Design and development of a comprehensive data management platform for cytomics: cytomicsDB

Larios Vargas, E.

Chapter 1

Introduction

In this chapter the thesis scope as well as the current key concepts related to this dissertation are introduced. In addition, the thesis structure is outlined.

1.1 Thesis scope

One of the main challenges in software engineering is to develop systems capable to adapt to the continuos changes in their environment. This problem is present in software systems for biomedical research in cell systems. In biomedical research, the study of cellular systems on large scale is called Cytomics, High-Throughput screening (HTS) is one of the most common techniques in Cytomics for target validation. HTS has to face that challenge due to the diversity and the emergence of new technologies in the context of data, roles, hardware, software tools, techniques, algorithms, protocols, etc. Additionally, the extremely large and growing volumes of data produced in these experiments complicate the exploration, interoperability, understandability, sharing and reusability of the data. Therefore, there is an urgent need to design and develop a comprehensive data management platform for experiments in Cytomics. This thesis focuses on the study of how to organize the data managed in *Cytomics* in an optimal and yet flexible manner, so that these challenges can be tackled while taking care of an appropriate organization of the heterogeneous data and then providing the necessary tools and software to facilitate the extraction of meaningful data from these large repositories.

1.2 HTS experiments workflow

In drug discovery, an HTS workflow is a sequence of operations and activities required to identify starting points for drug design or also called hits. These sequence consist of five stages: (YLL+11): (1) *Experiment Design*, (2) *Wet-lab*, (3) *Image acquisition*, (4) *Image analysis* and (5) *Data analysis*.

1.2.1 Experiment Design

The key component in HTS is the "plate"; this is basically a small container that features a grid of wells and it is considered the principal information carrier of the experiment (c.f. 1.2). A spreadsheet application is used to create a layout of the plate in order to store practical information about the treatment introduced for each well in the cell culture plate. Figure 1.1 displays the design of a 96 well plate; details of the metadata are linked to each well therefore this is also called "plate design". The "plate design" contains the input data for the plate e.g. metadata and allows to link the output of the plate e.g. image files. Figure 1.1 shows a brief description of the plate design. Wells are identified by a row from "A" to "H" and a column from "1" to "12". A colour-coded notation in the plate design is used which facilitates the readability of the basic metadata linked to the experiment. For this experiment the MCF7 cell line is used and per well specific genes are switched off; i.e. knocked down, by the use of siRNA transfection which are represented by the not colour-coded wells. The negative controls are shown in light yellow, the positive control for transfection is in green, and the positive control for knockdown and the noco-assay is in red e.g. Dynamin2. The cells from wells B1 to B6 have been transfected with siGFP (light yellow). Wells B1 and B2 were additionally treated with DMSO while wells B3 and B4 have been treated with Nocodazole.

MF_091004_HP_HK_2nd_DEBHA_211510

WP7	DMSO		Noco		Wash out		DMSO		noco		Wash out	
	1	2	3	4	5	6	7	8	9	10	11	12
A	No siRNA						Control #2					
B	siGFP						MAP4K2_01					
C	MAPK7_01						MAP4K2_02					
D	MAPK7_02						MAP4K2_03					
E	MAPK7_03						MAP4K2_04					
F	MAPK7_04						MAP4K2_pool					
G	MAPK7_pool						Dynamin2					
H	pax		siGlo		Kif 11		Dharmafect IV					

Figure 1.1: *Design of a 96 wells plate for an experiment using MCF7 wt cells*

1.2.2 Wet-lab

Based on the design stage, one or more plates are used during the experiment. During the experiment, the *in vitro* cell lines are prepared and put into a culture well of the

plate (c.f. 1.2). Plates for HTS can have either 24, 96, 384 wells, etc. i.e. multiples of 24. These wells will contain, depending on the type of experiment, culture medium and depending on the layout/design, some chemical compounds. These wells will also contain cell populations, and designed treatments are induced into them (YLL+11). Upon completion of the wet-lab procedure, the plate is prepared for the further readout during the image acquisition stage.

