Development of an integrated Information Technology System for management of laboratory data and next-generation sequencing workflows within a cancer genomics research platform

Catherine Voegele

Genetic Cancer Susceptibility Group

N° 95-2015

THESE de l'UNIVERSITE

délivrée par

l'Université CLAUDE BERNARD LYON I

Ecole Doctorale de *Biologie Moléculaire, Intégrative et Cellulaire* de Lyon (BMIC)

Spécialité : Bioinformatique

Diplôme de DOCTORAT
(arrêté du 7 août 2006)

présentée et soutenue publiquement le 27 novembre 2015

par

Catherine VOEGELE

Development of an integrated Information Technology System for management of laboratory data and next-generation sequencing workflows within a cancer genomics research platform

JURY :

Dr James MCKAY : Directeur de thèse
Pr Anke VAN DE BERG : Rapporteur
Dr Tom GAUNT : Rapporteur
Pr Joël LACHUER : Président
Pr Françoise GALATEAU-SALLE : Examinateur
Dr David COX : Examinateur

Titre en français

Développement d'un système informatique intégré pour la gestion des données de laboratoire et des étapes de séquençage de nouvelle génération au sein d'une plateforme de recherche en génomique du cancer.

Résumé en français

L'objectif de mon travail de thèse était de développer des outils bio-informatiques permettant d'améliorer la traditionnelle gestion de l'information scientifique au sein d'un grand centre de recherche et en particulier au sein d'une plateforme de génomique.
Trois outils ont été développés: un cahier de laboratoire électronique, un système de gestion de l'information de laboratoire pour des applications de génomique dont le séquençage de nouvelle génération, ainsi qu'un système de gestion des échantillons pour de grandes bio-banques. Ce travail a été réalisé en étroite collaboration avec des biologistes, épidémiologistes et informaticiens. Il a également inclus la mise en place d'interactions entre les différents outils pour former un système informatique intégré.
Les trois outils ont été rapidement adoptés par l'ensemble des scientifiques du centre de recherche et sont désormais utilisés au quotidien pour le suivi de toutes les activités de laboratoire mais aussi plus globalement pour les autres activités scientifiques du centre de recherche. Ces outils sont transposables dans d'autres instituts de recherche.

Titre en anglais

Development of an integrated Information Technology system for management of laboratory data and next-generation sequencing workflows within a cancer genomics research platform.

Résumé en anglais

The aim of my thesis work was to develop bioinformatics tools to improve the traditional scientific information management within a large research centre and especially within a genomics platform.
Three tools have been developed: an electronic laboratory notebook, a laboratory information management system for genomics applications including next generation sequencing, as well as a sample management system for large biobanks. This work has been conducted in close collaboration with biologists, epidemiologists and IT specialists. It has also included the setup of interactions between the different tools to make an integrated IT system.
The three tools have been rapidly adopted by all the scientists of the research centre and are now daily used for the tracking of all the laboratory's activities but also more globally for the research centre's other scientific activities. These tools are transposable in other research institutes.

Développement d'un système informatique intégré pour la gestion des données de laboratoire et des étapes de séquençage de nouvelle génération au sein d'une plateforme de recherche en génomique du cancer.

Résumé

L'évolution et la modernisation des sciences de laboratoire ont entraîné la production de grandes quantités de données et en particulier de données électroniques qu'il faut gérer. Les avancées technologiques en génomique ont également impliqué la mise en place de nouvelles procédures de laboratoire plus complexes et permettant des analyses à haut débit en augmentant la capacité des instruments et en réduisant les coûts. Ces procédures nécessitent un suivi strict et soigneux à chaque étape du travail de laboratoire. L'augmentation du débit des analyses a favorisé en parallèle le recrutement de grandes collections d'échantillons biologiques pour lesquelles il est également indispensable de stocker les données associées de façon performante et sûre. De ces progrès et avancées technologiques a découlé le défi de trouver des solutions informatiques adaptées pour une gestion appropriée de toutes ces données.

L'objectif général de mon travail de thèse était donc de développer des outils bioinformatiques permettant d'améliorer la traditionnelle gestion de l'information scientifique au sein d'un grand centre de recherche et en particulier au sein d'une plateforme de génomique. Pour cela, j'ai travaillé en étroite collaboration avec les chercheurs biologistes, épidémiologistes et informaticiens du Centre International de Recherche sur le Cancer (CIRC) pour implémenter des outils permettant le suivi des échantillons, des expériences et des résultats, suivi indispensable à la conduite d'études à grande échelle.

Le premier outil développé est un cahier de laboratoire électronique destiné à remplacer le cahier papier qui permettait jusqu'à présent d'enregistrer les activités expérimentales journalières des laboratoires et les investigations scientifiques pour références futures. Malgré sa grande valeur, le cahier papier présente des limitations en particulier concernant l'insertion des données électroniques générées par les

laboratoires modernes, la recherche d'informations qui peut s'y avérer laborieuse ou le partage d'informations qui reste compliqué.

Le cahier électronique développé est un outil gratuit et basé sur le système de gestion de contenu WordPress qui est un logiciel de publication libre (« open-source »). Il est accessible depuis n'importe quel ordinateur, tablette ou smartphone par l'intermédiaire d'un navigateur internet grâce à un identifiant et un mot de passe propre à chaque utilisateur. Il intègre les fonctionnalités d'un cahier papier en permettant l'enregistrement des données expérimentales de manière flexible et sécurisée. Il est possible en outre d'attacher des documents électroniques ou des liens vers des ressources extérieures. L'outil facilite la recherche d'information et permet son partage grâce à la possibilité de définir des droits d'accès en lecture ou écriture très précis que ce soit sur des cahiers personnels ou communs pour la gestion de projets ou d'instruments spécifiques. Il présente ainsi des moyens de communication supplémentaires qui permettent d'améliorer et de renforcer les échanges et les collaborations entre les chercheurs. Il est particulièrement adapté aux activités de laboratoire mais également à la prise de notes de travaux d'épidémiologie, de bioinformatique ou de biostatistiques. Ce cahier de laboratoire électronique est approprié à la fois pour des petits laboratoires car requérant de faibles ressources informatiques, et pour des grands instituts de recherche car multi-utilisateurs et multidisciplinaire. Enfin, la reconnaissance de la valeur juridique des données électroniques devrait favoriser dans les années à venir l'adoption des cahiers électroniques par l'ensemble de la communauté scientifique.

Le deuxième outil développé est un système de gestion de l'information de laboratoire (LIMS) pour des applications de génomique dont le séquençage de nouvelle génération. En effet, ces nouvelles technologies ont apporté dans les laboratoires des procédures complexes à plusieurs étapes variables, certaines critiques, d'autres optionnelles et souvent associées à du haut débit. Une automatisation et un suivi rigoureux de chaque manipulation d'échantillons et de réactifs sont donc nécessaires incluant un stockage de toutes ces informations dans un environnement robuste et sécurisé.

L'outil mis en place repose sur une base de données et une interface internet qui permet d'enregistrer chaque étape des procédés de laboratoire. La définition des besoins étant cruciale, une connaissance approfondie des procédures de laboratoire est

nécessaire. La stratégie de design et de modélisation est importante pour une utilisation correcte de l'outil et une gestion performante du laboratoire. Elle doit être flexible et prendre en compte une estimation de la variabilité ainsi que des évolutions et exigences futures. L'interface doit être facile à utiliser, basée sur un vocabulaire adapté et des menus suivants chaque étape dans un ordre chronologique. L'outil est également connecté à des imprimantes à code-barres, robots et autres instruments de laboratoire pour faciliter l'échange d'information et optimiser l'automatisation des tâches.

Le troisième outil développé est un système de gestion des échantillons pour de grandes biobanques. Ces dernières représentent une ressource importante en recherche médicale pour des études dans des domaines variés et notamment en génomique et en épidémiologie génétique. En particulier, la biobanque du CIRC regroupe des échantillons appartenant à de nombreuses collections variées et hétérogènes qui étaient jusqu'alors gérées individuellement par les investigateurs. Ceci avait pour conséquence un manque de visibilité globale des ressources disponibles pour la recherche. Une augmentation du nombre de collections liées à de nouvelles études à grande échelle associée à une volonté croissante de partager ces ressources avec la communauté scientifique a nécessité la mise en place d'un outil informatique robuste pour gérer les données épidémiologiques basiques, la localisation au sein des infrastructures de stockage ainsi que le mouvement des échantillons.

L'outil développé repose sur une base de données relationnelle et une interface internet avec accès sécurisé par un identifiant et un mot de passe. L'outil est « open-source » et applicable à la gestion de différents types de biobanques même très hétérogènes car adaptable à une grande variabilité d'échantillons et d'équipements de stockage (cuves à azote, congélateurs, armoires) grâce à un système de hiérarchies de contenants imbriqués tels des poupées russes. L'outil a été optimisé pour permettre l'import de grande quantité de données par l'intermédiaire de fichiers formatés dont la cohérence par rapport au contenu attendu dans la base de données est vérifiée pour assurer l'intégrité des données. Des fonctionnalités avancées de recherche et d'extraction des données permettent de générer des tableaux, arbres, images et graphiques.

Les trois outils ont été rapidement adoptés par l'ensemble des scientifiques du CIRC et sont désormais utilisés au quotidien pour le suivi de toutes les activités de laboratoire mais aussi plus globalement pour les autres activités scientifiques du centre de recherche. Le cahier électronique compte plus de 100 utilisateurs, le LIMS gère toutes les activités de laboratoire et notamment celles concernant le séquençage à haut débit. Le système de gestion de la biobanque, quant à lui, inclut les données et la localisation de plus de 5 millions d'échantillons.

Ces trois outils complémentaires permettent de gérer différents types d'information de laboratoire. L'accès à ces données doit être sécurisé de façon appropriée que ce soit du point de vue de la propriété intellectuelle dans le cas des cahiers électroniques de laboratoire ou du point de vue de la confidentialité des informations personnelles associées à des échantillons humains stockés dans les biobanques. Afin de réduire la charge de travail liée à l'enregistrement des données et faciliter la recherche d'information, j'ai également mis en place des interactions entre les différents outils pour former un système informatique intégré. Il pourrait être intéressant par la suite de connecter ces outils à des logiciels d'analyses bioinformatiques pour stocker, de façon similaire au stockage des étapes d'expériences de laboratoire, toutes les étapes du traitement bioinformatique en répertoriant les références, les versions de logiciels utilisées et les fichiers de résultats intermédiaires. En effet, ces traitements prennent une place de plus en plus importante dans les laboratoires de génomiques et la traçabilité et la reproductibilité de ces analyses bioinformatiques sont essentielles.

Les trois outils développés ont été publiés avec les codes sources pour permettre à d'autres instituts de recherche de les installer voire de les adapter à leurs besoins spécifiques. Le développement de ce type d'outil devient aujourd'hui incontournable pour faciliter le partage de connaissance encouragé par les politiques récentes de certains journaux scientifiques prônant l'accès au plus grand nombre pour la dissémination des recherches et l'émulation des investigations scientifiques.

Remerciements

Tout d'abord merci à James de m'avoir permis de faire ma thèse en parallèle de mon travail de bio-informaticienne au sein de son équipe et de m'avoir fait confiance. Merci aux membres de mon comité de suivi de thèse pour leurs précieux conseils ainsi qu'à tous les membres du jury qui ont accepté d'évaluer mon travail.

Merci à toutes les personnes avec qui j'ai eu le plaisir de travailler sur les différents projets de ma thèse :

- Lucile, Baptiste, Nivo, Brigitte et tous les testeurs de l'ELN ;
- Nathalie, Amélie, Florence, Maroulitsa et Geoffroy pour les neurones perdus dans d'interminables discussions destinées à définir les besoins du LIMS ;
- Lucile et Elodie pour SAMMMMY ;
- tous les membres de l'équipe GCS et ITS pour leur soutien, leur aide au quotidien et les joyeuses pauses déjeuner et pauses café autour de grilles de mots fléchés destinées à maintenir nos cerveaux au top de leur forme.

Un merci tout particulier à :

- ma famille qui m'a toujours soutenu et en particulier à mes parents et mon frère ;
- Sean pour la confiance qu'il m'a accordé il y a 10 ans et pour tout ce qu'il m'a appris ;
- Florence avec qui c'est un plaisir de partager mon bureau et mes idées ;
- Fabienne pour son aide précieuse et ses bons conseils ;
- Nath pour son amitié et les pauses « thé » ;
- tous les gentils relecteurs qui ont corrigé mon manuscrit: Florence, Fabienne, Lucile, Sébastien, Vincent, Didier, Elo, Nath, Antoinette et Mathieu ;
- tous mes amis… danseurs… Mimi, Anne-Eli, Val, Elo, Betty, Brigitte, Flo, Pierre-Yves, JP, Johan, Francis, Didier… chanteurs…Jflm… musiciens… Vincent, Jean-Jacques … et badistes … Sylvain, Charles, Seb… et tous les autres…

… pour leur soutien, pour tous les bons moments partagés et pour les innombrables fou-rires passés et à venir…

Enfin un énorme merci à Lucile mon super binôme de travail sans qui je ne me serais peut-être pas lancée dans cette folle aventure et qui m'a tant apporté personnellement et professionnellement.

Thanks

First of all, thanks to James for having let me do my PhD in parallel with my bioinformatician duties within his team and for his trust. Thanks to the members of my thesis committee for their precious advice as well as to all the members of the jury who agreed to evaluate my work.

Thanks to all the people with whom I had the pleasure to work on my different PhD projects:

- Lucile, Baptiste, Nivo, Brigitte and all the ELN testers;
- Nathalie, Amélie, Florence, Maroulitsa and Geoffroy for the neurons lost in never-ending discussions intended to define the LIMS needs;
- Lucile and Elodie for SAMMMMY;
- all the members of GCS and ITS teams for their support, their help and for the happy lunch-breaks and coffee-breaks around crossword puzzles aimed at maintaining our brains in good shape.

Special thanks to:

- my family who has always supported me and in particular to my parents and my brother;
- Sean for the trust he granted me 10 years ago and for everything he taught me;
- Florence with who it is a pleasure to share office and ideas;
- Fabienne for her precious help and good advice;
- Nath for her friendship and for the tea breaks;
- all the nice proof-readers that corrected my manuscript: Florence, Fabienne, Lucile, Sébastien, Vincent, Didier, Elo, Nath, Antoinette and Mathieu;
- all my friends… dancers … Mimi, Anne-Eli, Val, Elo, Betty, Brigitte, Flo, Pierre-Yves, JP, Johan, Francis, Didier … singers… Jflm… musicians… Vincent, Jean-Jacques… and badminton players … Sylvain, Charles, Seb … and all the others…

… for their support, for all the good moments shared and for the uncountable past and to-come giggles…

Finally an enormous thank to Lucile my great working partner without whom I probably would not have launched into this crazy adventure and who brought me so much personally and professionally.

Development of an integrated Information Technology System for management of laboratory data and next-generation sequencing workflows within a cancer genomics research platform

Foreword

The work presented in this thesis is the result of researches started in 2007 and carried alongside my professional duties as a bioinformatician. It includes thoughts and discussions on scientific information management systems that matured with experience acquired since the first Laboratory Information Management System I put in place in 2003 for sequencing at the European Molecular Biology Laboratory (EMBL) Genomics Core Facility up to the recent Electronic Laboratory Notebook developed for the International Agency for Research on Cancer (IARC).

Table of Contents

Abbreviations

API: Application Programming Interface

BD2K: Big Data to Knowledge

BBMRI-ERIC: Biobanking and BioMolecular resources Research Infrastructure

BC: Barcode

BCNet: Biobank and Cohort Building Network

BRC: IARC Biological Resource Center

CENSA: Collaborative Electronic Notebook Systems Association

CMS: Content Management System

CPU: Central Processing Unit

CSS: Cascading Style Sheets

CSV: Comma-separated values

DNA: Deoxyribonucleic acid

DBMS: DataBase Management System

dNTP: Désoxy-N-Tri-Phosphate (N being Adénine, Cytosine, Guanine or Thymine)

EBI: European Bioinformatics Institute

ELN: Electronic Laboratory Notebook

EMBL: European Molecular Biology Laboratory

EPIC: European Prospective Investigation into Cancer and Nutrition

ER: End-repair

ES: Enrichment of Spheres

ExAC: Exome Aggregation Consortium

FDA: Food and Drug Administration

GB: Gigabytes

GCS: Genetic Cancer Susceptibility

GSP: Genetic Platform

GUI: Graphical User Interface

GWAS: Genome-wide association study

HPC: High Performance Computing

HRM: High Resolution Melting Curve Analysis

HTML: HyperText Markup Language

IARC: International Agency for Research on Cancer

IBB: The IARC Biobank

ICGC: International Cancer Genome Consortium

ID: Identifier

IP: Internet Protocol

ISO: International Organization for Standardization

IT: Information Technology

LAMP: Linux Apache MySQL PHP

LIMS: Laboratory Information Management System

LMICs: Low- and Middle-Income Countries

LOV: List Of Values

LSC: Laboratory Steering Committee

LSST: Large Synoptic Survey Telescope

LTS: Long Term Support

MB: Megabytes

NASA: National Aeronautics and Space Administration

NGS: Next Generation Sequencing

NIH: National Institute of Health

OT: Ion OneTouchTM

PC: Personal Computer

PCR: Polymerase Chain Reaction

PDF: Portable Document Format

PHP: Hypertext Preprocessor (originally Personal Home Page)

PLN: Paper Laboratory Notebook

R&D: Research and Development

RAM: Random-Access Memory

RBC: Red Blood Cells

RNA: Ribonucleic acid

SAMI: SAmple Management system for IARC biobank

SDK: Software Development Kit

SMS: Sample Management System

SOP: Standard Operating Procedures

SQL: Structured Query Language

SSL: Secure Sockets Layer

TB: Terabytes

TCGA: The Cancer Genome Atlas

TOC: Table of Contents

UK: United Kingdom

URL: Uniform Resource Locator

US: United States (of America)

VPN: Virtual Private Network

WHO: World Health Organization

WP: WordPress

Introduction

Context

The laboratory

The International Agency for Research on Cancer (IARC) was established in 1965 as an extension of the World Health Organization (WHO) to work on the identification of cancer causes and on the establishment of preventive measures for improvement of health through a reduction in the incidence and mortality from cancer.

The Agency is interdisciplinary, bringing together skills in epidemiology, biostatistics and laboratory sciences. As an independent entity, it promotes and establishes international collaborations in cancer research including partners from low and middle-income countries (LMIC). A core part of the Agency's mission is also education and training of cancer researchers worldwide through fellowships, courses and publications with a priority given to researchers from LMIC in the areas of cancer epidemiology and cancer registration.

Among all the topics studied, genetic susceptibility plays an important role in the development of many types of cancer. So far findings only explain a minor proportion of familial clustering, and most of the genetic risk remains to be discovered.

The main objective of the Genetic Cancer Susceptibility (GCS) group is to further explore and describe genetic susceptibility to human cancers by evaluating the inherited genetic factors involved in this susceptibility and how the variants are related to events in the tumour tissue itself. This goal is achieved through several subprojects:

- conducting genetic studies to identify rare familial mutations conferring high risk of developing cancer, and more common genetic variations associated with more modest increase in risk;

- studying the functional effect of susceptibility alleles and mechanisms involved also in interactions with environmental factors and viruses;

- assessing how genetic susceptibility may be mediated by considering how susceptibility alleles relate to gene expression levels;

- and more recently, exploring the potential of genomic techniques to evaluate novel non-invasive biomarkers for early detection and surveillance of cancer, such as circulating tumour DNA or exosomal RNAs from plasma samples.

Although the group does work on common cancers such as lung ([1] - Wang et al., 2014;[1] - Wang et al., 2014) or breast cancers ([2] - Park et al., 2014), its research tends to focus on rarer cancers such as those of upper aerodigestive tract ([3] - Delahaye-Sourdeix et al., 2015), nasopharynx ([4] - Fachiroh et al., 2012), kidney ([5] - Purdue et al., 2011), lymphomas ([6] - Cozen et al., 2014) and melanomas.

GCS thus contributes to better understanding of the biological pathways involved in cancer development and indirectly to the application of clinical screening methods for cancer risk prevention and management at health level.

Investigations are done partly through imputation techniques in genome-wide association study (GWAS) and mainly through laboratory techniques such as next-generation sequencing (NGS) using a variety of study designs including both familial and case-control studies. GCS also coordinates the development of the Agency′s genomics and bioinformatics capacity.

The group hosts the Genetic Platform (GSP), a key component of the Agency, which develops, maintains and provides a suite of genetic and genomics laboratory techniques and expertise for all Groups within IARC through collaborative projects. It includes large-scale molecular epidemiology projects and many other genomics-based projects within IARC and outside collaborators.

The GSP makes cutting-edge genomic techniques accessible to IARC scientific groups and collaborators. It provides support through the complete project life-cycle, including study design, planning, execution, quality control, and subsequent bioinformatics analysis. In the laboratory context, GSP strives to develop semi-automated workflows for tailored, flexible, and cost-effective genomic analysis of IARC′s large and heterogeneous biological sample collections. GSP incorporates quality control measures throughout these processes to ensure the highest data quality.

GCS has been especially focusing since years on successive sequencing technologies for mutation screening using gel electrophoresis sequencer to capillary sequencer. The platform integrated NGS technologies from July 2011 on by acquiring three instruments from Life Technologies: a SOLiD5500XL ([7]), ([8] - Mardis, 2008), an Ion-Torrent Personal Genome Machine (PGM) ([9]) and an Ion Proton ([10]) with associated informatics support including three servers and one cluster. These instruments are used for sequencing at different scale and depth (whole exome sequencing on SOLiD and Proton, targeted sequencing on PGM and Proton, RNA sequencing on SOLiD and Proton).

The GSP includes also several pipetting robots and different instruments for different types of genotyping, high resolution melting curve analysis (HRM), gene expression assays and methylation profiling. These instruments all play an important role for achieving laboratory research work.

The associated IT infrastructures

In order to manage the laboratory aspects of the work and also the analyses of the experiments' results, GCS owns a quite large panel of computers all connected to the internal network protected behind a firewall (**Figure 1**):

- several desktop machines for each user and for piloting each robot, the Taqman ([11]) , the LightScanner ([12]) and the Illumina Beadstation 500 GX platform ([13])

- three performant Linux servers for each of the three NGS instruments (SOLiD, Ion PGM and Proton)

- a Hewlett-Packard high performance computing (HPC) cluster for all next generation sequencing analyses. The latter is composed of one head node and twelve computing nodes linked to a storage of 170 terabytes (TB). Each node has 2 Central Processing Units (CPUs) with 12 cores each totalling 288 cores.

- several Linux servers hosting the databases and web applications that make up the tools that I will described in this thesis.

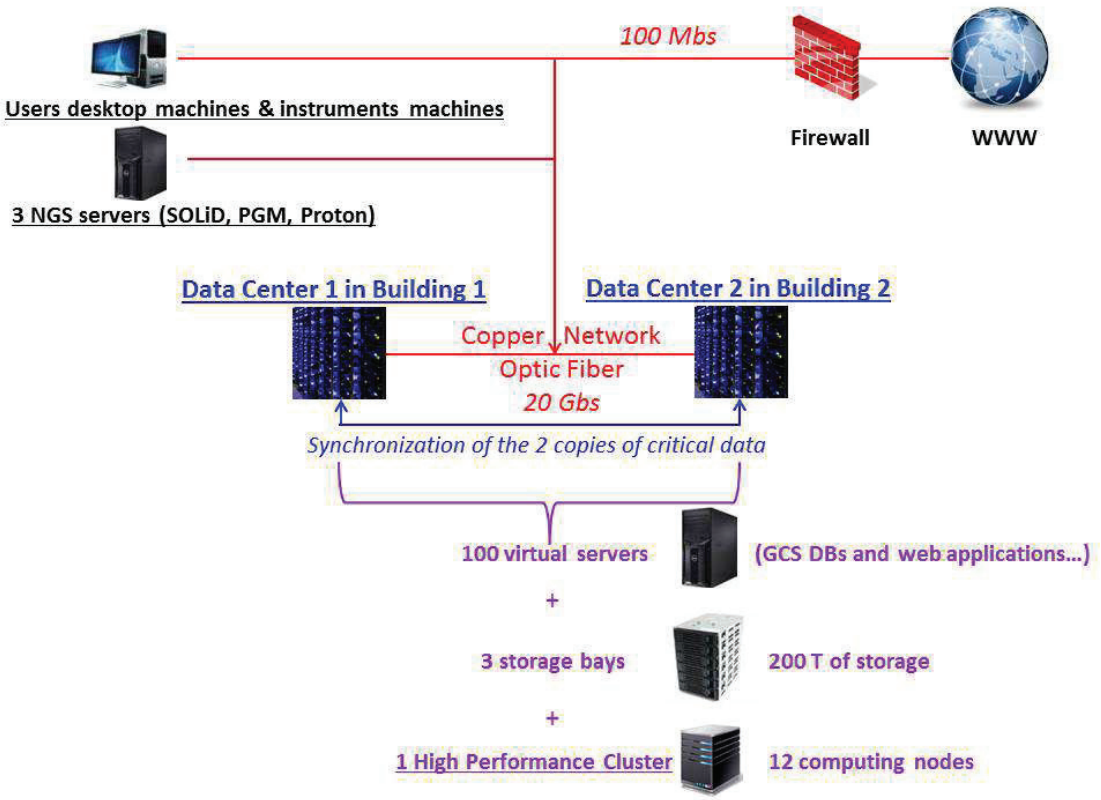


Figure 1: IARC IT network and infrastructure. *They include GCS IT material: the users' desktop machines, the instruments' machines, the servers for the databases and web applications, the storage bays and the high performance cluster.*

Bioinformatics

Definitions

GCS activities include a large part of bioinformatics as is most common in many research institutes and especially genomic laboratories. Bioinformatics is a broad discipline which uses by definition both biology and computer science to answer biological questions. Everyone does not agree on what kind of work and skills it implies, and even inside the scientific community opinions are divided. For some people it includes computational biology, biostatistics and biomathematics. For others, bioinformaticians design computer programs to solve biological problems whereas computational biologists apply computer programs and use computer tools to solve biological problems, *i.e.* they have knowledge in biology and use computers. In other words if you are writing software without understanding the biology, you are a programmer. If you are interpreting the biology without developing the code, you are a biologist or a computational biologist. Only if you are developing and writing tools or building databases to interpret the biology while simultaneously understanding the biology, one might consider oneself as a bioinformatician (**Figure 2**).

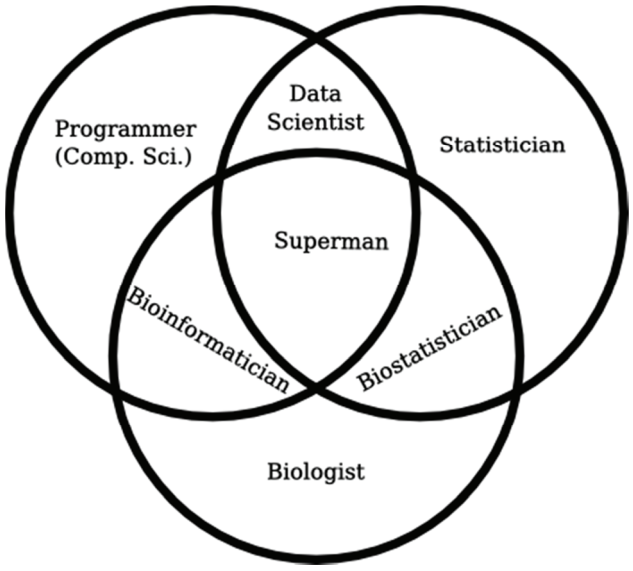


Figure 2: Chart from Anthony P. Fejes - PhD in Bioinformatics - describing his vision on frontiers in the biology and computer science world *([14]).*

Bioinformatics covers many various areas out of which we can distinguish the following main ones:

- biological databases and information systems aimed at data collection and query - this branch represents the first application of bioinformatics to store biological data in a structured and organized environment enabling easy retrieval of information for further investigations; it includes institutes' private databases but also public online databases such as Ensembl genome database hosting sequences and annotation from human, mouse and other vertebrate and eukaryote species ([15]). These latters play an important role in the sharing of knowledge.

- web applications and web technology - a concept introduced in 1999 enabling the interactive use of information technology (IT) tools through networks without having to install software on users' computers making them very attractive and popular. The web tools are hosted on servers and accessed by users via their own computers' web browser being compatible with the different operating systems (Linux, Windows and Macs).

- algorithmics and programming aimed at biological analyses and particularly genomics analyses to facilitate understanding of complex data sets (including phylogeny and sequence annotations)

- biostatistics which is the application of statistics to the design of biological experiments and the analyses and interpretation of their results

- modelling and imaging of biology structures such as proteins

Overall bioinformaticians have one major and common goal which consists in applying computational resources of information technology – tools or methods for different kind of applications - to the understanding of research questions.

Applications

Bioinformatics touches and has become essential in many scientific disciplines such as genetics, genomics and metabolomics. It is therefore very central to GCS activities for both data management and data analyses. As an example, my activities as bioinformatician focus on these two different aspects:

- the design, development, implementation and maintenance of databases and web applications tools which will be described in this thesis.

- genomic analyses for both the GCS group and other collaborators within the Agency ([16] - Vaca-Paniagua et al., 2015); ([17] - Kim et al., 2014); ([2] - Park et al., 2014); ([18] - Damiola et al., 2014); ([19] - Le Calvez-Kelm et al., 2012); ([20] - Ahmad et al., 2012) ; ([21] - Park et al., 2012); ([22] - Le Calvez-Kelm et al., 2011) ; ([23] - Tavtigian et al., 2009).

The group also plays a major role in overseeing and coordinating the whole bioinformatics capacities of the research centre by providing technical advices regarding how bioinformatics can be used to support and advances scientific researches. This requires as well taking care of associated information technology aspects and informatics infrastructures.

The growth of scientific data and the resulting technical challenges

One of the main bioinformatics' applications is the management of data generated from biological studies as are chemo-informatics and physico-informatics used for chemistry or physics data management. Indeed knowledge and data management is a general concern in many areas and particularly in science whether in biology, physics or other scientific disciplines. It implies adapted policy for production, dissemination, accessibility and use of information. Scientific data collected represent the primary source for scientific research. It has the particularity of being complex, incomplete, error-prone and in very high-demand for variety of applications requiring high performance informatics platform to support collection, curation, collaboration, exploration, and analysis of massive datasets. Indeed new technologies are creating large amount of data that requires efficient and adapted IT systems to be managed. In 2005, Jim Gray estimated in his report on global "scientific data management in the coming decade" that data volumes would approximately double each year ([24] - Gray et al., 2005), which was even an under-estimation for some areas.

Indeed, genomics in particular has become a "Big Data" science with the rise of NGS. Big data is defined as any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data is characterized by its large volume which cannot be handled by standard database management systems, as well as its velocity (high speed of data flow, change and

processing) and variety (generated by different sources). The European Bioinformatics Institute (EBI) in Hinxton, United Kingdom (UK), part of the European Molecular Biology Laboratory (EMBL) and one of the world's largest biology-data repositories, used to store 2 petabytes (10^{15} bytes) of genomic data in 2012, a number that is actually exponentially growing (**Figure 3**).

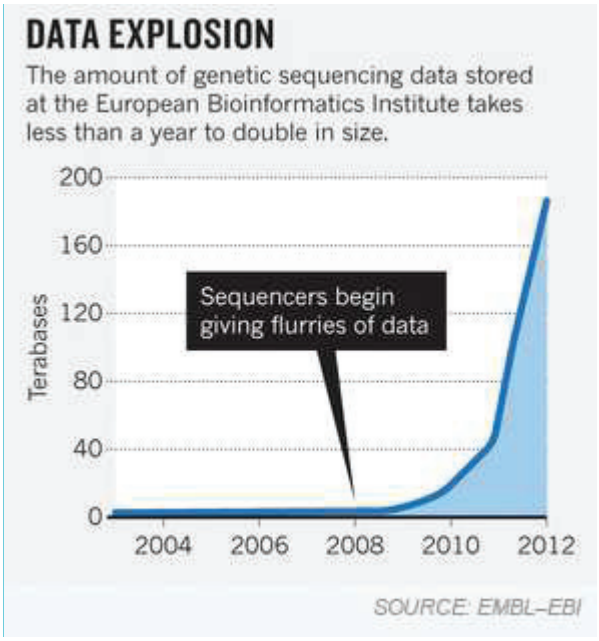


Figure 3: Growth of genetic sequencing data stored at EBI *([25] - EMBL - European Bioinformatics Institute, 2013).*

Following the increase in sequencing instrument capacity from 10^2 kilo-basepairs (kbp) output per run in 2001 to more than 10^{14} kbp in 2010 ([26] - Mardis, 2011), genomic data is predicted to take in 10 years' time the lead as the biggest data provider domain in the world creating more digital information than astronomy, particle physics and even popular Internet sites like Youtube ([27] - Stephens et al., 2015). 2.5 million plant and animal genomes are expected to be sequenced by 2025 because of the promise of genomic medicine to revolutionize the diagnosis and treatment of disease encouraging countries to sequencing large portions of their populations like England ([28]) or Saudi Arabia ([29]) which have announced plans to sequence 100,000 of their citizens. Also one-third of Iceland's 320,000 citizens have donated blood for genetic testing ([30] - Sulem et al., 2015), and researchers in both

the United States (US) ([31] - Kaiser, 2015) and China ([32] - Zhu, 2012) aim to sequence 1 million genomes in the next few years (**Figure 4**).

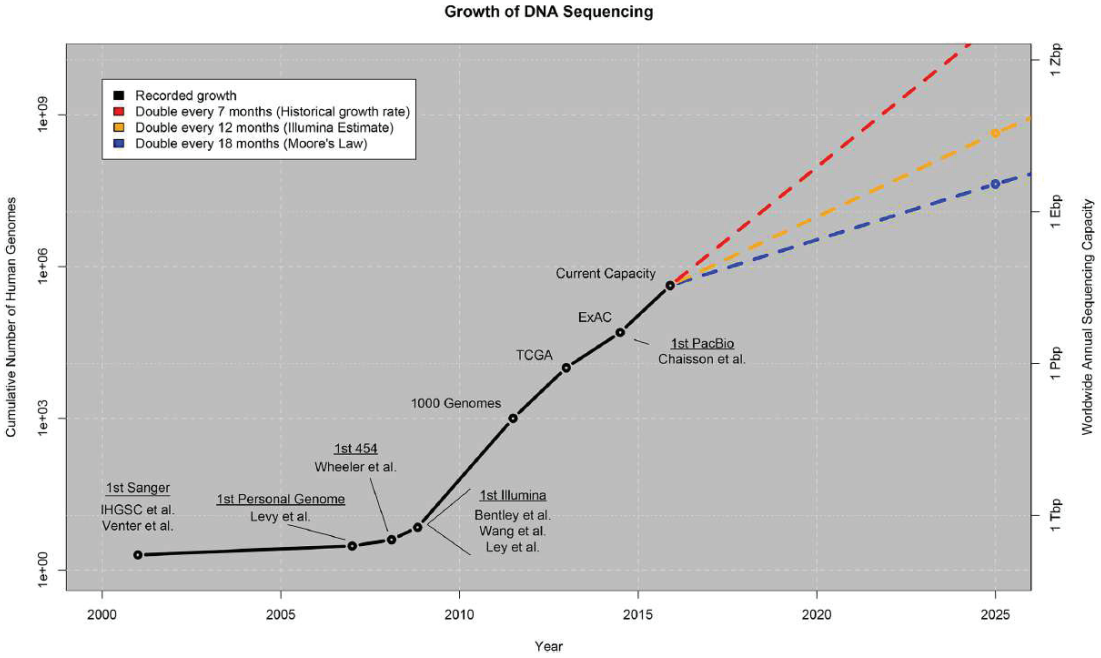


Figure 4: Growth of DNA sequencing *(from ([27] - Stephens et al., 2015). "The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values through 2015 are based on the historical publication record, with selected milestones in sequencing (first Sanger through first PacBio human genome published) as well as three exemplar projects using large-scale sequencing: the 1000 Genomes Project, aggregating hundreds of human genomes by 2012; The Cancer Genome Atlas (TCGA), aggregating over several thousand tumor/normal genome pairs; and the Exome Aggregation Consortium (ExAC), aggregating over 60,000 human exomes. The values beyond 2015 represent our projection under three possible growth curves as described in the main text".*

Technological advances and instruments accuracy have also increased data quality. In order to take benefit from this amount of better quality data, the challenge is to build smart tools that unlock the data and make it easy to capture, organize, analyse, visualize and disseminate to the scientific community. Indeed, in front of the

exponential growth of these data, managing the data and its metadata (descriptive information about the data and how it was generated) is challenging especially because biological data is complex, being collected from many places and in many different formats while requiring to be integrated from the different data sources. Indeed the Omics-Maps catalogue of all known sequencing instruments in the world ([33]) reports that there are currently more than 2,500 high-throughput instruments, manufactured by several different companies, located in nearly 1,000 sequencing centres in 55 countries in universities, hospitals, and other research laboratories ([27] - Stephens et al., 2015). Tracking the provenance and relationships of data is therefore also essential since biological discovery depends, to a large extent, on the presence of a clean, up-to-date and well-organised dataset ([34] - Koh et al., 2004). Appropriate management throughout the processing of data would allow gaining full value from these 'big data'. After acquisition, there is a challenging multi-step process starting with information extraction and cleaning, continuing with integration and aggregation up to analysis, interpretation and deployment ([35] - Jagadish et al., 2014).

GCS and more broadly IARC but any research institutes are generating a lot of scientific data of diverse nature. In my thesis, I will explain how we developed IT systems to manage these data while enabling traceability of the origin and production process as well as storage of meta-data and annotations to reach high level quality data as informative as possible.

Large sample size and high throughput challenges

The management of laboratory information is especially important when like GCS working with large sample size in high-throughput studies generating these big data. Knowing accurate details on how it was generated, from which sample with which features, using which technique and which reagents is essential for its exploitation and analyses to be able to take benefit of the potential of these (big) data. Indeed, the increase in instruments' capacity and throughput has enabled conduction of very large studies. As an example, ([1] - Wang et al., 2014) genotyped 10,246 cases and 38,295 controls to search for rare variants in BRCA2 and CHEK2 affecting risk of lung cancer. As another example, The Breast Cancer Association Consortium (BCAC) combines data from many studies to identify genes that may be related to the

risk of breast cancer. The BCAC recently identified so FGFR2 as susceptibility gene for breast cancer using 53,835 cases and 50,156 controls from 49 different studies ([36] - Agarwal et al., 2014). This requires as well appropriate tools to follow each case and each control through their analysis, store all the data and be able to export it in standards formats enabling further combined analyses.

Existing scientific data management systems and approaches

Scientific data management has traditionally been performed using file systems and later on databases systems which have progressed to provide powerful data definition tools and associated research capacities ([24] - Gray et al., 2005). Currently, biological data management is largely conducted within specialized bioinformatics databases (warehouses) which aimed at containing cleaned information with detailed annotations (such as functional and structural properties, expert-enriched information like in GenBank ([37]) or UniProt ([38])) and integrate data from multiple sources as well as searches and analysis tools, qualities essential for the in-depth study of the relevant data ([34] - Koh et al., 2004).

Traditional database management system (DBMS) has its shortcomings such as long data loading and retrieving time and lack important features to meet the needs of increasingly data rich sciences ([39] - Cudre-Mauroux et al., 2009). The problems that bioinformatics now faces regarding the management of large volumes of data have been faced by other scientific disciplines. The DBMS community has indeed been working on large science databases for years building prototypes such as Sequoia 2000 for earth science (with Postgres ([40] - Dozier et al., 1994)) or the MonetDB extended and used for the Sloan Digital Sky Survey warehouse ([41]) giving the general public and the astronomers access to partial map of the universe (116TB from latest release in July 2014). This transformed their work from field-based to computer-based work ([42] - Ivanova et al., 2007). Biology is on its way to follow the same tendency.

Still these DBMS cannot efficiently handle extremely large data (such as the petabytes produced each year in astronomy by telescopes). “Big data is any data whose scale, complexity and diversity require new architecture, techniques, algorithms and analytics to manage it as well as to extract value and hidden

knowledge from it" ([43] - Kim, 2015). Thus other novel high performance solutions based on multi-dimensional nested-array model are being explored in the multi-institution SciDB project ([44]) which aims to develop an open-source data management platform for various data-intensive scientific applications, including astronomy and computational biology ([39] - Cudre-Mauroux et al., 2009). For instance, version 1 has been successfully used for the Large Synoptic Survey Telescope (LSST ([45])) generating 15 TB of raw images every observing night with a growing rate for whole data of 15 PetaBytes/year ([46] - Kantor et al., 2006).

Also to overcome the data scale issue, distributed storage and distributed processing of very large data sets on computer clusters are another type of solution. As an example Doug Cutting has created Hadoop an open-source software framework based on distribution to allow fast processing ([47]). Yahoo and Facebook are handling their large amount of data with Hadoop clusters.

Finally scientists whether from big research centres with already powerful IT resources or smaller institutes are investigating an emerging technology known as cloud computing based on grid computing which combine different computer resources hosted in different places. Genome databases from Ensembl ([15]), GenBank ([37]) and data from the 1000 Genomes Project, TCGA, International Cancer Genome Consortium (ICGC) are accessible via clouds ([48] - Baker, 2010). Bioinformaticians have turned their attention to writing software that can analyse genomic data directly on clouds. Nevertheless transferring data between institutes and clouds is problematic, networks being quickly saturated with large amount of data ([49] - Marx, 2013). In addition, confidentiality issues have to be considered. Scientists should ethically and legally fulfil their obligations to protect the privacy of human subjects if data are housed by a third party.

To summarize, there are different possible solutions for data management in fields like genomics where amounts of data are growing while requiring movements, reformatting and integration for advanced analytics. Neither unique conventional relational database management system nor Hadoop-based systems readily meet all the workflow, data management and analytical requirements. We will explore in this thesis the possibility to have several data management systems such as ELN, LIMS and sample management systems dedicated to specific types of laboratory information and communicating together for a global information management approach.

Because relational databases based-tools showed their ability in many domains, we worked on the development of three different web-interfaced database tools for three laboratory specific applications and their integration associated with best practice working guidelines on how, when and for which purpose to use which tool in order to know where to store and find which data.

Aim of the thesis

Evolution and modernisation of laboratory science through sophisticated and connected instruments has led to an increasing amount of data and especially electronic data. They also brought new complex laboratory workflows and enable higher throughput by increasing the capacity of the instruments and reducing the cost of the experiments ([50] - Studt, 2014). This requires therefore strict and careful follow-up of the laboratory work. The increase of the throughput favoured in parallel the setup of larger samples collections for which information needs as well to be stored efficiently. Following these progresses and advances in technologies, IARC and GCS in particular, are concerned by the challenge of finding solutions for appropriate IT management of all these data and the associated meta-data. The latter are often unavailable in science because not kept together in secured systems but left on paper notes, personal computers of unlinked files.

The overall aim of my thesis was therefore to develop bioinformatics tools to facilitate the laboratory activities and improve the traditional information management within the research centre. To achieve this aim, I have worked closely with researchers to implement tools to track samples, experiments and results essential for management of large-scale studies. The three specific aims of my thesis work are detailed below.

Aim 1: Development of an Electronic Laboratory Notebook

A decade ago, the relatively limited scale of research made it easy and practical for scientists to detail the laboratory and analytical processes in conventional paper laboratory notebooks (PLN) and when problem arose, consulting the relevant procedure, protocol, and reagents' lot numbers could solve the problem. This was dependent on the researchers' ability to record, extract and assemble the notebooks information. The PLN has been around for many years and, despite its compelling value, has its limitations. Like all manual systems, finding relevant information efficiently and sharing it is difficult and challenging.

In line with IARC's mission to provide leadership in cancer research and cutting-edge tools, we thus explored the possibility of using a modern electronic laboratory notebook (ELN). This kind of tool appeared recently to follow the

technology and communication advances and is nowadays more and more used by organizations in medicine, research, industry and education to store, share and search for information ([51] - Nussbeck et al., 2014). The first task of my thesis project has consisted in developing an open-source ELN for IARC to replace the paper notebook while providing a tool adapted to a large variety of research activities from laboratory to epidemiology through bioinformatics and enabling complex user permissions management required by large institutes with different research groups working on collaborative projects. This work is detailed in Chapter 1.

Aim 2: Development of a Laboratory Information Management System

Specifically in genomic platforms, the setup of modern and rapidly evolving and laboratory technologies brought new complex multi-step laboratory procedures for which automation was required to enable high-throughput applications of these methods to the large bio-repositories housed within IARC. Especially for next generation sequencing technologies workflows which include many variable, optional and sensitive stages for samples' preparations, there is the need to keep track of each manipulation of each sample with the reagents' lot numbers used at each step in a safe and secure environment. For this, a second aspect of my thesis work has consisted in developing the Laboratory Information Management System (LIMS) described in Chapter 2 including follow-up of laboratory processes but also some information on the analyses performed. I will particularly emphasize on the importance of the tool's design for a powerful management of modern laboratories.

Aim 3: Development of a Sample Management System for IARC Biobank

Like many large biobanks, IARC biobank groups samples from numerous and various collections that are available for the researches of the institute's scientists. The collections used to be managed in individual tools from Excel sample sheets to small Access databases disseminated in several places leading to a lack of global visibility of the biobank resources. Several recent large-scale research projects have increased the number and variety of samples to be handled and archived in the Agency's biobank. This exponential growth has required the set-up of strict procedures and of a computerized tool to manage the storage, the use and the movements of samples. We therefore developed a SAmple Management system for IARC biobank (SAMI) aimed at hosting all IARC samples collections' basic

epidemiologic and location data. We took care of designing an open-source tool that could be used for management of any type of biobank. This system is presented in Chapter 3.

Finally, I have proposed solutions to minimize the work load needed to feed these different tools by setting up connections between them. The aim was to facilitate the daily tasks of the laboratory technicians and scientists in charge of the different projects by trying to avoid duplication of information recording and improve information search.

Chapter I – The IARC Electronic Laboratory Notebook (ELN)

I] 1. Introduction

I] 1.1 Definition and background

Paper laboratory notebooks (PLNs)

IARC has established the use of official laboratory notebooks like most scientific institutions following research best practice rules ([52] - Kanares, 1985). The first diary-style notebook found was Leonardo Da Vinci's who wrote down his experiments among which those on combustions 500 years ago. Since then most scientists among whom Faraday, Darwin, Einstein but also technicians and students have been recording more or less carefully their experiments in paper notebooks ([53] - Bird et al., 2013). These laboratory notebooks intended to document research, experiments and procedures performed in the laboratory or field studies, remain crucial and essential to the activities of the research community.

Indeed they provide a permanent record of the daily research work and scientific investigations from the initial phases to conclusions by tracing and preserving the experimental data and observations for future references. They assist future researchers with the understanding, the interpretation and the reproduction of experimental findings and therefore are a valuable resource for writing a scientific article, a report or a thesis. They can also be referred to in patent prosecution or intellectual property litigation as written evidence of ideas and inventions because they are legal documents dated and signed as they act as legally certified records of what was done, when, by whom and how.

Though the guidelines vary between institutions, the use of laboratory notebook follows general rules adopted by the entire scientific community regarding the format and expected content ([54] - Ryan, 2015). All data must be recorded in chronological order with the date of entry and should be recorded in ink (no pencil or felt-tip pens should be used). The records should be clear and accurate and should reflect the work progress. They should be concise and give sufficient details so that any associates can reconstruct and/or repeat the experiments. They should systematically record every aspect of their research including negative results and

nothing should be erased. The more information is provided, the easier and faster difficulties can be overcome. Laboratory notebooks are therefore important written records of knowledge, experience and observations for scientific organizations.

PLNs at the IARC

At the IARC, the paper notebooks were provided to the newcomers by the administration staff that assigned a tracking number, date of issue and signed its first page. The scientists receiving notebook were also required to sign on the first page and to follow the instructions contained in the preface. The information on the notebook including tracking number, user name, research group and date were recorded in a small database. When a notebook was completed or when a user left the agency, the notebook was inspected, signed and dated by the supervisor and either kept in the respective laboratory if still needed or archived in the anti-fire strongroom. The database was then updated with projects names and keywords to facilitate later searching.

The content of paper notebooks generally fitted the following structure:

- First page with instructions and rules of use (**Figure 5**)

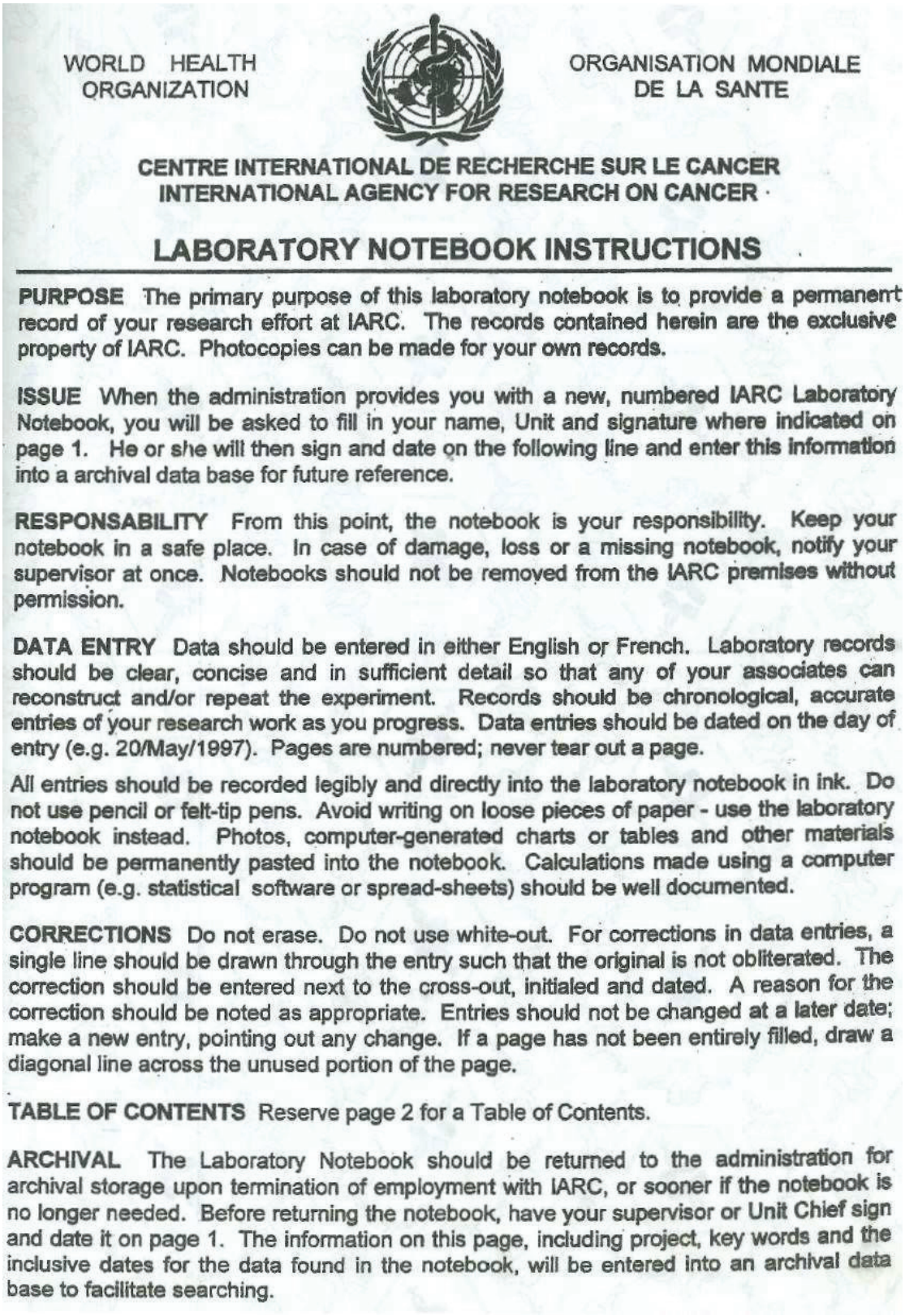
- 2nd page with notebook number, notebook owner, research group, dates, project, keywords, dates and signatures

- 3rd and 4th pages with the table of contents (TOC) containing the list of experiments with their title, their date and the corresponding pages. An introduction may also be included before the first lab records as well as a list of abbreviations containing common laboratory terminology.

