

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.



.....Alcatel-Lucent 

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".



HCS run: 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	Time: - Image processing time (depend on core processors CPU) - Acquisition time	Number of Images	Number of records	Storage size
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 Total: 140,402 GB
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 Total: 97,3 GB
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 Total: 1312.84 GB = 1.3 TB
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 Total: 875.96 GB = 0.8 TB
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 Total: 3156,3 GB =3.2TB
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	1 363 968	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 Total: 2 009.8 GB =2TB

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.



Maastricht University *Leading in Learning!*

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

Top master's programmes

- 33rd place Financial Times worldwide ranking: MSc International Business
- 1st place: MSc International Business
- 1st place: MSc Financial Economics
- 2nd place: MSc Management of Learning
- 2nd place: MSc Economics
- 2nd place: MSc Econometrics and Operations Research
- 2nd 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



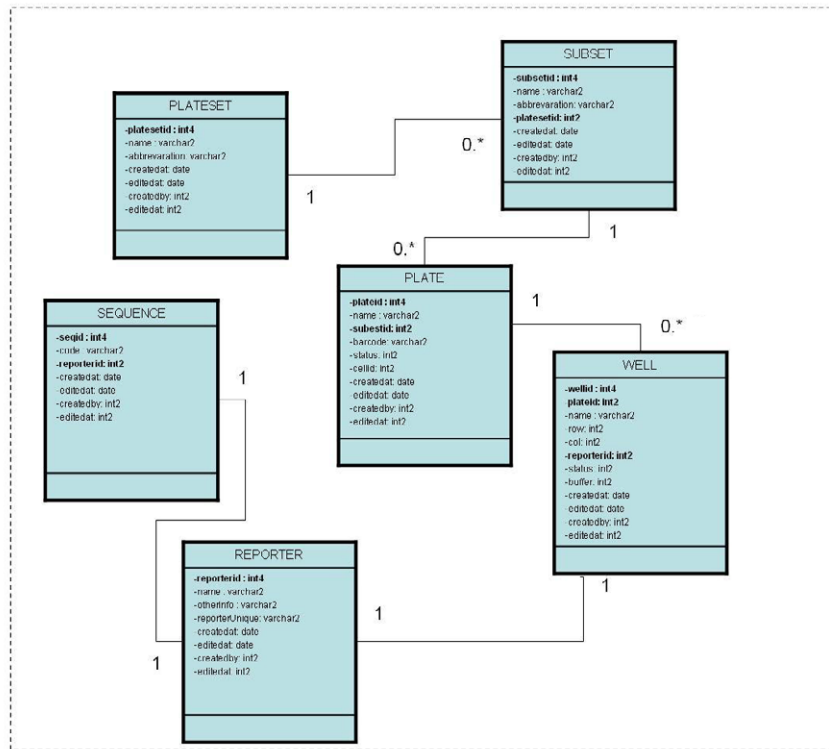


Fig 7: Data model for a Library Database.

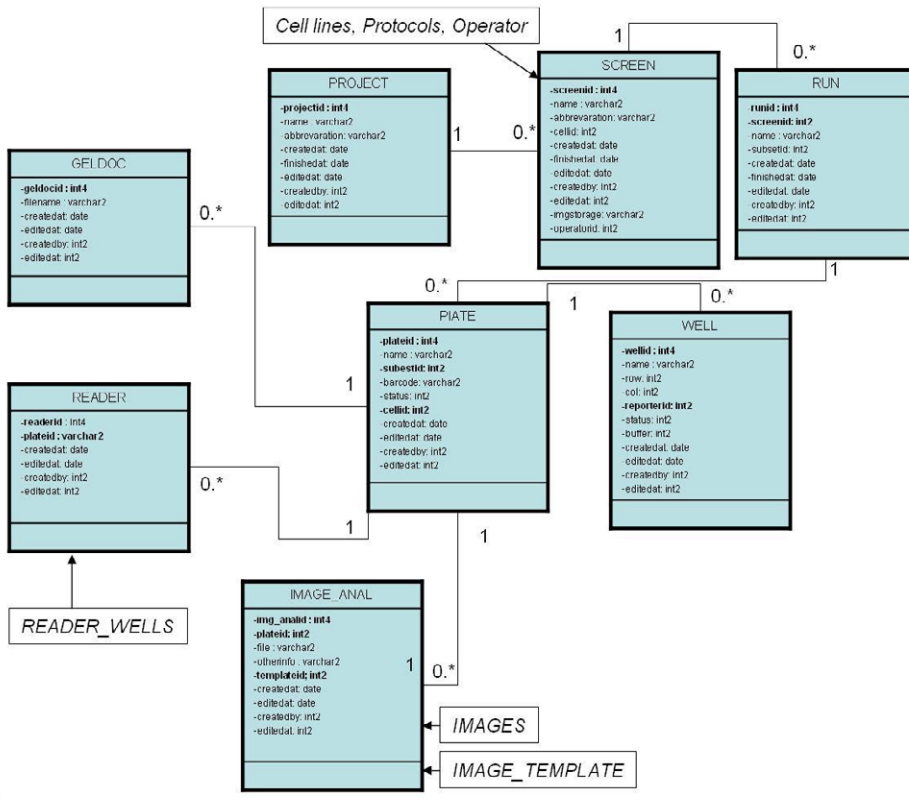


Fig 8: Data model for HCS Results Database.

