Pisa takes a stand for responsibility in healthcare and medical technology
6th Annual HCTM Conference -HOF- Scuola Superiore Sant'Anna
3-5 October 2007, Pisa, Italy

Telematic integration of health data: the INTESA project

Remo Bedini, Lorenzo Guerriero, Matteo Dalle Luche, Silvia R. Viola, Ivan Porro, Angela Testi

Institute of Clinical Physiology, National Research Council, Via Moruzzi 1, 56124, Pisa, Italy. E-mail: bedini@ifc.cnr.it

Abstract: Following an approach based on the methods of basic research, the INTESA project has developed a complete architecture of health information system, capable to guarantee a smart and safe storing of the essential information, an effective and personalized retrieval of data, and some innovative models to compare the results of clinical and medical activities of all the "actors" of the health care process. Together with other metropolitan repositories based on HL7 messages and applications able to examine the data stored, the developed archive will contribute to keep a check on every citizen's health history, clinical examinations and cure therapies, but, above all, it will allow to verify the efficacy and efficiency of the health care processes related to particular pathologies.

Keywords: electronic health record; data storing and integration; data warehousing; data mining.

Biographical notes: Remo Bedini is Senior Investigator in Pisa Institute of Clinical Physiology of the Italian National Research Council. He obtained his degree in Electronics Engineering at the University of Pisa. He participated to EU projects on Standard Communication Protocol for Computerised ECG, was member of the Project Team 007 CEN/TC 251 for ECG transmission standards and was responsible of units in the Camarc and Briter EU Projects. He was member of CEN/CENELEC and UNI for biomedical standards. He was responsible for Italian projects concerning automatic ECG, telemedicine, multifunctional telemonitoring and assisted gas-therapy. He leads research groups on Medical Electronics, Telemedicine and biotelemetry in endurance sport and high risk professional activities. He has organized the II level University Master in Underwater and Hyperbaric Medicine. He is senior member of the International Society on Biotelemetry. He is author of more than 170 scientific papers and nine patents in Italy, USA and EU.

Lorenzo Guerriero works as researcher for the National Research Council in the Institute of Clinical Physiology of Pisa. He graduated in Telecommunications Engineering at the University of Pisa. He received his master's degree in Information Technology from the CEFRIEL (Center of Excellence For Research, Innovation, Education and industrial Labs partnership) of Milan. His research areas of interest include Information & Communication Technologies, web communication and design, Telemedicine and e-health.

Matteo Dalle Luche works as researcher for the National Research Council in the Institute of Clinical Physiology of Pisa. He graduated in Computer Science at the University of Pisa. He received his master's degree in Biostatistics from the department of Statistical Sciences of the University of Bologna. His

research areas of interest include Computer Science and Biostatistics.

Silvia R. Viola is currently with the "Università Politecnica delle Marche" of Ancona as Research Associate and with the University for Foreigners of Perugia as Lecturer for the Course of Knowledge Engineering. She graduated in Philosophy at the University of Pisa. She got her PhD in E-Learning at the "Università Politecnica delle Marche" of Ancona. Her research interests are in mathematical models of the WWW data, learners' profiling by data driven approaches, data driven approaches for monitoring and quality assurance in E-Learning and mathematical bioengineering modelling. She is currently IEEE, IEEE CS and ACM Member, Member of the Program Committee of different Conferences and Workshops on Learning Technologies, Executive Peer-Reviewer for Educational Technology & Society Journal and for International Journal of Emerging Technologies in Learning, and is currently co-chairing two International Workshops on Educational Data Mining and Data Mining for E-Learning.

Ivan Porro, Msc Eng in Biomedical Engineering, currently works at the Laboratory for Bioimages and Bioengineering - BIOLab - in the Department of Communication Computer and System Sciences (DIST) of the University of Genoa. He received his PhD in Bioengineering from the University of Genoa. His primary research interests are in the field of medical informatics and bioinformatics data integration, Grid computing and medical image processing. He is co-author of about 20 papers and conference proceedings. He is member of the Italian Bioinformatics Society (BITS) and of the National Group of Bioengineering (GNB).

Angela Testi is Associate Professor of Political Economy in the School of Economics, University of Genoa. She teaches courses of Microeconomics and Health Economics in the School of Economics, Engineering, Medicine, Law in Genoa. Her research areas of interest include quantitative evaluation methods applied to healthcare delivery and social and equity issues such as deprivation indexes, quality indicators, appropriateness of levels of care. She has published in journals such as the Health Care Management Science, Journal of Evaluation of Clinical Practice, International Journal of Simulation. She is responsible for many research projects funded by Italian Health Ministry and is member of many scientific committees.

