Enabling Business Intelligence Functions over a Loosely Coupled Environment

Giampaolo Armellin[1], Leandro Paulo Bogoni[1,2], Annamaria Chiasera[1,2], Tefo James Toai[1], Gianpaolo Zanella[1]

[1] GPI SpA, Via Ragazzi del '99, 13 - Trento, Italy
[2] Information Engineering and Computer Science Department - University of Trento, Trento, Italy
bogoni@disi.unitn.it,
{garmellin, achiasera, ttoai, gzanella}@gpi.it

Abstract. Planning effective and well targeted actions to manage and improve the local and national healthcare services requires institutions to understand and analyse the real needs of the population based on reliable and timely statistical analysis on citizens' health state. This is particularly important in developing countries in which healthcare facilities lack ICT infrastructures and network connectivity, making data collection and analysis particularly difficult with a considerable manual effort leading to potentially unreliable or incoherent information. In this scenario, we propose a generic communication infrastructure, developed in the SIS-H project for Mozambique hospitals to capture, communicate and analyse clinical events. Our solution enables the exchange of data amongst healthcare facilities over all the different aggregation levels of the hierarchical healthcare system of Mozambique regardless of the availability of communication media (e.g., compact disk, usb stick, web-internet). The plugin-based solution adopted supports reporting and Business Intelligence analysis for exploring data at different granularity levels.

Keywords: Business Intelligence, e-Governance, plugin-based architecture

1 Introduction

Improving accessibility and quality of health services is one of the outcomes of the Mozambican Health Sector Strategic Plan (PESS 2007-2012 6, 5). Despite public and external founds, medical assistance is not accessible to all the population owing to the difficulty in coordinating and controlling the actions performed and a chronicle deficiency of human resources. The World Health Organization points out the need of monitoring systems to control resource flows, progresses and outcomes finalized to monitor and evaluate health services [5], [9]. In the Strategic Plan for the Health Sector [6] the role of such monitoring system is essential for decision makers to manage and supervise the maintenance of hospitals. Another pre-requisite to evaluate and improve health quality is to encourage community participation with the active communication of the results of the analysis promoting transparency and accountability.

In this paper we will report our experience in the development of a communication infrastructure applicable at the national level for the monitoring of the quality of healthcare system in Mozambique. We propose a lightweight and robust infrastructure that requires minimum resources to operate with poor or no connectivity to support people in the collection, processing and communication of statistical information with a simple and intuitive tool usable even by not trained personnel.

1.1 The Context

The structure of the Mozambican Healthcare System is level-oriented and hierarchical – i.e. operational units and hospitals are grouped according to managed services and geographical areas. The central hospitals of Beira and Maputo are at the top levels of the hierarchy (level IV hospitals) and refer directly to the ministry of health. Pemba provincial hospital is at level III and refers to the Provincial Health Office.

The ministry of health and the provinces provide economical support to operational units and hospitals and plan actions and projects to prevent and deal with chronicle and highly contagious diseases (like Malaria, HIV/AIDS, Cholera). These activities include, among others, economical investments to improve the infrastructures (hospitals' facilities like emergency rooms, intensive units, infective divisions), staff training, and education of the population for prevention.

Governing bodies need to know which are the real needs of the population to identify and deal with sanitary emergency, to plan an effective set of activities (e.g. the continuous and periodic surveillance required by HIV epidemic [7]) and to reduce waste of resources (staff and financial). In [10] it is highlighted how this requires at first the surveillance of the morbidity and mortality causes. The analysis of such statistical data allows on one side to understand the health state of the population and on the other side to monitor hospital resource utilization and in particular costs of hospitalization and staff employed. Based on these analyses the ministry of health can better plan public health interventions, for example to grant a suitable number of beds in hospitals or health posts facing particularly critical sanitary crisis.

