National Physical Laboratory

NPL REPORT MS 25

ACQUISITION & MANAGEMENT OF HIGH CONTENT SCREENING, LIGHT-SHEET MICROSCOPY AND MASS SPECTROMETRY IMAGING DATA AT ASTRAZENECA, GLAXOSMITHKLINE AND NPL

SURVEY REPORT

COOKE, E

HAYES, M A

ROMANCHIKOVA, M

APRIL 2020

Department for
Business, Energy
& Industrial Strategy

FUNDED BY BEIS

Acquisition & Management of High Content Screening, Light-Sheet Microscopy and Mass Spectrometry Imaging Data at Astrazeneca, Glaxosmithkline and NPL

NPL authors:
Elizabeth Cooke, Mark Hayes, Marina Romanchikova
Data Science

Alex Dexter, Rory Steven, Spencer Thomas
National Centre of Excellence in Mass Spectrometry Imaging

Mike Shaw
Biometrology

AstraZeneca contributors:
James Pilling, Stephanie Ling, Nicole Strittmatter, Delyan Ivanov, Yinhai Wang, Alan Race

GlaxoSmithKline contributors:
Donna Fraser, Jo Francis, Carla Newman, Andy West, Joseph Lavelle

ISSN 1754-2960

https://doi.org/10.47120/npl.MS25

National Physical Laboratory

Hampton Road, Teddington, Middlesex, TW11 0LW

This work was funded by the UK Government's Department for Business, Energy and Industrial Strategy (BEIS) through the UK's National Measurement System programmes.

Approved on behalf of NPLML by

Louise Wright, Data Science team Head of Science.

CONTENTS

GLOSSARY/ABBREVIATIONS

Assay: an experimental procedure with the aim of determining a single property (e.g. concentration) of a target, often in response to a reagent.

Biomarker: a molecule, gene, or characteristic by which a biological process or condition can be identified.

Channel: in microscopy, an imaging channel is identified by the wavelength of light recorded (c.f. red/green/blue channels in an RGB image) or the name of the fluorescent dye associated with that wavelength.

Confocal Microscopy: a technique that raster scans across a sample by illuminating only a very small area at a time through a pinhole aperture of variable size. This allows a controlled depth of field, reduction of light emitted outside the focal plane and the ability to collect a series of two-dimensional images from a three-dimensional sample.

Da/dalton: a unified atomic mass unit, approximately equivalent to a mass of a single nucleon (proton or neutron).

DESI: “Desorption ElectroSpray Ionization”, an ionisation technique used in mass spectrometry whereby a sample is sprayed with an electrically charged solvent mist, causing the emission of secondary ions from the sample which are used to detect the mass spectrum of the sample.

Field of View: the diameter of the visible portion of a sample under a microscope. It is inversely proportional to magnification.

High Content Screening (HCS): a method used to identify substances that alter the phenotype of cells within a sample in a desired manner, using imaging across many simultaneous samples under differing conditions or across a timeseries.

Light Sheet Microscopy (LSM): a technique that uses a laser to illuminate a sample in a thin sheet arranged perpendicularly to the detector. In comparison to confocal microscopy, which requires raster scanning across the imaging plane, LSM results in faster scan times and lower phototoxicity.

MALDI: “Matrix Assisted Laser Desorption/Ionization”, an ionisation technique used in mass spectrometry whereby a laser energy absorbing matrix material is applied to or mixed with a sample before the laser is applied. The matrix material enables a greater concentration of unfragmented ions from the sample to enter the gaseous phase and enter the mass spectrometer than would be the case for an untreated sample.

Mass Range: the upper and lower bounds on the mass (or mass/charge ratio) of ions that a mass spectrometer is capable of detecting.

Mass Resolution: a measure of the ability of a mass spectrometer to distinguish two peaks of slightly different mass/charge ratio. A higher resolution implies a better separation of peaks.

Mass Spectrometry Imaging (MSI): a method of visualising the spatial distribution of a molecule of interest across a biological sample. Traditional mass spectrometry is used to collect a mass spectrum at each point across a sample.

Phenotype: a set of observable characteristics of a biological sample or organism resulting from its interaction with the environment. For a sample analysed in a high content screening experiment, this could include observed number of cells, cell shape and behaviour over time.

Protocol: an explicit, detailed, often standardised and well documented plan for an experiment, procedure or test.

SIMS: “Secondary Ion Mass Spectrometry”, an ionisation technique used in mass spectrometry whereby the ions released from a sample by irradiation with a high energy primary ion source are collected and analysed. For mass spectrometry imaging purposes, SIMS has the highest spatial resolution of the three common techniques but is only useful for thin slices and over small areas.

Target: in High Content Screening, the molecule or molecular complex within a sample that is tested for its reaction against a reagent under many different, simultaneous environmental conditions.

Well Plate: a rectangular plate or tray, often made of plastic, with multiple wells that act as small receptacles for samples and reagents in a High Content Screening experiment. Commonly used well plates have arrays of 96, 384 or 1536 wells. Also known as micro-titre plates.

Widefield Microscopy: widefield microscopy refers to the simultaneous illumination and imaging of a whole sample, rather than a smaller area of interest or two-dimensional slices through a sample (as possible with confocal or light sheet microscopy).

XML: eXtensible Markup Language format to store and annotate data developed by W3C consortium.

EXECUTIVE SUMMARY

This survey was conducted between October 2018 and April 2019 as a part of a collaborative project named “Federation of Imaging Data for Life sciences research” (FIDL). During the survey, the National Physical Laboratory (NPL) Data Science team interviewed scientists from AstraZeneca (AZ), GlaxoSmithKline (GSK) and NPL bio-metrology group to describe the landscape of the instruments and data management methods in three imaging domains: high content screening, mass spectrometry imaging and light-sheet microscopy.

The purpose of the survey is to identify the key factors and challenges in bioimaging data management in scientific, operational and business context. The gained knowledge of bioimaging methods and instruments in the inspected domains will enable identification of measurement and knowledge management needs that can benefit from NPL’s metrology expertise.

1 HIGH CONTENT SCREENING

High-content screening, or HCS, is a method that is used in biological research and drug discovery to identify the effects of a reagent (e.g. a potential new drug) on “targets” such as small molecules or antibodies. Interventions such as gene knockout or RNA interference may also be tested. The resulting changes to the phenotype of a cell are observed in a controlled manner. Phenotypic changes may include an increase or decrease in the production of cellular products such as proteins and/or changes in the morphology of the cell. High content screening includes any method used to analyse individual cells or components of cells with simultaneous readout of several parameters and includes wide-field and confocal imagers (Zock, 2009), as well as laser-scanning cytomers. The goal of HCS is to detect and quantify critical features such as number of objects, their shape, texture, colour, size or intensity from a cell population in a short time interval (Buchser, 2014).

Cell samples are normally arranged in an array of several hundred wells on a plate, so that the wells containing variable concentrations of a reagent, for example, can be analysed in parallel across the whole plate. Each well can have multiple Fields of View (FoVs), the number of which will depend on the imaging system used. The images are acquired using different filter settings with variable exposure times or laser power. Fluorescent dyes are used to highlight subcellular structures or protein complexes.

