BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 21 2010, pages 2798–2800
doi:10.1093/bioinformatics/btq502

Databases and ontologies

Advance Access publication August 31, 2010

A sample storage management system for biobanks

C. Voegele[1,*], L. Alteyrac[2], E. Caboux[3], M. Smans[2], F. Lesueur[1], F. Le Calvez-Kelm[1] and P. Hainaut[4]

[1]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer (IARC), [2]Information Technology Services, IARC, [3]Laboratory Services and Biobank Group, IARC and [4]Molecular Carcinogenesis Group, IARC, Lyon, France

Associate Editor: Alex Bateman

ABSTRACT

Summary: Establishment of large-scale biobanks of human specimens is essential to conduct molecular pathological or epidemiological studies. This requires automation of procedures for specimen cataloguing and tracking through complex analytical processes. The International Agency for Research on Cancer (IARC) develops a large portfolio of studies broadly aimed at cancer prevention and including cohort, case–control and case-only studies in various parts of the world. This diversity of study designs, structure, annotations and specimen collections is extremely difficult to accommodate into a single sample management system (SMS). Current commercial or academic SMS are often restricted to a few sample types and tailored to a limited number of analytic workflows [Voegele *et al.* (2007) A laboratory information management system (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening. *Bioinformatics*, **23,** 2504–2506].

Thus, we developed a system based on a three-tier architecture and relying on an Oracle database and an Oracle Forms web application. Data are imported through forms or csv files, and information retrieval is enabled via multi-criteria queries that can generate different types of reports including tables, Excel files, trees, pictures and graphs. The system is easy to install, flexible, expandable and implemented with a high degree of data security and confidentiality. Both the database and the interface have been modeled to be compatible with and adaptable to almost all types of biobanks.

Availability and implementation: The SMS source codes, which are under the GNU General Public License, and supplementary data are freely available at 'http://www-gcs.iarc.fr/sms.php'

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: voegele@iarc.fr

Received on July 23, 2010; revised on August 24, 2010; accepted on August 26, 2010

1 INTRODUCTION: SCOPE OF SINGLE SAMPLE MANAGEMENT SYSTEM

The International Agency for Research on Cancer (IARC) Biological Resource Center hosts about 1 million samples from over 50 different studies. It contains both population-based and disease-based collections consisting of samples of various nature and origins.

The first requirement of single sample management system (SMS) is to harness the inherent variability and heterogeneity in the types of storage units and in the storage levels. At the inception of the project, a review of collections and facilities was conducted, providing a basis for defining extensive specifications incorporating the whole range of IARC sample diversity. The storage infrastructure is made up of three main categories of containers: liquid nitrogen tanks, freezers and fridges and dedicated humidity-controlled rooms, at various temperatures and each including several subtypes of containers. A comprehensive and strict hierarchy from single tube up to room defines the precise position of each sample location within the system. The exact capacity of each storage level is determined in order to monitor unused storage capacity. These hierarchies are relatively stable but the system has the capacity to evolve by incorporating new hierarchical levels, new storage devices within each level as well as movements and transfer of samples within each hierarchical level (See Supplementary Material: SD-1).

The second requirement for SMS is to keep track of all samples or containers movements and status changes, including the management of dynamic processes from specimen retrieval to extraction and aliquoting of by-products and transfer to in-house analytical platforms as well as shipment to other research centers.

The third requirement for SMS is to allow for simple and rapid import of existing information from various databases, documents and sheets and to accommodate the full range of information available for each particular specimen collection. In this way, SMS does not replace the study-specific databases developed by clinicians and epidemiologists but interfaces with them to provide a dynamic sample and data management system.

Finally, the SMS needs to comply with levels of data safety and confidentiality compatible with the high ethical standards of research conducted on human biospecimens.

In this article, we describe how our SMS addresses all these requirements, and how it operates to handle large datasets through a simple and user-friendly interface.

2 RESULTS: DEVELOPMENT OF SMS

The SMS is based on a three-tier architecture with one Linux database server, one Windows 2003 application server and multiple clients (either PC or Mac) integrated into the Agency's internal network, secured within a strict firewall.

2.1 Database architecture

Given the size of the datasets, the relational database was developed under Oracle 10g R2, which can manage a high volume of data and transactions, providing high query performances. It includes

[*]To whom correspondence should be addressed.

powerful tools to ensure safety and reliability with a high degree of recoverability and audit.

Taking into account the needs and specifications laid out before, the database model was designed to provide large flexibility to accommodate evolution of storage facilities. The model of the IARC biobank includes major tables identifying uniquely projects, samples, aliquots, containers, container types, sample movements and container movements (see model in Supplementary Material: SD-2).

2.2 Web application (interface)

The database is managed through a web-enabled graphical user interface developed with Oracle Forms Builder and using Oracle JInitiator plug-in.

The interface, platform independent, is compatible with most of the current web browsers. The configuration of the interface is user- and group of user-dependent: it is dynamically and automatically adapted to the user depending on his role and permissions.