- Records of the daily laboratory ideally with objective, material and method, results, conclusion.

- The content of the PLN may also be summarized in the last page.

The pages are numbered and it is forbidden to tear any page or erase any content. Corrections are allowed but following specific rules (**Figure 5**).



WORLD HEALTH ORGANIZATION

ORGANISATION MONDIALE DE LA SANTE

CENTRE INTERNATIONAL DE RECHERCHE SUR LE CANCER
INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

LABORATORY NOTEBOOK INSTRUCTIONS

PURPOSE The primary purpose of this laboratory notebook is to provide a permanent record of your research effort at IARC. The records contained herein are the exclusive property of IARC. Photocopies can be made for your own records.

ISSUE When the administration provides you with a new, numbered IARC Laboratory Notebook, you will be asked to fill in your name, Unit and signature where indicated on page 1. He or she will then sign and date on the following line and enter this information into a archival data base for future reference.

RESPONSABILITY From this point, the notebook is your responsibility. Keep your notebook in a safe place. In case of damage, loss or a missing notebook, notify your supervisor at once. Notebooks should not be removed from the IARC premises without permission.

DATA ENTRY Data should be entered in either English or French. Laboratory records should be clear, concise and in sufficient detail so that any of your associates can reconstruct and/or repeat the experiment. Records should be chronological, accurate entries of your research work as you progress. Data entries should be dated on the day of entry (e.g. 20/May/1997). Pages are numbered; never tear out a page.

All entries should be recorded legibly and directly into the laboratory notebook in ink. Do not use pencil or felt-tip pens. Avoid writing on loose pieces of paper - use the laboratory notebook instead. Photos, computer-generated charts or tables and other materials should be permanently pasted into the notebook. Calculations made using a computer program (e.g. statistical software or spread-sheets) should be well documented.

CORRECTIONS Do not erase. Do not use white-out. For corrections in data entries, a single line should be drawn through the entry such that the original is not obliterated. The correction should be entered next to the cross-out, initialed and dated. A reason for the correction should be noted as appropriate. Entries should not be changed at a later date; make a new entry, pointing out any change. If a page has not been entirely filled, draw a diagonal line across the unused portion of the page.

TABLE OF CONTENTS Reserve page 2 for a Table of Contents.

ARCHIVAL The Laboratory Notebook should be returned to the administration for archival storage upon termination of employment with IARC, or sooner if the notebook is no longer needed. Before returning the notebook, have your supervisor or Unit Chief sign and date it on page 1. The information on this page, including project, key words and the inclusive dates for the data found in the notebook, will be entered into an archival data base to facilitate searching.

Figure 5: IARC PLN Instructions for scientists performing experimental work.

The past decade has seen a progressive modernization of the laboratories with arrival of high-throughput platforms like NGS and microarrays generating more and more electronic data which printing and pasting into the paper notebooks became impractical. Computers are now ubiquitous in science and technology. They control most instruments, enable us to capture, analyse, and annotate our data through in-silico tools ([53] - Bird et al., 2013).

These new technologies may help to facilitate daily laboratory work and prove potentially more efficient manners of communication and data sharing. We therefore explored the possibility of using electronic notebooks at the Agency.

Electronic laboratory notebooks (ELNs)

ELNs are software or "systems to create, store, retrieve and share fully electronic records in ways that meet all legal, regulatory, technical and scientific requirements; records being collections of information or data associated with an experiment to enable a suitably skilled person to repeat it". *(*Definition from the "Collaborative Electronic Notebook Systems Association (CENSA) – ([55] - Lysakowski, 1997).* It has a functioning equivalent of a paper notebook, but in a more advanced and efficient way by providing added value that will be detailed in this chapter.

In order to develop the most appropriate solution, at least as robust and scientifically sound as the paper notebook, we defined very precise and mandatory specifications in collaboration with IARC laboratory steering committee which is in charge of overseeing all Agency's laboratory activities and which discussed and reviewed the critical requirements for the ELN. While designed for IARC scientists' needs, we also attempted to keep the tool as open as possible to allow its application in any other research institutes.

I] 1.2 Specifications

The requirements for development of ELNs consist in integrating all the features existing in the paper-based laboratory notebooks together with additional

functionalities to enable easy and intuitive navigation and improve information search capacities upon the notebooks ([56] - Wright, 2009).

With this in mind, the following specifications were defined in order to be able to evaluate different options and guide us through the best solution for the development, whether it was more appropriate to design a new tool or to modify an existing tool to adapt it to our research centre's needs.

a) General specifications

The main objectives of the ELN are:

- to record experimental data in a flexible way through a common platform suitable to the various research activities of the institute while maintaining the rigorous scientific practices required for day-to-day note keeping (traceability of each single modification), all within a highly secure environment

- to improve laboratory efficiency and work productivity by facilitating protocol and data sharing, record keeping, particularly in terms of limiting repetitive laboratory tests or optimizations

b) Mandatory requirements

The ELN should:

- be free and open-source to follow Agency's willingness to encourage use, development and sharing of open access tools;

- be web-based requiring no installation of software on user's computer (therefore accessible from anywhere in the offices or laboratories as soon as users have a connection to the network);

- enable easy reading of the web interface that should be as simple as possible and understandable by non-IT people (meaning including a basic editor for formatting of the text like underlining or highlighting);

- enable insertion and visualisation of tables, various types of images (screenshots, plots), hyperlinks, audio and video clips;

- enable attachment of documents containing experimental results or outputs from scientific instruments to provide a link between the laboratory records and the data produced or the data outputs;

- provide efficient means by which to search for archived material (search for keywords in all the notebooks the users has access to (cf. management of users' permissions);

- enable the users to decide how to fill the notebooks: either one page per day or one page per experiment. Both possibilities should be available. In the same way the users should be able to sort the pages of their notebook by title or date;

- enable the users to own different types of notebooks: personal notebooks specific to one user and shared ones specific to a project or to maintenance of specific instruments (with easy distinction between the different types of notebooks);

- generate automatically a table of contents containing titles assigned to each page, author's identification as well as date of publication;

- provide a pre-defined template to use to format each page with (i) a title shortly describing the experiment to help tracking it through the table of contents (ii) an objective, (iii) the material and methods (iv) the results, discussions and conclusion whenever appropriate (v) attached documents or links to related documents (vi) references whenever appropriate (templates should be available within the editing tool and should not be mandatory);

- have the same scientific and juridical value as paper notebook which means complying with research best practice rules by:

    - enabling storage of all modifications with userstamp and timestamp;
    - providing long-term and secure storage in basic electronic format with correct and appropriate back up so that there is no risk to lose any data;

- provide support to collaboration by facilitating communication among scientists in different geographic locations for joint experiments and research programs (eliminate physical barriers to sharing information);

- last but not least - the most important feature of the ELN - manage user permissions: each notebook should be writable and readable by only specific selected users. There can be one (personal notebooks) or more editors (notebooks dedicated to specific projects or instruments).

c) *Other optional requirements*

Ideally the ELN should also permit:

- numbering of pages to ensure no page is removed;
- to help the user to fill correctly his notebook (*i.e.*: material used, conclusion), spelling check;
- auto-completion for long name of molecules or reagents (inclusion of dictionary);
- its potential use with smartphones and tablets although these device have not yet entered our laboratory. It should be compatible with the largest possible number of different browsers.

To summarize, for wide adoption of the ELN by staff, the tool should be beneficial in terms of time saving compared to the PLN by simplifying the data recording, improving data search, storage and backup as well as helping sharing valuable knowledge like protocols or laboratory inventories among users at any time and from everywhere with safe and secured accesses.

I] 2.Tool installation, configuration and developments

We performed a state-of-the-art review of free and open-access existing tools that had either been implemented in other institutes or that could be used to build on our ELN. We also looked at the main functionalities that the large number of commercial tools provide especially in "A review of electronic laboratory notebooks available in the market today" ([57] - Rubacha et al., 2011). The number and variety of solutions have increased with the growing interest from both research and routines laboratories making the selection of the right notebook a cumbersome and complicated process. Some of the existing solutions are for specific applications area such as biology (Studylog Systems from Studylog ([58]) or Gene Inspector from Textco BioSoftware ([59])) or chemistry (i.e. Benchware Notebook from Tripos/Certara ([60])). Others are tailored to research and development (i.e. NoteBookMaker from NoteBookMaker ([61]) or Agilent ELN/Kalabie ELN from Agilent Technologies ([62])) or designed to meet quality assurance and quality control needs (Nexxis ELN from PerkinElmer-Labtronics ([63]) or SmartLab from SoftGroup-VelQuest ([64])). Many include laboratory workflow management making them looking like small LIMS but only a few are multi-disciplinary and generic (E-Notebook from PerkinElmer/CambridgeSoft ([65]) or E-WorkBook from IDBS ([66])).

For our project, we did not consider commercial tools for three main reasons: (i) they are not flexible enough in terms of adaptation to specific needs to specialized scientific institutes such as IARC (ii) we wanted to stay independent from any company for future developments and modifications, commercial tools being mostly black boxes for which we cannot have a full understanding of their programming code which would prevent us from performing ourselves customizations (iii) we had the internal human resources and expertise that could be dedicated to the development of the tool.

During our state-of-the-art review, we found only a few freely available tools, which we explored. Most of them did not allow the management of several users with setup of precise access permissions or were already obsolete providing no support and/or not updated since several years (ORNL Electronic Notebook ([67]); CyNote ([68]); The Monster Journal ([69]); Electronic Laboratory Notebook ([70]); LabBook

([71])) – Cf. Appendix 1. As these two features (the possibility to setup very precise access permissions for multi-users and the support) were crucial ones for us, we excluded these tools and did not further investigate them.

Nevertheless, two tools upon which an ELN could be developed appeared suitable to suit our needs: WordPress ([72]) and Drupal ("MyLabBook" ([73])), both being Content Management Systems (CMS) – open source publishing software, modern and with a large number of developers and users around the world. We favoured WordPress for the following reasons:

- Written in PHP (a server side scripting language designed for web development but also used as software programming language) and based on a MySQL database (the most widely used open-source relational database management system), it was well known in-house and used for other projects, which was not the case of Drupal;
- WordPress has a strong basement while providing freedom in evolution and modifications; It is indeed easy to design and to customize and quite friendly for both developers and users;
- It is mostly used as 'blog generator' but its functionalities enable management of any type of website - Many of the functionalities needed for our ELN were already available in WordPress implying flexibility which is an important consideration for our ELN;
- The WordPress developers' community is very active. It is a modern and well-adopted CMS. A lot of plugins (software components that adds a specific feature to an existing software application enabling customization) exist for a multitude of functionalities. It is a living system with regular updates and bug corrections.

We evaluated further WordPress adaptability for the setup of our ELN through the study of its basic features and functionalities as well as the existing plugins and their associated documentation. We came to the conclusion that WordPress was flexible enough to be modified to cover all our requirements and especially the critical management of multi-user's permissions with tight access controls. All these developments are described in the following paragraphs.

I] 2.1 Implementation

Strategy

We set up two instances of WordPress and thus of the ELN: one for test and one for production. This implementation strategy enables to do new developments and perform updates or upgrades on the test environment without any risk of disturbing users working on the production environment, losing data or even crashing the system. Once the tests on the development environment are validated, the modifications are reproduced on the production environment.

Hardware requirements

The system requires minimal IT and memory resources which make it largely implementable in small laboratories. This was an important aspect regarding its availability as open-source tool ensuring that all users could install it.

Indeed we deployed WordPress on a LAMP server that hosts both the web application and the database. The operating system is a **L**inux Ubuntu 8.04.4 LTS (free and open-source operating system) including **A**pache 2.2.8 (the world most widely used and multi-platform web server software), **M**ySQL 5.0.51a and **P**HP 5.2.4 which are the only requirements for the installation as it does not need an expensive powerful machine to run. Our server has an Intel Xeon CPU at 2.53 GHz, 4 GB of RAM and a hard disk of 100 GB. The tool in itself is only around 200 MB.

The number of users and the way of using the ELN - writing a lot of text or attaching a lot of files - are determining the growth of the associated database as well as the upload directory and therefore enable to evaluate the disk space necessary in long-term perspectives. As an example after two years of use with creation of 5000 pages, our ELN database content remains minimal, with storage in the order of 600 MB and attached documents in the order of 5 GB.

Configuration

After installation of both WordPress and the database instance, the configuration was done in a file (wp-config.php) which specifies the MySQL database name used to store the users' information, pages, the administrator name(s) and password(s), the address of the database server and some keys.

Once these settings are specified, the interface theme, logo and home page can be customized along the lines of the research centers' corporate designs using CSS (Cascading Style Sheets).

All the developments to set up the ELN have been done through one of the four following means:

- modification of WordPress native code. This required studying carefully the system to have a comprehensive understanding of the coding and find the proper PHP files to modify in the core of WordPress. We limited these customizations to minimize the time spent on this at each update of WordPress. It concerned authorized MIME types, maximum upload file size and disablement of the uncategorized posts.

- installation of original plugins: mandatory ones to have an appropriate functioning of the ELN (Category Reminder, Collapsing Categories, Login Configurator, Role Scoper, WP-PageNavi) as well as useful but not crucial ones (CKEditor, My Link Order, Post2PDF Converter, Post Notification, Post revision display, Search By Category, WP-Print, WP Favorite Posts) and finally completely optional ones (Clean WP Dashboard, CSS Column, Dynamic to Top, Last Updated Posts Widget, ThreeWP Activity Monitor, WassUp Real Time Analytics).

For this, we searched, selected and installed more than one hundred of plugins through a process of trial and error until finding the most adapted ones.

- creation of a specific IARC plugin for some features that were not available through any plugin.

- modification of existing plugins.

All these configurations and codes are detailed in the supplementary data of our article published in Bioinformatics ([74] - Voegele et al., 2013).

I] 2.2 Transposition of WordPress model into ELN model

The correspondence between the WordPress terminology and the laboratory notebook has been defined as following:

- WordPress "Users" are any ELN users (laboratory technicians and scientists). Users can belong to several groups of users.
- WordPress "categories" correspond to notebooks or groups of notebooks.
- The WordPress term "post" designates the page of the notebooks.
- The WordPress term "page" is kept for static website pages.
- WordPress "comments" are users' annotations or comments added to the pages.
- WordPress "Media" are files related to the laboratory activities such as instruments images or documents.
- WordPress "Links" are links to ELN menus, ELN pages or other web pages.

I] 2.3 Management of users' permissions and administration

An important concern regarding ELNs is which data a particular user or group of users is able to access whether in reading or in writing. It is defined by what we call users' "permissions". The management of these permissions with specific access rules to selected notebooks or groups of notebooks was a crucial step since enabling to have one single tool for the whole research centre. The ability to specify very precisely multi-users data accesses based on permissions is a feature which could not be found in any of the open-source ELN available to the scientific community.

Set up of users' permissions

Users' and groups of users' permissions are managed through WordPress administrators' dashboard by the plugin "Role Scoper". When a new user is created, it is possible to assign the user to one or more groups. The permissions can be defined at the user level or for an entire group and for a specific notebook or for a group of notebooks that we called workspaces, using the option "assign for selected and sub-categories". They correspond to roles which can be specified very precisely: "Post reader" or "Post author" (**Figure 6**).

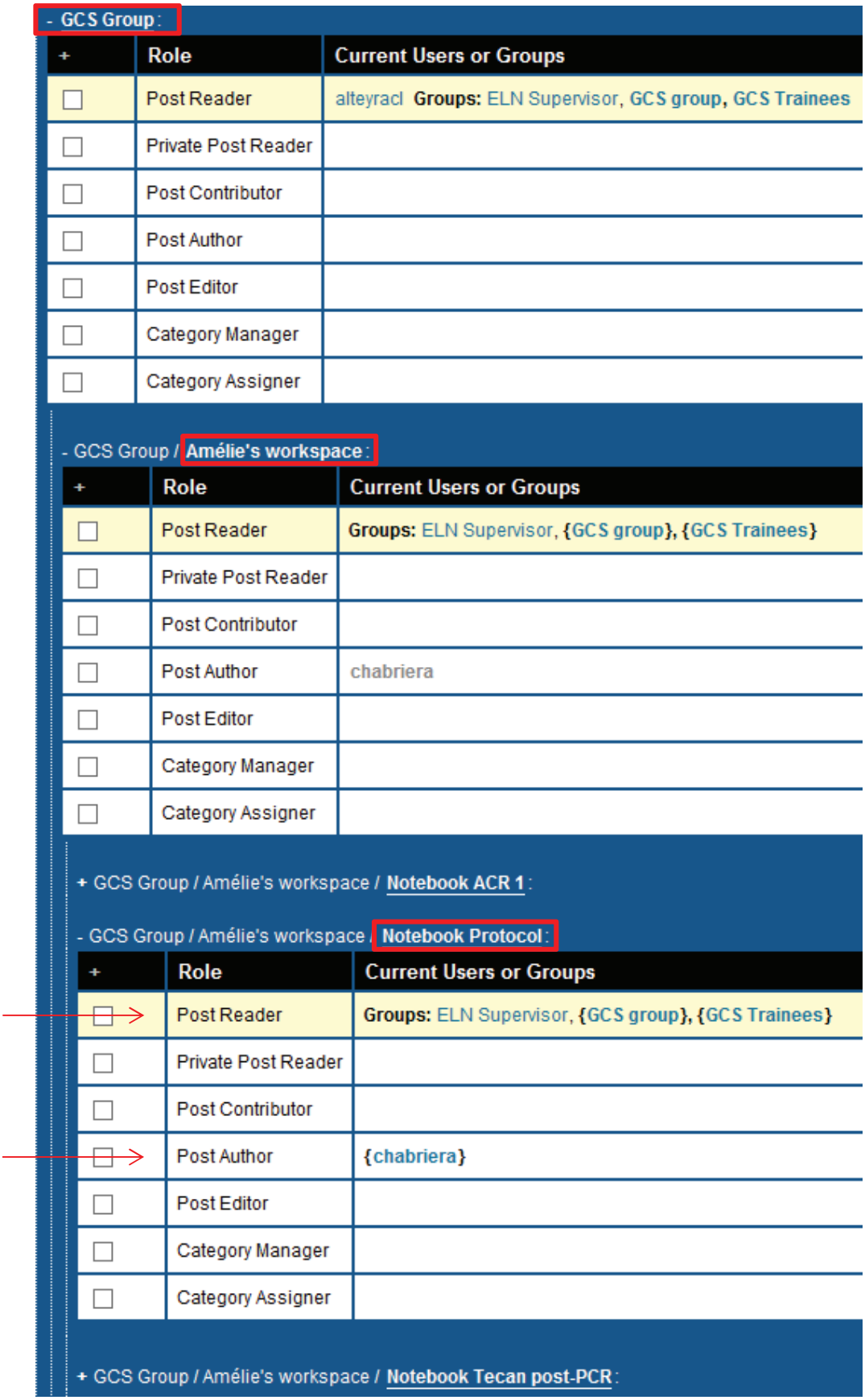


Figure 6: Example of definition of read and write permissions on a workspace and notebook.

The figure shows the different levels of permissions: GCS group workspace is readable by user "alteyrac" and the 3 groups: "ELN Supervisor", "GCS group" and "GCS trainees". Then the specific research assistant's workspace and notebook "Protocol" are readable by the groups "ELN Supervisor", "GCS group" and "GCS trainees" and writable only by the research assistant himself.

Read and write accesses to the ELN were oriented around the following rules after discussion with the Laboratory Steering Committee's scientists:

- the control of the attribution of the ELN and of reading/writing access limitations is defined by the persons responsible for the oversight of the laboratories in agreement with the user's supervisor. For notebooks shared between scientific groups, the permissions are defined jointly by the involved responsible scientists.
- two administrators of the system are responsible for its maintenance and for providing password-protected access to all users.
- as for the paper notebook a person representing laboratory services management carries out regular checks of the notebooks to confirm compliance with the IARC Standard Operating Procedures (SOP) regarding ELN work practise.
- 3 ELN workspaces are available for each group of users:

  * ELN group: dedicated to personal notebooks. It contains one sub-workspace for each member of the group which hosts one or more notebooks (for example one user can have one notebook per project).

  The usual permissions are read-only access for all group members on all notebooks and writing access on sub-workspaces only for the owner of the sub-workspace.

  * ELN share: This workspace contains notebooks with specific user access, for instruments, robots, shared projects between different users. Writing can be done by all authorised users, from the same group or from different groups.

  * ELN archives: This workspace contains all closed archived notebooks from students or staff members who have left IARC or from finished projects.

The usual permissions are: read-only access for all group members, no write access by anyone (**Figure 7**).

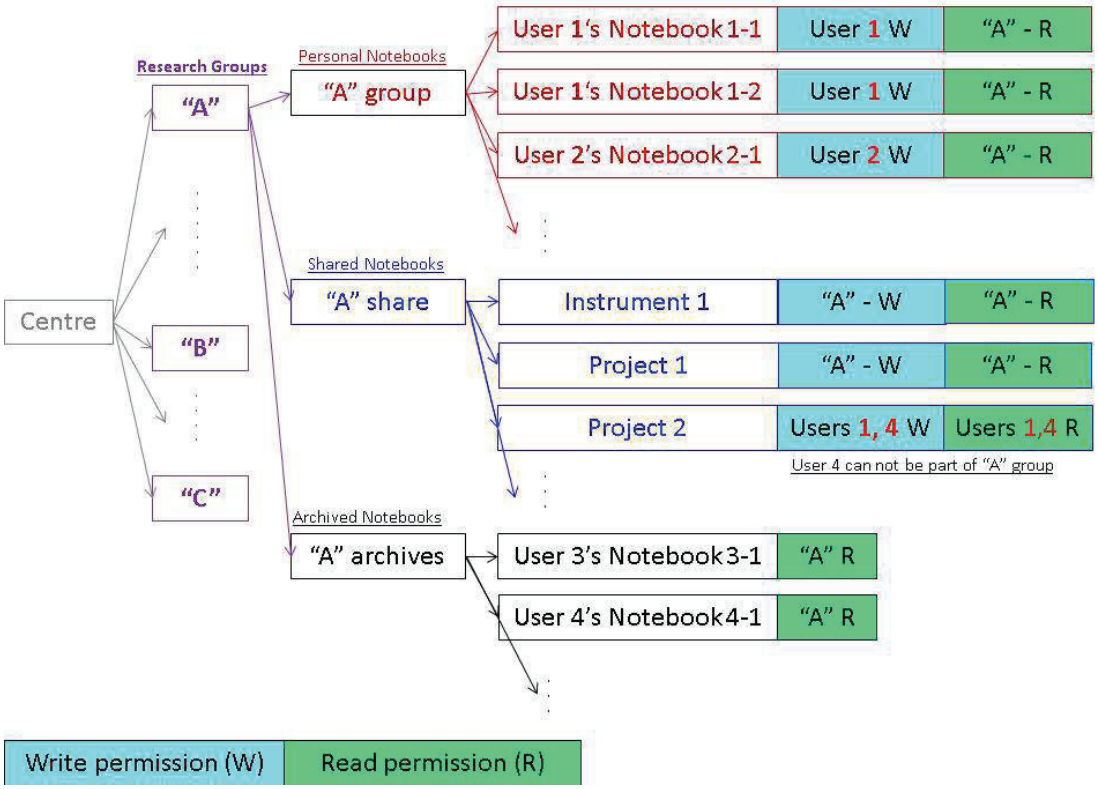


Figure 7: Hierarchical diagram of precisely defined ELN permissions. *The research center includes several research groups which have 3 types of notebooks: (i) personal ones writable by each owner and readable by the group; (ii) shared ones: some writable and readable by all members of the group (Instrument 1 and Project 1) and some writable and readable by only a few specific members of the group (Project2); (iii) archived ones: no more writable but readable by all the members of the group.*

The permission to comment is linked to the read access. A basic rule of WordPress implies that anyone who is allowed to read a page can comment on it. This was not an issue for the ELN since the original content of the pages is not affected, comments being completely "separated". Anyone trying to access a notebook which he/she is not allowed to see – for example by typing an URL manually – will be returned an error message.

Quality controls of the ELNs

Proper use of the ELN being a key component for its value, functionality was added for storage of information on the controls done by the staff members in charge of checking the quality of the users' inputs in the ELN to ensure appropriate usage.

For the tracking of these inspections, we developed our own plugin: "iarc-forms" and created a new table "IARC_FORMS" in the database containing the notebook ID (category ID), the date of issue, date of checking, date of archival and remark.

The plugin is made up of different functions enabling the storage, querying and display of the notebooks' checks' information:

-*getNotebooks*: SQL query (structured query language designed for managing data held in a relational database management system) for the "Search by Notebook' to get information restricted to the ones belonging to users in the laboratory staff group. It returns the owner names, and notebook title;

-*getAuthors*: SQL query to give the list of authors of unique personal notebooks (which have only one author). It returns the first name, last name, user ID and group name;

-*getDetails_callback*: gives details of the notebook to be displayed in the form: category_name, firstname, lastname, group, min_post_date (date of first entry), max_post_date (date of last entry);

-*getEnteredInfos_callback*: returns the information on notebooks already entered in the table wp_iarc_forms in order to pre-fill the form (date_of_issue, date_of_checking, date_of_archival and remark);

-*submitFormPhp_callback*: launches the update of the table WP_IARC_FORMS;

-*getNotebooksTable*: returns all the notebooks having only one author with information of checking (table "WP_IARC_FORMS");

-*displayIARCForms*: main function that redirects to the display of the form or the table depending on the choice (parameter);

-*displayNotebooksTable*: main function displaying the table of notebooks;

-*displayNotebooksForm*: main function displaying the form for updating the information on notebooks' checks.

Archival of the ELNs

For long-term management, notebooks are archived at departure of staff members or students. A notebook can also be closed upon request to the administrators: for example at the end of a project or when a notebook is becoming too large to navigate easily into its table of contents. The administrator will remove any ‘write’ permissions on these archived notebooks. Users will still be allowed to consult them if they were allowed before the archival, but no one will be authorized to modify them anymore. Commenting the archived notebooks will remain possible.

The management of users’ permissions from arrival to departure of a staff member as described here enables a safe use of the ELN and contributes to the overall security of the system, security which will be detailed in the following paragraph.

I] 3. Results: IARC ELN tool

Within this section I present how we managed to fulfil all the requirements and rules of a PLN providing additional features like storage of electronic data and easy sharing thanks to our ELN.

I] 3.1 Login

The developed tool is entirely web-based and therefore accessible from any "web-device" (Windows or Macintosh PCs, tablets and smartphones) connected to the internal network making it available in both laboratories and offices but also outside the Agency via Virtual Private Network (VPN). The access is controlled by personal login and password (**Figure 8**).

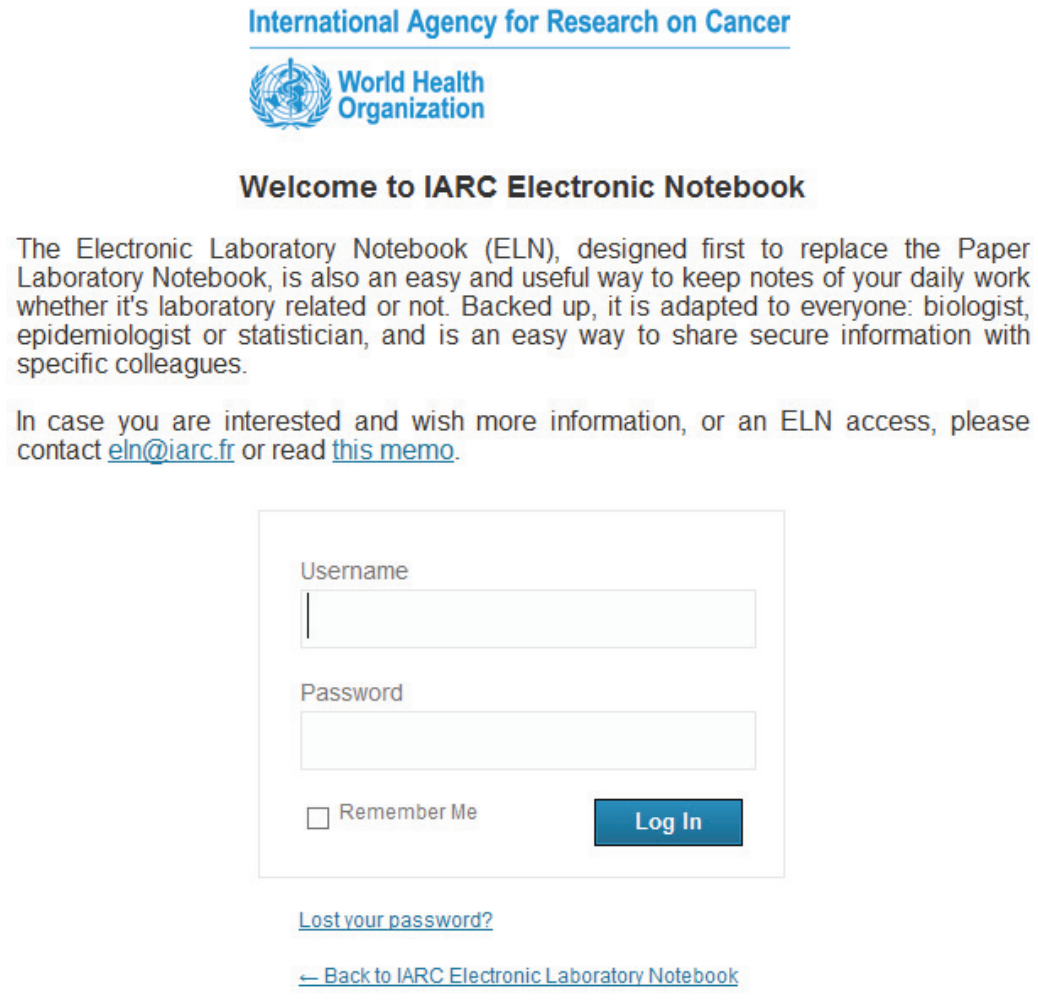


Figure 8: Screenshot of ELN login interface requiring username and password.

I] 3.2 Menus and main features

a) A user-friendly interface

Our aim was to develop an interface adapted to all users irrespective of their informatics skills. This involved facilitating the use and intuitive entry of information while not deviating from the ideals and usage of PLN which are the keys for wide acceptance of the ELN in the scientific community.

Figure 9 shows the welcome page which we have designed with clear menus on top, left and right sides containing quick access links to useful information to guide users through the recording of data.

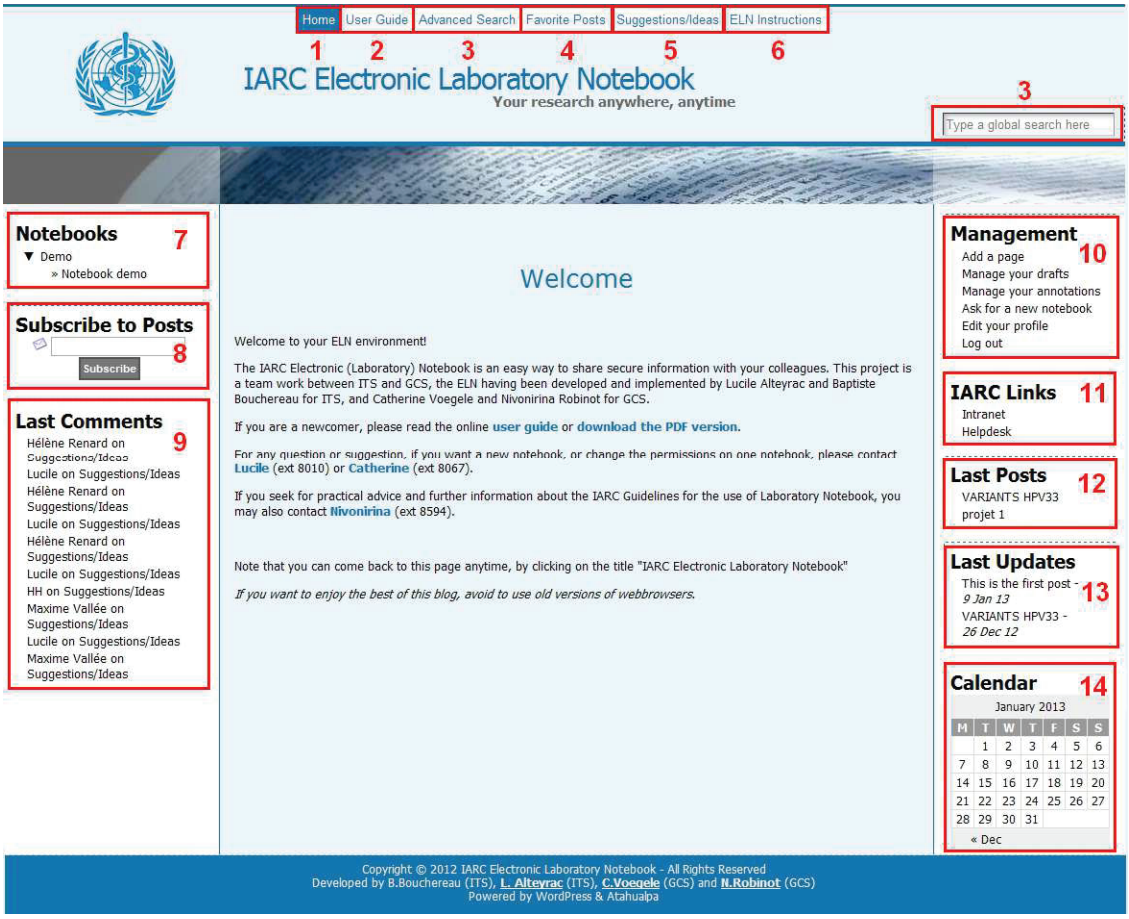


Figure 9: Screenshot of ELN welcome page with navigation links and tool boxes.

The main menu at the top of the ELN was set up directly in WordPress and includes:

- *1. The link to welcome page (home)*
- *2. The link to the online user guide (SOP for use of the ELN)*

- *3. The "Advanced Search" button*
- *4. The link to the "Favorite posts" (Customized WP Favorite Posts plugin)*
- *5. The link to Suggestion/Ideas page: useful for centralizing comments from users proposing new functionalities or improvements.*
- *6. The ELN instructions button (page with purpose of ELN, responsibility, data entry, corrections, table of contents and backups)*

The left sidebar menu displays:

- *7. The list of workspaces accessible by the user logged in collapsing the lists of notebooks (plugin "collapsing categories").*

  By default when the administrators create a notebook, they add the first page that explain the permissions associated with the notebook (who is allowed to read the notebook)
- *8. The "Subscribe to post" function corresponding to the "post notification" plugin. When activated, it enables users to receive emails each time a new page is added to one of the notebooks they have access to.*

  The plugin had to be modified so that the users can only subscribe to notebooks for which they have read access permissions. A subscription can be cancelled as requested.
- *9. The list of last comments that contains the 10 latest comments done in the notebooks the user has access to (with the name of the commentator and the concerned notebook).*

On the right sidebar, we show:

- *10. The Management tool box that enables the user to add a new page, manage its drafts, check the comments he received, ask for a new notebook or edit his profile (to change his password for example and a few other personal settings)*
- *11. Links specific for each group (for example bookmarks to web tools like the Sample Management System for IARC Biobank, the Laboratory Information Management System or the software to book PCR machines and other laboratory instruments). The links are added upon request to the administrators and are accessible depending on the users' permissions.*
- *12. The list of the last posts/pages that contains the titles of the 5 last posts written in the notebooks the user has access to, ordered from most recent to oldest.*

- *13. The list of the last updates that contains the titles of the 5 last posts updated in the notebooks the user has access to.*
- *14. A calendar in which the user can see which days posts have been published (in bold). Title(s) are displayed in a tooltip when you mouse-in over the calendar and the complete list of all posts published on a specific date can be viewed when clicking on the specific date.*

b) Advanced text Editor

The plugin 'CKeditor for WordPress' was installed to offer advanced editing features common to widely used word processors such as Microsoft Office™ (**Figure 10**). These features are essential for comprehensive formatting of the information stored in the ELN and thus easier reading and retrieval of this information:

- column formatting (specific plugin 'CSS Column')
- copy, cut and paste
- cancel and restore actions
- find and replace text, select all
- insert line, special characters or table specifying the number of rows and columns
- insert anchors or hyperlinks to cross reference other pages from the same notebook or from other ones or to link to any website. This includes linking to other documents in specific controlled location (on file sharing server for instance)
- format in bold, italic, underline, strikethrough, subscript, superscript
- insert bullets or numbering
- insert tabulations
- specify alignment of text (left, right, middle, justified)
- specify font and size of text plus colour of text and colour of the highlighting.

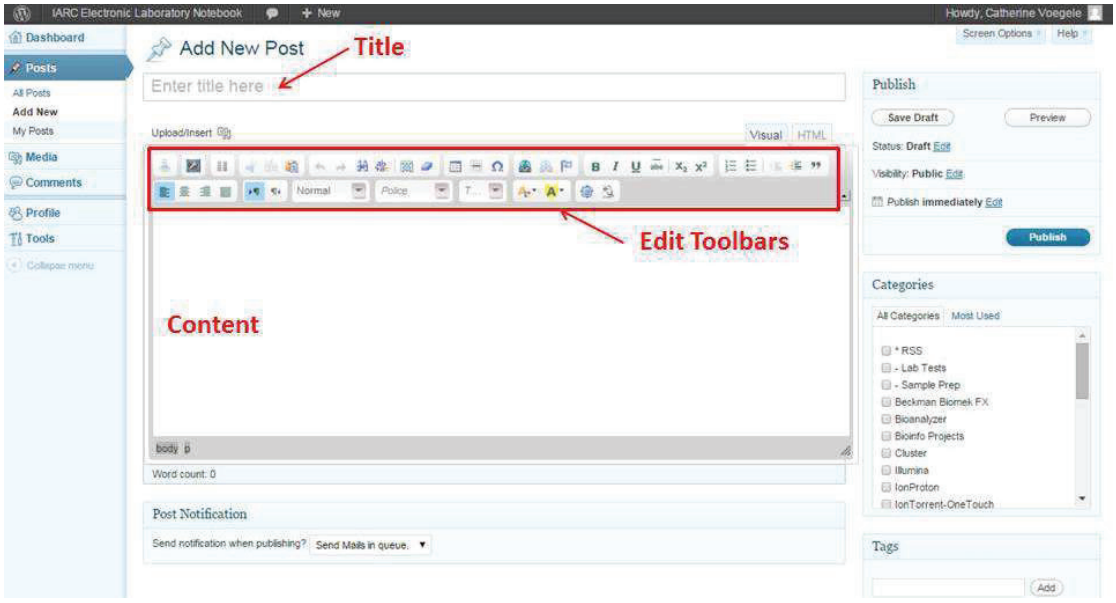


Figure 10: Screenshot of ELN text editor. *It includes a toolbar for editing features. Its size can be changed by the users by dragging the cursor of the button on the right bottom of the window or by pressing the maximise button in the editor toolbar.*

Page content is of course specific to each user and his work but follows most of the time the science outline with objective, material, method and results of the experiments. Therefore two specific templates for both laboratory's work (**Figure 11**) and instruments' maintenance (**Figure 12)** were developed and provided in the CKeditor toolbar through two specific buttons that enable their insertion  . Those are available as guidance on what kind of information is expected to be recorded especially for laboratory students and trainees.

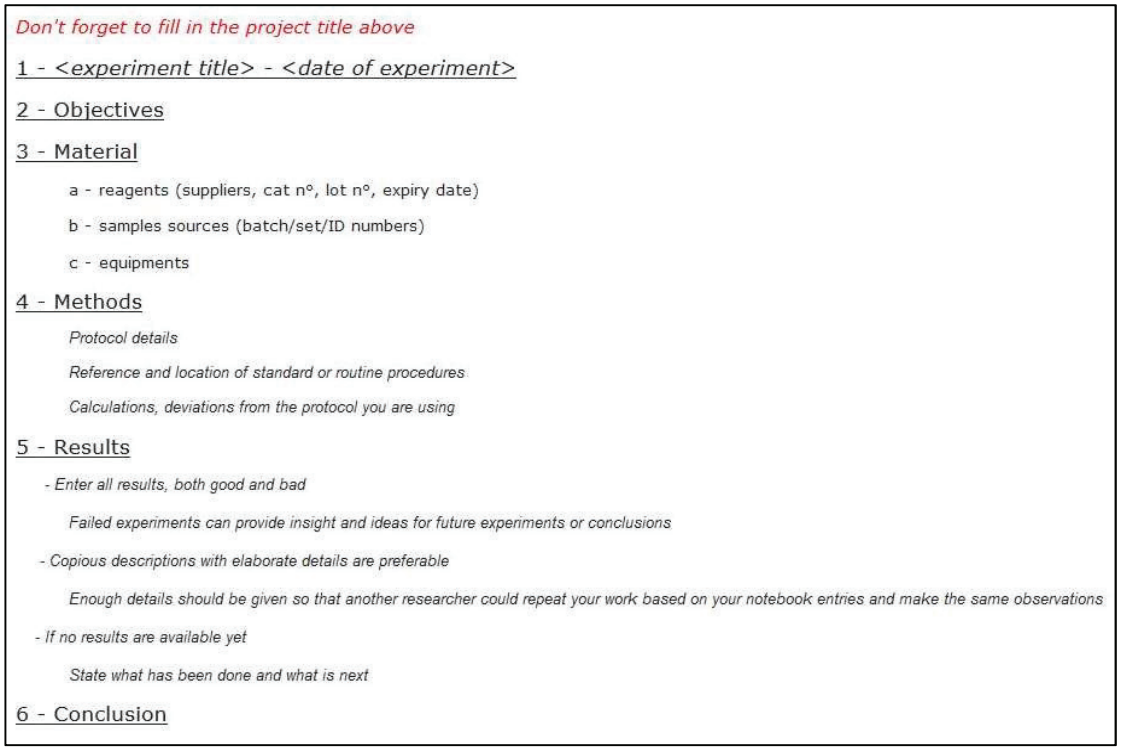


Figure 11: Template for laboratory experiments.

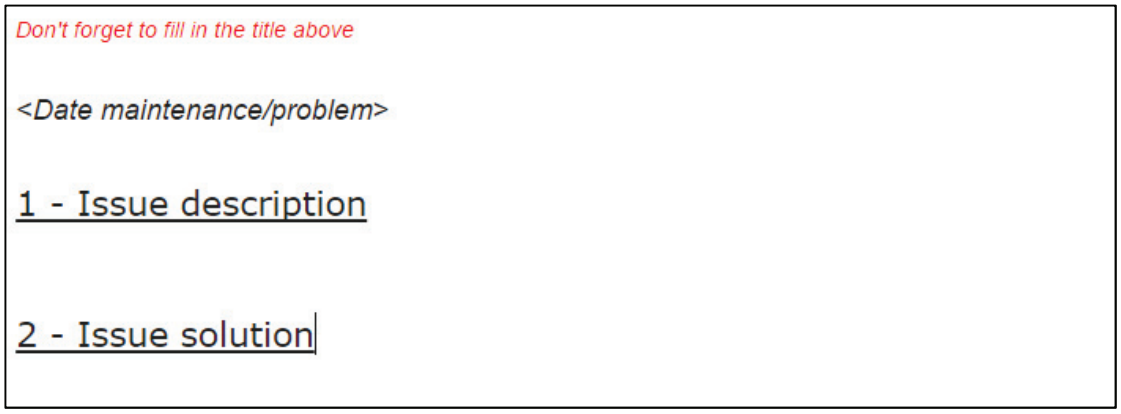


Figure 12: Template for instrument maintenance and troubleshooting.

Associated data management

One of the most important features of the ELN editor is the possibility to upload files generated in the laboratory for either insertion in or attachment to the pages. Different kinds of files are supported like Word documents, scripts, Excel files, PDF, pictures (for example screenshot) and also multimedia files. This allows incorporation of data directly outputted from instruments or computers avoiding

printing and pasting to PLN. For pictures, the users have the choice to either insert it in a page specifying the alignment and the size (only a thumbnail or in full size) or to attach the file that will then be available through a link. It is also possible to provide a title, a caption, a description and to associate a link to an URL.

The maximum upload size per file was set at 5 MB. Users can specify links on specific storage servers for files exceeding this size. All the uploaded files are stored in user's media library database with the timestamp and can be re-used anytime in any page of any notebook.

The CKeditor plugin manages as well the edition of comments with basic formatting and possibility to insert links. The text editor permits also to associate tags to each page by typing them or by selecting from pre-recorded most used ones. This enables to group posts with a same theme and thus facilitates further search for information.

All these editor functionalities together help users to save time without losing any information when they repeat some work using the same procedures.

c) Publishing usage and options

A user writes in his notebook by adding and editing pages. Users just have to type in or copy/paste some content, format it if they wish, attach files and then either save the page as draft to come back to it later or publish it which will make it visible by everyone who has the appropriate reading permission. Indeed by pressing on the publish button, the record is passed into the database as formal record. Subsequently all edits of this record will be stored in the database.

Each time a new page is created, it is automatically dated with the current date. For proper and accurate record tracking, this date could be changed to the date of experiment when users write up their records at a later date.

The auto-save functionality of WP has been adapted to be available for all posts saved as draft or published. This functionality is important in case some ELN users forget to actively save their work.

When publishing data, users have to select the notebook in which they want to publish by checking in the box on the right the "Category" (name of the notebook) among those they have access in writing (**Figure 13**). Because the list of notebooks can be quite long there is the option to visualise the most frequently used ones marked to facilitate their selection. The selection of the notebook is of course mandatory so in case users forget to select one, an error message will appear. This is permitted by to the plugin "Category reminder" that we installed specifically to prevent publication in "Uncategorized" notebook.

We also enabled a page with relevant permissions to be shared in several notebooks, function which can be very useful for some shared projects.

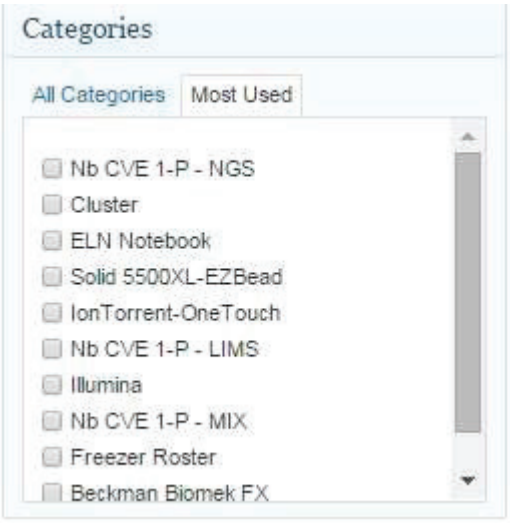


Figure 13: List of notebooks for selection at page publication. *The user has write permissions for all the listed notebooks.*

Though they are provided with general guidelines, users are left to define the manner of use of their notebook whether they prefer to create one page per day of experiment or one page per activity. Users can also record experimental research carried out over several days on a same page, in which case, they must take care of appropriate time stamping in the text in addition to the records automatic time stamping.

For shared ELNs – catalogs of scientific activities for a specific project or a specific laboratory instrument –, each allowed user can either create a new page to record his activities or continue filling an existing page even if created by someone

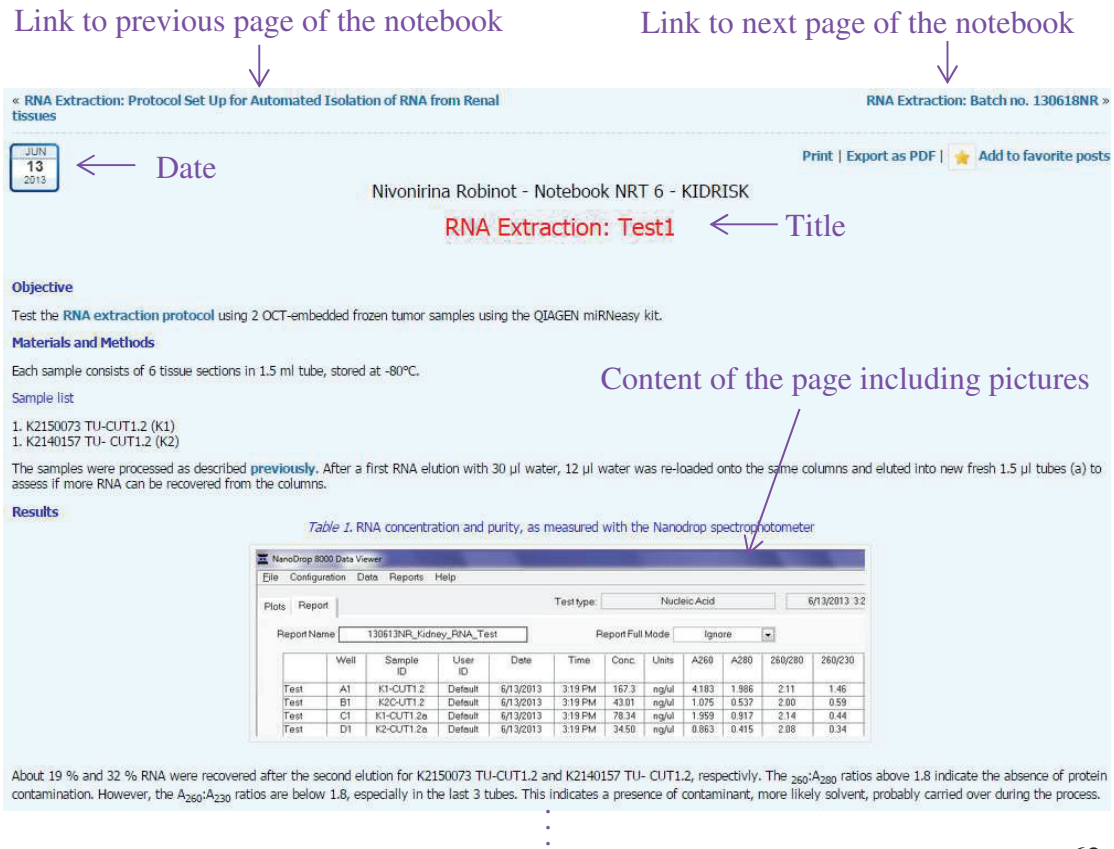
else. The recommendation is then to indicate clearly the name and date before the edition of the text.

d) Pages revisions

Once published, posts can be edited but not deleted and each edition is automatically stamped – at the bottom of the post, under the "Post Revision" section – displaying the name of the person that did the modification as well as the date and time of modification. The previous text is, in addition, always stored in the database. This function is enabled by the plugin "Post revision Display".