As pointed out in the analysis of Campione et al. [10] the social, economical and technological environment of Mozambique poses some strict constraints to realize such analysis that in developed countries are normal Business Intelligence activities [2]. In particular [10] identifies the following challenges:

- lack of IT skills: qualified personnel, able to operate IT devices and tools is scarce and consequently the data collected is not trustworthy;
- lack of data: the diagnostic capability of the lower level point of care is rather poor compared to more specialized level of care (hospitals of level III or below have less diagnostic facilities than hospitals of level IV);
- territory highly distributed and sparsely populate: in order to get a comprehensive picture of the health state of population it is necessary to have a good coverage of the territory at the national level;
- infrastructure limitations: no electricity, no information communication technologies, no connectivity and no computers available;

- lack of information systems: there are neither the applications to manage health records nor the personal information of patients and consequently it is difficult to identify patients in the long term and to maintain detailed information on their health story.

Data is typically collected on paper with significant manual effort and difficulties of communication and analysis by other parties at higher level of the hierarchy where planning and resource management is performed. The lack of computerization in the point of care, especially at the periphery, makes the data collection particularly difficult; with a considerable amount of time and manual effort required from the personnel. Furthermore, people are not motivated to spend effort in such a boring task as they do not see any useful results in short time coming out. These factors leave the data collection process incomplete and error-prone.

Indeed a great step forward has been performed with the adoption of a standardized list of diagnosis selected from the ICD-10 standard [4]. However, an IT solution capable of effectively supporting the collection of data, and is easy to install in critical environment with lack of infrastructure and is simple to use even for poorly trained people; is still missing. This solution should be general enough to be usable at any level of the national health care system with a great degree of flexibility and customizability to collect as much information as possible.

In this work we present the solution we provide in the context of the project "SIS-H – Módulo Internamentos" launched with internal tender by the Mozambican Ministry of Health (MISAU) according to the Mozambique eGovernment Strategy.
The main project goal is to rapidly overcome the lack of statistical information for hospitals of level III and level IV, devising software applications to collect and aggregate data in the clinical area, with specific focus on admissions and discharges of patients. Those aggregated data concern Maputo Central Hospital, Beira Central Hospital and Pemba Provincial Hospital and are being periodically sent to Provincial Health Offices and to the Ministry of Health, enabling analysis and planning on the health care system.

We provide a platform that effectively answers the challenges identified above both technical (lack of IT infrastructure) and organizational (data sources distributed on the territory with a hierarchical organization, few qualified human resources)[1].

We believe that a typical solution to perform business intelligence analysis realized for use in developed countries cannot be applied as-is in developing countries because the problems, the needs and the constraints are completely different. For that reason, we choose to work in strict collaboration with local companies and governmental organizations to bring our experience in the development of similar solutions in such critical context: which only domain experts living directly in contact with these problems could help us to fully understand.

[1] Our solution brings new opportunities for the use of the same platform also in other domains, such as economy and education.

2 Related Work

Mature technologies and methodologies are available in literature [2][13] for the design and development of data warehouse (DWH) systems that may serve our data analysis problem. However, these approaches make some assumptions that are not valid in our context. First of all, they assume data can be extracted from IT systems at the data sources and centralized in a data warehouse. The problem is that in our case such IT systems are not always available (as in most of the cases they are not even present) and there is not (yet) a central data collector that could host a DWH.

Furthermore, classical DWH approaches assume an enterprise is highly motivated to create and maintain active a DWH sending timely information. In our case people acting as data collectors may be: not only employees of an organization but also volunteers, that should be motivated to collect timely and correct information; even if they do not understand or see immediately the positive effect of their work.

Another limitation is that these solutions typically assume a DWH is loaded basically in a real time mode, but in our case data may arrive with delays of weeks or not arrive at all.

Solutions for Business Intelligence available in the open source community like SpagoBI[2] or Talend[3] are interesting but too complex for our case as they provide many functionalities that in these contexts are not useful, and they risk to transform the solution into an unmanageable tool that nobody is able to use. Instead, we need to aim at simplicity with few and intuitive functionalities. Furthermore, these systems are general purpose BI applications that cannot be easily customized and cannot be adapted easily to different contexts. Besides, they require an amount of resources that in our case are not available; such as memory occupancy, local database and trained people to configure and customize the Extract, Transform and Load (ETL) [14] and reporting phases.