1 Introduction

The integration of heterogeneous and scattered data is actually seen as a growing problem of modern health systems and represents a crucial aspect if related to the development perspectives of Information & Communication Technologies (ICT) in Hospitals and, in general, of Telemedicine. Nowadays, health data integration has become an important field of study and many researchers are focusing their attention on this emerging challenge (Yoo, Kim, Park, Choi and Chun, 2003; Hanzlicek, Spidlen and Nagy, 2004; Nardon and Moura, 2004; Orphanoudakis, 2004; Poulymenopoulou and Vassilacopoulos, 2004; Schabetsberger, Gross, Haux, Lechleitner, Pellizzari, Schindelwig, Stark, Vogl and Wilhelmy, 2004; Snee and McCormick, 2004; Gerdsen, Müeller, Jablonski and Prokosch, 2005; Müller, Uckert, Bürkle and Prokosch, 2005; Orlova, Dunnagan, Finitzo, Higgins, Watkins, Tien and Beales, 2005; Blobel, Engel and Pharow, 2006; Clark, Müller, Gao, Lin, Lehmann, Thom, Inchingolo and Chen, 2006; Glaser and Lo, 2006; Harno and

Ruotsalainen, 2006; Knaup, Garde, Merzweiler, Graf, Schilling, Weber and Haux, 2006; Spidlen, Hanzlícek, Ríha and Zvárová, 2006).

The reasons of this big interest are multiple: the principal ones are the greater availability of heterogeneous sources of clinical data and the necessity to integrate such data inside of coherent and accurate patient profiles. Last but not least, the possibility to keep a day by day check on the evolution of patient's health status is seen as a necessary condition to achieve the continuity of care of the citizen in all its aspects, including follow up and home care, both for short and long periods.

The digitization process of all data concerning health care, the development of standards for e-health and the increase of computer education for all clinical personnel, are the fundamental requirements for the development of projects aimed to exploit new techniques in order to solve the problem of data fragmentation and supply new and effective health care services to the National Health Systems (NHS).

The INTESA project (Telematic Integration for the continuity of the citizen health care process) has recently used some new techniques in order to develop a new reference model of health data storage. INTESA was one of the greatest projects of the Italian strategic public programme named "New Medical Engineering", which was partially financed by the funds of the Italian ministry of research. This triennial project was coordinated by the Institute of Clinical Physiology of the National Research Council (CNR) of Pisa and involved five public universities, two research institutions, and three of the principal biomedical companies in Italy.

Following an approach based on the methods of basic research, the final purpose of the project was to develop a complete architecture of health information system, capable to guarantee a smart and safe storing of the essential information, an effective and personalized retrieval of data, and some innovative models to compare the results of clinical and medical activities of all the "actors" of the health care process (the citizen, the general practitioner, the specialist, etc.).

The fundamental results of the INTESA project concern the dynamic selection of a minimum data set of information necessary to most of the medical fields, and the modular infrastructure for health data communication that guarantees the storing and retrieval of such information, according to recognized medical and ICT standards like Health Level 7 (HL7), Clinical Document Architecture (CDA), Digital Imaging and COmmunication in Medicine (DICOM), etc.

Moreover, we investigated the usability of modern data mining methods to harmonize and classify the amount of available medical data concerning the cardiac pathologies, in order to purposely supply well-suited responses to the different "actors" enquiring the system through its telematic infrastructure. Finally, a pilot study has been performed as a first validation of the proposed architecture; this study was focused on the heart failure disease and included a specific investigation on the estimated socioeconomic benefits of the integration of the health care processes related to this pathology.

2 Overview of Health Information Systems and data warehousing

Currently available Health Information Systems (HIS) have been designed to assess basics needs in terms of data flow and monitoring from the peripheral to regional and national centralized government. They are at most developed and maintained in order to provide "production" information and are focused on identifying, e.g., the amount of

some kinds of procedures and health services provided to patients before or during hospital admission, to check, verify and monitor drug therapies and so on. These information are key information in the economic management of the hospital, or, more generally, for what is generally identified as the Enterprise Resource Planning (ERP). Starting from 1995 Italian health policy makers introduced rules and formalizations for patients check out (the SDO, a Patient Hospital check-out Form) now reporting the ICD9CM coding of diagnosis and Diagnosis Related Group coding. This was the first informative data flow from hospitals to government, very useful to monitor costs and health-care providers performances but poor in terms of clinical data. The patient health record, which contains all the clinical information about a patient admission at hospital, is in most cases a paper document, or a mix of paper and electronic data. Where the electronic health record is fully implemented, it happens often that patient data are spread across different ICT systems, provided by different vendors and under different responsibilities, making integration really difficult.

Today HIS are a mixture of relational database based applications, devoted to the collection of data in different areas of intervention, such as social records, emergency room, hospital admission, instrument diagnosis (e.g. medical imaging, laboratory analysis), pharmaceutics, rehabilitation, home-based assistance. There are, however, two main drawbacks. On one hand they are all "one-shot" systems: they can provide a really detailed view and exhaustive analysis based on data collection for a given clinical event, but they can not correlate it to the overall clinical history of the patient, including other more or less recent events. On the other hand, they may be really precise and accurate in collecting information about a pathology but they are not usually able to consider the correlated pathologies such as co-morbidity, making sometimes impossible to calculate costs due to chronic conditions affecting about 20% of the population, but absorbing more than 60% of total expenditure. The "history depth" richness currently available in nowadays HIS is an invaluable source of information and makes an opportunity for researches and stakeholders to analyze, evaluate and assess social and economics benefits, given adequate hypothesis.

Data warehousing is a fundamental tool to assess knowledge management in health care, since data is spread across heterogeneous and distant (along the three axes: space, technology, purpose/context) systems and data sources. Data warehousing can be also the preferred decision maker tool, since it allows deeper data analysis and cross-checks on costs and economic data, and a useful clinical analysis tools, since it allows to integrate clinical data, fit missing data, avoid errors (i.e. with a simple and fast direct check on patient allergies or clinical history against drugs to be prescribed during an emergency situation).