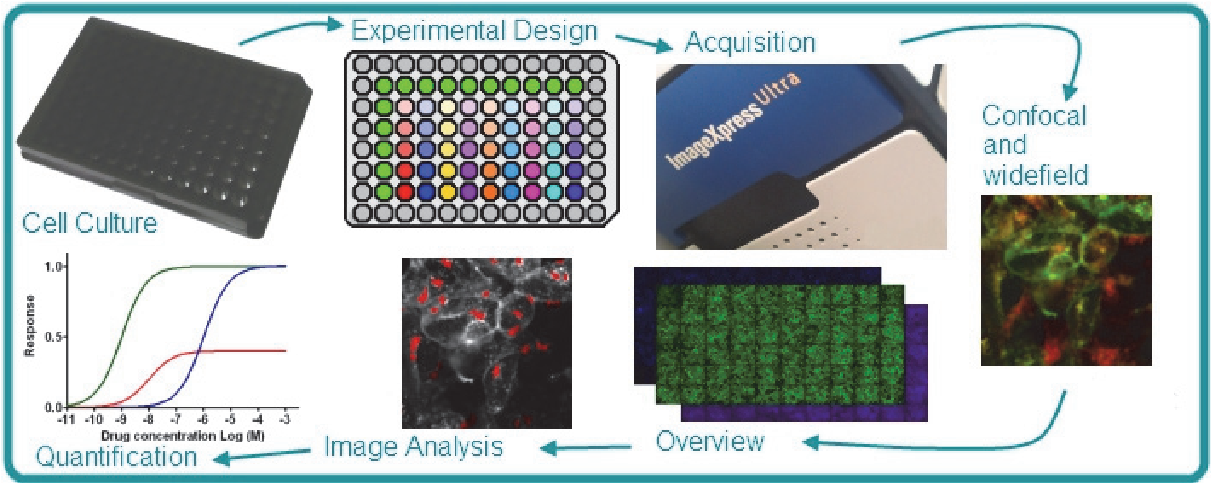

Figure 1: Overview of High Content Screening experimental processes
Source: https://www.nottingham.ac.uk/life-sciences/facilities/slim/cell-signalling-imaging/high-content-screening/

Typical image acquisition scenarios include repeated imaging of a FoV over multiple timepoints or at multiple depths in a sample to capture 3D information. During an experiment, images are acquired at different wavelengths, sometimes described as channels. This allows the response of various fluorescent dyes or structures in the sample to be evaluated, identifying different biomarkers. A low-resolution image may be taken prior to an experiment to identify the wells or FoVs of interest and to reduce the imaging time.

1.1 HIGH CONTENT SCREENING INSTRUMENTS

This section provides an overview of the instruments used at AZ Cambridge and GSK Stevenage. There is currently no HCS instrumentation at NPL.

1.1.1 HCS instruments at AZ Cambridge

The HCS facilities in Cambridge run CV7000 & CV8000 Yokogawa instruments, a CellInsight ThermoFisher instrument, an ImageXpress XL instrument by Molecular Devices and an IncuCyte Zooms instrument by Essen BioScience. Beyond these devices, AZ Cambridge scientists also have access to other high content cellular imaging systems such as the Perkin Elmer Operetta CLS located in Boston, USA and ImageXpress Confocal instruments located in Gothenburg, Sweden.

Yokogawa CV7000 and CV8000

The Yokogawa instruments comprise a high-end high throughput confocal imager, with an environmental chamber for temperature and CO2 control. For live cell imaging and z- plane acquisition there are 4/5 UV and optical lasers, 2 cameras and up to 6 automatically switched objectives with magnifications of 4x, 10x, 20x, 20x Ph-LWD[1], 40x in water and 60x in water. They are used for biological applications requiring high resolution, 3D imaging and live cell capability. They export data as ".tiff" images and metadata files in plain text or XML formats within a folder-based structure.

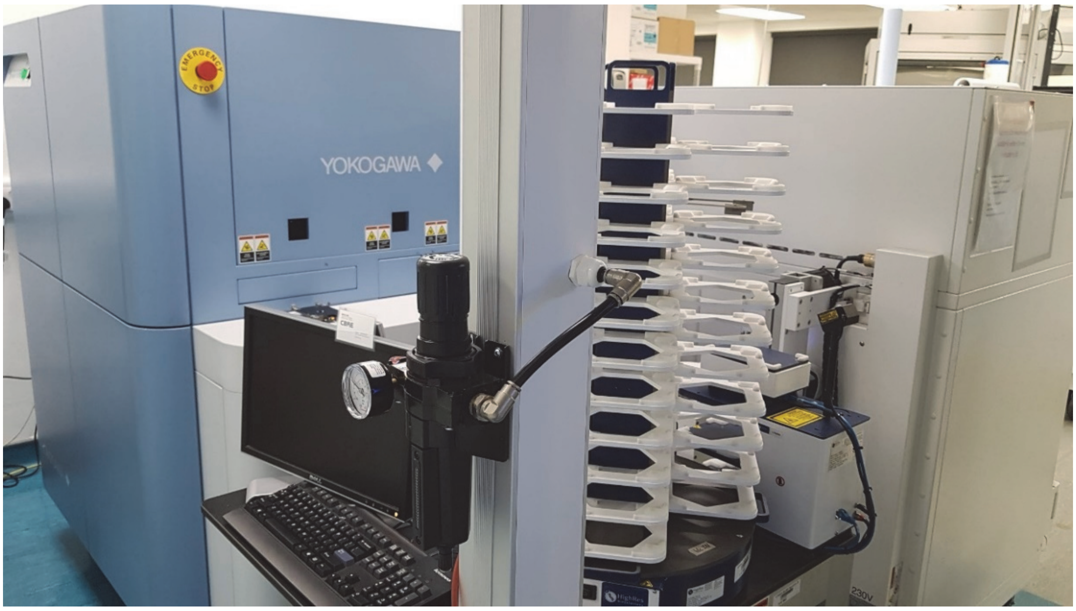

Figure 2: A Yokogawa High Content Screening platform. Left to right: microscope, instrument computer, well plate holder, incubator.

[1]: Phase-contrast long working distance objective

CellInsight

The CellInsight ThermoFisher platform is an automated widefield imaging system with built in image analysis capability. It has an LED light source and manually switched objectives. It is typically used for applications requiring a fast data turnaround, e.g. Design-Make-Test-Analyse drug discovery cycle and high throughput screening. Images are stored in a shared location and metadata is stored within a MS-SQL based database called Cellomics Store.

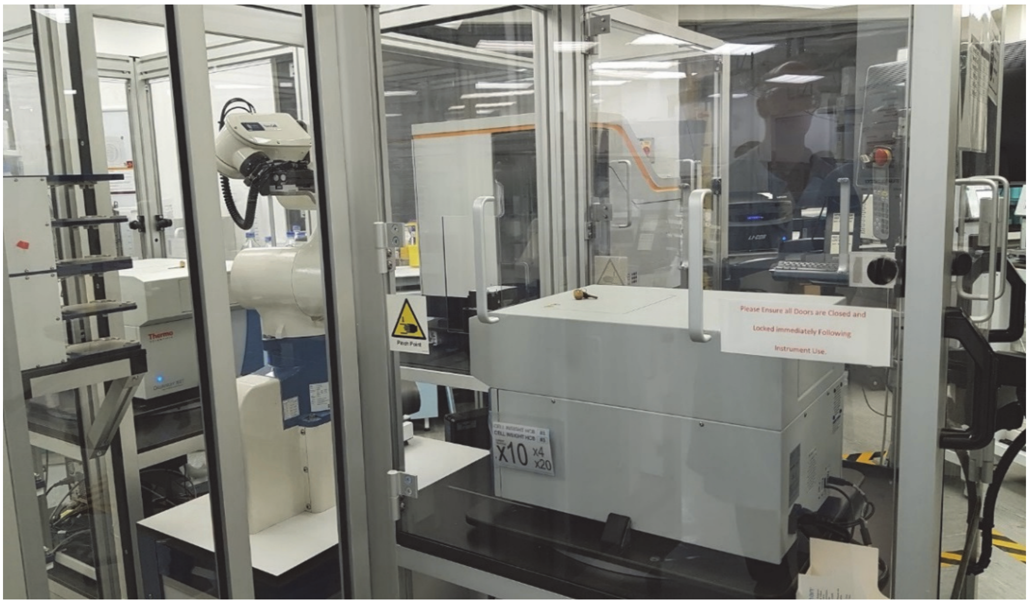
Please Ensure all Doors are Closed and
Locked Immediately Following
Instrument Use.
X10 x4
x20
Thermo

Figure 3: A CellInsight High Content Screening platform.

