

# Large Scale Data Handling in Biology

Karol Kozak



Karol Kozak

# Large Scale Data Handling in Biology

---

Large Scale Data Handling in Biology

1<sup>st</sup> edition

© 2014 Karol Kozak & [bookboon.com](http://bookboon.com)

ISBN 978-87-7681-555-4

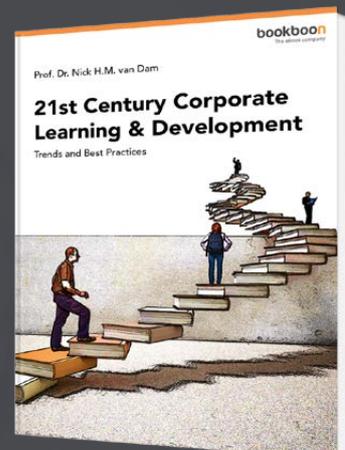
# Contents

	<b>Large Scale Data Handling in Biology</b>	<b>6</b>
<b>1</b>	<b>What to Do with All the Data?</b>	<b>7</b>
<b>2</b>	<b>Data Storage, Backup and Archiving Architecture</b>	<b>9</b>
2.1	Organization of HCS Informatics Infrastructure	9
2.2	Hardware and Network Infrastructure	10
2.3	Do we need robust Data Movers (DM) in High Content Screening for data-flow automation?	13
<b>3</b>	<b>Workflow Systems</b>	<b>19</b>
3.1	Why is a workflow system important?	22
3.2	Visualization in workflow systems	23
3.3	Architecture of workflow systems	24
3.4	Public Domain Workflow Systems	27
3.5	Commercial Workflow Systems	28
3.6	Summary and Vision	34

## Free eBook on Learning & Development

By the Chief Learning Officer of McKinsey

[Download Now](#)



<b>4</b>	<b>Database Development: Laboratory Information Management Systems and Public Databases</b>	<b>35</b>
4.1	What Type of HCS Data Have to Be Managed in the Database?	35
4.2	Database Schema	39
4.3	LIMS Architecture	42
4.4	LIMS and User Management System	43
4.5	Type of Users	44
4.6	Integration and Public Databases	49
<b>5</b>	<b>References</b>	<b>54</b>



www.sylvania.com

We do not reinvent the wheel we reinvent light.

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM

**OSRAM SYLVANIA** 

# Large Scale Data Handling in Biology

**Karol Kozak**

Data Handling in Biology – the application of computational and analytical methods to biological problems – is a rapidly evolving scientific discipline. Written in a clear, engaging style, Large Scale Data Handling in Biology is for scientists and students who are learning computational approaches to biology. The book covers the data storage system, computational approaches to biological problems, an introduction to workflow systems, data mining, data visualization, and tips for tailoring existing data analysis software to individual research needs.

## **Chapters:**

1. What to do with all the data?
2. Data Storage, Backup and Archiving Architecture
3. Workflow Systems
4. Database Development: Laboratory Information Management Systems and Public Databases

## **Abstract**

“High-throughput” in High Content Screening is relative: although instruments that acquire in the range of 100 000 images per day are already marketed, this is still not comparable to the throughput of classical High Throughput Screening. Assays get more and more complex, consequently assay development times become prolonged. Further, standardization of cell culture conditions is a major challenge. Informatics technologies are required to transform HCS data and images into useful information and then into knowledge to drive decision making in an efficient and cost effective manner. Major investments have to be made to gather a critical mass of instrumentation, image analysis tools and IT infrastructure. The data load per run of a screen may easily go beyond the one Terabyte border, and the processing of the hundreds of thousands of images applying complex image analysis software and algorithms requires an extraordinarily powerful IT infrastructure. This chapter will give an overview of the considerations that should be kept in mind while setting-up the informatics infrastructure to implement and successfully run large-scale high-content experiments. In this chapter we describe some of the challenges of harnessing the huge and growing volumes of HCS data, and provide insight to help toward implementing or selecting, utilizing a high content informatics solution to meet organization’s needs and give an overview of informatics tools and technologies for HCS.

# 1 What to Do with All the Data?

High-content screening can easily generate more than one Terabyte in primary images and metadata per run, that have to be stored and organized, which means an appropriate laboratory information management system (LIMS) has to be established. The LIMS must be able to collect, collate and integrate the data stream to allow at least searching and rapid evaluation of the data. After image acquisition and data transfer, image analysis will be run to extract the metadata. Further evaluation includes testing for process errors. Heat maps along with pattern recognition algorithms help to identify artefacts such as edge-effects, uneven pipetting, or simply to exclude images that are not in focus. All plates should be checked so that the selected positive and negative controls exhibit values in a pre-defined range. Further, data may be normalized against controls before further statistical analysis is run to identify putative hits. Known proteins of the pathway being screened should score, and are a good internal control for the accuracy of the assay and workflow. Hits have to be verified by going back to the original images. Further, results have to be compared between independent runs. After this, an appropriate hit verification strategy has to be applied as discussed above. Target gene expression should be confirmed, for example, by running a microarray analysis of gene expression for the given cell line. Finally, data will be compared to other internal and external data sources. Cluster analysis will assist in identifying networks and correlations.

A critical aspect of high content screening is the informatics and data management solution that the user needs to implement to process and store the images. Typically multiple images are collected per microplate well at different magnifications and processed with pre-optimised algorithms (these are the software routines that analyse images, recognize patterns and extract measurements relevant to the biological application, enabling the automated quantitative comparison and ranking of compound effects) to derive numerical data on multiple parameters. This allows for the quantification of detailed cellular measurements that underlie the phenotype observed. From an image analysis perspective the following should not be overlooked when reviewing vendor offerings: the breadth of biology covered; how the software is delivered, does it run quickly, or open a script; is analysis done on-the-fly or offline; have the algorithms been fully validated with biology; the ease of exporting image files to other software packages; and access to new algorithms, is the user dependent on the supplier or is it relatively easy to develop your own or adapt existing algorithms?

The key theme and piece of information repeated throughout this chapter is “partnering”. Scientific research and informatics must work together for the mutual benefit of screening like the drug discovery process. To really be part of the winning team in any organization, all areas must bring their collective expertise together and make the extra effort to understand one another and defer where there is lack of knowledge to those on the team with the experience and expertise or to seek external advises. It is necessary to start off by setting the stage concerning where laboratory computing, which includes the data management (we will discuss a bit later in the chapter), has progressed in order to gain the necessary understanding of where it currently is and where we anticipate it will be going in the HCS area in the future.

A goal of this chapter is to provide an overview of the key aspects of informatics tools and technologies needed for HCS, including characteristics of HCS data; data models/structures for storing HCS data; HCS informatics system architectures, data management approaches, hardware and network considerations, visualization, data mining technologies, and integrating HCS data with other data and systems.

## 2 Data Storage, Backup and Archiving Architecture

HCS systems scan a multiwell plate with cells or cellular components in each well, acquire multiple images of cells, and extract multiple features (or measurements) relevant to the biological application, resulting in a large quantity of data and images. The amount of data and images generated from a single microtiter plate can range from hundreds of megabytes (MB) to multiple gigabytes (GB). One large-scale HCS experiment, often resulting in billions of features and millions of images that needs multiple terabytes (TB) of storage space. High content informatics tools and infrastructure is needed to manage the large volume of HCS data and images.

### 2.1 Organization of HCS Informatics Infrastructure

There are many rules that are common for the image based HCS informatics infrastructure in academic or non academic organization. Answering the following questions analyzed by entire organization tells one exactly which strategy and organization setup has to be taken and what type of work has to assign to experts and researchers. In choosing the strategy and organization setup one needs to answer the following questions:

- Is the required analysis software available off-the-shelf or must it be written in-house? This decision has to be taken in collaboration between IT and scientists, based on the defined requirements.
- What kind of data will be acquired (how many screens in year)?
- How is the data stored, managed, and protected for short-, medium-, and long-term use?
- What type of desktop clusters and servers are required for HCS computing? (brand, type, speed, and memory)
- How do the computer systems interface with the necessary data collection instrumentation and connect to the network and servers at the same time?
- Can allowances and accommodations be made for external collaborations and programs shared among scientists?
- Are we interested in setup a safety buffered zone outside of our firewalls to allow this external data exchange?

After analysis of those questions one would think to have dedicated IT person from IT department working together with the scientists to allow IT professionals to take over responsibility for informatics tasks. The side-by-side person would allow the informatics organization to understand needs of HCS unit. For example the servers processes could be placed inside of HCS pipeline or infrastructure and not be placed as usual and forced to add extra steps to the workflow. It is also important to decide what will be operated by informatics department and what by HCS unit within organization. It makes better sense for informatics department to own, operate, and manage a data center because they have overview on this and they can provide the service for the researchers also. Some advantages of placing the data center in a central IT department:

- Physical data center
- Facility management (electricity, air conditioning, fire protection, physical security).
- Networking Infrastructure (rules for sockets, switches)
- Security infrastructure
- Existing backup and recovery mechanism.
- Standards for PCs and peripherals, servers, desktop applications, middleware applications, web standards.
- Investment in informatics infrastructure elements.

All those items allows HCS unit to take advantage also of economies of scale.

## 2.2 Hardware and Network Infrastructure

It is obvious that the High Content Screening units have very specialized computing needs, which are very different from other facilities, research labs of the academic institute or companies. Research laboratories which are using an HCS unit are not able to use off-the-shelf items in most cases because of their specific requirements. They have various instruments so there are requirements for special hardware to connect to instruments and special software for collecting, manipulating, synthesizing and managing the data coming from those instruments.

There are a wide variety of ever evolving options for server hardware, storage hardware, and networking capabilities. The number of HCS instruments, number of users, the number of sites, and the network bandwidth within a site (i.e., Local Area Network), are a few of the key factors impacting the hardware requirements for an informatics solution. Sizing and scoping the optimal hardware for an informatics solution is an area where professional IT support is critical. Nevertheless, each organization is unique in their HCS usage scenarios, which directly impacts the requirements put on an informatics solution. Academic and industrial units have completely different setups. In general, it is best to identify an informatics solution with a system architecture that can scale as the organization's HCS needs development over time. The best way for a new unit is to start with one reader (microscope or scanner) and one robotic platform which will be used in ongoing experiment. The setup for one microscope (reader) has to be organized and connected together in one data flow pipeline (Fig 1). Follow items are demand for pipeline:

- intermediate server (buffer) for data acquisition
- image storage server
- image processing application
- data mining and visualization
- archiving procedure
- databases or laboratory information management systems (LIMS)



Discover the truth at [www.deloitte.ca/careers](http://www.deloitte.ca/careers)

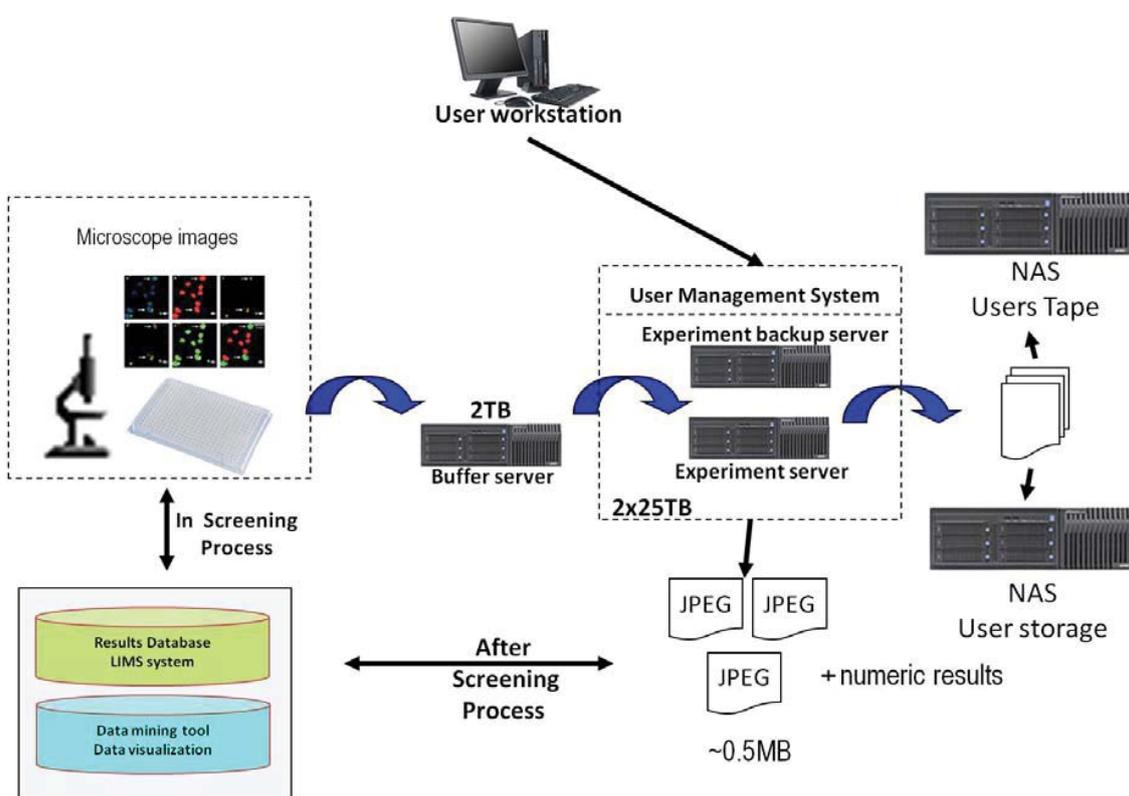
**Deloitte.**

© Deloitte & Touche LLP and affiliated entities.



Click on the ad to read more

Later on, the number of readers may be increased. In non-academic institutions the entire architecture is mostly based on external vendor's informatics applications, so it should be flexible and scalable to fit a variety of hardware configurations and usage scenarios. In academic units the setup is usually less complicated and very often the connection between components is based on a customized development. For both, academic and not academic institution one of the key factors is the bandwidth of the network that connects multiple computers with various operating systems. HCS instruments typically generate data and images at a rate of 1–20 GB (or more in the future) per hour and there are limits to current network and server technology that can support writing this amount of data across networks at multiple sites. The scale of experiments determines the network infrastructure including buffer servers, storage systems and a high speed network. In case of one experiment with one single plate only the microscope local storage space and an external hard drive are required. That is why, balance between networks bandwidth, server, and storage system configurations, and each organization's unique determines how information will be accessed and shared, all need to be taken into account in order to optimize overall system performance.



**Fig. 1.** Example of High Content Screening Informatics system architecture.

To manage a tremendous amount of HCS data collected over a period of experiment, an effective and automated data-flow must be developed. During the data-flow setup, the following questions may arise: who is allowed to view and manage the data, how the data will be backed up, and when are data archived or deleted.