Much like a sentence may be crossed out in a PLN, corrections of published data are allowed using the "Strike Through" functionality of the ELN text editor. The accurate data can then be edited right after the cross-out. Deletion, contrariwise, is not permitted and each modification is automatically stored in the database and traced in the revisions. The latters can be viewed using the link displayed at the bottom of the post, in the "Post Revisions" section (**Figure 14**).

Additional information subsequent to publication can also be added in the comment editor section and are called annotations. They can be added by the owner or any editor of the notebooks but also by "simple" readers (authorized persons).



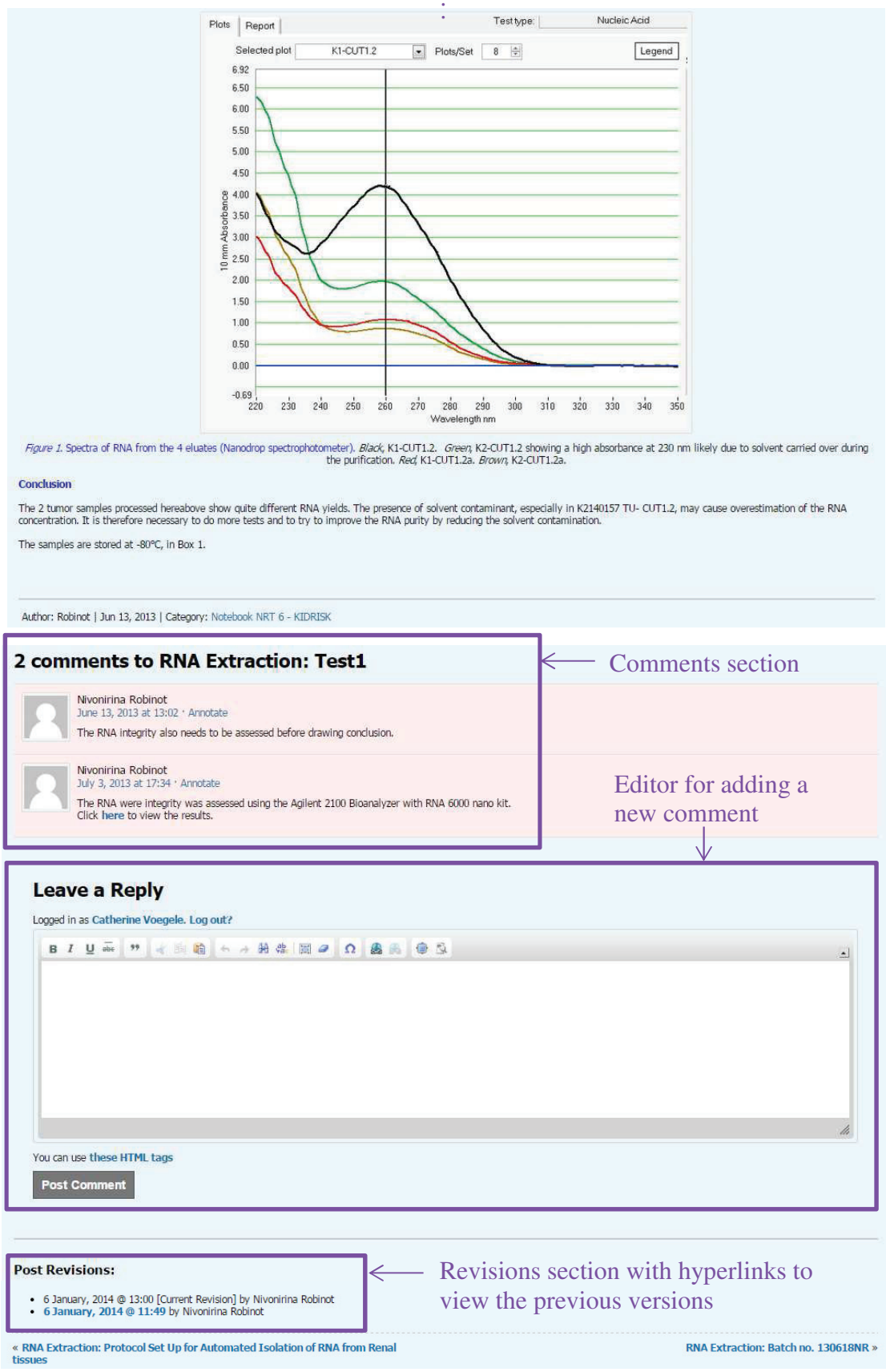


Figure 14: Example of a page containing the title, the content of the page, the comments and the hyperlinks to revisions.

I] 3.3 Other important features

Table of contents

WordPress generates automatically a table of contents (TOC for each notebook (**Figure 15**). For easier navigation within the table of contents when the number of entries is growing, we additionally customized the Atahualpa code and the plugin: "WP-Pagenavi to separate the table of contents in pages of maximum 25 entries. We also added the possibility to sort the table of contents by title, date of creation or date of last update descendant or ascendant.

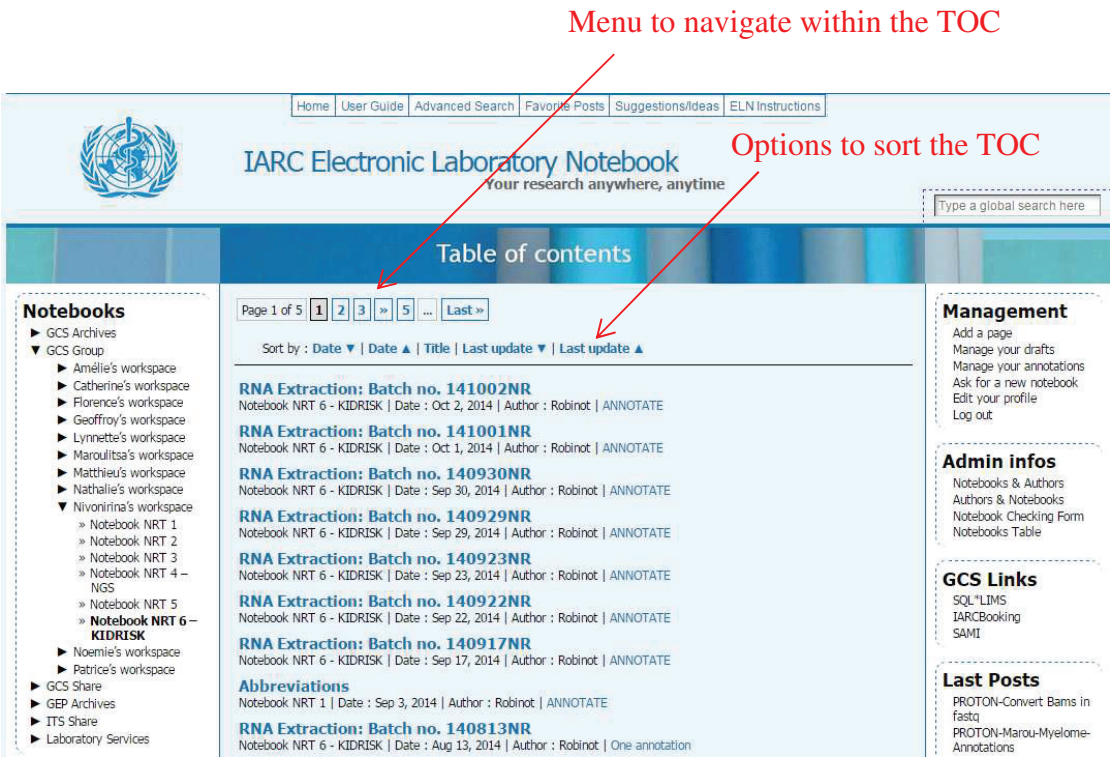


Figure 15: The ELN's table of contents. *The TOC displays the titles, date and author of each page of a notebook. It can be sorted by title, date of creation or date of last update.*

Search functions

The plugin "Search by category" enables users to make a search by keywords in one specific category (*i.e.* notebook or groups of notebooks).

There are different ways to search in the ELN:

- a simple search on one page using the classic search function of all web browsers (CTRL+F) that will search the entire web page for the specified keyword
- a global search by keywords using the text field on the top right of the screen – this search will look for the specified keyword in all the notebooks the user has access to
- a more advanced search that enables to restrict the search to a specific notebook or a specific group of notebooks. This function is accessible from the menu on top of the ELN pages
- a more specific search by tag which will return all the pages that had been associated with the selected tag independently from the content of the pages

Export in pdf

The plugins "WP-Print" and "Post2PDF enable the export of ELN pages to pdf. However the export is for the moment only possible page by page and not for an entire notebook.

I] 3.4 Security

Anyone owning an official IARC ELN, either individual or shared, is granted a controlled access password protected. The appropriate recording and secure use of the ELN credentials remains the responsibility of the user which is a critical point to convey during user training to be provided to each new user.

Nevertheless, the security of the system is additionally maintained through the following points:

- prevention from scientific fraud, falsification of data and protection of intellectual property: the very precisely defined access controls of who is able to see or edit the data and when, make it more secure than any paper notebook; all amendments of records are stored in the database; a detailed documentation – Standard Operating Procedure file – describes how to use the ELN with very precise instructions and guidelines.

In a manner analogous to how the PLN numbered pages can be noted if removed/teared, users cannot delete any post or any comment and deletion of part of text can be visualized.

- longevity: the html format of the raw data combined with appropriate daily backups on distant servers make it long-lasting and prevent from damage or deterioration; the ELNs cannot be misplaced, lost or accidently destroyed and at departure of staff, the ELNs are closed by the administrators and archived to read-only status; all the users' permissions are removed.

- protection from unauthorized and/or external access: the system is hosted on a secure server within a secure network behind firewall. We added the plugin "Login configurator" to force the login and to prevent any unauthorized connection to the ELN. Also, all the accesses and attempts of accesses are recorded.

I] 3.5 Evaluation

Evaluation of appropriateness for use at IARC

As the installation and uptake of an ELN are a significant change in scientific practice within a research institute, a trial period and consultation with the users were necessary in order to assess the systems' suitability for recording the scientific activities of the institute. The role out of the ELN at the Agency involved a two-step process.

Firstly a prototype of the ELN was presented to the IARC Laboratory Steering Committee (LSC) which is composed by the senior scientific staff leading the Agency's laboratory groups. The role of this committee is to oversee the IARC core laboratory facilities and to advise the Director on their most efficient use. During the presentation of the ELN, the LSC provided feedback regarding the ELN structure and

function as well as appropriateness for its expected use across the different laboratories. The primary concerns raised by the members of the LSC were:

- the definition of the accesses and permissions;
- the guidelines for recording of the scientific activities in different types of notebooks;
- the attachment of files;
- the tracking of modifications;
- the archiving.

As seen previously all these issues were addressed in our ELN prototype. The committee thus decided to launch and coordinate the second step of the ELN implementation which was the testing of the ELN beta-version by the main users, in this instance, laboratory research assistants. The LSC coordinated this pilot test by selecting a minimum of two testers in each research group for a period of six months. This test involved the "Epigenetics" group (EGE), the "Molecular Mechanisms and Biomarkers" group (MMB), the "Molecular Pathology" group (MPA), the "Infections and Cancer Biology" group (ICB), the "Biomarkers" group (BMA), the "Genetic Cancer Susceptibility" group (GCS) and the "Laboratory Services and Biobank" group (LSB). A few bioinformaticians were also volunteers to test the ELN. We had thus testers representing of a large panel of laboratory and non-laboratory activities.

The trial was thus performed by 32 users in total, which constituted a working group that met to give their feedbacks regarding their first experiences with the tool and discuss any aspects for further needs or improvements. The team reported no major problem and in contrast, considered it a main advance to have permanent and organized electronic records of experiments and other related matters that can be searched easily and shared. The answers brought to the questions raised as well as the rapidity of adapting the prototype to fulfil all requests such as the attachment of files and archiving were estimated satisfying.

The results of this trial were reassessed and discussed within the LSC. The exploratory phase successful and particularly that the ELN developed was judged conformed to IARC rules and regulations on laboratory notebooks **(Figure 5)** as well as standards relating to electronic data protection and computer back-up issued by the Food and Drug Administration (FDA) Code of Federal Regulations Title 21 ([75]).

The LSC thus provided a report regarding the testing and implementation of the ELN to the IARC Director recommending the official introduction of the ELN in

the Agency as from January 2013 and emphasizing on the main advantages which were summarized as follow:

"ELN cannot get lost, which is the case for 10 per cent of the 1400 PLNs; it can be used to link to data from specific laboratory equipment; all kinds of attachments, hyperlinks and comments from supervisors can be included obtaining the same level of detail in documenting as with PLNs; and it is easy to read, to search, and to share with colleagues. It may as well encourage a sharing attitude and unlike PLNs can evolve".

The Director approved the introduction of the ELN. In addition it was decided that all new staff would be required to use the ELN for scientific record keeping and that the paper notebooks would be progressively phased out in favour of the ELN for the existing staff members. The ELN was then expanded Agency wide and proposed to all staff irrespective of their type of work providing a common platform for data capture across all disciplines, whether laboratory, epidemiology, biostatistics or bioinformatics-oriented.

Summary of current use

The system is indeed currently used by more than 100 persons within the Agency. Initially developed for the wet laboratory users, it is actually used at 50% by non-laboratory staff for keeping track of their daily work and sharing information with their colleagues. After 2 years of use, we have had a total of 172 users including 70 left staff. We have 110 living notebooks, 58 shared notebooks and 60 already archived notebooks. 12 research groups are using the ELN out of which 7 are laboratory research groups. Inside GCS, we have 16 shared notebooks and 12 archived notebooks. From the 16 shared notebooks, 12 are for robots, instruments and HPC maintenance and troubleshooting and 1 for the storage of the various laboratory protocols. They are all shared in reading and writing with all GCS group members. The remaining 3 shared notebooks are for specific projects with limited access to the participants in the projects.

Figure 16 provides statistics on usage of the ELN of a given time period demonstrating that the tool is actively used. We continue receiving feedbacks from users, generally positive and constructive which appears consistent with the tool serving the community for which it was designed.

Finally we have also tested the recovery of the backups to be sure to be able to restore the complete ELN database in case of problem without losing any piece of information.

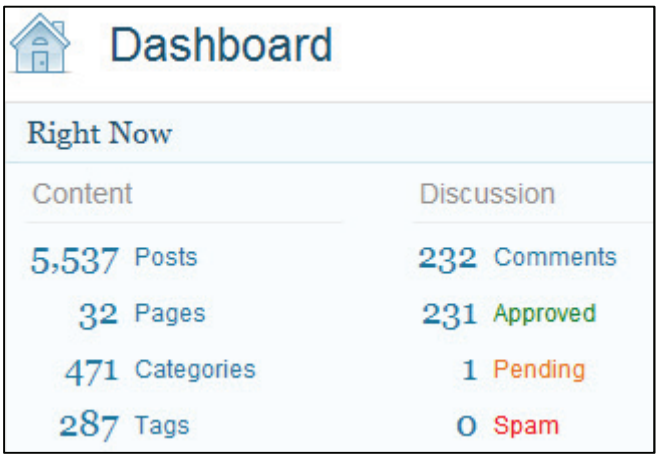


Figure 16: Statistics of use of the ELN.

On the top the screenshot shows basic statistics available to ELN administrators through the WordPress Dashboard: the IARC ELN contains 5,537 pages and 232 comments (up to July 2015). The page on ELN ideas and suggestions is the most commented one with 23 comments. One third of the comments are made by people who did not write the page showing that the sharing is effective.

On the bottom the screenshot from WassUp plugin shows the number of ELN visits and page views that occurred between the 8th and 20th of July 2015: respectively 421 visits and 1452 page views. Unfortunately these statistics have some limitations as we cannot distinguish data entry visits from data visualization visits.

Evaluation by user survey

To go further in providing the best possible tool in the future, we have set up a forum for suggestions and ideas and we conducted a formal satisfaction and feedback survey with all active users asking for frequency of use, type of use, advantages, additional functionalities that may be useful and drawbacks plus any general comment. For this purpose, we developed a questionnaire using Surveygizmo tool ([76]) and asked all regular users by email to answer the survey providing them with a web link to access it. The regular users were defined as having access the ELN at least once in the last 6 months. The survey was anonymous. Though, we gave the possibility to users to give their name in case they would like to be contacted regarding a specific request or comment.

Users had two weeks to complete the survey with two reminders. 35 people responded over 71 invited. The response rate being only 50%, we recognize that there may be some bias in the results and conclusions we can draw from the survey. Statistics on the use are summarized in **Figure 17**.

There is almost unanimity within ex-PLN users to say that the ELN has improved their work documentation. Just one person regrets the fact that he cannot take the ELN with him in the laboratory when there are no computers, smartphone or tablet available. The main advantages reported are “easiness of use”, “clarity for reading” and “work reduction by time saving”. The latter is probably linked to the copy/paste functions. The most useful features are the table of contents, attachment of documents, linking, search functions and sharing. Interestingly, almost half of the respondents pointed out a lack of more advanced editor features for example to draw directly on, format embedded pictures, perform calculations or make graphs.

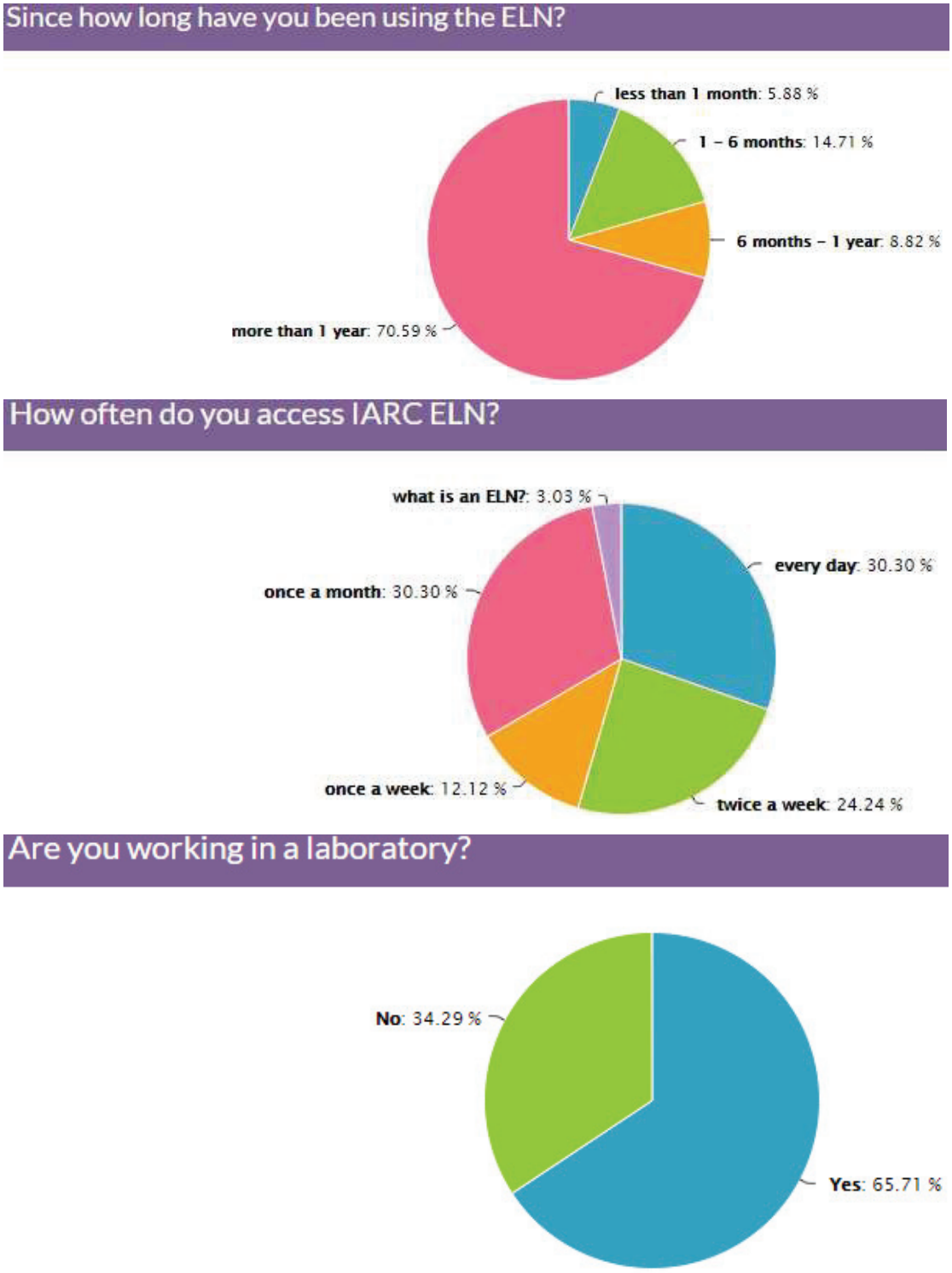


Figure 17: ELN survey results. *70% of the 35 respondents have been using the ELN for more than one year. 30% use it every day and 67% at least once a week, which gives an additional weight to the answers. The survey confirmed a strong use by non-laboratory staff (35%) so we would like to investigate further whether they would have additional personal or technical specific needs.*

I] 4. Discussion

I] 4.1 PLN vs ELN: advantage to the ELN

The table below compares and contrasts the advantages and disadvantages of the PLN and ELN (**Table 1**).

	PLN	ELN
Use	Easy	User-friendly interface, easy to understand even for persons with limited computer skills
Modifications and corrections	Strike-through and rewrite	Each update is stored automatically with userstamp and timestamp
Table of contents	Manual	Automatic but no page number
Laboratory template (material and methods...)	Manual	Button to insert text template automatically
Insertion of documents	Print and paste – problematic with voluminous data	Attach files (faster)
Pages numbered	Yes	No but time-stamped
Sorting	Unavailable	Possible by date of creation, title and date of last update
Links within the pages	Indication of page number	Direct hyperlinks to pages also between different notebooks
Removal of pages	Forbidden but could happen	Impossible
Annotation	Possible	Possible with automatic userstamp and timestamp
Reading	Handwriting can be problematic	Easy
Sharing	Possible "by hand" but limited	Easy especially for collaborative projects
Searching in	Long and laborious	Easy thanks to the advanced search tool
Accessibility	In one specific place	Everywhere through connection to the internal network (directly or via Virtual Private Network)
Access safety	Low control of who reads the notebook	Specific user permissions for reading and editing
Long-term storage	Cupboards	Raw data in long-lasting HTML format Backups in a secure place => easier to archive and preserve, and less likely to misplace, lose, or accidentally destroy
Cost	Quite high (23,32 euros per notebook)	It is open-source with no licensing cost. It requires only a small server and some time for installation and administration (around 2 hours per week for an institution of 300 people with quite high staff turn-over)

Table 1: Comparison between PLN and ELN.

As seen above, ELNs provide significant advantages over PLNs. Data can be quickly entered, retrieved, located, shared and are easy to archive and backup; data including text, images, attached documents, links. They allow direct incorporation of data from instruments, replacing the laborious practice of printing out data to be stapled or pasted into paper notebooks. Captured information is organized and available in a common and safe platform. This is particularly important in institutions with high personnel turnover.

The ability to reuse or repurpose details of experiments recorded previously in ELNs, to include electronic data and embedded hyperlinks as well as the easy-to-use search functions have proven to bring higher quality documentation and gain of time for data recording and retrieving ([77] - Butler, 2005). Glyn Williams specialist in software Research and Development (R&D) for large biotechnology and pharmaceuticals companies estimate the efficiency gains of 15-25% for biology ([78] - Dutton, 2006).

IARC ELN tool has been designed so that an ELN can belong to one single person or be shared within and/or between research groups when being specific to one common project or one instrument or equipment. It is a multidisciplinary tool which is flexible enough to accommodate different types of data including centrally updated manuals or SOPs as well as less traditional data such as proposal developments or meetings notes. It is therefore not only applicable to laboratory work but also useful to epidemiologists, statisticians or any other researchers for their daily notes.

Our ELN offers thus additional means of communication which should improve and strengthen collaborations and exchanges between researchers providing an important advantage for a global work recording in multidisciplinary research organizations. This is quite a big change for scientists switching from a personal recording activity where the notes and data were rather private to a capture of work's details for them to be read and completed by collaborators.

Finally, the control of very precisely defined accesses together with appropriate backups, make it more secure than paper notebooks.

I] 4.2 Cost

Several factors come into evaluation of the ELN cost: the material, the time of installation, the time of development and time of maintenance.

Material cost

As seen previously, the tool requires only a small basic server (an Intel Xeon CPU at 2.53 GHz, 4 GB of RAM and a hard disk of 100 GB). There are 3 different possible options:

- installation on a virtual web server. That is the solution we opted for, using a web server already in place for other web applications implying no additional specific cost for IARC.
- installation on a physical machine with the minimal WordPress requirements. The standard price would be in the order of 1,200 euros (Dell price for a basic server) but can vary depending on contracts with suppliers, partnerships, academic discounts.
- installation on a cloud which would cost around 12 euros per month (price from OVH.com the 3rd worldwide internet service provider).

Related material costs such as the required network and associated secure firewall are assumed to be in place and are therefore not factored into the current cost evaluation.

Installation cost for other research institutes

While the tool is completely free and open-source, the set-up in other research institutes requires some IT specialist time which cost is difficult to estimate as well since dependant on the skills and the salary of the responsible which can be highly variable between institutions and countries. However for one IT specialist following our instructions it could be performed in 3 days including server installation, WordPress and plugins installation and configurations plus creation of first users, groups and notebooks. With an estimated salary of 2,500 euros per month, it would thus cost around 350 euros.

Development cost for IARC

We evaluated the person time spent on the development of the ELN tool at a total of 5 Full Time Equivalent (FTE) months:

- 1 FTE month for the definition of the needs and specifications as well as the choice of the tool to use for the development of the ELN (WordPress)
- 3 FTE months for the configuration and adaptation of WordPress to create the ELN
- 1 FTE month for testing of the appropriateness of the ELN relative to IARC needs as well as assessment of the tool performance.

Assuming a base salary of 2,500 euros/month, this gives an estimated cost of 12,500 euros.

Maintenance cost

Our system as set up requires indeed some administrative tasks for adding new users and new notebooks when needed and specifying the access permissions. This represents a variable work-load depending on the staff and trainees turnover knowing that it can take between 5 to 10 minutes per new user. It can be quite light for small laboratories but may be heavier for large research centres involving several hundreds of employees. This has an IT time cost too but maintenance is required equivalently when using commercial tools and has therefore an equivalent cost.

From the developers points of view, WordPress being an evolving CMS it shows an advantage but also a drawback. Indeed new versions are coming out regularly for both WordPress and the many plugins we have installed requiring significant additional work for the administrators to re-test all the features at each upgrade. It is crucial to ensure that the security is preserved and also to perform tests of non-regression. Indeed in some cases upgrade can lead to a loss of functionalities and therefore should not be applied. Finally since web-based the ELN needs to stay compatible with the new versions of the browsers continuously evolving. We estimated the time spent on the upgrades (which are optional but recommended to keep an up-to-date tool) at 20 days per year since they imply re-testing of all the plugins to ensure they are all still compatible and functioning as they should and used to.

Total cost estimation for set-up

Adding the cost of material, installation and development, we end up with a total cost of 12,500 euros for the implementation of the ELN at IARC and between 350 and 1,550 euros for its implementation in other institutes (**Table 2**). We could have probably found a commercial tool less expensive especially since prices went down since the start of our project thanks to the competition. We chose to privilege the possibility to have a tool completely adapted to our needs and to stay independent from vendors for future modifications, evolution or yearly per-users licences that could imply a much higher long-term cost. It is also the choice for stability in order not to have an obsolete unmaintained tool in a few years.

COST	Cost for IARC (in euros)	Cost for other institutes (in euros)
Material	0	From 0 to 1,200
Installation	NA	350
Development	12,500	NA
TOTAL	**12,500**	**From 350 to 1,550**

Table 2: Evaluation of the cost of ELN implementation at IARC and at other institutes.

As a comparison, commercial tools prices can vary a lot from 200 dollars (i.e. low-cost Accelrys BIOVIA Notebook) up to thousands of dollars depending on their functionalities and applications ([79] - Smith, 2014). Like our ELN, some are also flexible across different disciplines (ie. Agilent's OpenLAB Electronic Lab Notebook). Clive Higgins, vice president of marketing and informatics at PerkinElmer believes that "a higher-priced notebook should be purchased when requirements do not match common scientific workflow, data and analysis procedures or it should be purchased if you are in a highly-regulated environment where custom software development is required to exactly match regulators processes and procedures" (i.e. very complex certification norms).

Otherwise for simple and limited use, it is possible to find now specific smartphones and tablets' applications for scientists to take laboratory notes. Examples

includes Notebooks, Lab assistants, My Lab, Sparklix, LabArchives, eCAT and Wingu ([80] - Smith, 2012).

I] 4.3 Certifications

Some laboratories would need to comply with ISO norms or other certifications. We did not investigate this point much in detail but the tool is conforming to the international practice guidelines and data security and protection standards (FDA – Code of Federal Regulations Title 21 ([75])). Indeed in order to address the issue of fraud and falsification of data and ensure authenticity and integrity, our ELN follows specifically the rules dealing with control and management of electronic records:

- Access to data is controlled and unauthorized access is not possible
- The records are protected, author-stamped and time-stamped as well as appropriately backed-up. Amendments of records are tracked and stored.
- It includes procedures for identification, collection, indexing, access, filing, storage, maintenance and disposal of records
- Data is stored in appropriate IT environment to prevent damage or deterioration and data is easy to retrieve.
- Users are trained for correct use of the ELN and have permanent access to detailed documentation. They are informed of their responsibility for their records.

Signatures are an issue in case of need to meet legal and juridical values of the records. Written signatures on paper are relatively hard to forge, copy or delete. For electronic ones, public-key cryptography guarantees their security while transiting on networks by using a combination of two keys that are mathematically linked: one public for encryption and one private for decryption ([81]).

Evidence that a given researcher authored specific discovery or content is based on four assumptions: 1- the organisation is trustworthy, 2- the assurance of the identity of the researcher is effective, 3- the procedure to create certificate and signatures has not been compromised and 4- the researcher has kept his private key secured (not shared and not stolen). The first two ones are similar to those for paper

notebooks and the last two can be checked using secure sockets layer encryption (SSL) and other secure IT tips like protection of information flows and logs ([81]).

In our ELN all the records and comments are user-stamped in our ELN with the user's name based on his credentials but encryption of this information as e-signing may be interesting to investigate as well as an additional layer of security.

I] 4.4 Perspectives

Firstly adopted by the industrial laboratories from the 80's on ([82] - Figueras, 1987), the first web-based ELNs were only introduced in 1998 ([83] - Skidmore et al., 1998). Since then ELNs developed themselves towards modern tools which made them become more and more popular also in academic research fields. Indeed very recently the French "Institut Curie" has developed an interesting free and open-source ELN: eLabFTW ([84]) but the latter is still a beta version under testing and is lacking the fine-grained multi-users permissions definition that make our ELN a unique tool over the other open-sources ones available to the scientific community.

The completely paperless laboratory promised since early 1990's has not yet taken off but things are changing ([77] - Butler, 2005). Only the lack of clear requirements for legal defensibility of digital records is still slowing the ELN adoption. Some people are still concerned about the protection of their intellectual property regarding the legal value and/or possibility of falsification of ELN records but their number tend to decline ([85] - Elliott, 2007). Community efforts to define best practices such as those within CENSA and electronic signatures recognition are an important drive towards wide ELN acceptance. Indeed in 2000 the US government and other worldwide governments and authorities (e.g., UK government, European Central Court) issued new laws stating that all electronic records have the same validity and are subject to the same rules of evidence as paper records ([86] - Nehme and Scoffin, 2006).

As of 2008, year over year growth is still above 20 percent, making ELN one of the fastest growing informatics technologies ([87] - Elliott, 2008). This led to bigger choice for the customers from a variety of vendors associated with a decrease in price and increase in quality thanks to this commercial competition. The main drawback is the difficulty to find the right solution and a stable system over time

based essentially on the vendor information. Installing and testing tools is long and costly.

This growing demand has also led to the development of new open-source ELN since the start of our ELN development project. Mr Nicolas Carpi listed in **Table 3** the top free and open source lab notebook software available in April 2014 ([88]). 5 of them are multi-disciplinary: Bikalabs which is actually a LIMS, eLabFTW which is still a beta version under testing and is lacking the possibility to define precise multi-users permissions, MyLabBook which is based on Drupal an alternative to WordPress CMS, MonsterJournal which as well does not provide the possibility to define precise multi-users permissions and finally Labtrove which is a blog ELN similar to our ELN enabling multi-users accesses but which does not distinguish between read and write permissions.

ELN Software	License	Multiplateform	Type of Lab	Structure	Language
Bikalabs	GPL/AGPL	YES	Any	Server – Client	Python (plone)
eLabFTW	AGPL	YES	Any	Server – Client	PHP + MySQL
Indigo ELN	GPL	YES	Chemistry	Server – Client	Java
MyLabBook	GPLv2	YES	Any	Server – Client	PHP
Labtrove	Various	YES	Any	Server – Client	PHP
ELN by jdmyers	Other	YES	Biology	Local client only	Java
Cynote	GPLv2	YES	Bioinformatics	Server – Client	Python
LabJ-ng	GPLv2	YES	Chemistry	Server – Client	Perl
MonsterJournal	GPLv2	YES	Any	Server – Client	Perl
OpenEnventory	GPLv2	YES	Chemistry	Server – Client	PHP + MySQL

Table 3: Top ten open-source lab notebook software *(from ([88])).*

The tool we developed, since designed with the users for the users following their needs, has been well adopted as the latter could see an improvement relevant to their own work ([89] - Bruce, 2002). The survey we made with the users indicated that we should focus our future efforts on editor's improvement to provide an even better tool.

Finally, following the publication of this work in Bioinformatics to make the tool available for the scientific community ([74] - Voegele et al., 2013), we have been contacted by several potential users from other institutions all over the world (London, Munich, San Francisco, Alberta, Sao Paulo) who were interested in installing our ELN in their institutes. Though all steps are well explained in the

supplementary data to enable them to implement the ELN with appropriate IT support, it would be useful to make a ready-to-use package for the installation of the ELN or a virtual machine image.

Publication

Voegele C et al., Bioinformatics, 2013

“A universal open-source Electronic Laboratory Notebook”

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 13 2013, pages 1710–1712
doi:10.1093/bioinformatics/btt253

Databases and ontologies

Advance Access publication May 3, 2013

A universal open-source Electronic Laboratory Notebook

Catherine Voegele[1,*], Baptiste Bouchereau[2], Nivonirina Robinot[1], James McKay[1], Philippe Damiecki[2] and Lucile Alteyrac[2]

[1]Genetic Cancer Susceptibility group and [2]Information Technology Services group, International Agency for Research on Cancer (IARC), 69003 Lyon, France

Associate Editor: John Hancock

ABSTRACT

Motivation: Laboratory notebooks remain crucial to the activities of research communities. With the increase in generation of electronic data within both wet and dry analytical laboratories and new technologies providing more efficient means of communication, Electronic Laboratory Notebooks (ELN) offer equivalent record keeping to paper-based laboratory notebooks (PLN). They additionally allow more efficient mechanisms for data sharing and retrieval, which explains the growing number of commercial ELNs available varying in size and scope but all are increasingly accepted and used by the scientific community. The International Agency for Research on Cancer (IARC) having already an LIMS and a Biobank Management System for respectively laboratory workflows and sample management, we have developed a free multidisciplinary ELN specifically dedicated to work notes that will be flexible enough to accommodate different types of data.

Availability and implementation: Information for installation of our freeware ELN with source codes customizations are detailed in supplementary data.

Contact: voegele@iarc.fr

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

Received on March 1, 2013; revised on April 15, 2013; accepted on April 30, 2013

1 INTRODUCTION

The requirements for the development of International Agency for Research on Cancer (IARC) Electronic Laboratory Notebooks (ELN) consisted of integrating all the features existing in the paper-based laboratory notebooks (PLN) together with additional functionalities to enable easy and intuitive navigation and information search capacities upon the notebooks (Wright, 2009), following the definition of the Collaborative Electronic Notebook Systems Association: 'System to create, store, retrieve and share fully electronic records in ways that meet all legal, regulatory, technical and scientific requirements; records being collections of information or data associated with an experiment to enable a suitably skilled person to repeat it'. The specifications included (i) recording of experimental data in a flexible way through a common platform suitable for the various research activities of the institute, and at the same time maintaining the rigorous scientific practices required for day-to-day notekeeping (traceability of each single modification), all within a highly secure environment; (ii) improving lab efficiency and work productivity by facilitating protocol and data sharing, record keeping, particularly in terms of limiting repetitive tasks; (iii) providing an efficient means by which to search archived material; (iv) being free, open-source and web based requiring no installation of software on user's computer; and finally (v) enabling long-term and secure storage in basic electronic format (text).

*To whom correspondence should be addressed.

2 METHODS

Based on these specifications, we have tested several tools (cf. Supplementary Material) and finally chose WordPress 3.4.1 with Atahualpa 3.7.6. theme. WordPress is a modern well adopted Content Management System (CMS), easy to design and to customize, written in PHP and based on a MySQL database. This user friendly open-source publishing software was found to fit the best the Agency's needs mainly because it enables the set-up of multi-users with tight access controls and provides a lot of options for customization.

A simple Linux server hosts both the web application and the database. A large number of free plugins offering different functionalities have been tested to find the most adapted ones (cf. Supplementary Material). The system is currently being developed, maintained and administrated by an IT developer and a bioinformatician.

The correspondence between the CMS terminology and the laboratory notebook has been set up as following: categories are equivalent to notebooks or groups of notebooks, posts are equivalent to pages and comments to annotations.

3 RESULTS

ELNs provide significant advantages over PLNs. Data can be quickly entered, retrieved, located, shared and are easy to archive and backup: data including text, images, attached documents, links. An ELN can belong to one single person or be shared within and/or between research groups when being specific to one common project or one instrument or equipment (for example, for maintenance follow-up and troubleshooting issues). It can be used for proposal developments, meetings notes as well as centrally updated manuals or SOPs.

We will detail in the following chapters and also in the Supplementary Material how we managed to fulfill all the requirements and rules of a PLN (*) and the additional features making the ELN a far better alternative to PLN.

3.1 Access

The tool developed is entirely web based and therefore accessible through any web browser connected to the internal

firewall-protected network, making it accessible from both laboratories and offices. The access is controlled by personal login and password (notably thanks to the 'Login configurator' plugin). The IARC ELN is compatible with PC, Mac, tablets and smartphones. The access to data in reading and editing is defined very precisely and controlled by specific user permissions thanks to the 'Role Scoper' plugin (from completely private to public). The system offers the option for some notebooks to have more than one author or to be read only.

By default we put in place a 3-workspaces design for each research group in the Agency:

- 'group workspace' dedicated to personal notebooks and containing one sub-workspace for each member of the group, each sub-workspace containing one or more notebooks (each member of the group has write access to his own notebook and read access to the other members notebooks).
- 'share workspace' containing notebooks for instruments, robots or shared projects (with read and write access for the whole group).
- 'archives workspace' containing all closed and archived notebooks of staff who have left the Agency (set to read-only access).
- The users are offered the possibility to leave comments on all the notebooks that they have read access to.

3.2 Features

Our ELN system includes the following main functionalities:

- Easy to navigate interface and menus providing access to all users irrespective of their informatics skills - cf. Fig. 1 (thanks to widgets, custom code and 'Collapsing Categories' plugin).
- Advanced text editor similar to the well-known office ones and in which specific templates have been integrated for laboratory work including 'Title of experiment, objectives, material and methods, results and conclusion'* and instrument troubleshooting with 'Date, issue description, solution and conclusion' ('CKEditor' plugin with custom code).
- Insertion and attachment tool for different kind of files such as pictures, Word documents, Excel sheets, pdfs, screenshots, multimedia, hyperlinks (among which cross-referencing of notebook pages), plots, computer outputs (which allow incorporation of data from instruments avoiding printing and pasting into PLN).
- Automatic table of contents* of pages/records providing the possibility to sort by title or date (customization of Atahualpa code and 'WP-PageNavi' plugin).
- Search tool by keywords ('Search by Category' plugin).
- Userstamp and timestamp for each update or annotations of a page* (recorded by default by WordPress and displayed by 'Post Revision Display' plugin).
- Lists of favorite pages (Customized 'WP Favorite Posts' plugin).
- Annotations or comments.
- Post notification plugin that enables the user to be warned by email each time a new post is added into some selected notebooks ('Post Notification' plugin with custom code).
- Export to pdf ('WP-Print' and 'Post2PDF Converter' plugins).
- Management of users permissions by ELN administrators ('RoleScoper' plugin).

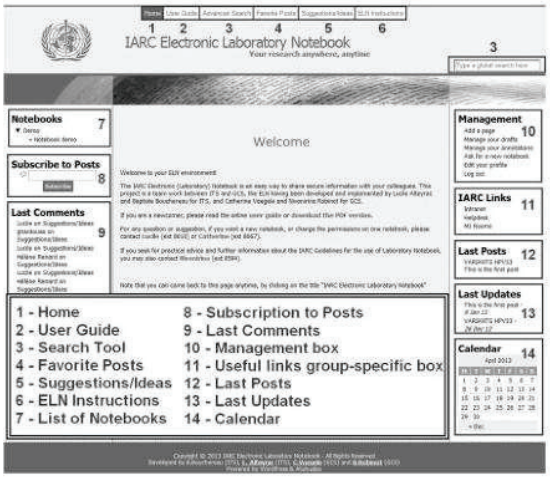


Fig. 1. ELN main page with navigation links and tool boxes

3.3 Security

Security is ensured in terms of (i) prevention from scientific fraud and protection of intellectual property: the fine-grained access controls make it more secure than any paper notebook; all amendments of records are stored in the database; a detailed documentation describes how to use the ELN, which should prevent any deviation from the protocol; (ii) longevity: the html format of the raw data combined with appropriate daily backups makes it long-lasting and prevented from damage or deterioration; the ELNs cannot be misplaced, lost or accidently destroyed; at departure of staff, the ELNs are closed and archived to read-only status; (iii) protection from external access (secured server and network, accesses and attempts of accesses are recorded).

4 CONCLUSION

After 6 months of testing and validation, the ELN accounts for around 50 regular users, 100 notebooks and over 1300 pages. This exploratory phase having been a complete success, the Agency has decided to switch from the PLN to the ELN since 1st of January 2013. As IARC's LIMS (Voegele, 2007) and IARC's Biobank Management System (Voegele, 2010), the IARC ELN has been largely adopted and is now intensively used within the agency, as it has proven to bring higher quality documentation, gain of time for data recording and retrieving as well as higher security. It should also strengthen collaborations and exchanges between researchers within the Agency. Last but not least it demands very

low resources and is scalable whatever the size of the institution or laboratory is (cf. Supplementary Material).

ACKNOWLEDGEMENTS

We would like to thank Philippe Boutarin, Brigitte Chapot, Florence Le Calvez-Kelm, Christopher Jack and all the ELN testers.

Conflict of Interest: none declared.

REFERENCES

Voegele,C. *et al.* (2007) A Laboratory Information Management System for a high throughput genetic platform aimed at candidate gene mutation screening. *Bioinformatics*, **23**, 2504–2506.

Voegele,C. *et al.* (2010) A sample storage management system for biobanks. *Bioinformatics*, **26**, 2798–2800.

Wright,J.M. (2009) Make it better but don't change anything. *Autom. Exp.*, **1**, 5.

Chapter II – Laboratory Information Management System (LIMS)

II] 1. Introduction

II] 1.1 Definition

LIMS are software packages used for tracking and monitoring laboratory work. They integrate tools, users and procedures for management of laboratory workflows, quality control, equipment configuration and maintenance, follow-up of consumables and reagents stocks and storage of experimental data. Originally dedicated to high-throughput workflows, they are now more and more commonly used for middle-throughput workflows by tracking samples following the same precise and fixed protocols. They provide guidance through the different steps from sample registration to data capture and reporting. Indeed LIMS aim at eliminating duplication and deviation in records, standardising record-keeping and monitoring project status. They are therefore more than just elaborate ELNs. They are complementary as optimised to handle more structured data following strict workflows.

We distinguish various types of LIMS, which are more or less complex regarding hardware, software, organisation and functionalities with particular solutions for specific applications. However, the central component of all modern LIMS is always a database intended to organize and store large amounts of data. A user interface -generally web based- enables users to insert and retrieve information in and from the central database. Usually LIMS are also connected to multiple entry components like barcode printers, software, robots and laboratory instruments.

II] 1.2 History and evolution

The first LIMS appeared in the seventies and were mostly simple storage devices, with little or no real-time analytical capabilities. They were custom systems designed by independent companies to be run in specific laboratories. The first commercial LIMS were introduced in early 80s and were proprietary systems (PerkinElmer, Digital Equipment Corporation, Hewlett-Packard, IBM). Their

evolution paralleled evolution of computer systems on which they run ([90] - Gibbon, 1996).

At the end of the 80s they became more widely spread and practical to capture and process data from analytical instruments. They could manage the entire study cycle, from data generation through processing, to analysis and reporting. The first client/server and command-line based systems with flat-file architecture (data is stored in ordinary non-indexed flat file read and written in its entirety for data manipulation) arrived on the market enabling access from multiple computers.

From the mid-90s onwards appeared more intuitive and user-friendly graphical user interfaces (GUI) as well as the development of new capabilities with connections to other tools and software (3rd part software).

The major improvements of LIMS through the years are:

- the use of relational databases (information is stored in tables with rows and columns, tables being referred to as a "relation" in the sense that it is a collection of objects of the same types: rows that can be linked to other objects) with standardized SQL language. This increased dramatically the amount of data that could be stored and the speed of information retrieval with a resulting performance gain;

- the apparition of web based interfaces allowing the LIMS to be accessed from different sites through a decentralised architecture with networking.

These broad spectrum developments helped in the migration from proprietary commercial systems towards more open ones offering higher flexibility and functionalities. LIMS were then increasingly deployed in smaller laboratories serving research centres and industries following the new needs to cope with the increased data output from modern, multi-channel analytical instruments requiring implementation of new powerful management tools. Nevertheless LIMS require considerable customisation to meet specific and often heterogeneous needs of the particular laboratory that it was designed for.

II] 1.3 The new genomic laboratory needs with the rise of NGS

The automated Sanger sequencing method developed by Frederick Sanger in 1977 and based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication, has

dominated the industry of sequencing for almost two decades and led to a number of monumental accomplishments including the completion of the human genome sequence ([91] - International Human Genome Sequencing Consortium, 2004).

At start, the relatively low throughput of Sanger sequencing projects didn't require absolutely a LIMS to follow the experiments but GCS group started in 2004 to develop laboratory high-throughput procedures for a large-scale multi-centric case-control mutation screening study that used a combination of high-resolution melting curve analyses (HRM) and capillary Sanger sequencing. The objective of the study was to identify germline DNA sequence variants - base substitutions or small insertions and deletions - in various candidate genes that may contribute to breast cancer susceptibility. This required the development and implementation of a LIMS because it implied complicated workflows and we aimed to reach a relatively high throughput with the screening of the coding regions of more than 15 genes and almost 400 amplicons (specific pieces of DNA that were amplified for screening) in more than 2500 subjects.

GSP transitioned from Sanger sequencing to Next-Generation Sequencing (NGS) by acquiring a LifeTech SOLID 5500XL in July 2011 which was upgraded to Wildfire technology for higher throughput in July 2013. In parallel we also purchased an Ion Torrent Personal Genome Machine (PGM) and in January 2014, acquired an Ion Torrent Proton to eventually replace the SOLID instrument. The major advance offered by these NGS methods is the ability to deliver fast, quite cheap and accurate genome information through a broad range of applications from whole genome sequencing to exome sequencing and RNA sequencing. NGS thus enables the re-sequencing of a large number of genomes quickly to enhance understanding of how genetic differences affect health and particularly susceptibility to development of cancers ([92] - Metzker, 2010). However the complexity of the laboratory process and the volume of data generated make it complicated to manage without an appropriate IT tool.

The platform was enlarged with these instruments to run several large sequencing projects which successes are partly dependant on the ability to adequately automate them through a well-adapted LIMS and ensure central reliable storage, retrieval and analysis of whole associated laboratory data.

In this chapter, I will explain what the strategies of design, modelling and development are for the implementation of a new laboratory workflow in a LIMS. As for any scientific data management system, the best solution is an evolutionary one starting for a specific application and gaining experience to implement new applications.

Keeping in mind that the performance of a LIMS is largely dependent on the capability to adjust the specific needs of a laboratory, I will focus here the description on the evolution of LIMS within the GCS platform and on the LIMS modules that I developed to keep track of the latest promising technologies set up in GSP: exome and targeted sequencing using Ion Torrent PGM and Proton instruments. NGS laboratory workflows are indeed particularly complex and variable with many choices and optional steps, which make them even more essential to be tracked. The more steps a workflow is composed of, the more possibilities of problems and errors there are and the more important it is to be able to carefully follow each stage. NGS laboratory workflows are as well rapidly changing. The LIMS should be therefore easily configurable to accommodate the changes in experimental procedures and instruments.

II] 2. Development

II] 2.1. Choice of platform and implementation

GCS LIMS platforms

In order to keep track of the large number of samples we aimed at processing for our specific mutation screening project, we started investigating which LIMS solution could suit our needs. As for the ELN, the first question to consider was whether to build in an in-house LIMS or buy a commercial package, both having economic costs. The first option is appropriate for non-standard laboratory workflow with very specific needs not available in existing tools and therefore requiring large customizations meaning considerable human resources and programming skills.

The main advantage of the second option is that it offers very large and optimized functions with many complex features that require large teams of developers to implement. This suits classic laboratories working with standard procedures and having standard needs. The main issue is that these commercial packages target essentially industries and especially pharmaceutical companies which are the main customers. This results in quite complicated vocabulary and screen design that discourage its use by laboratory assistants.

In our case, we first developed in-house a LIMS based on a MySQL database and customized for our particular narrow set of research by mutation screening approach. It included the following experimental and analytical steps: multiplexing of amplicons for nested PCRs, HRM analysis, cherry picking of samples with specific melting profile that needed to be processed for Sanger sequencing, sequencing, purification of the sequencing products, storage of resulting chromatograms, reading with a specific connected software and recording of results (genotypes). The system developed for the follow-up of the mutation screening workflow is described in our publication in Bioinformatics: "A Laboratory Information Management System (LIMS) for high throughput genetic platform aimed at candidate gene mutation" ([93] - Voegele et al., 2007).

This first LIMS used in combination with automation of laboratory processes and thanks to its connections with the laboratory's robots, instruments and analysis software, improved efficiency and quality of work by enabling sample tracking activities such as cherry picking that are very difficult perhaps even impossible to perform without error by hand, reducing potential for human mistakes and accelerating the throughput of analysis.