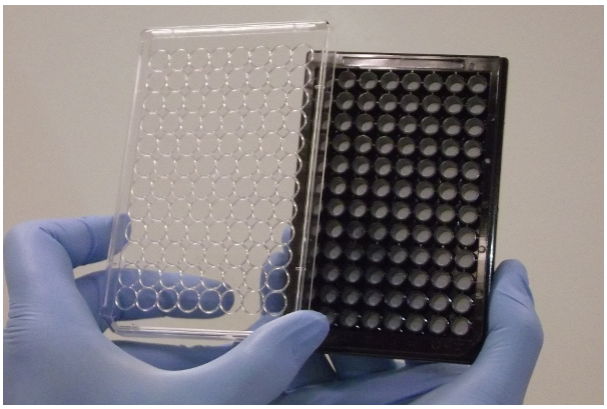


Figure 1.2: *96 wells plate for HTS.*
(Alv)

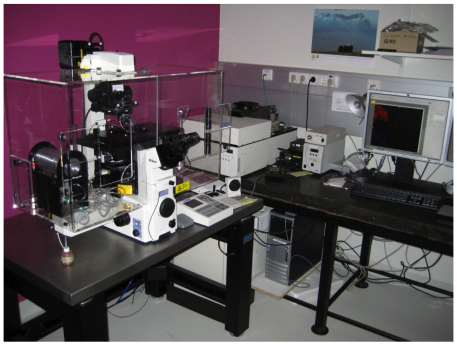
1.2.3 Image acquisition

In biomedical laboratories images are usually acquired by measuring photon flux in parallel, using a camera, or sequentially, using a point detector and equipment that scans the area of interest. However, capturing an image from a camera into the computer is straightforward, in most cases image acquisition needs to be tightly synchronized with other computer-controllable equipment such as shutters, filter wheels, x-y stages, z-axis focus drives and autofocus mechanisms; the controls for this are implemented in software and hardware. This automation is necessary to gather desired multiparametric information from a sample to allow the unattended acquisition of large numbers of images in time-lapse series, z-stacks, multiple spatial locations in a large sample or multiple samples in a large-scale experiment. Dedicated image acquisition software is therefore indispensable to communicate with these various components and coordinate their actions such that the hardware functions as quick and flawless as possible as well as allowing the researcher to easily execute the following sequence of acquisition events (EBG+12), i.e.:

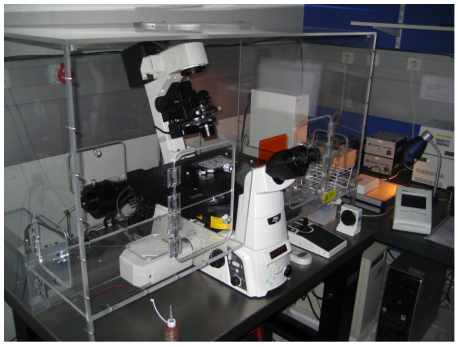
1. Manual or automated acquisition.
2. Single time point or time series.

3. Single focal plane or three-dimensional stack.
4. Single channel, multiple channels or hyperspectral.
5. Acquisition protocol is predefined or determined on the fly based on analysis of image data.

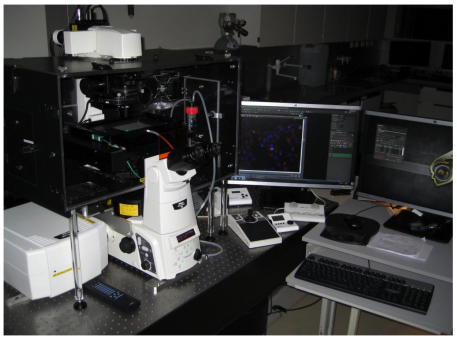
Figure 1.3 shows four different microscopes commonly used in Hight-Throughput Screening.



