National Physical Laboratory

NPL REPORT MS 24

FEDERATION OF IMAGING DATA FOR LIFE SCIENCES; CURRENT STATUS OF METADATA COLLECTION FOR HIGH CONTENT SCREENING, MASS SPECTROMETRY IMAGING AND LIGHT SHEET MICROSCOPY AT ASTRAZENECA, GLAXOSMITHKLINE AND NPL

BROCHU, F
BUNCH, J
COOKE, E
DEXTER, A
ROMANCHIKOVA, M
SHAW, M
STEVEN, R T
THOMAS, S A

MAY 2020

Department for
Business, Energy
& Industrial Strategy

FUNDED BY BEIS

Federation of imaging data for life sciences: current status of metadata collection for high content screening, mass spectrometry imaging and light sheet microscopy at AstraZeneca, GlaxoSmithKline and NPL

NPL authors:
Frederic Brochu, Elizabeth Cooke, Marina Romanchikova
Data Science

Josephine Bunch, 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, Yinhai Wang, Alan Race

GlaxoSmithKline contributors:
Donna Fraser, Jo Francis, Carla Newman, Vijaykumar Chavda, Colin Wood, Joseph Lavelle

ISSN 1754-2960

https://doi.org/10.47120/npl.MS24

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.

GLOSSARY/ABBREVIATIONS

AZ	AstraZeneca
BIOMET	NPL’s Biological Metrology group
DCMI	Dublin Core Metadata Initiative
DESI	Desorption Electro-Spray Ionization
EMBL-EBI	European Molecular Biology Laboratory’s European Bioinformatics Institute
FAIR	Findable, Accessible, Interoperable, Reproducible: data management principles
GSK	GlaxoSmithKline
HCS	High Content Screening
LSM	Light Sheet Microscopy
MALDI	Matrix-Assisted Laser Desorption/Ionisation
MSI	Mass Spectrometry Imaging
NiCE-MSI	The National Centre of Excellence in Mass Spectrometry Imaging
NPL	National Physical Laboratory
SIMS	Secondary Ion Mass Spectrometry
URI	Uniform Resource Identifier: a sequence of characters that uniquely identifies an information resource

EXECUTIVE SUMMARY

The rapid expansion of technology and computing processes is facilitating research on unprecedented scales. The multitude of proprietary file formats and data storage systems makes data location and sharing difficult. Many bioimaging modalities generate terabytes of imaging data per day and pose new challenges for data analysis and management. These challenges can be partially addressed by using standardised data descriptors (metadata) to capture scientific, regulatory and business-related features of a dataset. This report aims to highlight the importance of metadata, to provide an insight into different metadata types and to compare the current bioimaging annotation practices at three high profile partner sites: AstraZeneca, GlaxoSmithKline and the National Physical Laboratory. The report is focussed on three life science imaging techniques: high-content screening, mass spectroscopy imaging, and light-sheet microscopy.

1 INTRODUCTION

The Federation of Imaging Data for Life Sciences (FIDL) project is a collaboration between the National Physical Laboratory (NPL), AstraZeneca (AZ) and GlaxoSmithKline (GSK). The principal aims of the project are

1. To highlight the importance of metadata for increasing the long-term value of data, improving research reproducibility and enabling regulatory compliance.
2. To define the relevant metadata types and their roles in the bioimaging data management.
3. To summarise the current practices in the annotation of bioimaging data across the three high-profile partner sites.

The data management methods and the metadata annotations used were collected for three use cases:

1. Mass spectrometry imaging (MSI),
2. High content screening (HCS),
3. Light sheet microscopy (LSM).

These three cases all have high volumes of complex imaging data. These data are often generated in proprietary formats with little regard to the potential re-use of the data. While some initiatives have been undertaken to harmonise image data management for LSM and MSI domains (Goldberg et al. 2005; Linkert et al. 2010; Ellenberg et al. 2018; Huisman et al. 2020), they predominantly focussed on capturing the imaging device parameters rather than the experiment context, the business value of data or the research workflow. Across the imaging data landscape there is little consistency in the generation and format of metadata produced. Whilst details of the configuration of equipment may be automatically generated alongside raw data in human-readable form, other metadata relating to the sample analysed or experimental conditions is often hand-coded in filenames or directory structures, saved in unstructured text files & spreadsheets, entered free-form in an electronic lab notebook or stored in other databases.

The available metadata are often inconsistent between laboratories and scientists and do not follow a controlled vocabulary. As a result, the data annotated in this inconsistent manner are not easily searchable and are therefore limited in their utility. This practice inhibits the re-use of the data and slows down the scientific process. The aim of this work is to address the data management issues for life science imaging by offering recommendations for minimum metadata annotation of the imaging data to improve reproducibility and to encourage re-use and data sharing. For example, *experimental* or *contextual* metadata describing the configuration of an instrument, or the conditions and protocols of an experiment, is often stored in an XML or a text file in the same directory as the data itself, on paper, in electronic laboratory notebooks or other electronic documents. *Enterprise* metadata identifying the owner of the data and its provenance may only be represented within filenames or directory structures and is therefore vulnerable to unintentional change or misinterpretation.
A *data container* is intended to make the connection between a dataset and its metadata explicit by bundling them within one file. The metadata should be defined by a controlled vocabulary (ontology) and should be searchable. A data container should enable portability, be self-describing and therefore enhance the integrity of the dataset and the ability to share it effectively.

1.1. SCOPE

This document describes the metadata currently captured by three high-profile organisations in bioimaging. It concludes the three FIDL reports produced by NPL Data Science team:

1. “A review of file formats with data annotation capability for bioimaging”, 2018
2. “Acquisition and management of high content screening, light-sheet microscopy and mass spectrometry imaging data at AstraZeneca, GlaxoSmithKline and NPL”, 2019
3. This report “Federation of imaging data for life sciences: current status of metadata collection for high content screening, mass spectrometry imaging and light sheet microscopy at AstraZeneca, GlaxoSmithKline and NPL”, 2020.

The annotations presented in this document summarise the current metadata captured in the experimental setting, the sample handling and the project details. With respect to the FAIR data management principles, the initial focus of the metadata capture presented in this report is on “Findable”, i.e., being able to locate and categorise the data for re-use.
While this document is focused on three imaging modalities specified above, some of the metadata captured may be applicable in other bioimaging areas for example, digital histopathology.

1.2. OUT OF SCOPE

This report does not provide recommendations as to what metadata should be captured and what software tools should be used for image annotation. It is aimed to inform its reader on the current state of affairs at AZ, GSK and NPL. Future work will look at recommendations based upon this current state and the outlook for the bioimaging domain.
This report does not discuss in-depth the imaging equipment or the implementation of the metadata annotations. This information is contained within the other FIDL project reports.

1.3. DATA SHARING AND RESEARCH REPRODUCIBILITY

Collection and storage in digital formats allow high volumes of data to be used in advancing research, as well as sharing of data across research groups and organisations. Digital data are increasingly being used beyond the scope of the original project for which they were collected, allowing advances in new areas of study and the creation of larger datasets encompassing wider ranges of data across different scales and modalities. An added benefit of data sharing is more accountability; results are easier to reproduce, and research practices may be tested and improved.
Traditionally, scientific data was kept in-house, often within a single research group. This had competitive advantages for researchers and organisations. However, as data sharing was not a priority, there are now myriad proprietary file formats and systems for storing metadata, which does not facilitate cross-sharing of data.
In addition, estimates suggest that more than 50% of scientific findings are never published and remain in personal storage systems (Chan, et al., 2015). This has the effect of not only wasting the time and money required to produce the results, but also potential for duplication of research efforts as the results are not in the public domain.
The sharing of data in life sciences is crucial, not only to prevent this wasted effort by duplicating results, but also to enable new discoveries that would not be possible with any individual dataset. Increasingly, funding agencies and scientific journals are also now requiring data supporting an article to be made publicly available before it can be published.
In order to effectively share data across multiple institutions, there needs to be common file formats, or the ability to convert between file formats, centralised accessible data storage platforms, and access to metadata to allow interpretation of the data.

1.3.1 Implications for pharmaceutical industry

Drug discovery is currently performed on an ad-hoc basic, due to the lack of accessibility of data from previous experiments. In 2014, it was estimated that a typical Fortune 1000 company increasing their data accessibility by just 10% would generate more than $65 million in additional net income (McCafferty, 2014). Although various laboratory information systems (LIS) and scientific data management systems (SDMS) exist, they are typically designed by one vendor and have insufficient integration with non-vendor data sources. In the academic community, integration attempts have been undertaken (see Section 1.5), but these lack the enterprise-related metadata structures required to connect datasets between projects, laboratories and modalities.