As the laboratory transitioned to even higher through-put as well as new technologies and instruments, the system showed limitations in terms of database power. Extraction of data became excessively time-demanding. For example, generating a report with results of mutation screening of one gene in 2,500 samples could take several hours depending on the number of amplicons screened for the specified gene. We therefore continued our process of investigating existing solutions by evaluating:

- the overall technical capability of the biology-tailored LIMS that were available on the market

- the inclusion of a set of pre-programmed modules able to operate across different computer platforms and ideally web-based, as well as a Software Development Kit (SDK) to be able to modify ourselves the tool;

- the cost of the tool's licences and training.

We selected four companies (Amersham GE Healthcare, Sibio-Genomining-HP, Thermolab and Applied Biosystems,) which LIMS seemed to be able to fulfil those criteria and requested a presentation of their software as well as bids that should address 3 areas: (1) the base software and its database – (2) the set of pre-programmed modules for basic sequencing workflow and connection to instruments – (3) the SDK.

After study of each proposition, we "pointed out" that Amersham's Scierra LWS was more expensive, Sibio's SibioCLE had less functionality and had small development team and Thermolab's Nautilus database was closed for direct interaction. Applied BioSystem's SQL*LIMS ([94]) was the most appropriate and had the real advantage of giving us the option to buy the source code which would enable us to modify both the underlying database and the interface as well as to build our own screens. This guarantied some independence from the vendor for further developments, adaptations and customisations provided that we could understand the coding and modify it. In addition, this tool was already in place in other genomics

laboratories and appeared to be able to track samples and plates associated with variable types of data through different steps of PCR and sequencing. We therefore opted for this compromise between the commercial LIMS and the in-house developed LIMS.

Implementation

We installed the SQL LIMS on two servers: one for the Oracle Database running under Linux (RedHat ES4) and one for the Oracle application and the Apache web server running under Windows Server 2003. We installed two instances: one for the tests and one for the production. Each development is initially done on the test server before being replicated on the production one.

The computer clients are PCs running either Windows XP, 7 or 8, Linux or Macs OS X with the following compatible browsers: Internet Explorer 7,8; Firefox 3.0 et 3.6, and Safari 2.0.3, 3.0.

The tools used for the development and customisations are Oracle SQL Developer ([95]) regarding the database and Oracle Developer Suite: Forms for the web interface ([96]).

We started the implementation of workflows in the new system with the mutation screening by combination of HRM analysis and sequencing that was already managed in the previous LIMS using the same design type based on successive stages allowing some flexibility and optional steps in the workflow.

We then developed additional modules for management of Taqman and LightScanner genotyping, Illumina GoldenGate genotyping, Illumina Infinium genotyping, Whole genome expression profiling, SOLiD exome sequencing, PGM and Proton exome and targeted sequencing. Within this section, I will more precisely describe this most recent NGS workflow as a demonstration of the process by which the LIMS modules are developed.

II] 2.2 Scoping phase: study of the laboratory workflow

The first step before the modelling, configuration and development, of a new workflow is to study carefully each step of the laboratory procedures and estimate the variability within the different steps. This scoping phase is crucial to understand precisely what kind of data and information flow has to be tracked in order to design the required features. Close interactions with future users are therefore essential to define their expectations and translate their laboratory language into IT concepts as well as to understand the underlying molecular processes involved to have a view of the whole processes and thus appropriately design the LIMS tool.

The writing of specifications should include as much as possible likely future needs to facilitate addition of new functionalities and more customization. The resulting design should be highly flexible for both database scheme and the coding of the interface interacting with the database.

The sequencing process to be designed in the LIMS

The instance that I chose to demonstrate in this thesis is centred on the Ion Torrent sequencing. This technology like all NGS technologies rely on combination of template preparation (single DNA molecule or clonally amplified DNA), sequencing and imaging followed by data analyses.

The specific PGM ([9]) and Proton ([10]) instruments' chemistry is based on the detection of pH signals. It exploits the fact that addition of a specific dNTP to a DNA polymer releases one or several H+ ion. To summarize the principle, the input DNA is amplified; adaptors are added and each fragment is attached onto a bead. The molecules are clonally re-amplified using emulsion PCR and each bead is then placed into a single well of a slide containing thousands of wells. The instrument floods the slide with single species of dNTP along with buffers and polymerase, one NTP at a time. The pH is detected in each of the wells, as each H+ ion released will decrease the pH. The change in pH allows determining if that base was added to the sequence read (**Figure 18** and **Figure 19**). The process is repeated cycling through the different dNTP species which enable to get the whole sequence converted into a chromatogram.

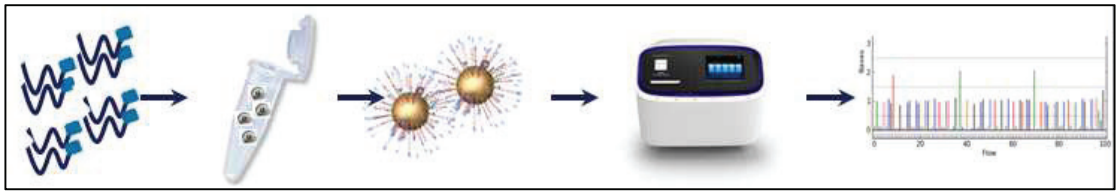


Figure 18: The "Ion" workflow with preparation of the library, clonal amplification, isolation of the spheres (beads), loading of the chip, sequencing and data analysis.

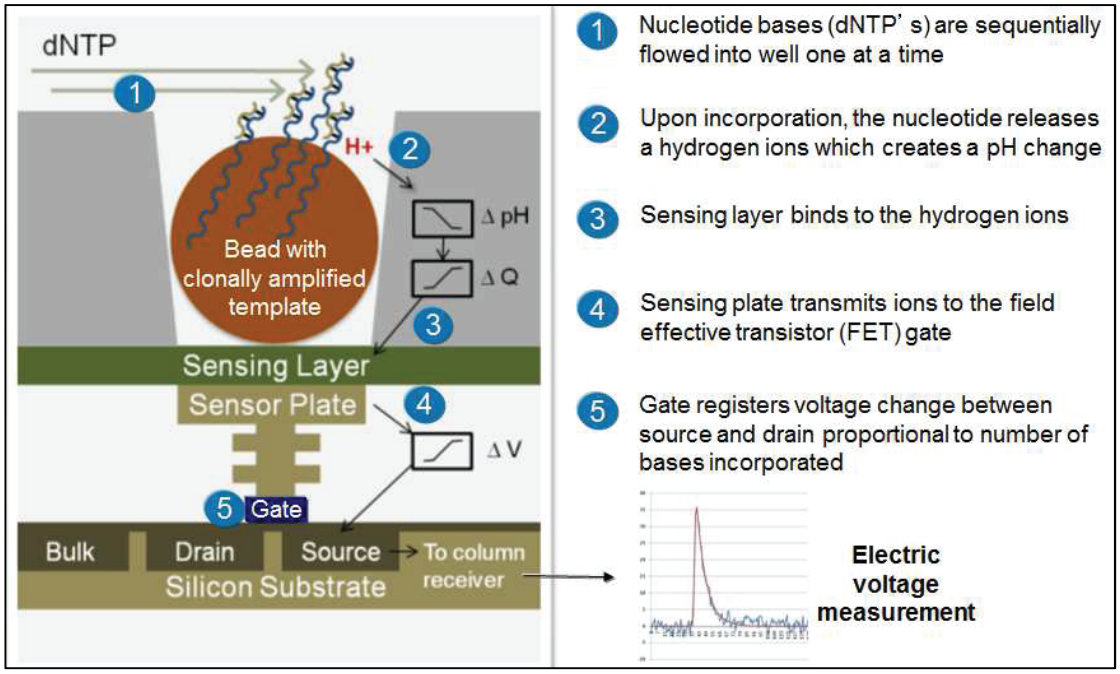


Figure 19: Principle of sequencing by PGM and Proton using pH signals.

The advantage of such method compared to the other platforms using fluorescence or chemiluminescence is the speed of the run (2 hours on PGM) and the possibility to push the throughput even more with new chips. The difference between PGM and Proton is the output capability: up to 1 Gigabases for the PGM (so more dedicated to small targets sequencing) up to 10 Gigabases for the Proton (designed for whole exome sequencing). The choice of the instrument and chips is indeed dependent on the experimental design that will take into account the expected coverage, the number of samples to be sequenced on the chip and the size of the sequencing target. It is thus easily possible to reach a throughput of 384 samples screened for a few genes in 3 days which would be complicated to manage without appropriate LIMS tool.

The laboratory workflow could therefore be divided into four main steps (**Figure 20**):

- 1. DNA sample amplification by PCR, pooling and purification of the PCR products.

- 2. Library preparation including end-repair, adaptor ligation and barcoding, purification, amplification and pooling of the amplified barcoded libraries into one tube.

- 3. Template preparation through emulsion amplification of pooled libraries with beads (One touch) and selection of spheres with libraries (Enriched Spheres).

- 4. Run on the instrument and primary analyses statistics

We distinguished two different applications: one for targeted sequencing that focus on specific regions of interest in the genome and which can be performed on either PGM or Proton depending on the scale of sequencing; and one for whole exome sequencing by Ion Ampliseq^{TM} - based on ultrahigh-multiplex PCR - which can be performed only on the Proton (**Figure 20**).

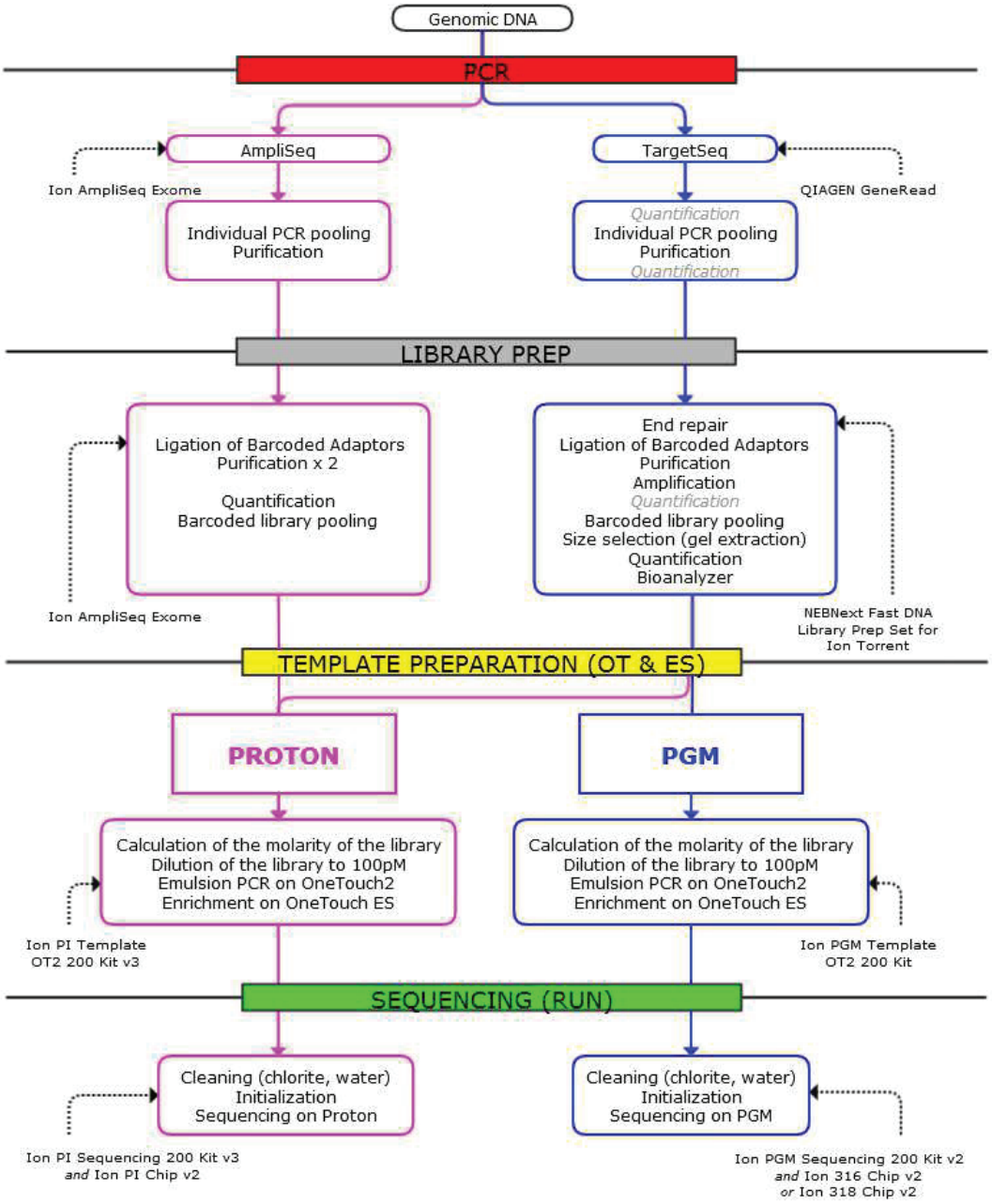


Figure 20: The laboratory workflow for whole exome sequencing (in pink) and targeted sequencing (in blue).

II] 2.3 Specifications

The global specifications for our LIMS were defined at the beginning of the setup of the genomic platform and then more dedicated features were defined as new applications were setup. Our LIMS should include the following standard and more advanced LIMS features:

- Sample login, identification and tracking through complex multi-steps laboratory workflow

- Barcoding of the processed samples and plates

- 96 and 384 well plate tracking and re-arraying

- Procedure and test assignment with storage of laboratory data at each step (method, reagents used, userstamp and timestamp)

- Quality control

- Reporting using different and complex queries criteria and generating different output formats (html, xls)

- Management of consumables stocks (plates, reagents quantities and volumes)

- Equipment configuration and maintenance tracking

- Interactions with robots and instruments to control and to collect data

LIMS should also allow flexibility for upgrade when technologies change or when deviations from initial workflows occur with possible addition or deletion of sub-steps. Adaptability of the LIMS over time is indeed a key requirement to cope with rapid evolution of laboratory workflows.

Particularly for our specific sequencing workflow, the design should take into account:

- Both Ion Torrent instruments (PGM and Proton)
- Both whole exome and targeted sequencing technologies
- Variable number of samples (from 1 to 384)
- Various types of pooling of samples and libraries at different steps
- The need for optional steps

II] 3. Results

II] 3.1 Strategy of modelling and specific developments

The most important impact of a properly designed LIMS is its ability to integrate many sub-processes. It should be flexible; especially the underlying database must provide an architecture flexible enough to capture the fine details and history of a multiplicity of protocols and procedures ([97] - Strass, 2008).

After installation of the tool, we rapidly realized how complex the standard interface was, using specific vocabulary for concepts and objects with which our users were not familiar at all. We spent some time evaluating how we could adapt the LIMS to our objectives and finally decided to develop an over-interface above the standard one to guide the users along the processes through new own simplified screens we developed based on the software forms templates for an easier integration within the system. We gave users access to only those screens following their vocabulary and their habits. This strategy enabled us also to constraint the users to perform precise steps in a specific order imposing strict adherence to the laboratory protocols, so reducing misuse and errors.

Within the LIMS the following concepts are used: a sample (instance level) is a type of material (template level), a submission (instance level) is a group of samples or a study (template level), and a plate is a container which is associated with a container template and container maps for re-arraying. One container template is created for each step of the workflow with attached information on the procedure followed at the specified step to store in the database. Doing so it is possible to add easily sub-steps by creating a new container template and modifying the container maps.

We had also to create many specific objects which we all prefixed by “IARC” to easily distinguish them from the standard LIMS objects. For the Ion exome sequencing workflow it includes four specific database tables, 3 database views among which the general view on all “ION” plates: IARCV_PLATES_ION containing the following plates attributes: container ID, parent container ID, template

name, container name, datagroup, description, class, status, project, origin, comments, userstamp and timestamp.

In total all workflows together we created 31 specific tables (for example for genes, SNPs, amplicons and their links, assays, ION reagents, NGS runs, NGS analyses), 54 views (result sets of stored queries on a mixture of standard and custom tables), 15 indexes (data structures that are copy of selected information to improve speed of data retrieval operations), 3 packages with around 35 procedures, 1 library and 55 forms corresponding to 55 different screens of the LIMS' interface.

II] 3.2 An intuitive Graphical User Interface (GUI)

This strategy adopted with the development of our own screens enabled to keep a quite intuitive interface similar to GCS previous LIMS with which the users were familiar. Also the native screens were too focused on industrial and more precisely pharmaceutical companies as do all LIMS providers using terms that are not part of the classic genomic laboratory vocabulary.

The access to the LIMS is permitted through a login and password window that lead the users to their dashboard limited to the 'MySQL LIMS' tab displaying the different workflows they are allowed to work with (**Figure 21**). The security is ensured by selection of specific user roles (system manager or user with restricted job type) and separation of data including sample information and workflows in datagroups.

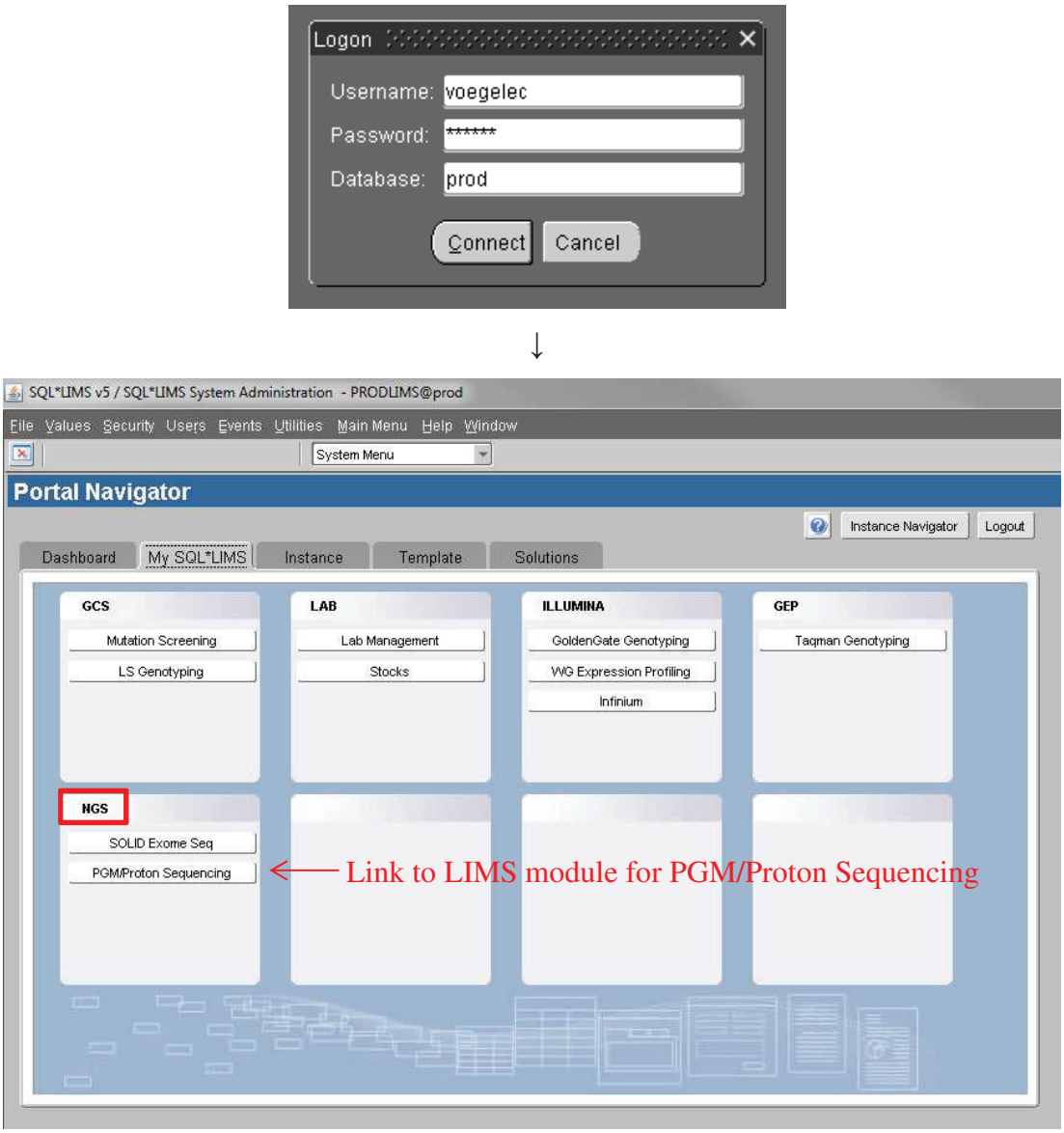


Figure 21: Logon window and My SQL*LIMS dashboard. *The LIMS dashboard presents the different workflows managed in the LIMS including NGS SOLiD and PGM/Proton sequencing accessible through the menu's buttons.*

Within each laboratory workflow a welcome page shows its purpose and a menu on the left which opens forms or screens specific to one object (projects, reagents and samples) or to one main step of the workflow (PCR, library preparation, "One Touch and ES" and Sequencing run) – (**Figure 22**).

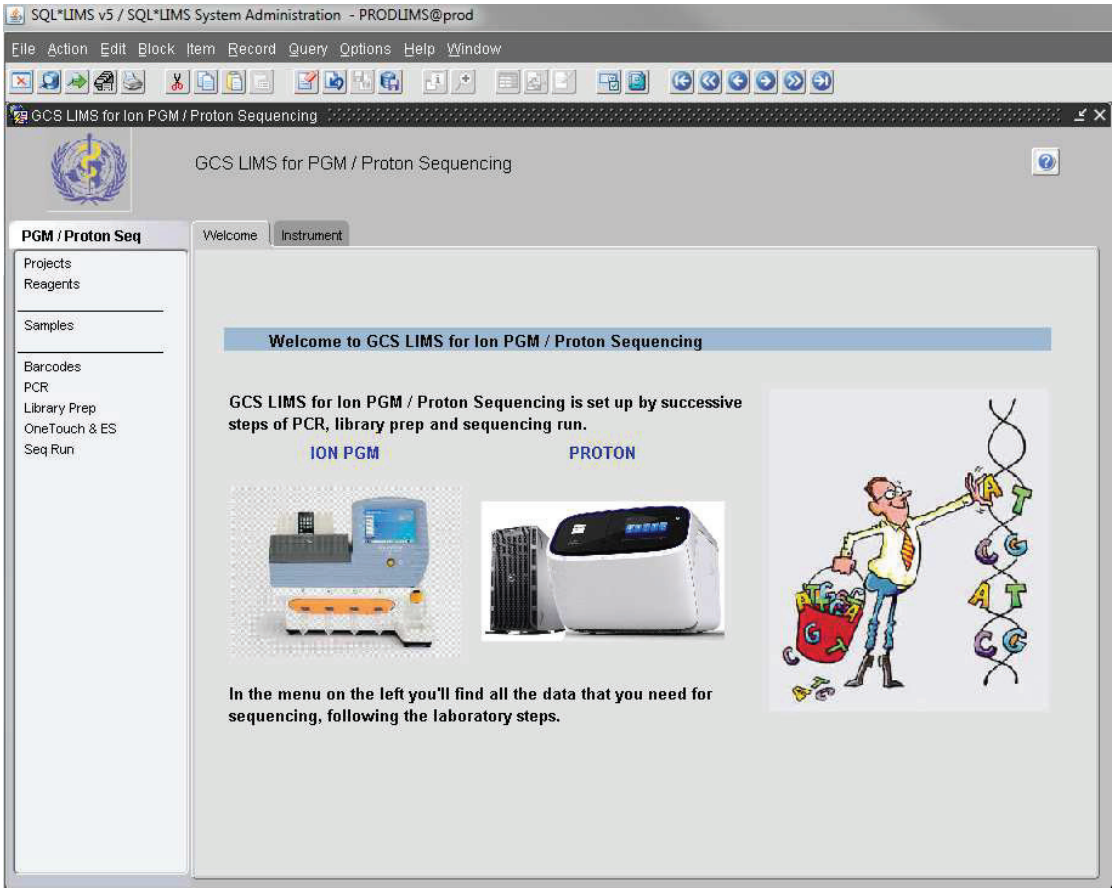


Figure 22: LIMS main page for Ion sequencing. *The menu on the left shows the different objects and steps managed for the PGM/Proton sequencing workflow.*

Within each form, tabs are specific to one action (adding data, updating data, search for data and listing data) or to one sub-step of the workflow (for example PCR or pooling). Users simply fill the forms with the information of their work and then submit it by clicking on the save button  (**Figure 23**). This launches a validation script specific to each object or each step that will check both if users filled all the required fields and the data entered: its type and integrity with respect to what is expected and to the database tables formats. If data is correct and sit the checks, data transaction is performed. Barcode scanning fields and pull-down list of values guide as well the users. Some triggers (procedural codes automatically executed in response to certain events on a particular database table like insertions or deletions) developed within the forms additionally enable to populate some fields depending on what the user enters in some other fields. Finally a specific colour code in the headers of the forms enables users to distinguish rapidly the steps dedicated to whole exome

sequencing (red), the steps dedicated to targeted sequencing (blue) and the steps required for both workflows (purple).

All this together prevents users to skip any step and minimizes the errors. Users are never allowed to delete any data. In case they make a mistake while entering data despite all the checks, they have to ask the LIMS administrator for deletion. This is to prevent unwanted loss of data.

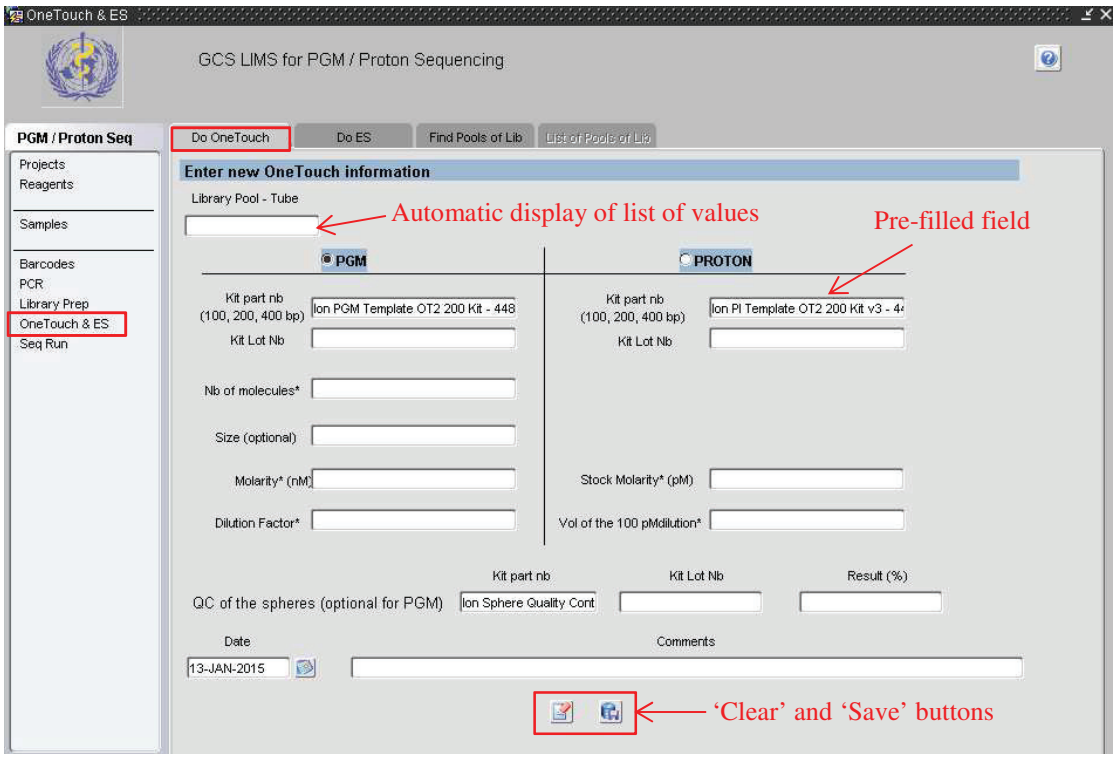


Figure 23: Example of a LIMS form for a specific step of the PGM/Proton sequencing workflow (OneTouch and ES).

Data retrieval is achievable from any form so for all objects or steps of the workflow through the 'Find' and 'List tabs. Different research criteria were defined enabling specific querying. Once the find form is filled, users click on the arrow button to get the results visible in the "List" tab. In some forms for which it was pertinent we additionally provided the option to export these results into Excel files.

II] 3.3 The ION exome sequencing workflow in the LIMS

The ION exome sequencing workflow is composed of seven specific forms following the experiments flow.

II] 3.3.1 Projects

Users start by entering information on their sequencing project: ID, name, description, cancer(s), gene(s), manager(s) and collaborator (s). The LIMS automatically stores as well the project's datagroup, the userstamp and timestamp. By default within the PGM/Proton sequencing workflow, the project name starts with 'NGS_ION' for easy association with the type of experiments.

II] 3.3.2 Reagents and management of stocks

Three main categories of reagents require careful tracking carefully during PGM/Proton sequencing experiments: Ion Serapure Beads, IonXpress Oligonucleotides and Ion Barcodes (for sample's barcoding when pooling). For this, two specific tables were developed storing information on the reagents' name, quantity, date of update and movements (in and out with the date to be able to retrieve the lots) – (**Appendix 2**).

A specific screen enables the users to enter the reagents data both by filling the form or using an Excel file more adapted when having a different quantity used for each of the 384 oligonucleotides (**Figure 24**).

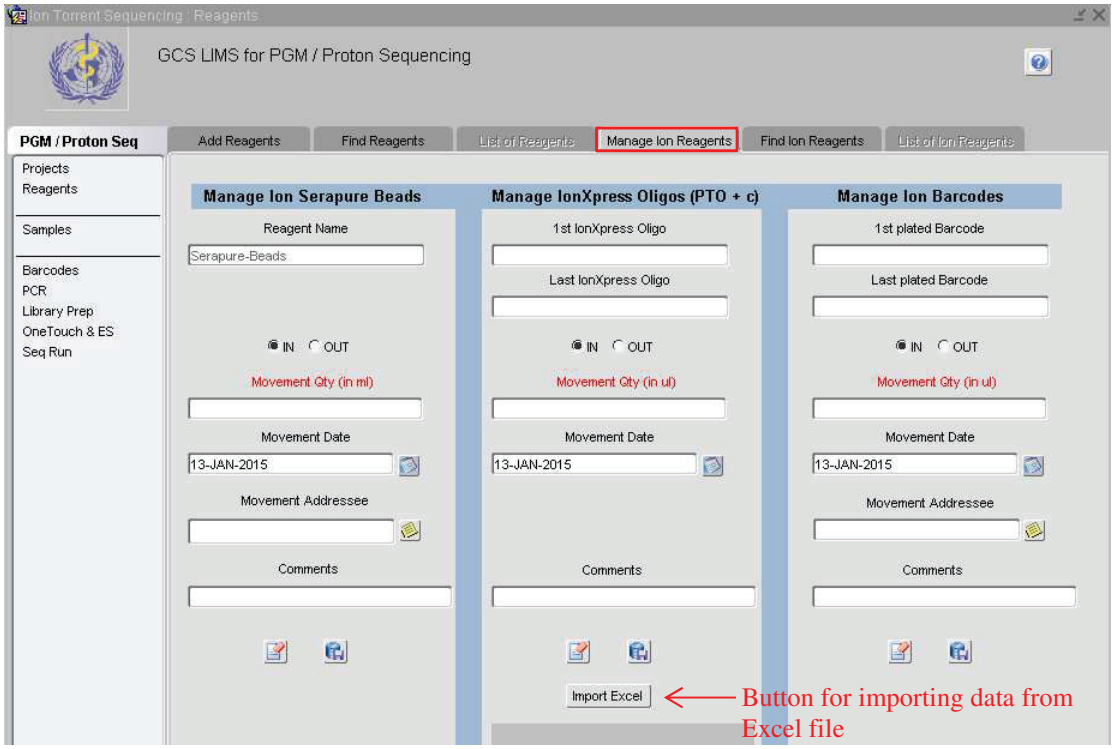


Figure 24: Form for management of reagents stocks. *The form enables to store for each of the three types of reagents the "in" and "out" movements with the corresponding date of re-supplying (in) or use (out).*

Within the 'Find' table, it is possible to launch specific excel reports showing the list of IonXpress oligonucleotides with low remaining quantities (below 80µl) and of Barcodes with very low quantities (below 50µl) for a proper management of the stocks and re-orderings.

The tab listing the reagents displays their name, current quantity, comment last update timestamp and last update userstamp as well as the detail of their movement (type 'IN' or 'OUT', date, quantity used, addressee, userstamp and timestamp).

Tracking reagents batch numbers can be really helpful to identify possible failure in experiments.

II] 3.3.3 Samples

The users enter the samples associated to their project and which they would like to sequence. The following samples' information is stored in the LIMS database: ID, name, barcode, project, type, status, concentration, origin, delivery date, IARC

database reference and ID, external name 1, external name 2, comments, datagroup, userstamp and timestamp. They have the possibility to enter the samples' information one by one within the form or to upload an excel file with specific format in which case a specific PL/SQL procedure parses the excel file data to fill in the database.

In the "find" tab, the users can search for samples using various criteria: project, sample ID, sample name, barcode, IARC ref, IARC ID or origin (all fields having pre-filled lists of values (LOV)).

The "list" tab displays basic information about the sample like its name, concentration, type or origin with a counter for the number of samples matching the selected criteria but also more detailed information with userstamp and timestamp as well as the list of DNA plates in which we can find the selected sample together with its position. A function enables to sort the list of samples by name, type, sample ID or origin. Finally, a magnifying glass button opens a screen with even more information on the samples like the PCR plates containing these samples and some results from different analyses performed on these samples (**Figure 25**)

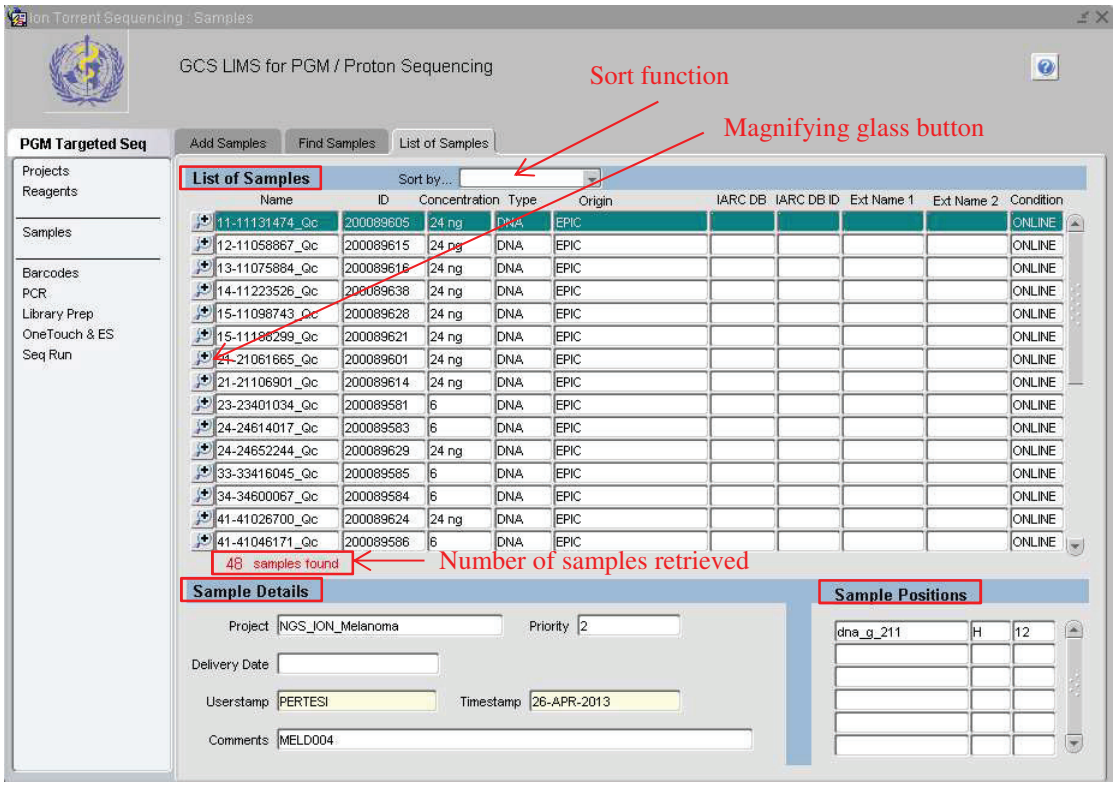


Figure 25: LIMS' screen listing samples. *The screen displays information on samples matching the query made to the database: basic information on the top and more detailed information on the bottom.*

IIJ 3.3.4 PCR

1. Users first enter their DNA plate using the 'Do DNA' tab for uploading an Excel file containing from 1 to 384 samples with their position (**Table 4**).
2. The second sub-step is for doing the PCRs which number can vary from from 1 up to 12.
3. The third step is the pooling of PCRs. This step can generate from one (pooling of the 384 by quadrant) up to four 96 well plates.

   Two different container maps enable the re-arraying of the PCR products depending on the number of samples:

   - "iarc_ion_pool_384" transfers the 384 samples from "ion_pcr" plate to four different "ion_pool_pcr" plates by quadrants (**Figure 26**).

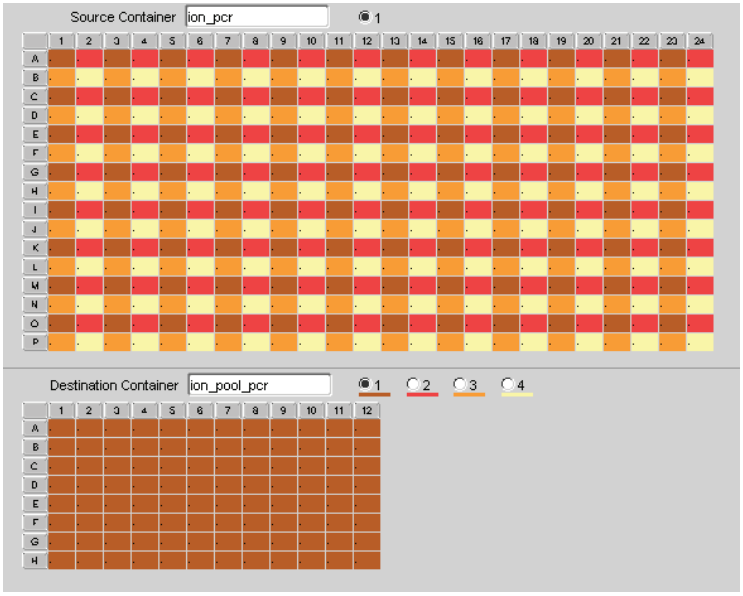


Figure 26: "iarc_ion_pool_384" container map for re-array.

- "iarc_ion_pool_96" transfers the 1st quarter of samples from one "ion_pcr" plate to one "ion_pool_pcr" plate (**Figure 27**). This is of course only possible when the number of samples is below or equal to 96.

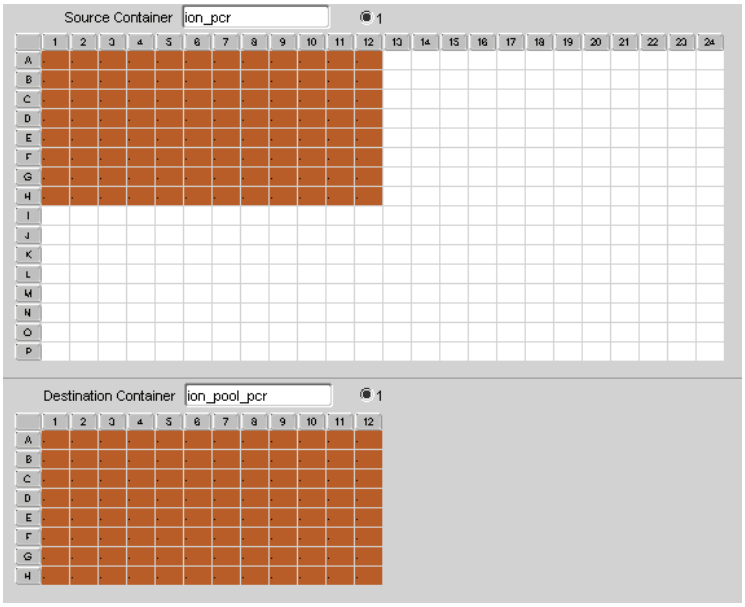


Figure 27: "iarc_ion_pool_96" container map for re-array.

4. The last step is optional and is dedicated to purification of the pools of PCRs.

All specific objects created for each step of the workflow are detailed in **Table 4**. All plates have the classic following attributes: container ID, barcode, status, userstamp and timestamp plus additional specific attributes (**Table 4**). All along the workflow the derivate plates' barcodes are automatically defined with the same identifier than the original DNA plate to enable easy following of the steps.

Step (Form)	Sub-step (Tab)	Sub-sub-step	Status	Container template	Container map	Specific plates attributes
PCR	1. Do DNA		Required	Ion_dna (384-well format)	None	Project, date of creation, comments
PCR	2. Do PCR		Required	Ion_pcr (384-well format)	Ion_dna_pcr transfers samples from ion_dna to ion_pcr at same well position	Origin plate (DNA plate), date of creation, type (GeneRead, AmpliSeq or Custom), number of PCRs, kit name, kit lot, comments, PCR1 reaction vol up to PCR12 reaction vol, PCR1 program up to PCR12 program, PCR1 comments up to PCR12 comments
PCR	3. Pool PCRs		Required 2 options	Ion_pool_pcr (96-well format)	a) Ion_pool_384 b) Ion_pool_96 (cf. Figures 22 and 23)	Origin plate (ion_PCR plate), type of pooling (equimolar, equivolume or custom), pooling quantity, pooling volume, pooling technique (manual or robot), robot program, PCR1 volume up to PCR12 volume, date of creation, comments
PCR	4. Purification		Optional	Ion_pool_pcr_pur (96-well format)	Ion_pcr_pool_purif transfers samples from ion_pool_pcr to ion_pool_pcr_pur at same well position	Origin plate (ion_pool_pcr), type of purification (manual or robot), robot program, beads ratio, final vol of elution, date of creation, date of serapure lot, Fupa part, Fupa lot, Fupa PCR_program, AmPure part, AmPure lot, EXOSAP, comments

Table 4: Specific LIMS objects created for "PCR" step.

II] 3.3.5 Library Preparation

1. Once the PCRs are done, the preparation of the libraries starts with an end-repair step but only for targeted sequencing (GeneRead). It is followed by an optional purification (**Table 5**).
2. Common to both PGM and Proton sequencing workflows are then the adaptor ligation and barcoding steps followed by purification.
3. After ligation users perform amplification but only for the targeted sequencing workflow. It is followed by an optional purification step.
4. Instead of the amplification, there is a step of quantification by qPCR for exome sequencing by AmpliSeq.
5. The next step is the pooling of the libraries followed by an optional size selection on gel.
6. Finally users perform quantification for the targeted sequencing workflow.

II] 3.3.6 OneTouch (OT) and Enrichment of Spheres (ES)

Once the libraries are prepared they go through two last steps before the run: the "One Touch" step (template preparation through emulsion amplification of pooled libraries with beads) and the "ES" step that corresponds to the selection of spheres with libraries (Enriched Spheres) (**Table 6**).

Step (Form)	Sub-step (Tab)	Sub-sub-step	Status	Container template	Container map	Specific plates attributes
Library Prep	1. End-repair	a) End-repair	For Targeted Sequencing	Ion_er (96-well format)	Ion_pcr_pool_pur_er transfers samples from ion_pool_pcr_pur to ion_er at same well position	Origin plate (plate_pcr_pool_pur), DNA quantity, volume of samples, ER kit type (predefined options), ER kit part, ER kit lot, reaction volume (1, ½, ¼), date of creation, comments
Library Prep	1. End-repair	b) Purification	For Targeted Sequencing but optional	Ion_er_pur (96-well format)	Ion_er_purif transfers samples from ion_er to ion_er_pur at same well position	Origin plate (er_plate), AMPure part, AMPure lot, date of creation, comments
Library Prep	2. Ligation and barcoding	a) Ligation and barcoding	Required	Ion_lig (96-well format)	Ion_pcr_pool_lig, ion_pcr_pool_pur_lig, ion_er_lig, ion_er_pur_lig transfers respectively samples from ion_pool_pcr, ion_pool_pcr_pur, ion_er, ion_er_pur to ion_lig at same well position	Origin plate (plate_pcr_pool, plate_pcr_pool_purif, plate_er or plate_er_pur), DNA quantity, volume (1, ½ or¼), type of manip (manual or robotic), robot program, barcodes type (IonXpress or custom), barcodes range (1-96, 97-192, 193-288, 289-384, successive BC in vertical or creative number of BC), barcodes kit part, barcodes kit lot, first barcode, last barcode, date of creation, comments Sample's barcode (=plates' well attribute)
Library Prep	2. Ligation and barcoding	b) Purification	Required	Ion_lig_pur (96-well format)	Ion_lig_purif transfers samples from ion_lig to ion_lig_pur at same well position	Origin_plate (ion_lig), type of manip (manual or robotic), robot program, date of serapure lot, AMPure part, AMPure lot, number of purif, beads ratio, final vol of elution, date of creation, comments
Library Prep	3. Amplification	a) Amplification	For Targeted Sequencing	Ion_amp (96-well format)	Ion_lig_pur_amp transfers samples from ion_lig_pur to ion_amp at same well position	Origin plate (lig_pur), number of cycles, PCR program, volume of samples, date of creation, volume (1, ½, ¼), comments
Library Prep	3. Amplification	a) Amplification	For Targeted Sequencing	Ion_amp (96-well format)	Ion_lig_pur_amp transfers samples from ion_lig_pur to ion_amp at same well position	Origin plate (lig_pur), number of cycles, PCR program, volume of samples, date of creation, volume (1, ½, ¼), comments
Library Prep	3. Amplification	b) Purification	For Targeted Sequencing but optional	Ion_amp_pur (96-well format)	Ion_amp_purif transfers samples from ion_amp to ion_amp_pur at same well position	Origin plate (amp plate), type of manip (manual or robotic), robot program, AMPure part, AMPure lot, final vol of elution, date of creation, comments
Library Prep	4. Quantif - qPCR		For Exome Seq by AmpliSeq	Ion_lig_pur (96-well format)	Ion_lig_purif transfers samples from ion_lig to ion_lig_pur at same well position	Quantif type (qPCR), quantif kit part, quantif kit lot, date of quantif, ELN qPCR results, quantif comments
Library Prep	5. Pooling in tube	a) Pooling	Required	Ion_lib_pool_tube (tube)	None	Origin plate 1 to origin plate 4 (lig_pur plate, amp plate or pur_amp plate), type of pooling (equimolar or equivolume), pooling quantity, pooling volume, nb of pools per plate, nb of samples per pool, pooling set, date of pooling, pooling comments
Library Prep	5. Pooling in tube	b) Size selection on gel	Optional	Ion_lib_pool_tube (tube)	None	Extraction kit, date of size selection, size selection comments
Library Prep	6. Quantification		For Targeted Sequencing	Ion_lib_pool_tube (tube)	None	Quantif type (bioanalyzer or qPCR), quantif kit part, quantif kit lot, average molarity, date of quantif, ELN qPCR results, quantif comments

Table 5: Specific LIMS objects created for "Library Preparation" step.

Step (Form)	Sub-step (Tab)	Sub-sub-step	Status	Container template	Container map	Specific plates attributes
OneTouch & ES	1. Do One Touch		Required	Ion_lib_pool_tube (tube)	None	Ion instrument (PGM or Proton), OT kit part, OT kit lot, OT pgm nb molecules, OT pgm size, OT pgm molarity, OT proton stock molarity, OT pgm dilution factor, OT proton vol dilution, OT Spheres QC kit part, OT spheres QC kit lot, OT spheres QC result, date of OT, OT comments
OneTouch & ES	2. Do ES		Required	Ion_lib_pool_tube (tube)	None	ES streptavidine beads part, ES streptavidine beads lot, ES vol beads (13 ul for PGM or 100 ul for Proton), date of ES, ES_comments

Table 6: Specific LIMS objects created for "OneTouch and ES" step.

II] 3.3.7 Run

The last step is the run on the instrument. For this, a specific table was created including information on the run itself with the library, the chip and reagents used but also information on the program launched for the run and the first statistics of the analyses performed on the machine which gives quality feedback (for example total number of reads and mean depth) – (**Appendix 3**).

II] 3.4 Interactions with printers, robots and instruments

Instrument interfacing is a chief benefit of implementing a LIMS but most of the time the software that pilot these instruments are proprietary and therefore neither use or modification of the code neither direct interaction are possible. Companies proposed Application Programming Interfaces (APIs) but the cost is high and source codes are not available leaving customers dependent on companies for any changes requests. So a workaround needed to be found using files that could be uploaded within the software. LIMS vendors provide this kind of option through plugins or drivers but at very expensive price. That is why we looked for an in-house alternative.

Bi-directional communications with printers, instruments and robots have been set up and mediated by sending and reception of different types of files through the network.

First, the LIMS has been connected to our Zebra barcode printer. For this a PL/SQL (procedural language extension for SQL and Oracle databases) specific package IARCP_BARCODES was created with the following 5 procedures:

- create_plate_barcode that is specific to plate and uses the plate name to build a barcode file

- create_virtual_label that is specific to ephemera plates used between 2 steps of the workflow

- create_virtual_barcode that is to create any type of barcode specified in the corresponding field of the barcode form

- print_barcode that sends the barcode file to the printer

- reprint_barcode that re-sends an existing barcode file to the printer

So at each creation of plate in any workflow in the LIMS, a file containing some codes in a proprietary language and specifying the barcode label, size and type is built and sent to the printer through the network. It is thus instantaneously printed. A specific form was also developed for (re-)printing any barcode (**Figure 28**).