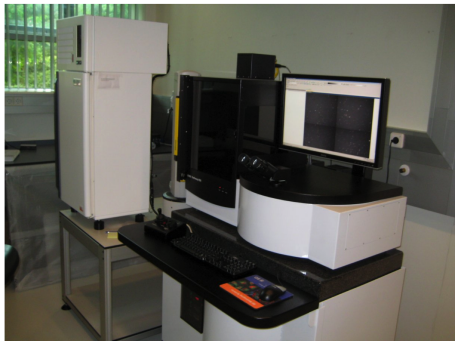
(a) Nikon 1



(b) Nikon 2



(c) Nikon 3



(d) BD Pathway

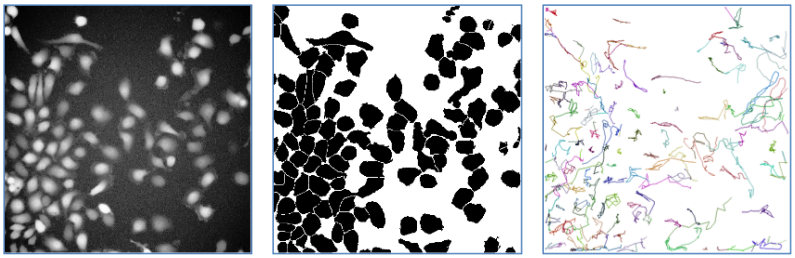
Figure 1.3: *Microscopes for HTS.*
(Tox)

1.2.4 Image analysis

Eliceiri et. al. (EBG+12) highlight the fact that biologists are increasingly interested in using image analysis to convert microscopy images into quantitative data (GM07) (LC09). Image analysis, in particular, is a necessary step for experiments in which hundreds or

thousands of images are collected by automated microscopy, whether for screening multiple samples, collecting time or z-series data, or other technologies that generate vast volumes of image data. In addition to image analysis in a high-throughput context, image analysis is important for many biological studies: for example, quantifying the amount and localization of a signaling protein, measuring changes in structures over time, tracking invading cancer cells or looking at nonspatial data such as fluorescence-lifetime data (Lak99). Image analysis facilitates data reduction and can help to ensure that results are accurate, objective and reproducible.

For instance, during the image analysis process, raw images from the image acquisition (c.f. 1.4(a)) stage are converted to auxiliary image results (c.f. 1.4(b) and 1.4(c)). Quality enhancing filters and segmentation algorithms are executed to extract region of interests (ROIs). Moreover, other types of processing based on ROIs such as object tracking are also applied. For each image sequence, phenotypic measurements are gathered from ROIs and trajectories. Trajectories are not available for static images (YLL+11).



(a) Original image (b) Binary masks of (a) (c) Trajectories of (a)

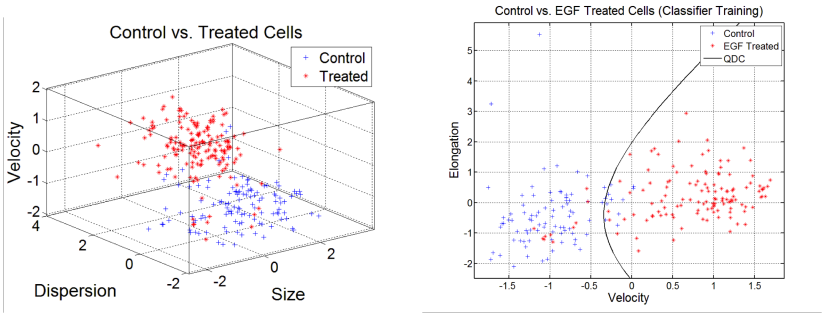
Figure 1.4: *Segmentation and subsequent image analysis* (YLL+11)

1.2.5 Data analysis

Eliceiri et. al. (EBG+12) conclude that the interpretation of the results from a high-content imaging application requires proper presentation and visualization of the image data. Many data display options are currently available according to the type of microscope used. Image analysis results can be displayed graphically as heat maps, line graphs, bar charts, dose response curves, or scatter plots e.g. (c.f. 1.5(a)) and 1.5(b)). In addition, to interpret data within a plate or set of plates at a glance, thumbnail images of the entire plate can also be displayed.

