A Grid-enabled web platform for integrated digital biobanking in paediatrics

Massimiliano Izzo[1]✉, Andrea Schenone[1], Sara Barzaghi[2], Fabiola Blengio[2], Marco M Fato[1], Luigi Varesio[2]

[1]Biolab, Department of Informatics Bioengineering Robotics and System Engineering, University of Genoa, Genoa, Italy
[2]Molecular Biology laboratory, Giannina Gaslini Institute, Genoa, Italy

Motivation and Objectives

A solid and integrated biobanking framework is an absolute requirement for high quality investigation in paediatric tumours. The overall goal of our activity is to design and develop a centralized Digital Biobank prototype able to integrate and interconnect an increasing number of local biobanks situated in various centres across Europe. As a first step, we are designing a web-based repository to store all tissue and genomic data from paediatric tumours collected by the G. Gaslini Children's Hospital, in Genoa. The repository satisfies flexibility and extensibility criteria, and is being deployed on a data Grid architecture (Bote-Lorenzo et al., 2004).

Methods

The repository is designed to contain data from all the tissue and blood samples obtained from infants and children affected by paediatric tumours, such as primary bone tumour and neuroblastoma. Many samples may be extracted from the same patient in a single visit or surgical operation; moreover from a single sample, nucleic acids (i.e. DNA and RNA) may be extracted for further analysis. These extractions could happen more than once, even at a distance of months or even years, if required.

In order to satisfy the strict requirements above and ensure the extensibility of the repository, we have adopted a process/event model, already used for designing data and image repositories in Neuroscience (Corradi et al., 2012). The process/event model is a multipurpose taxonomic schema composed by two main generic objects: processes and events. An event can be any 'atomic' operation that is performed on patients or samples, or any processing of data or everything else related to the repository administration and management. A process is defined as a group of sequential events or sub-processes related to an activity, allowing the creation of a sort of hierarchical structure. As an example, the storage of a DNA sample in a specified location within a -80°C freezer and a post-processing step (such as differential expression, survival or correlation/anti-correlation analysis on microarray data) are single events, pertaining respectively to the more general 'Nucleic Acid Extraction' and 'Data Mining' processes.

Platform Architecture

The repository has a client-server architecture and it is composed by three main components, as shown in Figure 1:

- Repository portal
- Database
- Grid storage

The repository portal is designed to make the storage and the navigation of data and information easy, through a simple and transparent web interface. It is a Java Enterprise Edition web application based on several existing open source tools for the development of web applications. The basis of the portal consists in a framework that relies on an Apache Tomcat web application container. It incorporates a database interface layer built through MyBatis, a persistence framework that automates the mapping between SQL databases and objects in Java. To provide users with highly interactive interfaces, some components are designed using the Asynchronous Javascript and XML (AJAX) programming technique. Wherever possible information is exchanged in XML or JavaScript Object Notation (JSON) format. The web portal represents the main access point to all the functionalities available through the overall integration platform, and exposes both user and administrator interfaces.

The repository itself is based on a MySQL database. The database design is fundamental in order to make the repository highly flexible and easily extensible. The core of the database is formed by the two previously described entities: processes and events and their relationships to data and metadata. Existing processes and events are contained in two homonymous tables. Each element in the event table refers to an element in the data table. The information