1.4. FAIR PRINCIPLES AND THE NEED FOR METADATA CAPTURE

The guiding principles of scientific data are known as the FAIR principles: data must be findable, accessible, interoperable and reusable (Wilkinson et al. 2016). The metadata attached to images are crucial in interpreting the data as well as in managing, sharing and finding the information. The metadata should provide sufficient information to understand the data and reproduce the experiment's results. In the context of clinical trials, the metadata are key to help comply with regulatory requirements and manage the data lifecycle. **Figure 1** illustrates the variety of users wanting data access and the metadata types required to locate bioimaging data.

Currently, the amount of metadata associated with bio-images varies greatly. The exporting and converting of file formats provide an additional layer of complexity as they may cause the partial loss of metadata. There are also image processing steps which may be taken, and thus need to be accurately recorded.

It is therefore important to have a consistent list of metadata elements across institutions to allow compatibility, interoperability and improve the quality of the dataset. Ideally, there would be a standard that may be applied to all life sciences imaging metadata. Following such standards also makes it possible to develop public repositories, centralised databases and common data analysis tools for use across the industry.

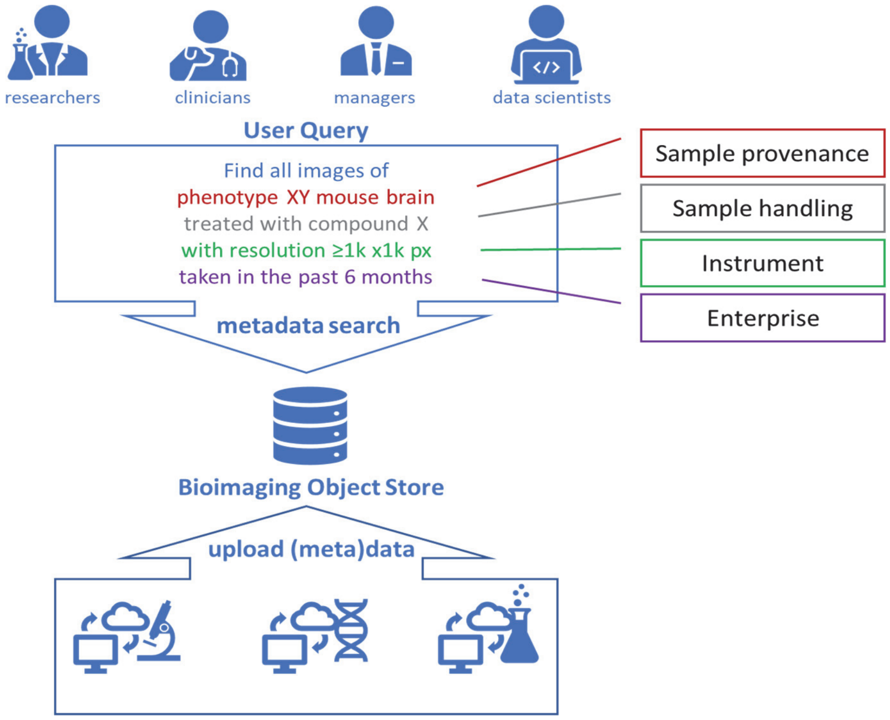


Figure 1: Metadata elements required for bioimaging data retrieval for a typical query scenario.

1.5. STATE OF THE ART IN BIOIMAGING DATA MANAGEMENT

Integration of data enables new trends and links between studies to be identified. In genetics and genomics, combining studies of different populations has rapidly increased the knowledge of genetic determinants in health and disease (McCarthy et al. 2008).
The image data resource (IDR) links data from several imaging modalities, including high content screening, with public genetic and chemical databases. Combination of several different studies can lead to new discoveries such as novel representations of gene networks (E. Williams et al. 2017). This shows the power of data integration and searchability.
There are several tools to facilitate data harmonisation expanded upon in the following section.

1.5.1 Tools for information management

Interpretability of scientific data relies upon the use of consistent domain terminology for data annotation. There is a number of semantic tools that can be used to organise domain terms, including controlled dictionaries, taxonomies and ontologies. A (controlled) dictionary is a collection of agreed terms and their definitions. A taxonomy allows one to present a set of terms and their hierarchical relationships: For example, mouse is a subclass of mammal. An ontology is a method of representation that links terms and their definitions within a field. It describes both terms and the relationships between those terms.
Table 1 shows an example of why common terminologies and ontologies are crucial for data sharing: the same concept may be expressed using different terms at different institutions. Even within a single institution, terms and acronyms can vary between laboratories, individuals and over time.

Currently, several ontologies for imaging exist within the biomedical sector. Most of these describe a specific domain in biomedicine, rather than spanning a broad range of domains or methodologies. Most of the ontologies in healthcare may be accessed through the web service BioPortal. The most common formats in biomedical ontologies are the Web Ontology Language (OWL) and the Unified Medical Language System (UMLS) (Bodenreider 2004).

Table 1: Variation in terminology across pharmaceutical industry. Source: C. Wood, RSC 2014

Domain Entity	**AstraZeneca**	**GSK**	**Novartis**	**Pfizer**	**PubChem**	**FDA**
Compound	Compound	Parent	Substance (Parent)	Parent	Compound	Molecular Entity
Substance (Compound Substance)	Substance	Version	Salt	Salt or Compound	Substance	Substance or Active Ingredient
Batch	Sample	Preparation or Lot	Batch	Batch	N/A	N/A
Lot	Sample	Sample	Sample		N/A	N/A

In biomedicine, there are a few projects and data sources that allow for browsing and searching through existing data. However, the interfaces may be complicated, unintuitive, and provide limited functionality. Most of the time, some knowledge of a query language, such as SPARQL, is required. More functional interfaces include the Semantic Web Portal (Ding et al. 2010) and Linking Open Data (Tilahun et al. 2014).

1.5.2 Data standards and formats

In order to share data across platforms and industries, there is a need for file formats to be readable outside of the parent company. This is addressed using data standards and open file formats. These can be domain-specific formats listed in the examples below or domain-agnostic such as HDF5.

Within clinical medicine, the issue of data formatting 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 in data formats have been undertaken by the OME consortium through the introduction of Bio-Formats and OME-TIFF.

In diagnostic radiology, the radiology lexicon (RadLex) provides a vocabulary of terms used in clinical practice, research and education. RadLex is not yet an ontological framework as further steps are required for it to become an ontology. For example, issues still exist with relationships between terms and with gaps in definitions, however this may become the basis for an ontology in the future.

1.5.3 Bioimaging repositories

There is an increasing realisation that sharing of well-annotated bioimaging data aids research reproducibility, accelerates discovery and enables effective collaboration. Several efforts have been undertaken to date to provide resource for upload, management and sharing of bioimages.

The Open Microscopy Environment (OME) is a community of academics and commercial partners which produces tools for data management in microscopy. Relevant to this report are their OMERO and Bio-Formats tools. OMERO is an image-sharing platform that allows uploading, processing, analysis and sharing of microscopy image data (Allan et al. 2012). OMERO supports over 140 image file formats, including metadata for images. Bio-Formats is a software plug-in that is used to convert file formats into OME-compatible formats (Linkert et al. 2010). It is able to read and write both proprietary images and their metadata and convert these to the OME data model, i.e., into the OME-TIFF format (Goldberg et al. 2005).
The *Journal of Cell Biology* (JCB) *DataViewer* (E. H. Williams, Carpentier, and Misteli 2012) is an online repository for original image data in the life sciences. The JCB is based on OMERO and Bio-Formats and stores both original binary data and metadata. However, the available metadata varies significantly between images due to the lack of a minimum guideline on required metadata.
The European Bioinformatics Institute (EBML-EBI), part of the European Molecular Biology Laboratory (EMBL), is developing databases and software tools designed to store, search and visualise molecular data. In 2019, EBML-EBI announced an expansion of its remit to include bioimaging data by founding a new dedicated resource called BioImage archive in collaboration with the EuroBioimaging initiative and the OME consortium (Ellenberg et al. 2018).

2 IMAGING MODALITIES

In this report we aim to describe the metadata captured and used at three institutions within three life science imaging modalities:

- Mass spectrometry imaging,
- High content screening,
- Light sheet microscopy.

These three case studies are all characterised by high volumes of complex data, which may be generated and stored in several different proprietary formats. In some cases, there is little consistency in the generation and format of metadata produced. Whilst details of the configuration of equipment may be automatically generated alongside raw data in XML format, other metadata relating to the sample analysed or experimental conditions is often hand-coded in filenames or directory structures or entered free form in an electronic lab notebook. Metadata is often not drawn from a well-defined and unique vocabulary or ontology. Indeed, the wide variety and complexity of the samples and imaging experiments undertaken make it difficult to define controlled ontologies which cover all conditions which may be used. As a result, these metadata are not easily searchable and therefore limited in their utility. Critically, the practice of poor data engineering inhibits re-use of the data and the scientific reproducibility of the experiment. In the following sections we briefly describe the imaging modalities.

