* perform basic and advanced *user management* tasks (manage users, assign user privileges and permissions, perform access control on data), *(iii) record, organise and manage* clinical/patient data by resorting to a number of built-in and customisable data entry forms, and *(iv) search and filter* information by using a powerful yet simple point-and-click mechanism that poses restrictions on the stored data and extracts the requested information in a number of formats/outputs including raw data, pie/column charts, and ready-to-process spreadsheets. Due to the cloud infrastructure, computational resources are allocated on demand, providing *elasticity* and *fault-tolerance* in a way that is transparent to the end-user.

The contributions of this work are twofold:

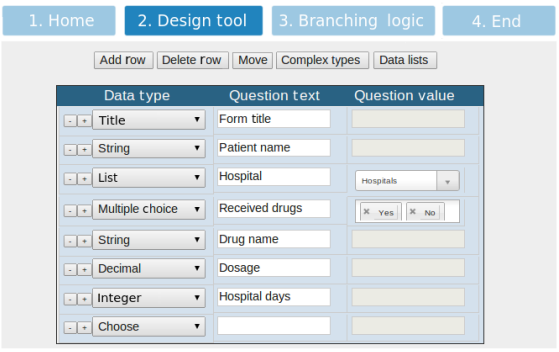

Figure 1: Platform creation and editing.

- We propose a *cloud-based, zero-cost, zero-administration* tool that offers both fundamental and advanced user and data/knowledge management functionality for multi-centre studies. To the best of our knowledge, this is the first cloud-based system that focuses on multi-centre studies and allows users to deploy their own platforms within minutes, alleviating the need to rely on expensive custom-made solutions that require IT infrastructure and skills to maintain.
- We present the architectural considerations and solutions behind the proposed tool, and describe a number of *novel services* that allow users without any prior IT knowledge to create, administer, launch, and use personalised platforms.

Our system is currently under beta testing for multi-centre studies led by the Hellenic Society for Chemotherapy and the University Hospital Attikon, and has already been used by more than 12 public hospitals in Greece.

The rest of the paper is organised as follows. Section 2 describes the implemented services and functionality, outlines the system architecture and describes a demonstration plan to be presented at the conference, while Section 3 discusses related work.

2. SYSTEM OVERVIEW

Our system supports three different types of users that correspond to three data access privileges. More specifically, we have *(i) IT administrators* (ITAs) that are responsible for the update and maintenance of the system and have access (edit, delete, or filter) to all data in the database, *(ii) study administrators* (SAs) that are typically in charge of an ongoing study and may access all data related to the specific study, and *(iii) participants* in an ongoing study that may access only the data added by them.

2.1 Supported Functionality

Questions and platforms. *Questions* are the backbone of multi-centre studies. To support the set of questions of a particular multi-centre study, we introduce the concept of *platforms*. Technically, a platform is a custom-made set of questions, together with all necessary user administration and data/knowledge components of a study. Platforms are typically created and administered by the SA. More specifically, the SA is able to *(a)* design and launch a new platform by inputting a platform name and defining a number of study questions and *(b)* rearrange, or delete questions through drag-and-drop actions. For each question, the SA specifies three key elements: the *data type*, the *question text*, and the *question values*. This process is performed with the design tool of Figure 1. Our system supports all standard

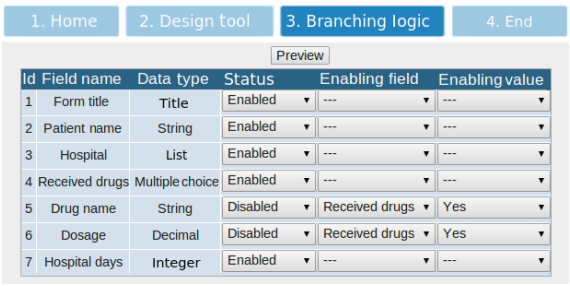

Figure 2: Platform branching logic.

data types such as *strings*, *integers*, *dates*, *decimal*, etc. Additionally, it supports *titles*, *multiple choices*, *lists* and *complex data types*.

Titles are used to introduce new questionnaire sections.

Multiple choices restrict possible answers to a fixed set.

Lists allow data input from dynamic, custom-made drop-down menus that the SA dynamically creates, stores, and edits during the platform design or maintenance. These lists may be shared across different studies of the same user and are used to ensure data consistency, enhance data integrity, and enforce validation of input. To define a new list, the SA specifies its unique name and defines its contents. Subsequently, when specifying a question, the SA sets the data type to list and selects one of the stored drop-down lists. For example in Figure 1, the third question specified is a user-defined list of hospitals participating in a study.