3 INTESA archive and its modules

The organizational and technical solutions provided by the INTESA project are here described.

3.1 Patient's centrality in INTESA

INTESA project aimed to integrate, sort and guard within a logically centralized archive, every citizen's health-related data generated by various providers over time, in order to

make them available to the very patient and the healthcare supervisors and health specialists (family physician, doctors, call centres, social services, researchers, etc.). This means that this archive must place the patient at the centre of the healthcare information.

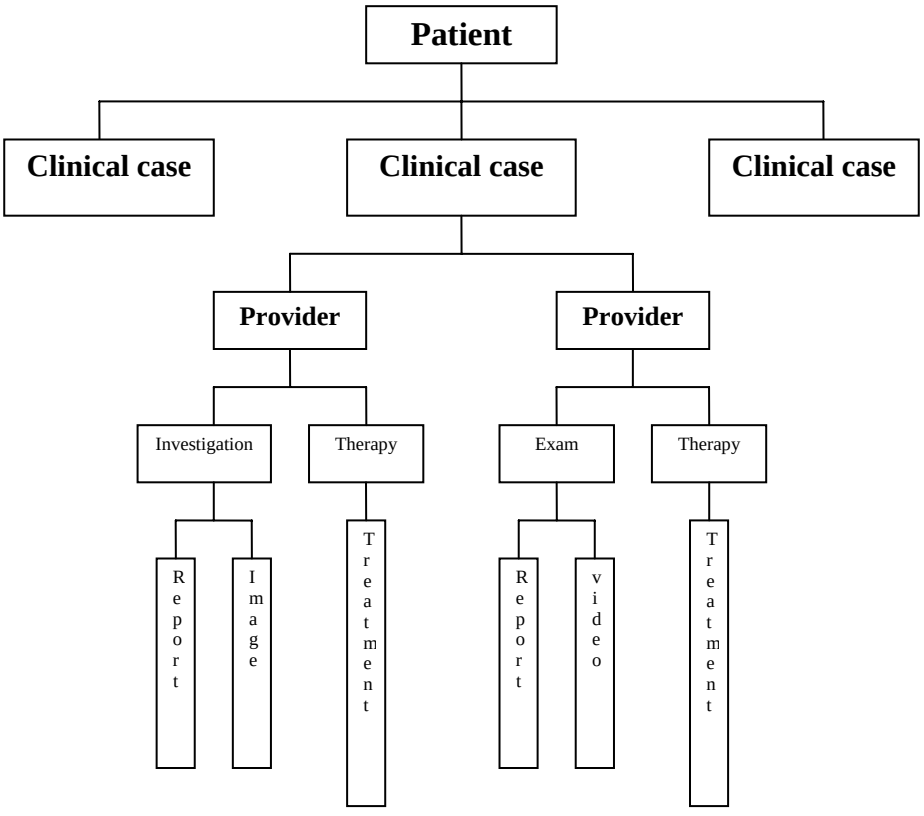
Moreover, the health information (e.g. charts, clinical data, prescriptions) is usually tied to a specific period of the patient, that is to the context that has carried the patient to start a certain number of exams, therapies, investigations, hospital admissions. This temporal period, in which one or more pathologies are dealt with, is identified as a "clinical case". During the treatment process, the patient presents to whom we can call "providers" of services (e.g. physician, laboratory, hospital, etc); however, also the patient's house could be a place from which the health-related information is collected.

Thus, in INTESA archive, named ARC, the health information has been organized on three levels:

1. patient;
2. clinical case;
3. providing organization.

This architecture is shown in Figure 1.