Rules or procedures for storing HCS data need to be determined by each organization. Policies and communication between IT experts must be formulated by organizations. In many cases they just decide to play it safe and store everything. Regarding who can view or manage the data, some forethought must occur just to get the system up and running. This is the most difficult part of the entire setup.

HCS unit operation teams and users usually ask the following questions: Are my data secured with backup? Do I have full and easy access to my data from my workstation? Can we protect our data against loss? Therefore having access to professional IT personnel with experience in assigning and managing permissions is extremely important. Managing permissions is also a key point that reveals why user management (UM) applications are so important. The most powerful and commonly used is a UNIX user management system including groups, users where read/write permission set on folder level. Thanks to this type of UM architecture users can be assigned permissions to the file storage. UM also has to be organized to access relational databases which are used as storage for LIMS applications. UM greatly simplifies deploying and managing the system, especially when trying to share data across multiple sites or different domains. Backing up the data is another area where professional IT support is very valuable. In large organizations a dedicated IT department usually helps with archiving of data. The key feature of a successful HCS backup strategy has is preventing the volume of data that needs to be backed up from growing beyond the manageable range of the backup solution. One of the best approaches to achieve this, is to store the HCS data in different locations based on time. A location's time may then be used to determine whether the data has already been backed up. Once a particular location is no longer having data added, a final backup of this location may be completed. A backup policy depends on internal agreement. The data may be archived based on simple criteria such as creation date, storage location, or creating user based. Other option, if biological metadata like projects, compounds, or hits could drive the archive process, then scientists will need the ability to archive data. Regardless of who actually performs the archiving, coordination among users, knowledge of IT department rules, understanding communication between different IT experts and IT staff is vitally important to effectively manage HCS data.

### 2.3 Do we need robust Data Movers (DM) in High Content Screening for data-flow automation?

Manual data movement between hardware elements of HCS informatics is challenging task. Can a data-flow to be automated? Programs running in the background, that take care of moving file-based raw data produced by readers or any other measurement device to a (remote) central storage are called "Data Movers". Here is a list of data movers program freely available and used in HCS:

- DataMover, ETH Zurich, [http://www.cisd.ethz.ch/software/Data\\_Mover](http://www.cisd.ethz.ch/software/Data_Mover) (Windows, Linux)
- Rsync, available on every Linux/Unix operating system (Linux)
- RoboCopy, SH Soft, <http://www.sh-soft.com> (Windows)
- Allwaysync, Allway Sync, <http://allwaysync.com/> (Windows)

In very simple scenario DM would just copy a single plate/run/screen/experiment directory or can recursively copy a directory and its subdirectories. Such tools classify files by whether they exist in the source directory, in the destination directory, or both. Such tools should classify files by comparing time stamps and file sizes between the source file and the corresponding destination file. Users must be able to specify those copies that are restarted in the event of a failure which saves even more time when your network links are unreliable. In general DM should fulfill the following functionality:

- Selectively copy data files.
- Copy data files with full accuracy
- Deletion of plate files and directories from microscope computer after copying is possible
- Allowing the user to control the number of times the program retries an operation after encountering a recoverable network error.
- Allowing the user to control recovery program handle the state when program retries an operation after encountering network error.
- Usage of plate file names, wildcard characters, paths, or file attributes to include or exclude source files as candidates for copying.
- Allow plate file names, wild card characters, paths or file attribute to be included or excluded in source file as candidates for copying.
- Able to exclude directories by name or by path.
- Allowing the user to schedule DM jobs to run automatically.
- Allowing the user to specify when copying is to be performed.
- Monitor directory tree for changes to detect new plate data
- Report in form of logging mechanism errors appeared during data movement

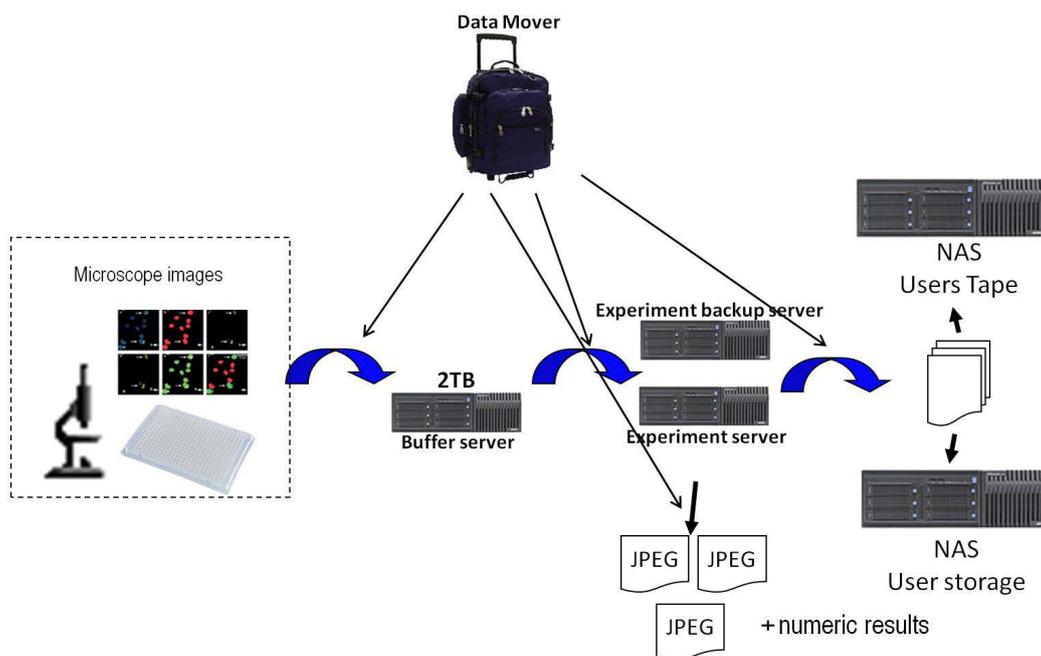


Fig. 2. Illustration of Data Mover function in High Content Screening informatics infrastructure.

A role of DM is illustrated on Figure 2. Usually data movement procedures will be repeated in a configurable period of time. Since the buffer server or central storage is connected via a network, the copy process can create network trouble, i.e.: get terminated or stuck. The DM take cares of these problems to the certain extent and let the user know when the problem persists. What other problems have to be addressed in data movement process and what other options DM provides?

### Protection against disk capacity being exceeded.

For the storage or buffer system, it should be possible to specify a “high-water mark”. A high-water mark is the lowest level of free disk space for a given directory. Once the high-water mark is reached (the available free disk space lies below the specified high-water mark), the administrators should be automatically notified (for example via email) and the DM should stop moving files, waiting until sufficient disk space is available again.

SIMPLY CLEVER

ŠKODA



We will turn your CV into  
an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand?  
We will appreciate and reward both your enthusiasm and talent.  
Send us your CV. You will be surprised where it can take you.

Send us your CV on  
[www.employerforlife.com](http://www.employerforlife.com)



Click on the ad to read more

### **Data completion check**

Data Mover should contain an option which determines if an item (multiwell plate folder) on the microscope local storage has been completed and ready to be moved. This mechanism should start after an incoming file or folder that has not changed during the specified quiet period (see quiet-period option). With the option “data completed timeout” one can specify a time-out (in seconds) for the data movement process. If the script does not finish before the time out it will be killed. Using a “handshake” robustness of the system can increase as it eliminates the need for Data Mover to “guess” when an incoming item is ready to be moved.

### **File Cleansing on Microscope Side**

Data cleansing the feature that the removes certain files before moving experiment images to the storage system or buffer. The rationale behind this feature is that sometimes user cannot prevent the microscope from creating certain files that are not needed but would eat up time and network bandwidth when moving to the central storage. It is quite important to remove these files before moving the directory that contains them to the remote side.

### **Prefixing Incoming Data Sets**

The prefix option allows setting a prefix for all imaged plates that have to be moved to storage system for analysis and later for archiving. Two examples when this would be needed are given:

1. Assuming that a microscope could produce files or directories with the same name (e.g. derived from the same barcode), prefixing for example with the time point will make it unique.
2. If a screening unit has more than one microscope of same type running in parallel, where data moved to the same outgoing directory and one would still know from which microscope the data are derived

### **Extra Local Copy of Produced Images**

If there is a need to access the raw data from an intermediate server (buffer server) or additional storage and the user does not want to transfer these data from the central storage, it is useful to automatically generate an extra temporary copy in a specified a target directory. However it is very useful in extra copy option to use hard links to save disk space, if the file system supports it.

### **Dealing with Failures**

When a plate can not be successfully copied to the storage or buffer system even after a specified number of retries. It is important to have a notification system in which, depending on the logger configuration, will send an alarm to an administrator per email or sms. This failure notification enables timely intervention by an administrator.

### Robustness with Respect to Clock Mismatch

When the microscope computer is located on a different host than the final storage system, there the clocks of the two hosts might be not synchronized. In order to avoid this problem, Data Mover should use an algorithm that ensures that this condition does not lead to a premature transfer and deletion processes (which might even lead to data loss). To this end, the Data Mover should never compare time from the microscope computer and the storage system directly. Instead, the last modification time of a plate (which may be an image file or a directory) is compared to the last modification time of the same item at an earlier time, where the time difference that decides on when to compare last modification times is determined from the storage system clock.

### Robustness with Respect to Program Restarts

What happened if the program will be restarted and at the same during movement of data?? The directory-based communication has to be preferred over a memory-based one, because it is more robust with respect to restarting the program. This is because with the directory-based approach information is kept on the file system instead of the memory. Thus, restarting the program, or even the server, will restart an operation where it was terminated. It has to be ensured, that after restarting the program recovers properly, finishing all operations that were stopped in the middle. Such mechanism is called a *recovery cycle* which is run automatically after each program start.



**e-learning for kids**

- The number 1 MOOC for Primary Education
- Free Digital Learning for Children 5-12
- 15 Million Children Reached

**About e-Learning for Kids** Established in 2004, e-Learning for Kids is a global nonprofit foundation dedicated to fun and free learning on the Internet for children ages 5 - 12 with courses in math, science, language arts, computers, health and environmental skills. Since 2005, more than 15 million children in over 190 countries have benefitted from eLessons provided by EFK! An all-volunteer staff consists of education and e-learning experts and business professionals from around the world committed to making difference. eLearning for Kids is actively seeking funding, volunteers, sponsors and courseware developers; get involved! For more information, please visit [www.e-learningforkids.org](http://www.e-learningforkids.org).

## Conclusion

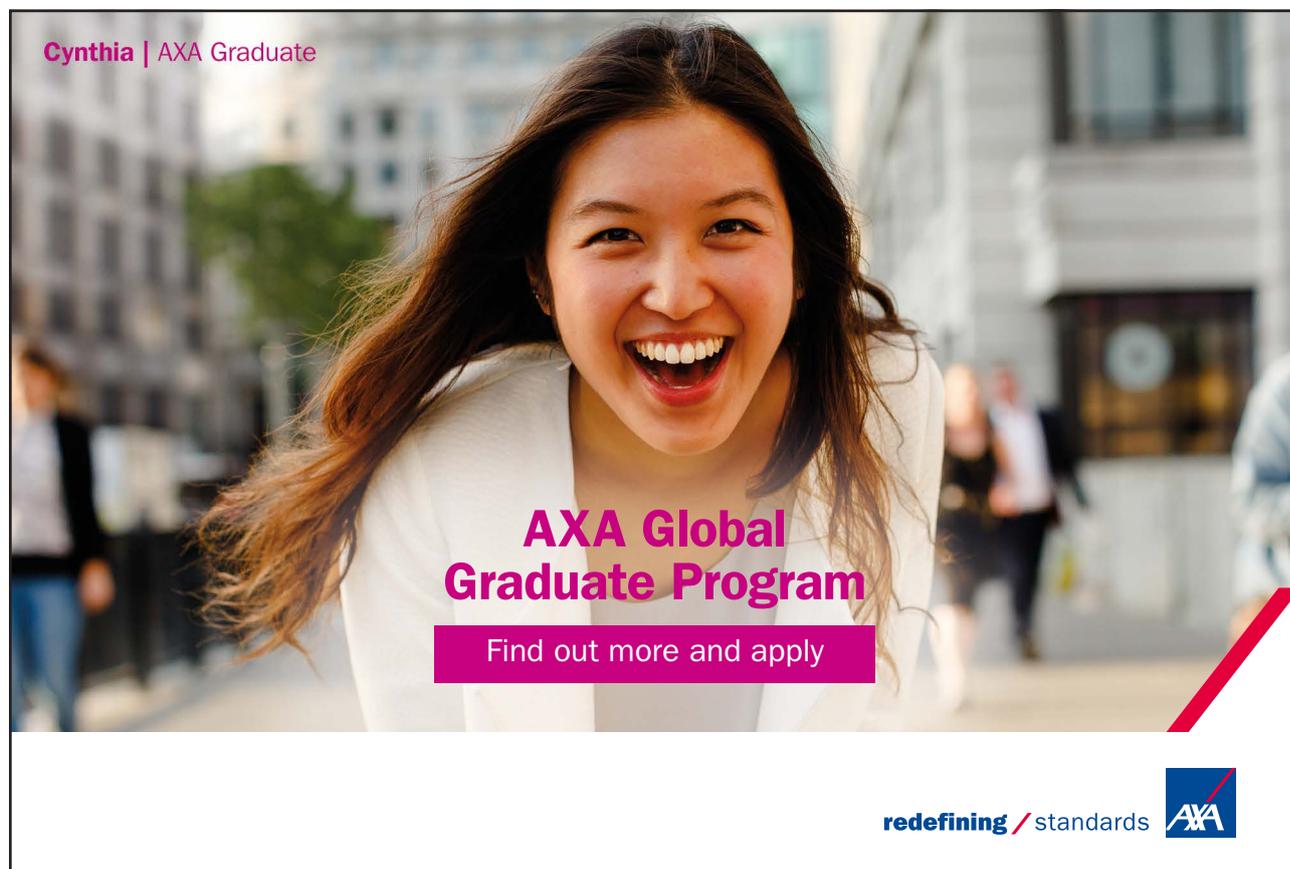
The organization of on HCS informatics infrastructure and management of highly IT skilled personal is the most difficult part of the entire screening operation. For example, informatics people never thought it would be possible to gain the acceptance, respect and ultimate responsibility to take computing away from the scientists. On the other hand, the scientists did not believe that anyone could do as good a job as they had been doing, no matter what level of expertise people had. There are a number of people in the organization that were PhD biologists in HCS units and are now system administrators or Matlab, R, java programmers, IT project managers. Below is a summary of general recommendation for running HCS IT infrastructure:

- Scientific hardware/software suppliers should be partnering with other computer hardware/software vendors
- All storage systems should be located in the corporate data center and monitored by dedicated expert.
- Data storage system should be part of the enterprise storage program used by the entire institution.
- Backups should be done on a scheduled “off hours” basis to minimize work disruption.
- Side-by-side partnership between research informatics and corporate informatics department with a new understanding of requirements and demands in both parts of the organization.
- Placing proper equipment to proper role of HCS unit: super users are equipped with very high-end pc or workstations, normal users, users traveling/working from home are equipped with laptops or notebooks.
- Standardized desktop systems (hardware and software) should be in place including the same application suite for all (i.e., Microsoft Windows XP, Microsoft Office, antivirus software, and Web access and portals).
- Evaluation of availability of computer equipment located on the actual HCS laboratory (flat panel monitors, 100 megabyte or gigabyte network connections).