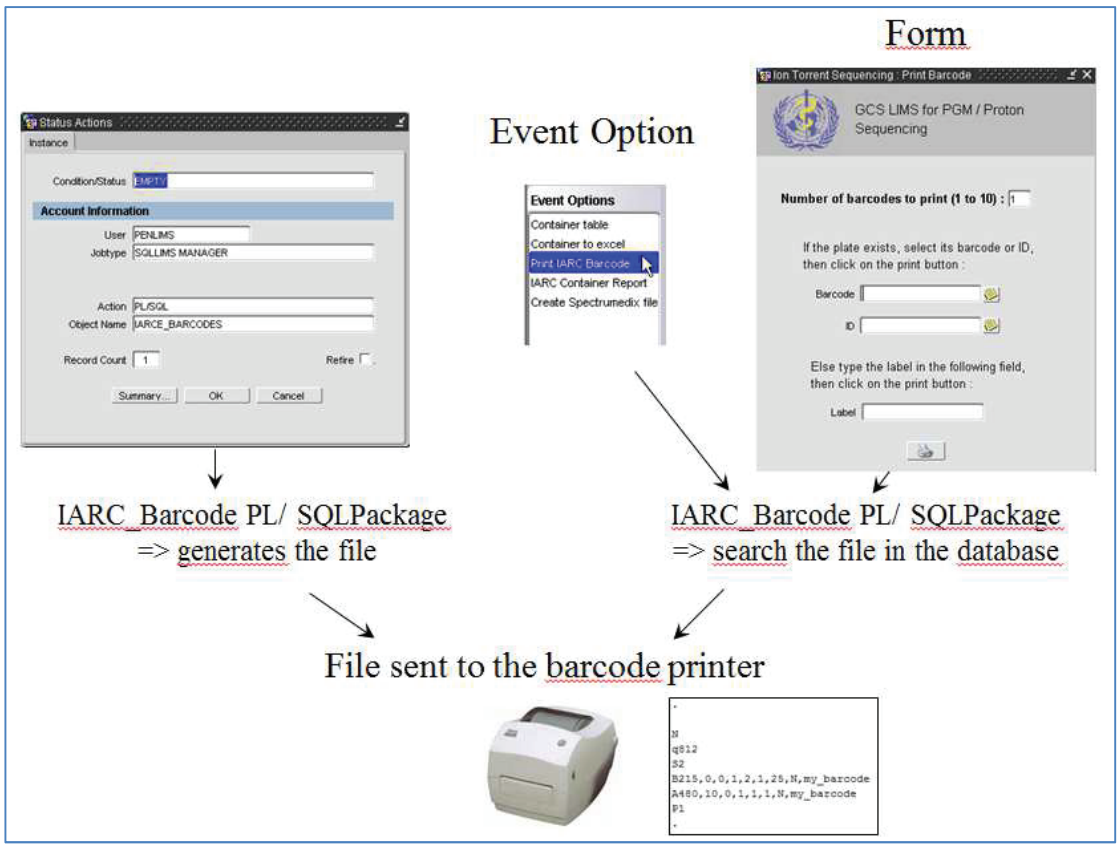


Figure 28: How LIMS can launch barcodes printing. *The forms launch generation or retrieval of a file containing instructions for the barcode printer, file which is sent through the network to the printer.*

For our Beckman and Tecan pipetting robots piloted by completely different software, communication is mediated by worksheets or worklists containing plates' barcodes and positions on the robots desks. For the other instruments sample sheets containing the sample ids or names and their positions on the plates are used (**Figure 29**).

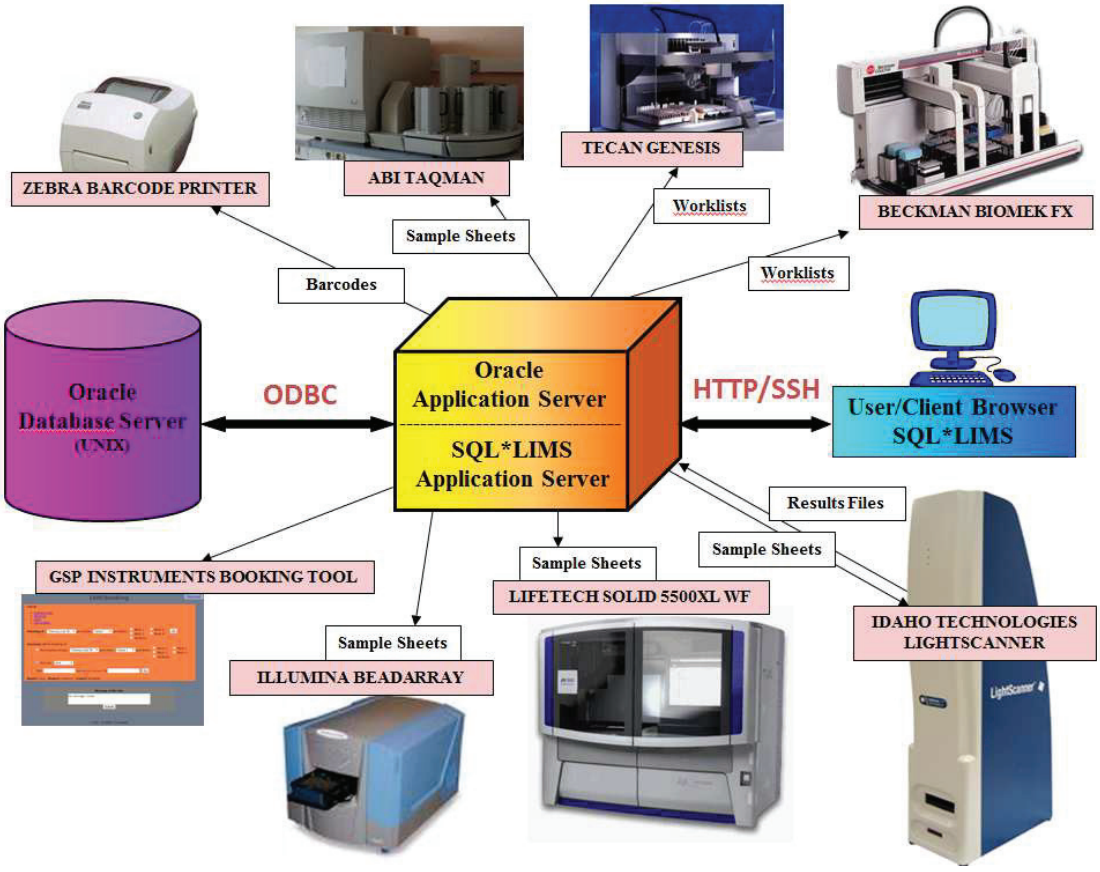


Figure 29: The LIMS architecture and connections. *The LIMS is currently connected to two pipeting robots: a Tecan Genesis and a Beckman Biomek FX, to a Zebra barcode printer, a Taqman, an Idaho LightScanner and an Illumina Beadarray as well as the SOLID sequencer and a tool for instruments booking.*

Another specific package IARCP_INSTRUMENTS was developed in that purpose with procedures to generate these different files. The method used is identical in all cases: the files are sent via samba to the specified IP address of the robots or instruments (**Figure 30**) in a specific directory where the piloting software can upload them before run.

Hand-coding is needed for each added instrument since they have all specific file formats.

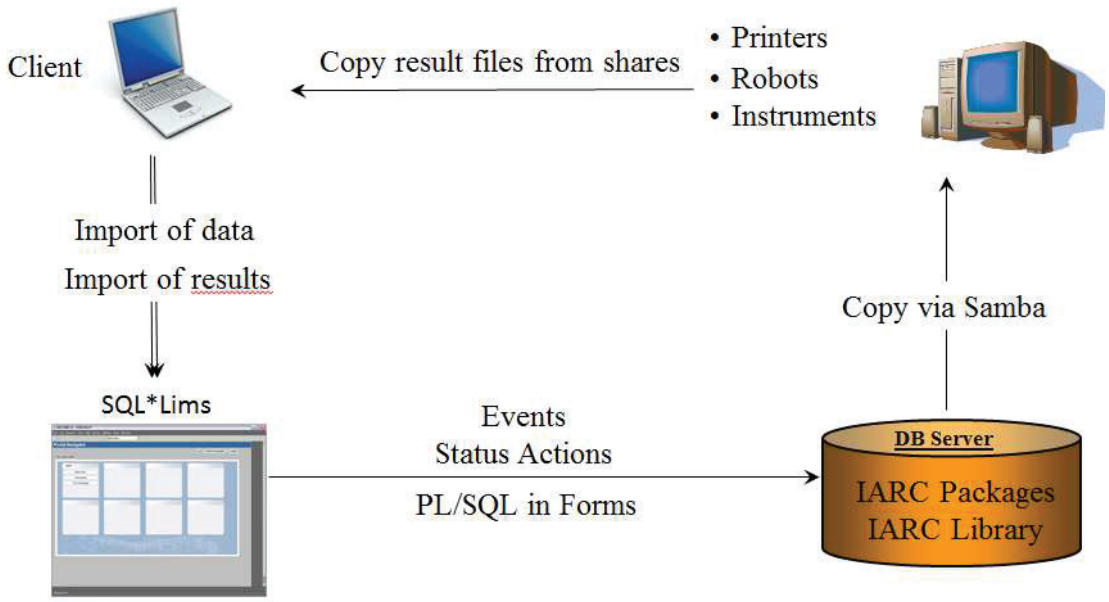


Figure 30: Communication between the LIMS and the different robots and instruments. *The communication is mediated by sending and receiving text files which are managed by PL/SQL procedures.*

In the same way, users could use robots or instruments output files to extract data, parse and store them in a database for further easy search. Data is imported from the computers via shares. The lines of text, csv or excel files are parsed based on specific formats and the data is directly inserted into the database.

Since the platform is used by several people within GCS group but also from other IARC groups, we additionally included a Java tool developed by Arnaud Dumont for booking of PCR machines, robots and the instruments of the platform. This tool, accessible from the LIMS and installed on a Linux server includes a small MySQL database and a web interface that presents a calendar for each of our PCR machine and instrument. Laboratory research assistants can subscribe for a specific time slot during which they are sure to be able to use the instrument they needed.

II] 3.5 Summary and evaluation

Summary

The LIMS has become essential in the laboratory for an efficient daily management and quality control of the genomic platform. Every single laboratory action is now tracked from sample log-in to clear reports of experimental results. It improved work quality and productivity by considerable reduction of human potential errors by guiding the users through the NGS workflows: PGM and Proton sequencing but also Taqman, LightScanner (HRM) or Illumina genotyping and expression profiling. The step by step design and the possibility of in-house developments with appropriate database capacity and customisations make it adaptable to any kind of biological workflow and various applications.

Well-designed LIMS adapted to the specific needs of laboratories are the key component for an optimal laboratory management especially when it concerns complex and high throughput workflows that need strict tracking of experiments, as it is the case for NGS. As described in this chapter, high performing systems to be used at their maximum capacity should fulfil two important conditions:

(1) be user-friendly and deliver strict guidelines for a proper and harmonized recording of laboratory data

(2) be designed to allow flexibility for future evolution.

LIMS use evaluation

The LIMS is used in GCS by both the laboratory assistants and the managers. Some of them need the LIMS every day, others once a week depending on the frequency of the laboratory experiments. The laboratory assistants are using it for recording the samples that undergo specific laboratory analyses following each step of the pre-defined workflows. They additionally use it for managing the laboratory stocks storing every entry and exit of consumables (plastics, tips, films, various reagents and kits). They receive thus automatic e-mails alerts when one of the consumables reaches the re-order point limit. They also use the LIMS when looking for some information on a previous experiment or to know the remaining quantity of samples, assays or plates. As for the managers they use the LIMS to query for reports on the experiments performed or the results of the experiments or analyses. The main advantage seen by the users is the fact that they are guided through the workflows

preventing entry errors, step skipping and performing automatic quantity calculations. The fact that all the information about specific types of experiment is stored in one and the same place is crucial to have a global view on studies and the associated work achieved.

As a result, we have now stored in our LIMS ~100,000 samples that underwent various sequencing and genotyping analyses tracked on ~20,000 plates resulting in almost ~650,000 individual genotypes (A, C, G, T) at specific positions on the human genome (**Table 7**).

Since the development of the LIMS for NGS, we have been running several projects using both the SOLID and the PGM and Proton instruments. 815 samples have already been sequenced with NGS technologies. All the library preparations and runs have been recorded in the LIMS and I additionally recorded basic information on the analyses that I performed on data generated for more than one hundred and fifty exomes and RNA seq experiments.

Some statistics of the global and workflow-specific data stored in the LIMS are shown in **Table 7**.

LIMS performance evaluation

Each form and each type of transaction (insertions or updates) to the databases through the web interface is tested manually and carefully to ensure all the information is correctly stored in the database.

Despite the large amount of data in the LIMS database (6.8G in total in July 2015), the query time for reporting is still relatively rapid: 50 seconds for retrieving 500 samples with the associated location and results; 10 seconds for retrieving a 384-well plate schema. For the other objects the retrieval is almost instantaneous. Query time is a good performance indicator for databases as it can be the bottleneck when the design has not been appropriately optimized to handle the volume of data aimed at being stored.

As for the ELN, restore of backups have also been validated to ensure data stored in the LIMS can be recovered in case of server failure.

Workflow	Object	Number stored
All	Projects	88
All	Samples	86,502
All	Genes	575
All	Consummables references (Stocks management)	142
Mutation screening	Samples	17,447
Mutation screening	Primers	1,515
Mutation screening	Plates	8,225
Mutation screening	Genotypes	627,114
Mutation screening	Chromatograms (Paths)	45,311
Samples quantities management	Samples	46,400
Samples quantities management	Samples movements	40,216
Taqman genotyping	Samples	69,055
Taqman genotyping	SNPs	1,545
Taqman genotyping	Assays	1,387
Taqman genotyping	DNA Plates	8,259
Taqman genotyping	PCR Plates	2,152
NGS	Samples	815
NGS	Reagents	1,156

Table 7: LIMS database content statistics

II] 4. Discussion

II] 4.1. The choice of a LIMS

Similarly to ELNs, a large number of LIMS is nowadays available on the market. It is thus complicated and time-consuming to select the right one for his laboratory to such a point that companies offer to evaluate and choose LIMS for customers. The global market is expected to reach $1,323.6 million by 2019 from $848.5 million in 2014, growing at an annual rate of 9.3% from 2014 to 2019 ([98] - marketsandmarkets.com, 2014). The market is segmented on the basis of product (commercial off-the-shelf or legacy LIMS), component (software or services), delivery mode (on-premise LIMS, cloud-based LIMS, and remotely hosted LIMS), end user (healthcare industries, petrochemical refineries and oil and gas industries, chemical industries, food and beverage and agricultural industries, environmental testing laboratories, metals and mining industries, and others), and region (America, Europe, Asia). Healthcare industries are further segmented into pharmaceutical and biotechnology industries, biobanks/biorepositories, contract services organizations (CROs/CMOs), and academic research institutes. This segment is expected to grow at the highest rate from 2014 to 2019. The increasing number of clinical trials, the increase in globalization and outsourcing activities as well as the increasing burden to comply with regulatory guidelines, are major factors that will drive the growth of this segment in the next five years. Key players in the market include LabWare, Inc. (U.S.), LabVantage Solutions, Inc. (U.S.), STARLIMS Corporation (U.S.), Thermo Fisher Scientific, Inc. (U.S.), Core Informatics (U.S.), Autoscribe Informatics (U.S.), PerkinElmer, Inc. (U.S.), LabLynx, Inc. (U.S.), Computing Solutions, Inc. (U.S.) and GenoLogics (Canada).

They are a few points to have in mind when it comes to choose a LIMS:

- narrow your research to your segment of interest based on the latter categories (basis of product, component, delivery mode, end user and region)
- perform a detailed study of the needs engaging the right people early in the process of definition of the requirements and thinking both short and long term.

- know if the LIMS should handle only one specific application such as “Proton” sequencing (there may be some commercial ones tailored specifically to your exact application requiring no costly customization) or various workflows (in this case you would need a tool that enables you to design your own workflows)
- know if the LIMS should manage only the laboratory experiments or the whole life of the laboratory including for example purchase and tracking of consumables or interfacing with instruments.
- be aware that small companies are less likely to survive on the long-term and that requests for modifications may take longer especially if they appear not to be reusable by other customers.
- know your budget as price can be very variable

Finally, as we have seen for the ELN, there are also many advantages in developing internally his own tool if the human resources can be allocated to this: an easier adaptation to the needs, no yearly licence cost and a longer support after over the years. This should be balanced with the main drawback: large human resources are needed to develop highly sophisticated and powerful tools and there is always a risk that these human resources leave taking with them the knowledge of the systems developed.

II] 4.2 Challenges

As all IT tools, LIMS are continuously evolving and improving which leads to the availability of major upgrades. This is the case with our LIMS. Although we are completely satisfied with the actual tool, the LabVantage company which in the meantime has bought Applied Biosystem LIMS sector, proposed that we migrate our LIMS to a new web technology providing all standard screens and functionalities in the new version. The constraint is that all our 60 custom forms would have to be redesigned, which would represent a considerable amount of work. Indeed, the original LIMS has been highly modified and not only customized which enabled us to meet our very precise laboratory needs but required a significant investment of human resources, time and therefore implied an additional cost. We may need to consider the

option of rewriting the interface in the coming years since the company will not keep on supporting our version for very long.

LIMS require internal human resources with at least one administrator to check the servers, database, data integrity, security and take care of backups and archiving.

II] 4.3 Perspectives

The development of a LIMS is never "finished". LIMS are evolving concepts and living systems designed for and in collaboration with the users. There are always changes in the laboratory workflows that need to be echoed or improvements suggested by the users that need to be made. From GCS work short-term perspectives we would need to additionally develop a RNA sequencing module for the Proton, an application which we are currently testing in the laboratory.

Also with previous generation of sequencers, one single person could lead his/her own project, from wet-lab to computer analysis and data exploitation. This new generation of massively parallel sequencers produces an enormous amount of raw data that is impossible to mine without a dedicated set of tools going through complex analyses pipelines. I therefore envisage developing additional LIMS modules that would manage both:

- an automatic import of the run statistics through connections with the PGM and Proton server by parsing their outputs.

- our in-house developed NGS analyses pipelines with appropriate storage for all intermediary and final output files.

Similarly to the importance and advantages of keeping track of each single step of laboratory workflows for quality control, it is as important to keep track of each single step of bioinformatics pipelines, the scripts and their versions to be able to ensure the same quality control and investigate possible errors or weaknesses. We could therefore plan to integrate details on bioinformatics analyses' pipelines and results coming from the different tools we use either on our HPC or on other personal workstations. This could include the parameters used for the successive steps of the pipeline i.e. mapping of the reads to the human genome, variants calling, annotation as well as quality and functional filtering (for example read depth, mapping quality,

unique start position, strand bias, repetitive regions, pathogenicity predictions or frequencies in the variants databases) plus statistical analyses.

The NGS field being an extremely dynamic field, we will have to be able to adapt the laboratory workflows and also to perform “technology watch” for all analyses tools and pipelines.

Publication

Voegele C et al. , Bioinformatics, 2007

"A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening"

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 18 2007, pages 2504–2506
doi:10.1093/bioinformatics/btm365

Databases and ontologies

A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening

C. Voegele[1,*], S.V. Tavtigian[1], D. de Silva[1], S. Cuber[1], A. Thomas[2] and F. Le Calvez-Kelm[1]

[1]International Agency for Research on Cancer (IARC), Lyon, France and [2]Department of Medical Informatics, University of Utah, Salt Lake City, USA

Received on April 16, 2007; revised on June 20, 2007; accepted on July 6, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Summary: High throughput mutation screening in an automated environment generates large data sets that have to be organized and stored reliably. Complex multistep workflows require strict process management and careful data tracking. We have developed a Laboratory Information Management Systems (LIMS) tailored to high throughput candidate gene mutation scanning and resequencing that respects these requirements. Designed with a client/server architecture, our system is platform independent and based on open-source tools from the database to the web application development strategy. Flexible, expandable and secure, the LIMS, by communicating with most of the laboratory instruments and robots, tracks samples and laboratory information, capturing data at every step of our automated mutation screening workflow. An important feature of our LIMS is that it enables tracking of information through a laboratory workflow where the process at one step is contingent on results from a previous step.

Availability: Script for MySQL database table creation and source code of the whole JSP application are freely available on our website: http://www-gcs.iarc.fr/lims/.

Contact: voegele@iarc.fr

Supplementary information: System server configuration, database structure and additional details on the LIMS and the mutation screening workflow are available on our website: http://www-gcs.iarc.fr/lims/

1 LIMS ARCHITECTURE

1.1 The LIMS components (Fig. 1)

The LIMS is hosted on two secured Linux servers running a Debian Woody Operating System. One server is dedicated to the MySQL relational database and the other to the web application. The latter enables interactions with the laboratory's instruments, robots and analysis software. The whole system is integrated into an internal network and secured within a strict firewall.

The MySQL database embodies a complex data model that integrates different types of data from sample features to result reports including all plates and reagents that are used along the mutation screening process. The database schema mirrors the workflow and is sufficiently flexible to allow evolution in workflow subprocesses. The content is secured by a combination of different kinds of daily and incremental backups.

1.2 User interface

The LIMS presents a friendly and intuitive user interface; users navigate within the application in the same manner as they are accustomed to navigate web sites, using menus that follow the laboratory workflows. A series of ~250 Java Server Pages (JSPs) enable database management by linking the user interface to the database. The interface includes one or more display screens that are specific for each step in our process and in which the related data are presented in tables that list the queried database content. Each laboratory step has a corresponding LIMS transaction (mostly plate barcode transactions) that triggers addition(s) or update(s) of single or batches of database records. All of these functions are managed through 'restrictive' forms that use pull-down menus, whenever possible, to minimize form-filling errors. Specific JavaScript form validations check the data type and integrity with respect to other tables in the database and guide the users through the workflows by prompting them to fill the required fields of one step before moving on to the next step. Plate barcodes are entered using barcode scanners to avoid miswriting, and each database transaction is stored in the database with the user's name and the date.

Information retrieval is managed through two types of queries. The first type focuses on the content of individual database tables. The second type, much more complex, includes relationships between the tables to retrieve the current content and process history of plates as well as the final experimental results. These queries generate multi-entry tables that are easily converted to tab delimited text for display or analysis with other software.

This web-based approach results in a zero-installation for the users and enables access from any computer within the LAN. The system supports unlimited simultaneous connections;

*To whom correspondence should be addressed.

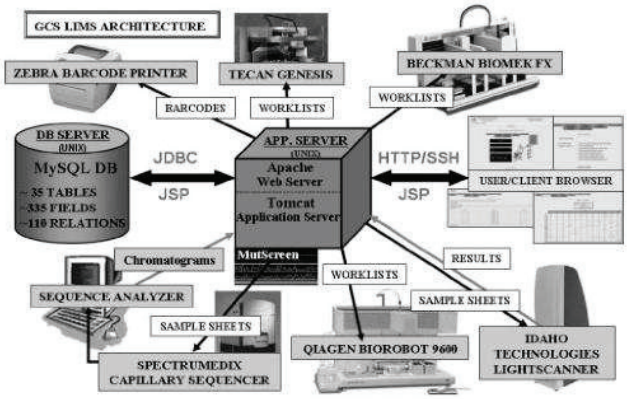


Fig. 1. LIMS architecture.

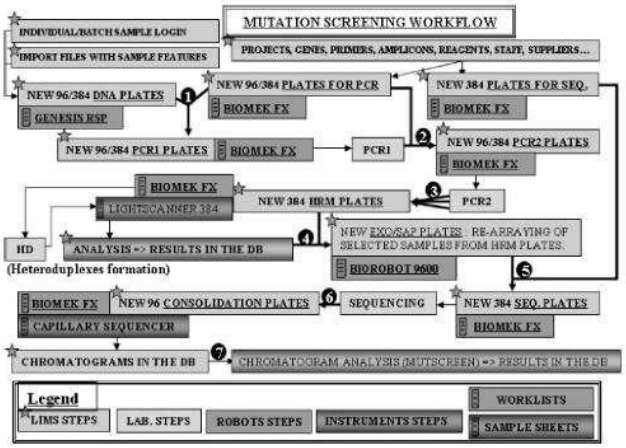


Fig. 2. Overview of the mutation screening workflow.

operates equally well whether the individual user is working in a Windows, Mac OS or UNIX environment, and is compatible with most popular web browsers.

1.3 Creating and deploying new modules

The LIMS is integrated into an Ultimate Bulletin Board (UBB), which is a customized application based on version 6.02 of the UBB from Infopop, Inc. The application contains a framework for database table creation and standard basic web page generation. Each database table generally corresponds to one entity or one step in the workflow and has at least seven associated standard files for listing, adding, updating and deleting entries. In addition, a large number of pages have been created for specific functionalities and for more sophisticated modules such as plate pictures, results requests or scripts for sending files to printers and robots.

User authorization access has been implemented for different levels of functionality through the UBB. LIMS users are identified with personal logins and passwords stored encrypted in the database together with the authorization codes for all activities. Thus, an integrated and safe environment is maintained for the whole application.

2 IMPLEMENTATION

To detect new rare sequence variants, our mutation screening strategy relies on a combination of High Resolution Melt curve analysis (HRM) and resequencing (Chou *et al.*, 2005; Margraf *et al.*, 2006; Tavtigian *et al.*, 1997). All laboratory steps of the process are mirrored in and managed through the LIMS.

Over the next few paragraphs, we outline the workflow displayed in Figure 2. The description of step 4 is expanded to highlight LIMS management of the contingent process occurring at this step. After primary (1) and secondary (2) DNA amplifications, the products are consolidated to a 384-well HRM plate (3). The HRM plate is then queued for formation of hetero-duplexes and collection of HRM profiles on a barcode reader equipped Idaho Technologies 384-position LightScanner.

(4) The HRM curve results are imported in the LIMS database, and based on these results, a subset of samples, including all of those with HRM profiles suggestive of the presence of a sequence variant, are selected for sequencing. In the LIMS, the barcodes for 1–3 384-well HRM plates are specified for re-arraying and a new 96-well sequencing plate is created. The specification process launches the display of the content of the HRM plates color-coded by HRM result into 3 groups:

(Group 1) samples that must be resequenced because their HRM curve was indicative of the presence of a sequence variant—these are already pre-selected—(Group 2) samples that can be resequenced because their curves were at the edge of the distribution of apparently homozygous wild-type samples—(Group 3) samples that are generally not resequenced except to provide a few wild-type controls because their curves were clearly indicative of homozygous wild-type.

In principle, the samples that will be selected for resequencing (all of Group 1 and some of Groups 2 and 3) are randomly distributed across the source HRM plates. Selection by the user of up to 96 samples launches a JSP function that: (a) creates the transfer pattern for re-arraying and consolidating the selected samples from the HRM plates into a 96-well plate (called EXO/SAP because PCR products are subsequently digested with exonuclease I and shrimp alkaline phosphatase). The rearrangement is done automatically in a specific order so that samples derived from an individual amplicon are transferred to consecutive positions on the plate. (b) Inserts the new plate's features and well content in the database. The identities of each sample queued at this step are thus contingent on results from the mutation scanning, their process history being recorded in the LIMS. (c) Sends a worksheet to the Qiagen BioRobot 9600 to launch the re-arraying program as specified by the LIMS.

PCR samples are then purified and sequenced in both senses (5). The 4 individual (A, C, G, T) dye-primer sequencing reactions are consolidated before being cleaned up and injected on to a 96-capillary electrophoresis sequencer (6). The run produces a pair of forward and reverse SCF format chromatograms for each sample. These files are deposited in the

database linked through their process history to the samples from which they originated and the amplicon for which they were generated. The pairs of forward and reverse chromatograms are then ready for sequence analysis (7).

3 COMMUNICATIONS WITH INSTRUMENTS

The LIMS interacts with most of our laboratory robots and instruments, all of which are run by computers integrated into the internal network. Communications are bidirectional and are mediated by sending and receiving text files. The transfer is ensured by a JSP tag that launches a general loading function, sending specific files to specified computers by calling a Perl script. Below, we describe the interactions with the barcode printer, the LightScanner and the BioRobot. Although not described here, worklists specifying the positions and barcodes of plates to be processed are also sent to the Tecan and Beckman robots and to the sequencer.

Correct plate tracking and therefore barcode printing is fundamental to our workflow. Each creation of plates triggers the creation and sending of a text file to a specific directory on the computer that manages the barcode printer, which launches instantaneous printing. The LIMS generates the barcodes by incrementing the highest plate id of the same type stored in the database, thereby guaranteeing barcode uniqueness. The files sent to the barcode printer contain these barcodes to be printed, encoded in the barcode printer language EPL2.

Communication with the LightScanner is another key element because downstream process decisions are contingent on HRM results. Sample sheets containing the barcode id of the HRM plate to be loaded in the LightScanner, as well as the positions, ids and amplicons of every sample on that plate, are sent to the computer that runs the instrument. Users associate each HRM plate with its sample sheet by barcode. After analysis, the mutation screening results are exported into a new specific sample sheet named with the HRM plate barcode. In the LIMS, the contents of these sample sheets are parsed to transfer each datum to the appropriate tables and fields of the database.

Based on the HRM results, samples are selected for sequencing. A worksheet containing the (1–3) source plate barcode(s) and the one destination plate barcode is sent to the Qiagen BioRobot 9600. This worksheet also contains the source and destination well coordinates that specify the required sample transfers.

4 CONCLUSION

LIMS has now become essential to our laboratory activities. Its use in combination with automation of laboratory processes has improved the efficiency and quality of the work by reducing potential for human errors, accelerating the throughput of analysis, and enabling sample tracking activities that are very difficult to perform without error by hand. The first two projects relying on this system in its entirety have been complete mutation scanning of the promoter, exons and introns of TP53 in a series of 50 subjects and mutation scanning of all of the coding exons of CHEK2 in a series of 1300 subjects. These studies resulted in the generation of ~20 000 chromatograms; because the mutation screening is equivalent to genotyping every base pair screened, the analyzed results correspond to 4.2 million genotypes (Garritano *et al.*, and Le Calvez *et al.*, manuscripts in preparation). We routinely maintain throughput of approximately 50 PCR plates per week.

While several commercial software packages now also provide analogous web-based architecture, phases of specification to adjust these packages to specific laboratory needs are still laborious and suffer limitations. Our LIMS programmed *de novo* can be flexible and expansible, providing options for continuous improvement and providing a template for the development of new modules governing new workflows.

Conflict of Interest: Dr. De Silva is currently employed by Idaho technology Inc., and holds stock in the company.

REFERENCES

Chou,L.S. *et al.* (2005) A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductor regulator gene as a model. *Am. J. Clin. Pathol.*, **124**, 330–338.

Margraf,R.L. *et al.* (2006) Mutation scanning of the RET protooncogene using high-resolution melting analysis. *Clin. Chem.*, **52**, 138–141.

Tavtigian,S.V. *et al.* (1997) Genomic organization, functional analysis and mutation screening of BRCA1 and BRCA2. In: Fortner, J.G. and Sharp, P.A. (eds.) *Accomplishments in Cancer Research 1996*. Lippincott-Raven Publishers, New York, pp. 189–204.

Chapter III – A SAmple Management System for the IARC biobank (SAMI)

III] 1. Introduction

III] 1.1 Biobank

A biobank is a repository of biological materials that collects, processes, stores and distributes biospecimens and associated information to support future scientific investigation. Since the late 1990s biobanks have become an important resource in medical research supporting contemporary studies in many fields and particularly in genomics and personalized medicine. The value of these biobanks has increased as laboratory technologies and especially genomic technologies have advanced.

Scientists used to collect biological specimens for their own experiments but quickly realized the interest in sharing the specimens for cross purpose research studies to acquire sufficient samples for powerful value. Now biobanks are opening their access to worldwide researchers following both the increase in number and size of samples collections as well as demand for use. As an example, in 2008 United States hosted in their biobanks 270 million biospecimens with a rate of new collection at 20 million per year. Such large biobanks require careful management of the samples and their associated data with tight control of accesses to personal and medical information for ethical and privacy issues ([99] - Haga and Beskow, 2008).

III] 1.2 The IARC Biobank

The IARC develops a large portfolio of studies including cohort, case-control and case-only studies with samples from various parts of the world. It implied the establishment of a large-scale biobank essential to conduct molecular pathological and/or epidemiological studies.

The IARC Biobank (IBB) hosts one of the largest, most varied and richest international collections of samples in the world. The IBB currently contains 5 million of biological samples from over 50 different studies. Four million of them have been drawn from 370,000 individuals enrolled in the European Prospective Investigation into Cancer and Nutrition (EPIC) study ([100] - Riboli and Kaaks, 1997). More than half a million of participants were recruited across 10 European countries and

followed for almost 15 years for investigating the relationships between diet, nutritional status, lifestyle and environmental factors, and the incidence of cancer and other chronic diseases. Close to one million of samples are coming from other smaller collections; some of them being extremely precious as originated from isolated ethnic groups or rare cases.

The IBB contains both population-based collections from research projects focusing on gene-environment interactions (as in the EPIC study) and disease-based collections aiming at exploring some specific biomarkers. Study designs include case-series, prevalence studies, case-control and cohort studies.

The IARC Biological Resource Center (BRC) hosts these millions of samples of various origins and nature (**Figure 31** and **Figure 32**) within a large and heterogeneous infrastructure. Indeed, storage facilities include 48 liquid nitrogen tanks at -196°C, 68 freezers at -80°C, -40°C and -20°C, fridges, cold rooms and dedicated cupboards at room temperature for paraffin blocks, slides, hair and nail samples and dried blood spots.

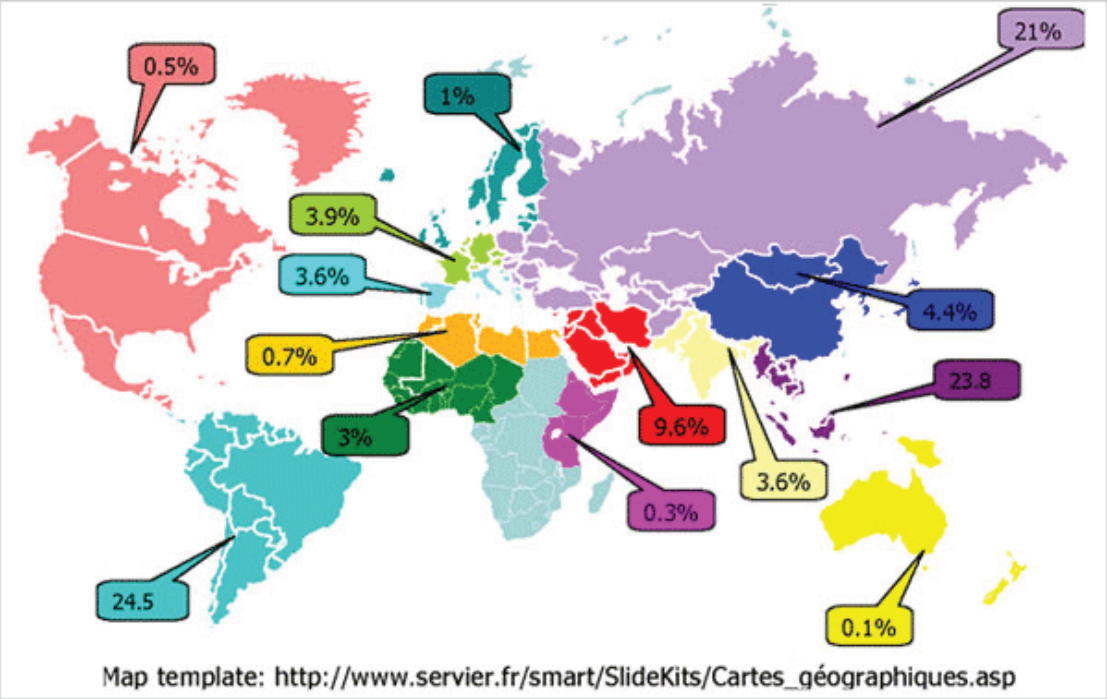


Figure 31: Geographical origin of IARC sample collections.

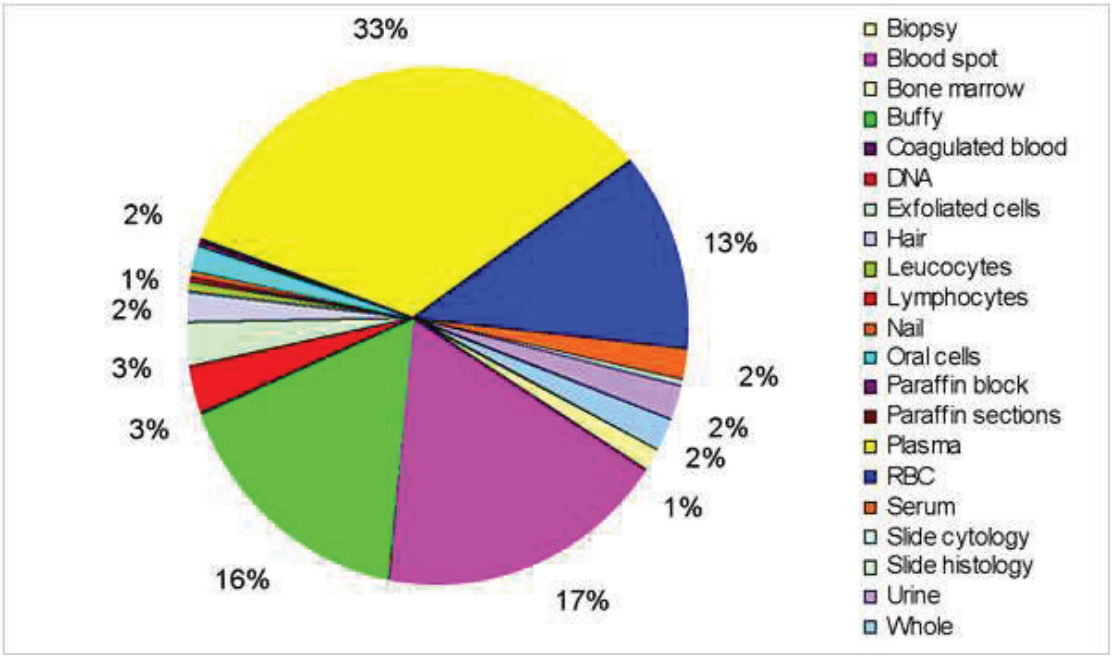


Figure 32: Break down of the type of samples hosted at IARC *(RBC stands for Red Blood Cells).*

III] 1.3 Challenges

Before the setup of common biobank management and infrastructure a decade ago, samples used to be spread in different places at the IARC as well as their information. The tools used to keep track of the samples varied from small databases to simple paper or electronic documents not all maintained up-to-date and spread within the different research groups leading to a lack of global visibility regarding IARC's sample collections and storage.

Moreover since IARC is continuously developing new research projects implying the arrival of new biological collections, very precise standard operating procedures have recently been set up for the standardisation and automation of all new sample reception, aliquoting, barcoding, DNA extraction or shipment in order to facilitate correct cataloguing of the specimens ([101] - Caboux et al., 2007). Some of the samples undergo analytical processes which need to be traced to avoid loss of information.

One of the major roles of IARC being to promote scientific collaborations between countries and research centres, the samples or some aliquots often leave IARC to sometimes return. It is also important to be able to follow these movements.

The diversity of study designs, structure, annotations and specimen collections are extremely difficult to accommodate into a single sample management system (SMS). The challenge was to build a system that could accommodate with all these different types of data whenever available, record their position, track their movements and allow queries to facilitate usage of the samples as needed.

III] 1.4 Specifications of the needs and requirements for the SMS

At the inception of the project in 2010, we had in-depth discussions with the biobank manager to get precise description of the state of the collections, the storage infrastructures and the needs. The consultation process started by a review of the existing collections and facilities, providing a basis for defining extensive specifications incorporating the whole range of sample and storage units' diversity. The first requirement of SMS was indeed to cope with the inherent variability and heterogeneity of both sample's primary data and storage infrastructure.

Samples specifications

The samples of the biobank are biological specimens of different types. They are defined by the project they belong to and several associated individual and epidemiological data as well as qualitative and quantitative features. In collaboration with the biobank managers and the biobank steering committee which aims at providing general guidance and best-practice in the broad area of biospecimen collection and use, a minimal data set was set up for efficient cataloguing following international standards from Biobanking and BioMolecular resources Research Infrastructure (BBMRI-ERIC) ([102];[103] - Wichmann et al., 2011).

This standard minimum required information on collections enables harmonization to facilitate pooling of samples in combined analyses from large scale studies ([104] - Norlin et al., 2012). Two types of data can be distinguished: the mandatory one and the optional one that adds value to the samples.

Storage facilities specifications

The samples are stored in various types of containers which are pieces of equipment going from the buildings' rooms (first level) up to the small tubes or plate wells in which the sample is placed (last level) going through the freezers' drawers and boxes. The containers are linked in a fashion analogous to a Russian doll (**Figure 33**) and form hierarchies of containers. We distinguished three main categories of hierarchies following the same types of sub-containers:

- liquid nitrogen tanks which are of two kinds (tanks with racks or tanks with canisters)
- freezers and fridges at various temperatures containing drawers and boxes
- cupboards with shelves and boxes

In total IARC uses over 70 different kinds of containers among all 6 types of freezers, 11 types of racks , 22 types of boxes, 5 types of bags, and 4 types of tubes. A comprehensive and strict hierarchy from the room to the single tube containing the sample 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 should

have the capacity to evolve by incorporating new hierarchical levels, new storage devices within each level as well as movements and transfer of containers within each hierarchical level.

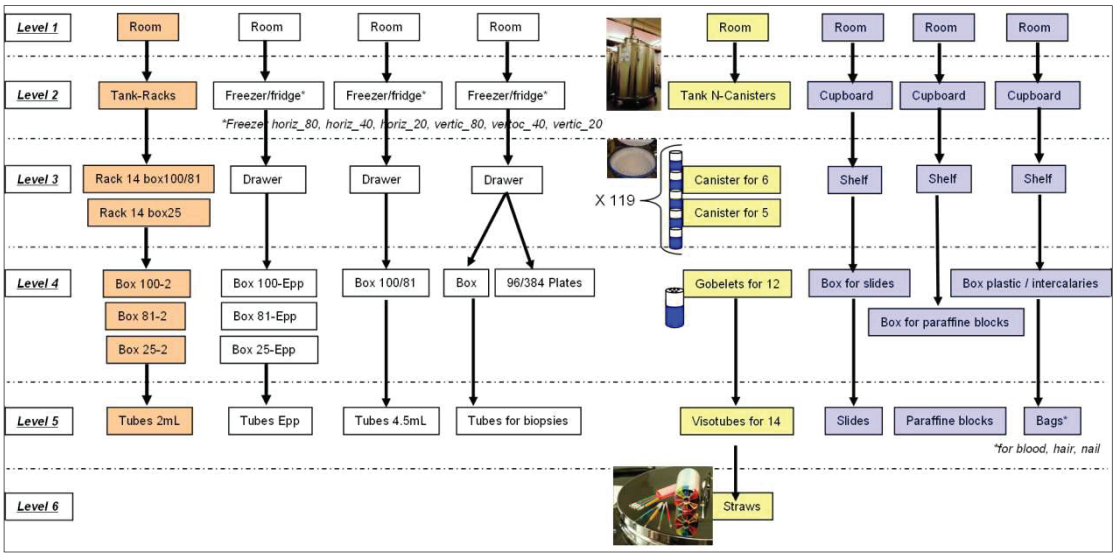


Figure 33: Global containers hierarchies. *There are up to 8 levels of containers with 3 different types of hierarchies: those using liquid nitrogen tanks (orange and yellow), those using freezers or fridges (white) and those using cupboards (blue).*

Samples' and containers' movements' specifications

The second requirement for SMS was to keep track of all samples' or containers' movements and status or quantity changes, including the management of dynamic processes from specimen retrieval to transformation, extraction and aliquoting of by-products, transfer to in-house analytical platforms and shipment to other research centres. It should improve the traceability of samples and their derivative products as well as the quality assessment based on the number of defrosting and re-freezing.

Data import specifications

The third requirement for SMS was 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 did not aim at replacing the study–specific databases developed by clinicians and epidemiologists but at interfacing with them to provide a dynamic sample and data management system.

Security and use specifications

Another quite critical requirement for the SMS is the need to comply with levels of data safety and confidentiality compatible with the high ethical standards of research conducted on human biospecimens ([105] - Hansson, 2009), by providing appropriate security in terms of access and backups.

Finally, the SMS aims at being used by people with different background and expertise, from laboratory technicians who need to pick samples for their manipulations to the data managers and researchers who need to know what they have at their disposal for their studies. The tool should therefore be easy to use offering an intuitive interface and be based on simple and universal concepts.

This chapter will explain how the system we developed addressed all these requirements while enabling 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 transformation, aliquoting or shipment to in-house or external analytical platforms.

III] 2. Developments

Although there is a large and growing body of literature on the design and implementation of biobanks, the vast majority of these publications focus on sample acquisition and on practical aspects of sample storage ([106] - Elliott and Peakman, 2008). In contrast, very few publications are specifically dedicated to information management systems that handle and keep track of biospecimen collections. Existing literature, as well as most of currently available commercial softwares, are addressing specific types of collections and their associated analytic workflows – for instance tumor banks – ([107] - Owen and Woods, 2008). To our knowledge, there is no description of freely available systems that are large and flexible enough to accommodate a wide diversity of biospecimen types and storage conditions.

Like for the ELN we decided not to purchase an existing software because it would have been impractical to adapt any commercial tool to the extremely specific and diversified needs of our large and heterogeneous biobank. Also one of the Agency's research group had purchased a few years ago a software for managing the samples of one large collection they were developing. The tool was very rigid and customizations, which had to be made by the company were long and expensive. They were thus looking for an alternative solution and collaborated with us for the definitions of the needs in order to be able to switch from their system to the one we aimed at develop for the whole agency.

Finally we had acquired experiences and expertise with the development of the LIMS that could be used for SMS. We thought for a moment of having only one tool for tracking both the sample collections storage and the samples that underwent laboratory analyses but the needs are really different and thus the technical requirement as well.

It was therefore decided to use the internal human resources and knowledge to develop our own system based on a three-tier architecture with one Linux database server, one Windows 2003 application server and multiple clients (either Windows or Mac PCs) integrated into the Agency's internal network, secured within a firewall.

III] 2.1 Database

Given the large size of the datasets as well as previous experience and expertise, we decided to develop the relational database under Oracle 10g R2, which can manage a high volume of data and transactions, and provide high query performances. Oracle Suite also includes powerful and appropriate tools to ensure interfacing, safety and reliability with a high degree of recoverability and audit.

Three distinct databases "samidevdb", "samitestdb" and "samidb" have been set up. The first two ones respectively dedicated to developments and users' tests have been installed on one Linux server; and the last one hosting the data in production is installed on a separate Linux server.

Taking into account the needs and specifications laid out previously, the database model was designed to provide large flexibility so that it can easily accommodate evolution of storage facilities and samples' movements or transformations. This crucial step of design was one of the most complex steps as we needed to find common consensus definitions of the data types and features to setup standards despite the existing heterogeneity of the samples collections and storage infrastructures.

III] 2.1.1 Tables

Database tables are collections of related data consisting of fields (data attributes) and rows (data). The SMS database model includes 42 tables in total out of which we distinguish:

- 9 major tables identifying uniquely projects (sami_projects), samples (sami_samples), aliquots' specific attributes (sami_aliquots_attr), DNAs' specific attributes (sami_dna_attr), cell lines' specific attributes (sami_cell_lines_attr), containers (sami_containers), containers' types (sami_containers_types), samples' movements (sami_samples_movements) and containers' movements (sami_containers_movements) (**Figure 34**);

- 8 smaller tables for more static data management: sami_users, sami_users_datagroups, sami_contacts, sami_jobtypes_details, sami_menus, sami_constants, sami_samples_locations (stores the samples location hierarchies),

sami_samples_traceability (stores the samples projects assignments) and sami_glossary. The latter contains useful information on more than 50 different container types with explanations in both French and English languages, together with a photo of a container of each type to ensure everyone communicates effectively with the same and appropriate terminology.

- 11 tables for temporary data upload during the imports which are deleted once the data integrity is checked and the data is inserted into the standard tables (for samples, aliquots, DNA, cell lines, containers, samples movements, containers movements - cf. III] 3.1)
- 7 tables for imports and errors management and a few other tables for administrators' tasks.

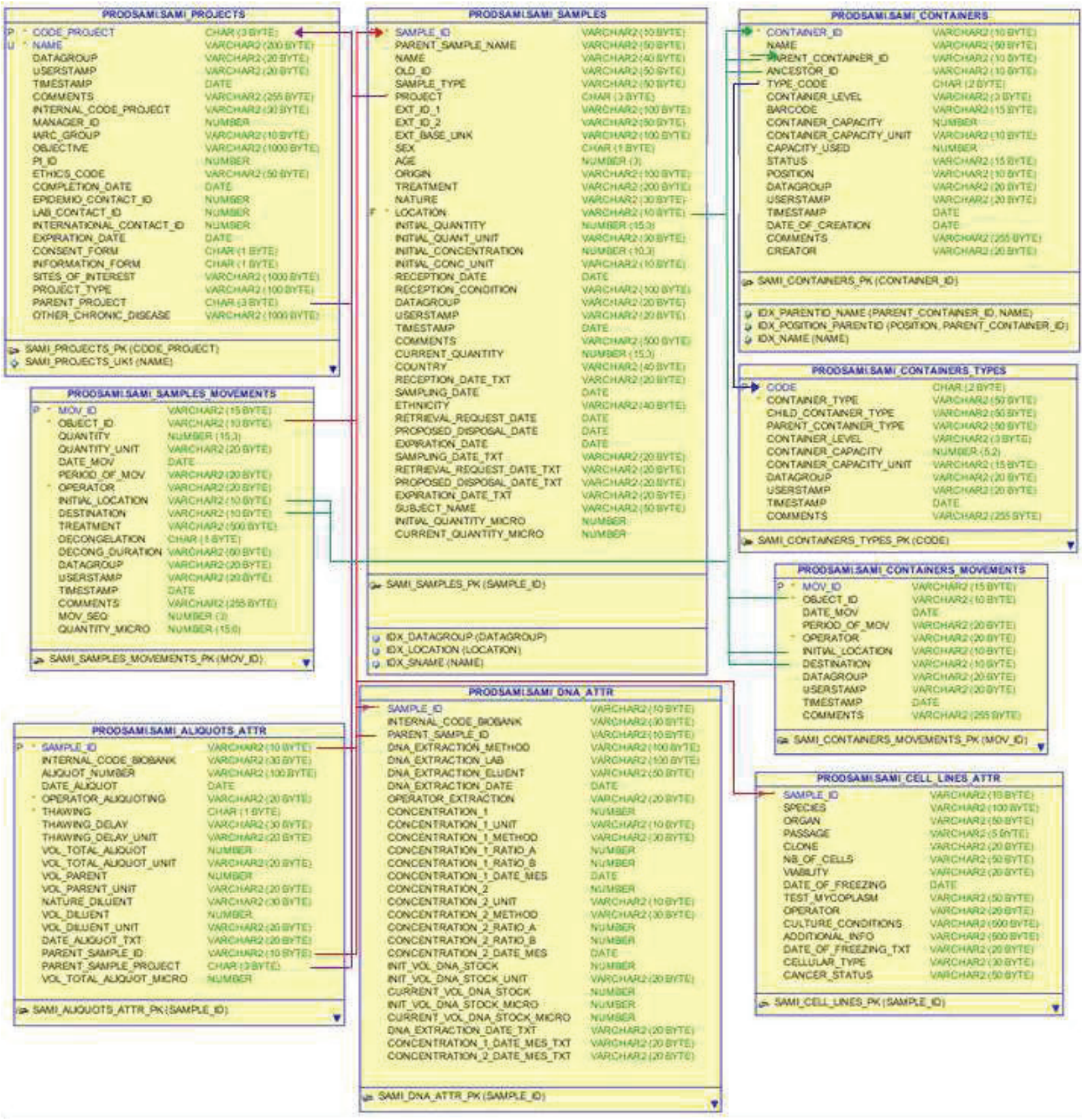


Figure 34: Relational schema of the main tables.

** Projects are at the top of the pyramid and there can be subprojects (permitted by the "parent_project" field). Each project is defined by a unique 3 characters code which enables 17576 potential combinations for projects identifiers and includes information on the managers at different level:*

project, epidemiology and laboratory as well as ethics consent and information forms.