Figure 1 The architecture of health data in ARC



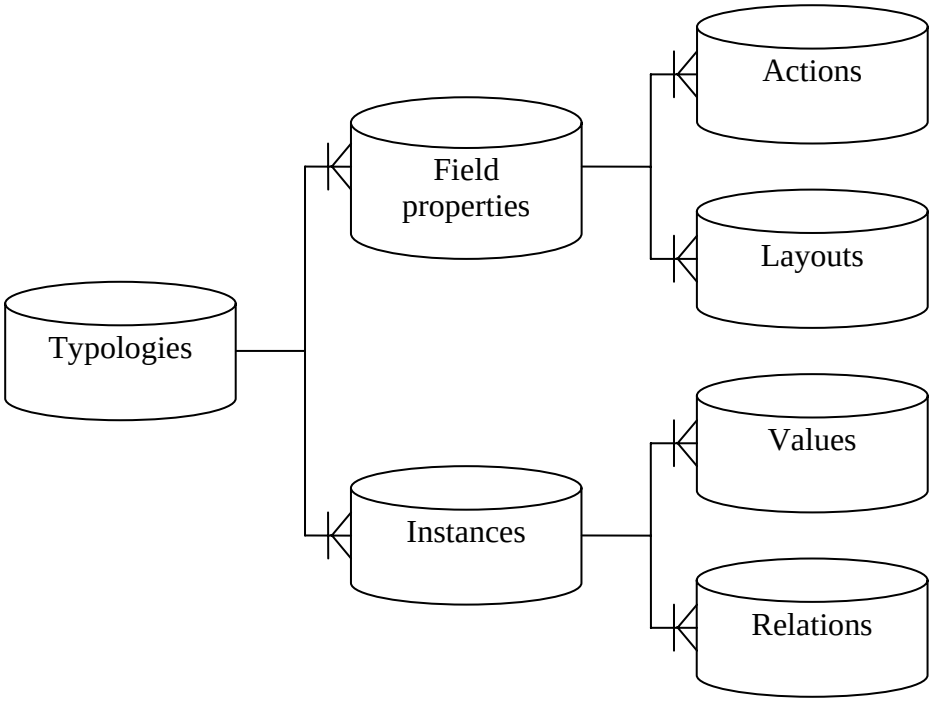
3.2 The dynamicity of ARC

As stated above, the aim of ARC is to collect, sort and guard in a dynamic data structure, all the clinical information of the patient. With the term dynamicity of the data structure, we refer to an organized data base built so that new groups of information can be registered, sorted and put in relation with those already existing, without taking part on the structure of the database or the interfaces of access, but acting through an appropriate interface of configuration for the system administrator.

The demand for dynamicity arises from the previewed increment of the typology of the data sources in time. Initially, it is foreseen that the systems feeding the archives will belong to few disciplines (one of which is the cardiological one of the pilot study), but other categories will be adding on in the future, creating the need of other fields in which to record the transmitted data.

The structure on which ARC is based follows a master-details logic, where some tables are used to define the property of the category, the property of the single fields that compose it, the relations between each other and the links with the clinical case/structure, while other tables are destined to contain the information. The relations between the tables are shown in Figure 2.

Figure 2 The relations between the tables of ARC



The content of each table is the following:

- Typologies: it contains the list of the categories of information (e.g. "electrocardiographic examination" or "laboratory data");
- Field properties: it describes the fields that belong to the typology (e.g. number, text, image);
- Actions: the validation activities that must be made at the arrival of new information. The actions can be tied to the single field or to all the record;
- Layouts: property that defines the visualization of the typology;
- Instances: the main properties for every new archived piece of information pertaining to the category (provider, date of insertion);
- Values: the values assumed by the fields of the instance;
- Relations: the ties of the instance to the patient/case/provider.

3.3 Interfaces, standards and technologies used in ARC

According to the target of the project, ARC must be able to interact with various peripheral subsystems of access, which will try to interact with ARC through different communication protocols. For this reason, we have developed a software platform based on a MOM system (that is, a Message Oriented Middleware) where the information in transit from an application to an other, travels inside messages. In the MOM, the integration logic is separated from the applicative one, and is remapped and remodelled in the terms of the elements constituting the infrastructure of messaging, that are:

- Messages: contain the information to be integrated;
- Queues of messages: contain the messages on hold;
- Message routing: allows the transit of the information from one application to the others.

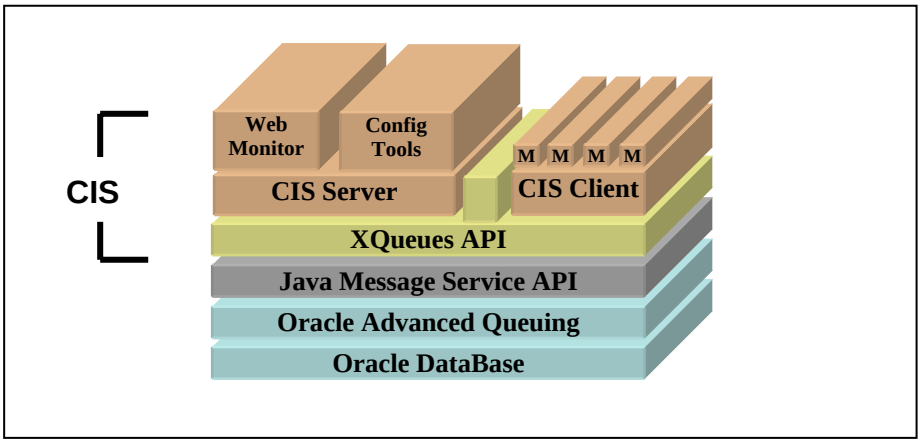
The content of a message can be transformed before its delivery, so as to be able to adapt its format to that one of the receiving application; this is one of the main tasks of the Clinical Integration Service (CIS). The fundamental aspect of CIS is the ability to reproduce and to implement whichever workflow used inside of the context in which it has to operate. Such ability arises from the architecture with which it is constructed: with interoperating modules, everyone of which implements the minimal entirety of functionality, but highly flexible and configurable. Configuring and appropriately connecting such "blocks of base", it is possible to model and to recreate the operations of whichever process. CIS server and CIS client modules are the higher applicative levels of this stratified system; the various levels are shown in Figure 3.

The infrastructure of messaging on which CIS is based, is represented by Oracle Advanced Queuing, to which the necessary operations to the memorization of the messages, the management of the queues and the exchange of the messages between the applications are all delegated. Oracle Advanced Queuing puts on hand of CIS all its characteristics of emergency, performances, stability and integrity of the data, absolutely indispensable in this type of applications.

XQueues is the Application Programming Interface (API) through which CIS composes and uses the instruments offered by Oracle Advanced Queuing, giving place to the architecture that characterizes it, based on the expandability and the modularity. XQueues API is written in Java language and is based on Java Message Service (JMS) standard. That, therefore, renders CIS independent from the infrastructure below and

allows anyone needing to use it to be able to take advantage of the already acquired know-how with the acquaintance of standard Java API.