It should be mentioned that completion of all rules and setting up an IT infrastructure can take approximately 2–3 years to put in place. Setup period can depend on the numbers of involved IT experts and input from HCS researchers. It is very clear that there is a need for different experts in different areas and although scientists are highly skilled and learned, they do not have the specific professional knowledge that people in the computer industry have about making the correct business decisions related to IT.

### 3 Workflow Systems

Within the last few years a large number of tools and softwares dealing with different computational problems related to HCS have been developed. Incorporating third party or new tools into existing frameworks needs a flexible, modular and customizable workflow framework. Workflow (Pipeline) systems could become crucial for enabling HCS researchers doing large scale experiments to deal with this data explosion. The workflow is termed abstract in that it is not yet fully functional but the actual components are in place and in the requisite order. In general, workflow systems concentrate on the creation of abstract process workflows to which data can be applied when the design process is complete. In contrast, workflow systems in the life sciences domain are often based on a data-flow model, due to the data-centric and data-driven nature of many scientific analyses. A comprehensive understanding of biological phenomena can be achieved only through the integration of all available biological information and different data analysis tools and applications. In general, an ideal workflow system in HCS can integrate nearly all standard tools and software. For example, for an HCS using small molecules, the workflow system must be able to integrate different image processing software and data mining toolkits with flexibility. The possibility that any single software covers all possible domains and data models is nearly zero. No one vendor or source can provide all the tools needed by HCS informatics. So it is suggested that one uses specialized tools from specialized sources. Also not all softwares components can be integrated with all workflow systems.



Cynthia | AXA Graduate

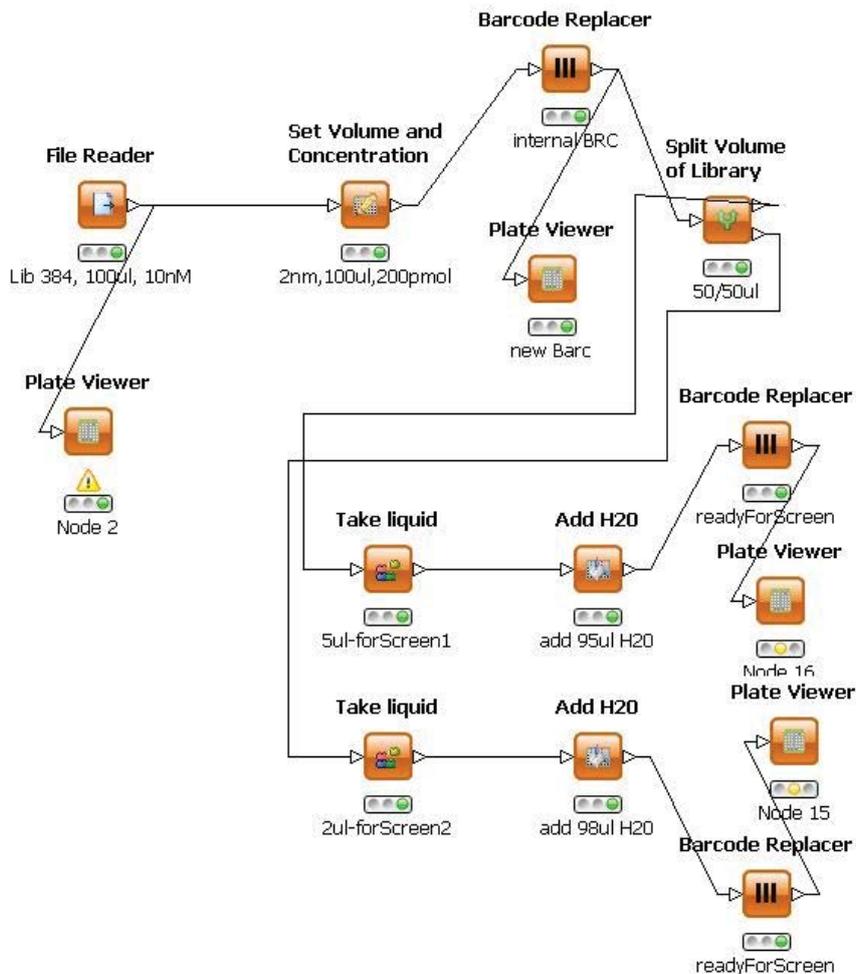
**AXA Global Graduate Program**

Find out more and apply

redefining / standards AXA

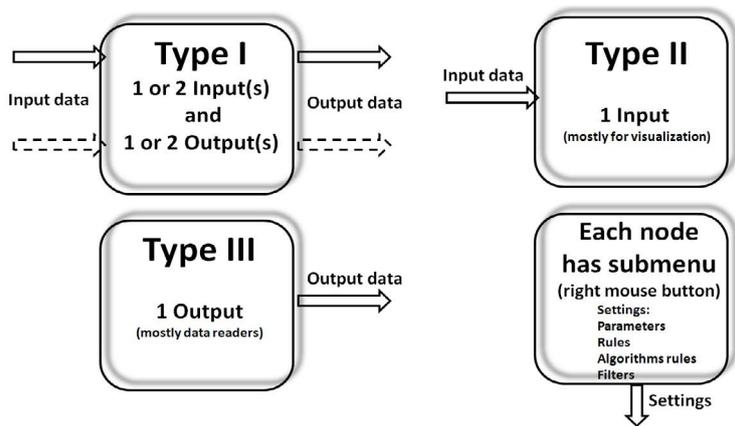
Workflow environment helps also HCS researchers to perform the integration themselves without involving of any programming. A workflow system allows the construction of complex in silico experiments in the form of workflows and data pipelines. Data pipelining is a relatively simple concept. Visual representation of the workflow process logic is generally carried out using a Graphical User Interface where different types of nodes (data transformation point) or software components are available for connection through edges or pipes that define the workflow process. Graphical User Interfaces provide drag and drop utilities for creating an abstract workflow, also known as “visual programming”. The anatomy of a workflow node or component (Fig. 3) is basically defined by three parameters: input metadata, transformation rules, algorithms or user parameters and output metadata. Nodes can be plugged together only if the output of one, previous (set of) node(s) represents the mandatory input requirements of the following node. Thus, the essential description of a node actually comprises only in-and output that are described fully in terms of data types and their semantics. The user can create workflows using any combination of the available tools, readers, writers or database connections in workflow system by dragging/dropping and linking graphical icons. The component properties are best described by the input metadata, output metadata and user defined parameters or transformation rules. The input ports can be constrained to only accept data of a specific type such as those provided by another component. An HCS workflow design is best carried out in phases. In the first phase, a conceptual workflow is generated. A conceptual workflow, as the name suggests, is a sequential arrangement of different components that the user may require to accomplish the given task. It is possible that some of those steps may in turn be composed of several sub components. The next phase converts the conceptual workflow into an abstract workflow by performing a visual drag and drop of the individual components that were figured to be a part of the workflow in the first phase. The workflow is termed abstract in that it is not yet fully functional but the actual components are in place and in the requisite order. In general, workflow systems concentrate on the creation of abstract process workflows to which data can be applied when the design process is complete. HCS screening workflows are based on a dataflow which integrate most of the available, standard software tools (either commercial or public domain) along with different classes of programmable toolkits. As an example, Figure 3 shows a workflow designed to be run by the HCDC-KNIME Workflow Management System (<http://hcdc.ethz.ch>). This workflow is used by HCS facilities. It obtains RNAi from databases, annotates them, make dilutions steps, barcode handling, split volume. In this case, the tasks, also known as steps, nodes, activities, processors or components, represent either the invocation of a remote Web service (the databases), or the execution of a local recalculation. Data-flows along data links from the outputs of a task to the inputs of another, is prepared according to a pre-defined graph topology. The workflow defines how the output produced by one task is to be consumed by a subsequent task, a feature referred to as orchestration of a flow of data.

Any computational component or node has data inputs and data outputs. Data pipelining views these nodes as being connected together by ‘pipes’ through which the data flows (Figure 4).



**Fig 3:** A HCDC-KNIME workflow that simulates the library handling process of multiwell plates including barcode handling

Workflow technology is a generic mechanism to integrate diverse types of available resources (databases, microscopes, servers, software applications and different services) which facilitates data exchange within screening environment. Users without programming skill can easily incorporate and access diverse instruments, image processing tools and produced data to develop their own screening workflow for analysis. In this section, we will discuss the usage of existing workflow systems in HCS and the trends in applications of workflow based systems.



**Figure 4:** General concept of a pipeline node. The component properties are described by the input metadata, output metadata and user defined parameters or transformation rules. The input and output ports can have one or more incoming or outgoing metadata or images.

### 3.1 Why is a workflow system important?

Many free and commercial software packages are now available to analyse HCS data sets using statistical method or classification, although it is still difficult to find a single off-the-shelf software package that answers all the questions of HCS analysis. Statistical open source software packages such as BioConductor ([www.bioconductor.org](http://www.bioconductor.org)) provide large collections of methods suitable for HCS data analysis.

I joined MITAS because  
I wanted **real responsibility**

The Graduate Programme  
for Engineers and Geoscientists  
[www.discovermitas.com](http://www.discovermitas.com)



Real work  
International opportunities  
Three work placements



**Month 16**  
I was a construction supervisor in the North Sea advising and helping foremen solve problems



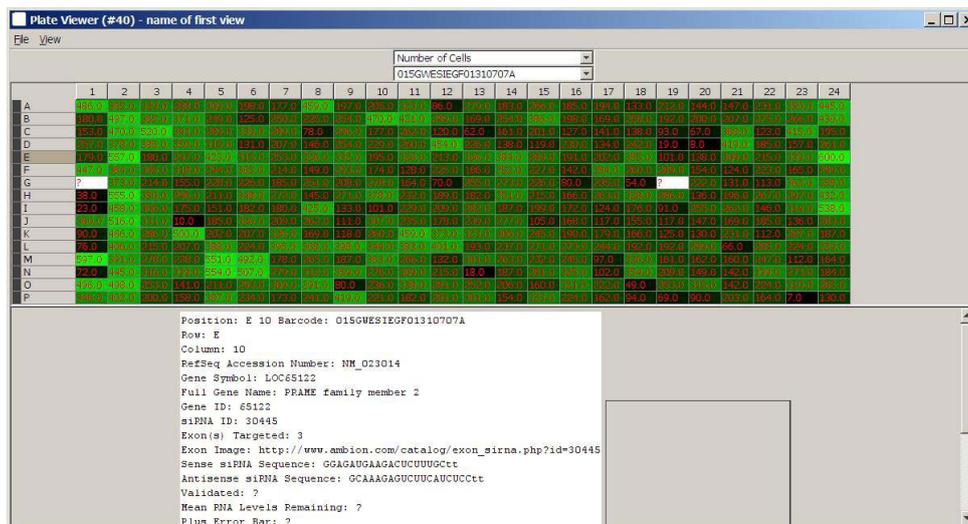


However, their command-line usage can be too demanding for users without adequate computer knowledge. As an alternative, software packages where users can upload their data and receive their processed results are becoming increasingly common: Weka<sup>25</sup>, CellAnalyzer<sup>4</sup>, CellHTS<sup>3</sup>, TreeView<sup>21</sup> have all been published within the last year. Unfortunately, these services often allow only limited freedom in the choice and arrangement of processing steps. Other, more flexible tools, such as Eclipse<sup>6</sup>, KNIME<sup>13</sup>, JOpera<sup>2</sup>, operate either stand-alone or require considerable computer knowledge and extra software to run through the web. In order to make use of the vast variety of data analysis methods around, it is essential that such an environment is easy and intuitive to use, allows for quick and interactive changes to the analysis process and enables the user to visually explore the results. To meet these challenges data pipelining environments have gathered incredible momentum over the past years. These environments allow the user to visually assemble and adapt the analysis flow from standardized building blocks, which are then connected through pipes carrying data or models. An additional advantage of these systems is the intuitive, graphical way to document what has been done.

In a workflow controlled data pipeline, as the data flows, it is transformed and raw data is analyzed to become information and the collected information gives rise to knowledge. The concept of workflow is not new and it has been used by many organizations, over the years, to improve productivity and increase efficiency. A workflow system is highly flexible and can accommodate any changes or updates whenever new or modified data and corresponding analytical tools become available.

### 3.2 Visualization in workflow systems

Visualization tools are one type of workflow systems that provide a quick and effective means to interrogate HCS data and images stored in a secure repository. Users want to view the data, share it with colleagues, and compare results. Visualization tools in workflow system should provide powerful search and navigation tools to rapidly locate plate, well, cell, and image data. Rich search functions should be available to find data based on various metadata and derived data parameters (e.g., user name, dates/times, assay type, features, and so on). The most basic form of any HCS data visualization node should provide interactive tools for reviewing data with drill-down capabilities from the plate, well, and cell level together with links to images, and any graphical image overlays. Various forms of viewing the data should be provided including tables/spreadsheets and graphs (bar charts, scatter plots, and so on). Various views should also be provided for different types of users (e.g., managers, scientists, operators, IT personnel, and so on). Capabilities should be provided for comparing data within a plate, across plates, and so on. Additional capabilities should also be provided for generating statistics on groups of data (e.g., groups of wells, cells, and so on). The data should be displayed in ways that allow the user to explore patterns and recognize patterns and outliers. Users want to be able to save their analyses and visualizations as well as build reports and save these. Making annotations on the data is also very important. Common uses for visualization in HCS include assessing the quality of the dataset (e.g., identifying outliers and false positives), and identifying hits. There are many possibilities for visualization of HCS data and one important visualizer is a plate viewer.



**Fig 5:** Plate viewer plug-in. Visualization of image processing parameters in a heatmap with access to library metadata.

Plate Viewer (PV) guarantees the identification of library and well position of a specific compound on a plate. The history of the location of each compound in the screen, run and replicate along with reformatting information are recorded and reconstructed by PV. Within the GUI the user may select the library, plate and if desired, compounds data derived from a specific 96, 384 or 1536-well plate. Once a plate is selected, a window is opened in a plate viewer that provides functions for easy navigation within the plate that helps extracting comprehensive information about particular compounds (Figure 5).