Developing an information integration solution using poor data quality is totally useless and also risky as it may lead to wrong results from the analysis and consequently wrong decisions. Conscious of this problem, we are currently studying approaches to trace and improve data quality and in particular to: keep the consumers of BI analysis (BI consumers) informed on the quality of the results they are using; trace and identify the sources and causes of mistakes with the help of the BI consumers by colleting suggestions on how to correct the mistakes; minimize the occurrence of the mistakes with tools to support critical phases of the data lifecycle, like data entry which is a task essentially with manual effort which is the primary cause of inconsistent data.

In this regard, we will pay a particular attention to the data entry and the cleaning phases and to the solutions already proposed in the research arena in the hope to reuse and extend them. For example, in [11] is proposed Usher, an end-to-end system to design forms that dynamically adapt to reduce the probability of errors during data entry (e.g. providing feedback in real-time to guide the data enterer toward more likely values).

---

[2] http://www.spagoworld.org/xwiki/bin/view/SpagoBI/

[3] http://www.talend.com/

The work in [1] extends the traditional Functional Dependencies into *Conditional functional dependencies* (CFDs) to capture the notion of "correct data" and to improve Data Cleaning tools in the detection and correction of inconsistencies and errors in the data.

As pointed out in [3] dirty data[4] often violates some integrity constraints reflecting organization's policies related to the quality of the data. To deal with this, [3] presents an approach that suggests possible rules and identifies conformant or non-conformant records (context-dependent rules).

3 Solution

Currently, healthcare services are monitored by collecting (often manually) data on the clinical process and sending hardcopy reports to districts, provinces and finally to the healthcare office (MISAU). Owing to some lacks on IT-systems and connections, data capture and analysis operations cannot always guarantee reliable and coherent information.

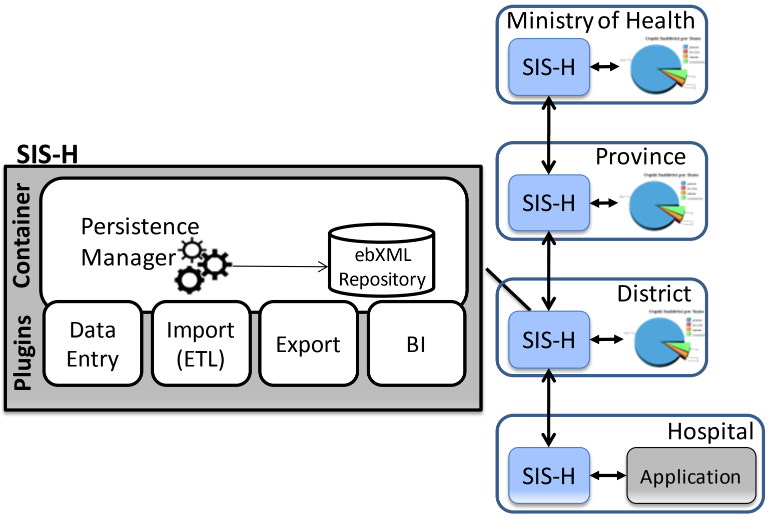

Fig. 1. Deployment of the SIS-H module in the healthcare hierarchical organization of Mozambique.

We provide a generic communication infrastructure, SIS-H, to exchange data amongst facilities, across the whole healthcare system hierarchy – from operational units, hospitals, districts, provinces and up to the ministry – regardless of the available communication media (e.g., compact disk, usb stick, web-internet).

The SIS-H module could be replicated at the different sites of the Mozambique Healthcare organization and be applied also in other domains (e.g. to support the

[4] Not suitable for data analysis, with data problems such as inconsistency and incompleteness.

Ministry of Education and of the Interior). The caption in (Fig. 1) shows the main components of the module divided into a container managing data persistence (ebXML Repository) and a series of plugins providing the applicative functionalities and in particular: Data Entry, to support users to type in the system aggregated statistics on patient admissions and discharges that are collected on paper forms; Import (ETL), to load statistic data collected from other sites, typically from other levels of the organization; Export, to prepare an export of statistical information to be imported and used in another site, typically at a higher level of the hierarchy; BI, a module to produce business intelligence reports.

The next section presents a more detailed description of how the architecture works and in particular how the SIS-H module can produce an export that is understandable to installations in other sites allowing their interoperability.