2.1 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 quantitative information about metabolites, lipids, peptides, proteins and glycans in a single experiment on a label-free sample with minimal preparation. An optical image of the sample is often taken of a sample to provide image guidance and to identify high level structures. An example of the workflow for MSI data is shown in **Figure 2**.

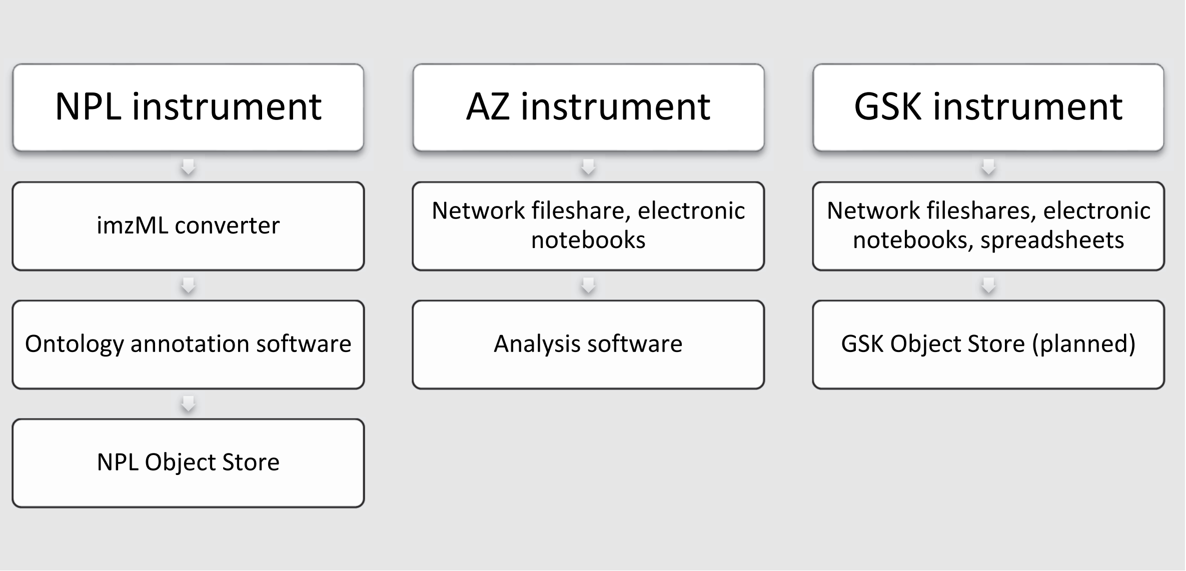

NPL instrument
imzML converter
Ontology annotation software
NPL Object Store
AZ instrument
Network fileshare, electronic notebooks
Analysis software
GSK instrument
Network fileshares, electronic notebooks, spreadsheets
GSK Object Store (planned)

Figure 2: MSI data workflow from instrument to long-term storage.

2.2 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 phenotype of a cell are observed in a controlled manner. Phenotypic changes may include increases or decreases 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 the number of objects, their shape, texture, colour, size or intensity from a cell population in a short time interval (Buchser et al. 2004).

2.3 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 and 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, although single objective variants also exist (Dunsby 2008). 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; and a high image acquisition rate (typically tens of frames per second).

The LSM image data processing workflow at NPL is shown in **Figure 3**. GSK does not run LSM facilities at the moment.

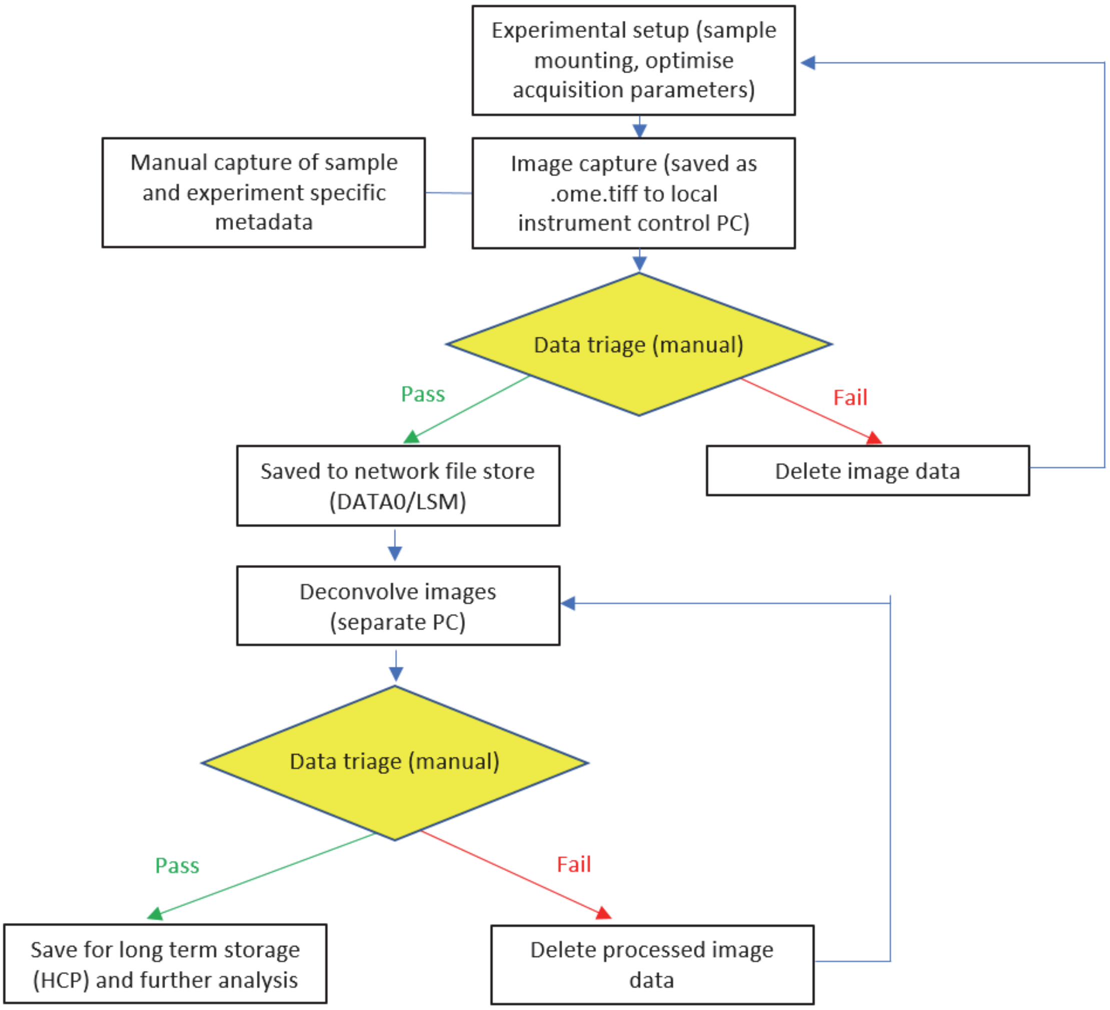

Experimental setup (sample mounting, optimise acquisition parameters)
Manual capture of sample and experiment specific metadata
Image capture (saved as .ome.tiff to local instrument control PC)
Data triage (manual)
Pass
Fail
Saved to network file store (DATA0/LSM)
Delete image data
Deconvolve images (separate PC)
Data triage (manual)
Pass
Fail
Save for long term storage (HCP) and further analysis
Delete processed image data

Figure 3: LSM data processing at NPL.

3 METADATA CATEGORIES

The metadata captured in bioimaging varies according to organisation, laboratory, and equipment operator. Consistent metadata capture enables greater interoperability between organisations and research groups, increases transparency in reporting of results, facilitates access to data, and adds value to the dataset and the group producing the data. In this section we list and compare the metadata currently captured in the three organisations. We categorise metadata into the three categories:

1. Domain-agnostic enterprise level metadata,
2. Domain-specific experiment level metadata,
3. Storage-specific storage level metadata.

The following sections describe the metadata categories and list the associated metadata entries.

3.1 ENTERPRISE METADATA

The purpose of the enterprise metadata is to provide administrative and business-level insight into the data to inform the user about the ownership, legal requirements for retention and management, origin and purpose of the dataset. A substantial part of administrative metadata for information assets has been covered in the Dublin Core Metadata Element Set developed by Dublin Core Metadata Initiative (DCMI). A summary of enterprise metadata entries is given in **Table 2**, whereby the entries already present in DCMI terminology are marked as "Dublin Core".