### 3.3 Architecture of workflow systems

The design of a typical architecture of workflow system is based mostly on plugin framework. The entire application is a functional set of nodes, working together. For example, a plug-in for opening and processing HCS files (library, numeric results and images) in HCDC-KNIME was developed within the KNIME environment. All those open source components (Eclipse environment, KNIME, R-Project, Weka and ImageJ) were chosen for their platform-independence, simplicity, and portability. The pipeline model describes the exact behaviour of the workflow when it is executed. The nodes are usually designed with the following main principles:

- *Resource type for primary data:* The source of data can be collection of high level images familiar to the user or single image. Software should support as many as possible image types.
- *Computation:* Dataflow pipelines dictate that each process be executed as soon as its input data are available. Node processes that have no data dependencies amongst each other can be executed concurrently. They are used for analysis pipelines, data capture, integrating data from different sources, and populating scientific models or data warehouses. Control flows directly dictate the flow of the process execution, using loops, decision points etc.

- *Interactivity*: Node execution could be fully automatic or interactively steered by the user. Data flows are combined by a simple drag&drop process from a variety of processing units. Customized applications can be modeled through individual data subpipelines.
- *Adaptivity*: The nodes and workflow design or instantiation can be dynamically adapted “in flight” by the user or by automatically reacting to changed environmental circumstances.
- *Modularity*: Processing units and containers should not depend on each other in order to enable an easy distribution of computation and allow for independent development of different image processing algorithms.
- *Easy expandability*: There should be easy ways to add new hardware (e.g. microscope), data analysis, or image processing software nodes or views. The distribution of new item should be easy, through a simple plug-in mechanism without the need for complicated install/reinstall procedures.

Figure 6 shows in a schematic way an example of an HCS data analysis flow and the corresponding nodes used in the HCDC-KNIME workflow system.

**ie** business school

93%  
OF MIM STUDENTS ARE  
WORKING IN THEIR SECTOR 3 MONTHS  
FOLLOWING GRADUATION

**MASTER IN MANAGEMENT**

- STUDY IN THE CENTER OF MADRID AND TAKE ADVANTAGE OF THE UNIQUE OPPORTUNITIES THAT THE CAPITAL OF SPAIN OFFERS
- PROPEL YOUR EDUCATION BY EARNING A DOUBLE DEGREE THAT BEST SUITS YOUR PROFESSIONAL GOALS
- STUDY A SEMESTER ABROAD AND BECOME A GLOBAL CITIZEN WITH THE BEYOND BORDERS EXPERIENCE

Length: 10 MONTHS  
Av. Experience: 1 YEAR  
Language: ENGLISH / SPANISH  
Format: FULL-TIME  
Intakes: SEPT / FEB

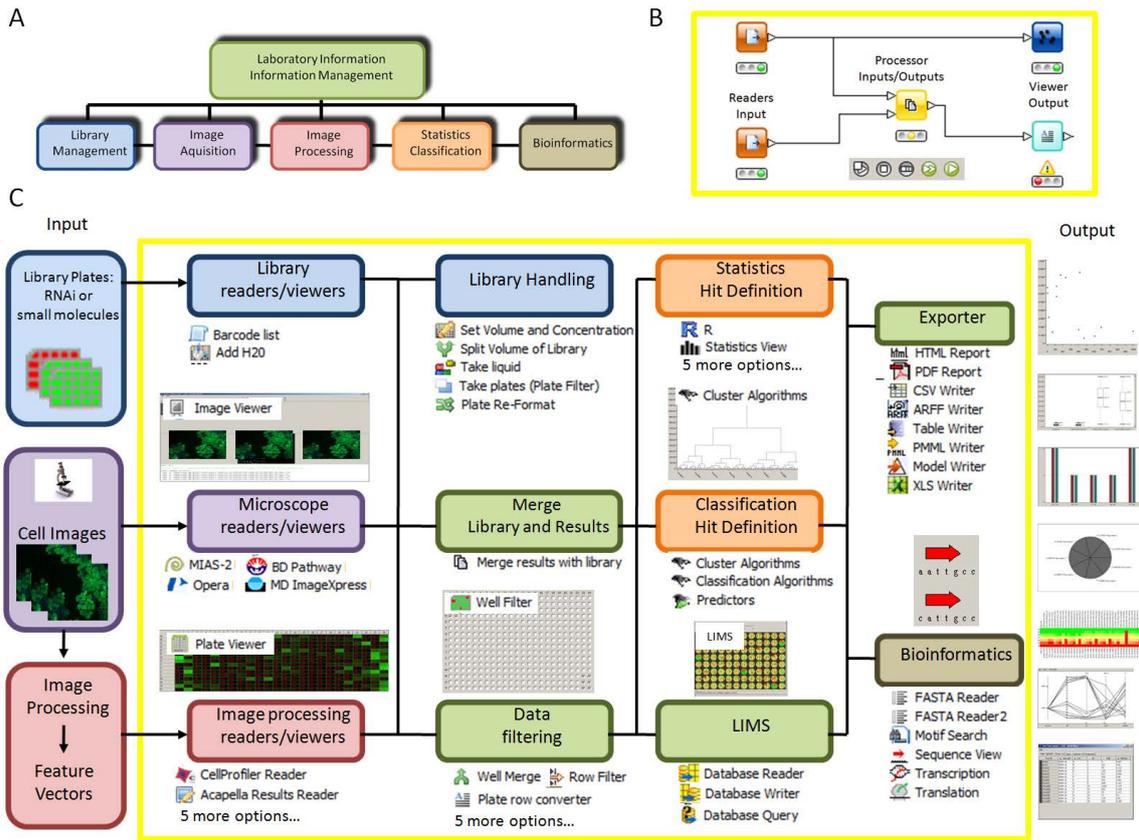
5 SPECIALIZATIONS  
PERSONALIZE YOUR PROGRAM

#10 WORLDWIDE  
MASTER IN MANAGEMENT  
FINANCIAL TIMES

55 NATIONALITIES  
IN CLASS

[www.ie.edu/master-management](http://www.ie.edu/master-management) | [mim.admissions@ie.edu](mailto:mim.admissions@ie.edu) | [f](#) [t](#) [in](#) Follow us on IE MIM Experience





**Fig 6:** HCDC Platform. **a:** Informatics elements behind High Content Screening. **b:** Illustration of a workflow environment with nodes managing the data flow. **c:** Summary of some functionality of HCDC.

### 3.4 Public Domain Workflow Systems

The choice of tools optimized for HCS from entire collection of open-source workflow systems in the life sciences domain is limited. Open-source workflow systems provide major advantages for an the academic environment, not just because they are free of license charges, but also because open-source workflow systems are based on community models of development in which people from diverse background actively contribute to the application. Interestingly, many commercial life sciences workflow products make heavy use of open source and publicly available programs for pre- and post-processing analysis of screening data using packages like R-Project, CellHTS, Weka, BioJava, BioPerl, Chemistry Development Kit (CDK), EMBOSS, etc. HCDC-KNIME (<http://hcdc.ethz.ch>, <http://knime.org>) from the ETH Zurich and University of Konstanz is based on the Eclipse platform provides an excellent data mining platform for drug discovery informatics, bioinformatics and chemistry research. HCDC-KNIME workflow recently released specific nodes for HCS including library handling, quality controls, connection to microscopes, image processing tools, plate visualizations. For small molecule screening the tool already includes CDK nodes and other plug-ins to incorporate existing data analysis tools, such as Weka, the statistical toolkit R, Python scripting and JFreeChart. Tripos Inc. and Schroedinger support KNIME for chemoinformatics and drug discovery nodes or plug-ins. For chemoinformatics and QSAR studies HCDC-KNIME includes nodes that can be used to visualize and transform molecular structures, compute QSAR descriptors and molecular properties, generated fingerprint(s), perform data mining, implement machine learning algorithms (Support Vector Machines, Regression and Bayesian Modeling, Principal Component Analysis), and search for functional groups (substructure and similarity searching). There are other open-source workflow systems in the life sciences domain which can be used in HCS but they needs optimization and specific node development which will support screening standards, formats, hardware, third-parties software. A list of other available workflow systems for life science which can be used in post-processing or used to find molecule correlation between other large scale technologies is provided below:

- Pegasus (Planning for Execution in Grids<sup>6</sup>) is a workflow mapping engine which maps complex scientific workflows onto distributed resources. The Genome Analysis and Database Update (GADU) system, uses Pegasus to perform high-throughput analysis and annotation of the genomics information.
- Core Kepler tool has General Atomic and Molecular Electronic Structure System (GAMESS) as an ab initio quantum chemistry package.
- Kepler<sup>15</sup> provides full support for computational chemistry (nodes for Babel, OpenBabel, and GAMESS) and related workflow for statistical analysis.
- Pegasys<sup>23</sup> developed at UBC provides a specialized workflow management for high-throughput sequence data analysis and annotation. Pegasys can incorporate new tools into existing frameworks.

- The DiscoveryNet<sup>19</sup> platform is a system that integrates bioinformatic tools based on grid computing technologies. Applications of DiscoveryNet are reported for high throughput genomics, proteomics, chemoinformatics, large-scale genotyping data analysis, realtime drug resistance studies and integrative life science analysis.
- SOMA workflow<sup>14</sup> is used to handle different molecular modeling problems related to computer-aided drug design processes developed at genomic biology techniques such as microarrays with bioinformatics tools such as BLAST to identify and characterize eukaryotic promoters.
- myGrid<sup>24</sup> project is a tool for developing semantically enabled grid middleware for supporting bioinformatics and drug discovery applications and, is regarded as the most powerful workflow system.
- Taverna<sup>18</sup> which addresses problems beyond the capabilities of the present system to improve many areas including Data flow centric model, Scalability and Data streaming. Taverna includes services based on SOAP, BioMoby<sup>26</sup>, Biomart<sup>8</sup>, Soaplab<sup>22</sup>, SeqHound<sup>16</sup> and R for numerical analysis.
- Wildfire<sup>20</sup> is a distributed, grid-enabled workflow construction and execution environment developed at the Bioinformatics Institute (A\*STAR).
- Biopipe<sup>12</sup> is a workflow framework based on BioPerl which also allows for execution of workflows across clusters.

### 3.5 Commercial Workflow Systems

- ChemSense is a commercial package which provides high range of chemoinformatics solutions for HCS using small molecules ranging from the analysis and visualization of chemical libraries to the development of combinatorial chemistry libraries, and includes a wide range of QSAR, ADME-Tox prediction, molecular modeling and evaluation methods.
- With their Open Workflow Partner Network, InforSense provides the best-of-breed tools for specific scientific analytic needs.
- Accelrys Pipeline Pilot<sup>11</sup> is one of the very first workflow systems in life sciences optimized for HCS. Pipeline Pilot is chemically intelligent and possesses a robust and highly scalable environment that can run on a multiprocessing environment. Pipeline Pilot is widely used to process High Content Screening and drug discovery data and it comes with specialized solutions for computational chemistry, chemoinformatics and bioinformatics. Pipeline Pilot covers chemistry (compound library acquisition, combinatorial library design, molecular property calculators, filters, and manipulators), ADME/Tox, Decision Trees, Modeling, R Statistics, Reporting, sSequence Analysis, BioMining, Text Analytics and Integration Collection (flexible mechanisms to link external applications and databases).

Below is a list of other commercial available workflow systems for life sciences which can be used in a post-processing or used to find molecule correlation between other large scale technologies:

- BioLog's BioLib is an open architecture Informatics System that creates a unique set of drug discovery IT tools. BioLib covers Protein Modeling, Small Molecules and Peptide Analysis, Database Access, Sequence Analysis and Data Mining.
- UeberTool is a software system for the integration and analysis of molecular biological data with over 200 types of bioinformatics methods. It also enables access to public biological databases and proprietary data including all UeberTool results.

Definitely InforSense KDE and SciTegic's Pipeline Pilot are state of the art workflow systems widely used in HCS operation and applied in academic and non-academic organizations. Table 1 summarizes all available workflow systems (open-source and commercial) used in life science data analysis.

Software	License type	Vendor and URL	Features	Integration with other software
HCDC-KNIME	Open source	LMC, ETH Zurich <a href="http://hcdc.ethz.ch">http://hcdc.ethz.ch</a>  University Konstanz, <a href="http://knime.org">http://knime.org</a>	<ul style="list-style-type: none"> <li>• Provides interactive views of data and models</li> <li>• Based on the Eclipse platform with extensible modular API</li> <li>• Modular data exploration platform</li> <li>• Plate Viewer</li> <li>• Library Handling</li> <li>• QualityControl</li> <li>• Classification</li> <li>• Statistics</li> <li>• Pattern recognition</li> <li>• Library Investigation</li> <li>• HeatMaps</li> <li>• Molecular structure</li> <li>• Image processing</li> <li>• LIMS integration</li> </ul>	CellProfiler Acapella Extrenal tool execution MD Micro microscope BD Pathway microscope Opera microscope MAIA Scientific microscope Matlab R-Project Weka Java API
myGrid	Open Source	UK e-Science <a href="http://www.mygrid.org.uk">http://www.mygrid.org.uk</a>	<ul style="list-style-type: none"> <li>• High level, knowledge-enabled middleware based on web services to support personalized <i>in silico</i> experiments in bioinformatics on the grid</li> <li>• creation, discovery and enactment form a central feature of myGrid services</li> </ul>	

Software	License type	Vendor and URL	Features	Integration with other software
Taverna	Open Source	EBI <a href="http://taverna.sourceforge.net">http://taverna.sourceforge.net</a>	<ul style="list-style-type: none"> <li>Built-in support for web services, local Java functions, BioMoby, and Soaplab</li> <li>workflow language (XScufl)</li> </ul>	
MIGenAS	Open Source	Max-Planck-Society <a href="http://www.migenas.org/">http://www.migenas.org/</a>	<ul style="list-style-type: none"> <li>Integrated bioinformatics workflow engine for web-based sequence analysis</li> <li>Focused on research with microbial genomes</li> </ul>	
Kepler	Open Source	UC Berkeley <a href="http://kepler-project.org/">http://kepler-project.org/</a>	<ul style="list-style-type: none"> <li>Scientific workflow system built on top of the Ptolemy II system</li> <li>XML based workflow definition – MoML</li> <li>Actor prototyping tool</li> </ul>	
GeneBeans	Open Source	UNC Wilmington <a href="http://www.uncw.edu/csc/bioinformatics/">http://www.uncw.edu/csc/bioinformatics/</a>	<ul style="list-style-type: none"> <li>Uses a three-layer architecture</li> <li>An engine, with Graphical User Interface, that models bioinformatics queries as dataflow graphs Discovery (Net Imperial College London), <a href="http://ex.doc.ic.ac.uk/new/">http://ex.doc.ic.ac.uk/new/</a></li> <li>System is a middleware that allows service developers to integrate tools based on existing and emerging grid standards</li> <li>Supports high throughput genomics, proteomics and chemoinformatics</li> <li>Uses discovery process mark-up language (DPML)</li> </ul>	