Complex datatypes support medical operations that involve groups of recurring questions – as in treatment plans or therapeutic protocols which usually consist of a set of medicines with periodically recorded data (e.g., name, start/end dates, outcome). The advantages of creating complex data types include better data modelling and knowledge capture, and thus richer query possibilities, and flexibility in the design of studies with complex/repeating medical processes.

Branching logic. SAs may also specify the *branching logic* of questions. This functionality allows SAs to specify the values in questions that are required to enable or disable following ones. For example in Figure 2, the SA specified that if the answer to the question Received drugs (Id = 4) is yes, the questions Drug name and Dosage (Id = 5 and Id = 6, respectively) should be enabled.

Templates. Typically, specific parts of a study may be used (as they are or with minor modifications) in other studies. For example, demographic data are a typical resuable part of many biomedical multi-centre studies since they hardly change among different studies. To support this reusability, the system offers the SA the possibility to create *platform templates* that may be used across different studies. Figure 3 shows a platform template meant for demographic data.

Editing an existing template involves two different scenarios: *(i) simple editing* that includes adding, deleting and modifying questions or their branching logic, and *(ii) structural editing* that relates to the modification of question order or data types. Such editing could affect the consistency of stored data/knowledge or cause compatibility problems between the stored data/knowledge and the new template and is thus limited to actions that do not cause such issues. For instance in Figure 1, the SA is not allowed to change Patient name to integer.

Figure 3: An example template for demographic data.

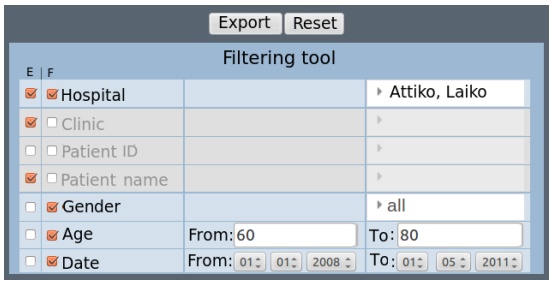

Figure 4: The filtering tool.

Filtering and chart tools. SAs are able to extract information from a study, while participants may only extract information added by them. Such information may be retrieved and presented as *(i)* a set of tuples (using the *filtering tool*), and *(ii)* a graph (using the *chart tool*).

The filtering tool (Figure 4) uses a powerful yet easy-to-use query issuing mechanism that allows users to *(i)* filter and retrieve and *(ii)* export to a third-party application stored records by applying constraints with simple point-and-click interactions. Data filter and export involves a *two-step process*. In the first step the user defines the query output by checking the questions that will be used for *data projection* (i.e., data to be exported). In the second step, he applies one or more *filtering conditions* on the data. The filtering conditions are introduced by presenting *all distinct values* stored for a specific question and allowing the user to define the ones that satisfy the filtering criteria. In this way, he may define *conjunctions* and *disjunctions* both on the questions and on the stored data. The query result may then be exported in a spreadsheet, or presented as a list of tuples. In Figure 4, the SA has selected to export the results for Hospitals, Clinics, and Patient names (notice the checked boxes in the first column) for all records that were input between 01/01/2008 and 01/05/2011 in Attiko or Laiko hospital and involve patients between 40 and 60 years of age (notice the specified constraints).

The filtering tool is able to capture the complex data types described earlier, allowing the user to *(i)* set more than one filters for every complex data type and *(ii)* include constructs like concurrent episodes of a diagnosis or treatment. Examples of supported queries include:

- Show patients older than 40 years of age with temperature greater than 38, who were diagnosed with microorganisms (Pseudomonas and Candida). Notice that the names of the microorganisms are values already stored in the database.
- Show patients between 40 and 60 years of age, who live in (England or France or Greece), have been diagnosed with microorganisms ((Pseudomonas and Candida) or (Pseudomonas and Salmonella) or (Citrobacter)), have undertaken therapy with (Ampicillin or Cefotaxime), and had progress as the eventual outcome of their condition. Notice the application of Boolean operators both on the questions, and on the data stored for each question.
- Show patients who suffer from Massive hemoptysis, had an initial episode of Bacteremia and accepted a dose between 2 and 4mg of antibiotic Amikacin, and subsequently had a second episode of Bacteremia. Notice that the query is applied to a complex data type (recurring episodes/treatments as described above).