The purpose of data analysis varies from experiment to experiment. One basic purpose is to identify "hits", meaning to identify genes or compounds which play a functional role in the cellular process that is under investigation (genes), or may positively or negatively affect the process (compounds). This is determined by comparing

each sample with control treatments that are carried out under the same condition but induce no change in the celullar process. Before "hits" are identified, quality control and data normalization are performed to remove systematic errors and to allow comparison and combination of samples from different plates (Di13).



(a) Cell migration (b) Structure dynamics

Figure 1.5: *Data analysis* (YLL+11)

1.3 Data management in Cytomics

Finding "hits" contributes to identify targets for drugs. *Drug discovery* has a fundamental intrinsic dedication in increasing the number of novel medicines. However, in order to accomplish this goal, high costs are involved from discovery through clinical trials to approval (SKE07). Thus, the requirements for a more sophisticated IT infrastructure used to assist the drug development process has been exponentially increased during the last decade. *Cytomics* studies introduce information related to the behaviour of the cell systems, exploiting this information is a key factor for the drug development process. High-Throughput Screening (HTS) is one of the most common techniques used in *Cytomics*. In HTS the image dataset size per plate is approximately between 20Gb to 40Gb and depending on the type of experiment performed multiple plates can be used in one experiment and the number of images generated can reach around 100000 images per plate. The large volume of data generated in these experiments make its management a challenge and a necessity.

The current data management challenges in *Cytomics* can be summarized as following categories: (1) *Heterogeneity of the data*, (2) *Data exploration*, (3) *Interoperability*, (4) *Data sharing and reusability*, (5) *Understandability* and (6) *Transport of data*. In the next paragraphs we will discuss the categories.

1.3.1 Heterogeneity of data

In HTS experiments the diversity of data is one of the main issues to be tackled for the cytomics-based systems. The variety in the data is a consequence of the diversity of harware and software components involved in HTS (c.f.1.6). A summary of these components are:

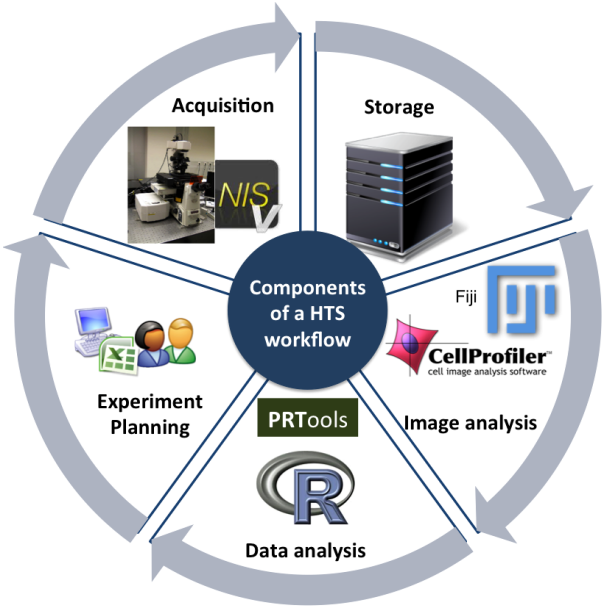


Figure 1.6: *Hardware and Software components in an HTS workflow*

Types of microscope. In modern biology, one of the most basic tools is the visual inspection of cells using a microscope and in modern optical microscopy for HTS is commonly used the following types of microscope: (1) Confocal microscopes, (2) Multiphoton microscopes (3) Multispectral microscopes and (4) Fluorescence microscopes. Each one managing different type of variables and parameters.

Image formats. The different types of microscopes have different file formats associated with them. This means that metadata is sometimes only available in propietary format. Consequently the image files generated will use a propietary format as well. However, these propietary software also includes plugins or tools for exporting or visualizing the images using more standarized format conventions e.g. tiff files.

Image and data analysis tools. There is a large number of open source tools and inhouse applications developed for image and data analysis in HTS. Depending on the programming language for the development and operating systems supported, it is necessary to create additional scripts or import data for further processing according to the file types supported by those tools. The output can also vary according to the

analysis performed, for instance plain text files or binary data.