Figure 3 The functional levels of ARC



Note: M = Module (gateway)

3.4 CIS Server and CIS Client modules

In CIS, the higher applicative level is constituted by CIS Server and CIS Client, which are contituted by instruments of configuration, server applications and applicative modules.

In detail, CIS Server is composed by:

- Instruments of management and configuration;
- CIS Message Router.

CIS Client is composed by disjointed and interoperating modules, distributable also on various servers. In order to maintain an unified control of this scenario, CIS Server has a registry, that is populated at the moment of the start of every module, containing the list of the modules configured and qualified to the operations, and a protocol of remote management of the modules, based on the use of Remote Method Invocation and Internet Inter-ORB Protocol (RMI/IIOP).

In particular, the Gateways represent the speciality portals of access to CIS; that is, they are its interfaces with the external world. Through the Gateways, CIS obtains the access to devices, servers, systems of communication, and systems of recording, and interprets their logic of operation, remapping it in its own architectural and operational model.

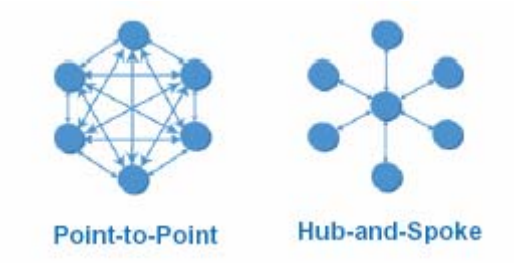
Every application integrated via CIS, communicates with it through dedicated CIS-Gateways. The communication can occur in two ways:

- Synchronous: preferable from the point of view of the performances and the general reliability of the system, but not always possible;
- Asynchronous: to be used when the synchronous modality is not applicable.

3.4.1 Message routing

From the point of view of the distribution of the messages, the structural model to which CIS makes reference is the so-called "Hub and Spoke", that previews the existence of a main centre (hub) and of peripheral centres (spokes). A comparison between Point-to-Point and Hub-and-Spoke models is shown in Figure 4.

Figure 4 Communication models: Point-to-Point versus Hub-and-Spoke



While the role of the Spokes is covered by several modules, the role of Hub is carried out by a centralized system of identification, shunting and routing all the messages, called exactly "Message Router".

The routing of the messages towards the peripheral modules is based on rules of routing defined in phase of configuration. The location of the destination is carried out both based on the source of origin of the message and based on the content. For this purpose, CIS Router has Content Managers, in a position to identify the format of the informative content of a message (MIME type) and to manage it.

3.4.2 HL7 Gateway

Through this HL7 adapter, CIS is in a position to communicate with systems that use this standard, implementing different profiles of communication.

The HL7-Gateway realizes both the synchronous communication functionality, implementing the MLLP (Minimal Lower Level Protocol) based on TCP/IP connection, as specified in the definition of the standard, and the asynchronous one, through the management of various physical supports (e.g. file system, RDBMS) as systems of persistence and spooling of the messages. CIS, with this gateway, acquires therefore the complete range of functionalities within HL7 connectivity.

3.4.3 DICOM Gateway

DICOM gateway allows CIS to connect to RIS/PACS (Radiology Information Systems/Picture Archiving and Communication Systems) or to diagnostic devices and equipments that support and implement DICOM protocol and to exchange information with them.

Through DICOM connectivity, CIS is in a position to manage the transfer of personal and clinical information, every daily work list, images or complete series of images and studies. Moreover, through the association of the patient/case coordinates related to the different systems, it allows to specialized applications to retrieve the data and the images

related to a specific patient and to carry out the necessary elaborations, like, as an example, the visualization of images for diagnostic purposes by a whichever DICOM viewer.

3.4.4 FileSystem Gateway

FileSystem Gateway allows CIS of interact with whichever file system, managing the encapsulation of the content of the files inside of messages and vice versa. FileSystem Gateway is in a position to feed CIS through files extracted in an independent way from various paths of the file system, without limits of format and dimension.

The pathname with which the files will be identified could be constructed beginning from static values defined in the configuration or through expressions estimated at runtime and based on the content of the file.

Moreover, the pathname can derive from the format of the file, identified and managed through the Content Manager system described in the paragraph about the Message Router. FileSystem Gateway turns out extremely useful both for the realization of simple profiles of integration and for the creation of printouts or log files.

3.4.5 XML Translation Engine

XML Translation Engine (XEngine) represents the heart of CIS, because transforms XML documents from a structure to another, allowing the adaptations and the conversions of format that are the base of every operation of integration.

Through a system of rules opportunely defined during the configuration, XEngine is in a position to identify XML documents in input and apply the right style sheet of transformation. The style sheets of XEngine can contain calls to Java functions contained inside of XQueues API, through which accomplish operations of adaptation of format or conversion of type.

Moreover, XEngine can merge and split XML documents, that is the fusion of several documents in a unique file and the production of various files from one file in input. XEngine possesses predefined style sheets for the particular production of text files (e.g. HTML, HL7) with a standardized semantic.