Software	License type	Vendor and URL	Features	Integration with other software
Pegasys	Open Source	UBC Bioinformatics Centre <a href="http://bioinformatics.ubc.ca/pegasys/">http://bioinformatics.ubc.ca/pegasys/</a>	<ul style="list-style-type: none"> <li>• High-throughput sequence data analysis workflow</li> <li>• Tools for pair-wise and multiple sequence alignment, gene prediction, RNA gene detection, masking repetitive sequences in genomic DNA</li> <li>• Easy integration with Atlas biological data warehouse and its API ProGenGrid (University of Lecce), <a href="http://datadog.unile.it/progen">http://datadog.unile.it/progen</a></li> <li>• Service Oriented Architecture (SOA)</li> <li>• Provides services for drug discovery, access to distributed data and data sharing</li> <li>• Uses gSOAP Toolkit for web services and Globus Toolkit as grid</li> <li>• Middleware</li> </ul>	
Wildfire	Open Source	A*STAR <a href="http://wildfire.bii.a-star.edu.sg/">http://wildfire.bii.a-star.edu.sg/</a>	<ul style="list-style-type: none"> <li>• Distributed, grid-enabled workflow construction and execution</li> <li>• Borrows user interface features from Jemboss</li> <li>• Uses GEL as underlying workflow execution engine Orange (University of Ljubljana), <a href="http://www.ailab.si/orange/">http://www.ailab.si/orange/</a></li> <li>• Component-based framework</li> <li>• Seamless integration within Python</li> <li>• Components for Functional Genomics</li> </ul>	
Biopipe	Open Source	OBF <a href="http://www.biopipe.org">http://www.biopipe.org</a>	<ul style="list-style-type: none"> <li>• Collection of Perl modules for constructing workflows from BioPerl applications</li> </ul>	
Pegasus	Open Source	USC <a href="http://pegasus.isi.edu/">http://pegasus.isi.edu/</a>	<ul style="list-style-type: none"> <li>• Provide abstract workflow and maps it to the available grid resources</li> <li>• Supports a deferred mode</li> <li>• Well-defined APIs and clients</li> </ul>	

Software	License type	Vendor and URL	Features	Integration with other software
Triana	Open Source	Cardiff University <a href="http://www.trianacode.org">http://www.trianacode.org</a>	<ul style="list-style-type: none"> <li>Part of GridOneD project for creating Java middleware for grid applications</li> <li>Pluggable architecture</li> <li>Peer-to-peer implementation based on the Sun's JXTA protocols</li> </ul>	
WsBAW & BioWBI	Open Source	IBM <a href="http://www.alphaworks.ibm.com/tech/wsbaw">http://www.alphaworks.ibm.com/tech/wsbaw</a>	<ul style="list-style-type: none"> <li>WsBAW is Java client application through which users are able to send batch requests to a specific bioinformatics workflow execution engine BioWBI, by using a web service AdaptFlow (University of Leipzig), <a href="http://informatik.uni-leipzig.de">http://informatik.uni-leipzig.de</a></li> <li>Rule-based dynamic workflow adaptation based consultation system</li> <li>Supports the handling of the complex trial therapy processes</li> </ul>	
BioAgent	Open Source	O2I <a href="http://www.bioagent.net">http://www.bioagent.net</a>	<ul style="list-style-type: none"> <li>Supports the biomedical and clinical research</li> <li>Oncology over Internet (O2I) context</li> </ul>	
SOMA	Open Source	Finnish IT Center for Science <a href="http://www.csc.fi/proj/drug2000">http://www.csc.fi/proj/drug2000</a>	<ul style="list-style-type: none"> <li>Workflow for small molecule property calculations</li> <li>Supported by the core workflow program Grape</li> <li>Includes programs: CORINA, ROTATE, BRUTUS, GOLD, SYBYL,</li> <li>VOLSURF, XSCORE</li> </ul>	
Pipeline Pilot	Commercial	Accelrys <a href="http://www.scitegic.com/">http://www.scitegic.com/</a>	<ul style="list-style-type: none"> <li>Visual programming HCS optimized</li> <li>plate viewer</li> <li>HeatMaps</li> </ul>	<p>Widely applied in drug discovery and HTS</p> <p>Visual programming Integration of third party applications</p> <p>Wide range of Bioinformatics, QSAR, molecular modeling and evaluation methods</p>

Software	License type	Vendor and URL	Features	Integration with other software
ChemSense	Commercial	InforSense <a href="http://www.inforsense.com/">http://www.inforsense.com/</a>	<ul style="list-style-type: none"> <li>Built on InforSense KDE core provides seamless integration with third party tools</li> <li>Wide range of Bioinformatics, QSAR, molecular modeling and evaluation methods</li> </ul>	<ul style="list-style-type: none"> <li>Wide range of Bioinformatics, QSAR, molecular modeling and evaluation methods</li> <li>Full integration of Perl, R/Bioconductor and Matlab</li> <li>HCS instrumentation support</li> </ul>
VIBE	Commercial	INCOGEN <a href="http://www.incogen.com/">http://www.incogen.com/</a>	<ul style="list-style-type: none"> <li>Can interface with a variety of environments, including high throughput platforms such as Sun Microsystem's Grid Engine</li> <li>supports GRID computing used in image processing</li> </ul>	<ul style="list-style-type: none"> <li>Extensive use of XML for configuration, data exchange, data storage and communications</li> <li>Extensible Java API</li> </ul>
ueberTool	Commercial	Science Factory <a href="http://www.science-factory.com/">http://www.science-factory.com/</a>	<ul style="list-style-type: none"> <li>can be used in HCS for post-analysis, hits evaluation</li> <li>Integration and analysis of molecular biological data</li> <li>Graphical user interface makes constructing bioinformatics workflows</li> <li>Blast</li> <li>supports FASTA format</li> </ul>	<ul style="list-style-type: none"> <li>Integrated programming language for extending core functionalities</li> </ul>
BioLib	Commercial	BioLog	<ul style="list-style-type: none"> <li>Support Bio-IT data warehousing</li> <li>Comprehensive Methods Library and a customizable Algorithm Library</li> <li>Open-architecture Informatics System</li> </ul>	

**Table 1.** Workflow systems in life sciences domain. Green highlighting indicate workflow systems optimized for High Content Screening.

### 3.6 Summary and Vision

Large-scale HCS data analysis needs flexible workflow based integration of different components and sub-processes from diverse formats (library, image readers, microscope nodes, image processing results, data mining) which can provide *in silico* experimental design through visual programming and execution on grids. A pipeline system in HCS is a very new concept and still evolving. The final goal is a distributed and ubiquitous environment which can integrate all automated microscopes, all available image processing packages, bioinformatics databases and data from other large scale experiments (proteomics, microarray, flow cytometry, new sequencing, etc). Workflow systems can be data-intensive, computation intensive, analysis intensive, visualization-intensive, process-intensive. Problem of service composition is how to compose simple services to perform complex tasks. The scalability of a workflow system is an important factor which helps in large-scale HCS data analysis in a high performance parallel and distributed computing environment<sup>5</sup>. One very important aspect is the option to run workflows in the background on a remote server which is, especially advantageous in case of long running workflows. In this situation only the control of the workflow should be presented in the remote GUI (desktop or web client). Present workflow systems in life sciences which can be applied for HCS need to integrate several resources like web technology (LIMS systems), grid services (for powerful image processing) and web services (access to bioinformatics sources). Web and grid services provide access to distributed resources, while workflow techniques enable the integration of these resources to perform *in silico* experiments. Most of the HCS database systems (LIMS) and very often all external compound information (RNAi or small molecules) are accessible over the web. After finalizing the experiment and retrieving the hits it is necessary to investigate (compare) results with external information. Semantic web services will help accessing this biological knowledge in a distributed, heterogeneous environment by adding semantics, defining common ontologies and applying them to software tools and databases. Semantic web technology can provide more generic solutions that can be re-used between related workflows. “Web services” is a distributed computing technology that provides software services over the web. Over the past few years the evolution of web services in bioinformatics<sup>17</sup> has shown tremendous impact on the sharing of data and tools. With the intention directed towards execution in a heterogeneous and often distributed environments, the interoperability of web services has become much more important.

# 4 Database Development: Laboratory Information Management Systems and Public Databases

How best to archive and mine the complex data derived from HCS experiments that provides a series of challenges associated with both the methods used to elicit the RNAi response and the functional data gathered? To enable effective data retrieval for HCS experiments, data and images and associated information must be stored with high integrity and in a readable form. HCS data should be stored in a form that takes advantage of the characteristics of this type of data to enable full access, analysis and exploitation of the data. A key factor is the database model which represents data in logical form. The data model (or database structure or database schema) should be flexible to handle the various HCS data types (i.e., compound information, results: image data and derived metadata), experiment simulation and a wide range of changes in the data (e.g., different numbers of wells, cells, features, images, different image sizes and formats, different number of time-points, and so on).

The structure of the data model provides a way of describing data and the relationships within the data, enabling data to be organized, cataloged, and searched effectively. Databases where a database model is implemented enable joining of related data to allow meaningful visualization, analysis, and data mining. The data model is also important for integration with other systems.

## 4.1 What Type of HCS Data Have to Be Managed in the Database?

HCS data are containing three types of data:

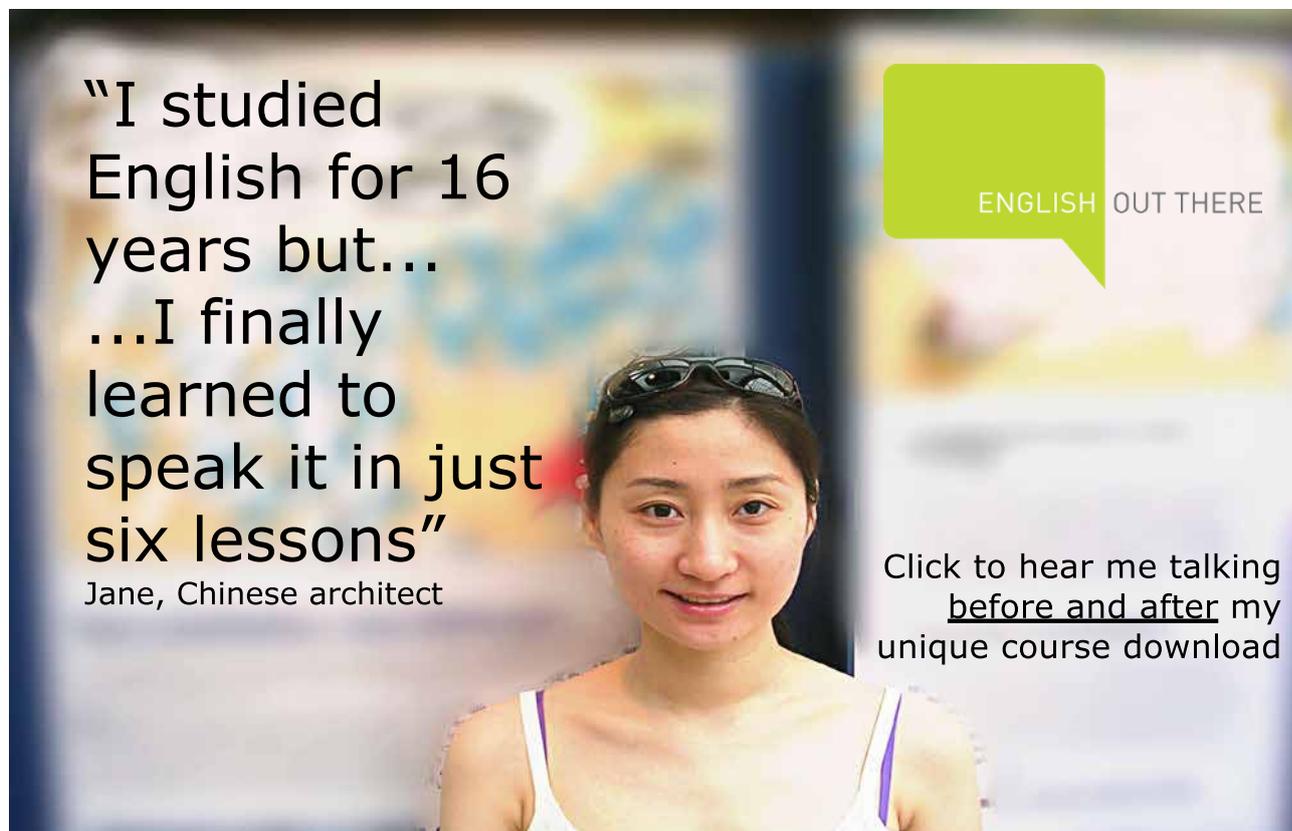
1. Database of compounds (RNAi or small molecules).
2. Numbers of images that require significant amounts of storage.
3. Numbers of files including image processing parameters.
4. Meta-data.

Thus, a large amount of data is collected for just one well of a single plate. In addition, other associated information about the assay or experiment, such as protocol information, is also typically recorded.

Having four types of data is easy to define three general categories of HCS data:

- **Image data:** These are the images acquired at each channel for each field within a well and produced thumbnails for visualization purposes
- **Numeric Results data:** these are the measurements that result from performing an analysis on an image with image analysis algorithms.
- **Metadata:** These are the associated data that provide context for the other two categories of data (i.e., metadata are data that describes other data). Examples are: well – compound annotation, assay type, plate information, protocols, operators, calculated data such as dose–response values, as well as annotations imported from other systems.

Let's try to understand how data are produced. HCS microscopes typically scan multiwell plates. These plates typically have 96, 384, or 1536 wells. Each well is a container in the plate that contains an individual sample of cells. Each well is divided into multiple fields. Each field is a region of a well that represents an area to image. Each field consists of multiple images, one for each individual wavelength of light (referred to as a “channel”, “staining”), corresponding to the fluorescent markers/probes used for the biology/dye of interest (e.g., DAPI). There are typically between two and four channels per field (e.g., each channel shows different elements of the cells: 1 channel nuclei, 2 channel: cell membranes, 3 channel: endosomes, and so on). The images produced are immediately analyzed using automated image processing. Experiment results are produced.