ImageXpress

This is an automated widefield imaging system with LED light source and automatically switched objectives. Images are stored in a shared location and metadata is stored within a proprietary Oracle-based database called MDCStore.

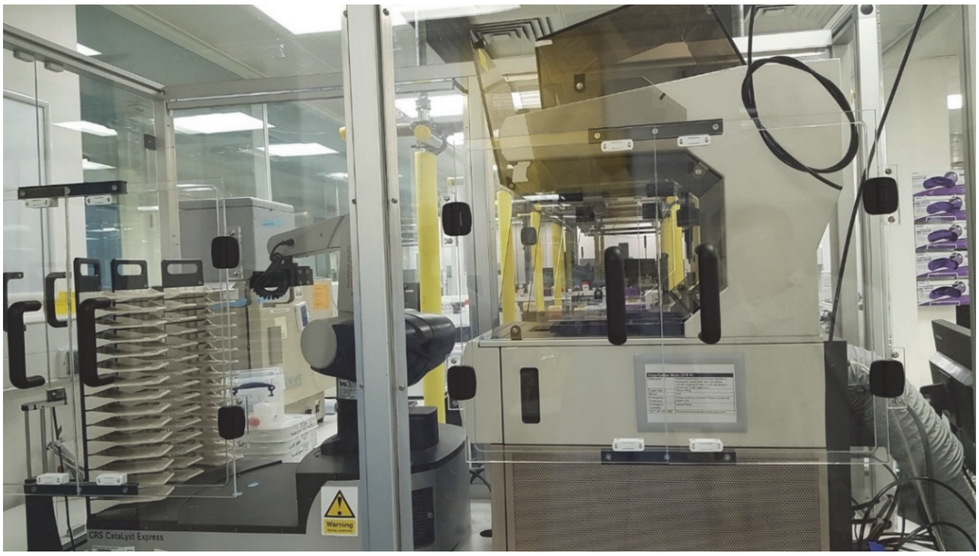

Figure 4: An ImageXpress High Content Screening platform

1.1.2 HCS instruments at GSK Stevenage

While the GSK laboratory in Stevenage runs numerous HCS devices, this report focusses on the InCell systems by GE Healthcare and IncuCyte Essens BioScience instruments.

IncuCyte

The IncuCyte device is a standalone fluorescence microscope for imaging live tissue culture flasks and microtitre plates. The imaging system is enclosed within an incubator to maintain constant temperature and humidity. The *IncuCyte Zoom* used at GSK Stevenage includes a robotic sample delivery mechanism, while the AZ device does not have one. IncuCyte devices are typically used for kinetic assays and quality control of tissue cultures, for example plotting growth curves and generation of cellular reagents. Images and metadata are stored locally on the system hard drive.

InCell systems versions 2000, 2100 and 6000 (GE Healthcare)

The 2000 and 2100 devices have similar functionality, whereas the 6000 device has a confocal imaging feature that allows z-stack imaging. These devices have a self-calibration feature using the well plate as a test object for flat field intensity correction, compensating for the higher intensity in the image centre. shows an InCell 6000 device featuring a well plate rack on the left, a robotic arm in the middle and an enclosed microscope on the right.

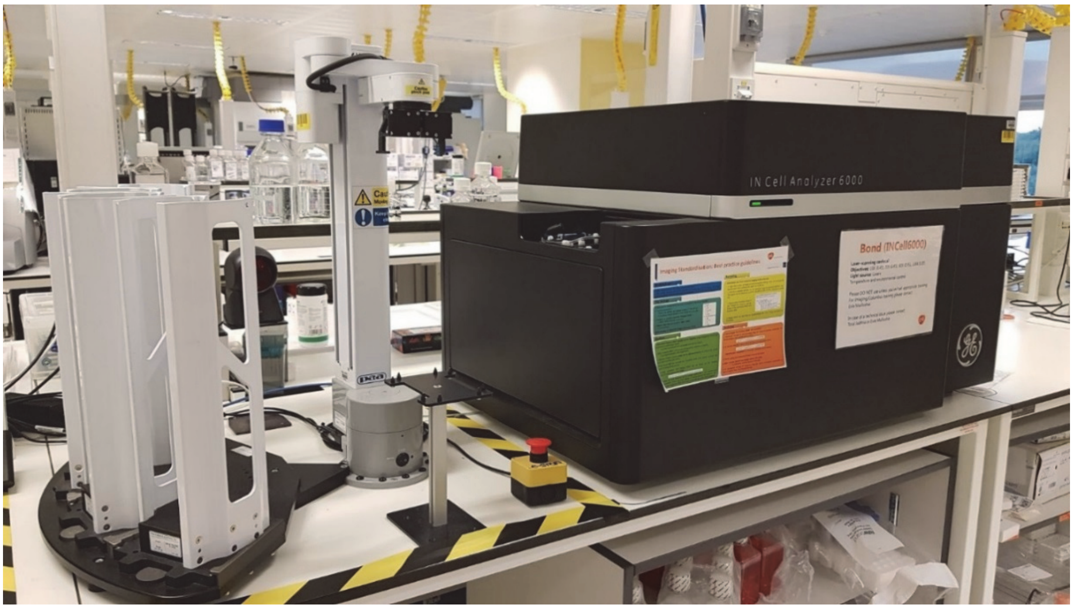

Figure 5: An InCell 6000 High Content Screening platform

Both devices are fitted in an open-access lab and are used by multiple GSK teams.

1.2 HCS DATA AND METADATA MANAGEMENT

HCS datasets are characterised by their large size, often amounting to thousands of gigabytes per assay. The size of a dataset from an individual experiment depends on the data format used to store the image data, on the number of plates, wells, imaging time-points, wavelengths, fields of view and the number of imaging planes across each well.

The number of images generated during an HCS experiment will be dependent on the experimental setup which can be characterised by the parameters listed below:

- Number of wells, typically 24, 96, 384 or 1536 per plate [W]
- Number of fields of view imaged, typically three, theoretically it could be hundreds [FOV]
- Number of channels, typically four [C]
- Number of time points [T]
- Number of images taken through a 3D object, i.e. sections through the object images [Z]

For example, if 3 imaging planes at different depths were taken through each well, using 16 fields of view, in every well of a 384 well plate for 4 channels at 16 different time points, this will create Z × FOV × C × W × T = 1,179,648 files for one plate. If there are 200 plates in the assay, the total data volume amounts to 450 terabytes (*number of files × number of plates × file size = 1,179,648 × 200 × 2 = 471,859,200 bytes*).

These considerable data volumes can increase in size if an assay is acquired as a time series, where plates are imaged over longer periods of time to track slower changes such as differentiation or cell division. The acquisition time will vary depending on the number of FOVs, wells, channels and other experiment settings and can range from 2 minutes to several hours per plate.

Metadata is captured in a variety of ways, both automatically by the instruments and manually by the operator. For example, instruments typically output their configuration settings automatically as part of every dataset. Folder structures and filenames often contain information on the structure and experimental context of a dataset. Details of the protocol and samples used are captured, often in free text form, within an Electronic Lab Notebook or ELNB.

Table 1 summarises the software packages in use at AstraZeneca and GSK for the management and post-processing of HCS data. Both companies use Columbus for image archival & retrieval and post-processing, alongside proprietary software supplied by the instrument manufacturers, freeware and in-house software developed in MATLAB or Python.

Table 1: Software in use at AstraZeneca Cambridge & GSK Stevenage.

Software categories	**AstraZeneca**	**GSK**
Multi-purpose image analysis software	Columbus	Columbus
Electronic laboratory notebook	E-Notebook (PerkinElmer)	E-Notebook (PerkinElmer)
In-house bespoke software	MATLAB (MathWorks)	Python
Commercial software	Screener (GeneData) HCS Studio (Cellomics)	Analyzer, Developer (GE)
Freeware	OMERO	Fiji

1.2.1 HCS data management and post-processing at AZ

AstraZeneca uses Columbus software to import individual images and read the embedded equipment metadata from Yokogawa, Cellomics and ImageXpress devices. However, the proprietary metadata which describes the context of the HCS experiment, i.e. the specific treatment that is applied to each well and FoV, is not implemented within Columbus.