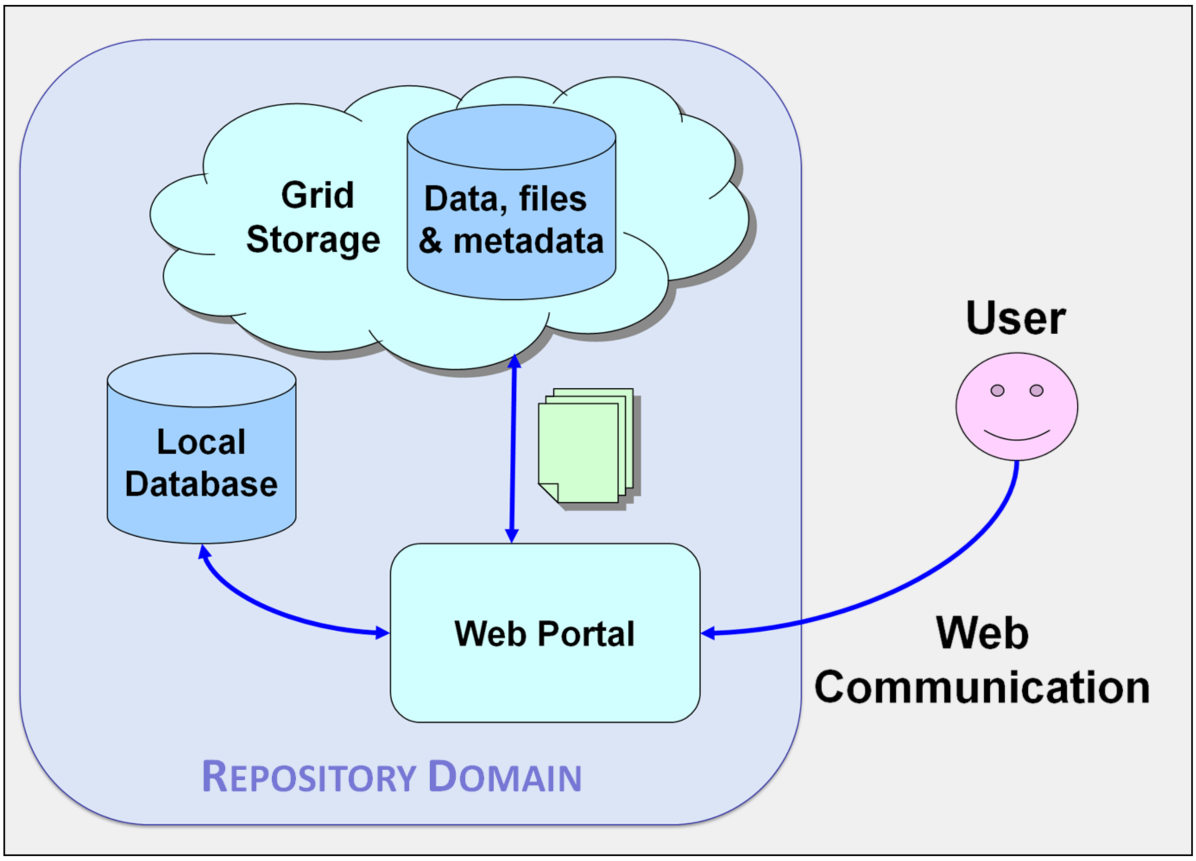

Figure 1. Repository overall architecture

inside the latter represents all the data inserted in the repository. These data can be associated with one or more files accordingly to their data type. The file table contains the logical path of all the stored files. The repository can be configured to store the metadata totally or partially within the database. In this latter case, the metadata are stored as XML descriptions inside the data table, to display the data in a rapid and dynamic way using XSL Transformations, and as records of specific metadata tables, to perform complex queries in an easier way. All data files are contained in the Grid storage, so the database doesn't really have to deal with hundreds of GB of data. Moreover, the number of operators should be quite small, thus making MySQL a reasonable choice as a database. The storage subsystem has been built around the iRODS data grid software (Rajasketar et al., 2010), chosen among others because it allows building a federated and distributed data storage system without the need of central components. Being able to deal with a huge amount of metadata, iRODS is widely used by the research community, also for Next Generation Sequencing Projects (Chiang et al., 2011).

Careful attention has been given to security and privacy issues. All data are anonymised and cannot be linked in any way to patients' names, since the connection between clinical and personal data is done using unique identifiers managed exclusively by clinicians. Administrators are able to control users' access by creating groups and their association with pages and functions, define processes, events and all their relationships, define new data types and related metadata, associate them with the related events and manage available ontologies. Normal users, according to their assigned permissions, can insert new data, retrieve patients' information and view all the related data, download stored information, explore processes together with all the related events, data and metadata to have a global picture.

The integrated system we envision at a European level will take advantage of the data Grid features provided by iRODS. Each hospital or biobank involved in the virtual community may have a local database and a dedicated separated iRODS system (called iRODS zone) where its own metadata and files can be saved. All the iRODS zones in the community will be federated. Federated iRODS zones are administered separately, but the users in the multiple zones, if given permission, will be able to access data stored in the other zones. If more hospital or research groups are working on the same project or using the same data structure, they may share a single iRODS zone and database. To provide access to the various local databases, federated database systems will be taken into account.

Results and Discussion

A first prototype of the repository is currently being tested at the Giannina Gaslini Institute, in Genoa. Information on over 1300 tissue samples, with their related DNA and RNA purified samples, have been stored together with administrative and clinical data from more than 700 patients. Three kinds of genomic analyses (i.e. event types) are currently provided, two for DNA samples - Comparative Genomic Hybridization (CGH) array and Multiplex Ligation-dependent Probe Amplification (MLPA) - and one for RNA - microarray analysis. For each analysis it is possible to store one or more files and user customized metadata. New data types can be configured via administrator interface, without additional programming, when new types of analyses or processing are required. The extensibility of our data model with user-defined data types and metadata is a crucial aspect of our implementation.

As mentioned before, future developments will comprise the integration of our local biobank at the Gaslini Institute, with similar digital structures located across Europe. We are currently testing a distributed storage configuration, implementing data management policies expressed as rules that are interpreted by the iRODS Rule Engine.

Acknowledgements

Our research activity is performed in the framework of the 'European Network for Cancer Research in Children and Adolescents' (ENCCA) European project.

References

1. Bote-Lorenzo ML, Dimitriadis YA and Gomez-Sanchez E (2004) Grid characteristics and uses: a grid definition, Proceedings of the First European Across Grids Conference, ACG'03, Springer-Verlag, LNCS 2970, 291-298. doi:10.1007/978-3-540-24689-3_36
2. Chiang GT, Clapham P, Qi G, Sale K and Coates G (2011) Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute BMC Bioinformatics 2011, 12:361. doi:10.1186/1471-2105-12-361
3. Corradi L, Porro I, Schenone A, Momeni P, Ferrari , Nobili F, Ferrara M, Arnulfo G and Fato MM (2012) A repository based on a dynamically extensible data model supporting multidisciplinary research in neuroscience, BMC Medical Informatics and Decision Making (in press).
4. JSON (JavaScript Object Notation), [online], http://www.json.org/.
5. MyBatis, [online], http://www.mybatis.org.
6. Rajasketar A, Moore R, Hou C et al. (2010) iRODS Primer: Integrated Rule-Oriented Data Systems. Morgan & Claypool. doi:10.2200/S00233ED1V01Y200912ICR012
7. XSL Transformations [online], http://www.w3.org/TR/xslt.