3.1 Architecture

Each SIS-H plugin serves a dedicated purpose independently from the rest of the components so that modules can be changed or replaced without any impact on the rest of the system. This ensures greater flexibility, scalability and maintainability of the system. The main components of the architecture are the container and the various plugins for data Entry, data import (ETL), data export, and BI purposes.

The core functionalities of the system are given by the container which encapsulates a persistence manager (see Fig. 1). The Container is decomposed into five pillar features namely: *Plugin Manager*, *Workflow Manager*, *Data mapping/conversion Manager* and *Plugin Installation Manager*.

The *Plugin Manager* is responsible for associating user actions to dedicated plugins as well as loading configurations of the selected plugin. The *Workflow Manager* executes the workflow of the plugin by calling exposed methods through java reflection. The *Data mapping/conversion* manager transforms plugin internal data objects into the ebXML format (defined in [8]). Finally, the Plugin installation manager deals with addition and removal of plugins into the system. The behaviour of a plugin in terms of functionalities accessible by a certain user, data accessible and output formats are configured in the Plugin installation manager and can be easily updated to adapt to different contexts.

The SIS-H modules could be deployed in three flavours: centralized, client-server and embedded. In the *centralized* configuration, the basic layer contains the ebXML repository for data persistency while in the *distributed* case the ebXML repository is hosted in a remote application server as in Fig. 2. Finally, the *embedded* configuration is released with an embedded database (like Apache Derby[5]) for a fully portable system. The system can work both in a loosely connected environment and in a fully networked area[6].

---

[5] http://db.apache.org/derby/

[6] On one hand, in remote areas where the network connectivity may be poor or even not be present, a fat client approach in which the application and persistency layer reside on the local machine or are embedded in the application would be suitable. On the other hand, in larger cities such as Maputo (the capital of Mozambique), equipped with network connectivity, the system can run in a client-server mode with LAN or internet configuration.

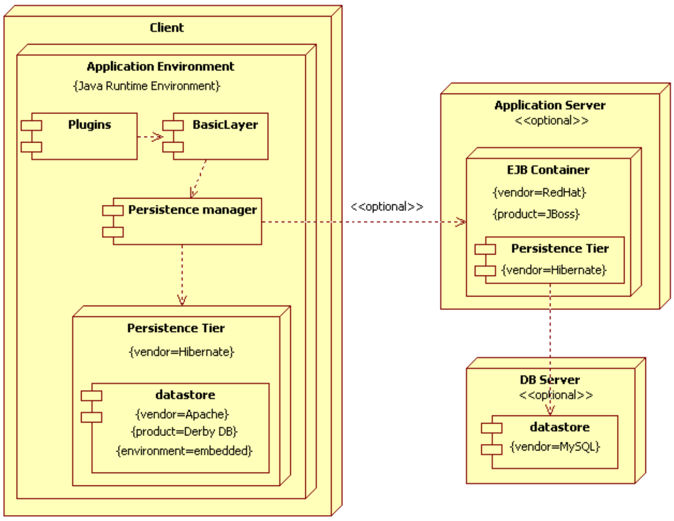

Fig. 2. Deployment diagram of the SIS-H application.

It is worth mentioning that the application can support different languages and geographic regions through the use of ResourceBundle [12] residing outside the architecture, thus new languages can be added without changing the overall architecture of the system.

3.2 Conceptual Model

Fig. 3 shows the organizational model of the hierarchical structure of the Mozambican health system. The healthcare facilities are organized in Health Care Units (health centers and clinics) and hospitals of different kinds (Specialized Hospital, Rural Hospital and so on). Each hospital is divided into departments which offer specific services. Each healthcare facility refers to an entity of the Government (e.g. the Province or MISAU) to which it must submit reports (known as "*ficha*") on the activities performed. The data structure of such reports reflects at a conceptual level the internal operating procedures of hospitals and it is mapped internally and transferred from one level of the organizational hierarchy to the next in XML according to the ebXML standard.

Since there is no centralized database in which all the data is stored, it is vital not only to identify each *ficha* uniquely across different health units all the way up to the government administrative domain, but also to capture the organizational structure and the hierarchical level from which the *ficha* originates. In essence, each SIS-H module is aware of its location in the hierarchy of the health system. In this way it is simple to map other hierarchical organizations (for example, to add another hierarchical level, like an internal department or another top level administrative body like another ministry).