Raw data are stored on their Storage Area Network, or SAN, which includes the instrument computers. AZ has three synchronised SAN nodes located in the UK, Sweden and USA, while the Medimmune infrastructure in Boston is run independently. The current storage systems offer little to no search functionality. The AZ Scientific Computing Platform is being rolled out to address these issues.

Next Generation Sequencing is the first AZ team to use the Scientific Computing Platform, or SCP and the SwiftStack Object Store with Elasticsearch function to store and locate data, followed by the Cellular Imaging division including HCS. The network hardware infrastructure has been upgraded in 2018, and the bandwidth between Cambridge and Manchester is sufficient to support the migration to the Object Store.

The Discovery Science group at AZ is mandated to use the Bio-ELN Electronic Notebook software by CambridgeSoft owned by PerkinElmer. The data in the notebooks can include Excel, Microsoft Word, PDF, images and other documents. The notebook records contain several mandatory fields to capture the minimum metadata required, including project codes, experiment dates, contributors, targets and substances used. The ELN software allows users to search for common terms within the minimum metadata populated by links to other internal databases.

Data retention policies are in place at AZ but are difficult to enforce as data owners are responsible for deleting their own data after its useful lifetime. Typically, experimental raw data is kept for a period of one year after the acquisition. Important refined datasets and summary documents are expected to be kept for 75 years.

AZ is aware that HCS datasets may have considerable value that supports the case to store them for a longer period. An example of unlocking the value from historic raw data is the generation of “virtual plates” to infer the knowledge about a cell type, gene or target from multiple experiments. This method is akin to a meta-study in the clinical domain, and the quality of the prediction relies on the completeness and the accuracy of the experimental metadata.

In recognition of the increased use of HCS data for knowledge inference and machine learning applications, AZ are working towards a 10-year retention policy for certain reference datasets.

1.2.2 HCS data management and post-processing at GSK

The workflow for an HCS experiment at GSK starts with an Electronic Lab Notebook (ELNB) record with the experiment name and protocol details. The experiment on the HCS machine will then be attributed to an existing ELNB record and will include well plate barcodes. Other GSK groups and labs may not follow this workflow.

HCS data at GSK is organised in a hierarchical folder structure: "experiment -> plate -> FOV images for each well". The FOV images are sorted by well number and FOV number. A typical single FOV file is 8 MB, although the file size can be reduced to 2 MB by changing the underlying number type and number of bytes stored per pixel.

The IncuCyte instruments export imaging data and the associated metadata in various proprietary formats and rely on the manufacturer’s proprietary software to analyse the images. The image and metadata are challenging to export and analyse in third party platforms such as Columbus software, as the metadata are spread across several files.

The InCell instruments used at GSK generate metadata in “.xdce” files in XML format which is supported by Columbus. The fields within the “.xdce” file include exposure time, time at the start of the experiment, FOVs and other device information. However, in the InCell user interface some entries such as laser power levels are user editable. Thus, the saved metadata might not reflect an actual experimental setup and should be analysed with great care. It should be possible to deduce the laser power level from other metadata recorded by the device. This is not the case for other HCS systems used across GSK.

Other relevant metadata are widely distributed across unlinked databases. At GSK these include compound registration within the Mosaic database, protocols and methods via PACT and many others. The PACT database was originally developed by GSK and contains all metadata essential to reproduce the experiment. However, the data stored in PACT are not standardised, do not follow controlled dictionaries or ontologies and contain mainly free text descriptors. Furthermore, data may be associated with several project IDs over its lifetime. Therefore, the use of a master database to keep track of the data provenance and lifecycle is essential.

GSK operate a laboratory-wide SAN and are testing data transfers to a centralised data centre in Stevenage with a view to use Microsoft Azure cloud for data processing. GSK's data centre is based on a Dell ECM Object Store that hosts all raw imaging data and ELNB data collected from instrument computers from Stevenage laboratories.

2 MASS SPECTROMETRY IMAGING

Mass Spectrometry Imaging (MSI) provides the capability to detect, quantify and visualise the spatial distribution of thousands of molecules across a tissue sample by collecting a mass spectrum at multiple points of a user-defined grid (Buchberger, 2018). MSI methods can deliver high-resolution information about metabolites, lipids, peptides, proteins and glycans in a label-free manner with minimal sample preparation. MSI is also compatible with many other imaging modalities and form of histological and immunostaining.

MSI instruments are classified by a) the ionization technique used within the device and b) the methods for performing mass-to-charge separation. The three primary ionization techniques are listed below.

1. MALDI: Matrix-assisted laser desorption ionisation. A laser is used to desorb / ablate material from thin tissue sections after deposition or mixing with a ‘matrix’; a chemical employed to enhance desorption and ionisation of compounds of interest. Data are typically acquired with 10-50 µm pixel size, though recent studies have shown acquisition of lipid ion images at 500 nm pixel size. MALDI may be used for analysis of small metabolites, drugs, lipids, peptides, proteins and other compound classes.

2. DESI: Desorption electrospray ionisation. Here the sample is probed with an electrically charged solvent spray (electrospray), whereby the ions are extracted from a thin tissue section through a combination of a solubilisation and ballistic ejection. DESI is an ambient technique requiring minimal sample preparation and so is suitable for sample forms not readily compatible with MALDI or SIMS as well as more rapid or pseudo on-line analysis. This comes with a cost of reduced image resolution where analytical pixel sizes are typically in the region of 50 – 150 µm. DESI is particularly strong in highly sensitive analysis of small metabolites, drugs and lipids.

3. SIMS: Secondary Ion Mass Spectrometry. In SIMS, a high energy primary ion source is used to produce secondary ions from the sample. The secondary ions are collected and analysed. SIMS, like MALDI is best suited to thin tissue slices and is particularly useful for detecting small molecules or atomic species. It has the highest spatial resolution of the three techniques, at between 50 nm and 5 µm, depending on the SIMS instrument and modality employed. SIMS also has a long track record for analysis of inorganic materials with a 3D depth profiling mode commonly used to look at interfaces between layers in, for example, electronic devices.