Table 2: Enterprise metadata summary

Term	Format	Description
Rights	Dublin Core: Rights	A statement about various property rights associated with the resource, including intellectual property rights. In NPL this field corresponds to "Heads of Terms" agreement.
Creator	alphanumeric	Person, device or IT system responsible for generation or contribution to the data.
Owner	alphanumeric	Link to master database with the data owners.
Group	URI	Link to master database with business unit names (team, department, division).
Format	URI	Link to master database with pre-defined data formats used in the organisation.
CreationDate	Dublin Core: Date	The date upon which the data was first made available for use in its original form.
Identifier	Dublin Core: Identifier	Externally generated identifier in the master storage system and uniquely identifies the content.
ConfidentialityCategory	alphanumeric	E.g., "classified", "open access", "internal use". Categories should be populated from a pre-defined list.
InformationType	Dublin Core: Type	The nature of the resource to uniquely identify the purpose of the record. For example, "Image", "Dataset" or "Text".
Language	alphanumeric	ISO 639-1 language code.
MaterialEntityID	URI	Link to master database with compound, target or material names. This record is pharma-industry specific.
OriginatingOrganization	URI	Link to master database with organisation names.
PersonalData	boolean	Personal information includes all information types that can be used to identify an individual.
LegalHoldNotice	alphanumeric	Name of the preservation notices which requires the continued preservation of the record until the preservation notice(s) is cancelled.
RetentionCategory	URI	Link to master database with the company's current retention policies.
RetentionStartDate	Dublin Core: Date	Defines the category of the information resource, and thus the amount of time it must be retained for before it can be discarded, providing it is not subject to a preservation notice.
ReviewDate	Dublin Core: Date	Date assigned for when the record is scheduled to delete
ProjectCode	URI	Link to the master database containing numbers of a project or a study for which the data was acquired. At NPL, this database contains project and task codes in the format <project code:task code>.
Title	Dublin Core: Title	A free-text description of the data aimed to help external readers.

3.2 EXPERIMENTAL METADATA

Experimental, or “context” metadata describes both the biological sample used and details of its preparation and treatment during the experiment. Experimental metadata would typically identify the following types of information: species, tissue type, cell line, temperature, reagents used and their concentration. These will be of critical importance to all potential users of the data and affect the reproducibility of the data. Experimental level metadata often has the widest variance in quality with respect to its consistency and completeness. As shown in **Figure 4**, the experimental metadata may be sub-divided into four groups: instrument settings, sample provenance, sample handling, and data processing. The domain-specific lists of experimental metadata are summarised in **Table 3**, **Table 4** and **Table 5**.

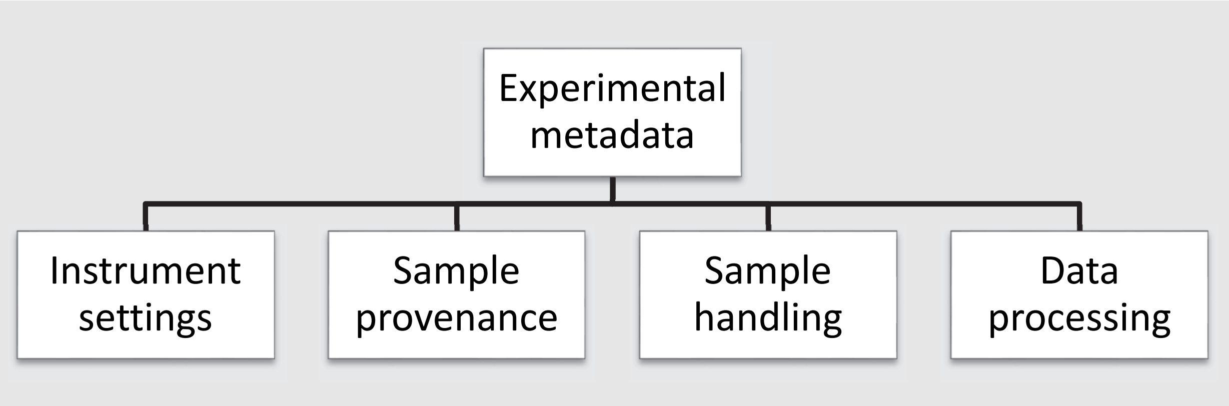


Figure 4: Metadata groups to describe the imaging pipeline.

3.2.1 Instrument settings

These metadata elements describe the equipment configuration and specification used in the experiment. The types of information captured will be very specific to the instruments and their usage within the laboratory or the project. However, we can identify certain generic fields which will be common to all microscopy or all mass spectrometry experiments. Although this metadata is often automatically generated by the instruments, it may not be recorded with the exported images or may be stored in undocumented proprietary formats, hindering its accessibility and future re-use.

While the instrument metadata is a backbone for experiment reproducibility, it has little or no function to other end-users such as project managers, database administrators or non-specialists.

3.2.2 Sample provenance

These metadata describe each step of the imaging process and the sample being imaged, for example the cell type. This metadata is usually already recorded and stored; however, it should be checked that it remains with the relevant images when they are transported or shared. The provenance of the sample and imaging data is crucial for reliable comparisons with future experiments, analysis by other scientists, and ensuring experiments are not duplicated unnecessarily.

3.2.3 Sample handling

Recording of these elements are essential for experiment reproducibility. Currently a lot of this information is stored in formats which are not tied to the data, such as paper-based or electronic laboratory notebooks. As with the instrument settings, the types of information captured will be specific to the reagents, instruments and techniques used within the laboratory, but we may still identify generic fields common to most experiments.

3.2.4 Data processing

These metadata entries are used to describe the image format, as well as the type and sequence of various operations applied to the raw data from the instrument to obtain the image. Capturing these elements enables planning for future file storage, sharing and comparison with other images. These metadata include image properties such as pixel size, image scale, file format, number of colour channels as well as compressions and transformations that have been applied to the data. The latter is important to capture, as raw instrument data are frequently post-processed to save storage space and remove unnecessary data such as images of empty wells in HCS or data with signals below a set threshold in MSI.

Beyond research, these metadata may be useful for other user groups such as information architects, database administrators and information technology support staff.

3.3 STORAGE METADATA

The purpose of these data elements is to describe the storage-related rules and requirements. These metadata may, and should, be extended and customised to provide the information about the location of the intermediate and long-term storage, encoding and compression mechanisms, chunking rules for large datasets, bucketing rules for Object Store systems etc.

Although much of the storage metadata depends on the underlying system architecture, there are some system-agnostic metadata elements that may encompass:

- Data custodian(s)
- Creation & last modified timestamps
- Version information
- Data type/format & chunking/splitting rules
- Read-only flag
- Retention policy of the storage system (that may differ from the retention policy applied to imaging data)
- Review date

4 EXPERIMENTAL METADATA FOR MASS SPECTROMETRY IMAGING

MSI file sizes vary considerably and may generate between megabytes and hundreds of gigabytes per dataset. Tens or potentially hundreds of variables are either explicitly or implicitly chosen and implemented in each experiment, some of which the experimentalist may not interact with directly. However, many of these variables have the potential to contribute significant variance to the results obtained, and so comprehensive and FAIR recording of metadata is highly desirable.

MSI begins with a sample whose spatial and chemical heterogeneity require mapping, and for molecular MSI this will typically be a biological sample. It is here that potentially relevant metadata collection (e.g. animal model, age, sex) should begin. The tissues are then processed in some way with variables related to preservation, storage and material transfer coming into play. Sample preparation for MSI is then carried out which may involve tissue embedding, sectioning, slide mounting and additional chemical treatment, all of which can significantly influence the final results. Finally, the sample, having been prepared for analysis, is transferred to the mass spectrometer, and this instrument is set up and calibrated ready for measurements to be acquired. At this point there are many voltage, pressure and analytical parameters which may be changed or monitored, again playing a big role in the nature of the data obtained. There have been some attempts to move towards minimum reporting (McDonnell et al. 2015; Gustafsson et al. 2018) and (meta)data validation (Race and Römpp 2018), but there remains significant room for development in these areas across the MSI research field.