** Samples belong to a specific project. For each sample, we record basic individual data (sample type, sex, age, origin, cancer status, different IDs to link them with more complex epidemiological databases), information on sample's quantity and quality (reception date and conditions, concentration, treatment), and location within the IARC biobank infrastructures. A database trigger enables to generate unique samples IDs for ensuring correct cataloguing (Cf. III] 2.1.2).*

One sample can be derived from another sample and will have a "parent" sample. This is the case for aliquots, DNAs extracted from plasma, or tissue and cell lines. These samples can have specific attributes recorded in separate tables such as:

- date and procedure of aliquoting which are stored in the table "sami_aliquots_attr"

- date and procedure of isolation, concentration which are stored in the table "sami_dna_attr"

- procedure of cell culture which are stored in the table "sami_cell_lines_attr"

** Any modification of quantity or position of the samples is stored in the table "samples_movements" with information on the type of modification, initial location, destination, operator and date. The history of movements for each sample can therefore be reported.*

** Containers are pieces of equipment for which we store in the database information on the type (with specific criteria defined in the table "sami_containers_types"), the position, the capacity, the capacity used and the parent container to be able to reconstruct the hierarchies.*

*We created the concept of "ancestors" which are the "top" containers from level 2 (**Figure 33**) in which all the "child" (sub) containers are named uniquely to define the exact samples' positions within the biobank infrastructure when providing the ancestor and the last level container (tube, plate well, straw, bag). They are for example the tanks, freezers, fridges, or cupboards in which we have only one drawer 1. This distinction enables to have several "drawer 1" in different hierarchies.*

The notion of hierarchies defined by ancestors themselves made up of different types of sub-containers makes it easy to add new hierarchies in the system.

** Containers can be moved within or outside a hierarchy and this is stored in the table "containers_movements" with initial location, destination and date of movement.*

The design of the table of containers which defines uniquely the location of the container with the parent container enables easy movement of any container since the global hierarchy can be instantaneously reconstructed from the new parent container.

III] 2.1.2 Indexes, views, triggers, packages and procedures for data management

We developed several other database objects for easier data management. First indexes were set up on the main tables to provide quick access to the rows in the tables. This is crucial for large tables containing millions of rows such as the samples or the container ones.

A few views were also created to facilitate access to only the most requested and relevant pieces of information avoiding long querying of the database. This is particularly useful for specific research groups' collections containing specific information.

In parallel, we set up some triggers and procedures for generating unique IDs for the samples (tg_get_sample_id) the containers (tg_get_container_ID), the samples' movements (tg_get_sample_mov_ID) and the containers' movements (tg_get_container_mov_ID). Samples IDs are created by combination of the 3 letters code of the associated project and 5 characters made of 10 digits and 24 letters excluding the "I" and "O" to avoid any confusion with "1" and "0" (**Appendix 4**). This enables generation of more than 45 million of different IDs per project. Similarly containers IDs are created using a combination of the 2-letters container type code, one numeric and 5 alphanumerics characters excluding also "I" and "O". This enables to store 454 million containers for each container type. Movements IDs are the concatenation of the sample ID and the incremented number of movements.

We also linked the system to an Eltron Zebra barcode printer and to a Brady 2D barcode printer for printing the samples and containers IDs, providing more efficient tracking of both types of objects (samples and containers).

To ensure samples collections data entered in the system follow the standards that we set up for it to be informative despite its heterogeneity, we developed:

- a few other triggers that send emails to the administrators whenever someone tries to insert incomplete or uncommon data, for example if the nature of a sample is missing or unknown (tg_check_sample_nature) or if the type of a container is missing or unknown (tg_insert_container).
- several PL/SQL packages that survey 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 major errors are found. Otherwise an email is sent to both the user which launched the import and the system manager, asking to check the returned errors.

  For example the "Sami_samples_checker package" is made up of the 19 checking procedures called according to the type of import and reading the temporary tables on which checks are applied if needed (**Table 8**):

Name of procedure	Checks performed
Check_Projects (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	**All the samples project's codes and names exist.**
Check_Natures (p_file_name IN VARCHAR2, p_filter IN VARCHAR2, v_res IN OUT CLOB);	**All the samples' natures exist and are not null (mandatory field);** **All the samples belong to the same datagroup.**
Check_Sample_Types (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	**All the samples' types exist.**
Check_Datagroups (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	**All the samples' datagroups exist, that they are not null (datagroup is mandatory);** **The user belongs to these datagroups.**
Check_Locations (p_file_name IN VARCHAR2, p_ancestor_id IN VARCHAR2, v_res IN OUT CLOB);	**The containers names are not null (mandatory field);** **The parent containers exist.**
Check_Location_Types (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	**The containers types codes exist.**

:
:

Name of procedure	Checks performed
Check_Location_Datagroups (p_file_name IN VARCHAR2, p_ancestor_id IN VARCHAR2, v_res IN OUT CLOB);	All location datagroups exist and that they are not empty (mandatory field); The user belongs to the provided datagroup; The samples' container's datagroup is the same as The parent container datagroup or is 'global'.
Check_Location_Capacities (p_file_name IN VARCHAR2, p_ancestor_id IN VARCHAR2, v_res IN OUT CLOB);	The parent containers is not full (by comparing the number of samples already in the specified container with its capacity).
Check_Units (p_units_list IN VARCHAR2, p_table_name IN VARCHAR2, p_column IN VARCHAR2, p_file_name IN VARCHAR2, v_res IN OUT CLOB);	All the initial quantities units exist; The initial quantity unit is not null if the initial quantity is not null; Both checks are also applied to vol_parent and vol_total_aliquot.
Check_Volume_Units (p_units_list IN VARCHAR2, p_table_name IN VARCHAR2, p_column IN VARCHAR2, p_file_name IN VARCHAR2, v_res IN OUT CLOB);	All the volumes units exist; Volume unit is not null if volume is not null; Both checks are also applied to initial volume of DNA stock and volume of diluent.
Check_Parents (p_file_name IN VARCHAR2, p_table_name IN VARCHAR2, p_ancestor_id IN VARCHAR2, v_res IN OUT CLOB);	The aliquots parent samples are not null; The parent sample project is not null; The column 'Parent used entirely' is either Y or N; The parent sample ID exist if it is specified.
Update_Parents_Ids (p_file_name IN VARCHAR2, p_table_name IN VARCHAR2, v_res IN OUT CLOB);	It retrieves parent sample ID using parent sample name and parent sample project.
Check_Thawing (p_file_name IN VARCHAR2, p_table_name IN VARCHAR2, p_ancestor_id IN VARCHAR2, v_res IN OUT CLOB);	The aliquots thawing is either 'Y' or 'N.
Check_Nature_Diluent (p_file_name IN VARCHAR2, p_table_name IN VARCHAR2, p_ancestor_id IN VARCHAR2, v_res IN OUT CLOB);	The aliquots' natures of diluent used exist.
Check_Aliquots_Volumes (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	The aliquots' total volume is the sum of the parent volume and the diluent volume.
Check_Passage (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	The cell lines' assage value is either null, "None" or "P>00<".
Check_Cell_Lines_Nature (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	All cell lines have the same datagroup; The cell lines natures exist.
Check_Cell_Lines_Type (p_file_name IN VARCHAR2, v_res IN OUT CLOB);	The cell lines types exist.
Check_Operators (p_file_name IN VARCHAR2, p_table_name IN VARCHAR2, p_col_name IN VARCHAR2, v_res IN OUT CLOB);	The operators for aliquots, DNAs and cell lines exist and are not null.

Table 8: Samples checks procedures.

We established similar verifications for the imports of containers' hierarchies and samples' and containers' movements.

As for the performance of the imports, the system was optimized to be able to check and load information on 120,000 samples in about 15 minutes.

III] 2.2 Web application

The content of the database is managed through a web-enabled graphical user interface developed with Oracle Forms Builder. It is platform-independent – except for Excel interactions only possible with Windows based computers - and compatible with all the web browsers supporting Java 6. 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 his permissions on datagroups.

In the same way as for the LIMS (Cf. II]3.2), users have a personal login and password to connect to SAMI web interface. A total of 47 forms makes up the interface and enable interaction with the SAMI database excluding any deletion of any kind of data (**Figure 35**).

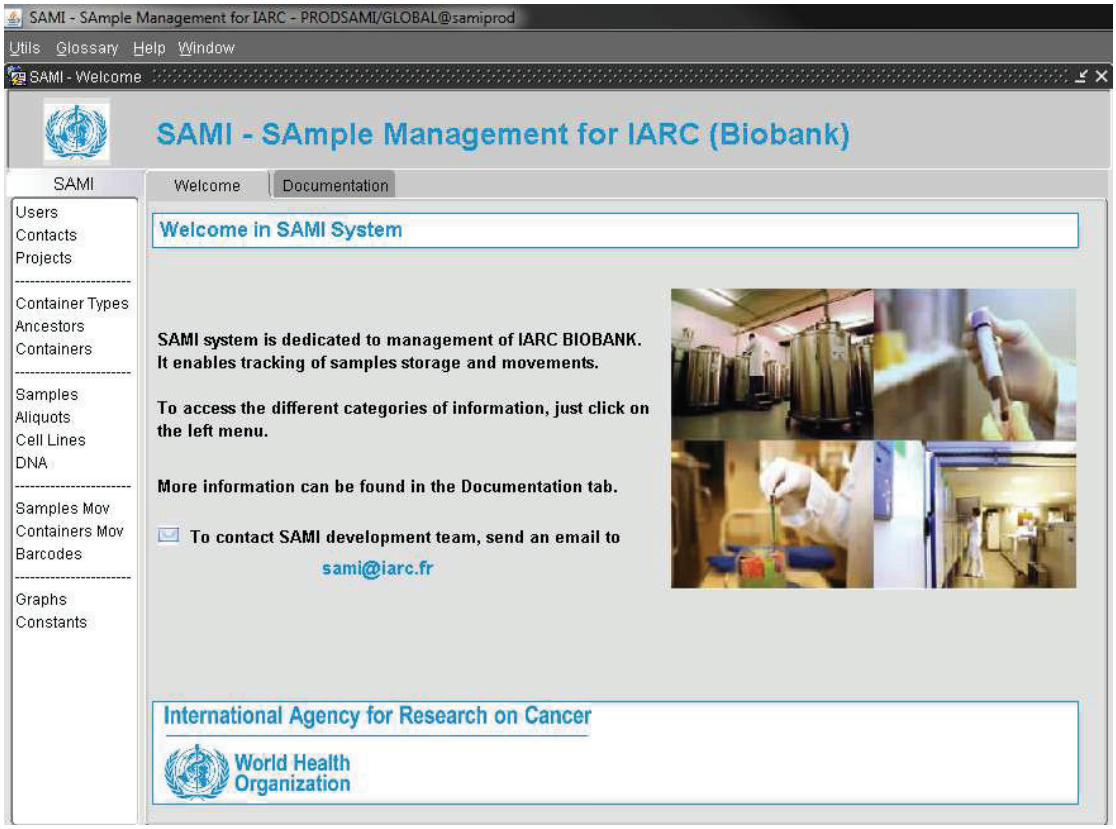


Figure 35: Welcome page of SAMI. *A menu on the left shows the different objects and opens specific forms dedicated to each object with different tabs to add information, search for information and list results of queries.*

III] 3. Results: features of the tool

As seen above, both the database and the interface were modelled to be compatible with and adaptable to almost all types of biobanks regardless of the nature of samples and the type of storage infrastructures. Data can be easily imported through forms or excel 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.

III] 3.1 Data import: example of samples' movements

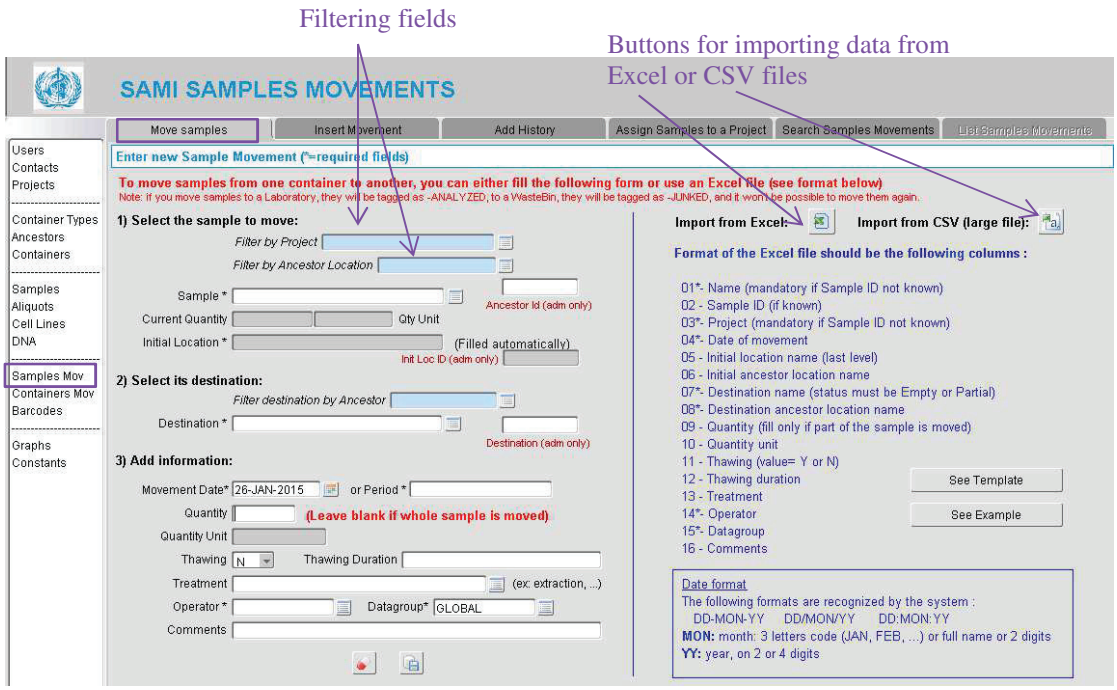


Figure 36: Form for entering samples movements' information.

*Data can be inserted in the database through filling of forms (**Figure 36** – left part) or upload of pre-formatted csv files that are read and checked by a set of PL/SQL procedures (**Figure 36** – right part). Users are guided through the forms as prompted to fill the required fields and to select values from pull-down menus that they can pre-filter on certain criteria (**Figure 36** – blue fields), in order to minimize form-filling errors. Samples' and containers' barcodes can also be entered using*

*scanners to avoid miswriting and some fields are dynamically populated depending on the information the user already entered (**Figure 36** - dark grey fields retrieving automatically the samples' current quantities and initial locations).*

Similar forms are available for the samples (among which specific ones for aliquots, DNA or cell lines), the containers and the containers' movements.

For both ways of data insertion (form filling and upload of file), the PL/SQL validations check data type and its integrity with respect to the tables in the database. E-mails are automatically sent to users and system managers in case of errors during data imports or other problems.

The imports use Oracle SQL Loader tool ([108]) that enables the loading of any flat file type into the database thanks to its powerful data parsing engine. It uses control files which define the loading rules: where to find the data, how to parse and interpret it with the link to our PL/SQL procedures and where to insert it. Traces of imports are registered in "LOG" files and invalid data are kept in "BAD" files.

We developed also within samples' and containers' forms a function that enables to build automatically from a list of samples or containers a file for import of movements containing the ID, initial location and so on.

It is also possible to add in a retroactive way a history to a sample which went through one or several movements before it was recorded in SAMI database and for which the initial location stored in SAMI is not actually its real initial location. A specific module developed in the tab 'add history' enables to store this additional information (**Figure 37**).

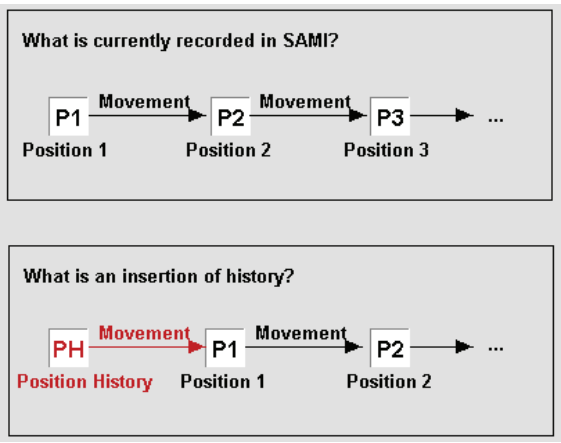


Figure 37: Retroactive movements.

Finally, a sample can be moved from one project to another thanks to the tab "Assign Samples to a project". This is particularly useful when a sample need to be used in several projects. Since one sample can only belong to one project, the solution is to create a master project grouping several projects and to move the sample to the master project.

III] 3.2 Reporting

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

Screens for searching and listing data (Figure 38)

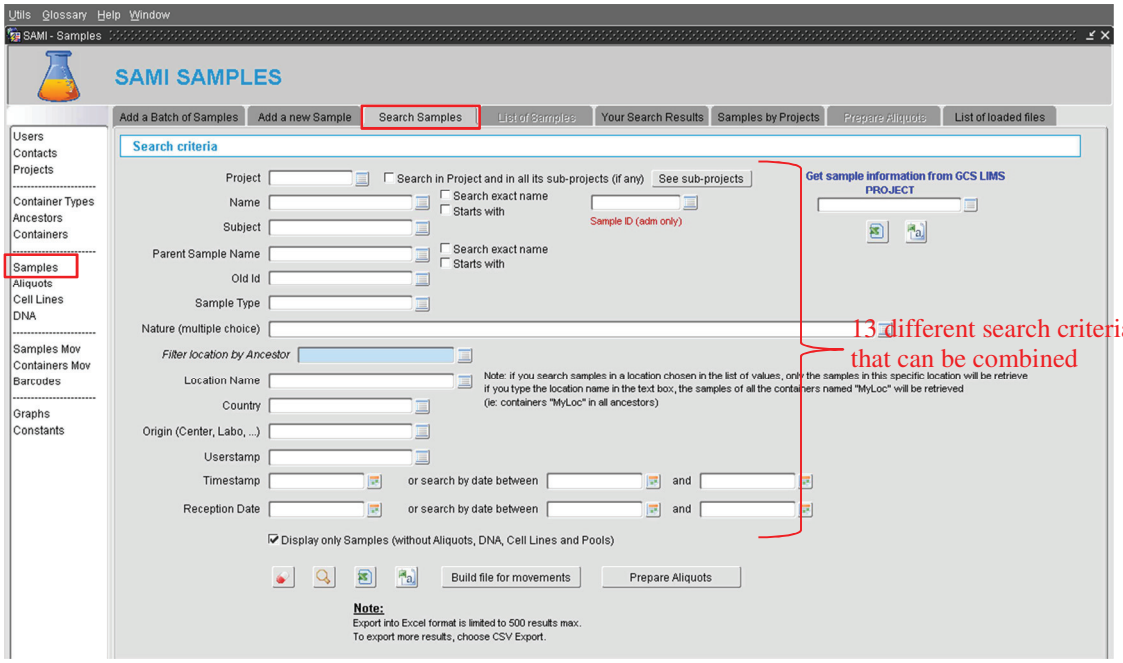


Figure 38: The form to search for sample information. *Each form has a specific "Search" tab with large set of search criteria (). Results of the queries can either be displayed within the interface in the "List" tab by clicking on the magnifying glass button*  *or be exported directly in Excel or csv file by clicking respectively on the Excel button*  *or csv button*  *. However, the exports in Excel are limited to search generating less than 500 results for performance issues since the file is generated on the fly. Otherwise users are requested to export the search results in csv format (the file is created and stored on the user computer).*

The tab displaying the result of queries is divided in two main parts: on the top there is a list of the objects responding to the search criteria with basic information (Figure 39 - A) and on the bottom there are more details on each object of the list (Figure 39 - B).

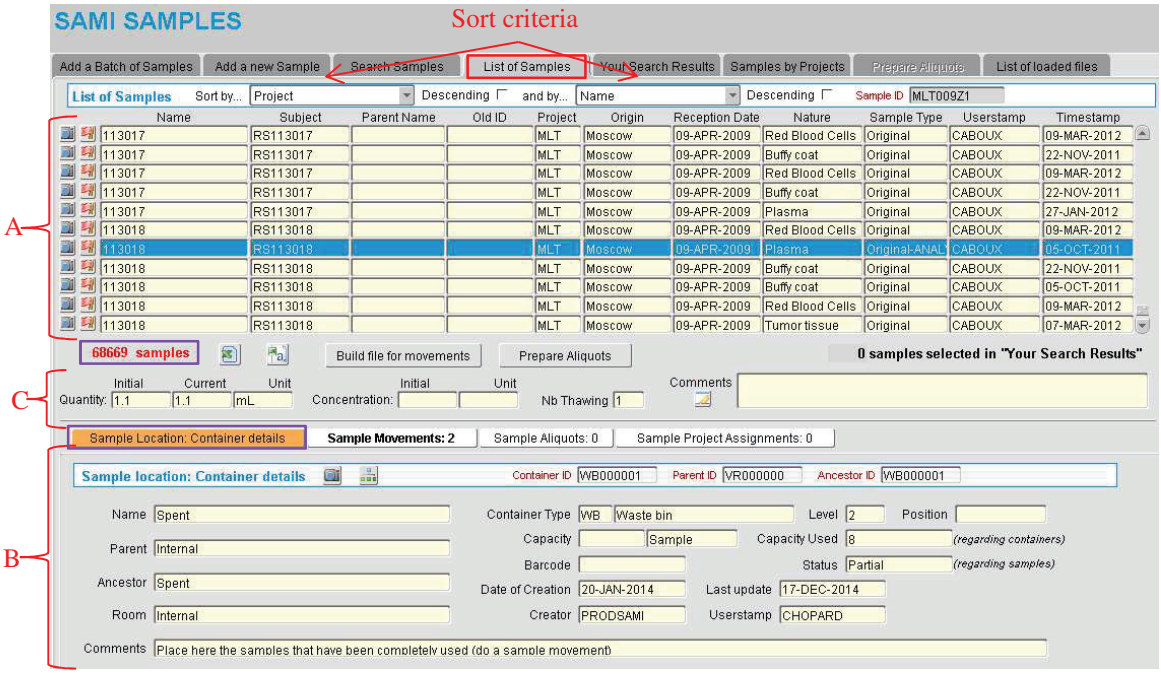


Figure 39: Screens showing results of samples query for a specific project.

The list of samples contains in particular information on subject name, project, origin, reception date, nature and sample type. The list can be sorted on project and/or sample name ascending or descending and a counter shows the number of samples responding to the search criteria.

On the left part of the list, a first button enables to show all the sample's details in a pop-up window. A second button enables to select only some samples for further export or to build file for import of movements.

Additional information on sample quantity, concentration and number of thawing is displayed just below the list (Figure 39 - C). Users can also find four sub-tabs at the bottom with:

** basic information on sample location (with a button to access details and a button to access the whole location hierarchy)*

** information on sample movements with their number (initial location, destination, date, quantity, treatment, thawing, operator)*

** information on sample aliquots with their number (name, date, method, quantity and volume as well as location)*

** information on project changes for the specific sample*

Container tree

A container tree displays all the hierarchies of containers. Navigating within the tree provides details on each container, its parents and childen (**Figure 40** - A) with the possibility to directly retrieve the list of samples that are stored in the selected containers either directly in the interface or through an export in Excel or csv file (**Figure 40** - B)

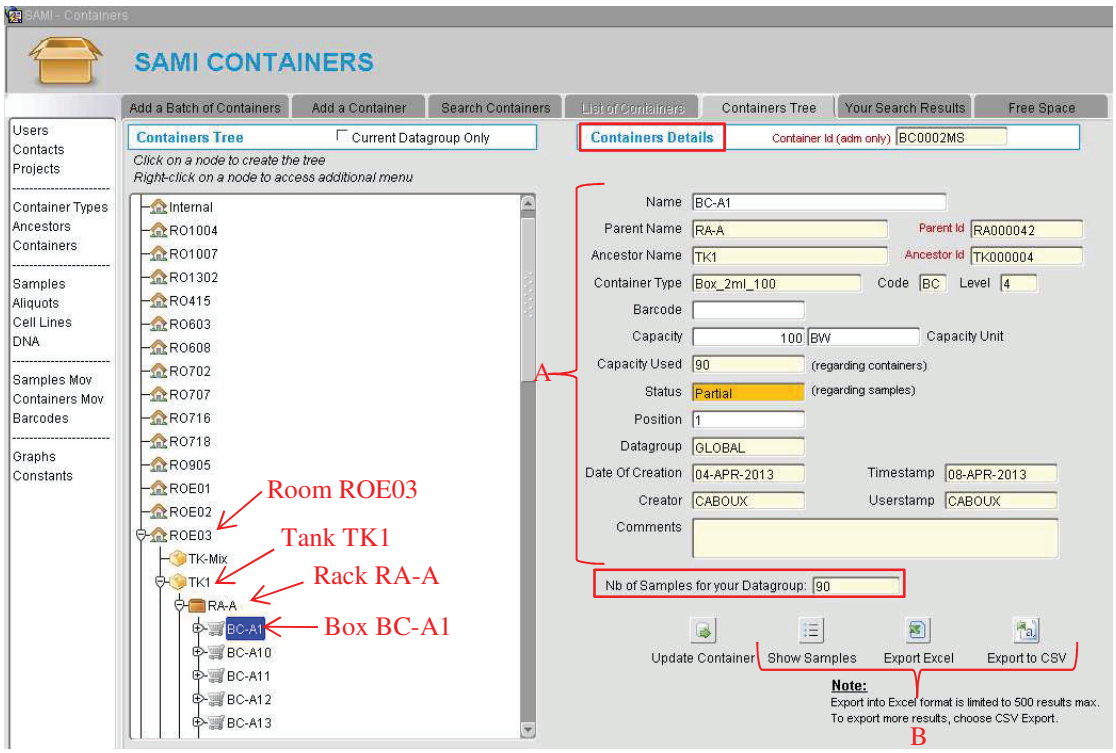


Figure 40: Containers tree. *The container tree shows the status of box A1 in rack A of the tank 1 which is stored in room E03. It is partially filled with 90 samples for a capacity of 100.*

Container pictures

For containers, a function enabling generation of specific pictures for visualization of the containers' content has been created. The type of picture depends on the type of containers with colour codes based on storage status (green when empty, orange when partially filled and red when full – **Figure 41**).

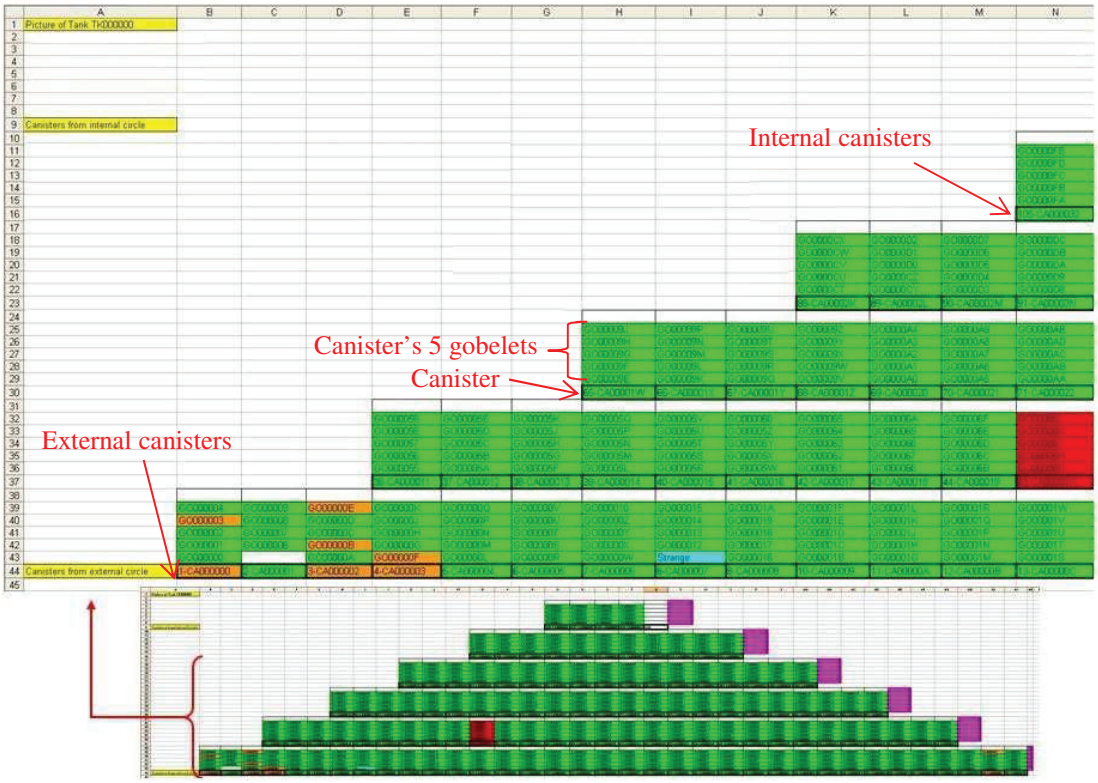


Figure 41: Picture of a tank. *By clicking on the "camera" button, the pictures of tanks are generated dynamically in Excel. Each level of the pyramid represents one circle of canisters from the more external (on the base) to the more internal (on the top). Above each canister are the 5 gobelets that make up the canister.*

Pictures of containers enable visualization of the content of the sub-containers, thus providing the possibility of displaying two levels of content (**Figure 42**).

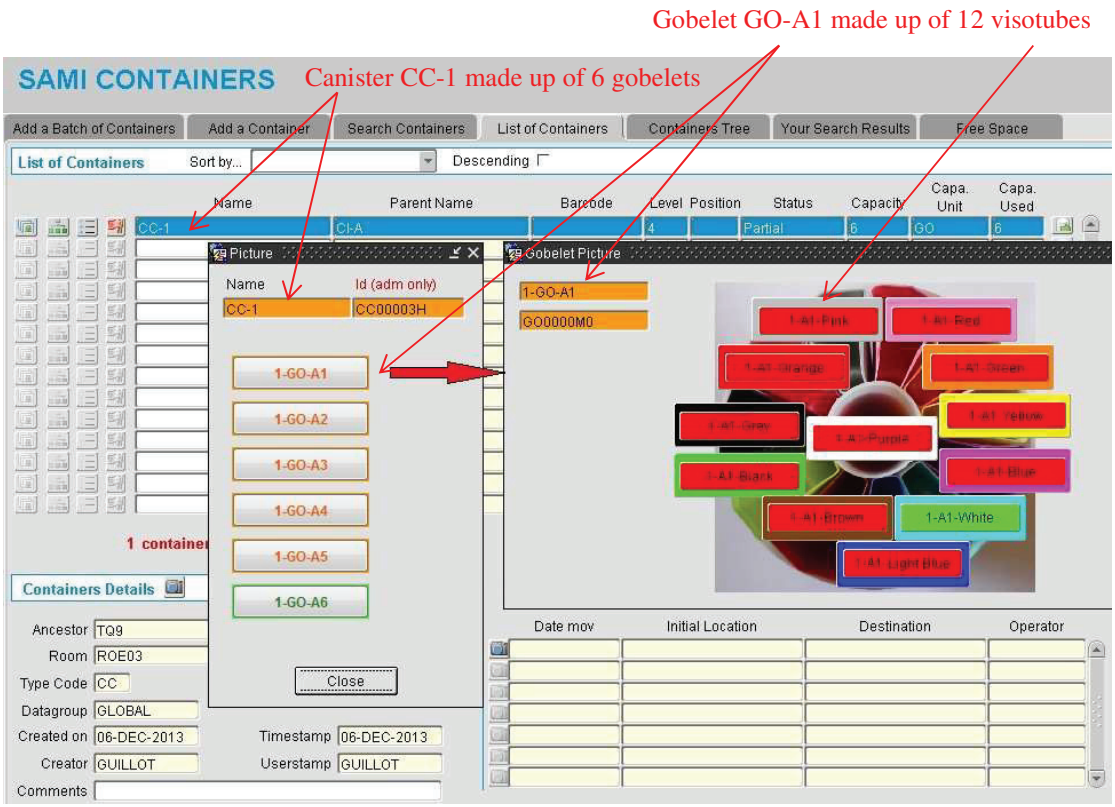


Figure 42: Picture of a canister made up of 6 gobelets with each 12 visotubes containing straws.

Graphs

For easy, practical and rapid visualisation of samples' data, graphs can be generated dynamically from any data stored in the database. Different types of graphs like bars or pies are directly designed in the form's code using Java beans based on a Java component provided by Oracle: the FormsGraph class.

Within the "Graph" form, available to administrators, a first tab shows an overview of the data stored in SAMI: number of samples, number of containers, number of samples movements and number of container movements per datagroup. The other tabs display graphs of relevant samples (**Figure 43**) and container information. Finally, the last "report' tab displays some other relevant statistics like number of samples or sample movements per user.

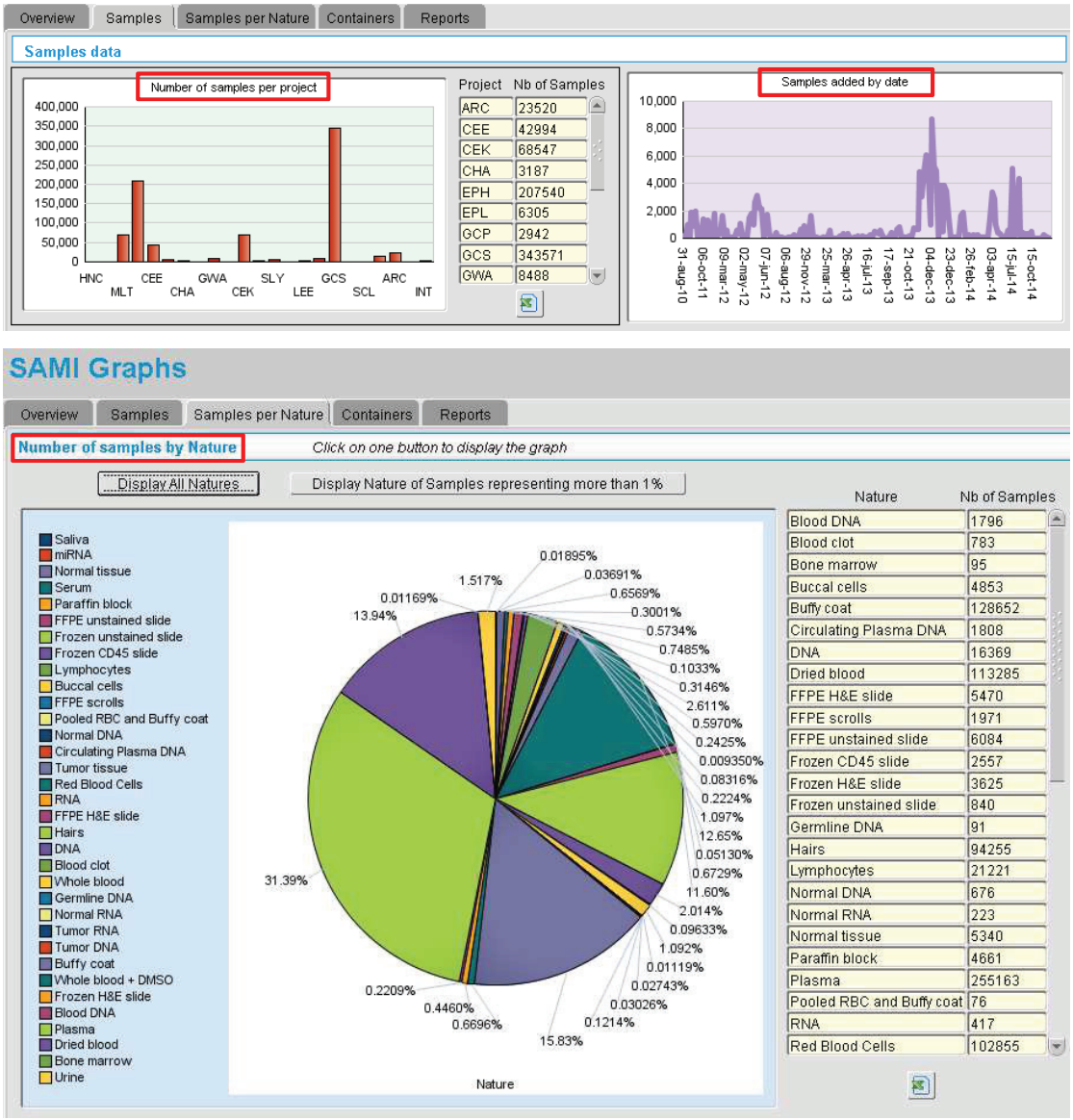


Figure 43: SAMI graphs. *This screenshot shows examples of graphs that can be displayed in SAMI (Number of samples per project, number of samples added by date or number and percentage of samples by type.*

III] 3.3 Security

Data management systems for biobanks dedicated to human specimens must comply with strict standards for confidentiality and protection of personal data ([109] - Eder et al., 2012). We therefore took particular care of anonymizing all specimens and of controlling the database access. Different IT solutions exist to achieve this aim. First, a trigger prevents direct connection to the database. User connection is only

possible through the LIMS interface. Secondly, all connection attempts are stored and unsuccessful attempts are reported to administrators via e-mail. Furthermore, the information stored in the database is divided into datagroups that can be specific to a research group, a study or a sub-study depending on the need to classify or restrict access to the data. Each sample collection belongs to one datagroup, and each user has permission to access data from 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 datagroup 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. In addition, each data modification is stored automatically in a secure field in each table with the user name and the date (userstamp and timestamp).

SMS's safety is also ensured by the implementation of a combination of automatic mirrored backups across servers. The result is sent by email to the administrator to ensure that no error appears during the process.

- The Oracle engine configuration and the Linux system are backed up once a week. In case of important changes made in Oracle or on the system, a complete image backup is done.
- The windows operating system of the Application server is also backed up once a week. The oracle configuration is saved into archive files once a week with a combination of batch and perl scripts. The application files such as forms, libraries or menus are saved into an archive file once a week.
- The full database is backed up using Oracle RMAN (Recovery Manager: ([110])) once a week, incremental backups being done every day. The database is never stopped thanks to the "archivelog" mode, which allows live backups. The frequency of incremental backups can easily be adjusted, depending on the volume of transactions.

III] 3.4 Summary and evaluation

Summary

The Sample Management System for IARC's biobank (SAMI) was officially implemented in 2011. It is now managing more than 5 million of samples from 321 projects and stored in 1,5 million of containers. We have already registered more than 185,000 samples movements but the inventory of the biobank sample collections is still on-going and new projects will bring new collections so the database will continue to grow.

At present six data managers from six different research groups are using SAMI for managing their sample collections which are divided in 8 different datagroups with specific read access permissions. These data managers can insert, update and delete data whereas additional "simple" users can but not delete any data. Insertions are done when new collections arrive at the Agency and updates are done when samples are removed from their storage for undergoing experiments or for sending to collaborators. Finally, several "readers" have only access in reading to what is stored in SAMI database. These are mainly researchers who are looking for samples with specific cancer types that would be available for enrolment in a particular study.

This results in daily querying, transactions and updates of the system which has now become essential for the management of IARC Biobank. SAMI also enables to provide summarized information on samples collections available that are listed on IARC biobank website.

Evaluation

To ensure the system is functioning properly, we tested each form, making imports and updates of data plus queries on the development environment using real data to verify that each piece of program was running as expected. We manually checked the consistency of small data sets entered or updated through the web interface during these tests.

We also developed PL/SQL procedures to control automatically the insertion and update of data reporting errors and warnings to both the users and administrators by emails. We receive an average of 5 of these emails per week (one third of the imports generate warning or errors but this is mostly because the data loaders rely on

the system for checking their data). In addition some of these checking procedures are launched automatically on daily, weekly or monthly basis through Oracle scheduling jobs. Indeed, the check of the table of samples locations is launched every day (checks that the sample and the location still exist), the check of the containers hierarchies and their status is launched once a week (respectively checks that a container has the same "ancestor" than its parent and that the status of the container (empty, partial or full) is correct regarding its capacity and actual content), and the check for containers duplicates (same parent and same location) is launched once a month. We get email alerts when these jobs reports errors (~ 2/months).

The optimization of the databases and the various checking procedures enables the storage of 120,000 samples in only 15 minutes. As for the accuracy of samples locations stored, samples are being removed from their storage every day for analyses and we have never been informed of a sample not being in the right place. However despite all the checks it is possible that a few of the 5 million samples may not be located where they should be (where the database indicates they are). But in that case, the reason is more probably linked to a human error during the imports or updates than a malfunctioning of SAMI.

Thanks to an appropriate set up of the database with a combination of indexes on the main tables, the search for data is still relatively rapid despite a global database content representing 5.7G of data (July 2015). It is difficult to give the exact times of querying since they depend greatly on the complexity of the queries (number and type of criteria) as well as on the number of concurrent connected users and what they are doing. However we performed some "benchmarking" within the interface using the querying forms and the results are:

- Retrieval within the interface of the list of all the samples with basic information (name, subject, parent sample, project, origin, reception date, nature, sample type) and links to get details on one sample's location, movements, aliquots, quantity, concentration, treatment…) took:
    - 3 seconds for a large study enrolling 207,540 samples (example from EpiHealth study)
    - 1 seconds for a medium study enrolling 9,649 samples (example of SAS study from South America)
    - Less than 1 sec for a small study enrolling 1,897 samples (example of LEE study from Leeds)

- Export of samples complete information including detailed features and location (Sample ID, name, subject, parent sample, type, project, external IDs, country, origin, nature, initial quantity, current quantity, concentration, reception date, sampling date, comments, location) in csv file took:
    - 45 min for the large study
    - 2 min for the medium study
    - 25 sec for the small one study

Finally we also considered the worst case of stress test which would be the crash of the server. For this, we performed complete restores of backups and were able to recover the full database.

III] 4. Discussion

III] 4.1. Cost

The sample management system is running on 2 servers with 2 CPU Intel Xeon 2.74GHz, 4G of RAM and 140G of storage. One hosts the Oracle database and the other the Oracle web application. The cost of the servers could be estimated to 2 x 1,500 euros so 3,000 euros.

To this should be added the cost of Oracle database, Oracle web application and per user licences which are annual licences. Oracle costs are greatly depending on the status of the institute whether public or private and on possible agreements. Commercial prices in June 2015 are around 30,000 euros in total:

- 21,350 $ (~19,000 euros) for a standard edition database including updates, licenses and support;

- 12,200 $ (~11,000 euros) for a standard web application including updates, licenses and support.

The time for installation of the database and the application server could be estimated to 1 week for an Oracle database administrator. It would cost around 600 euros based on a salary of 2,500 euros/month.

Finally the time already spent up to now on the development of SAMI which include the specifications, the coding, the tests, the servers administration, the data management, the documentation and the "helpdesk" to users, has been estimated to 900 days in total (~ 3,5 years full-time) for a cost around 100,000 euros based on a salary of 2,500 euros/month.

COST	**Cost for IARC (in euros)**	**Cost for other institutes (in euros)**
Material (servers)	NA	From 0 to 3,000
Oracle packages	Confidential	From 0 to 30,000
Installation	NA	600
Development	100,000	NA
TOTAL	**>100,000**	**From 600 to 33,600**

Table 9: Evaluation of the cost of SAMI implementation at IARC and at other research institutes.

Adding all costs, we end up with a total cost of more than 100,000 euros for IARC and between 600 and 33,600 euros for other research institutes (**Table 9**). 100,000 euros may seem quite high and in the range of similar commercial tools prices but as a result we have a tool specifically adapted to the management of IARC millions of samples from diverse origin and nature in heterogeneous infrastructures as well as a tool for which we have a complete knowledge.

III] 4.2. Advantages

As shown, 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 like management of tumour banks in hospitals (VitroPath), 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 ([101] - Caboux et al., 2007). This is the key requirement for efficient performance and correct use of this powerful material for all coming studies.

The system is relatively easy to install, flexible, expandable, and implemented with a high degree of data security and confidentiality system fulfilling stringent data protection standards. We took particular care of making available all the source codes under GNU General Public Licence (the most widely used free software license which guarantees users the freedoms to run, study, share and modify the software) and deployment information on GCS website. All the details to reproduce the system have been published in our Bioinformatics paper in the supplementary data ([111] - Voegele et al., 2010).

III] 4.3 Challenges

Though the system is very flexible, a set of rules has been put in place to ensure users fill correctly mandatory information essentially through the PL/SQL check packages. The main drawback is that importing existing data into the SMS requires particular formatting to standardize these data. This can be potentially time consuming for the user but the subsequent gain of time when searching for relevant information is so large that it is worth spending a few hours on the formatting. Also templates and data error reports via emails assist the users in performing the data imports.

Another issue is the Oracle and especially the Oracle versions and Operating Systems dependence for both database and web application points of view. The system could easily be installed in any type of biobanks but it is based on now quite old and obsolete tools versions and implies costs for licences. However the database could be easily reproduced in non-Oracle Database Management Systems (DBMS) such as MySQL but the interface would then need to be redesigned using non-Oracle technologies.

Regarding the ethical and privacy concerns raised by the access and sharing of data from biobanks, the tool we develop provide quite high security through specific user permissions and the separation in datagroups as described in III] 3.3. In addition personal patients data is not stored in our system but in external epidemiological databases with which it is possible to make a link if needed but not directly in the SMS.

III] 4.4. Other biobank's management systems

Different biobanks may have different needs for IT management so the software available on the market (**Table 10**) address different requirements from sample laboratory management (storage, processing, quality controls) to associated patient data management going through interfacing with laboratory equipment.

Commercial	Vendor/Provider	Country	Internet
BIGR Solutions Suite	GulfStream Bioinformatics	USA	www.gulfstreambio.com
BioTracer	CloudLIMS	USA	https://cloudlims.com/lims/overview.html
BSI	Information Management Services	USA	http://www.bsisystems.com/
CentraXX	Kairos	Germany	http://www.kairos-med.de/systemloesungen/centraxx-biobank-software/
C-line® control	ASKION GmbH	Germany	http://www.askion.com/
Cresalys	ALPHELYS, Alphametrix	France	http://www.alphelys.com
Easy Track 2D	TWIN HELIX S.R.L	Italy	http://www.twinhelix.eu/?act=products&sub=prodotto&prodotto_id=2435
eBioControl	Gefat-IT	Germany	http://www.ebiocontrol.de/ebc_gb.html
eurocryoDB	Fraunhofer IBMT (Custom Code)	Germany	http://www.eurocryo.de/
FreezerPro®	RuRo, Inc.	USA	http://ruro.com/software/freezerpro/overview
Freezerworks	Dataworks Development, Inc.	USA	http://www.freezerworks.com/
Investigate™	RemedyMD	USA	http://remedyinformatics.com/
ItemTracker	Itemtracker	UK	http://www.itemtracker.com/
Lab OS	soventec	Germany	http://www.soventec.de/company/index.php/de/lab-os.html
Labmatrix	BioFortis	USA	http://www.biofortis.com/next-generation-biobanking-platform/
LabVantage	Labvantage Solutions, Inc.	USA	http://www.labvantage.com/lims/biobanking
LabWare LIMS	LabWare, Inc.	USA	http://www.labware.com
Modul-Bio	Modul-Bio	France	http://modul-bio.com/
mysamples	myData	Germany	http://www.mysamples.de/
Nautilus	UpToData / Thermo Fischer Sceintific	Germany	http://www.uptodata.com/lims-en/products-en.html
noraybanks	noraybio	Spain	http://www.noraybio.com/en/products/noraybanks-en
ordoSYSTEM	Avalas	Germany	http://www.avalas.de/ordoSYSTEM-BioBanking_22.0.html
PathXL Biobank	PathXL	Ireland	http://www.pathxl.com/pathxl-research/pathxl-biobank
Pro-curo	Pro-curo Software Ltd	UK	http://www.pro-curo.com/
QuaLIS LIMS	Agaram Technologies	India	http://www.agaramtech.com/product/qualis-lims/overview.html
SampleNavigator	itemscope	Netherlands	http://www.samplenavigator.com/
Samples	Ziath inc.	UK	http://www.ziath.com/index.php/products/sample-management-software/
SampleWare	Biomatrica	USA	http://www.biomatrica.com/sampleware.php
SCARAB-LIMS	Karolinska Institute Stockholm	Sweden	http://ki.se/forskning/scarab-lims
SmartBiobank	AstridBio	Cananda	http://www.smartbiobank.com/
STARLIMS	Starlims Corporation (Abbott)	Germany	http://www.starlims.com/de-de/home/
Swisslab	Roche	Germany	http://www.swisslab.de/
TD Biobank	Technidata	France	http://www.biobank-network.com/en-us
Track-IT	Micronic	USA	http://micronic.com/product/track-it-sample-management-software
Open Source	**Vendor/Provider**	**Country**	**Internet**
ATiM	Canadian Tumor Repository Network	Canada	http://www.ctrnet.ca/atim
CAISIS	caisis.org	USA	http://www.caisis.org/
OBiBa BIMS	Obiba Consortium	Canada	http://www.obiba.org/
OpenSpecimen	Krishagni Solutions	India	http://www.catissueplus.org/

Table 10: List of available Biobank software *(from ([112] - Kersting et al., 2014)).*

Looking at the systems in place in other biobanks (**Table 11**), we noticed that most of them are not only managing samples collections storage but also include associate clinical data and questionnaires data (for example SmartBiobank ([113]) and Databiotec from Oriam used for the "Tumorothèque de Caen Basse Normandie" ([114])). In our case these information are in separated epidemiological databases for confidentiality and security reasons.

Some of them also include automation and informatization of processes for samples handling, storage and retrieval within the storage facilities thanks to robots. That is for example the case of the Karolinksa Institute Biobank in Sweden which is managed using different IT tools among which ELSA for external login of new collections and SCARAB-LIMS developed in collaboration with LabWare company for keeping track of samples' data and location within the biobank's freezers and tanks ([115]).

The UK biobank ([116]), one of the largest European biobanks which has collected since 2007 about 500,000 subjects for 20TB of data is using three different systems for data management. Commercial Nautilus LIMS from Thermo Scientific manages anonymized sample locations and is run at Cheadle ([117] - Downey and Peakman, 2008). Identification data (patient name, address, appointments…) and anonymized clinical data (including images and genomics) are stored separately in two systems run by the Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), Oxford University. Both are based on the open source Ingres RDBM ([118]) hosted on servers within the same network allowing transferring of data between the 2 systems thanks to Ingres built-in systems. These tools have been developed in-house largely in C++ by a team of 30 full-time programmers. Transfer of data between Nautilus and the patient and clinical data management systems are done using csv files via https and advanced encryption standards (Dr Young, Director of Information Science at CTSU, Oxford University, personal communication).

Which software are large european biobanks using?