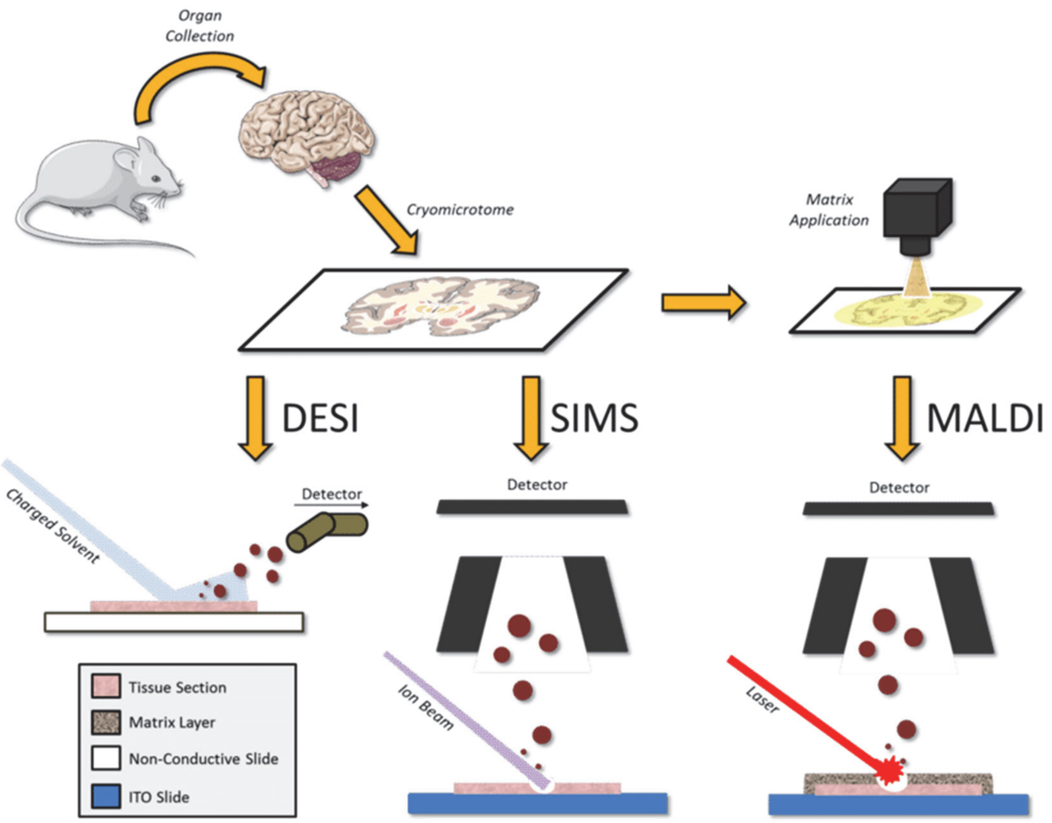

Figure 6: The three primary mass spectrometry imaging techniques (Barré, 2017)

Thin tissue sections are the most commonly analysed sample type. Frozen samples are cut into thin sections of 5 – 15 µm thickness in a cryo-microtome. Prior to an MSI scan, an optical image guidance scan of the entire slide is taken to determine which areas of the slide are to be imaged. In some laboratories, consecutive sections may be taken for correlative imaging or analysis by other MSI modalities or other analytical techniques including light microscopy or liquid-chromatography mass spectrometry (LC-MS) to provide complementary information.

The size of the data acquired as an MS image is dependent on the image resolution, the organ or tissue section size, the image ROI size selected, as well as the number of slides in the batch. Depending on these factors, the duration of an image acquisition can range from several minutes to several days. Consequently MSI data analysis presents a considerable challenge due to large data sizes, use of thousands of variables and inter-laboratory variations between datasets acquired using similar or identical samples, experimental protocols, and instrumentation (A. Buck, 2018).

2.1 MASS SPECTROMETRY IMAGING INSTRUMENTS

Table 2: A concise summary of the mass spectrometry instruments in use at AstraZeneca Cambridge, GSK Stevenage and NPL Teddington sites. Note: human readable meaning via common software such as Notepad or similar, not via proprietary or MS specific software. Some values may be approximate or specific to only certain iterations of an instrument or software version. *Courtesy of Rory Steven & Alex Dexter (NPL NiCE-MSI team).*

Site	Instrument	Modalities enabled	Primary mass analyser & approx. range	Data & metadata	Proprietary data processing software	imzML export
AZ	2x Q-Exactive ThermoFisher	ESI, LESA	Orbitrap *m/z* <6000 *m/Δm* <240,000	File format: .raw (Thermo format) Metadata included: yes Human readable: no	Xcalibur: spectral visualisation and basic analysis, no image analysis provision	No
	1x Q-Exactive Plus ThermoFisher	DESI, ESI	Orbitrap *m/z* <6000 *m/Δm* <240,000	File format: .raw (Thermo format) Metadata included: yes Human readable: no	Xcalibur (Thermo) [as above]	No
	1x Rapiflex TissueType Bruker	MALDI	ToF *m/z* > 200,000 *m/Δm* <50,000	File format: Metadata included: yes Human readable: no	flexImaging: image + spectral visualisation and limited analysis SCiLS Lab: MS image data analysis, univariate and MVA provision	Yes
GSK	1x SolariX Bruker	MALDI, ESI	FT-ICR *m/z* <3000 *m/Δm* <10,000,000	File format: . Metadata included: yes Human readable: no	flexImaging. SCiLS Lab [as above]	Yes
	2x UltraFlex Bruker	MALDI	ToF *m/z* < 200,000 *m/Δm* < 40,000	File format: Metadata included: yes Human readable: no	flexImaging. SCiLS Lab [as above]	Yes
	1x RapiFlex Bruker	MALDI	ToF *m/z* < 200,000 *m/Δm* < 50,000	File format: Metadata included: yes Human readable: no	flexImaging. SCiLS Lab [as above]	Yes

NPL	3x Synapt G2 Si Waters	MALDI, DESI, ESI, LESA	QToF m/z <6000 $m/\Delta m$ < 50,000	File format: .raw (Waters format) Metadata included: yes Human readable: yes (partially)	MassLynx (Waters): spectral visualisation and basic analysis HDI (Waters): MS image data analysis, some MVA provision	Yes
	2x Xevo G2 XS Waters	DESI, ESI	QToF m/z <6000 $m/\Delta m$ < 40,000	File format: .raw (Waters format) Metadata included: yes Human readable: yes (partially)	MassLynx (Waters) HDI (Waters) [as above]	Yes
	1x Orbitrap Elite ThermoFisher	AP-MALDI, ESI, LESA	Orbitrap m/z <2000 $m/\Delta m$ < 240,000	File format: .raw (Thermo format) Metadata included: yes Human readable: no	Xcalibur (Thermo) [as above]	No
	1x Q-Star XL Sciex	MALDI, ESI	QToF m/z <4000 $m/\Delta m$ < 15,000	File format: .wiff Metadata included: yes Human readable: no	Analyst (Sciex): spectral visualisation and basic analysis, no image analysis provision	No
	IONTOF IV	SIMS	ToF m/z < 14,000 $m/\Delta m$ < 10,000	File format: IONTOF .itmx or .itax Metadata included: yes Human readable: yes (if exported to .gdr)	SurfaceLab (IONITOF): spectral visualisation and basic analysis image data analysis, some MVA provision	Yes
	IONTOF V	SIMS	ToF: m/z < 14,000 $m/\Delta m$ < 10,000	File format: IONTOF format [as above]	SurfaceLab (IONITOF) [as above]	Yes
	3D OrbiSIMS	SIMS	Orbitrap: m/z <6000,	File format: Thermo format, IONTOF format [as above]	SurfaceLab (IONITOF) Xcalibur (Thermo)	Yes

			m/Δm < 240,000 ToF: *m/z* <14,000 *m/Δm* < 10,000		[as above]	

2.1.1 MSI instruments at AZ Cambridge

AZ has two mass spectrometry imaging groups in Cambridge and Gothenburg. There are two device types at AZ Cambridge which are used most commonly across all projects: 1) Prosolia DESI ionisation sources coupled to ThermoFisher quadrupole Orbitrap mass analysers and 2) a MALDI Time-of-Flight (MALDI-TOF) Rapiflex TissueType device by Bruker. The Bruker Rapiflex TissueType is a MALDI-TOF device. It comes as a fully integrated machine and is characterised by lower mass resolving power, higher scan speed and higher spatial resolution with pixels of 5 - 50 µm diameter compared to the 30 - 150 µm pixel sizes employed in DESI MSI. AZ also works with a collaborator who uses a MALDI-Fourier transform-ion-cyclotron-resonance (MALDI-FTICR) Bruker instrument.

2.1.2 MSI instruments at GSK Stevenage

GSK uses a Bruker Solarix MALDI instrument that features a Fourier-transformed ion cyclotron resonance detector (Figure 7), as well as MALDI-TOF Bruker UltraFlex (2x) and a MALDI-TOF Bruker RapiFlex. The file sizes produced by the Solarix are of the order of 100 GB or larger.