3.4.6 SQL Translation Engine

SQL Translation Engine (SQL2XML), developed in PLSQL (Procedural Language/Structured Query Language), allows to execute whichever SQL command on Oracle databases and to give back to CIS a XML message containing the result of such operation. The adopted technique is that one to show to CIS database the tables from which to capture (SELECT) or modify (INSERT, UPDATE, DELETE) the records. In case the remote database is on Oracle platform, the technique adopted is to create links to the source database.

Moreover, the instructions to be executed can be often archived in CIS and called passing some parameters. This module is able also to execute procedures and functions predefined "ad hoc". If the source database is not on Oracle platform, the communication is however possible acquiring the Oracle Transparent Gateway licences or activating the Heterogeneous Service - generic connectivity procedure using ODBC (Open Database Connectivity).

3.4.7 Text Translation Engine

Text Translation Engine loads data contained inside of files or read from byte stream, directly inside of XML files whose structure can entirely be specified during the configuration. Taking advantage of the rules and the parameters defined during this phase, Text Translation Engine is able to create style sheets (XSL) through which to carry out the transformation in order to create a new XML file starting from the old file.

The operation is based on the conversion in structured format of the data contained in the files through the identification of records and logical fields. The values therefore identified can be used as keys of identification of the logical records or converted and used in conditional expressions or in functions. These functionalities allow Text Translation Engine to work both in an interactive way, that is to create XML structures on direct demand of the user, and in batch way, that is, transforming the identified logical records in opportune XML documents based on rules of association defined at the moment of the configuration.

The parsing engine imposes negligible limitations to the format of the files treatable, and through an intuitive graphical interface it allows a simple definition of the rules of mapping Text to XML; moreover, it allows to:

- load in an unique session the data contained in various files;
- load in an unique session the data of various XML documents;
- specify the character set with which the data files in input are codified;
- load the data selectively.

3.4.8 DICOM Translation Engine

Dicom Translation Engine works in collaboration with the analogous gateway and allows to create XML documents starting from DICOM dataset and vice versa. The structure of the created documents reflects the Object Oriented nature and the recursive models that are the base of the representation of the information which DICOM standard uses (Information Object Definition, Service Object Pairs), and thus enable the transformation of whichever possible dataset, since it is based on a architectural approach rather than on the definition of static mapping rules. Therefore, this module can be employed also to carry out queries to systems that support DICOM standard, in order to respond to requests and in order to transport information.

The conversion/mapping of the typologies of data and the adaptation of the contents based on the different semantic rules is managed automatically, referring to models defined in the standards and by now universally recognized as reference, like Dicom Structured Reporting and Clinical Document Architecture (HL7-CDA).

Given the enormous size of data that normally composes DICOM studies or series, Dicom Translation Engine operates a first separation between the structured data and the data that constitutes the images, the videos or whichever other type of non-textual information, recording it in different archives, but maintaining explicit inside of a XML document the coordinates necessary to their retrieval.

4 Data mining approaches to support clinical decisions

Besides of providing an infrastructure for handling clinical data, data mining approaches have been experimented within the INTESA Project. Such experiments were aimed at

verifying which methods can be considered more effective in supporting the clinical decisions within personalized applications. Principal Component Analysis (PCA), Multiple Correspondence Analysis (MCA), Bayesian Networks (BN) and back propagation Neural Networks (NNs) have been tested and compared.

The first three approaches have been tested on a dataset coming from the UCI Machine Learning Repository, named Cleveland, and available at the URL http://www.ics.ici.edu/mlearn/. The dataset contains 297 instances without missing values.

It is to notice that clinical data – as well as Cleveland data – are in general quite heterogeneous. For heterogeneous data is here meant data that come from different sources and are expressed on different scales, and therefore needs appropriate techniques to be treated.

PCA (Jackson, 1991; Jolliffe, 1986) is a well known statistical technique. It consists of finding a basis that maximizes the total variance of data on which data are projected. That basis is found usually by solving an Eigenvalue or Singular Value (SVD) problem (Kalman, 1996). From this projection a subset of dimension – at most equal to the rank of the matrix – is subsequently extracted for low dimensional representation. These projected data can be then clustered or classified. Because of heterogeneity of data, different vector norms (1-norm, 2-norm, infinity norm) for normalizing and classifying (Perlibakas, 2004) data have been tested; moreover, the impact of the coding format of data has been investigated.

MCA (Greenacre, 1984; Lebart, Morineau and Warwick, 1984) is also a well known multivariate technique, and is usually used for graphical data exploration and for data dimensionality reduction. It consists of a projection of data on a basis such that the whole projection error is minimized together in columns and in rows space (Michaidis and De Leeuw, 1998), according to chi-square metrics (measuring the degree of association between values of two or more variables).

BN (Abrams and Myles, 2004) are well known techniques for discovering associations and dependencies between data and for classification. BN employ the bayesian model of inference to represent knowledge on a given domain by means of directed acyclic graphs (Ghahramani, 2001). In such a graph, each node represents a variable while arcs represent the statistical dependencies between variables. Statistical dependencies to be investigated can be both inferred from data and be defined as a priori knowledge, and subsequently refined by data analysis. Here the mutual information metrics for estimating dependencies has been used (Chen, Bell and Liu, 2002). It gives a measures of the informative entropy of each co-occurrence of values of two variables.