“I studied English for 16 years but...  
...I finally learned to speak it in just six lessons”  
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

<b>HCS run:</b> assuming follow parameters - produced image size (single image, not packed in stack, one field from one well, no montage) ~1.5MB - produced thumbnails 200kB - metadata file for each well (average 25 features) – 200kB - 500 cells per well - cell segmentation and nuclei detection - off line image processing	<b>Time:</b> - Image processing time (depend on core processors CPU) - Acquisition time	<b>Number of Images</b>	<b>Number of records</b>	<b>Storage size</b>
12 x 384-well plates (very often used for Kinome) 9 fields per well 2 channels No time lapse	Image processing: 16 CPUs = 18 h 32 CPUs = 9 h 128 CPUs = 3 h 1000 CPUs = max 1 h  Acquisition 48 h	82944	Cell based= 2 304 000 Image based = 82944 Well based= 4 608	Images = 124 GB Thumbnails = 8.2GB Metadata: Well based = 2.4MB Image based = 8.2GB  <b>Total: 140,402 GB</b>
12 x 384-well plates (very often used for Kinome) 4 fields per well 3 channels No time lapse	Image processing: 16 CPUs = 16 h 32 CPUs = 8 h 128 CPUs = 2 h 1000 CPUs = max 1 h  Acquisition 44 h	55296	Cell based= 2 304 000 Image based = 55296 Well based= 4 608	Images = 83 GB Thumbnails = 6.2GB Metadata: Well based = 1.8MB Image based = 6.3GB  <b>Total: 97,3 GB</b>
100 x 384-well plates one run of genome experiment) 9 fields per well 2 channels No time lapse	Image processing: 16 CPUs = 300 h 32 CPUs = 150 h 128 CPUs = 40 h 1000 CPUs = 10 h  Acquisition 400 h (17days)	691 200	Cell based= 19 200 000 Image based = 691 200 Well based= 38 400	Images = 1 036.8 GB Thumbnails =138 GB Metadata: Well based = 76.4MB Image based =138GB  <b>Total: 1312.84 GB = 1.3 TB</b>
100 x 384-well plates one run of genome experiment) 4 fields per well 3 channels No time lapse	Image processing: 16 CPUs = 300 h 32 CPUs = 150 h 128 CPUs = 40 h 1000 CPUs = 10 h  Acquisition 400 h (17days)	460 800	Cell based= 19 200 000 Image based = 460 800 Well based= 38 400	Images = 691.2 GB Thumbnails =92 GB Metadata: Well based = 76.4MB Image based =92GB  <b>Total: 875.96 GB = 0.8 TB</b>
296 x 384-well plates (very often used for Genomewide) 9 fields per well 2 channels No time lapse	Image processing: 16 CPUs = 900 h = 38days 32 CPUs = 444 h = (18days) 128 CPUs = 222 h = 9days 1000 CPUs = 48 h = 2 days  Acquisition 1184 h 49 Days	2 045 952	Cell based= 56 832 000 Image based = 20 459 52 Well based= 113 664	Images = 3 068.9 GB Thumbnails = 40GB Metadata: Well based = 7.4MB Image based = 40GB  <b>Total: 3156,3 GB =3.2TB</b>
296 x 384-well plates (very often used for Genomewide) 4 fields per well 3 channels No time lapse	Image processing: 16 CPUs = 900 h = 38days 32 CPUs = 444 h = (18days) 128 CPUs = 222 h = 9days 1000 CPUs = 48 h = 2 days  Acquisition 1184 h 49 Days	1 3 63 968	Cell based= 56 832 000 Image based = 1 363 968 Well based= 113 664	Images = 2 045.9 GB Thumbnails = 27GB Metadata: Well based = 5MB Image based = 27GB  <b>Total: 2 009.8 GB =2TB</b>

**Table 2.** Examples for Acquisition Time, Processing Time and Data Volumes for Different HCS Run Scenarios.

Each well is seeded with a certain number of cells which has to be detected by image processing algorithms. The cell number counted is a basic parameter used for the quality control of automation, microscopy or assay performance. The number of cells per well varies depending on the experiment, but typically ranges between 5 and 10000 cells. Very often images from well fields are merged into one image using montage function. For each cell, multiple object features (or measurements) are calculated by automated image processing. The cell features include measurements such as size, shape, intensity, and so on.

The number of cell features calculated varies depending on the assay, but typically ranges between 5 and 500. Those features have to be carefully investigated, filtered and only parameters should be considered for hit definition. In addition, cell features are often aggregated to the well level to provide well level statistics (one well one row labeled with plate annotation and position as unique identify). The total storage size for experiments is primarily based on the acquired image data, image thumbnails, library information and the numeric results data. The amount of data, acquisition and processing time varies depending on a number of factors including the type of assay, the number of plates, the type of the screen (primary, secondary), available computational hardware, the throughput of the instrument or analysis application and the number of instruments which can work parallel. Table 2 demonstrates example experiments and summarizes necessary time, number of records and require for storage space. The size of the library information and numeric results data are counted in megabytes. Numeric results are estimated by the number of feature records (lines in tables). Image storage depends on the number of images acquired. The number of images depends on plate number, plate type (96, 384, 1536 ), number of fields, number of channels, confocality levels and eventually time points in case of kincetic studies. The typical image size acquired ranges between 500KB and 2 MB (single slice, single tiff file without montage). Thumbnails of those images often are generated using jpeg compression, their size range between 150–300 kb. For numeric results data are categorized in three types of outputs: cell based, image based and well based. The number of image based record should be equal to the number of acquired images which is also equal to the number of thumbnails produced. The record number of well based results data should be equal to the number of all wells in screening experiment.

In high content informatics, the numeric data are supported by the images and the validation of numeric data is based on the visual inspection of images. Any high content informatics solution therefore needs to be able to efficiently handle the relationships between the various levels of numeric results data, library information and the associated images. In the next subsection we will describe a database model (schema) and a database solution for handling library data, images and numeric results data.

## 4.2 Database Schema

Defined relations in HCS data model allow the stored data to be broken down into smaller logical and easier maintainable units (mostly tables) in relational database system. To create, modify, and query data stored in tables of the database, the Structured Query Language (SQL) has been developed. The usage of relational database management systems (DBMS) leads to the following advantages:

- Reduction of duplicate data: Leads to improved data integrity.
- Data independence: Data can be thought of as being stored in tables regardless of the physical storage.
- Application independence: The database is independent of the systems, microscopes and programs which are accessing it.
- Concurrency: Data can be shared with many users.
- Complex queries: Single queries may retrieve data from more than just one table.

The data model installed on relational database is accessed from the application specific drivers which open communication for programs with graphical user interfaces, designed client applications or through the web server using standard browsers. The relational core database integrates and stores data from experimental results obtained from the HCS assays and from the bioinformatics analysis referenced by a common identifier. The database can be queried with the client tools supplied by the database system. These applications offer a high degree of flexibility but lack visualization features (e.g. results are shown in one table only). A more comprehensive user interface can be provided by custom-made client applications (developed in a high-level programming language) as stand-alone programs or applications on a web server. They wrap the appropriate SQL statements and process the returned results.

The data integration in a relational database represents a major advantage since the high-level structured query language (SQL) can be used to access all data regardless of their origin. The user can pose arbitrarily complex queries to crosscheck experimental results within compound features by simply executing appropriate SQL statements. Other queries can be made to check for compliance between predicted results and experimental compound features in order to confirm results for further annotations.

The idea of the HCS data model is provide structure able to store the data required to both reproduce the samples and experiments involved in compound production and to inform subsequent work. The HCS data model should be designed around the themes of sample, experiment, target, and experimental objective. Figure 7 illustrates a very basic model which can be used as a base for an HCS library database. Figure 8 illustrates general models which can be used for an HCS results database. Both data models can be combined into one data model by merging and by using identifiers from the plate and well table.

For an HCS data model several key capabilities are required:

- The model must enable the description of the following:
  - (1) the composition of a sample
  - (2) the physical location of a sample
  - (3) the involvement of a sample in an experiment
  - (4) experiment protocols
  - (5) experiment results (images, metadata)
  - (6) the sequence of work performed to produce a sample
  - (7) the relationship between sample, target, and experimental objective
  - (8) the ownership of samples and experiments
- The model must be sufficiently flexible to cope with unexpected products from experiments.
- The model must be extensible and maintainable.

Excellent Economics and Business programmes at:



university of groningen




**“The perfect start of a successful, international career.”**

**CLICK HERE**  
to discover why both socially and academically the University of Groningen is one of the best places for a student to be

[www.rug.nl/feb/education](http://www.rug.nl/feb/education)



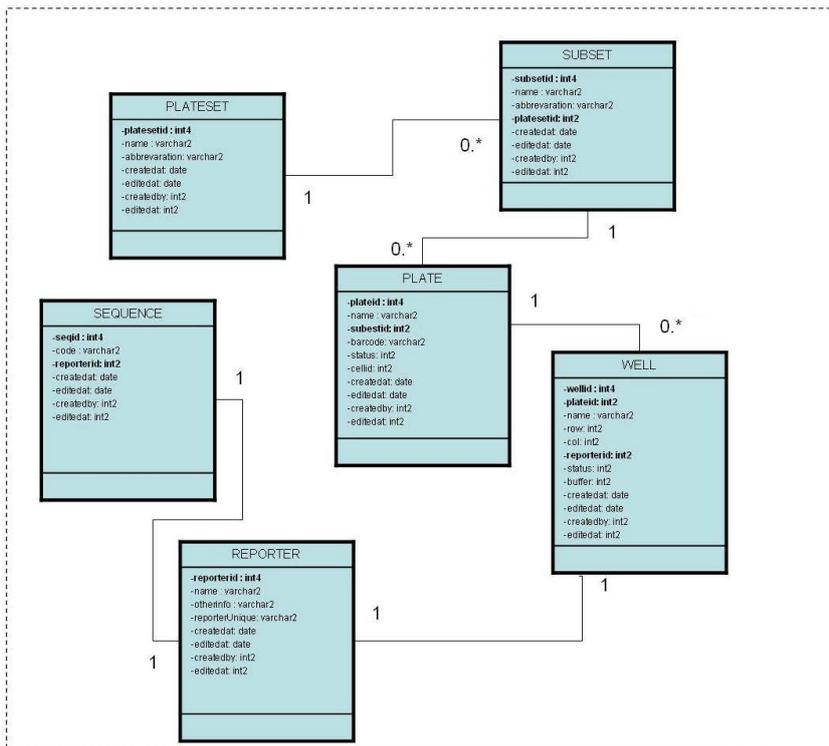


Fig 7: Data model for a Library Database.

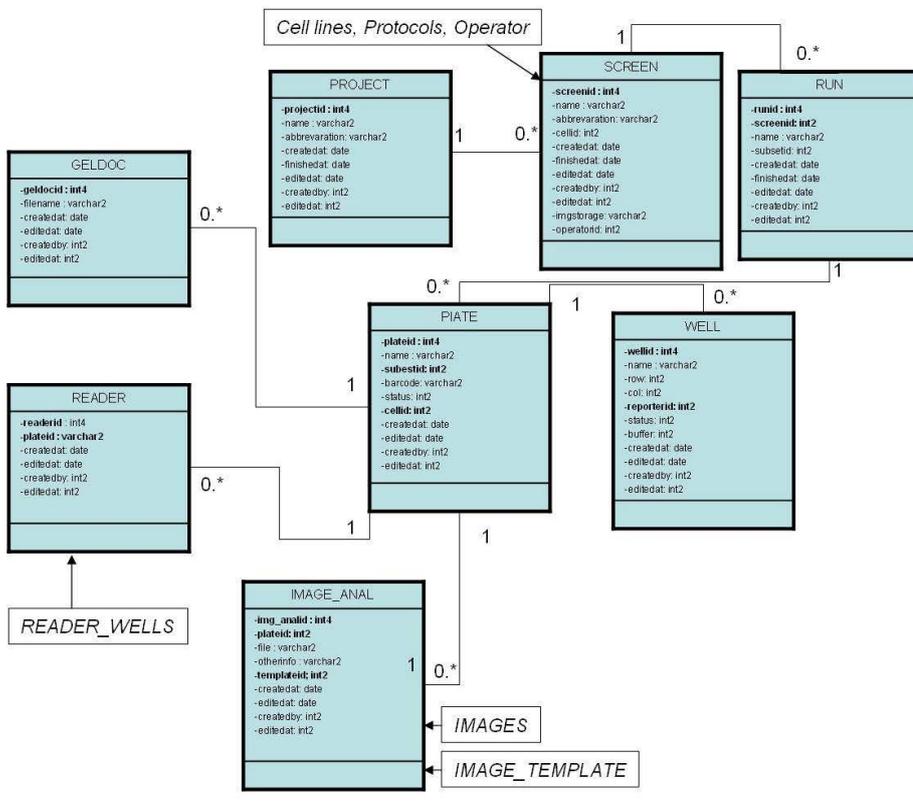


Fig 8: Data model for HCS Results Database.

### 4.3 LIMS Architecture

Data model design belongs to the first development phase of a Laboratory Information Management System (LIMS). After model design, LIMS should be developed to enable a flexible integration of the heterogeneous data types mentioned above, data sources, and applications. Such systems should provide well defined user and data interfaces and fine grained user access levels.

Consequently, following the specific aims must be considered for LIMS development:

- Design and development of the LIMS including:
  - An integrated laboratory notebook to store the necessary information during biomaterial manipulation.
  - A laboratory information management system to keep track of the information that accrues during production in multiwell plates and the screening.
  - Well defined data interfaces for importing, exporting, and handling data.
  - An Plug-in Architecture (PA) to connect other bio applications and link to its data without amending the LIMS code.
  - A web-service interface to allow external applications such as data mining tools to query and read the stored data and to write back results.
  - The management of experimental data coming from various types of investigations.



**American online**  
**LIGS University**  
is currently enrolling in the  
Interactive Online **BBA, MBA, MSc,**  
**DBA and PhD** programs:

- ▶ enroll **by September 30th, 2014** and
- ▶ **save up to 16%** on the tuition!
- ▶ pay in 10 installments / 2 years
- ▶ Interactive **Online education**
- ▶ visit [www.ligsuniversity.com](http://www.ligsuniversity.com) to find out more!