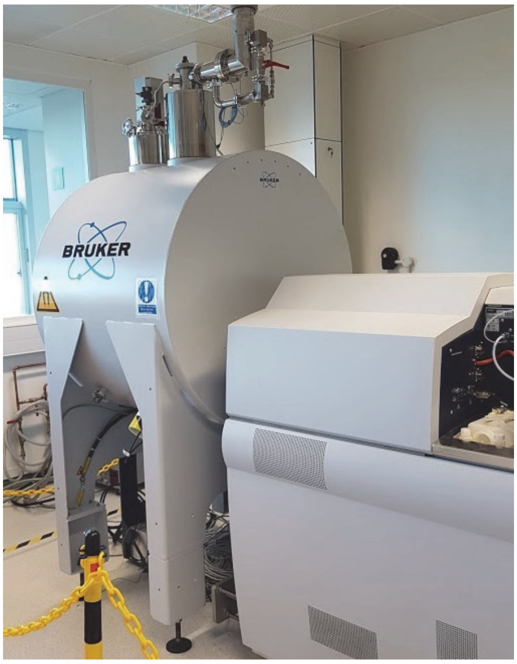

Figure 7: A Bruker Solarix MSI device at GSK.

2.1.3 MSI instruments at NPL

The NiCE-MSI group at NPL has two MSI laboratories with both vacuum (MALDI, SIMS) and ambient (LESA, ESI) ionisation techniques being employed. The terminology of a vacuum or ambient technique here primarily refers to the sample analysis chamber (ion source) being either under vacuum or ambient pressure. MALDI and DESI ionisation sources are in use coupled to Waters and Thermo instruments. The three Synapt instruments have a standard MALDI source, a prototype MALDI source and a DESI sources. The two Xevo instruments are coupled to a REIMS source and DESI source. The Synapt instruments have ion mobility separation capability which is used to further characterise ions by their collision cross section. The Synapt and Xevo instruments are otherwise comparable in terms of mass resolution and configuration. The Orbitrap instrument is coupled to a developmental atmospheric pressure MALDI MSI ion source, a Prosolia DESI ion source and an ESI source. There is also a liquid extraction surface analysis LESA ion source which can be fitted to either the Orbitrap, Xevo or Synapt instruments.

2.2 MSI DATA AND METADATA MANAGEMENT

Across all vendors in MS, data is modified prior to being displayed and made available. The exact processing implemented within a given instrument's hardware and software is typically not made explicit by the manufacturer and so is often unclear to the operator. Commonly a noise reduction or removal step is carried which both reduces the data size and, in theory, removes unwanted signal prior to the data being made available to the operator.

2.2.1 MSI data management and post-processing at AZ

Fully integrated MSI solutions such as Bruker and Waters export data in imzML, while ThermoFisher devices export data in .raw format, and converters into mzML and imzML formats are available for the user to access and download themselves. The imzML format aims to harmonise data handling in mass spectrometry imaging. Unfortunately, not all vendors are exporting MSI data in this format, and the imzML files produced by the proprietary converters lack even the most essential device settings. There is also a lack of standardisation on the measurement units: for example, some devices export millimetres squared, while others use micrometres.

The DESI devices at AZ send the experimental settings in the image metadata for each spectrum. They generate comparatively small imzML datasets of the order of several gigabytes. The reason for the smaller size is the common practice of 'peak-picking' data prior to export.

To store the experiment description, AZ is running the PerkinElmer Electronic Lab Notebook BioELN software. Experimental metadata is manually entered in unstructured form as Excel and Word documents, which are stored within BioELN. Scientists also use paper-based instrument logbooks to record their experiments, logging AZ staff name and some device settings for each day. The experimental metadata collection is hindered by the fact that the DESI devices do not keep track of some parameters that influence the resulting data such as nebulising gas pressure. By contrast, the Bruker MALDI devices capture many instrument settings in the image header. AZ scientists noted that MALDI device geometries are patented, and this might present a challenge for metadata export.

The MSI raw data are currently stored locally on the instrument computers. The processed MSI data is uploaded to network project folders. The project data does not contain instrument or raw data, but only the summary of the data analysis in Word, Excel or even PowerPoint formats. The analysis data lacks adequate experiment description, so that it is difficult or impossible for another scientist to interpret the experiment findings. For post-processing of MSI data, AZ uses the ProteoWizard toolbox as well as SCiLS software by Bruker. The latter can import and export both imzML and SCiLS data formats.

AZ believes that the most important metadata entries for image interpretation are mass range (m/z range), polarity, mass resolution (these can change between experiments) and spatial resolution.

2.2.2 MSI data management and post-processing at GSK

At the date of this survey, there are no company-, site- or laboratory-wide conventions on how the MSI data are organised and what metadata are captured. A considerable number of metadata entries such as image acquisition date, animal number, organ or tissue type, section number, section thickness and Electronic Lab Notebook number (ELNB) are recorded in the folder name. Furthermore, some metadata are captured in file names. These may include

- mass resolution
- spatial resolution
- target
- experiment date
- study ID
- slide number.

Sometimes file names also feature ion mode, mass range and non-standard settings used for DESI devices. An example of the experiment metadata encoded in the file name is

“date—ELNB—animal—organ—GSK compound number—resolution in um”.

The ELNB number can be used to find and update the image data reference in the corresponding ELNB. GSK has been using the PerkinElmer BioELN electronic laboratory notebook software for over 10 years. They are planning to replace this software in the future, although the exact date is not yet unknown. Alongside BioELN, some GSK sites use an in-house developed Excel spreadsheet to capture imaging device and sample metadata, whereby the device settings such as image resolution, laser settings, polarity etc. are extracted automatically from the *.method file exported by Solarix.

An ELNB record can be defined using customisable templates, and can contain external documents such as Word, Excel, text or PDF. The ELNB templates do not use consistent terminology, controlled vocabularies or ontologies. Each ELNB must contain the sample location and storage conditions, e.g. freezer number, temperature etc., as well as the location of the imaging data. The latter is not automatically updated if the data are moved.

At GSK MALDI-MS imaging data are stored on a 70 TB network share organised in experiment folders. These contain the optical image guidance scan, H&E stained tissue and the MALDI scan from the same section. The Bruker MSI devices at GSK save experimental metadata per run only, although settings may differ between the individual spectra. In 2019 GSK initiated a project on cloud processing of MSI data, whereby the Ion-TOF vendor software SurfaceLab will be made available via an Azure cloud container to unify the data processing across GSK sites and potentially to simplify the GSK-NPL MSI data transfer and processing workflow.

2.2.3 MSI data management and post-processing at NPL

The instruments in the NiCE-MSI laboratory output data in the vendor's proprietary formats such as the ".raw" and ".pat" files. An in-house developed publicly available software (Race, Styles, and Bunch 2012) on a laboratory server automatically converts the proprietary data to .imzML format, whereby the metadata are stored in ".imzml" header files and the spectral data in ".ibd" binary files (Schramm, 2012). The server is linked to instrument PCs and monitors these for new incoming data. Although imzML format provides a useful step to MSI data harmonisation, not all equipment metadata is captured in imzML files, for example, MALDI laser energy and the laser repetition rate are not measured or recorded for any commercial instrument. Recent efforts have been made to remove errors and validate imzML stored data (Race and Römpp 2018).

In the MSI laboratories at NPL, no commercial Electronic Lab Notebook software is used. However, the group has developed its own metadata capture system based on shared Microsoft Excel forms. This is intended to capture all metadata for an experiment that is not already output by the instrument. Metadata is entered either as free form text or selected from drop down menus. The metadata captured includes information on the sample, instrument configuration and data post-processing.

Specific examples include:

- sample type
- origin of sample
- instrument and ionisation method
- mass range
- polarity
- any non-standard settings not reported by the instrument
- number of peaks retained during post-processing
- whether PCA is performed or peaks matched against databases

Presently, NPL does not hold a license to handle human tissues, but a separate sample handling system for human samples is being installed to satisfy the additional regulatory requirements. The data collection in MSI team is largely driven by CRUK Grand Challenge requirements to preserve data for five years after the project end date in 2022. Well curated and documented datasets will be kept indefinitely and made public. The raw MSI data is transferred from the lab server to the NPL's medium-term central storage facility called 'Data Zero'. The system already holds hundreds of terabytes worth of MSI data. No data is currently deleted. 90% of the post-processing done within NiCE-MSI is done using MATLAB using both in-built functionality and the 'SpectralAnalysis' software package (Race et al. 2016). The rest consists of custom-built R or Python scripts. There is no commercial post-processing software in use, although the lab does have the Waters HDI software suite available.