A Neural Network (Haykin, 1994) is a mathematical model or computational model based on Biological NNs and consists of an interconnected group of artificial neurons. In most cases, a Neural Network is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. The data base used to train and test the performances of NNs was constituted by two parts: 100 records of patients with heart failure, described by 50 variables, and 50 records of patients without heart failure, described by 22 variables. For the type of data proposed, we experimented a supervised neural network, that is a network trained through the use of examples (patients) classified correctly and on the base of these classifications, it has to try to classify some new patients. The training of the neural network was based on a back-propagation algorithm (Mitchell, 1997) and the adopted mathematical model was parametric, for the type of data normalization, for the activation functions of the single

neurons, for the stop criteria of the training, for the update parameters of the weights and for the topology of the network.

The performances of these methods have been as follows:

- *Multivariate methods (PCA & MCA):* For PCA, the results show that both the chosen norm, and the coding format impact on the performances. The best performances have been obtained by using 1-norm and avoiding to use zero as coding values. In such a case, the classes according to which individuals are classified – especially intermediate ones- are more separable, and more evident polarizations appear. Moreover, a more clear interpretation of data can be done. Eventually, the models drawn using different norms tend to recognize the same classes and the same discriminant values. For MCA, a good performance for graphical data exploration can be appreciated.
- *Bayesian Networks:* BN have shown to be effective for classification and clinical decision support. In fact, a percentage of 84,9% of right classification has been obtained by using ten folds cross validation method. Therefore, it can be concluded that BN are robust for classification and able to assist in defining a good model for supporting clinical decisions.
- *Neural Networks:* The target to achieve consisted in classifying some patients through a value comprised in a fixed interval; that value should indicate how much the patient was affected by heart failure. Although the number of records was very small, the values of some internal parameters have allowed the neural network to converge and to classify the new patients in an acceptable way for the clinicians.

Although the health data were quite heterogeneous, the results obtained by data mining methods have been satisfactory, thus a wide scale exploitation of such methods can be foretold for both short and long clinical monitoring in order to support clinical decisions in every moment of the health care process.

5 Final assessment and conclusions

The developed archive, being connectable to the information systems of every hospital and clinic, is an instrument able to integrate single medical activities into health care processes associated to the patients.

In the pilot experiment, the applications involved in the integration process were a Radiology Information System/Picture Archiving and Communication System (RIS/PACS), a Hospital Information System (HIS) and a private application that managed data of patients with heart failure. Thanks to the integration infrastructure, these three systems were able to exchange the personal data of the patients, in order to maintain themselves "aligned" to each other.

After every visit of a patient with heart failure in the doctor's office, the new data were collected by the integration system that elaborated them, archived them in ARC and then sent them to the HIS too. In the same way, the RIS/PACS sent to the integration system every clinical report and the references to the images associated to it. Once this information had been archived in ARC, the clinical report in pdf format and the

references to the images were supplied to the HIS, so that they could be visualized directly by the HIS through a web browser.

The technical and organizational solutions provided by the INTESA project are designed in order to optimize both available resources and personnel usage, since they avoid data loss and overlapping, and make easy data integration for data-mining and data-warehousing analysis.

With tools like ARC, it will be possible to analyze overall patient health-care history, monitor the health services quality, cross-check care costs with patient fiscal status, insurance positions and social benefits to avoid frauds, monitor resource usage and plan resource improvement and re-location on the territory.

It is clear that benefits for patients, health-care personnel and policy makers are all relevant: costs can be defined for a given pathology, a given clinical event or a given clinical pathway, allowing to define and assess/validate new protocols.

In particular, the proposed methodology addresses a chronic disease (heart failure) that represents a huge burden of ill health all over the world and in particular in Italy, due to the aging of population, and therefore a large cost to the NHS. Costs are uncertain because for many years government policy has focused on acute care and still has no agreed model for managing all chronic diseases, so differences in treatment exist between citizens.

In detail, the proposed methodology will allow:

- To estimate health care costs for heart failure patients;
- To identify different health care consumption patterns, if any, by age and gender;
- To provide information for developing disease management guidelines;
- To explore potentials of record linkage of administrative and demographic databases.

Consequently, more efficiency would be attained with the same resources, because diffuse information can stimulate each stakeholder in utilizing better community and individual resources. In particular, it would be possible to assess if clinical care is consistent with scientific evidence and to evaluate benefit of alternative clinical pathways. It is also an essential tool to support manager decisions such planning staff and organizational requirements.

From a societal point of view, it will be possible to check for appropriateness, focusing on patient's needs instead on vendors expectations. Besides, the proposed methodology is very cheap, because it is based on administrative data and can be implemented with continuity without requiring any particular application.

Together with other metropolitan repositories based on HL7 messages and applications able to examine the data stored, the developed archive will contribute to keep a check on every citizen's health history, clinical examinations and cure therapies, but, above all, it will allow to verify the efficacy and efficiency of the health care processes related to particular pathologies.

Acknowledgements

We thank Stefano Polacci, Stefano Dalmiani and all those who contributed to the INTESA project. We also thank Manuella Walker for her assistance in preparing this paper.

References

Abrams, K., and Myles, J. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley, New York.

Blobel, B., Engel, K. and Pharow, P. (2006) 'Semantic interoperability--HL7 Version 3 compared to advanced architecture standards', *Methods of Information in Medicine*, Vol. 45, No. 4, pp.343-353.