**Note: LIGS University is not accredited by any nationally recognized accrediting agency listed by the US Secretary of Education. More info [here](#).**

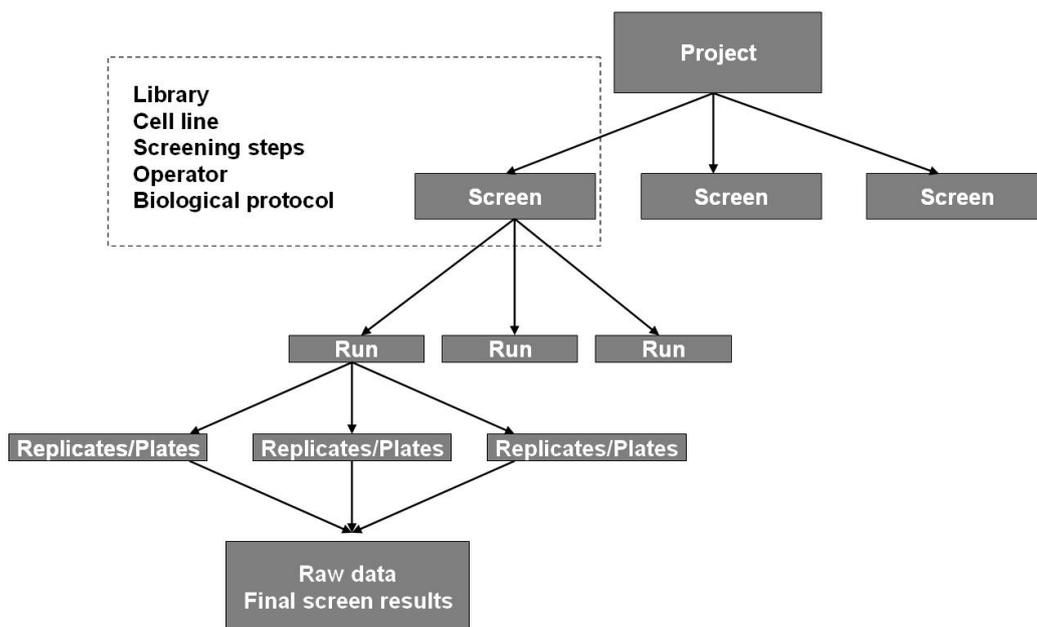


- Initiation, design and implementation of a user management system that provides libraries and interfaces which can be integrated in any application to facilitate user authentication and authorization.
- Initiation of database and a web portal to browse and upload screening results and screening datasets in order to analyze the compound image analysis in the context of several biological assays.

Currently, there are many LIMS available in life sciences (Table 3). The LIMS is a customizable software package and analysis platform designed to be installed in HCS laboratory and to serve many users simultaneously via the web or desktop client. LIMS should be able to import data into the database, group plate data together into experiments, and in a uniform and streamlined fashion, apply filters and transformations and run analyses. To facilitate online collaboration, users can share almost any object within the database with another user. Data can be exported in a multitude of formats for local analysis and publication. Compounds of a library stored in a library database can be interactively linked with the next module called HCS Results Database. The entry results data can begin with the definition of a project, screen, run and all experimental protocols presented in Figure 9, goes through definitions of biomaterials used, cell culture conditions, experimental treatments, experimental designs, definition of experimental variables, to definition of experimental and biological replicates and finally ends with the selection of the compound library for the screen. The user of the LIMS should easily simulate the project hierarchy via additional GUI interfaces which simulate cases that exist in a real screening process. The database should facilitate remote entry of all information concerning the screen, where users may create associations of labeled extracts and substances, scanned raw images from microscope and quantification matrices (files with results after image analysis). The user may wish to create associations of labeled extracts, scanned raw images, quantification matrices. As a single compound located in one well of a multiwell plate can be scanned in an automated screening microscope and/or under different settings.

#### 4.4 LIMS and User Management System

The researchers that use LIMS are in most cases organized in groups and each user belongs to one or more groups. The purpose of the groups is to define a set of users with common permissions over elements of the system, in other words, the subsets of plates that a group of users can view or use. The groups allow the assignment and management of permissions very easily, but also provide enough granularity to control access of the different users to the subsets and plates. A typical HCS unit and their users are composed by different groups and laboratories, each of them working in different projects. The manager is able to control privileges and is able to create at least one group for LIMS users or research group. A specific research group will work with a set of plates and the rest of laboratories should not have access to those plates.



**Fig 9.** Typical screening hierarchy. Screening parameters are defined on a “screen” level and shouldn’t be modified in sublevels.

#### 4.5 Type of Users

In many cases, there are three types of users or level access in LIMS systems:

- **Manager:** This type of user is the responsible of introducing, maintaining and updating the data about plates and reporters in the database system. Additionally, the manager defines the screen, protocols, robots and external databases and assigns the adequate permissions to the rest of users for visualizing the subsets of plates. The manager has total access to the application and can do any modification within the database.
- **Researcher:** The researcher represent the most general user of the LIMS. This access is limited to the visualization and searching of the data from plates. A researcher typically corresponds to a scientist of the institute or the laboratory.
- **Guest:** This user access has the same privileges as the researcher, the difference is that it should be used by different people to access LIMS. The manager must carefully handle the permissions of subsets, and allow the visualization of these elements to the guest only if the data are to be published.

Software	Supplier	Scope	Description
HCDC-LIMS	ETH Zurich <a href="http://hcdc.ethz.ch">http://hcdc.ethz.ch</a>	Data storage and management	<ul style="list-style-type: none"> <li>- Storage of library handling data from pipeline into database</li> <li>- Read and organize library in database</li> <li>- Store image processing results in database</li> </ul>
SapphireTM	LabVantage <a href="http://www.labvantage.com">http://www.labvantage.com</a>	LIMS	<ul style="list-style-type: none"> <li>- LIMS with an open architecture enabling free definition of workflows.</li> <li>- Integrates external compound repository databases.</li> </ul>
Metamorph1 and AcuityXpress	Molecular Devices <a href="http://www.moleculardevices.com">http://www.moleculardevices.com</a>	HCS – image management and analysis	<ul style="list-style-type: none"> <li>- Integrates with Molecular Devices HCS readers and AcuityXpress Image storage, analysis and mining software suite for cellular images with open image database API.</li> <li>- Includes management tools for multi-user environments.</li> </ul>
Genedata Screening data analysis	Genedata <a href="http://www.genedata.com">http://www.genedata.com</a>	Screening data analysis, information management	<ul style="list-style-type: none"> <li>- Screening application supports quality control and analysis of interactively managed early-stage and large volume screening datasets.</li> <li>- Provides exhaustive interactive visualizations based upon a broad range of statistical analyses to help prioritize compound sets for follow-up work.</li> <li>- Phylosopher1 integrates metadata from drug discovery projects ranging from genomics to pathway data and mode of action (MOA) studies.</li> </ul>
Cellenger	Definiens <a href="http://www.definiens.com">http://www.definiens.com</a>	Image analysis and data management for high-content screening (HCS) and biomedical applications	<ul style="list-style-type: none"> <li>- Cellenger Developer Studio and Enterprise for automated (pre-defined work flows using Cellenger Server) object-oriented image analysis, uses structural and relational information in images (morphometric quantization) and realizes an image</li> <li>- object hierarchy.</li> <li>- Based upon 'Cognition Network Technology' aiming to mimick human perception of objects.</li> </ul>

Software	Supplier	Scope	Description
Acapella™ Columbus	PerkinElmer <a href="http://www.perkinelmer.com">http://www.perkinelmer.com</a>	High-content data analysis	<ul style="list-style-type: none"> <li>- Columbus is a convenient and easy-to-use solution for high volume data management, storage, retrieval, visualization and protection of images and analyzed results.</li> <li>- Designed as a complementary product for the Opera™ platform, Columbus can import, export and manage image formats from a wide variety of sources, providing a central repository and solution for all your microscope imaging requirements.</li> <li>- Interactive, fully scriptable and compatible with 3rd-part platform environments.</li> <li>- Upgradeable with user libraries.</li> <li>- Provides high-level language to reduce coding overhead for main applications in image analysis (HCS): object recognition, grouping and segmentation, morphologic filtering, image arithmetic.</li> <li>- Libraries available also for Photon Statistics or specific applications like FLIPR kinetics analysis.</li> <li>- SDK available.</li> </ul>

**DON'T EAT YELLOW SNOW**

What will your advice be?

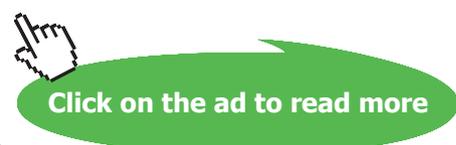
Some advice just states the obvious. But to give the kind of advice that's going to make a real difference to your clients you've got to listen critically, dig beneath the surface, challenge assumptions and be credible and confident enough to make suggestions right from day one. At Grant Thornton you've got to be ready to kick start a career right at the heart of business.

Sound like you? Here's our advice: visit [GrantThornton.ca/careers/students](http://GrantThornton.ca/careers/students)

Scan here to learn more about a career with Grant Thornton.

**Grant Thornton**  
An instinct for growth™

© Grant Thornton LLP. A Canadian Member of Grant Thornton International Ltd



Software	Supplier	Scope	Description
HCITM	ThermoFisher <a href="http://www.thermofisher.com">http://www.thermofisher.com</a>	HCS – image management and analysis	<ul style="list-style-type: none"> <li>- Multi-tier integrated environment for large volumes of HCS data.</li> <li>- Middle-layer manages data level (image store) and presentation (user) level with plug-in interfaces for additional functionalities like user data I/O, visualizations, workflow management and QA (vHCS Discovery Toolbox).</li> </ul>
CellMine™ and SIMSTM	Biolmagene (SciMagix) <a href="http://www.biolmagene.com">http://www.biolmagene.com</a>	HCS – image management	<ul style="list-style-type: none"> <li>- Multi-tier architecture for fast image-I/O of large volume HCS data from various instrument sources.</li> <li>- Supports workflows for reorganization, aggregation and visualization of image and metadata for further analysis.</li> </ul>
ActivityBase™	IDBS <a href="http://www.idbs.com">http://www.idbs.com</a>	HTS data management and analysis	<ul style="list-style-type: none"> <li>- Biological assay data- and experiment-management platform.</li> <li>- All data processing via a central relational database as the store and Microsoft Excel for data analysis (analysis workflows defined via Excel templates).</li> <li>- Has chemistry cartridge and deals with drug metabolism and pharmacokinetics (DMPK) data specifics.</li> </ul>
IN Cell Miner	GE HEALTHCARE <a href="http://biacore.com">http://biacore.com</a>	High-Content Manager (HCM) for the effective management of complex data generated by cellular high-content screening and analysis systems.	<ul style="list-style-type: none"> <li>- IN Cell Miner HCM is designed to help increase scientists' productivity by offering:</li> <li>- Flexibility to import new as well as already-existing IN Cell Analyzer data</li> <li>- Functionality to view and retrieve data from plate to wells to cells</li> <li>- Tools to facilitate project annotation</li> <li>- Guided searches for easy data retrieval</li> </ul>
Pipeline Pilot™	Accelrys <a href="http://www.accelrys.com">http://www.accelrys.com</a>	Data analysis and mining	<ul style="list-style-type: none"> <li>- Data analysis and workflow management based on graphical programming (visual scripting): components are visually arranged to protocols.</li> <li>- Pipeline Pilot™ Publication of protocols for remote execution.</li> <li>- Configurable components for chemistry, statistics, sequencing, text mining as well as integration of 3rd party applications.</li> </ul>

Software	Supplier	Scope	Description
Genepattern (GP)	NIH grant project Genepattern (GP) <a href="http://www.broad.mit.edu/genepattern/">http://www.broad.mit.edu/genepattern/</a>	Data management and analysis	<ul style="list-style-type: none"> <li>- Workflow management system for data analysis and visualization.</li> <li>- Provides graphical IDE and object browser. GP comes along with plenty of modules for statistics, visualization, machine learning, etc. to be arranged as a sequential or parallel pipelined workflow. GP modules are also accessible from within R-project, Java and MATLAB1.</li> </ul>
Synapsia Informatics Workbench	Agilent <a href="http://www.chem.agilent.com">http://www.chem.agilent.com</a>	Knowledge management	<ul style="list-style-type: none"> <li>- Concurrent Synapsia provides the Discovery Manager desktop user interface: object hierarchies are mapped to a file- and directory-like structure whereby content and relationships can be displayed (e.g. with Spotfire1, as a SAR table or a phylogenetic tree).</li> <li>- The open architecture and documented APIs enable integration of external tools for (e.g. BLAST searches).</li> <li>- Together with Information Manager it represents a collaboration framework for cross-discipline R&amp;D projects.</li> </ul>
Foundation Server	TripotTM <a href="http://www.tripos.com">http://www.tripos.com</a>	Small Molecular Screening, Cheminformatics, computational chemistry	<ul style="list-style-type: none"> <li>- Application server integrates tools and provides access to third-party discovery informatics software.</li> <li>- Foundation Server SYBYL as an optional environment provides tools or molecular modeling and cheminformatics</li> </ul>
EMC2	Documentum <a href="http://www.documentum.com">http://www.documentum.com</a>	Content management	<ul style="list-style-type: none"> <li>- Managed collection of software tools to organize</li> <li>- unstructured information originating from sources like documents, spreadsheets, web pages or e-mail databases according to defined business rules.</li> <li>- Documentum Creates relationships, organizes metadata and provides tools for search, retrieval and presentation.</li> </ul>

**Table 3.** LIMS and Data Management Systems Used in HCS.

## 4.6 Integration and Public Databases

HCS data is usually exported from LIMS to third party systems, for either further analysis “warehousing” purposes or archiving. Linkage at the data level via an export is a simple means to deliver HCS data into the enterprise as well as integrate HCS data into laboratory workflows. The informatics architecture therefore needs to support the necessary relational data structures to permit annotation, such as sample identifiers for compounds. In order to push data into the enterprise and link it in, format neutral export tools are required. Over the past years XML (eXtensible Markup Language<sup>9</sup>) has arisen as the format of choice for data export, as it is self-explaining format (i.e., not only does it contain the data to export but a description of the data in the same file). Any software can interpret XML and it can be translated into other formats, if necessary. Data-level integration has certain advantages: It is relatively straightforward to implement, almost any data can be integrated, and few changes, if any, are required by either the source or target applications. Disadvantages are that an additional copy of the data is made and there may not be a way to actively link content (e.g., if one sees an interesting data point, one could see the associated image without further programming).



.....Alcatel-Lucent 

[www.alcatel-lucent.com/careers](http://www.alcatel-lucent.com/careers)

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".

Very often HCS data stored in LIMS are either directly integrated or published into a data warehouse with other discovery data sources, loader scripts or database views are used, and data are often cleansed or some middleware software is used as an abstraction layer to more loosely “federate” for example HCS LIMS with genomics, metabolic and cheminformatics databases. Middleware layers, often called metalayers, provide consumers of data with a single “view” on the data, independent of the native data format or schema. In this way a user application can query and work with data across perhaps dozens of data sources, be they relational databases or unstructured data such as text files and images<sup>10</sup>.

The integrated data warehouse approach to database integration have some advantages that it is relatively simple to implement and there are now sophisticated data warehousing tools for carrying this out. However, as the desire to integrate more data sources grows, the system has to scale and this requires hands on effort. The volume and complexity of HCS data is also a consideration when building a data warehouse/integrated data integration. Performance of the metalayer when querying across dozens of disparate data sources can also be an issue. If the schema of the source changes, the adapter has also to be updated.

How best to share published HCS data saved in LIMS or a warehousing application? The accelerating accumulation of HCS data from ongoing large-scale analyses projects calls for a public database system focused on phenotypes. Such a system should ideally be freely available, web-accessible, user-friendly, adhere to community standards and provide flexible query options and tools for analysis of the data between projects.