Spreadsheet applications. Spreadsheet applications are still very commonly used by biologist in the HTS workflow for bookkeping metadata related to the design of the experiment. The use of this type of applications have many drawbacks such as: susceptible to errors, difficult to maintain, limited security and accesibility, not shareable, and not suitable for querying.

Legacy systems. In the research groups, there are usually other systems which play a role in the management of laboratory data or other repositories which store logistic information of chemicals, compounds, or other components required in the lab. These systems work most of the time independent of the HTS workflow but the information contained needs to be synchronized with the experiment pipeline.

1.3.2 Data discovery

The drug discovery and development process depend on informed decision making. A Cytomics platform needs to increase the quality and pertinence of information for that decision-making process (SKE07). Therefore, identifying the location of meaningful data in the large datasets generated in HTS experiments is the most important challenge.

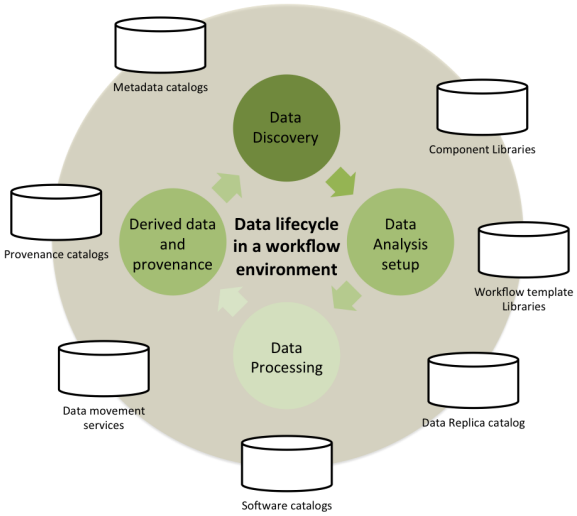


Figure 1.7: *Data Lifecycle in a workflow* (DC08)

E-Science represents the increasing global collaborations of people and shared resources that will be needed to solve the new problems in science and engineering. These e-Science problems range from the simulation of whole engineering or biological

systems, to research in bioinformatics, proteomics and pharmacogenetics (HT03). Hey et. al. address the imminent flood of scientific data expected from the next generation of experiments and the critical needs for analytical tools capable of exploiting and automating the process of going from raw data to information to knowledge.

Deelman et. al. (DC08) present from the point of view of data, the lifecycle in a workflow for e-Science expressed in four transformations. This includes the following transformations (c.f. 1.7):

1. Data discovery.
2. Setting up the data processing pipeline.
3. Generation of derived data.
4. Archiving of derived data and its provenance.

Data analysis is often a collaborative process or is conducted within the context of a scientific collaboration. Data discovery is highlighted due to the fact that scientists in a collaboration frequently submit workflows to process datasets and derive scientific knowledge. These collaborators may submit related workflows and build upon earlier work by other scientists. Thus, scientists need to be able to discover information about workflows that have been executed in the past, identify datasets of interest, and locate analysis code and workflow templates.

1.3.3 Interoperability

Interoperability represents the posibility to exchange information between repositories. Additionally, interoperability is a domain specific concept and the heterogeneity of ICT implementations is such that there exist different solution spaces depending on the combination of existing systems and in many cases such solutions are not directly transferable to other cases (Kos06). According to Chen and Doumeingts (CDV08), interoperability has the meaning of coexistence, and autonomy. If systems are interoperable, this mean that their components are connected by a communication network and can interact; they can exchange services while continuing locally their own logics of operation.

Each of the various types of data handled in the drug discovery process poses its own specific problems for integration into the information systems and decision-making processes (SKE07). The role of standards is thereby becoming crucial. Standarization processes should be present in each stage of the scientific data lifecycle.

The development and wide adoption of common standards is extremely difficult despite the fact that they represent the most effective tool to deal with interoperability. In HTS, one of the main challenges is the standarization of the metadata i.e. naming conventions in order to facilitate the validation, verification, exchange with external platforms and the exploration of annotated information. Moreover, it is necessary to

create standards to represent scientific data such as data models and standards for query languages.