Chen, J., Bell, D., and Liu, W. (2002) 'Learning Bayesian Networks from data: An Efficient Approach Based on Information Theory', *Artificial Intelligence*, Vol. 137, No. 1-2, pp.43-90.

Clark, J., Müller, H., Gao, X., Lin, Q., Lehmann, T.M., Thom, S., Inchingolo, P. and Chen, J.C. (2006) 'Medical imaging and telemedicine - from medical data production, to processing, storing, and sharing: a short outlook', *Computerized Medical Imaging and Graphics*, Vol. 30, No. 6-7, pp.329-331.

Gerdsen, F., Müeller, S., Jablonski, S. and Prokosch, H.U. (2005) 'Standardized exchange of medical data between a research database, an electronic patient record and an electronic health record using CDA/SCIPHOX', *AMIA Annual Symposium Proceedings*, pp.963.

Ghahramani, Z. (2001) 'An Introduction to Hidden Markov Models and Bayesian Networks', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 1, pp.9-42.

Glaser, J.P. and Lo, H.G. (2006) 'Concepts for building inter-organizational systems in healthcare: lessons from other industries', *Journal of Healthcare Information Management*, Vol. 20, No. 3, pp.54-62.

Greenacre, M. J. (1984) *Theory and Application of Correspondence Analysis*, Academic Press, New York.

Haykin, S. (1994) *Neural Networks, A Comprehensive Foundation*, IEEE Press, New York.

Hanzlicek, P., Spidlen, J. and Nagy, M. (2004) ' Universal electronic health record MUDR', *Studies in Health Technology and Informatics*, Vol. 105, pp.190-201.

Harno, K. and Ruotsalainen, P. (2006) ' Sharable EHR systems in Finland', *Studies in Health Technology and Informatics*, Vol. 121, pp.364-370.

Jackson, J. E. (1991) *A user's guide to Principal Components*, Wiley, New York.

Jolliffe, I. T. (1986) *Principal Component Analysis*, Springer Verlag, New York.

Kalman, D. (1996) 'A Singularly Valuable Decomposition: The SVD of a Matrix', *College Mathematics Journal*, Vol. 27, No. 1, pp.2-23.

Knaup, P., Garde, S., Merzweiler, A., Graf, N., Schilling, F., Weber, R. and Haux, R. (2006) 'Towards shared patient records: an architecture for using routine data for nationwide research', *International Journal of Medical Informatics*, Vol. 75, No. 3-4, pp.191-200.

Lebart, L., Morineau, A., and Warwick, K.M. (1984) *Multivariate Descriptive Statistical Analysis*, Wiley, New York.

Michaidis, G., and De Leeuw, J. (1998) 'The Gifi system of Descriptive Multivariate Analysis', *Statistical Science*, Vol. 13, No. 4, pp.307-336.

Mitchell, T. M. (1997) *Machine learning*, McGraw-Hill, Boston.

Müller, M.L., Uckert, F., Bürkle, T. and Prokosch, H.U. (2005) 'Cross-institutional data exchange using the clinical document architecture (CDA)', *International Journal of Medical Informatics*, Vol. 74, No. 2-4, pp.245-256.

Nardon, F.B. and Moura, L.A. (2004) 'Knowledge sharing and information integration in healthcare using ontologies and deductive databases', *Medinfo*, Vol. 11, No. 1, pp.62-66.

Orlova, A.O., Dunnagan, M., Finitzo, T., Higgins, M., Watkins, T., Tien, A. and Beales, S. (2005) 'Electronic health record - public health (EHR-PH) system prototype for interoperability in 21st century healthcare systems', *AMIA Annual Symposium Proceedings*, pp.575-579.

Orphanoudakis, S. (2004) 'HYGEIAnet: the integrated regional health information network of Crete', *Studies in Health Technology and Informatics*, Vol. 100, pp.66-78.

Perlibakas, V. (2004) 'Distance measures for PCA based face recognition', *Pattern Recognition Letters*, Vol. 25, No. 6, pp.711-724.

Poulymenopoulou, M. and Vassilacopoulos, G. (2004) 'An electronic patient record implementation using clinical document architecture', *Studies in Health Technology and Informatics*, Vol. 103, pp.50-57.

Schabetsberger, T., Gross, E., Haux, R., Lechleitner, G., Pellizzari, T., Schindelwig, K., Stark, C., Vogl, R. and Wilhelmy, I. (2004) 'Approaches towards a regional, shared electronic patient record for health care facilities of different health care organizations--IT-strategy and first results', *Medinfo*, Vol. 11, No. 2, pp.979-982.

Snee, N.L. and McCormick, K.A. (2004) 'The case for integrating public health informatics networks', *IEEE Engineering in Medicine and Biology Magazine*, Vol. 23, No. 1, pp.81-88.

Spidlen, J., Hanzlícek, P., Ríha, A. and Zvárová, J. (2006) 'Flexible information storage in MUDR(II) EHR', *International Journal of Medical Informatics*, Vol. 75, No. 3-4, pp.201-208.

Yoo, S., Kim, B., Park, H., Choi, J. and Chun, J. (2003) 'Realization of real-time clinical data integration using advanced database technology", *AMIA Annual Symposium Proceedings*, pp.738-742.