Starting from 2019, NPL NiCE-MSI, Data Science and IT teams have been developing an in-house software system for standardised semi-automated MSI data annotation via a web browser-based user interface. The annotated data will be validated and uploaded to NPL Object Store for fast and sustainable data storage, retrieval and sharing.

3 LIGHT SHEET MICROSCOPY

Light sheet microscopy (LSM), or selective plane illumination microscopy (SPIM), techniques seek to overcome some of the limitations of conventional fluorescence microscopy by decoupling the illumination and detection systems in an optical microscope (Girkin & Carvalho, 2018). Fluorescence excitation is spatially confined to a thin sheet (or plane) within the sample, which is imaged onto the microscope camera. This is typically achieved using a pair of microscope objectives arranged at 90° to each other (Figure 8), although single objective variants also exist (Placeholder1). To build up a 3D image the sample is scanned through the sheet (or the sheet through the sample). The principal advantages of LSM when compared with other fluorescence microscopy methods include: a relatively low light dose, reducing photobleaching and adverse phototoxic effects; the absence of out of focus light when imaging 3D samples; a high image acquisition rate (typically 10's of frames per second). LSM techniques were originally applied in developmental biology, where their capacity for fast, minimally invasive imaging made them well suited for the study of dynamic processes such as embryogenesis in model biological systems such as *Drosophila* (Tomer, Khairy, Amat, & Keller, 2012). However, LSM methods are now used in a range of volumetric bioimaging applications, in particular for visualising tissues, 3D cell cultures and organoids and model organisms. Technical innovations continue to broaden the applicability of LSM and push the boundaries of spatio-temporal resolution (Chen, 2014) (Vettenburg, 2014).

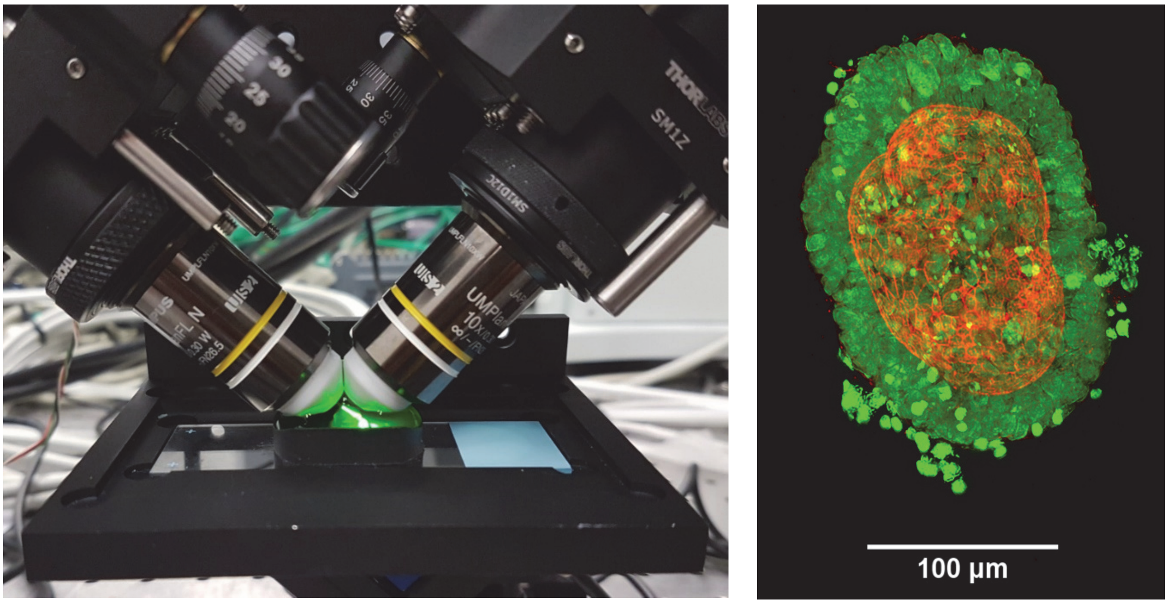

Figure 8. (left panel) Photograph of the light sheet microscope system at the National Physical Laboratory. The illumination objective lens (left) projects a thin sheet of light into the sample which is imaging onto the camera by the detection objective lens (right). The sample sits in the 'imaging pocket' between the white nose cones of the two objective lenses. (right panel) Tumour organoid stained for F-actin (red) and DNA (green) imaged at the National Physical Laboratory using light sheet microscopy.

The geometry of a light sheet microscope means that attention must be paid to sample preparation and mounting and most LSM systems are not designed for imaging samples in conventional cell culture chambers (e.g. multi-well plates) or mounted on coverslips. Instead samples are typically imaged within hydrogels such as agarose or Matrigel using water dipping objective lenses. A known challenge in LSM domain is the curation and processing of the resulting large volumes of image data. For example, a 3D single colour image from an LSM system is typically ~ 1-10 GB. A routine multicolour, time-lapse imaging experiment is likely to generate from hundreds of gigabytes to several terabytes of data. Standard processing and analysis operations such as deconvolution, spatial registration, multi-view fusion, segmentation, object detection and tracking and even image visualisation require significant computational resources. The development of new software tools for handling and analysing LSM data sets is an active area of research.

Presently, AZ are exploring whether plate-based LSM (Maioli & Chennell, 2016) provides an advantage over established techniques such as wide-field or confocal microscopy. To our knowledge, GSK does not conduct any LSM imaging. NPL bio-metrology team has their own light-sheet and confocal microscopy facilities. As a part of the FIDL collaborative project, this survey focusses on light-sheet microscopy devices. The confocal microscope devices are listed for completeness purposes only and are not discussed further.

3.1 LIGHT-SHEET MICROSCOPY INSTRUMENTS

3.1.1 LSM instruments at AZ

The team at AZ uses the same devices as HCS, including CellInsight by Thermo Fisher Scientific and Yokogawa CV 7000 and CV 8000 devices. In addition to these devices, a Zeiss Z.1 device is used that can rotate samples through 90° angles. To provide the directionality, the Z.1 device requires special sample presentation, whereby a sample is embedded in agarose tubes. This precludes the use of the Zeiss Z.1 system in plate-based screens. To that end, AZ and Chris Dunsby's Group from Imperial College London are exploring the use of Oblique plane microscopy (OPM) as an alternative approach that uses a single high numerical aperture microscope objective to provide both fluorescence excitation and detection, whilst maintaining the advantages of LSM (Dunsby C., 2008). The light sheet system in Imperial is designed for multi-well plates and can image 240 wells of spheroids in less than an hour. As CellInsight and Yokogawa devices are covered in section 1.1.1, this section will focus on the Z.1 device.

3.1.2 LSM instruments at NPL

The bio-metrology group at NPL use a variety of commercial instruments for both confocal and light-sheet microscopy, supplied by Leica, Olympus and M Squared Lasers. There are also two bespoke instruments built at NPL for structured illumination (SIM) super-resolution microscopy. The characteristics and usage of these instruments are detailed in **Table 3**.

Table 3: Light-sheet and confocal microscopy devices at NPL.

Microscope type	Manufacturer & model	Usage scenarios
Laser scanning confocal	Olympus FV1000	routine high-resolution imaging of live and fixed samples – generally adherent mammalian cells and bacteria
Light sheet (SPIM)	M Squared Lasers Aurora Alpha	Fast volumetric imaging of large samples (organoids, 3D cell cultures and model organisms)
Structured illumination (SIM) #1	Bespoke (NPL built)	Fast super-resolution imaging of cells (mammalian, bacterial) and proteins. Visualisation of subcellular morphologies, intracellular uptake, therapeutic response
Structured illumination (SIM) #2	Bespoke (NPL built)	High throughput screening of bacterial cells in a microfluidic chip

An additional Leica phase contrast microscope is used for routine examination of cultured cells to assess viability but is not otherwise used for quantitative imaging. The confocal instrument has the highest utilisation by the group: 75%, while the light-sheet and first SIM instrument has 50% utilisation. The second SIM instrument and phase contrast microscope have approximately a 25% utilisation.