Table 3: Metadata terms for mass spectrometry imaging

Category	Term	Description	Format & Unit	Used by
data processing	RawDataFileLocation	path or URI of raw data file	alphanumeric	AZ/GSK/NPL
data processing	CoordinateFileLocation	path or URI of the coordinate file	alphanumeric	NPL
data processing	AcquisitionStartTime	start time of image acquisition	Dublin Core: Date	GSK/NPL
data processing	AcquisitionCompletionTime	end time of image acquisition	Dublin Core: Date	GSK/NPL
data processing	ProcessingSoftware	name and version of the software used in conversion of data format	alphanumeric	AZ/GSK/NPL
data processing	ELNBNumber	electronic laboratory notebook number	alphanumeric	AZ/GSK
instrument settings	Instrument	instrument manufacturer and model number	alphanumeric	AZ/GSK/NPL
instrument settings	InstrumentLocation	physical location of the instrument on site, i.e. laboratory number or ID.	alphanumeric	AZ/GSK
instrument settings	InstrumentOperator	name or ID of the instrument operator	alphanumeric	AZ/GSK/NPL
instrument settings	NumberOfPixels	total number of pixels in the image	numeric	AZ
instrument settings	PixelSize	size of a pixel in microns	numeric, micron	AZ/GSK/NPL
instrument settings	IonBeamType	type of ion beam source	alphanumeric	AZ/GSK/NPL
instrument settings	PolarityAtSampleStart	positive/negative/alternating/dual	numeric	AZ/GSK/NPL
instrument settings	PolarityBlockSize	number of rows acquired with the same polarity (for alternating polarity)	numeric	NPL

instrument settings	MassAcquisitionMode	method used to detect and quantify ions in MSI; e.g. ToF, FT-ICR, Orbitrap, or MS or MS/MS modes	alphanumeric	AZ/NPL
instrument settings	MassRange	m/z	numeric	AZ/GSK/NPL
instrument settings	MassResolution	minimum peak separation	numeric	NPL
instrument settings	PointsPerSpectrum	number of data points per spectrum	numeric	AZ
instrument settings	BinWidth	m/z	numeric	NPL
instrument settings	NebulisingGasPressure	the pressure of the gas used to generate the matrix spray in bars or PSI	numeric, bars or PSI	AZ/GSK
instrument settings	LaserSpotSize	laser spot size in microns	numeric	AZ/NPL
instrument settings	LaserWavelenth	wavelength of laser in nm	numeric, nanometres	AZ
instrument settings	LaserFrequency	repetition rate of ablation laser	numeric	AZ/NPL
instrument settings	NumberOfLaserShots	total number of laser shots fired at tissue during run; is related to pixel number/ image size and gives an indication of potential laser aging over the course of acquisition	numeric	AZ
instrument settings	LaserEnergy	Energy of the laser delivered per pulse to the sample.	numeric, arbitrary units or Joules	NPL
instrument settings	SolventFlowRate	rate of solvent flow through the electrospray in ml/min (can impact spatial resolution and sensitivity)	numeric	AZ/GSK
sample handling	SprayTemperature	spray temperature in degrees Celcius	numeric	GSK
sample handling	SampleProcessing	any processing to sample that has occurred before acquisition e.g. frozen / PFA fixation / xylene wash / derivatisation / antibody staining: IF/IHC/IMC	alphanumeric	AZ/NPL
sample handling	CompoundID	name of the compound used in treatment. “MaterialEntityID” in GSK terminology.	alphanumeric	AZ/GSK
sample handling	CompoundDose	concentration of compound per weight per time period	numeric, mg/kg/day	AZ/GSK/NPL
sample handling	MatrixApplicationMethod	f.e., TM Sprayer, sublimation	alphanumeric	GSK/NPL
sample handling	MatrixType	f.e. DHB, CHCA, 9-AA	alphanumeric	GSK/NPL
sample handling	MatrixSolvent	f.e., Metanol:Water	alphanumeric	GSK
sample handling	MatrixSolventAdditive	additives used with DESI solvent	alphanumeric	AZ
sample handling	MatrixConcentration	milligram/millilitre	numeric	GSK/NPL

sample handling	MatrixDilutionRatio	f.e. 70:30	numeric	GSK
sample handling	MatrixDensity	matrix density in milligram per centimetre squared	numeric	GSK/NPL
sample handling	EmbeddingMedia	embedding media used, e.g., paraffin. “None” if no embedding.	alphanumeric	AZ/NPL
sample handling	MountingSubstrate	Material the sample is attached to e.g. glass slide, ITO, silicon wafer.	alphanumeric	NPL
sample provenance	SampleID	unique ID for sample, may link to an in-house database, “SampleType” in NPL terminology	alphanumeric	AZ/NPL
sample provenance	SpeciesID	unique ID for species - controlled dictionary	alphanumeric	GSK
sample provenance	StrainID	unique ID for the population of organisms that is genetically different from others of the same species and possessing a set of defined characteristics.	alphanumeric	GSK
sample provenance	CellTypeID	morphological or functional form of cell. F.e. “epithelial”, “glial” etc. “CellType” in AZ terminology.	alphanumeric	AZ/GSK
sample provenance	CellLineID	a cultured cell population that represents a genetically stable and homogenous population of cultured cells that shares a common propagation history	alphanumeric	GSK
sample provenance	TissueID	unique ID for tissue used in sample - controlled dictionary. “TissueType” in AZ terminology.	alphanumeric	AZ/GSK
sample provenance	TargetID	unique ID for gene/protein target - controlled dictionary	alphanumeric	AZ/GSK
sample provenance	SampleProvider	ID or name of the organisation that provided the sample	alphanumeric	AZ/GSK/NPL[1]
sample provenance	ContainerType	type of slide or vial	alphanumeric	GSK
sample provenance	SectionNumber	if multiple sections are cut, their number/order e.g. section 1	alphanumeric	AZ/GSK
sample provenance	SectionThickness	section thickness in microns	numeric, microns	AZ/GSK/NPL
sample provenance	SlideNumber	slide number	alphanumeric	GSK

[1] Identical to “OriginatingOrganisation” in the Enterprise Metadata section.

5 EXPERIMENTAL METADATA FOR HIGH CONTENT SCREENING

Although NPL does not own high content screening devices, the BIOMET group has confocal laser scanning and structured illumination microscopy facilities and has defined a set of Instrument Settings metadata reflected in Table 4.

Table 4: Metadata terms for high-content screening.

Category	Term	Description	Format & Unit	Used by
data processing	FileType	file extension, f.e. “xdce”	alphanumeric	AZ/GSK
data processing	TimeOfAcquisition	image acquisition timestamp, “creation time” in xdce files	Dublin Core: Date	AZ/GSK
data processing	ProcessingSoftware	Name and version of the software used in conversion of data format	alphanumeric	AZ/GSK
instrument settings	Instrument	instrument manufacturer & model no	alphanumeric	AZ/GSK
instrument settings	Modality	type of microscope in use, e.g., confocal/lightsheet/widefield	alphanumeric	AZ/GSK
instrument settings	ImageSizeX	X dimension of image in megapixels	numeric, megapixels	AZ/GSK
instrument settings	ImageSizeY	Y dimension of image in megapixels	numeric, megapixels	AZ/GSK
instrument settings	Objective	hardware objective lens in use - vendor name and model number	alphanumeric	NPL
instrument settings	Magnification	magnification of objective lens, e.g., '20' for 20x	numeric	AZ/GSK
instrument settings	CameraPixelSize	size of a camera pixel (um x um)	numeric, micron	NPL
instrument settings	NumericalAperture	aperture of objective lens	numeric	NPL
instrument settings	LightSourceType	type of source used, e.g., LED/laser/halogen	alphanumeric	GSK/NPL
instrument settings	LightSourcePower	excitation power for each channel, one value per channel	numeric	NPL
instrument settings	FilterType	type of emission filter, e.g., bandpass, blocking edge, etc.	alphanumeric	NPL
instrument settings	FilterModel	vendor and part number of the emission filter, e.g., Semrock FF01-260/16-25	alphanumeric	NPL
instrument settings	FilterCentreWavelength	emission filter wavelength in nm	numeric, nm	GSK/NPL
instrument settings	FilterFWHM	full width half maximum wavelength of emission filter in nm	numeric, nm	NPL
instrument settings	ExposureDuration	exposure time in milliseconds	numeric, ms	GSK/NPL
instrument settings	ChannelWavelength	wavelength of source for each channel, one value per channel	numeric, nm	GSK/NPL
instrument settings	ChannelID	index number identifying each channel, e.g., 1/2/3/4	numeric	AZ/GSK/NPL
instrument settings	ChannelName	text label corresponding to emission & excitation wavelengths (often the dye used), e.g., TL-Brightfield - dsRed, one value per channel. Can be called “FluorophoreName”	alphanumeric	AZ/GSK/NPL