The adopted encoding for uniquely indentifying the *ficha* follows the hierarchy of the structure up to the service point in which the data was transferred. For example, if the operator that performs data export is registered in Maputo under Medicine department in section Surgery the unique identity of the *ficha* transferred would be as shown by the path element in Fig. 4.

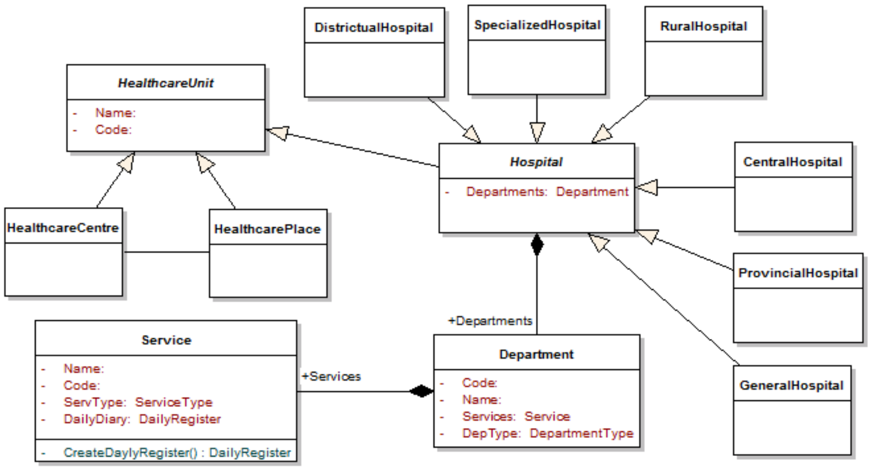

Fig. 3. Health system organizational model and relationships between units and structures of government.

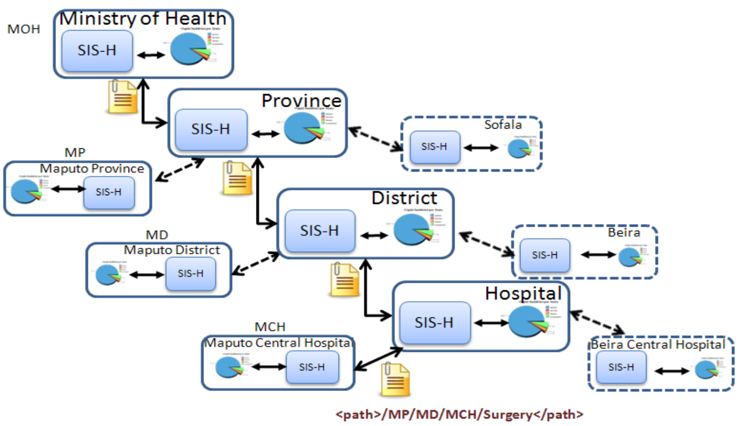

Fig. 4. *Ficha* flow and embedded organizational structure unique identity.

4 Deployed Prototype

All the system is designed and developed according to open source principles and technologies. Fig. 5 shows the main form of the application to access to the following features: the Data Entry plugin, used for the daily registration of admissions and discharges of patients; the Export Plugin used to export the "*fichas*" XML files to exchange data; the Import Plugin to import the data previously exported by a different location; the BI Plugin, used to produce BI reports, in which we adopted Jasper Intelligence[7] as BI tool; and System Management used to manage the users, the general system configuration, backup procedures, etc.