Public databases should fulfill the following goals:

- Make HCS technology available to the scientific community by providing a facility with the required infrastructure and expertise.
- Provide a common platform to exchange variables between screens, allowing for functional comparisons across studies.
- Create a database, in a standardized format, for the repository of results from all screens, which, upon publication, are made available to the public. The public database is divided into sections that offer researchers several basic data viewing options as well as a number of bioinformatics tools and links to other databases.
- The databases should contain a compendium of publicly available data and provides information on experimental methods and phenotypic results, including raw data in the form of images or streaming time-lapse movies.
- Phenotypic summaries together with graphical displays of compounds (small molecules or RNAi) to gene mappings allow for a quick intuitive comparison of results from different HCS assays and for a visualization of the gene product(s) potentially inhibited by each compound.

Public databases are usually searched using combinatorial queries using the novel tools, which rank compounds according to their overall phenotypic similarity. One of the ideas behind public sharing of genome information is the distributed public database model<sup>7</sup>, in which interconnected databases can also act as portals displaying specific types of information from other databases that are curated and developed by the community of people involved in populating them. Each public database contains on main home page a simple 'quick' search form for finding compounds using drop-down menus for selecting phenotypes by life stage and an optional text box for specifying screening experiments, genes, phenotypes, laboratories and experimental reagents by name. Screening experiments should also be searched by phenotype using either a simple menu driven form or an advanced phenotype search form that provides a combinatorial query builder. Additional search options should provide the ability to query any object represented in the database using either a simple class browser with an optional name or a wildcard pattern, a text/keyword search, or a Query Language statement. Related objects should be cross-referenced in the database, and these connections can be navigated via hyperlinked text.

Similarly, with the advanced search option users should be able to construct complex queries on specific characteristics of interest and can explicitly exclude undesired phenotypes. In essence this enables users to perform 'digital phenotypic screens' for specific objects of interest. For example, users can search for genes that display RNAi phenotypes indicative of defects in cytokinesis but not other aspects of mitosis. Such search, which would take months on the bench, is taking only minutes on the computer.



**Join the best at  
the Maastricht University  
School of Business and  
Economics!**

#### Top master's programmes

- 33<sup>rd</sup> place Financial Times worldwide ranking: MSc International Business
- 1<sup>st</sup> place: MSc International Business
- 1<sup>st</sup> place: MSc Financial Economics
- 2<sup>nd</sup> place: MSc Management of Learning
- 2<sup>nd</sup> place: MSc Economics
- 2<sup>nd</sup> place: MSc Econometrics and Operations Research
- 2<sup>nd</sup> place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

**Maastricht  
University is  
the best specialist  
university in the  
Netherlands  
(Elsevier)**

**Visit us and find out why we are the best!  
Master's Open Day: 22 February 2014**

[www.mastersopenday.nl](http://www.mastersopenday.nl)



Finally, the use of HCS in basic and applied research for example in drug discovery is only going to increase, but as these data sets grow in size, it is important to recognize that untapped information and potential discoveries might still be present in existing public available data sets (Table 4).

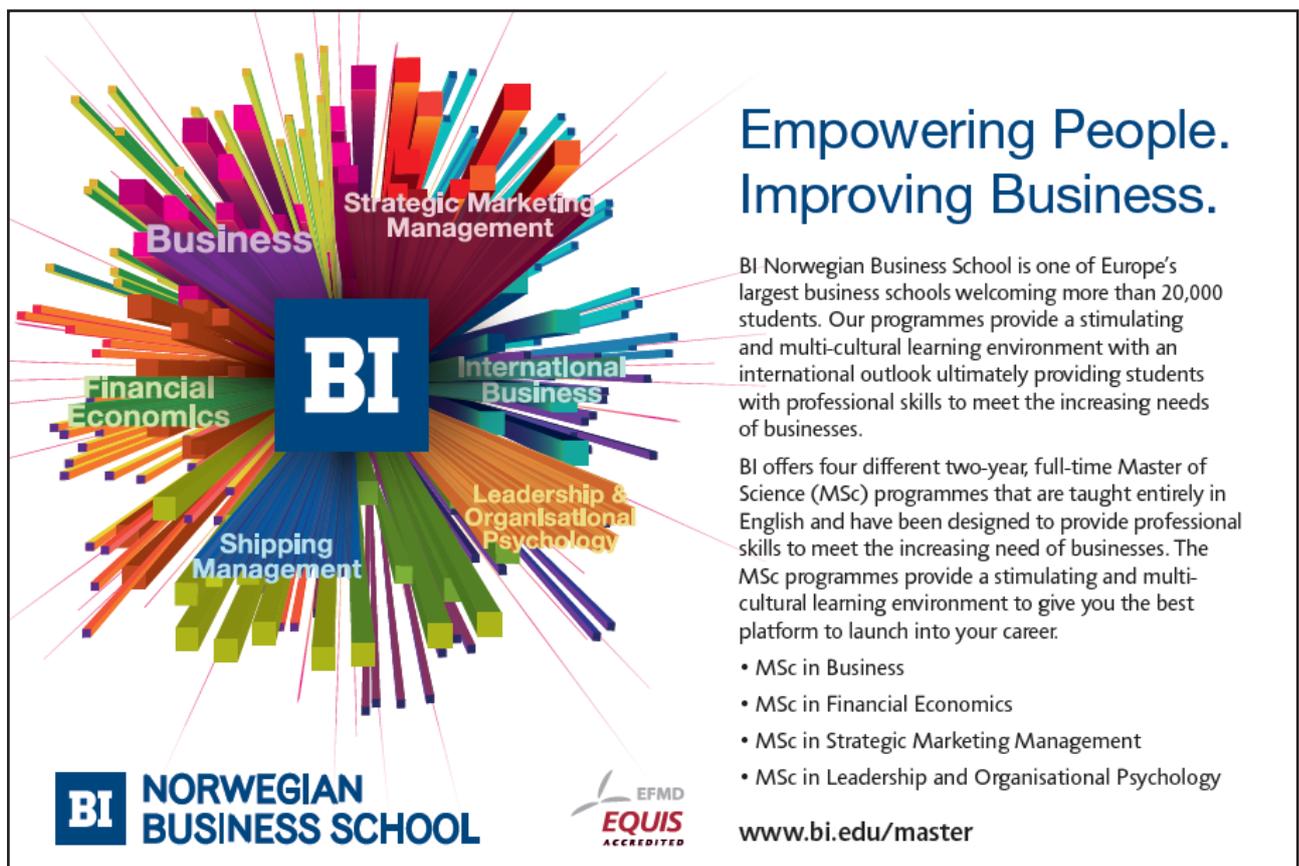
Name	Description	Source
FlyRNAi	Screens carried out in the <i>Drosophila</i> RNAi Screening Center between 2002 and 2006.	<a href="http://flyrnai.org/cgi-bin/RNAi_screens.pl">http://flyrnai.org/cgi-bin/RNAi_screens.pl</a>
DKFZ RNAi	Database contains 91351 dsRNAs from different RNAi libraries targeting transcripts annotated by the Berkeley <i>Drosophila</i> Genome Project	<a href="http://www.dkfz.de/signaling2/rnai/index.php">http://www.dkfz.de/signaling2/rnai/index.php</a>
FLIGHT	FLIGHT is a database that has been designed to facilitate the integration of data from high-throughput experiments carried out in <i>Drosophila</i> cell culture. It includes phenotypic information from published cell-based RNAi screens, gene expression data from <i>Drosophila</i> cell lines, protein interaction data, together with novel tools to cross-correlate these diverse datasets	<a href="http://www.flight.licr.org">http://www.flight.licr.org</a>
PhenoBank	Set of <i>C. elegans</i> genes for their role in the first two rounds of mitotic cell division. To this end, we combined genome-wide RNAi screening with time-lapse video microscopy of the early embryo	<a href="http://www.worm.mpi-cbg.de/phenobank2">http://www.worm.mpi-cbg.de/phenobank2</a>
PhenomicDB	PhenomicDB is a multi-organism phenotype-genotype database including human, mouse, fruit fly, <i>C.elegans</i> , and other model organisms. The inclusion of gene indices (NCBI Gene) and orthologues (same gene in different organisms) from HomoloGene allows to compare phenotypes of a given gene over many organisms simultaneously. PhenomicDB contains data from publicly available primary databases: FlyBase, <a href="http://Flyrnai.org">Flyrnai.org</a> , WormBase, Phenobank, CYGD, MatDB, OMIM, MGI, ZFIN, SGD, DictyBase, NCBI Gene, and HomoloGene.	<a href="http://www.phenomicdb.de/index.html">http://www.phenomicdb.de/index.html</a>
MitoCheck	RNA interference (RNAi) screens to identify all proteins that are required for mitosis in human cells, affinity purification and mass spectrometry to identify protein complexes and mitosis-specific phosphorylation sites on these, and small molecule inhibitors to determine which protein kinase is required for the phosphorylation of which substrate. MitoCheck is furthermore establishing clinical assays to validate mitotic proteins as prognostic biomarkers for cancer therapy.	<a href="http://www.mitocheck.org/cgi-bin/mtc">http://www.mitocheck.org/cgi-bin/mtc</a>

ZFIN	ZFIN serves as the zebrafish model organism database. The long term goals for ZFIN are a) to be <i>the</i> community database resource for the laboratory use of zebrafish, b) to develop and support integrated zebrafish genetic, genomic and developmental information, c) to maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) to facilitate the use of zebrafish as a model for human biology and f) to serve the needs of the research community.	<a href="http://zfin.org">http://zfin.org</a>
MGI	MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease.	<a href="http://www.informatics.jax.org">http://www.informatics.jax.org</a>

**Table 4.** Downloadable large data sets of HCS RNAi screening.

## 5 References

1. Berriman G.B., Good J., Laity A., Jacob J.C. and Katz, D.S.: Pegasus: a framework for mapping complex scientific workflows onto distributed systems. *Sci. Prog. J.* 13 (3), 219–237, 2005.
2. Cputasso C., Alonso G.: The JOpera Visual Composition Language, *Journal of Visual Languages and Computing*, Nov. 2004.
3. CellHTS the open-source Bioconductor/R package and cellHTS49. <http://www.dkfz.de/signaling/cellHTS>.
4. CellAnalyzer Project. <http://www.cellprofiler.org>
5. Chua C.L., Tang F., Lim Y.P., Ho L.Y. and Krishnan, A.: Implementing a BioinformaticsWorkflow in a Parallel and Distributed Environment. *Parallel and Distributed Computing*, 3320. Applications and Technologies of LNCS, Springer, pp. 1–4, 2005.
6. Deelman E., Singh G., Su M., Blythe J., Gil Y., Kesselman C., Mehta G. and Vahi K.: Eclipse Foundation, Eclipse 3.1 Documentation, <http://www.eclipse.org>.
7. Dowell R.D., Jokerst R.M., Day A., Eddy S.R. and Stein L.: The Distributed Annotation System. *BMC Bioinformatics*, 2, 7, (2001).
8. Durinck S., Moreau Y., Kasprzyk A., Davis S., De Moor B., Brazma A., Huber W.: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21 (16), 3439–3440, 2005.



**Empowering People. Improving Business.**

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

**BI NORWEGIAN BUSINESS SCHOOL**

EFMD **EQUIS ACCREDITED**

[www.bi.edu/master](http://www.bi.edu/master)

9. Harold E. and Means W. S.: XML in a Nutshell, Third ed., O'Reilly, Sebastopol, CA, 2004.
10. Dunlay R. T., J. Czekalski W. J. and Collins M. A.: Overview of Informatics for High Content Screening. A Powerful Approach to Systems Cell Biology and Drug Discovery. vol. 356
11. Hassan M., Brown R.D., Varma S., Brien O. and Rogers D.: Cheminformatics analysis and learning in a data pipelining environment. *Mol. Divers.*, 2006, 10 (3), 283–299.
12. Hoon S., Ratnapu K.K., Chia J., Kumarasamy B., Juguang X., Clamp M., Stabenau A., Potter S., Clarke L., Stupka E.: Biopipe: A flexible framework for protocol-based bioinformatics analysis. *Genome Res.* 13,1904–1915, 2003.
13. KNIME (University of Konstanz), <http://knime.org>
14. Lehtovuori P.T. and Nyronen T.H.: Workflow for small molecule property calculations on a multiplatform computing grid. *J. Chem. Inf. Model* 46, 620–625, 2006.
15. Ludascher B., Altintas I., Berkley C., Higgins D., Jaeger E., Jones M., Lee E., Tao J. and Zhao, Y.: Scientific Workflow Management and the Kepler System. *Concurr. Comput.: Pract. Exp.* 18 (10), 1039–1065, 2005.
16. Michalickova K., Bader G., Dumontier M., Lieu H., Betel D., Isserlin R. and Hogue C.: SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinf.* 3 (1), 32, 2002.
17. Neerincx P.B.T. and Leunissen, J.A.M.: Evolution of web services in bioinformatics. *Brief Bioinform* 6 (2), 178–188, 2005.
18. Oinn T., Addis M., Ferris J., Marvin D., Greenwood M., Carver T., Pocock M.R., Wipat A. and Li P.: Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20 (7), 3045– 3054, 2004.
19. Rowe A., Kalaitzopoulos D., Osmond M., Ghanem M. and Guo Y.: The discovery net system for high throughput bioinformatics. *Bioinformatics* 19 (Suppl. 1), 225–1225, , 2003.
20. Tang F., Chua C.L., Ho L.Y., Lim Y.P., Issac P. and Krishnan, A.: Wildfire: distributed, Grid-enabled workflow construction and execution. *BMC Bioinf.* 6 (69), 2005.
21. Saldanha A.J.: Java Treeview – extensible visualization of microarray data. *Bioinformatics* 2004 20(17):3246-3248; doi:10.1093/bioinformatics/bth349
22. Senger M., Rice P. and Oinn T.: Soaplab – a unified sesame door to analysis tools. In: Proceedings of the UK e-Science All Hands Meeting 2003.
23. Shah S.P., He D.Y., Sawkins J.N., Druce J.C., Quon G., Lett D., Zheng G.X., Xu T. and Ouellette, B.F.: Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinf.* 5 (40), 2004.
24. Stevens R.D., McEntire R., Goble C.A., Greenwood M., Zhao J., Wipat A., Li P.: myGrid and the drug discovery process. *Drug Discov. Today: BIOSILICO* 2 (4), 140–148, 2004.
25. Witten I.H. and Frank E.: *Data Mining: Practical Machine Learning Tools and Techniques (Weka)*, 2nd edn. San Francisco: Morgan Kaufmann, , 2005
26. Wilkinson M.D. and Links, M.: BioMOBY: an open-source biological web services proposal. *Brief. Bioinform.* 3 (4), 331–341, 2002.