instrument settings	NumberOfZPlanes	total number of z planes	numeric	AZ/GSK
instrument settings	ZPlanePosition	index identifying order of image in z plane	numeric	AZ/GSK
instrument settings	WellRowNumber	identifies the plate well imaged	numeric	AZ/GSK
instrument settings	WellColumnNumber	identifies the plate well imaged	numeric	AZ/GSK
sample handling	WellTreatmentCompound	compound used for treatment - controlled dictionary	alphanumeric	AZ/GSK
sample handling	WellTreatmentConcentration	concentration of treatment molecule	numeric	AZ/GSK
sample handling	WellType	test sample or control, e.g., positive control/negative control/test	alphanumeric	AZ/GSK
sample handling	PlateType	study number equivalent	alphanumeric	AZ/GSK
sample handling	StainID	the name of fluorophore/staining used, e.g., Hoechst/tubulin - controlled dictionary	alphanumeric	AZ/GSK
sample handling	ProtocolID	unique ID for treatment protocol - controlled dictionary	alphanumeric	AZ/GSK
sample provenance	SampleID	unique ID for sample, may link to an in-house database	alphanumeric	AZ/GSK
sample provenance	SpeciesID	unique ID for species - controlled dictionary	alphanumeric	AZ/GSK
sample provenance	StrainID	unique ID for the population of organisms that is genetically different from others of the same species and possessing a set of defined characteristics	alphanumeric	AZ/GSK
sample provenance	TissueID	unique ID for tissue used in sample - controlled dictionary	alphanumeric	AZ/GSK
sample provenance	TargetID	unique ID for gene/protein target - controlled dictionary	alphanumeric	AZ/GSK
sample provenance	SampleProvider	ID or name of the organisation that provided the sample	alphanumeric	NPL
sample provenance	MaterialEntityID	material of interest to pharmaceutical research & development. Can represent compound or product number	alphanumeric	GSK
sample provenance	SectionNumber	number of the tissue section from which the sample is taken	alphanumeric	GSK

6 EXPERIMENTAL METADATA FOR LIGHT SHEET MICROSCOPY

LSM imaging experiments typically generate from tens of gigabytes to several terabytes of data and require many processing and analysis steps that need to be captured in the metadata. LSM systems typically employ a pair of matched objective lenses, the nose cones of which form an imaging volume into which samples must be positioned. LSM samples are often embedded in a transparent hydrogel such as agarose for stability, with illumination and detection objective lenses dipping into the medium surrounding the sample (aqueous buffer, culture medium or clearing solution) (Reynaud et al. 2015). The LSM system at NPL uses a pair of water dipping objectives with samples typically mounted within agarose or Matrigel on top of a selective plane illumination microscope (SPIM), where a gel-embedded sample sits in a water-filled “imaging pocket”. Single objective variants, such as the oblique plane microscopy method which AZ is currently exploring in collaboration with Imperial College London (Maioli et al. 2016), allow imaging of samples mounted on glass slides and in multi-well plates.

These details of the instrument geometries and mounting method need to be captured in the image metadata to provide a meaningful insight into the experiment. Presently there are no known minimum reporting or metadata standards for LSM to capture these experimental details. **Table 5** presents the summary of LSM metadata captured at AZ and NPL.

Table 5: Metadata entries for light sheet microscopy. ‘NPL*’ refers to the entries that NPL intends to capture in the future.

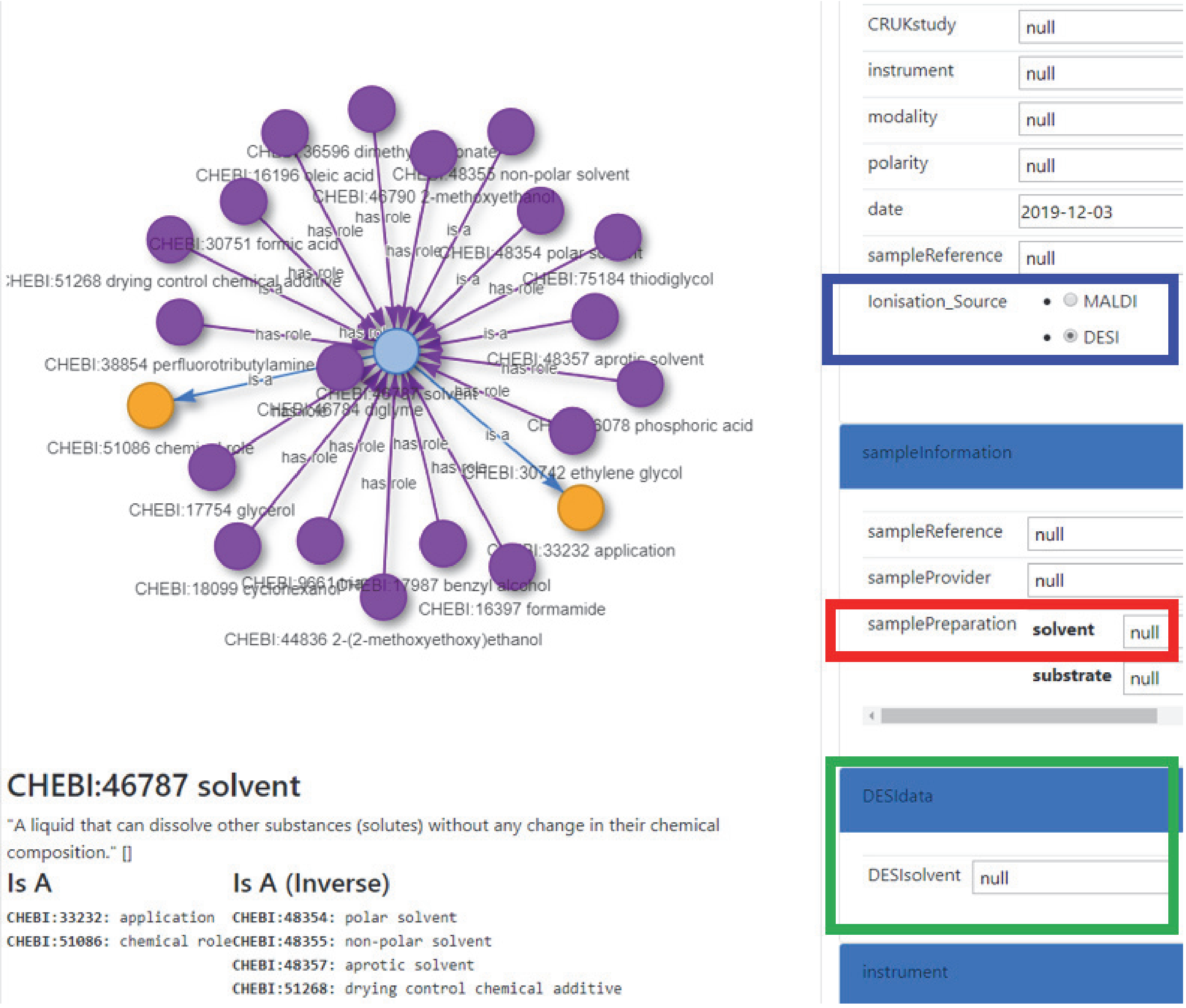
Category	Term	Description	Format	Used by
data processing	FileType	file format, f.e. “OME-TIFF”	alphanumeric	NPL
data processing	PSFFile	file that contains the Point Spread Function data used to deconvolve the raw images	alphanumeric	NPL
data processing	TimeOfAcquisition	image acquisition timestamp	ISO 8601	NPL
data processing	DeconvolutionSoftware	name and version of the deconvolution software used	alphanumeric	NPL
data processing	NumberOfIterations	number of iterations used by deconvolution software to produce the final image	numeric	NPL
data processing	GeometricCorrection	“yes/no” to whether geometric correction was applied during deconvolution	boolean, True/False	NPL
data processing	ImageAnalystName	name of the person who ran the image analysis. GDPR-sensitive entry.	alphanumeric	NPL
data processing	DimensionOrder	data organisation on pixel level, f.e. XYCZT (x/y-dimensions, channel, plane, timepoint)	alphanumeric	NPL
data processing	IsBigEndian	whether byte order for pixel values is most-significant-byte-first	numeric	NPL
data processing	NumberOfBits	number of bits per pixel, f.e. 16	numeric	NPL
data processing	NumberFormat	number format of pixel values, f.e. “uint16” for unsigned 16-bit integer	alphanumeric	NPL
instrument settings	Instrument	instrument manufacturer and model number	alphanumeric	NPL
instrument settings	InstrumentSerialNumber	serial number of the instrument	alphanumeric	NPL
instrument settings	Modality	type of microscope in use, e.g., confocal/lightsheet/widefield	alphanumeric	NPL
instrument settings	StageLabelName	LMS stage label name, f.e. “POS0”	alphanumeric	NPL