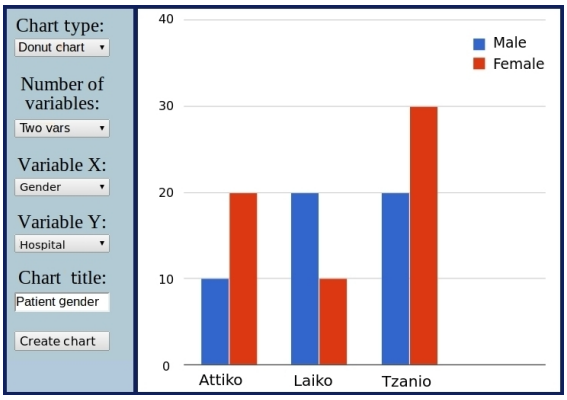

Figure 5: Using the chart tool with two variables.

The chart tool (Figure 5) may be used to extract and present information from stored data directly as a *pie, column, bar, or donut chart*. It supports the creation of single and multiple variable graphs depending on the constraints defined by the user. In this way the user may dynamically create graphs for queries with a single variable like "how many patients in the database are male/female", but also for queries with multiple variables like "how many patients are male/female in each hospital that participates in the study" (shown in Figure 5).

2.2 System Architecture

The main idea of the system is to allow users to design and build platforms, through a series of simple and adaptive processes. This can be done transparently through simple-to-follow wizards from users without any IT training, while the cloud-based architecture automatically adapts to the resources and infrastructure by relying on cloud elasticity.

The cloud functionality is provided by the open-source platform *ownCloud* (`http://owncloud.org/`), setup over a medium-sized computing infrastructure available at the university campus. Figure 6 presents a high-level view of the system architecture and the different types of modules implemented. The Cloud API is responsible for performing all necessary communication with the ownCloud platform and provides elasticity services, while the storage manager performs all necessary storage/retrieval operations to the data/knowledge base backend. Our backend implementation uses the LAMP framework as the backend infrastructure, while the rest of the modules have been developed using Javascript, PHP, and JQuery.

The Study Manager module is responsible for the creation, editing, and management of studies, and consists of a number of modules utilised to *(i)* manage the participants and the stored data associated with a study and *(ii)* filter/extract data requested by a SA. The Platform Manager

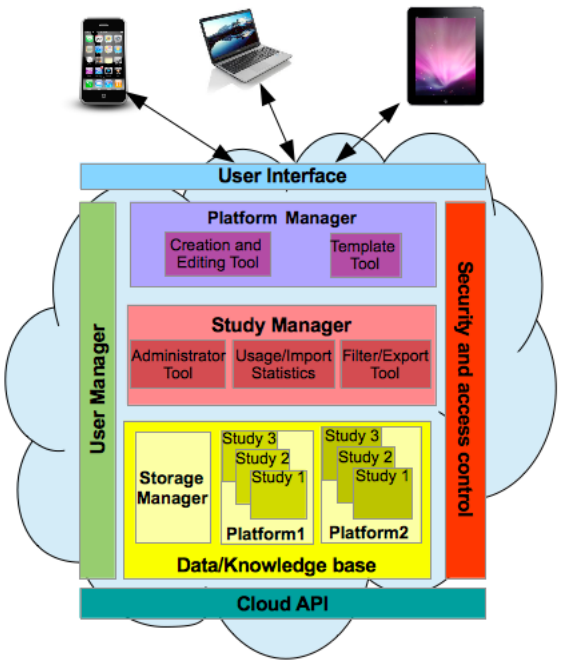

Figure 6: A high-level view of the architecture.

is used to create, edit, and manage platforms and templates utilised by different studies. The User Manager module is utilised by the ITA to create and manage the SAs, and also by the SAs to create and manage the participants (and their roles) in a specific study. The Security and Access Control Module enforces the security policies for the system and controls access privileges over the stored data. Security features include certificate and password-based authentication, single sign-on policy, and role-based user management. Finally, the User Interface module is responsible for identifying the hardware used to connect to the sytem (PC, tablet, smart phone) and adjust the viewing components accordingly.

Apart from the poster presentation we also plan to demonstrate the functionality of the system at the conference. The interested reader may find more information about the project at `www.uop.gr/~trifon/CloudStudy/`.