Developing comprehensive approaches is complex because data interoperability is a problem that goes beyond technical aspects. Data interoperability approaches, in order to be complete, should reconcile all the differences arising between data providers and data consumers with respect to organizational, semantic, and technical characteristics governing their "exchange" of data. The majority of existing solutions focus on technical aspects only with very limited efforts and guidelines being oriented to reconcile differences at the organizational level, probably because this domain presents more complexities than others do (PCC13).

1.3.4 Understandability

In e-Science, the capability of the user to understand the information or knowledge contained in the large repositories generated by each experiment is becoming a critical issue. An HTS dataset is very complex. Therefore there is a huge challenge in organizing all this information so that it remains interpretable. With respect to the mass of data produced, we do not believe that a solution is solely linking together databases containing all the screening, laboratory testing, and chemical information. The scientists whose collaborative work build the "road to the lead" need to efficiently communicate. Only clear-cut analyses at each stage can optimize the process and facilitate such communication (Hey02).

The approach proposed by Heyse et. al. (Hey02) consist of: (1) perform thorough quality control to include only reliable data in further processing stages, (2) standardize, process, condense, and annotate data at each stage as much as needed by the respective experts involved, (3) keep data related to its context using metadata on assays, analyses, and compounds, and (4) provide tools to analyze and interpret large datasets following statistical and biophysical models and standard processes. This makes the overall screening process transparent and traceable.

1.3.5 Data sharing and reusability

Demchenko et. al. (DGDLM13) emphasize that the emergence of computer aided research methods is transforming the way how research is done and scientific data are used. Four types of scientific data are defined:

- Raw data collected from observations and from experiments according to an initial research method.
- Structured data and datasets that have gone through data filtering and processing.
- Published data that support one or another scientific hypothesis, research result or statement.

- Data linked to publications to support the wide research consolidation, integration and openness.

Collaboration is possible thanks to the contribution of new results as a consequence of the validation performed by scientists after scientific data has been published. Collecting information about the processes involved in the transformation of raw data to published data becomes an extremely important aspect of scientific data management.

There is a close link between interoperability and collaboration. The interoperability achieved by the use of standards for the metadata in HTS experiments, facilitates the understandability, discoverability and thus promotes the collaboration among scientists. Reseachers will be able to improve their contributions thanks to the use of domain-specific metadata information standards.

Reusability of published data within the scientific community is still another aspect to take into consideration. Understandability of the semantics of published data is a key factor for reusability and this has been done manually. However, given the volume of the data, this becomes, so far, impractical in the scale of Big Data science.

1.3.6 Transport of data

Kaisler et. al. (KAEM13) describe the storage and transport issues for big data. Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, the current systems do not allow to directly attach the requisite number of disks. Access to those data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 116 days, if we assume that a sustained transfer could be maintained. It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it.

Two solutions manifest themselves. First, process the data “in place” and transmit only the resulting information. In other words, “bring the code to the data", vs. the traditional method of “bring the data to the code." Second, perform triage on the data and transmit only that data which is critical to the downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data, this will ensure traceability and validation of the data before the processing stage.

Cytomics based platforms must take special care to the six categories of the current data management challenges in Cytomics in order to facilitate the interaction of researchers with their data. This means that platforms should ensure that research data is FAIR (Findable, Accessible, Interoperable and Reusable).

1.4 LIM Systems

A typical approach to the management of large volume of data in a laboratory is a Laboratory Information Management System (LIMS). LIMS is a computer application designed for the analytical laboratory that is designed to administer samples, acquire and manipulate data, and report results via a database. It automates the process of sampling, analysis and reporting (Mah91) (DAJ12).

A more complete definition is provided by Skobelev et. al. (SZK+11), Laboratory information management systems belong to the class of application software intended for storage and management of information obtained in the course of the work of the laboratory. The systems are used to control and manage samples, standards, test results, reports, laboratory staff, instruments, and work flow automation. Integration of laboratory information management systems with the enterprise's information systems will make it possible to promptly transmit required data to the laboratory and the enterprise administration.