Biobank	Country	Software
Biobank Suisse	Switzerland	CAISIS
BMBH: BioMaterialBank Heidelberg (cBMB)	Germany	Starlims
Hannover Unified Biobank (HUB)	Germany	MySamples / CentraXX
IBBL: Integrated BioBank of Luxembourg	Luxembourg	LABVANTAGE
IBDW: Interdisziplinäre Biomaterial- und Datenbank Würzburg (cBMB)	Germany	CentraXX
KI Biobank (Karolinska)	Sweden	SCARAB-LIMS + ELSA
Nationale Kohorte	Germany	CentraXX
P2N: PopGen 2.0 Netzwerk (cBMB)	Germany	CentraXX
RWTH: Aachen zentralisierte Biomaterialbank (cBMB)	Germany	Starlims
UK Biobank	UK	Nautilus
ZeBanC: Zentrale Biomaterialbank der Charité (cBMB)	Germany	CentraXX

Table 11: List of software used by the large European biobanks *(from ([112] - Kersting et al., 2014)).*

There is also now an evolution towards integration of biobanks by combining their individual resources to reach a higher number of samples and data for specific studies. As an example the GenomEUtwin project is an international collaboration between eight Twin Registries which aim is to identify genetic variants associated with common diseases. They constructed a federated database infrastructure for connecting genotype and phenotype information collected from different sources in Netherlands, Denmark, Norway, Sweden, Finland, Italy, UK and Australia ([119] - Muilu et al., 2007). They have agreed on common standards for all stored genotype and phenotype data, which are maintained in the local operational databases at the data providers sites and then transferred to a data-collecting centre where data is checked and loaded into a common database. The benefit of this approach making data available using database federation is that data management work is distributed to the most experienced personnel and that the data providers can retain control over the data and make it available as needed.

III] 4.5 Perspectives

III] 4.5.1 Future improvements

Flexible and expandable, our SMS model provides opportunities for continuous improvements and for integration of new features, new types of samples or new types of containers in the future. We are currently working on a module that would enable the management of pools of samples. Like for the development of the basic features, each new development is done in close collaboration with the main users with regular working group meetings and tests to get feedbacks.

Also we have just set up a “sharepoint” space dedicated to SAMI to enhance user interactions which are the key for a system well-adapted to the users’ needs. It includes a document library with SOPs and guidelines, a discussion board to share ideas, tips and ask for some additional developments and a news board listing the system updates which go from a new querying criterion to a brand new module or a new type of report.

III] 4.5.2 A SAMI for low- and middle-income countries (LMICs)

The IARC has recently initiated and coordinated the setup of a LMIC Biobank and Cohort Building Network (BCNet) ([120]) in line with IARC's mission to contribute to worldwide cancer research and as an opportunity for LMICs to work together to address the many biobanking challenges. Among them is the local IT management of the samples' collections in formats that allow the information to be shared between centres. We have been consulted to bring our expertise in the field and we conclude that our system would perfectly suit the needs of these LMICs biobanks though most of them would have difficulties affording the extra cost of the Oracle licences.

Hence we are thinking of reproducing the system on completely open source tools based on our experience. Indeed the assets of our SMS do not necessarily rely on the platform on which it has been developed (Oracle) but rather on the way it has been designed and the ways it has been structured and organized for management of data. The database on its own, adaptable to any kind of sample collections and heterogeneous storage structures, is easily reproducible in any other relational database engine and the interface can serve as model for quicker development of another type of web application. We envision as well in the future moving the whole system to more recent and well supported technologies like Microsoft .NET (partially open source freeware software framework developed by Microsoft to make applications easily portable on the internet).

Publication

Voegele C et al., Bioinformatics, 2010

“A sample storage management system for biobanks” 2010

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.

Conclusion

Discussion

Bioinformatics has become an integral part of laboratory science. It promised to improve the manner in which the laboratory information, essential to its correct functioning, is recorded, accessed and archived. Digital technology has transformed how we as individuals and researchers handle and manage information, particularly as the use of smart devices means that computer access becomes even more ubiquitous for researchers.

A common feature to all research laboratories is that different types of information handling are needed from generic and global work to very specific and detailed workflows. For these reasons we have set up three different bioinformatics tools aimed at facilitating researchers' missions and activities. They all have been well adopted at IARC and are now integrated into the daily laboratory life.

ELNs

Traditional way of information management is based on classic laboratory notebook and standard file systems which used to work well for independent researches but were less convenient when needing to search for former data which sometimes could not be reconstructed because of loss of the PLN or impossibility to read or interpret the scientist' writing ([121] - Kühne and Liehr, 2009).

To overcome these issues, ELN has replaced PLN facilitating the search and sharing of information which is properly backed-up and archived. The ELN we set up at IARC is now intensively used by more than 100 persons including non-laboratory scientists recording in the ELN their epidemiological, bioinformatics or biostatistics methods and results.

Though ELNs are becoming an increasingly popular tool for scientific research, most of the advanced tools available remain commercial ones ([57] - Rubacha et al., 2011). The ELN we developed, conversely, is free, open-source and implementable in small laboratories as requiring low resources as well as in larger laboratories as adapted to structures composed of several research groups and permitting fine-grained access controls. In addition, research is more and more

achieved through inter-disciplinary approaches and our ELN favours sharing of information and cross interactions between laboratory scientists, bioinformaticians and epidemiologists as adapted to these various disciplines.

LIMS

LIMS have been central to laboratories for experiments management since years. They are as well very popular tools. They even have a dedicated on-line magazine ([122]) and a LinkedIn group with more than 70,000 members exchanging ideas on the subject. A large number of open-access LIMS are now available to scientific community for many different genomics applications including NGS ([123] - Grimes and Ji, 2014) ([124] - Venco et al., 2014), Sanger sequencing ([125] - Troshin et al., 2011) or microarrays technology ([126] - Cho et al., 2007).

They are essential for tracking large-scale projects but for this, they need to be properly designed to be efficient from both users' and managers' points of views; whence the importance of well-defined specifications and collaborative discussions with future users throughout the development process ([127]).

Since end of 2008, our LIMS has been tracking all GCS platform's major and routinely used workflows; it now also includes the next generation sequencing applications allowing efficient tracking of the large number of samples processed to robustly identify new cancer genes and variants. The LIMS has been highly customized to closely match our genomic laboratory workflows requirements, to be user-friendly and to store the most important and relevant laboratory information. Its design was adapted to comply with those NGS complex workflows including many sub- and optional steps and to be flexible to allow addition of supplementary tasks as well as upgrading of the actual ones.

Sample management systems for biobanks

Following advances in sampling, storage capacities, and bio-analyses technologies, biobanks have been growing quickly in the past years requiring robust IT tools to store the samples data. Since mid-2010 the SAMI has been progressively providing rigorous follow-up and management of the millions of samples of IARC's

biobank involved in various and heterogeneous studies, including basic epidemiological data, storage location information and movements from use for biological analyses to shipments to collaborators. The description of SAMI's database that has been designed to cope with any type of sample and any type of storage infrastructures allowing movements of these infrastructures has been published to allow other biobanks to take it as example for managing their own biospecimens and the associated data ([111] - Voegele et al., 2010).

In the coming years, IARC will favour new opportunities to establish unique cohorts through collaboration with different parts of the world including rich associated dataset to be maintained in SAMI and other appropriate epidemiological databases. IARC will further promote sample collection and sharing of both the biospecimens and their standardized data facilitating wider external access to IARC biobank. Emphasis will be placed on availability of well annotated samples which is a foundation to studying the causes of cancer – as well as data confidentiality and security.

IARC will also prioritize duplicate storage of LMICs collections and advice on best biobanking practises in the LMICS through the BCNet. In consequence, SAMI database will keep on growing and will increase visibility of IARC biobank providing a homogenous structure for data, facilitating querying and thus assisting IARC biobank to achieve its goals of sharing.

In addition, we will keep on working in close collaboration with the BCNet members to help LMICs setting up an adapted version of SAMI for the management of their local biospecimen collections and provide possibilities for sharing their resources as well with the scientific community as they can be of great value for worldwide studies and potential discoveries.

Sharing of biobank resources

Biobanks are an important resource in medical research. One aim of their development is indeed to maximize the value of the repositories by sharing them worldwide for developing ideas to advance research into a variety of health issues. As a consequence and to encourage greater coordination and promote harmonization, consortiums of biobanks have been set-up such as the UK National Cancer Research Institute (NCRI) Confederation of Cancer Biobanks (CCB) ([128]) or the BCNet

([120]). They enable to raise awareness of the samples collections within scientific community.

In order to achieve sharing of biobank's samples following ethical best practise, access policy and rules have been set up at IARC. The exact procedure on how to ask for human biological material is explained on the biobank website where a list of available sample collections is provided together with the name of the person responsible for the collection ([129]). This list is extracted from the SAMI database including only relevant sample nature and geographical origin information, which are essential for deciding of the appropriateness for a particular study. The requestor should fill and sent to the biobank a project application form and a partner profile form that will be reviewed by the principal investigator of the collection and the biobank steering committee. If the request is approved, an application should be made to the ethics committee who will decide of the sample transfer agreement and shipment including associated quality data extracted from SAMI database. Similar procedures are in place in many other large biobanks like the UK Biobank which authorized in March 2012 researchers from all over the world - whether they work in the public or private sector, for academia, industry or a charity - to apply to use its resources and associated anonymous data under the condition that the research is health-related and in the public interest ([130] - BBC, 2012). The EuroBioBank also provides on its website a catalogue of human DNA, cell and tissue samples available to the scientific community conducting research on rare diseases. It includes materials from 21 biobanks hosted by 9 different countries in Europe as well as Israel and Canada (e.g. Généthon, Myobank, Galliera Genetic Bank) ([131]). Finally, an online catalog has been established by the BBMRI for the collection and presentation of data describing the majority of European biobanks. "By March 2011 the catalog included data from 63 population-based and 219 clinical biobanks located in 27 countries, together representing more than 20 million samples" ([103] - Wichmann et al., 2011). It includes a search function allowing selection of multiple combined criteria (e.g. type of material, disease group, key publications…)

More broadly the keys for efficient sharing of biobanks materials are standardisation and harmonization of procedures of collection and annotation, common nomenclatures ([132] - Fransson et al., 2015) and compliance with ethics rules concerning sensitive "patient" data.

Long-term management and surveillance of the tools

Once tools are developed, implemented and users satisfied, they should still be followed-up regarding different "issues" in order to bring over time the maximum of their value:

- Correct usage should be regularly checked. Best practises and responsibilities should be well defined. If data is not correctly recorded or updated, it may not be a problem of the tool not being adapted but of the users not assuring their duties
- The data quality should be as well regularly checked. Some triggers and procedures were set-up to automatically verify the consistence of data recorded in the databases. For example the number of samples in one container should not exceed the capacity of the container. It is also possible to perform manual checks by picking randomly one piece of information and verify it.
- Data persistence should be ensured despite technology obsolescence. For our tools data is stored in long-lasting formats within relational databases.
- Improvements should be possible and provided upon request: the developers should stay in close interaction with the users to be able to answer new needs or requests for modifications. This has been done and is on-going regarding the three applications developed.

Importance of data security

The three tools described in this thesis all aimed at storing different types of scientific data which are sensitive and therefore needs to be safe and secured. This means defending data from unauthorized access, use, disclosure, disruption, modification, perusal, recording or destruction. This is ensured by a number of measures taken including access through logins and passwords with specific read and write permissions, usage guidelines and regular backups.

We can distinguish two types of concerns regarding data access. First, research work is the intellectual property of the researcher or its institute and should therefore be protected whether aimed at publication or not. ELNs by storing in

electronic format research investigations and/or results facilitate the sharing of information but only with appropriate collaborators. It is therefore important to be able to define very precisely the access permissions of every type of data stored, which is a key feature of our ELN but also our LIMS and SMS.

Whether ELN as electronic recording device, has sufficient value to protect intellectual property has been debated since the creation of the first ELN. The acceptance of electronic records with equal weight of other forms of evidence in late 2006 by US law has answered partially this concern ([133] - Elliott, 2011). However, proper attention must be paid to organizational policies and practises to minimize litigations risks. Appropriate user- and time- stamping as well as appropriate long-term archiving for future access of research data are as well required and not yet correctly set-up in all ELNs. We took particular care of implementing these crucial functionalities in the ELN for securing researchers' investigations.

The second concern is about privacy and ethical challenges regarding data associated to human biospecimens. This concern has increased with the opening of the biobanks to whole research community and led to development of guidelines and even legislations for research on human biological materials ([99] - Haga and Beskow, 2008). Indeed a study of public perception about biobanking spanning all Canadian provinces was conducted through online survey. Most people expressed willingness for their data to be shared with the entire scientific community and not only their country's institutions and for being well-informed on the research projects conducted but their main concern was about possible access and misuse by insurers, the government and other third parties ([134] - Joly et al., 2015). This opinion seems widely shared: the large majority (72,7%) of 285 participants in genomic studies conducted between 2008 and 2009 at Baylor College of Medicine in Houston, US, consider that they are more benefits (enabling research and medical knowledge progresses for themselves and others) than risks (private data access by unauthorized persons and finding out unwanted information about themselves) in data sharing ([135] - Oliver et al., 2012). Access to complete individual information should be restricted to relevant investigators.

In this respect, we set up different levels of access in our biobank management system and divided data into compartments. In addition samples are completely anonymized to users and coded for keeping the possibility of linkage to other data

repositories containing more personal data resulting from epidemiological questionnaires. This is to prevent identifiability of the samples and thus ensure protection of individual identity. To follow international biobanking best practises ([136] - Vaught et al., 2010), consent forms and IARC ethics committee approvals associated with biospecimens collections are also stored within our biobank management system.

Integration of the tools: all together

There are nowadays many software systems available to manage laboratory data such as ELN, LIMS and sample management solutions but there is also a need for hybrids attempting to coordinate all three. Indeed, while it is important to have various up-to-day tools well-adapted and dedicated to specific uses and purposes, an integrated solution is highly desirable to get the various and complementary tools communicating together to avoid duplication of work at record keeping. The need of building collaborative networks is enhanced by the fact that large amount of data is available in a variety of formats from heterogeneous sources which leads to need for integration to access, assemble, combine all these data and thus enhanced their value.

As information exploitation is crucially dependant on the effective integration of data and tools, we not only set up the tools but worked on linking them to go towards a 'connected' laboratory solution. It is essential for this linking to know perfectly well the design and content of the different databases to be able to define the information which it is relevant to cross-reference.

ELN with the LIMS and with SAMI

The frontier between ELN and LIMS could sometimes be blurred especially in the industrial domain where ELNs are often small LIMS which constraint users to fill pre-defined and very rigid forms. However in public research, ELNs are generally aimed at handling unstructured or partially structured data whereas LIMS manage essentially well structure data ([90] - Gibbon, 1996). The way we developed and implemented these tools followed this concept that is our ELN dealing with meta-data of experiments and tests including preparation instructions and specifications while our LIMS dealing with tracking of batches of samples through workflows as well as management of consumables stocks and/or costs.

In GCS laboratory, the main criteria on which we decided what to store in the ELN or the LIMS is the possible automation of the workflows and the throughput. Tests and small experiments are stored in the ELN whereas the automated workflows are just mentioned in the ELN with a reference to the LIMS experiment number while

the detail of the laboratory work required by these automated workflows and high-throughput workflow is stored in the LIMS. In terms of content, ELN stores unstructured text and files whereas the LIMS stores well-defined objects in database tables.

Both have shown to be essential and very complementary in terms of aim, use and type of data stored ([137] - Bolton, 2009). Indeed, the benefits of ELNs increase when they are effectively integrated with other laboratory informatics tools such as laboratory information management systems or other scientific data management systems ([138] - Machina and Wild, 2013). This will result in less time spent on manual data entry, data aggregation and manipulation as well as a gain of visibility into research. Information sharing between these two systems is a potentially added value.

We therefore undertook pilot studies that connected the ELN and the LIMS by implementing in several forms of the LIMS a search function that queries directly the ELN. The challenge was to define the relevant type of information useful to provide from one system to the others and to match the security models of the two systems. We thus selected the search variables such as “Project name”, Sample ID or Sample Name, Barcode, Sample origin (country). This query within the LIMS interface returns all the ELN pages containing the specified criteria (**Figure 44**).

Using this same linkage approach, we enabled querying the ELN from SAMI based on sample name or container name. Theses connections are providing convenient way to access and regroup information on projects or samples for which tests experiments may have been recorded in different ELNs by different laboratory assistants (**Figure 45**).

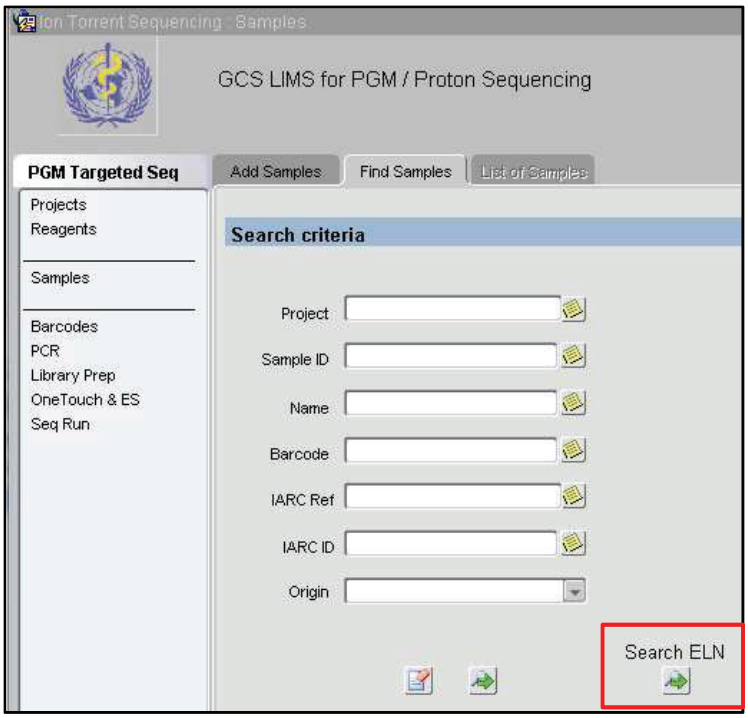


Figure 44: LIMS screen enabling search in ELN. *Within the LIMS NGS workflow we have added the possibility to query the ELN for a specific project or sample (through ID, name, barcode or origin).*

The LIMS with the sample management system

The LIMS stores analytical information on samples that went through some laboratory studies whereas SAMI stores basic epidemiological information as well as position and movements of all the samples hosted in IARC biobank facilities. So it may happen that some subsets of samples have complementary information stored in the two different systems. It was therefore important to make a connection between them.

As both tools rely on Oracle databases, we created an Oracle "public database link". This type of database object enables to query another database instance that can be physically on the same or on a distant server, the latter option being the one we chose. It is like a local pointer to access the remote database objects. We used the LIMS manager account as owner of this link to be able to return all the data stored in the LIMS and we restricted the access to authorized personnel within SAMI interface.

For non-Oracle databases such as MySQL databases, direct linkage is not possible but there are workarounds like virtually importing tables from the distant database through the FEDERATED storage engine ([139]).

In the LIMS we developed a module to manage specifically the location and quantities of samples that undergo laboratory analyses. Three specific tables enable to store in particular samples' initial and current positions on plates, initial and current concentrations and volumes as well as the type and date of movements, type including various kind of internal experiments as well as shipment to collaborators (**Appendix 5**). This information is important for the biobank management and to avoid duplication of the records in both the LIMS and SAMI, we developed in parallel within the SAMI's samples form a specific function to get information stored in the LIMS on the samples that underwent laboratory investigations. The query button is only visible by the group to whom the samples belong. Its activation launches the display of samples availability information on the fly in an Excel or CSV report, using the Oracle "public database link".

These connections are possible as the two databases have common consistent elements such as projects or samples names (**Figure 45**). These common variables are also searchable in the ELN and therefore retrievable across the three systems. One of the interesting potential applications of these connections is to be able to use the information stored in one system to improve information stored in another system and thus global knowledge. Especially, scientific experimental results stored in the ELN or the LIMS may help to diagnose samples problems enabling then to flag in SAMI these samples as being of less good quality. For example, it happened that all the samples from one specific origin centre had systematically bad results at Taqman genotyping meaning that they are of less good quality. Statistical analyses of assays performance or any other laboratory experiment success rates can establish associations with samples issues and therefore protocols of sample preparation in some centres. This type of information is essential for not selecting these same samples for some other sensitive studies.

Global integration

Scientific data management systems' (SDMS) aim is to provide users with single access point to all data by integrating with existing informatics systems such as ELN or LIMS ([140] - Machina and Wild, 2013). They typically act as wrapper for other data systems (*i.e.* system calling other systems) to facilitate knowledge management. Indeed data integration consists of wrapping data sources (getting data from somewhere and translating it into a common integrated format ([141] - Lacroix, 2002). In the age of scientific computing, the availability of data from varied and heterogeneous sources, coupled with the desire to build knowledge bases and collaborative networks, has driven this need for global integration at the fundamental data level. Due to the volume and heterogeneity of data, this can be a daunting task.

Open-source and commercial data management solutions are often tailored for a particular scientific application. The problems with these solutions is that they do not work well together due to proprietary data formats, and they lack interconnectivity to other systems short of reporting entire result sets ([142] - Hobbie et al., 2012). That is why the Oregon State University Superfund Research Centre developed a system that integrates environmental monitoring data with analytical chemistry data and downstream bioinformatics and statistics to enable complete "source-to-outcome" data modelling and information management. It includes commercial software for operational laboratory management (X-LIMS) and sample management (FreezerPro) in addition to open-source custom-built software for bioinformatics and experimental data management. To achieve this, they developed internally APIs and sets of SQL procedures to connect the different systems to the central LIMS database. This integration of data systems across all research projects at remote collaborators institutions and support cores appeared to improve research and collaboration. ([142] - Hobbie et al., 2012).

IARC already started a few years ago to set up such global management systems with the implementation of "Microsoft Active Directory" service (MS AD) which is a software system that stores, organizes and provides access to all computers within a Windows domain type network. Its controller authenticates users and manages the permissions assigning and enforcing security policies. The system enables also easy installation or update of software.

More recently IARC IT group has deployed "MS SharePoint" to facilitate communication, collaboration, sharing and search of information within the Agency. It is a web application framework and platform that integrates intranet portals, enterprise content and document management, internet sites, social networks promoting interactions between staff and rapid feedbacks from users while developing tools for them. It provides central management, governance and security controls.

In the future, our objective will be to integrate as well the ELN, the LIMS and SAMI within this global agency IT package and provide this single access point to whole research data whether laboratory of bioinformatics associated (**Figure 45**).

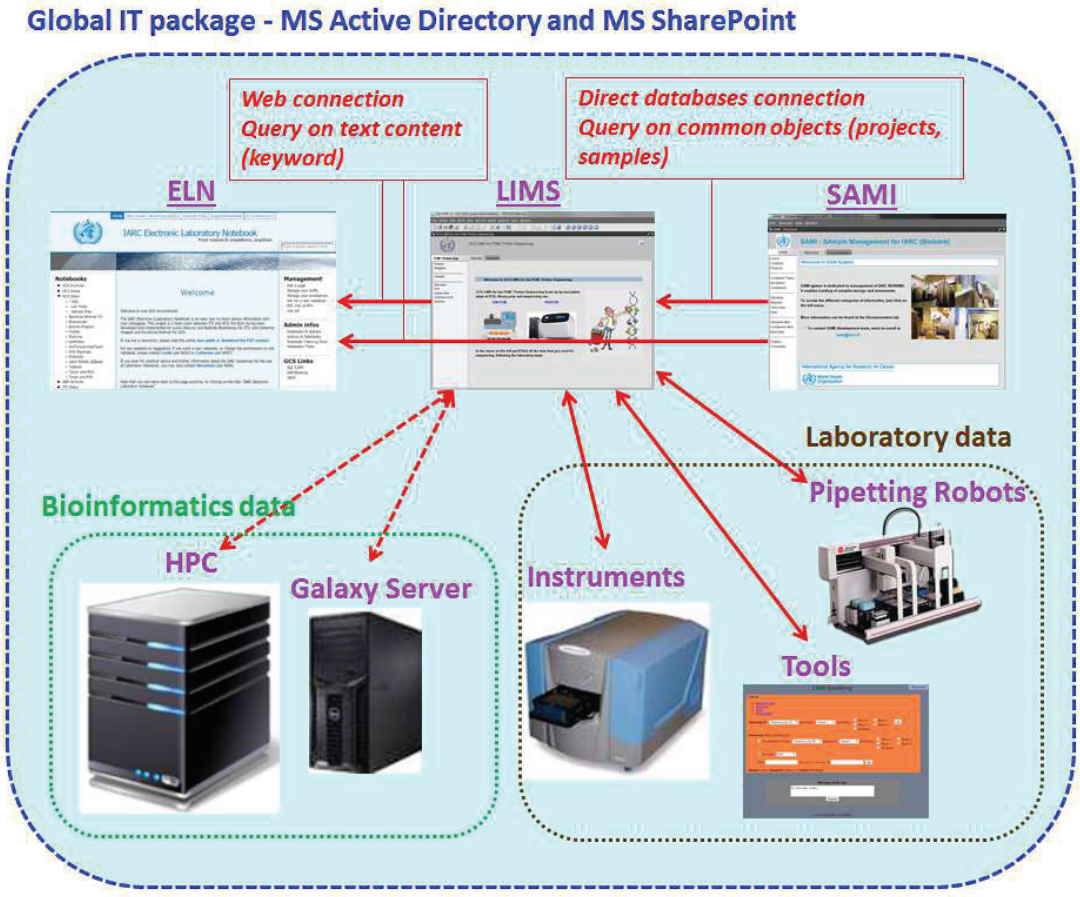


Figure 45: Global IT tools integration with existing connections in plain arrows and future connections in dashed arrows.

Perspectives

The common goal of all these electronic IT tools is to increase the quality of documentation of laboratory work while reducing the time spent on this documenting. They also aim to facilitate the searching and retrieval of information for the researchers. As effective use of time is a critical element in daily work activities, finding the appropriate balance between documenting too poorly and too much is essential to avoid being overwhelmed by the flow of information and work, taking into account the constraints on the laboratory assistants and the benefits.

While information management systems can be extremely well designed to provide the best possible tools including automatic checks of data and proper guidance, their efficiency is completely dependent on the users' commitment to follow the instructions. It is therefore also important to provide clear and well documented user guide specifying what type of information should be recorded in which system to avoid repetition of the work limiting therefore the burden on the laboratory assistants and waste of time. These systems can then warrant consistent, complete and accurate information about many aspects of the laboratory work.

However improvements are always possible and we will keep on working on this objective in the coming years whether for the ELN, the LIMS or SAMI.

LIMS and NGS bioinformatics challenges

In my opinion, one of the largest coming challenges within genomics laboratories will be to set-up "Bioinformatics Information Management Systems". Indeed the volume of data generated among others by NGS applications is continuously increasing ([48] - Baker, 2010) requiring powerful tools to keep the data available while reducing the size of the records and archive them with associated relevant information details. Documenting and tracking of the origin and features of the data will be crucial for proper exploitation, interpretation and secure access ([143]).

As an example, large and ambitious sequencing projects like the one announced in August 2014 by the UK government : the sequencing of 100,000

whole genomes ([143]) would be difficult without sophisticated multi-stage processing, strict tracking of the experiments and reporting of results in appropriate and well-designed information management tools. The UK has planned to provide the whole scientific community with all the data to enable further and cross investigations. The success of those types of project is totally dependent on novel strategies for storage, quality control and analyses but also on computational tools and high performance computing requiring a comprehensive system approach ([144] - Hunkapiller and Hood, 1991).

Recording of data annotations with versions of bioinformatics software and packages used for the analyses is also important for reproducibility issue ([145] - Ioannidis et al., 2009;[146] - Malone et al., 2014). This is the reason why scientific journals are now asking for providing the underlying data and details on tools and parameters used for data analysis: to ensure reproducibility ([147] - Peng, 2011). Details of computational experiments should be recorded with the same level of care and scrupulousness as those of 'wet' laboratory investigations ([148] - Nekrutenko and Taylor, 2012).

We envision therefore to store within our "LIMS for Bioinformatics" both:

- the origin, features and location of each raw data and analysed data files with appropriate access permissions. It would therefore need a connection with our High Performance Cluster and IT storage infrastructures
- each step of the bioinformatics analyses pipelines for our wide range of NGS applications. We are therefore thinking of including as well a connection with the Galaxy server we are just setting up. Galaxy is a scientific workflow platform specifically designed for non-bioinformaticians for analysis of NGS data but also other biological data. It acts as a wrapper for a large number of analysis tools and enables to store bioinformatics analyses pipelines for reuse and sharing between Galaxy users ([149];[150] - Goecks et al., 2010).

A few open source bioinformatics managements systems have recently been published for NGS analyses : Orione ([151] - Cuccuru et al., 2014), Galaxy LIMS ([152] - Scholtalbers et al., 2013) or NG6 ([153] - Mariette et al., 2012). Most of these tools are targeted to genomic facilities and clinical laboratories having predefined and fixed workflows such as SMITH developed at the Genomic Unit of the Italian Institute of Technology ([124] - Venco et al., 2014). The latter includes basic

laboratory information as well as analyses workflows with connection to their own HPC and Galaxy server. We would like to extend our LIMS by developing a similar system but providing more NGS applications and more flexibility in definition of various bioinformatics pipelines which are in constant evolution.

Sharing knowledge beyond the group, beyond the institute

Research advances are highly dependent on sharing, sharing of ideas, sharing of knowledge, sharing of biobanks, sharing of data, sharing of tools. Today making research data openly accessible to the scientific community is one of the main priorities for the global research system. In fact, there is wide consensus that data sharing may help scientific progress allowing a better exploitation of data and an optimized use of resources in a climate of scientific openness and transparency ([154] - Fischer and Zigmond, 2010).

"In human health, the major needs are driven by the realization that for precision medicine and similar efforts to be most effective, genomes and related 'omics' data need to be shared and compared in huge numbers. If we do not commit as a scientific community to sharing now, we run the risk of establishing thousands of isolated, private data collections, each too underpowered to allow subtle signals to be extracted" ([27] - Stephens et al., 2015).

From a bioinformatics point of view, collaborative open source programming is a key for the future of this discipline and its main goal: answering rapidly biological research questions. Our philosophy follows this concept of open sharing. We took indeed special care of making all our tools freely available to the scientific community through our publications in Bioinformatics and our website providing architecture, description of the design processes and source-codes for easy installations or even further improvements. Other scientists would therefore be able to implement these tools in their own research centres and further customize them as needed.

Especially for the ELN we have been contacted by various research centres worldwide: a microbiology laboratory in London, another one in Virginia (US), a R&D cell diagnostics laboratory in San Francisco (US), a cell imaging laboratory in Alberta (Canada), an hospital in Sao Paulo (Brazil), the Scottish National Blood

Transfusion Service, an institute of molecular toxicology and pharmacology in Munich (Germany), an immunology laboratory in India and also scientists from French public research (CNRS) and French scientific police. They all showed interest in installing a local version of the ELN.

Sharing of knowledge: opening access to information

Open access to information does not make the unanimity. Publishers of scientific journals have fought against a large group of scientists including Richard Roberts (Nobel laureate for his work on nucleic acids) and Harold Varmus (former director of US National Institute of Health (NIH)) who wished to make PubMed Central (PMC- the NIH- sponsored online repository of scientific literature) posting freely full texts of any article ever published and not only abstracts freely available. They proposed that the scientists pay for being published and not the readers for having access to the publications and also that the cost of publication should be included in grants applications. PMC represents one more example of a growing battle over information freedom ([155] - Thompson, 2001).

Nevertheless things have been moving towards opening access to information over the last few years. PLOS One ([156]) has recently announced changes in their publishing policies by supporting submissions of open source software following their aim to promote openness in research. Nature Publishing Group has also launched in 2014 “Scientific Data” a new online open-access peer-reviewed publication dedicated to the description of scientifically-valuable data sets ([157]). Finally, the Open Knowledge Foundation – “a worldwide non-profit network of people passionate about openness” aims at promoting the sharing of knowledge using advocacy, technology and training to unlock information ([158]).

Following this tendency, IARC has set up on 1st of January 2015 a new open access policy for its publications. Communication and dissemination of research is a key objective of the IARC and from my point of view should be of all researchers. One of the pillars of IARC's mission is to be an authoritative and unbiased "global reference for cancer information." But beyond ensuring quality and integrity of its publications, IARC as a publicly-funded international agency recognizes its obligation to share knowledge broadly and openly, in ways that are free of cost barriers and use

restrictions. “Agency authors are strongly encouraged to publish under a license such as the Creative Commons licenses ([159]) that accord with the principle of Open Access being free of most copyright barriers.” IARC became also in July 2014 a participating publisher in HINARI, a programme set up by WHO together with major publishers to enable institutions in LMICs to gain access to one of the world’s largest collections of biomedical and health literature (up to 13,000 journals, 29,000 e-books and 70 other information resources).

Sharing of data: opening science and collaborations for new discoveries

Open access and open sharing come together in “Open Science” which Nielsen defined in a blog devoted to the relationships between science and Web 2.0: “Open science is the idea that scientific knowledge of all kinds should be openly shared as early as in practical in the discovery process”([160]). Already in 2002 an article from Nicholas Thompson in Washington Monthly explored this “new” trend: biology in open source wondering whether biologists who share data freely out-innovate corporate researchers ([155] - Thompson, 2001). Putting in common findings as quickly and widely as possible with free circulation of data increases the chance that other scientists can see, improve or use them in ways the original discoverer did not foresee. It promises to be a research accelerator.

One good example of successful sharing is coming from Alfred Gilman -the 1994 Nobel Prize in medicine who founded the Alliance for Cellular Signaling made up of seven core labs serving as central coordinators and nearly 500 scientists worldwide who have lined up to design descriptive Web pages for molecules key to the inner workings of cells, following the development model of Linux operating system (source code is available for download, use and improvement by the entire developers community). They achieved work that would not have been possible for just one laboratory.

Another good example is the development by Sage Bionetwork - an association of promotion of scientific open source - of Synapse: an informatics platform dedicated to supporting the large-scale pooling of data, knowledge and expertise across institutional boundaries to solve some of the most challenge problems

in biomedical research ([161]). Indeed over 200 scientists collaborated through Synapse as part of the TCGA Pan Cancer Analysis Working Group tracking provenance and metadata, stable digital object identifiers for data referencing and flexible methods for data access ([162] - Weinstein et al., 2013).

The world wide web offers an easy way with which one can now transfer information as an interactive interface to scientific data facilitating exchange of information and supporting visualization and distribution with publicly accessible www-based data systems in various scientific fields (i.e. "COHOWeb" hosted by the National Aeronautics and Space Administration (NASA) provides access to heliospheric magnetic field, plasma and spacecraft position data from 14 spacecrafts ([163]) or "Genbank" lists nucleotide sequences and their protein translations from more than 100,000 distinct organisms ([37])).

However sharing is sometimes complicated especially when it concerns big data because of the limited transfer capacities of the network. Moving data from research centers to clouds are long. The solution yet is still to send hard drives but new technologies for transfer are being explored ([49] - Marx, 2013).

Sharing of IT capacities

With the rapid growth of data over the last few years, there was a growing need for large-scale computational resources, e-infrastructures and IT tools ([164] - Duarte et al., 2015). We have thus seen the development of projects aimed at sharing IT resources for scientific research. Like for data, the sharing of computers processors from 75,000 volunteers Internet users in 2001 for the "Décrypthon" - a project launched by the "Association Française contre les Myopathies" and IBM - have led to great achievements: the first cartography of the proteome was generated in less than 2 months instead of 1170 years with only one computer ([165]). Following this scientific success, the World Community Grid has been set up to create the world largest public computing grid to tackle scientific research projects. It uses the idle time of Internet-connected computers to perform research calculations from now almost 700,000 contributors ([166]).

Similarly, the European Grid Infrastructure is a publicly funded e-infrastructure put together to give scientists access to more than 530,000 logical CPUs, 200 PB of disk capacity and 300 PB of tape storage to drive research and innovation in Europe. Resources are provided by about 350 resource centres which are distributed across 56 countries in Europe, the Asia-Pacific region, Canada and Latin America ([167]). Over 21,000 researchers from 15 different disciplines carry out 1.4 million computing jobs a day for their intensive data analysis across over 15 research disciplines ([164] - Duarte et al., 2015).

As another example, the Broad Institute of MIT and Harvard (biomedical and genomic research centre located in Cambridge, US) has recently announced a partnership with Google Genomics aimed at providing cloud services for scientists combined with a toolkit that can be used to analyse the data ([168] - Weisman, 2015). The concept is to put part of the world's genomic data on Google's servers from where scientists from all over the world can collaborate and explore the data.

The NIH also launched the Big Data to Knowledge (BD2K) initiative in 2012, which focused on managing large data sets in biomedicine with elements such as data standards and handling, informatics training and software sharing. ([169] - Paten et al., 2015). Similarly, the EBI campus houses the hub - the technical command centre - of ELIXIR, a project to help life scientists across Europe safeguard and share their massive amount of data generated every day by publicly funded research thanks to pan-European insfrastructures.

In addition, there are initiatives for specifically sharing of IT tools whether for bioinformatics analyses such as the web based platform Galaxy that enables to execute and share bioinformatics pipelines ([149];[150] - Goecks et al., 2010) or more general like Taverna Workflow Management System: a suite of tool used to design, execute and share scientific workflows of web services and aid *in silico* experimentation ([170];[171] - Wolstencroft et al., 2013).

Finally, the Center for Open Science – a non-profit technology company fostering openness, integrity and reproducibility of scientific research – provides freely the Open Science Framework which is “part network of research materials, part version control system, and part collaboration software” ([172]). Its aim is to “support the scientist’s workflow and help increase the alignment between scientific values and scientific practises”. It could be, in addition to the publication in scientific journals, a

place where to deposit the latest versions and documentations of the tools we developed in order to share them with the scientific community.

Balancing the benefits of sharing, there are also barriers to be overcome, such as the inherent time and economic costs, possible data misuse, ethical issues and conflicts of interest with patenting discoveries ([173] - Anagnostou et al., 2015). The danger of sharing too much is also to decrease the quality of information available. Social networks are giving opportunities to researcher to get a public visibility and sharing freely and rapidly opinions and results ([174] - Rinaldi, 2014) with the risk of not being able to distinguish what is true from what is not in this flow of information available.

That is why it is equally important to encourage the scientific community to control, retest and peer-review this information to maintain high quality standards rather than to just simply encourage sharing. Researchers need to take care of making the shared information meaningful by documenting and disseminating work in ways that they can be reproduced and reused.

As convinced that open up one's work can bring a lot to research by letting new discoveries serve as bridges rather than endings, "open science" is the aim I pursued during my thesis work and I will keep on pursuing in the future while developing tools to share scientific information within research centres in appropriate, efficient and adapted manner.

List of first author publications (3)

Voegele C., Bouchereau B., Robinot N., McKay J., Damiecki P., Alteyrac L. "**A universal open-source Electronic Laboratory Notebook**". *Bioinformatics.* 2013 Jul ; 29(13), 1710-1712.

Voegele C., Alteyrac L., Caboux E., Smans M., Lesueur F., Le Calvez-Kelm F., Hainaut P. "**A sample storage management system for biobanks**". *Bioinformatics.* 2010 nov ; 26(21), 2798-2800.

Voegele C., Tavtigian S.V., De S.D., Cuber S., Thomas A., Le Calvez-Kelm F. "**A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening**". *Bioinformatics.* 2007 Sep ; 23(18), 2504-2506.

List of co-authored publications (12)

NGS exomes analyses related publications:

Castells X, Karanović S, Ardin M, Tomić K, Xylinas E, Durand G, Villar S, Forey N, Le Calvez-Kelm F,Voegele C, Karlović K, Mišić M, Dittrich D, Dolgalev I, McKay JD, Shariat SF, Sidorenko VS, Fernandes A, Heguy A, Dickman KG, Olivier M, Grollman AP, Jelaković B, Zavadil J. **"Low-coverage exome sequencing screen in formalin-fixed paraffin-embedded tumors reveals evidence of exposure to carcinogenic aristolochic acid".** *Cancer Epidemiol Biomarkers Prev.* 2015 Sep 17. Epub.

Vaca-Paniagua F., Alvarez-Gomez R., Maldonado-Martinez H., Perez-Plasencia C., Fragoso-Ontiveros V., Lasa-Gonsebatt F., Herrera L., Cantu D., Bargallo-Rocha E., Mohar A., Durand G., Forey N., Voegele C., Vallée M., Le Calvez-Kelm F., McKay J., Ardin M., Villar S., Zavadil J., Olivier M. "**Revealing the molecular portrait of triple negative breast tumors in an understudied population through omics analysis of formalin-fixed and paraffin-embedded tissues**". *PLoS.One.* 2015 May 11; 10(5).

Kim Y.H., Ohta T., Oh J.E., Le Calvez-Kelm F., McKay J., **Voegele C**., Durand G., Mittelbronn M., Kleihues P., Paulus W., Ohgaki H. "**TP53, MSH4, and LATS1 germline mutations in a family with clustering of nervous system tumors**". *Am.J.Pathol.* 2014 Sep ; 184(9), 2374-2381.

NGS RNA seq analyses related publications:

Accardi R, Fathallah I, Gruffat H, Mariggiò G, Le Calvez-Kelm F, **Voegele C**, Bartosch B, Hernandez-Vargas H, McKay J, Sylla BS, Manet E, Tommasino M. "**Epstein - Barr virus transforming protein LMP-1 alters B cells gene expression by promoting accumulation of the oncoprotein ΔNp73α**" *PLoS Pathog*. 2013 Mar ; 9(3).

LIMS and mutation screening analyses related publications:

Park D.J., Tao K., Le Calvez-Kelm F., Nguyen-Dumont T., Robinot N., Hammet F., Odefrey F., Tsimiklis H., Teo Z.L., Thingholm L.B., Young E.L., **Voegele C**., Lonie A., Pope B.J., Roane T.C., Bell R., Hu H., Shankaracharya, Huff C.D., Ellis J., Li J., Makunin I.V., John E.M., Andrulis I.L., Terry M.B., Daly M., Buys S.S., Snyder C., Lynch H.T., Devilee P., Giles G.G., Hopper J.L., Feng B.J., Lesueur F., Tavtigian S.V., Southey M.C., Goldgar D.E. "**Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers**". *Cancer Discov.* 2014 Jul ; 4(7), 804-815.

Damiola F., Pertesi M., Oliver J., Le Calvez-Kelm F., **Voegele C**., Young E.L., Robinot N., Forey N., Durand G., Vallee M.P., Tao K., Roane T.C., Williams G.J., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Goldgar D.E., Lesueur F., Tavtigian S.V. "**Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study**". *Breast Cancer Res.* 2014 Jun ; 16(3), R58.

Le Calvez-Kelm F., Oliver J., Damiola F., Forey N., Robinot N., Durand G., **Voegele C**., Vallee M.P., Byrnes G., Registry B.C., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Tavtigian S.V., Lesueur F. "**RAD51 and breast cancer susceptibility: no evidence for rare variant association in the Breast Cancer Family Registry study**". *PLoS.One.* 2012 ; 7(12), e52374.

Ahmad J., Le Calvez-Kelm F., Daud S., **Voegele C**., Vallee M., Ahmad A., Kakar N., McKay J.D., Gaborieau V., Leone M., Sinilnikova O., Sangrajrang S., Tavtigian S.V., Lesueur F. "**Detection of BRCA1/2 mutations in breast cancer patients from Thailand and Pakistan**". *Clin.Genet.* 2012 Dec ; 82(6), 594-598.

Park D.J., Lesueur F., Nguyen-Dumont T., Pertesi M., Odefrey F., Hammet F., Neuhausen S.L., John E.M., Andrulis I.L., Terry M.B., Daly M., Buys S., Le Calvez-Kelm F., Lonie A., Pope B.J., Tsimiklis H., **Voegele C**., Hilbers F.M., Hoogerbrugge N., Barroso A., Osorio A., Giles G.G., Devilee P., Benitez J., Hopper J.L., Tavtigian S.V., Goldgar D.E., Southey M.C. "**Rare mutations in XRCC2 increase the risk of breast cancer**". *Am.J.Hum.Genet.* 2012 Apr; 90(4), 734-739.

Le Calvez-Kelm F., Lesueur F., Damiola F., Vallee M., **Voegele C**., Babikyan D., Durand G., Forey N., McKay-Chopin S., Robinot N., Nguyen-Dumont T., Thomas A., Byrnes G.B., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Tavtigian S.V. "**Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study**". *Breast Cancer Res.* 2011 Jan ; 13(1), R6.

Tavtigian S.V., Oefner P.J., Babikyan D., Hartmann A., Healey S., Le Calvez-Kelm F., Lesueur F., Byrnes G.B., Chuang S.C., Forey N., Feuchtinger C., Gioia L., Hall J., Hashibe M., Herte B., McKay-Chopin S., Thomas A., Vallee M.P., **Voegele C**., Webb P.M., Whiteman D.C., Sangrajrang S., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Chenevix-Trench G. "**Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer**". *Am.J.Hum.Genet.* 2009 Oct ; 85(4), 427-446.

Garritano S, Gemignani F, **Voegele C**, Nguyen-Dumont T, Le Calvez-Kelm F, De Silva D, Lesueur F, Landi S, Tavtigian SV. "**Determining the effectiveness of High Resolution Melting analysis for SNP genotyping and mutation scanning at the TP53 locus**" *BMC Genet.* 2009 Feb 17 ; 10:5.

List of figures, tables and appendices

Bibliography – References

Bibliography - References

1. Wang Y., McKay J.D., Rafnar T., Wang Z., Timofeeva M.N., Broderick P., Zong X., Laplana M., Wei Y., Han Y., Lloyd A., Delahaye-Sourdeix M., Chubb D., Gaborieau V., Wheeler W., Chatterjee N., Thorleifsson G., Sulem P., Liu G., Kaaks R., Henrion M., Kinnersley B., Vallee M., LeCalvez-Kelm F., Stevens V.L., Gapstur S.M., Chen W.V., Zaridze D., Szeszenia-Dabrowska N., Lissowska J., Rudnai P., Fabianova E., Mates D., Bencko V., Foretova L., Janout V., Krokan H.E., Gabrielsen M.E., Skorpen F., Vatten L., Njolstad I., Chen C., Goodman G., Benhamou S., Vooder T., Valk K., Nelis M., Metspalu A., Lener M., Lubinski J., Johansson M., Vineis P., Agudo A., Clavel-Chapelon F., Bueno-de-Mesquita H.B., Trichopoulos D., Khaw K.T., Johansson M., Weiderpass E., Tjonneland A., Riboli E., Lathrop M., Scelo G., Albanes D., Caporaso N.E., Ye Y., Gu J., Wu X., Spitz M.R., Dienemann H., Rosenberger A., Su L., Matakidou A., Eisen T., Stefansson K., Risch A., Chanock S.J., Christiani D.C., Hung R.J., Brennan P., Landi M.T., Houlston R.S., Amos C.I., 2014. "**Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer**". *Nat.Genet.* 46(7), 736-741.

2. Park D.J., Tao K., Le Calvez-Kelm F., Nguyen-Dumont T., Robinot N., Hammet F., Odefrey F., Tsimiklis H., Teo Z.L., Thingholm L.B., Young E.L., Voegele C., Lonie A., Pope B.J., Roane T.C., Bell R., Hu H., Shankaracharya, Huff C.D., Ellis J., Li J., Makunin I.V., John E.M., Andrulis I.L., Terry M.B., Daly M., Buys S.S., Snyder C., Lynch H.T., Devilee P., Giles G.G., Hopper J.L., Feng B.J., Lesueur F., Tavtigian S.V., Southey M.C., Goldgar D.E., 2014. "**Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers**". *Cancer Discov.* 4(7), 804-815.

3. Delahaye-Sourdeix M., Anantharaman D., Timofeeva M.N., Gaborieau V., Chabrier A., Vallee M.P., Lagiou P., Holcatova I., Richiardi L., Kjaerheim K., Agudo A., Castellsague X., Macfarlane T.V., Barzan L., Canova C., Thakker N.S., Conway D.I., Znaor A., Healy C.M., Ahrens W., Zaridze D., Szeszenia-Dabrowska N., Lissowska J., Fabianova E., Mates I.N., Bencko V., Foretova L., Janout V., Curado

M.P., Koifman S., Menezes A., Wunsch-Filho V., Eluf-Neto J., Boffetta P., Fernandez G.L., Polesel J., Lener M., Jaworowska E., Lubinski J., Boccia S., Rajkumar T., Samant T.A., Mahimkar M.B., Matsuo K., Franceschi S., Byrnes G., Brennan P., McKay J.D., 2015. "**A rare truncating BRCA2 variant and genetic susceptibility to upper aerodigestive tract cancer**". *J.Natl.Cancer Inst.* 107(5).

4. Fachiroh J., Sangrajrang S., Johansson M., Renard H., Gaborieau V., Chabrier A., Chindavijak S., Brennan P., McKay J.D., 2012. "**Tobacco consumption and genetic susceptibility to nasopharyngeal carcinoma (NPC) in Thailand**". *Cancer Causes Control* 23(12), 1995-2002.

5. Purdue M.P., Johansson M., Zelenika D., Toro J.R., Scelo G., Moore L.E., Prokhortchouk E., Wu X., Kiemeney L.A., Gaborieau V., Jacobs K.B., Chow W.H., Zaridze D., Matveev V., Lubinski J., Trubicka J., Szeszenia-Dabrowska N., Lissowska J., Rudnai P., Fabianova E., Bucur A., Bencko V., Foretova L., Janout V., Boffetta P., Colt J.S., Davis F.G., Schwartz K.L., Banks R.E., Selby P.J., Harnden P., Berg C.D., Hsing A.W., Grubb R.L., III, Boeing H., Vineis P., Clavel-Chapelon F., Palli D., Tumino R., Krogh V., Panico S., Duell E.J., Quiros J.R., Sanchez M.J., Navarro C., Ardanaz E., Dorronsoro M., Khaw K.T., Allen N.E., Bueno-de-Mesquita H.B., Peeters P.H., Trichopoulos D., Linseisen J., Ljungberg B., Overvad K., Tjonneland A., Romieu I., Riboli E., Mukeria A., Shangina O., Stevens V.L., Thun M.J., Diver W.R., Gapstur S.M., Pharoah P.D., Easton D.F., Albanes D., Weinstein S.J., Virtamo J., Vatten L., Hveem K., Njolstad I., Tell G.S., Stoltenberg C., Kumar R., Koppova K., Cussenot O., Benhamou S., Oosterwijk E., Vermeulen S.H., Aben K.K., van der Marel S.L., Ye Y., Wood C.G., Pu X., Mazur A.M., Boulygina E.S., Chekanov N.N., Foglio M., Lechner D., Gut I., Heath S., Blanche H., Hutchinson A., Thomas G., Wang Z., Yeager M., Fraumeni J.F., Jr., Skryabin K.G., McKay J.D., Rothman N., Chanock S.J., Lathrop M., Brennan P., 2011. "**Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3**". *Nat.Genet.* 43(1), 60-65.

6. Cozen W., Timofeeva M.N., Li D., Diepstra A., Hazelett D., Delahaye-Sourdeix M., Edlund C.K., Franke L., Rostgaard K., Van Den Berg D.J., Cortessis V.K., Smedby K.E., Glaser S.L., Westra H.J., Robison L.L., Mack T.M., Ghesquieres H.,

Hwang A.E., Nieters A., de S.S., Lightfoot T., Becker N., Maynadie M., Foretova L., Roman E., Benavente Y., Rand K.A., Nathwani B.N., Glimelius B., Staines A., Boffetta P., Link B.K., Kiemeney L., Ansell S.M., Bhatia S., Strong L.C., Galan P., Vatten L., Habermann T.M., Duell E.J., Lake A., Veenstra R.N., Visser L., Liu Y., Urayama K.Y., Montgomery D., Gaborieau V., Weiss L.M., Byrnes G., Lathrop M., Cocco P., Best T., Skol A.D., Adami H.O., Melbye M., Cerhan J.R., Gallagher A., Taylor G.M., Slager S.L., Brennan P., Coetzee G.A., Conti D.V., Onel K., Jarrett R.F., Hjalgrim H., van den Berg A., McKay J.D., 2014. "**A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus**". *Nat.Commun.* 5, 3856.