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.



> Apply now

REDEFINE YOUR FUTURE
AXA GLOBAL GRADUATE PROGRAM 2015

redefining / standards 

agence.cdg © Photonistop



- 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.

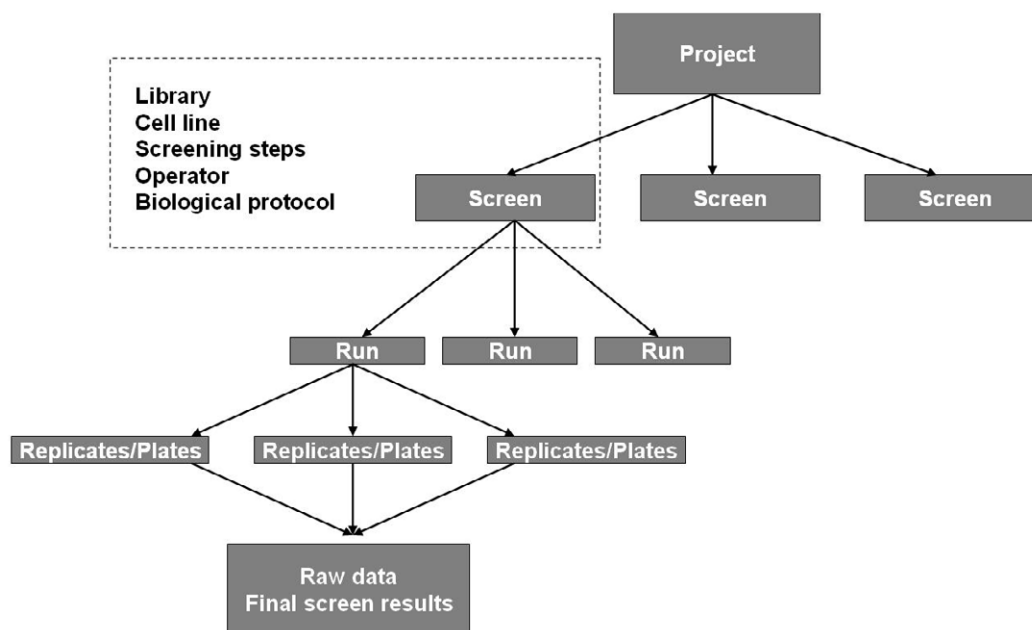


Fig 9. Typical screening hierarchy. Screening parameters are defined on a “screen” level and shouldn’t be modified in sublevels.

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

Software	Supplier	Scope	Description
Acapella™ Columbus	PerkinElmer http://www.perkinelmer.com	High-content data analysis	<ul style="list-style-type: none"> - Columbus is a convenient and easy-to-use solution for high volume data management, storage, retrieval, visualization and protection of images and analyzed results. - 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. - Interactive, fully scriptable and compatible with 3rd-part platform environments. - Upgradeable with user libraries. - Provides high-level language to reduce coding overhead for main applications in image analysis (HCS): object recognition, grouping and segmentation, morphologic filtering, image arithmetic. - Libraries available also for Photon Statistics or specific applications like FLIPR kinetics analysis. - SDK available.



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

www.bi.edu/master



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

Software	Supplier	Scope	Description
Genepattern (GP)	NIH grant project Genepattern (GP) http://www.broad.mit.edu/genepattern/	Data management and analysis	<ul style="list-style-type: none"> - Workflow management system for data analysis and visualization. - 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.
Synapsia Informatics Workbench	Agilent http://www.chem.agilent.com	Knowledge management	<ul style="list-style-type: none"> - 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). - The open architecture and documented APIs enable integration of external tools for (e.g. BLAST searches). - Together with Information Manager it represents a collaboration framework for cross-discipline R&D projects.
Foundation Server	TriposTM http://www.tripos.com	Small Molecular Screening, Cheminformatics, computational chemistry	<ul style="list-style-type: none"> - Application server integrates tools and provides access to third-party discovery informatics software. - Foundation Server SYBYL as an optional environment provides tools for molecular modeling and cheminformatics
EMC2	Documentum http://www.documentum.com	Content management	<ul style="list-style-type: none"> - Managed collection of software tools to organize - unstructured information originating from sources like documents, spreadsheets, web pages or e-mail databases according to defined business rules. - Documentum Creates relationships, organizes metadata and provides tools for search, retrieval and presentation.

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 Language⁹) 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).

Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

Get Help Now



Go to www.helpmyassignment.co.uk for more info



Helpmyassignment



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 images¹⁰.

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 model⁷, 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.



Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF

Download free eBooks at bookboon.com



Click on the ad to read more

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.	http://flyrnai.org/cgi-bin/RNAi_screens.pl
DKFZ RNAi	Database contains 91351 dsRNAs from different RNAi libraries targeting transcripts annotated by the Berkeley <i>Drosophila</i> Genome Project	http://www.dkfz.de/signaling2/rnai/index.php
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	http://www.flight.licr.org
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	http://www.worm.mpi-cbg.de/phenobank2
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, Flyrnai.org , WormBase, Phenobank, CYGD, MatDB, OMIM, MGI, ZFIN, SGD, DictyBase, NCBI Gene, and HomoloGene.	http://www.phenomicdb.de/index.html
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.	http://www.mitocheck.org/cgi-bin/mtc

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.	http://zfin.org
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.	http://www.informatics.jax.org

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