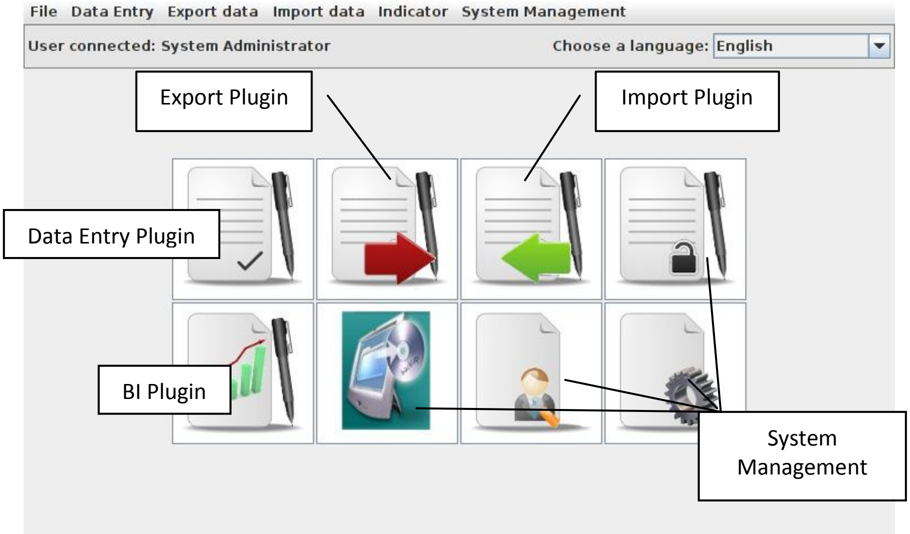

Fig. 5. Main Form sample, and the access to the Plugins.

5 Conclusion and future work

Our modular architecture divided into plugin modules allows composing a suite of functionalities on the base of the needs and characteristic of the installation site. Plugins can be easily updated and configured to customize their behaviour and functionalities, data accessible and output formats to adapt to different contexts. The modular architecture allows for simple integration of specific functions based on flow type, control functions and validation information.

In the future, our architecture could be extended to manage other protocols and media, enabling the architecture to be applied in other contexts like the ministries of

[7] http://www.jaspersoft.com/

finance and education. Other extensions could introduce the research results on the quality of the data managed inside our architecture.

Acknowledgments

The project is headed by the Mozambican Ministry of Health (MISAU), collaborating with the Department for Information and Health (DIS), MOASIS (closely related with the University E. Mondlane) and the Consortium Pandora Box Ltda – GPI Spa. Our special thanks to Dra. Célia Gonçalves, Dr. Amisse Momade, all team at MISAU, Dr. Alessandro Campione and all MOASIS team.

References

1. P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, "Conditional functional dependencies for data cleaning," Proc. ICDE, 2007.
2. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition. John Wiley & Sons, 2002
3. Fei Chiang , Renée J. Miller, "Discovering data quality rules", Proceedings of the VLDB Endowment, 2008.
4. World Health Organization. ICD-10: International Statistical Classification of Diseases and Related, Health Problems. 10th Revision. http://apps.who.int/classifications/apps/icd/icd10online/
5. World Health Organization. Mozambique, Mozambique's health system: Health and development. http://www.who.int/countries/moz/areas/health_system/en/index1.html
6. Ministério da Saúde, Plano Estratégico do sector Saúde, 2007 -2012, República de Moçambique. http://www.who.int/countries/moz/publications/pess_2007_2012.pdf
7. Ministry of health. Report on the revision of the data from HIV epidemiological surveillance. Round 2007. February 2008. Republic of Mozambique. http://www.misau.gov.mz/pt/content/download/4637/28101/file/Ronda2007EN.pdf
8. Standard ebXML: Registry Information Model. http://www.ebxml.org/specs/index.htm#technical_specifications
9. World Health Organization. Terms of Reference for Designing the Requirements of the Health Information System of the Maputo Central Hospital and preparation of the Tender Specifications. Technical Report. Version 2.0. January 2007
10. Use of ICD-10 for morbidity and mortality notification for in-patients, in recourse limited settings. The experience of Mozambique using reduced disease lists. Roberta Pastore, Alessandro Campione, Bernardina Gonçalves, Armando Melo, Celia Goncalves, Carla Silva Matos, Marcelino Mugai. Annual meeting of the WHO Family of International Classifications Network
11. Chen K, Chen H, Conway N, Hellerstein J. Usher: Improving data quality with dynamic forms. ICDE. 2010.
12. Sun Microsystems. Java Resource Bundle. [Online] http://java.sun.com/j2se/1.4.2/docs/api/java/util/ResourceBundle.html
13. Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. *ACM Sigmod record*. 1997.
14. R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004.