7. LifeTechnologies. "**5500XL SOLiD System**"
Online Source: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/5500xl-solid.html (Access Date: 1-3-2015)

8. Mardis E.R., 2008. "**Next-generation DNA sequencing methods**". *Annu.Rev.Genomics Hum.Genet.* 9, 387-402.

9. LifeTechnologies. "**Ion PGM System for Next-Generation Sequencing**"
Online Source: http://www.lifetechnologies.com/fr/fr/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing.html (Access Date: 1-3-2015)

10. LifeTechnologies. "**Ion Proton System for Next-Generation Sequencing**"
Online Source: http://www.lifetechnologies.com/fr/fr/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-proton-system-for-next-generation-sequencing.html (Access Date: 1-3-2015)

11. LifeTechnologies. "**7900HT Fast Real-Time PCR System with 384-Well Block Module**"
Online Source: https://www.lifetechnologies.com/order/catalog/product/4329001
(Access Date: 1-3-2015)

12. BioFire Diagnostics. "**LightScanner Instrument**"
Online Source:
http://www.biofiredx.com/pdfs/LightScanner/LightScanner%20InfoSht-0041.pdf
(Access Date: 1-3-2015)

13. Illumina. "**Illumina Beadstation 500**"
Online Source: http://www.geneworks.com.au/library/IL_Beadstation_sales-ROW.pdf (Access Date: 1-3-2015)

14. Fejes, AP. "**What is a bioinformatician**"
Online Source: http://blog.fejes.ca/?p=2418 (Access Date: 1-3-2015)

15. EMBL-EBI and Welcome Trust Sanger Institute. "**Ensembl Genome Database**"
Online Source: http://www.ensembl.org/index.html (Access Date: 1-3-2015)

16. Vaca-Paniagua F., Alvarez-Gomez R., Maldonado-Martinez H., Perez-Plasencia C., Fragoso-Ontiveros V., Lasa-Gonsebatt F., Herrera L., Cantu D., Bargallo-Rocha E., Mohar A., Durand G., Forey N., Voegele C., Vallée M., Le Calvez-Kelm F., McKay J., Ardin M., Villar S., Zavadil J., Olivier M., 2015. "**Revealing the molecular portrait of triple negative breast tumors in an understudied population through omics analysis of formalin-fixed and paraffin-embedded tissues**". *PLoS.One.*

17. Kim Y.H., Ohta T., Oh J.E., Le Calvez-Kelm F., McKay J., Voegele C., Durand G., Mittelbronn M., Kleihues P., Paulus W., Ohgaki H., 2014. "**TP53, MSH4, and LATS1 germline mutations in a family with clustering of nervous system tumors**". *Am.J.Pathol.* 184(9), 2374-2381.

18. Damiola F., Pertesi M., Oliver J., Le Calvez-Kelm F., Voegele C., Young E.L., Robinot N., Forey N., Durand G., Vallee M.P., Tao K., Roane T.C., Williams G.J., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Goldgar D.E., Lesueur F., Tavtigian S.V., 2014. "**Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study**". *Breast Cancer Res.* 16(3), R58.

19. Le Calvez-Kelm F., Oliver J., Damiola F., Forey N., Robinot N., Durand G., Voegele C., Vallee M.P., Byrnes G., Registry B.C., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Tavtigian S.V., Lesueur F., 2012. "**RAD51 and breast cancer susceptibility: no evidence for rare variant association in the Breast Cancer Family Registry study**". *PLoS.One.* 7(12), e52374.

20. Ahmad J., Le Calvez-Kelm F., Daud S., Voegele C., Vallee M., Ahmad A., Kakar N., McKay J.D., Gaborieau V., Leone M., Sinilnikova O., Sangrajrang S., Tavtigian S.V., Lesueur F., 2012. "**Detection of BRCA1/2 mutations in breast cancer patients from Thailand and Pakistan**". *Clin.Genet.* 82(6), 594-598.

21. Park D.J., Lesueur F., Nguyen-Dumont T., Pertesi M., Odefrey F., Hammet F., Neuhausen S.L., John E.M., Andrulis I.L., Terry M.B., Daly M., Buys S., Le Calvez-Kelm F., Lonie A., Pope B.J., Tsimiklis H., Voegele C., Hilbers F.M., Hoogerbrugge N., Barroso A., Osorio A., Giles G.G., Devilee P., Benitez J., Hopper J.L., Tavtigian S.V., Goldgar D.E., Southey M.C., 2012. "**Rare mutations in XRCC2 increase the risk of breast cancer**". *Am.J.Hum.Genet.* 90(4), 734-739.

22. Le Calvez-Kelm F., Lesueur F., Damiola F., Vallee M., Voegele C., Babikyan D., Durand G., Forey N., McKay-Chopin S., Robinot N., Nguyen-Dumont T., Thomas A., Byrnes G.B., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Tavtigian S.V., 2011. "**Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study**". *Breast Cancer Res.* 13(1), R6.

23. Tavtigian S.V., Oefner P.J., Babikyan D., Hartmann A., Healey S., Le Calvez-Kelm F., Lesueur F., Byrnes G.B., Chuang S.C., Forey N., Feuchtinger C., Gioia L., Hall J., Hashibe M., Herte B., McKay-Chopin S., Thomas A., Vallee M.P., Voegele C., Webb P.M., Whiteman D.C., Sangrajrang S., Hopper J.L., Southey M.C., Andrulis I.L., John E.M., Chenevix-Trench G., 2009. "**Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer**". *Am.J.Hum.Genet.* 85(4), 427-446.

24. Gray J., Liu D., Nieto-Santisteban M., Szalay A., DeWitt D., Heber G. "**Scientific data management in the coming decade**".34(4), 34-41. 2005. ACM. *SIGMOD Record.*

25. EMBL - European Bioinformatics Institute. "**EMBL - EBI Annual Scientific Report 2012**". 2013.

26. Mardis E.R., 2011. "**A decade's perspective on DNA sequencing technology**". *Nature* 470(7333), 198-203.

27. Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., Robinson G.E., 2015. "**Big Data: Astronomical or Genomical?**". *PLoS.Biol.* 13(7), e1002195.

28. Genomics England. "**The 100,000 Genomes Project**"
Online Source: http://www.genomicsengland.co.uk/the-100000-genomes-project/
(Access Date: 27-7-2015)

29. Briggs, H. "**Hundred thousand genomes to be mapped in Saudi Arabia**"
Online Source: http://www.bbc.com/news/health-25216135 (Access Date: 27-7-2015)

30. Sulem P., Helgason H., Oddson A., Stefansson H., Gudjonsson S.A., Zink F., Hjartarson E., Sigurdsson G.T., Jonasdottir A., Jonasdottir A., Sigurdsson A., Magnusson O.T., Kong A., Helgason A., Holm H., Thorsteinsdottir U., Masson G., Gudbjartsson D.F., Stefansson K., 2015. "**Identification of a large set of rare complete human knockouts**". *Nat.Genet.* 47(5), 448-452.

31. Kaiser, J, 30-1-2015. "**White House fleshes out Obama's $215 million plan for precision medicine**". *Science (News) Magazine* (http://news.sciencemag.org/biology/2015/01/white-house-fleshes-out-obama-s-215-million-plan-precision-medicine)

32. Zhu J., 2012. "**A year of great leaps in genome research**". *Genome Medicine* 4(1).

33. Hadfield, J and Loman, N. "**Next Generation Genomics: World Map of High-throughput Sequencers**"
Online Source: http://omicsmaps.com/ (Access Date: 27-7-2015)

34. Koh J., Krishnan S., Hong S., Tan P., Khan A., Lee M., Brusic V., 2004. "**BioWare: A framework for bioinformatics data retrieval, annotation and publishing**". *ACM SIGIR Conference 2004 - Workshop*.

35. Jagadish H., Gehrke J., Labrinidis A., Papakonstantinou Y., Patel J., Ramakrishnan R., Shahabi C., 2014. "**Big Data and Its Technical Challenges**". *Communications of the ACM* 57(7), 86-94.

36. Agarwal D., Pineda S., Michailidou K., Herranz J., Pita G., Moreno L.T., Alonso M.R., Dennis J., Wang Q., Bolla M.K., Meyer K.B., Menendez-Rodriguez P., Hardisson D., Mendiola M., Gonzalez-Neira A., Lindblom A., Margolin S., Swerdlow A., Ashworth A., Orr N., Jones M., Matsuo K., Ito H., Iwata H., Kondo N., Hartman M., Hui M., Lim W.Y., Iau P.T., Sawyer E., Tomlinson I., Kerin M., Miller N., Kang D., Choi J., Park S.K., Noh D., Hopper J.L., Schmidt D.F., Makalic E., Southey M.C., Teo S.H., Yip C.H., Sivanandan K., Tay W., Brauch H., Bruning T., Hamann U., Dunning A.M., Shah M., Andrulis I.L., Knight J.A., Glendon G., Tchatchou S., Schmidt M.K., Broeks A., Rosenberg E.H., van't Veer L.J., Fasching P.A., Renner S.P., Ekici A.B., Beckmann M.W., Shen C., Hsiung C., Yu J., Hou M., Blot W., Cai Q., Wu A.H., Tseng C., Van Den Berg D., Stram D.O., Cox A., Brock I.W., Reed M.W., Muir K., Lophatananon A., Stewart-Brown S., Siriwanarangsan P., Zheng W., Deming-Halverson S., Shrubsole M.J., Long J., Shu X., Lu W., Gao Y., Zhang B., Radice P., Peterlongo P., Manoukian S., Mariette F., Sangrajrang S., McKay J.,

Couch F.J., Toland A.E., Yannoukakos D., Fletcher O., Johnson N., dos S.S., I, Peto J., Marme F., Burwinkel B., Guenel P., Truong T., Sanchez M., Mulot C., Bojesen S.E., Nordestgaard B.G., Flyer H., Brenner H., Dieffenbach A.K., Arndt V., Stegmaier C., Mannermaa A., Kataja V., Kosma V., Hartikainen J.M., Lambrechts D., Yesilyurt B.T., Floris G., Leunen K., Chang-Claude J., Rudolph A., Seibold P., Flesch-Janys D., Wang X., Olson J.E., Vachon C., Purrington K., Giles G.G., Severi G., Baglietto L., Haiman C.A., Henderson B.E., Schumacher F., Marchand L.L., Simard J., Dumont M., Goldberg M.S., Labreche F., Winqvist R., Pylkas K., Jukkola-Vuorinen A., Grip M., Devilee P., Tollenaar R.A., Seynaeve C., Garcia-Closas M., Chanock S.J., Lissowska J., Figueroa J.D., Czene K., Eriksson M., Humphreys K., Darabi H., Hooning M.J., Kriege M., Collee J.M., Tilanus-Linthorst M., Li J., Jakubowska A., Lubinski J., Jaworska-Bieniek K., Durda K., Nevanlinna H., Muranen T.A., Aittomaki K., Blomqvist C., Bogdanova N., Dork T., Hall P., Chenevix-Trench G., Easton D.F., Pharroah P.D., Arias-Perez J.I., Zamora P., Benitez J., Milne R.L., 2014. "**FGF receptor genes and breast cancer susceptibility: results from the Breast Cancer Association Consortium**". *Br.J.Cancer* 110(4), 1088-1100.

37. NCBI. "**Genbank**"
Online Source: http://www.ncbi.nlm.nih.gov/genbank/ (Access Date: 1-8-2015)

38. UniProt. "**UniProt**"
Online Source: http://www.ebi.ac.uk/uniprot (Access Date: 1-8-2015)

39. Cudre-Mauroux P., Kimura H., Lim K., Rogers J., Simakov R., Soroush E., Velikhov P., Wang D., Balazinska M., Becla J., DeWitt D., Heath B., Maier D., Madden S., Patel J., Stonebraker M., Zdonik S., 2009. "**A demonstration of SciDB: a science-oriented DBMS**". *Proceedings of the VLDB Endowment* 2(2), 1534-1537.

40. Dozier J., Stonebraker M., Frew J., 1994. "**A next-generation information system for the study of global change**". *Mass Storage Systems*, 47-53.

41. SSDS. "**The Sloan Digital Sky Survey**"
Online Source: http://www.sdss.org/ (Access Date: 28-7-2015)

42. Ivanova M., Nes N., Goncalves R., Kersten M., 2007. "**MonetDB/SQL Meets SkyServer: the Challenges of a Scientific Database**". *SSDBM '07 Proceedings of the 19th International Conference on Scientific and Statistical Database Management*, 13.

43. Kim E., 2015. "**The Future of Molecular Medicine: Biomarkers, BATTLEs, and Big Data**". *Am Soc Clin Oncol Educ Book*, 22-7.

44. Brown University Data Management Research Group. "**SciDB – Data Management System for Large Scale Scientific Data**"
Online Source: http://database.cs.brown.edu/projects/scidb/ (Access Date: 27-7-2015)

45. LSST. "**Large Synoptic Survey Telescope**"
Online Source: http://www.lsst.org/lsst/ (Access Date: 28-7-2015)

46. Kantor J., Axelrod T., Becla J., Cook J., Nikolaev S., Gray J., Plante R., Nieto-Santisteban M., Hopkins J., 2006. "**Designing for peta-scale in the LSST database**". *Astronomical Data Analysis Software and Systems*.

47. The Apache Software Foundation. "**Apache Hadoop**"
Online Source: http://hadoop.apache.org/ (Access Date: 28-7-2015)

48. Baker M., 2010. "**Next-generation sequencing: adjusting to data overload**". *Nature Methods* 7, 495-499.

49. Marx V., 2013. "**Biology: The big challenges of big data**". *Nature* 498(7453), 255-260.

50. Studt, T, 5-7-2014. "**Technological Trends in the Lab**". *Laboratory Equipment Magazine* (http://www.laboratoryequipment.com/articles/2014/05/technological-trends-lab)

51. Nussbeck S.Y., Weil P., Menzel J., Marzec B., Lorberg K., Schwappach B., 2014. "**The laboratory notebook in the 21st century: The electronic laboratory notebook would enhance good scientific practice and increase research productivity**". *EMBO Rep.* 15(6), 631-634.

52. KANARES H., 1985. Writing the Laboratory Notebook.

53. Bird C.L., Willoughby C., Frey J.G., 2013. "**Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences**". *Chem.Soc.Rev.* 42(20), 8157-8175.

54. Ryan P. "**Keeping a Lab Notebook**".*National Institutes of Health, Office of Intramural Training and Education* . 2015.

55. Lysakowski R., 1997. "**The Collaborative Electronic Notebook Systems Consortium**". *IEEE*, pp. 2659-2661.

56. Wright J.M., 2009. "**Make it better but don't change anything**". *Autom.Exp.* 1, 5.

57. Rubacha M., Rattan A.K., Hosselet S.C., 2011. "**A review of electronic laboratory notebooks available in the market today**". *J.Lab Autom.* 16(1), 90-98.

58. Studylog. "**Studylog**"
Online Source: http://www.studylog.com/products/ (Access Date: 1-8-2015)

59. Textco BioSoftware. "**Gene Inspector**"
Online Source: http://www.textco.com/gene-inspector.php (Access Date: 1-8-2015)

60. Tripos/Certara. "**Benchware 3D Explorer**"
Online Source: http://www.certara.com/products/molmod/bw3de (Access Date: 1-8-2015)

61. NoteBookMaker. "**NoteBookMaker**"
Online Source: http://www.notebookmaker.com/ (Access Date: 1-8-2015)

62. Agilent Technologies. "**Agilent/Kalabie ELN**"
Online Source: http://www.chem.agilent.com/Library/brochures/5989-7278EN%20-%20ELN%20Brochure%20PRINT.pdf (Access Date: 1-8-2015)

63. PerkinElmer-Labtronics. "**Nexxis ELN**"
Online Source: http://www.cambridgesoft.com/literature/PDF/Nexxis%20ELN.pdf (Access Date: 1-8-2015)

64. SoftGroup. "**Velquest SmartLab**"
Online Source: http://softgroup.bg/en/velquestsmartlab (Access Date: 1-8-2015)

65. PerkinElmer. "**E-Notebook**"
Online Source: http://www.cambridgesoft.com/Ensemble/E-notebook/ (Access Date: 1-8-2015)

66. IDBS. "**E-WorkBook**"
Online Source: http://www.idbs.com/en/platform-products/e-workbook/ (Access Date: 1-8-2015)

67. Geist, A, Schwidder, J, Jung, D, and Nachtigal, N. "**ORNL Electronic Notebook Project**"
Online Source: http://www.csm.ornl.gov/~geist/java/applets/enote/ (Access Date: 1-3-2015)

68. Ling, MHT. "**CyNote - Cyber Laboratory Notebook for Biologists and Bioinformaticists**"
Online Source: http://cynote.sourceforge.net/ (Access Date: 1-3-2015)

69. Osburn, D. "**The Monster Journal**"
Online Source: http://monsterjournal.sourceforge.net/ (Access Date: 1-3-2015)

70. Myers, JD. "**Electronic Laboratory Notebook (ELN)**"
Online Source: http://collaboratory.emsl.pnl.gov/software/eln/ (Access Date: 1-3-2015)

71. Lang, S. "**LabBook**"
Online Source: http://sourceforge.net/projects/labbook/ (Access Date: 1-3-2015)

72. WordPress Community. "**WordPress**"
Online Source: http://wordpress.org/ (Access Date: 1-3-2015)

73. Kurdle and B. "**MyLabBook**"
Online Source: http://www.mylabbook.org/ (Access Date: 1-3-2015)

74. Voegele C., Bouchereau B., Robinot N., McKay J., Damiecki P., Alteyrac L., 2013. "**A universal open-source Electronic Laboratory Notebook**". *Bioinformatics.* 29(13), 1710-1712.

75. US Food and Drug Administration (FDA). "**CFR Code of Federal Regulations Title 21 Part 11 Electronic records; Electronic signatures**"
Online Source:
http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?CFRPart=11&showFR=1 (Access Date: 1-3-2015)

76. SurveyGizmo. "**SurveyGizmo**"
Online Source: https://www.surveygizmo.com/ (Access Date: 1-8-2015)

77. Butler D., 2005. "**Electronic notebooks: a new leaf**". *Nature* 436(7047), 20-21.

78. Dutton, G, 1-12-2006. "**Lab Notebooks Offer Efficiency Gains**". *Genetic Engineering & Biotechnology News Magazine* (http://www.genengnews.com/gen-articles/lab-notebooks-offer-efficiency-gains/1951/)

79. Smith, C, 25-3-2014. "**Go Paperless with These Electronic Lab Notebooks**". *Biocompare Magazine* (http://www.biocompare.com/Editorial-Articles/158438-Go-Paperless-with-These-Electronic-Lab-Notebooks/)

80. Smith, C, 2-10-2012. "**Lab Apps for Scientists**". *Biocompare Magazine* (http://www.biocompare.com/Editorial-Articles/122079-Lab-Apps-for-Scientists/)

81. Myers, JD. "**Collaborative Electronic Notebooks as Electronic Records: Design Issues for the Secure Electronic Laboratory Notebook (ELN)**"
Online Source:
http://collaboratory.emsl.pnl.gov/resources/publications/papers/seceln%28final1%291-22Nov.pdf (Access Date: 1-3-2015)

82. Figueras J., 1987. "**An Electronic Notebook for Chemists**". *American Chemical Society*, pp. 37-47.

83. Skidmore J., Sottile M., Cuny J., Malony A., 1998. "**A Prototype Notebook-Based Environment for Computational Tools**". *IEEE.*

84. Carpi, N. "**eLabFTW**"
Online Source: http://www.elabftw.net/ (Access Date: 1-3-2015)

85. Elliott, MH, 2007. "**The state of the ELN Market**". *Scientific Computing Magazine* (http://www.scientific-computing.com/features/feature.php?feature_id=50)

86. NEHME A., SCOFFIN R., 2006. Electronic Laboratory Notebooks. Computer Applications in Pharmaceutical Research and Development.

87. Elliott, MH, 30-9-2008. "**Electronic Laboratory Notebooks Enter Mainstream Informatics**". *Scientific Computing Magazine*
(http://www.scientificcomputing.com/articles/2008/09/electronic-laboratory-notebooks-enter-mainstream-informatics)

88. Carpi, N. "**Scientists and open source: the evolution of the lab notebook**"
Online Source: http://osdelivers.blackducksoftware.com/2014/04/16/scientists-and-open-source-the-evolution-of-the-lab-notebook/ (Access Date: 27-8-2015)

89. Bruce, S, 31-12-2002. "**A Look at the State of Electronic Lab Notebook Technology**". *Scientific Computing Magazine*
(http://www.scientificcomputing.com/articles/2002/12/look-state-electronic-lab-notebook-technology)

90. Gibbon G., 1996. "**A brief history of LIMS**". *Laboratory Automation and Information Management* 32, 1-5.

91. International Human Genome Sequencing Consortium, 2004. "**Finishing the euchromatic sequence of the human genome**". *Nature* 431(7011), 931-945.

92. Metzker M.L., 2010. "**Sequencing technologies - the next generation**". *Nat.Rev.Genet.* 11(1), 31-46.

93. Voegele C., Tavtigian S.V., De S.D., Cuber S., Thomas A., Le Calvez-Kelm F., 2007. "**A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening**". *Bioinformatics.* 23(18), 2504-2506.

94. ABI/Labvantage. "**SQL*LIMS**"
Online Source: http://www.sqllims.com/overview.php (Access Date: 1-3-2015)

95. Oracle. "**Oracle SQL Developer**"
Online Source: http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index-097090.html (Access Date: 1-3-2015)

96. Oracle. "**Oracle Developer Suite - Forms**"
Online Source: http://www.oracle.com/technetwork/developer-tools/forms/overview/index-098877.html (Access Date: 1-3-2015)

97. Strass H., 2008. "**Managing the lab - Laboratory Information Management System**". *GIT Laboratory Journal*(5-6), 24-26.

98. marketsandmarkets.com. "**Laboratory Information Management Systems/ LIMS Market by Product (COTS & Legacy), Delivery Mode (On-premise, Hosted, Cloud), Component (Software & Services), End User Industries (Healthcare, CRO, Petrochemical, Oil and Gas, Chemical) - Forecast to 2019**". 2014.

99. Haga S.B., Beskow L.M., 2008. "**Ethical, legal, and social implications of biobanks for genetics research**". *Adv.Genet.* 60, 505-544.

100. Riboli E., Kaaks R., 1997. "**The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition**". *Int.J.Epidemiol.* 26 Suppl 1, S6-14.

101. Caboux E., Plymoth A., Hainaut P. "**Common minimum technical standards and protocols or biological resource centers dedicated to cancer research**". 2007. *IARC Working Group Reports*.

102. BBMRI. "**MIABIS 2.0**"
Online Source: http://bbmri-wiki.wikidot.com/en:dataset (Access Date: 1-3-2015)

103. Wichmann H.E., Kuhn K.A., Waldenberger M., Schmelcher D., Schuffenhauer S., Meitinger T., Wurst S.H., Lamla G., Fortier I., Burton P.R., Peltonen L., Perola M., Metspalu A., Riegman P., Landegren U., Taussig M.J., Litton J.E., Fransson M.N., Eder J., Cambon-Thomsen A., Bovenberg J., Dagher G., van Ommen G.J., Griffith M., Yuille M., Zatloukal K., 2011. "**Comprehensive catalog of European biobanks**". *Nat.Biotechnol.* 29(9), 795-797.

104. Norlin L., Fransson M.N., Eriksson M., Merino-Martinez R., Anderberg M., Kurtovic S., Litton J.E., 2012. "**A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS**". *Biopreserv.Biobank.* 10(4), 343-348.

105. Hansson M.G., 2009. "**Ethics and biobanks**". *Br.J.Cancer* 100(1), 8-12.

106. Elliott P., Peakman T.C., 2008. "**The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine**". *Int.J.Epidemiol.* 37(2), 234-244.

107. Owen J.M., Woods P., 2008. "**Designing and implementing a large-scale automated -80 degrees C archive**". *Int.J.Epidemiol.* 37 Suppl 1, i56-i61.

108. Oracle. "**Database Utilities - SQL*Loader Concepts**"
Online Source:
http://docs.oracle.com/cd/B19306_01/server.102/b14215/ldr_concepts.htm#g1013706
(Access Date: 1-3-2015)

109. Eder J., Gottweis H., Zatloukal K., 2012. "**IT solutions for privacy protection in biobanking**". *Public Health Genomics* 15(5), 254-262.

110. Oracle. "**Oracle Recovery Manager (RMAN)**"
Online Source:
http://www.oracle.com/technetwork/database/features/availability/rman-overview-096633.html (Access Date: 1-3-2015)

111. Voegele C., Alteyrac L., Caboux E., Smans M., Lesueur F., Le Calvez-Kelm F., Hainaut P., 2010. "**A sample storage management system for biobanks**". *Bioinformatics.* 26(21), 2798-2800.

112. Kersting M., Prokein J., Bernemann I., Drobek D., Illig K. "**IT-Systems for Biobanking – A brief Overview**" (Slide). 2014.

113. SmartBiobank. "**SmartBiobank**"
Online Source: ww.smartbiobank.com (Access Date: 1-8-2015)

114. Oriam. "**DataBiotec**"
Online Source: http://www.oriam.com/oriam/Produits/DataBiotec.html (Access Date: 1-8-2015)

115. Karolinksa Institutet. "**KI Biobank**"
Online Source: http://ki.se/en/research/ki-biobank (Access Date: 10-8-2015)

116. UK Biobank. "**UK Biobank**"
Online Source: http://www.ukbiobank.ac.uk/ (Access Date: 30-7-2015)

117. Downey P., Peakman T.C., 2008. "**Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles**". *Int.J.Epidemiol.* 37 Suppl 1, i46-i50.

118. Actian Community. "**Ingres DBMS**"
Online Source:
http://community.actian.com/wiki/Ingres_DBMS_Home/Ingres_DBMS_Learn
(Access Date: 24-8-2015)

119. Muilu J., Peltonen L., Litton J.E., 2007. "**The federated database--a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe**". *Eur.J.Hum.Genet.* 15(7), 718-723.

120. IARC. "**BCNet**"
Online Source: http://bcnet.iarc.fr/ (Access Date: 1-3-2015)

121. Kühne M., Liehr A., 2009. "**Improving the Traditional Information Management in Natural Sciences**". *Data Science Journal* 8, 18-26.

122. Laboratory Informatics Institute. "**LIMSfinder.com the on-line interactive magazine**"
Online Source: http://www.limsfinder.com/ (Access Date: 15-4-2015)

123. Grimes S.M., Ji H.P., 2014. "**MendeLIMS: a web-based laboratory information management system for clinical genome sequencing**". *BMC.Bioinformatics.* 15, 290.

124. Venco F., Vaskin Y., Ceol A., Muller H., 2014. "**SMITH: a LIMS for handling next-generation sequencing workflows**". *BMC.Bioinformatics.* 15 Suppl 14, S3.

125. Troshin P.V., Postis V.L., Ashworth D., Baldwin S.A., McPherson M.J., Barton G.J., 2011. "**PIMS sequencing extension: a laboratory information management system for DNA sequencing facilities**". *BMC.Res.Notes* 4, 48.

126. Cho M., Kang J., Park H., 2007. "**A DNA Microarray LIMS System for Integral Genomic Analysis of Multi-Platform Microarrays**". *Genomics and Informatics* 5, 83-87.

127. Tagger, B. "**An Introduction and Guide to Successfully Implementing a LIMS (Laboratory Information Management System)**"
Online Source: http://www0.cs.ucl.ac.uk/staff/B.Tagger/LimsPaper.pdf (Access Date: 15-4-2015)

128. National Cancer Research Institute (NCRI). "**Biosamples for research**"
Online Source: http://www.ncri.org.uk/initiatives/biobanking (Access Date: 15-4-2015)

129. IARC. "**IARC Biobank**"
Online Source: http://ibb.iarc.fr/ (Access Date: 30-7-2015)

130. BBC, 30-3-2012. "**UK biobank opens to researchers**". *BBC News Newspaper* (http://www.bbc.com/news/health-17553931)

131. The EuroBioBank. "**The EuroBioBank**"
Online Source: http://www.eurobiobank.org/en/information/info_institut.htm (

132. Fransson M.N., Rial-Sebbag E., Brochhausen M., Litton J.E., 2015. "**Toward a common language for biobanking**". *Eur.J.Hum.Genet.* 23(1), 22-28.

133. Elliott, MH, 1-7-2011. "**New Debates over Intellectual Property Protection and ELN**". *Scientific Computing Magazine* (http://www.scientificcomputing.com/articles/2011/01/new-debates-over-intellectual-property-protection-and-eln)

134. Joly Y., Dalpe G., So D., Birko S., 2015. "**Fair Shares and Sharing Fairly: A Survey of Public Views on Open Science, Informed Consent and Participatory Research in Biobanking**". *PLoS.One.* 10(7), e0129893.

135. Oliver J.M., Slashinski M.J., Wang T., Kelly P.A., Hilsenbeck S.G., McGuire A.L., 2012. "**Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives**". *Public Health Genomics* 15(2), 106-114.

136. Vaught J.B., Caboux E., Hainaut P., 2010. "**International efforts to develop biospecimen best practices**". *Cancer Epidemiol.Biomarkers Prev.* 19(4), 912-915.

137. Bolton, S, 7-10-2009. "**LIMS and ELN: 1+1=3**". *Scientific Computing Magazine* (http://www.scientificcomputing.com/articles/2009/07/lims-and-eln-1-1-3)

138. Machina H.K., Wild D.J., 2013. "**Electronic laboratory notebooks progress and challenges in implementation**". *J.Lab Autom.* 18(4), 264-268.

139. MySQL. "**The FEDERATED Storage Engine**"
Online Source: http://dev.mysql.com/doc/refman/5.0/en/federated-storage-engine.html (Access Date: 21-4-2015)

140. Machina H.K., Wild D.J., 2013. "**Laboratory informatics tools integration strategies for drug discovery: integration of LIMS, ELN, CDS, and SDMS**". *J.Lab Autom.* 18(2), 126-136.

141. Lacroix Z., 2002. "**Biological data integration: wrapping data and tools**". *IEEE Trans.Inf.Technol.Biomed.* 6(2), 123-128.

142. Hobbie K., Peterson E., Barton M., Waters K., Anderson K., 2012. "**Integration of Data Systems and Technology Improves Research and Collaboration for a Superfund Research Center**". *Journal of Laboratory Automation* 17(4), 275-283.

143. Ettridge, L. "**UK to Become World Number One in DNA Testing with Plan to Revolutionise Fight Against Cancer and Rare Diseases**"
Online Source: http://finance.yahoo.com/news/uk-become-world-number-one-230500699.html (Access Date: 1-3-2015)

144. Hunkapiller T., Hood L., 1991. "**LIMS and the Human Genome Project**". *Biotechnology (N.Y.)* 9(12), 1344-1345.

145. Ioannidis J.P., Allison D.B., Ball C.A., Coulibaly I., Cui X., Culhane A.C., Falchi M., Furlanello C., Game L., Jurman G., Mangion J., Mehta T., Nitzberg M., Page G.P., Petretto E., van N., V, 2009. "**Repeatability of published microarray gene expression analyses**". *Nat.Genet.* 41(2), 149-155.

146. Malone J., Brown A., Lister A.L., Ison J., Hull D., Parkinson H., Stevens R., 2014. "**The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation**". *J.Biomed.Semantics.* 5, 25.

147. Peng R.D., 2011. "**Reproducible research in computational science**". *Science* 334(6060), 1226-1227.

148. Nekrutenko A., Taylor J., 2012. "**Next-generation sequencing data interpretation: enhancing reproducibility and accessibility**". *Nat.Rev.Genet.* 13(9), 667-672.

149. The Galaxy Team. "**Galaxy**"
Online Source: https://usegalaxy.org/ (Access Date: 1-8-2015)

150. Goecks J., Nekrutenko A., Taylor J., 2010. "**Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**". *Genome Biol.* 11(8), R86.

151. Cuccuru G., Orsini M., Pinna A., Sbardellati A., Soranzo N., Travaglione A., Uva P., Zanetti G., Fotia G., 2014. "**Orione, a web-based framework for NGS analysis in microbiology**". *Bioinformatics.* 30(13), 1928-1929.

152. Scholtalbers J., Rossler J., Sorn P., de G.J., Boisguerin V., Castle J., Sahin U., 2013. "**Galaxy LIMS for next-generation sequencing**". *Bioinformatics.* 29(9), 1233-1234.

153. Mariette J., Escudie F., Allias N., Salin G., Noirot C., Thomas S., Klopp C., 2012. "**NG6: Integrated next generation sequencing storage and processing environment**". *BMC.Genomics* 13, 462.

154. Fischer B., Zigmond M., 2010. "**The Essential Nature of Sharing in Science**". *Science and Engineering Ethics* 16(4), 783-799.

155. Thompson, N, 2001. "**Publisher Perish - The coming battle over putting all scientific journals online for free**". *Washington Monthly Magazine* (http://www.washingtonmonthly.com/features/2001/0110.thompson.html)

156. PLOS ONE. "**PLOS ONE Journal Information**"
Online Source: http://www.plosone.org/static/information (Access Date: 1-3-2015)

157. Nature Publishing Group. "**Scientific Data**"
Online Source: http://www.nature.com/sdata/ (Access Date: 1-8-2015)

158. The Open Knowledge Foundation. "**Open Knowledge**"
Online Source: https://okfn.org/ (Access Date: 1-8-2015)

159. Creative Commons. "**Creative Commons**"
Online Source: http://creativecommons.org/ (Access Date: 1-3-2015)

160. Nielsen, M. "**An informal definition of OpenScience**"
Online Source: http://www.openscience.org/blog/?p=454 (Access Date: 1-4-2015)

161. Sage Bionetworks. "**Synapse**"
Online Source: https://www.synapse.org/ (Access Date: 15-7-2015)

162. Weinstein J.N., Collisson E.A., Mills G.B., Shaw K.R., Ozenberger B.A., Ellrott K., Shmulevich I., Sander C., Stuart J.M., 2013. "**The Cancer Genome Atlas Pan-Cancer analysis project**". *Nat.Genet.* 45(10), 1113-1120.

163. NASA. "**COHOWeb**"
Online Source: http://omniweb.gsfc.nasa.gov/coho/ (Access Date: 1-8-2015)

164. Duarte A.M., Psomopoulos F.E., Blanchet C., Bonvin A.M., Corpas M., Franc A., Jimenez R.C., de Lucas J.M., Nyronen T., Sipos G., Suhr S.B., 2015. "**Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis**". *Front Genet.* 6, 197.

165. AFM. "**Décrypton**"
Online Source: http://www.afm-telethon.fr/projet-decrypthon-3149 (Access Date: 15-7-2015)

166. IBM. "**World Community Grid**"
Online Source: http://www.worldcommunitygrid.org/ (Access Date: 15-7-2015)

167. EGI. "**European Grid Infrastructure**"
Online Source: http://www.egi.eu/infrastructure/index.html (Access Date: 1-8-2015)

168. Weisman, R, 24-6-2015. "**Broad Institute, Google Genomics to develop online tools to analyze genetic data**". *The Boston Globe Newspaper* (http://www.betaboston.com/news/2015/06/24/broad-institute-joins-with-google-genomics-to-develop-online-tools-to-analyze-genetic-data/)

169. Paten B., Diekhans M., Druker B.J., Friend S., Guinney J., Gassner N., Guttman M., James K.W., Mantey P., Margolin A.A., Massie M., Novak A.M., Nothaft F., Pachter L., Patterson D., Smuga-Otto M., Stuart J.M., Van't Veer L., Wold B., Haussler D., 2015. "**The NIH BD2K center for big data in translational genomics**". *J.Am Med.Inform.Assoc.*

170. myGrid team. "**Taverna Workflow Management System**"
Online Source: http://www.taverna.org.uk/ (Access Date: 1-8-2015)

171. Wolstencroft K., Haines R., Fellows D., Williams A., Withers D., Owen S., Soiland-Reyes S., Dunlop I., Nenadic A., Fisher P., Bhagat J., Belhajjame K., Bacall F., Hardisty A., Nieva d.l.H., Balcazar Vargas M.P., Sufi S., Goble C., 2013. "**The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud**". *Nucleic Acids Res.* 41(Web Server issue), W557-W561.

172. Center for Open Science. "**Open Science Framework**"
Online Source: https://osf.io/ (Access Date: 1-8-2015)

173. Anagnostou P., Capocasa M., Milia N., Sanna E., Battaggia C., Luzi D., Destro B.G., 2015. "**When data sharing gets close to 100%: what human paleogenetics can teach the open science movement**". *PLoS.One.* 10(3), e0121409.

174. Rinaldi A., 2014. "**Spinning the web of open science: Social networks for scientists and data sharing, together with open access, promise to change the way research is conducted and communicated**". *EMBO Rep.* 15(4), 342-346.

175. Goble C.A., Bhagat J., Aleksejevs S., Cruickshank D., Michaelides D., Newman D., Borkum M., Bechhofer S., Roos M., Li P., de R.D., 2010. "**myExperiment: a repository and social network for the sharing of bioinformatics workflows**". *Nucleic Acids Res.* 38(Web Server issue), W677-W682.

Appendices

ORNL Electronic Notebook: http://www.csm.ornl.gov/~geist/java/applets/enote/

This software was used for personal notes by a few users at IARC before the ELN project started. It's written in Perl and enables management of a lot of users but the last update was done in 2010. We had preferred to look for a more modern tool.

CyNote: http://cynote.sourceforge.net/

Cynote is free, open source and written in Python. The tool has been tested at IARC but the generated notebook pages were not very clear. Moreover, Cynote project doesn't seem to be active anymore.

The Monster Journal: http://monsterjournal.sourceforge.net/

This open-source tool is web-based and written in different programming languages (C#, Java, JavaScript, Perl). It has been developed to replace the paper notebook, but didn't seem to enable definition of specific read and write accesses for multi-users.

Electronic Laboratory Notebook: http://collaboratory.emsl.pnl.gov/software/eln/

The development and support of this free and open-source tool has ended.

LabBook: http://sourceforge.net/projects/labbook/

This tool was developed with the Perl Catalyst Framework (a Perl web framework). A free old version is available on SourceForge, and the author provides the current version for free on his website (http://stefan-lang-bioinformatics.eu/?page_id=14) but he does not provide any support.

MyExperiment: http://www.myexperiment.org/

MyExperiment has been developed in Ruby. Its aim is to provide a collaborative environment where scientists can share workflows, experiment plans. It is not really an ELN but more a repository and social website for researchers who wants to share workflows ([175] - Goble et al., 2010).

MyLabBook: http://www.mylabbook.org/

This community website explores how to build electronic laboratory notebooks using Drupal, a free and open-source CMS. It also helps to implement LIMS. It is similar to WordPress are similar, but the wiser choice was to opt for the CMS we already knew at IARC: WordPress.

Appendix 1: Free open-source ELN considered and/or tested

```
CREATE TABLE "IARC_ION_REAGENTS"
   (  "REAGENT_NAME" VARCHAR2(50 BYTE) NOT NULL ENABLE,
      "CURRENT_QTY" NUMBER,
      "LAST_UPDATE_USERSTAMP" VARCHAR2(20 BYTE),
      "LAST_UPDATE_TIMESTAMP" DATE,
      "COMMENTS" VARCHAR2(500 BYTE),
       CONSTRAINT       "IARC_ION_REAGENTS_PK"       PRIMARY       KEY
("REAGENT_NAME"))

 CREATE TABLE "IARC_ION_REAGENTS_MOVEMENTS"
   (  "REAGENT_NAME" VARCHAR2(50 BYTE) NOT NULL ENABLE,
      "MOVEMENT_TYPE" VARCHAR2(20 BYTE),
      "MOVEMENT_DATE" DATE,
      "MOVEMENT_QTY" NUMBER,
      "MOVEMENT_ADDRESSEE" VARCHAR2(50 BYTE),
      "USERSTAMP" VARCHAR2(20 BYTE),
      "TIMESTAMP" DATE,
      "COMMENTS" VARCHAR2(255 BYTE)   )
```

Appendix 2: Database tables for management of PGM/Proton sequencing reagents stocks in the LIMS

```
CREATE TABLE IARC_ION_RUNS"
   (  "RUN_NAME" VARCHAR2(200 BYTE),
      "RUN_DATE" DATE,
      "LIB_POOL" VARCHAR2(50 BYTE),
      "CHIP_PART" VARCHAR2(200 BYTE),
      "CHIP_BARCODE" VARCHAR2(100 BYTE),
      "KIT_PART" VARCHAR2(200 BYTE),
      "KIT_LOT" VARCHAR2(100 BYTE),
      "COMMENTS" VARCHAR2(4000 BYTE),
      "RUN_STATUS" VARCHAR2(20 BYTE),
      "RUN_FOLLOW_UP" VARCHAR2(2000 BYTE),
      "USERSTAMP" VARCHAR2(50 BYTE),
      "TIMESTAMP" DATE,
      "RUN_NB" VARCHAR2(30 BYTE),
      "INSTRUMENT" VARCHAR2(30 BYTE),
      "BED_FILE" VARCHAR2(200 BYTE),
      "TEMPLATE_NAME" VARCHAR2(200 BYTE),
      "CALLER_WORKFLOW" VARCHAR2(500 BYTE),
      "TOTAL_BASES" VARCHAR2(20 BYTE),
      "ISP_LOADING" VARCHAR2(20 BYTE),
      "TOTAL_READS" VARCHAR2(20 BYTE),
      "USABLE_READS" VARCHAR2(20 BYTE),
      "ENRICHMENT" VARCHAR2(20 BYTE),
      "CLONAL" VARCHAR2(20 BYTE),
      "FINAL_LIBRARY" VARCHAR2(20 BYTE),
      "TEST_50AQ17" VARCHAR2(20 BYTE),
      "TOTAL_AQ17" VARCHAR2(20 BYTE),
      "TOTAL_AQ20" VARCHAR2(20 BYTE),
      "TOTAL_AQ_PERFECT" VARCHAR2(20 BYTE),
      "READS_ON_TARGET" VARCHAR2(20 BYTE),
      "AVERAGE_COVERAGE" VARCHAR2(20 BYTE),
      "UNIFORMITY" VARCHAR2(20 BYTE),
      "ANNOTATION" VARCHAR2(2000 BYTE),
      "FILTERING" VARCHAR2(2000 BYTE),
      "ADAPTER_DIMER" VARCHAR2(20 BYTE),
      "LOW_QUALITY" VARCHAR2(20 BYTE),
      "FIRST_REPORT_NB" VARCHAR2(20 BYTE)   )
```

Appendix 3: Database table for management of Ion sequencing runs

```
create or replace
TRIGGER TG_GET_SAMPLE_ID
BEFORE INSERT ON SAMI_SAMPLES
FOR EACH ROW

DECLARE

  s_id  VARCHAR2(10) := null;

  /* 00000 to ZZZZZ without letters I and O to avoid confusion with 1 and 0 (zero)
 * 10 digits + 24 letters = 34*34*34*34*34 ids available = more than 45 millions */
  char_list CHAR(34) := '0123456789ABCDEFGHJKLMNPQRSTUVWXYZ' ;

  v_first_char  NUMBER ;
  v_second_char NUMBER ;
  v_third_char  NUMBER ;
  v_fourth_char NUMBER ;
  v_fifth_char  NUMBER ;

BEGIN

  select substr(max(sample_id), 4, 5) into s_id from samis_samples where sample_id like (:new.Project||'%');

  -- First id
  if s_id is null then
    :new.sample_id := :new.Project || '00000';
    return;
  end if;

  v_first_char  := instr(char_list, substr(s_id,1,1));
  v_second_char := instr(char_list, substr(s_id,2,1));
  v_third_char  := instr(char_list, substr(s_id,3,1));
  v_fourth_char := instr(char_list, substr(s_id,4,1));
  v_fifth_char  := instr(char_list, substr(s_id,5,1));

  IF v_fifth_char < 34 THEN
    v_fifth_char := v_fifth_char + 1 ;
  ELSE
    v_fifth_char := 0 ;
    IF v_fourth_char < 34 THEN
      v_fourth_char := v_fourth_char + 1 ;
    ELSE
      v_fourth_char := 0 ;
      IF v_third_char < 34 THEN
        v_third_char := v_third_char + 1 ;
      ELSE
        v_third_char := 0 ;
        IF v_second_char < 34 THEN
          v_second_char := v_second_char + 1;
        ELSE
          IF v_first_char < 34 THEN
            v_second_char := 0;
            v_first_char := v_first_char + 1;
          ELSE
            v_first_char := 0;
          END IF;
        END IF ;
      END IF ;
    END IF ;
  END IF ;

  :new.sample_id := :new.Project
                    || substr(char_list, v_first_char, 1)
                    || substr(char_list, v_second_char, 1)
                    || substr(char_list, v_third_char, 1)
                    || substr(char_list, v_fourth_char, 1)
                    || substr(char_list, v_fifth_char, 1);

END;
```

Appendix 4: Trigger to generate SAMI samples' IDs

```
CREATE TABLE IARC_SAMPLES_QT_4_SAMI"
   (  "SAMPLE_NAME" VARCHAR2(50 BYTE) NOT NULL ENABLE,
      "PATIENT_NAME" VARCHAR2(50 BYTE),
      "SAMPLE_TYPE" VARCHAR2(50 BYTE),
      "PLATE" VARCHAR2(50 BYTE),
      "PLATE_ROW" VARCHAR2(2 BYTE),
      "PLATE_COLUMN" NUMBER,
      "PLATE_POSITION" NUMBER,
      "INITIAL_CONC" NUMBER,
      "INITIAL_VOL_UL" NUMBER,
      "INITIAL_QTY_UG" NUMBER,
      "CURRENT_VOL_UL" NUMBER,
      "CURRENT_QTY_UG" NUMBER,
      "USERSTAMP" VARCHAR2(20 BYTE),
      "INITIAL_TIMESTAMP" DATE,
      "COMMENTS" VARCHAR2(255 BYTE),
      "STUDY" VARCHAR2(30 BYTE),
      "SAMPLE_ID" NUMBER,
      "NANODROP_CONC" NUMBER,
      "NANODROP_RATIO" NUMBER,
      "INITIAL_PLATE" VARCHAR2(50 BYTE),
      "INITIAL_PLATE_ROW" VARCHAR2(2 BYTE),
      "INITIAL_PLATE_COLUMN" NUMBER,
      "INITIAL_PLATE_POSITION" NUMBER,
      "INITIAL_EXT_ID" VARCHAR2(50 BYTE),
      "BACKUP" VARCHAR2(1 BYTE),
      "POOLED" VARCHAR2(1 BYTE),
      "EXIST" VARCHAR2(1 BYTE),
      "UNIQUE_BARCODE" VARCHAR2(100 BYTE),
      "DNA_SOURCE" VARCHAR2(50 BYTE),
      "UPDATE_TIMESTAMP" DATE,
      "TRANSFER_TIMESTAMP" DATE,
      "PLATED" VARCHAR2(5 BYTE))

- CREATE TABLE "OPS$PRODLIMS"."IARC_SAMPLES_PROJ_MV_4_SAMI"
   (  "PROJECT" VARCHAR2(50 BYTE) NOT NULL ENABLE,
      "TYPE_MV" VARCHAR2(50 BYTE),
      "GLOBAL_QTY" NUMBER,
      "DATE_MV" DATE,
      "USERSTAMP" VARCHAR2(50 BYTE),
      "TIMESTAMP" DATE,
      "COMMENTS" VARCHAR2(255 BYTE)
   )
-  CREATE TABLE "OPS$PRODLIMS"."IARC_SAMPLES_MV_4_SAMI"
   (  "SAMPLE_ID" NUMBER NOT NULL ENABLE,
      "UNIQUE_BARCODE" VARCHAR2(100 BYTE),
      "PROJECT" VARCHAR2(50 BYTE),
      "MV_TYPE" VARCHAR2(50 BYTE),
      "MV_DATE" DATE,
      "MV_VOL_USED" NUMBER,
      "MV_QTY_USED" NUMBER,
      "USERSTAMP" VARCHAR2(50 BYTE),
      "TIMESTAMP" DATE,
      "COMMENTS" VARCHAR2(200 BYTE),
      "CNG_CEPH_BARCODE" VARCHAR2(50 BYTE)
   )
```

Appendix 5: Specific LIMS' tables to manage samples locations and quantities

Titre en français : Développement d'un système informatique intégré pour la gestion des données de laboratoire et des étapes de séquençage de nouvelle génération au sein d'une plateforme de recherche en génomique du cancer.

Résumé en français (max 1700 caractères) :
L'objectif de mon travail de thèse était de développer des outils bio-informatiques permettant d'améliorer la traditionnelle gestion de l'information scientifique au sein d'un grand centre de recherche et en particulier au sein d'une plateforme de génomique.
Trois outils ont été développés: un cahier de laboratoire électronique, un système de gestion de l'information de laboratoire pour des applications de génomique dont le séquençage de nouvelle génération, ainsi qu'un système de gestion des échantillons pour de grandes bio-banques. Ce travail a été réalisé en étroite collaboration avec des biologistes, épidémiologistes et informaticiens. Il a également inclus la mise en place d'interactions entre les différents outils pour former un système informatique intégré.
Les trois outils ont été rapidement adoptés par l'ensemble des scientifiques du centre de recherche et sont désormais utilisés au quotidien pour le suivi de toutes les activités de laboratoire mais aussi plus globalement pour les autres activités scientifiques du centre de recherche. Ces outils sont transposables dans d'autres instituts de recherche.

Titre en anglais : Development of an integrated Information Technology system for management of laboratory data and next-generation sequencing workflows within a cancer genomics research platform

Résumé en anglais :
The aim of my thesis work was to develop bioinformatics tools to improve the traditional scientific information management within a large research centre and especially within a genomics platform.
Three tools have been developed: an electronic laboratory notebook, a laboratory information management system for genomics applications including next generation sequencing, as well as a sample management system for large biobanks. This work has been conducted in close collaboration with biologists, epidemiologists and IT specialists. It has also included the setup of interactions between the different tools to make an integrated IT system.
The three tools have been rapidly adopted by all the scientists of the research centre and are now daily used for the tracking of all the laboratory's activities but also more globally for the research centre's other scientific activities. These tools are transposable in other research institutes.

Discipline : Bioinformatique

Mots-clés en français : Bioinformatique ; Cahier de laboratoire électronique (ELN) ; système informatique de gestion de laboratoire (LIMS) ; séquençage de nouvelle génération (NGS) ; système de gestion d'échantillons pour des biobanques.

Mots-clés en anglais : Bioinformatics; Electronic laboratory notebook (ELN); laboratory information management system (LIMS), next-generation sequencing (NGS); sample management for biobanks.

Intitulé et adresse du laboratoire :
Groupe d'étude des prédispositions génétiques au cancer (GCS) - Centre International de Recherche sur le Cancer (CIRC) – 150 cours Albert Thomas 69003 LYON, France