3. RELATED WORK

Over the years, many solutions aiming at the management and sharing of (bio)medical information have been proposed; in what follows we focus on approaches related to *electronic patient record (EPR)* systems specifically designed for multi-centre studies. Most of the proposed systems focus on a *specific study*, and put forward architectures and services tailored to the problem at hand. [10] presents an information system that may be used to manage multi-centre studies for cancer. Similarly, MSBase [3] introduces a web platform for collecting prospective data on patients with multiple sclerosis, while [2] presents a system for HIV/AIDS prevention and treatment. Finally, a number of EPR systems [4, 8, 9, 6], platforms (e.g., [5], HICDEP – `http://www.hicdep.org/`), and projects (e.g., CASCADE – `http://www.ctu.mrc.ac.uk/cascade/`, COBRED – `http://www.cobred.eu/`) have also been launched for supporting *(i)* different types of focused multi-centre studies [6] and *(ii)* clinical/biomedical data integration [7, 12, 11, 1].

The large number of existing specialised systems and the needs dictated by each different multi-centre study led researchers to the design of systems that are able to support *classes* of functionalities needed in multi-centre studies. The most prominent paradigms in this line of work are the REDCap project [5] and the Qure system [6]. REDCap provides a principled way of designing, constructing, and managing databases that are able to support multi-centre studies. However, to deploy a system for a specific study, the study coordinator needs to get in touch with the REDCap project, and inform the IT specialist on the specific needs and requirements of the study at hand. Subsequently, after an iterative and possibly long process of refinement and corrections in the database design, the REDCap IT expert will deploy the database and the users will be able to enter the data. Obviously, any subsequent changes in the specifications will result in the redesign of the database and the porting of the inserted data in the new database. On the other hand, the Qure system, while supporting online creation of study questionnaires, offers *(i)* limited data types (e.g., does not offer custom drop-down lists or complex data types), *(ii)* no data export functionality (e.g., pie/column charts, spreadsheets, SPSS compatible output), *(iii)* a query engine with limited expressiveness, and *(iv)* runs on dedicated hardware with no adaptation policy (e.g., elasticity of resources) and low quality-of-service guarantees.

4. REFERENCES

[1] Serge Abiteboul, Bogdan Alexe, Omar Benjelloun, Bogdan Cautis, Irini Fundulaki, Tova Milo, and Arnaud Sahuguet. An electronic patient record on steroids: Distributed, peer-to-peer, secure and privacy-conscious. In *VLDB*, 2004.

[2] A. Nucita et al. A global approach to the management of emr (electronic medical records) of patients with hiv/aids in sub-saharan africa: the experience of dream software. *BMC Medical Informatics and Decision Making*, 2009.

[3] H. Butzkueven et al. Msbase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Multiple Sclerosis*, 2006.

[4] H.S. Fraser et al. An information system and medical record to support hiv treatment in rural haiti. *British Medical Journal*, 2004.

[5] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J.G. Conde. Research electronic data capture (redcap) âĂŞ a metadata-driven methodology and workflow process for providing translational research informatics. *Biomedical Informatics*, 2009.

[6] M. Jager, L. Kamm, D. Krushevskaja, H. Talvik, J. Veldemann, A. Vilgota, and J. Vilo. Flexible database platform for biomedical research with multiple user interfaces and a universal query engine. In *DB&IS*, 2008.

[7] Toralf Kirsten, Jörg Lange, and Erhard Rahm. An integrated platform for analyzing molecular-biological data within clinical studies. In *EDBT Workshops*, 2006.

[8] Arvind Kumar, Amey Purandare, Jay Chen, Arthur Meacham, and Lakshminarayanan Subramanian. Elmer: lightweight mobile health records. In *SIGMOD*, 2009.

[9] Wen-Syan Li, Jianfeng Yan, Ying Yan, and Jin Zhang. Xbase: cloud-enabled information appliance for healthcare. In *EDBT*, 2010.

[10] M. Martinez, J.M. Vazquez, M.G. Lopez, F.M. Arnal, B. Gonzalez-Conde, J. Pereira, and A. Pazos. Semantic integration of data in an information system for multicenter epidemiological studies on cancer. In *MIE2008*, 2008.

[11] Rakesh Nagarajan, Mushtaq Ahmed, and Aditya Phatak. Database challenges in the integration of biomedical data sets. In *VLDB*, 2004.

[12] James F. Terwilliger, Lois M. L. Delcambre, and Judith Logan. Context-sensitive clinical data integration. In *EDBT Workshops*, 2006.