instrument settings	ImageSizeX	X dimension of image in pixels	numeric	NPL
instrument settings	ImageSizeY	Y dimension of image in pixels	numeric	NPL
instrument settings	ImageSizeZ	number of slices in stack	numeric	NPL
instrument settings	NumberOfChannels	number of channels per pixel value	numeric	NPL
instrument settings	NumberOfTimepoints	number of timepoints per pixel value	numeric	NPL
instrument settings	PixelSizeX	pixel size in x-dimension in µm	numeric, micron	NPL
instrument settings	PixelSizeY	pixel size in y-dimension in µm	numeric, micron	NPL
instrument settings	ZStepSize	separation between image focal planes in µm	numeric, micron	NPL
instrument settings	ZStepDirection	directionality of the movement between z- planes as “-“ or “+” sign	alphanumeric	NPL
instrument settings	Objective	hardware objective lens in use - vendor name and model number	alphanumeric	NPL*
instrument settings	Magnification	magnification of objective lens, e.g., '20' for 20x	numeric	NPL*
instrument settings	NumericalAperture	aperture of objective lens	numeric	NPL*
instrument settings	LightSourceType	type of source used, e.g., LED/laser/halogen	alphanumeric	NPL*
instrument settings	LightSourcePower	excitation power for each channel, one value per channel	numeric	NPL*
Instrument settings	EmissionFilterID	name of the emission filter, f.e. “CY5 647LP” or “RFP 600-52”	alphanumeric	NPL
instrument settings	FilterType	type of emission filter, e.g., bandpass, blocking edge, etc.	alphanumeric	NPL*
instrument settings	FilterModel	vendor and part number of the emission filter, e.g., Semrock FF01-260/16-25	alphanumeric	NPL*
instrument settings	FilterCentreWavelength	emission filter wavelength in nm	numeric, nm	NPL
instrument settings	FilterFWHM	full width half maximum wavelength of emission filter in nm	numeric, nm	NPL*
instrument settings	ExposureDuration	exposure time in milliseconds	numeric, milliseconds	NPL*
instrument settings	NumberOfChannels	number of colour channels per pixel	numeric	NPL
instrument settings	NumberOfTimePoints	number of time points per pixel per channel	numeric	NPL
instrument settings	ChannelWavelength	wavelength of source for each channel, one value per channel	numeric, nm	NPL*
instrument settings	LaserWavelenth	laser wavelength in nanometres	numeric	NPL*
instrument settings	SampleTemperature	sample temperature in degrees Celcius for live imaging experiments	numeric	NPL
instrument settings	GasComposition	gas composition (% of CO2 etc.) for live imaging experiments	numeric	NPL
sample handling	StainID	the name of fluorophore/staining used, e.g., Hoechst/tubulin - controlled dictionary	alphanumeric	NPL
sample handling	ProtocolID	unique ID for treatment protocol	alphanumeric	NPL
sample handling	MountingChamber	f.e. polydimethylsiloxane (PDMS) plinth in weigh boat, slide or vial	alphanumeric	NPL
sample handling	EmbeddingMediaName	f.e. “agarose”	alphanumeric	NPL
sample handling	EmbeddingMediaConcentration	f.e., % of agarose in the embedding media	numeric, percent	NPL

sample handling	ImmersionMediaName	f.e. phosphate buffered saline (PBS)	alphanumeric	NPL
sample handling	ImmersionMediaComposition	f.e. “66% TDE + 34% DI water”	alphanumeric	NPL
sample handling	SamplePreparatorName	name of the person who prepared the sample. GDPR-sensitive entry.	alphanumeric	NPL
sample handling	ImageOperatorName	name of the person who imaged the sample. GDPR-sensitive entry.	alphanumeric	NPL
sample handling	SampleHandlingNotes	other comments on sample preparation	alphanumeric	NPL
sample provenance	SampleID	unique ID for sample, may link to an in-house database	alphanumeric	NPL
sample provenance	SpeciesID	unique ID for species - controlled dictionary	alphanumeric	NPL
sample provenance	StrainID	unique ID for the population of organisms that is genetically different from others of the same species and possessing a set of defined characteristics.	alphanumeric	NPL
sample provenance	TissueID	unique ID for tissue used in sample - controlled dictionary	alphanumeric	NPL
sample provenance	SampleProvider	f.e. “University of Cambridge”	alphanumeric	NPL

7 TOWARDS FAIR BIOIMAGING DATA ANNOTATION AT NPL

In the effort to improve the reproducibility and the re-usability of bioimaging data, NPL's Data Science team is running two collaborative pilot projects with the NiCE-MSI and BIOMET teams with a long-term view to making software-assisted metadata capture an organisation-wide endeavour.


CHEBI:36596 dimethyl ...onate
CHEBI:16196 oleic acid
CHEBI:48355 non-polar solvent
CHEBI:46790 2-methoxyethanol
has role
has role
is a
CHEBI:30751 formic acid
has role
CHEBI:48354 polar solvent
:HEBI:51268 drying control chemical additive
has role
is a
is a
has role
CHEBI:75184 thiodiglycol
has role
has role
is a
CHEBI:38854 perfluorotributylamine
CHEBI:48357 aprotic solvent
is a
has role
CHEBI:46787 solvent
has role
CHEBI:46784 diglyme
has role
CHEBI:16078 phosphoric acid
CHEBI:51086 chemical role
is a
has role
has role
has role
has role
CHEBI:30742 ethylene glycol
has role
CHEBI:17754 glycerol
CHEBI:33232 application
CHEBI:18099 cyclohexanol
CHEBI:9661 ...
CHEBI:17987 benzyl alcohol
CHEBI:16397 formamide
CHEBI:44836 2-(2-methoxyethoxy)ethanol

CHEBI:46787 solvent

"A liquid that can dissolve other substances (solutes) without any change in their chemical composition." []

Is A	Is A (Inverse)
CHEBI:33232: application	CHEBI:48354: polar solvent
CHEBI:51086: chemical role	CHEBI:48355: non-polar solvent
	CHEBI:48357: aprotic solvent
	CHEBI:51268: drying control chemical additive

CRUKstudy	null
instrument	null
modality	null
polarity	null
date	2019-12-03
sampleReference	null
Ionisation_Source	○ MALDI ● DESI

sampleInformation

sampleReference	null	
sampleProvider	null	
samplePreparation	solvent	null
	substrate	null

DESIdata

DESIsolvent	null

instrument

Figure 5: MSI metadata entry form displaying a variety of DESI solvents following the user selection of "DESI" ionisation source (blue box, top) and highlighting the "solvent" field (red box, middle). The metadata prompts for "DESIdata" (green box, bottom) were automatically generated following the user selections.

Within these pilot projects, a dedicated web form interface has been designed to capture the sample handling and instrument-setup-specific metadata to enhance the metadata that is automatically captured by the imaging equipment.

The metadata entry forms are dynamically generated from a purpose-built ontology. Thus, each choice the user makes generates a form that contains a set of metadata fields to be completed that are relevant only within the context of the selected experiment type and imaging modality. Where applicable, entries are populated from the purpose-built ontology that combines ChEBI ontology (de Matos et al. 2010) and in-house data elements. **Figure 5** illustrates the MSI metadata capture with suggestions for DESI solvent presented to the user upon the selection of DESI device and the entry for DESI solvent.

In the BIOMET team, the microscopy laboratory generates hundreds of gigabytes of data yearly. In the absence of electronic laboratory notebooks, the metadata were captured within

Word document templates. The web framework developed for MSI data annotation is being extended and adapted to annotate the structured illumination microscopy, confocal laser scanning microscopy and LSM images with metadata on sample provenance and sample handling.

Once the annotation has been completed, the metadata annotations are saved in the XML file and bundled with the image data. The annotated image is automatically transferred for long term storage to NPL Object Store, where it is made findable and browsable via the Object Store search mechanism.

The BIOMET and MSI data annotation pilot projects will be used to inform the immediate work to take place in 2020-2021 that will focus on the development of domain-agnostic measurement ontology. This ontology will be high-level, with sufficient modularity and flexibility to be adapted to the needs of different measurement science domains.

8 SUMMARY

Currently in life sciences imaging, there is little consistency in the generation and format of metadata produced. Most of the metadata identified in this report is currently recorded in a variety of formats and is not necessarily integrated with the imaging data to which it refers. In addition, different naming conventions mean data sharing within or between groups or organisations has an extra layer of complication.

Here we have compiled a list of metadata currently captured and used at three high-profile organisations involved in bioimaging.

We have focused on three imaging modalities: high-content screening, mass spectroscopy imaging, and light-sheet microscopy, though the captured metadata may be applied to other bioimaging domains. Future work on a minimum recommended list of metadata entries that should be integrated alongside the imaging data should accelerate data recall, enable data re-use, make sharing between research groups or laboratories simpler and increase reproducibility & accountability for experiments.