High-throughput screening (HTS) due to the overwhelming amounts of data is pushing forward the need for sophisticated IT infrastructure in order to enhance the laboratory performance. LIM systems have been designed as a first step to face this dilemma. However, the continuous progress in tools, hardware, and heterogeneity of the components involved in HTS make still difficult to completely tackle this problem.

Under the term Laboratory Information Management System, industry-related IT solutions for research, development, service, and production in the fields of chemistry, biology, environmental protection, or medicine are combined. The interdisciplinary object of the life sciences meanwhile characterizes a broad, large class of laboratory information management systems and offers a potential for new general system concepts in laboratory automation before the background of innovative Internet technologies (TGDS04) (AMF00).

Figure 1.9 (DAJ12) describes how LIMS assist laboratories in tracking and managing its information resources, particularly the data that represents the laboratory product.

LIMS has contributed in improving the administration of labs data from the perpective of a production environment. However, with respect to the research environment, LIMS has limitations such as: (1) dealing with unstructured data, (2) ensure interoperability between components in an HTS environment, (3) facilitate experiment data sharing and (4) promote reusability.

1.4.1 LIMS environment

A LIMS is more complex than just a single application and it is better to consider the term environment to specify the set of elements who interact within a laboratory. These elements are shown in Figure 1.8 (DAJ12), i.e:

1. LIMS application.

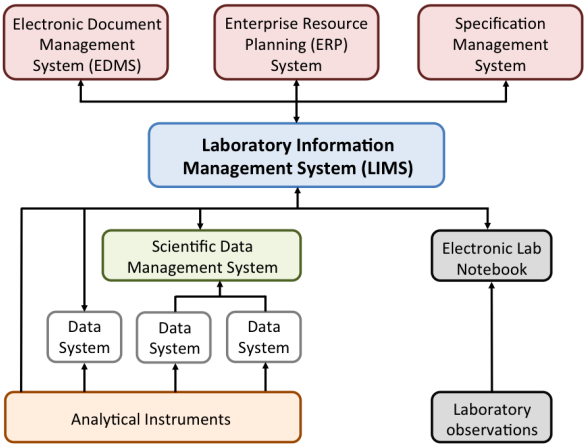


Figure 1.8: *Diagram of a tyical LIMS environment* (DAJ12)

2. Data Analysis tools interfaced directly with the LIMS.

3. Legacy systems and computer systems interfaced with the LIMS.

4. Applications external to the laboratory that are also interfaced to the LIMS e.g. an enterprise resource planning system.

1.4.2 How a typical LIMS works

The LIMS is in charge of the analysis and distribution of the data generated in the laboratory. It becomes the single point of interaction between the data and the different tools and systems used in the laboratory. The data is stored in a single repository and this allows to centralize its flow and access through the organization (DAJ12).

Once the sample is logged into the LIMS package, it automatically takes over the further operations and generates the work sheet/protocol sheet/analytical work record. The workflow contains all the diagnostics required to be carried out for the sample (Figure 1.9).

While carrying out the actual testing, the readings and observations are noted down in the workflow. After completing the diagnostics all data are fed into the LIMS package. It automatically calculates the results, compares the findings with the standards, and also decides about the compliance with the standards. After this it prints out the certificate of analysis in the prescribed format that is already fed into the software.

It is the LIMS that decides about the conformance and non-conformance of the sample with the standard and not the analyst. The technician just has to enter the

required inputs (readings/observations); the desired output format can be selected, and results can be generated in desired options without giving any external formulae for calculation (KGG10).