Key instrument settings common to all instruments include the numerical aperture and magnification of the objective lens, excitation wavelength and power, emission filter / detection bandpass, axial sampling (z step size), exposure time, pixel count and the pixel size of the camera. Sample preparation, which includes culture, fixation, mounting and staining, is a common source of experimental variability across all imaging modalities. Noise and imaging artefacts can give rise to computational errors in image correction or deconvolution. The bespoke structured illumination instruments also require additional calibration to determine the spatial frequency and orientation of the illumination pattern.

3.2 LSM DATA AND METADATA MANAGEMENT

An example LSM experiment generates tens to hundreds of gigabytes of imaging data, whereas a single 384 well plate image is approximately 7GB. At AZ, LSM images are frequently downsampled or “binned” to reduce the data volume, and 4x4 binning is frequently used on large samples. Binning results in a reduction of the image quality and should be used with care. Another technique to reduce the image size is cropping of the image to contain only the region of interest. At the moment, Zeiss systems do not provide automatic cropping. Zeiss devices export the imaging data in the CZI format that includes hundreds of metadata entries and is well supported by OMERO.

3.2.1 LSM data management and post-processing at AZ

The contextual metadata required to interpret the experimental results are comparable to confocal microscopy and include sample position, pixel size, binning as well as the microscope settings. The enterprise and some context metadata are recorded in BioELN software by PerkinElmer. The ELN record contains a Standard Operation Procedure (SOP) number that describes the assay validation and the experiment protocol. Some SOPs have been published. The SOPs are not machine-readable documents and contain some standard headings alongside unstructured text.

The experiment metadata recorded in the central database includes

- cell type or cell line
- species
- dyes or staining
- compounds and treatment, including concentrations used
- well type e.g. control or treatment
- channel type such as fluorescent or bright-field.

The folder and file names also contain metadata including well ID or row and column, experiment ID, plate ID or barcode.

The recorded analysis data includes plate ID and analysis results. These data are insufficient to reproduce the experiment. LM experiments are sometimes re-run as the data may not match expectations. A metadata quality test could be used, where the metadata captured during the experiment is used for another purpose, for example, a meta-analysis.

AZ use MATLAB scripts as well as the Fiji software for LM data post-processing. Due to very large file sizes, LM data cannot be imported into Columbus. They would be interested in using OMERO for LM data management. The raw data are usually deleted within three to six months, and no images are preserved at present. There is a need for improvement and further refinement of the data retention policy in LM at AZ. Similar to HCS, LSM data are post-processed and analysed with GeneData software. The use of this software is well regulated across AZ. Extensive compatibility testing is performed during the commissioning of new GeneData software versions.

3.2.2 LSM data management and post-processing at NPL

Raw image data from the light sheet microscope is in OME-TIFF format with associated equipment metadata in the header of these files. Some of the metadata which is not part of the OME data model, such as environmental conditions and calibration data, is recorded manually in a text document stored with the image files. Contextual metadata is recorded in unstructured form in a Microsoft Word document. Image post-processing is performed by a C++ application provided by the manufacturer and ImageJ or MATLAB. Data is managed in a similar fashion to the confocal microscope. There is no formal data retention policy. As raw light-sheet data is particularly large, only “valuable” data is typically retained.

Equipment metadata for the bespoke structured illumination microscopes is recorded in plain text “.txt” files by the LabVIEW software written in-house to control these instruments. Calibration metadata is recorded in MATLAB “*.mat” files. Again, contextual metadata is recorded in Microsoft Word documents. The raw images are in “.tiff” format and organised in folders on networked drives by YYYY/MM/DD date format. A combination of MATLAB, IMARIS and ImageJ is used for post-processing. There is no formal data retention policy and data from these instruments is generally kept indefinitely.

4 SUMMARY

The bio-imaging landscape in High Content Screening, Mass Spectrometry Imaging and Light-Sheet Microscopy is defined by large volumes of high-dimensional data that contains a wealth of information about tissue morphology, biological processes and drug-tissue interactions. Presently, this information has limited availability due to disseminated storage, lack of structured and standardised annotations, poorly designed interfaces between software packages used to store, access and analyse the data as well as lack of consensus on data formats and required annotations between equipment vendors, users and regulators. In the domain of clinical medicine, the latter issue has been addressed with an introduction of the DICOM standard and the associated DICOM file formats for various clinical imaging domains including computed tomography, ultrasound imaging and magnetic resonance tomography. In the domain of microscopy imaging, harmonisation attempts have been undertaken by the OME consortium by the introduction of Bio-Formats.

AZ, GSK and NPL have different data management strategies. Both AZ and GSK use laboratory notebooks that are linked to company-wide databases with compound names, gene descriptors and administrative data. The raw experiment data are stored in disseminated locations such as network folders and instrument computers. The imaging data are stored in vendor-proprietary formats that contain varying amounts of information about the device settings. The experiment descriptions are stored in electronic laboratory notebooks, non-standardised electronic documents, in the names of image files and folders as well as in paper documents. The scope and format of experimental data varies between laboratories and individuals. These storage structures do not permit the capture of the experiment description in machine-readable form. AZ and GSK are increasingly realising the need to unlock the value in the experimental data. Pilot projects to ingest annotated imaging data into centralised electronic storage are running at AZ’s Discovery Sciences and Next Generation Sequencing groups and at GSK as a part of Scientific Data Management System implementation.

At NPL, the MSI and LMS imaging data are harmonised by in-house developed software tools that convert vendor-proprietary formats to OME-TIFF for LMS images and to imzML for MSI. Both laboratories have pioneered the upload of annotated imaging data into the NPL Object Store introduced in 2018. At present, NPL does not deploy laboratory notebook software, and scientists keep the experiment notes as MS Office documents, and the structure and the content of the documents vary from project to project. To harmonise the annotation of the imaging data with the information about the sample handling, administrative project data and other relevant information, NPL is developing a web-browser based software for image search, and semi-automated annotation. The annotation software will interface with the NPL Object Store and image processing tools including OMERO and Fiji. The scope and content of bio-imaging annotations will be based on the findings of the FIDL project and this survey.

5 REFERENCES

A. Buck, B. H. (2018). Round robin study of formalin-fixed paraffin-embedded tissues in mass spectrometry imaging. *Analytical and Bioanalytical Chemistry*, 410:5969–5980.

Barré, F. &. (2017). Mass Spectrometry Imaging in Nanomedicine: Unraveling the Potential of MSI for the Detection of Nanoparticles in Neuroscience. *Current pharmaceutical design*(23).

Buchberger, A. e. (2018). Mass Spectrometry Imaging: A review of emerging advancements and future insights. *Anal. Chem., 90*(1).

Buchser, W. (2014). Assay Development Guidelines for Image-Based High Content Screening. *High Content Analysis and High Content Imaging*.

Chen, B.-C. e. (2014). Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *346*(1257998).

Dunsby C. (2008). Optically sectioned imaging by oblique plane microscopy. *Optics Express*, 20306-20316.

Girkin, J. M., & Carvalho, M. T. (2018). The light-sheet microscopy revolution. *20*(053002).

Maioli, V., & Chennell, G. e. (2016). Time-lapse 3-D measurements of a glucose biosensor in multicellular spheroids by light sheet fluorescence microscopy in commercial 96-well plates. *Nature Scientific Reports*.

Schramm, T. e. (2012). imzML — A common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of Proteomics, 75*(16).

Tomer, R., Khairy, K., Amat, F., & Keller, P. J. (2012). Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *9*(755).

Vettenburg, T. e. (2014). Light-sheet microscopy using an Airy beam. *11*(541).

Zock, J. (2009). Applications of high content screening in life science research. *Comb Chem High Throughput Screen., 12*(9).