The system enables the monitoring of the samples and their position in the hierarchy of containers as well as of the history of the movements from retrieval to extraction, aliquoting or shipment to in-house or external analytical platforms. The main features are as follows:

- Data insertion through forms or pre-formatted csv files that are checked by a PL/SQL procedure that surveys both data format and data integrity with respect to the requirements of the database. Data are parsed, stored in a temporary table and only transferred in the corresponding permanent tables if no errors are found. As for the performance of the imports, the system has been optimized to be able to check 6000 samples for 10 fields in about 5 min.
- Retrieval of information based on a large set of search criteria, generating either tables, Excel files or other types of reports.
- Storage of sample or container movements: the system not only retrieves and modifies automatically the positions but also updates the status of the initial and new containers.
- Automatic and instant barcode printing: the system is linked to linear and 2-dimensional barcode printers via a specific procedure and calling of a shell script. It enables selection of lists of barcodes to print and specification of format (Supplementary Material: SD-3).
- Automatic sending of e-mails to users and managers in case of errors during data imports or other problems.

A user-friendly and intuitive interface enables to access the different functions and to navigate within the forms and their multiple tabs. Each form has at least three tabs: one or two for adding information, one for searches and one for listing the results of queries.

Users are guided while navigating within the interface thanks to ‘restrictive’ forms that prompt them to fill the required fields and that restrict values with pull-down menus, in order to minimize form-filling errors. PL/SQL form validation checks data type and its integrity with respect to the tables in the database. Samples and containers barcodes can be entered using scanners to avoid miswriting. Each data modification, addition, or update is stored automatically in a secure field in each table with the user name and the date (userstamp and timestamp).

In addition, one of the key features of a reliable sample management system is powerful reporting with easy and quick ways of extracting information from the database. In this respect, we developed several different options:

- Each form has a specific tab with many search criteria. Results of the queries can either be displayed within the interface or be downloaded directly in Excel file (generated on the fly).
- For samples and containers, the queries list not only their features but also their location and their movement details.
- For containers, a function enables generation of specific pictures for visualization of the containers’ content. The type of picture depends on the type of containers with color codes based on storage status (green when empty, orange when partially filled and red when full: see Supplementary Material: SD-4).
- In the current SMS version, pictures of containers enable visualization of the content of the sub-containers, thus providing the possibility of displaying two levels of contents (Supplementary Material: SD-4 and SD-5).
- A container tree displays all the hierarchies of containers; navigating within the tree provides details on each container with the possibility to directly retrieve the list of samples that are stored in the selected container (Supplementary Material: SD-6).
- Graphs can be generated dynamically from data stored in the database. Different types of graphs are directly designed in the form’s code using Java beans based on a Java component provided by Oracle: the FormsGraph class.

2.3 Security

Data management systems for biobanks of human specimens must comply with strict standards for confidentiality and protection of personal data (Hansson, 2009). We therefore took particular care of anonymizing all specimens and of controlling the database access. First, a trigger prevents user connection from any other way than through the Oracle forms application. Secondly, all connection attempts are stored and unsuccessful attempts are immediately reported to the administrators via e-mail. Then the information stored in the database is divided into datagroups that can be specific to a research group, a study or a substudy depending on the need to classify or restrict access to the data. Each sample collection belongs to one datagroup, and each user has permissions to access data for one or more datagroups. This access is managed through the forms: wherever the user is in the interface, visible data are restricted to his current datagroup. However, if allowed, the user can switch from one data group to another using a specific menu. Actions permitted to the users are also restricted at database level, depending on their role and their level of responsibilities. Thus, users with read-only permissions can only navigate within the tabs for searching and listing data.

SMS safety is also ensured by the implementation of a combination of automatic backups of the two servers (Supplementary Material: SD-7).

3 CONCLUSION

We have developed a freely available and platform-independent sample management system (SMS) tailored to handle a wide variety of biospecimens and of storage conditions within an integrated biobank management model. While there are commercially available systems with high performance for a specific usage (e.g. management of tumor banks in a hospital setup), there is a lack of flexible systems capable of accommodating the wide diversity of the collections developed by a broad-based research institute such as IARC. Our SMS covers a large number of storage possibilities while monitoring closely the physical constraints of the containers and imposing strict adherence to the storage protocols (Caboux *et al.*, 2007).It is an essential piece for our biobank bridging annotation database with our LIMS for high-throughput analytical workflows (Voegele *et al.,* 2007). This is the key requirement for efficient performance and correct use of this powerful material for all coming studies.

Flexible and expandable, our SMS model provides opportunities for continuous improvements and for integration of new features in the future. The database on its own, adaptable to any kind of sample collections and heterogeneous storage structures, could serve as model for any research center's biobank and could be managed with any type of web interface.

Conflict of Interest: none declared.

REFERENCES

Caboux,E. *et al.* (2007) Common minimum technical standards and protocols or biological resource centers dedicated to cancer research. *IARC Working Group Reports*, pp. 1–38.

Hansson,M.G. (2009) Ethics and biobanks *Br. J. Cancer,* **100,** 8–12.

Voegele,C. *et al.* (2007) A laboratory information management system (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening. *Bioinformatics*, **23,** 2504–2506.