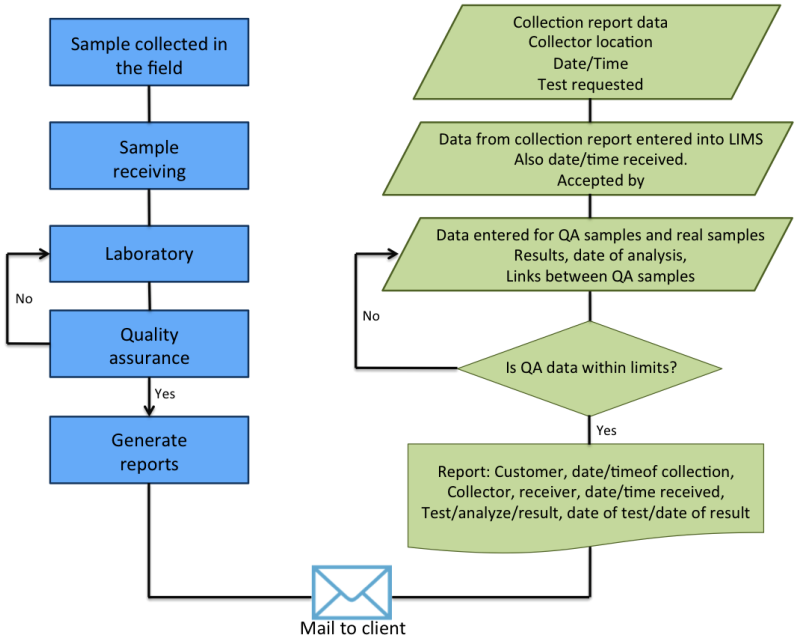


Figure 1.9: *Sample work flow diagram* (DAJ12)

1.4.3 Benefits of LIMS

A LIMS provides benefits for many of the users of a laboratory. However, a LIMS does represent an expense that must be considered. This expense will almost certainly have to be justified by a level of higher management. The following is a brief outline of several of the main benefits identified and realized from current users of LIMS (DAJ12).

1. Information can be obtained with the click of a button rather than having to dig through files.
2. Years of data can be kept easily without the need for traditional archiving.
3. The improvement of business efficiency.
4. Improvement of data quality (all the instruments are integrated).
5. Automated log-in, tracking, and management.

6. Automated customer reports (Turnaround Time, Work Load).
7. Automated Integration of Hand-held LIMS devices.
8. Automated Quality Control.
9. Daily Quality Reports.
10. Easily accessible data via the web.

LIMS is a useful tool to manage structured information produced in the laboratory. However, an HTS environment is constantly changing due to the diversity of technologies involved which produce continuosly more volume of data. Current datasets generated in HTS are commonly unstructured and not suitable for being stored in a rigid database system. The need to align research data to the FAIR paradigm force the urgent need to design new software architectures which are able to integrate the benefits of LIMS with platforms capable to manage unstructured data in Cytomics.

1.5 Thesis structure

This thesis consists of six chapters: (1) *Introduction*, (2) *CytomicsDB architecture*, (3) *Metadata management*, (4) *Metadata Validation*, (5) *Cluster Integration* and (6) *Conclusion and Discussion*.

Chapter 1 introduces the scope of thesis as well as its background. In addition, here we review the existing challenges in the management of large volumes of data in Cytomics. Thereafter, the lab management systems that currently manage lab data and their issues for dealing with high-throughput data is briefly described. Moreover, the typical workflow in High-Throughput Screening (HTS) experiments is presented.

Chapter 2 describes our research in developing CytomicsDB, a modern RDBMS based platform, designed to provide an architecture capable of dealing with the computational requirements involved in high-throughput content.

Chapter 3 introduces a semantic approach for organizing the metadata involved in High-Throughput Screening experiments. The main goal is to facilitate the exploration process in the HTS workflow, scientists are aware of semantics and they are pushing forward the need for new approaches in organizing the metadata according to which queries are mostly applied on the scientific data.

Chapter 4 provides a semantic-based metadata validation approach applied in CytomicsDB. The objective is to ensure the integrity, consistency and reliability of the data stored in the platform. This is a critical requirement for image and data analysis and further data exploration of the experiment's results.

Chapter 5 discusses how the integration with a computational cluster to CytomicsDB architecture can speed up the image analysis stage in the workflow of an HTS experiments.

Finally, **Chapter 6** provides the general conclusions and recommends future directions of research.