9 DISCUSSION

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 on the experiment settings, tissue morphology, biological processes and drug-tissue interactions. Presently, this information has limited availability due to siloed storage, lack of structured and standardised annotations, and has complex interfaces between software packages used to store, access and analyse the data as well as a lack of consensus on data formats and required annotations between equipment vendors, users and regulators.

To evolve from these currently captured metadata into a much-needed minimum metadata standard, further discussion with the wider community is required. It is important that such standards should be open to the community and are developed by a consortium of researchers, equipment vendors, users and regulators. Efforts are already being undertaken by international bodies that share bioimaging research data such as Image Data Resource (E. Williams et al. 2017) and the more recent EMBL-EBI initiative BioImage Archive that ran an international workshop on minimum metadata definition for bioimaging data for cryogenic electron microscopy, cellular light microscopy and correlative imaging in 2019.

There is huge value in having a consistent list of metadata elements across imaging in life sciences. Compatibility and interoperability improve not only the quality of the dataset, but also allow for the re-use of the dataset beyond the original study or project. Finally, machine-readable, harmonised analysis-ready data open the doors to big data analytics including conventional image processing or machine learning, allowing to gather new scientific evidence from cross-experiment and cross-imaging-domain data studies. Thus, the efforts required to develop the organisation culture and to implement the systems for data annotation lead to streamlining of operations, reduced time costs and unlocking the long-term value in data assets.

10 ACKNOWLEDGEMENTS

The Federation of Imaging Data for Life Sciences (FIDL) project is a collaboration between the National Physical Laboratory (NPL), AstraZeneca (AZ) and GlaxoSmithKline (GSK). The authors would like to thank all AZ, GSK and NPL scientists and information architects who provided materials for this report. Special thanks go to Claus Bendtsen & Yinhai Wang (AZ) and Simon Thornber & Joseph Lavelle (GSK) who led the collaboration from the partners' side, enabled the information exchange and provided business view into the bioimaging data management.

11 REFERENCES

Allan, Chris, Jean-Marie Burel, Josh Moore, Colin Blackburn, Melissa Linkert, Scott Loynton, Donald MacDonald, et al. 2012. ‘OME Remote Objects (OMERO): A Flexible, Model-Driven Data Management System for Experimental Biology’. *Nature Methods* 9 (3): 245–53. https://doi.org/10.1038/nmeth.1896.

Bodenreider, Olivier. 2004. ‘The Unified Medical Language System (UMLS): Integrating Biomedical Terminology’. *Nucleic Acids Research* 32 (Database issue): D267–70. https://doi.org/10.1093/nar/gkh061.

Buchser, William, Mark Collins, Tina Garyantes, Rajarshi Guha, Steven Haney, Vance Lemmon, Zhuyin Li, and O. Joseph Trask. 2004. ‘Assay Development Guidelines for Image-Based High Content Screening, High Content Analysis and High Content Imaging’. In *Assay Guidance Manual*, edited by G. Sitta Sittampalam, Abigail Grossman, Kyle Brimacombe, Michelle Arkin, Douglas Auld, Christopher P. Austin, Jonathan Baell, et al. Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences. http://www.ncbi.nlm.nih.gov/books/NBK100913/.

Ding, Ying, Yuyin Sun, Bin Chen, Katy Borner, Li Ding, David Wild, Melanie Wu, et al. 2010. ‘Semantic Web Portal: A Platform for Better Browsing and Visualizing Semantic Data’. In *Active Media Technology*, edited by Aijun An, Pawan Lingras, Sheila Petty, and Runhe Huang, 448–60. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-15470-6_46.

Dunsby, C. 2008. ‘Optically Sectioned Imaging by Oblique Plane Microscopy’. *Optics Express* 16 (25): 20306–16. https://doi.org/10.1364/OE.16.020306.

Ellenberg, Jan, Jason R. Swedlow, Mary Barlow, Charles E. Cook, Ugis Sarkans, Ardan Patwardhan, Alvis Brazma, and Ewan Birney. 2018. ‘A Call for Public Archives for Biological Image Data’. *Nature Methods* 15 (11): 849–54. https://doi.org/10.1038/s41592-018-0195-8.

Girkin, J. M., and M. T. Carvalho. 2018. ‘The Light-Sheet Microscopy Revolution’. *Journal of Optics* 20 (5): 053002. https://doi.org/10.1088/2040-8986/aab58a.

Goldberg, Ilya G, Chris Allan, Jean-Marie Burel, Doug Creager, Andrea Falconi, Harry Hochheiser, Josiah Johnston, Jeff Mellen, Peter K Sorger, and Jason R Swedlow. 2005. ‘The Open Microscopy Environment (OME) Data Model and XML File: Open Tools for Informatics and Quantitative Analysis in Biological Imaging’. *Genome Biology* 6 (5): R47. https://doi.org/10.1186/gb-2005-6-5-r47.

Gustafsson, Ove J. R., Lyron J. Winderbaum, Mark R. Condina, Berin A. Boughton, Brett R. Hamilton, Eivind A. B. Undheim, Michael Becker, and Peter Hoffmann. 2018. ‘Balancing Sufficiency and Impact in Reporting Standards for Mass Spectrometry Imaging Experiments’. *GigaScience* 7 (10). https://doi.org/10.1093/gigascience/giy102.

Huisman, Maximiliaan, Mathias Hammer, Alex Rigano, Farzin Farzam, Renu Gopinathan, Carlas Smith, David Grunwald, and Caterina Strambio-De-Castillia. 2020. ‘Minimum Information Guidelines for Fluorescence Microscopy: Increasing the Value, Quality, and Fidelity of Image Data’. *ArXiv:1910.11370 [Cs, q-Bio]*, January. http://arxiv.org/abs/1910.11370.

Linkert, Melissa, Curtis T. Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, et al. 2010. ‘Metadata Matters: Access to Image Data in the Real World’. *The Journal of Cell Biology* 189 (5): 777–82. https://doi.org/10.1083/jcb.201004104.

Maioli, Vincent, George Chennell, Hugh Sparks, Tobia Lana, Sunil Kumar, David Carling, Alessandro Sardini, and Chris Dunsby. 2016. ‘Time-Lapse 3-D Measurements of a Glucose Biosensor in Multicellular Spheroids by Light Sheet Fluorescence Microscopy in Commercial 96-Well Plates’. *Scientific Reports* 6 (1): 1–13. https://doi.org/10.1038/srep37777.

Matos, Paula de, A Dekker, M Ennis, Janna Hastings, K Haug, S Turner, and Christoph Steinbeck. 2010. ‘ChEBI: A Chemistry Ontology and Database’. *Journal of Cheminformatics* 2 (Suppl 1): P6. https://doi.org/10.1186/1758-2946-2-S1-P6.

McCarthy, Mark I., Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. 2008. ‘Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges’. *Nature Reviews. Genetics* 9 (5): 356–69. https://doi.org/10.1038/nrg2344.

McDonnell, Liam A., Andreas Römpp, Benjamin Balluff, Ron M. A. Heeren, Juan Pablo Albar, Per E. Andrén, Garry L. Corthals, Axel Walch, and Markus Stoeckli. 2015. ‘Discussion Point: Reporting Guidelines for Mass Spectrometry Imaging’. *Analytical and Bioanalytical Chemistry* 407 (8): 2035–45. https://doi.org/10.1007/s00216-014-8322-6.

Race, Alan M, and Andreas Römpp. 2018. ‘Error-Free Data Visualization and Processing through ImzML and MzML Validation’. *Analytical Chemistry* 90 (22): 13378–84. https://doi.org/10.1021/acs.analchem.8b03059.

Tilahun, Binyam, Tomi Kauppinen, Carsten Keßler, and Fleur Fritz. 2014. ‘Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation’. *JMIR Medical Informatics* 2 (2). https://doi.org/10.2196/medinform.3531.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship’. *Scientific Data* 3 (March). https://doi.org/10.1038/sdata.2016.18.

Williams, Eleanor, Josh Moore, Simon W. Li, Gabriella Rustici, Aleksandra Tarkowska, Anatole Chessel, Simone Leo, et al. 2017. ‘The Image Data Resource: A Bioimage Data Integration and Publication Platform’. *Nature Methods* 14 (8): 775–81. https://doi.org/10.1038/nmeth.4326.

Williams, Elizabeth H., Pamela Carpentier, and Tom Misteli. 2012. ‘The JCB DataViewer Scales Up’. *Journal of Cell Biology* 198 (3): 271–72. https://doi.org/10.1083/jcb.201207117.

Zock, Joseph M. 2009. ‘Applications of High Content Screening in Life Science Research’. *Combinatorial Chemistry & High Throughput Screening* 12 (9): 870–76. https://doi.org/10.2174/138620709789383277.

12 WEB REFERENCES

All online resources used in this report were valid as accessed in May 2020.