BioMedBridges

Deliverable D6.2

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	List of standards and ontologies in mouse image data sets	
WP No.	6	
Lead Beneficiary:	11: HMGU	
WP Title	Interoperability of large scale image data sets from different biological scales	
Contractual delivery date:	31 December 2013	
Actual delivery date:	23 December 2013	
WP leader:	Jan Ellenberg	1: EMBL
Partner(s) contributing to this deliverable:	11: HMGU	

Authors: Frauke Neff, Philipp Gormanns

Contents

1 Executive summary 3
2 Project objectives 3
3 Detailed report on the deliverable 4
3.1 Background 4
3.2 File Formats 5
3.3 Remote access 6
3.4 Image annotation 7
3.5 Conclusion 9
4 Delivery and schedule 9
5 Adjustments made 9
6 Background information 9

1 Executive summary

The WP6 use case addresses interoperability of large-scale image data sets, which is required for reusing and analysing data sets generated by different sources.

The following aspects of image data sets significantly influence their interoperability:

1. File formats
2. Accessibility of the data
3. Image annotation

Image datasets within the mouse scientist community consist mainly of proprietary raw image formats provided by the proprietary software running the slide scanning machines. This high diversity contributes to the low interoperability of the data coming from different sources. Especially the international mouse phenotyping consortium (IMPC) provides harmonisation approaches regarding the data sources and file formats. Here, the most commonly used file formats are ndpi-files as raw format which can be converted into jpg or tiff files. These image will be made visible within a public accessible database in an annotated form.

Finally, consistent annotation of mouse tissue image data sets is required for their interoperability. Although some ontologies for mouse anatomy and mouse pathology are available, none of them covers comprehensively the histological or the cytological phenotype. Therefore, image data sets are normally annotated using free text or a diverse set of ontologies. To bridge the gap between existing ontologies, WP6 partners are currently developing the cellular microscopy phenotype ontology (CMPO[1]).

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

[1] http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=CMPO

No.	Objective	Yes	No
1	Demonstrate the utility of the interoperability of large-scale image data sets from different biological scales (cell – tissue – organism)		x
2	Enable comparison of morphological image data of cellular phenotypes of individual genes with morphological image data of the diseased tissues in mouse and human	x	
3	Link imaging data with molecular data, including cancer genome sequences and cancer expression data		x

3 Detailed report on the deliverable

3.1 Background

Since their inception in the 17th century, light microscopes have been some of the most widely used instruments to study cells. While in the past, photographic film was the standard for image representation, the advent of digital imaging has simplified acquisition, storage and sharing of microscopy images. It has also opened the door to computational data analysis of images. However, these benefits also come with challenges associated with computational processing of this type of data. Since microscopy was always the tool for analysis of pathology samples and tissues, and becoming part of a broader analytical context, data processing of multiple datasets has become desirable. For this, interoperability of datasets is required - it is necessary to integrate image data sets resulting from different modes of acquisition and/or image data produced at different institutions.

As a test case of interoperability of image data sets from different sources, WP6 is set to look for ways to overcome obstacles created by this diversity. Here, we report the results of our analysis of the file formats and ontologies used for annotations by the cell biology, mouse and human pathology communities. This analysis is required to propose a solution for the integration of different image data sets.

3.2 File Formats

Images captured by microscopes and slide scanner are saved in one of many different file formats.

Two things contribute the most to the diversity of file formats:

1. Compression of images
2. Diversity of saved metadata

Images captured by digital cameras are matrices of numerical data. As raw image matrices generate large files, it is necessary to reduce file size by the application of compression algorithms. Therefore, reading and writing image files requires processing through coding/decoding algorithms. The diversity of compression methods has resulted in a proliferation of different file formats for the storage of images.

In addition, the diversity of experiments has delayed community agreement on what metadata should be included in image files.

As a consequence of both of these factors, a large fraction of image acquisition software use their own file formats to store image data and their selection of metadata. This results in more than a hundred different file formats being in common use and evolving with imaging technology. In Table 1 we list some of the most commonly used file formats for tissue imaging.

Table 1 *File formats commonly used in slide imaging*

Olympus .slide vsi
Aperio SVS
Zeiss MRXS
Zeiss CZI
Leica SCN
Hamamatsu NDPI

These file formats have the advantage of providing a dynamical picture, e.g. with the ability to zoom in and out of the scan, in contrast to the static file

formats tiff, jpg, etc. These dynamic formats are achieved by multiple compressed pictures being embedded within a frame by an algorithm. Many of these algorithms are patened, resulting in companies focusing on those formats for which they have a licence. Moreover, there is no comprehensive and up to date list of patent-free codecs available to researchers.

In order to deal with the diversity of existing file formats, each application for microscopy image processing must use decoders for several formats or use conversion software. Furthermore, the often undocumented variations in metadata stored in the image file contribute to the complexity of the process. As a result, conversion is usually done only for pixel data, which leads to the loss of associated metadata.

In recent years, a community effort has been made to promote standard open file formats for individual images and high throughput microscopy datasets. For the former, OME-TIFF[2] is the currently chosen standard. It standardises image compression and metadata in the image file, improving interoperability of image data sets. To this end, OME-TIFF is supported by a software library (Bio-Formats[3]) for converting dozens of different file formats to OME-TIFF. The other format promoted by the community is HDF5, which allows storing entire datasets and large amounts of metadata and derived data in one file, making it a favourable solution for high-throughout microscopy.

3.3 Remote access

Owing to the size of most image data, transfer of complete data sets over the internet is difficult. Existence of publicly accessible image repositories would enable researchers to work also with data they have not produced themselves, which would definitely promote re-use of datasets. However, publicly available repositories are extremely scarce and limited to a few datasets. Therefore, most image data sets are not generally made publicly available, and when they are, they are distributed by the producing institutions, each providing their own solution of image databases. These in-house image databases have become necessary to manage the large collections of images and associated data now routinely produced in many laboratories. However, for external users

[2] http://www.openmicroscopy.org/site/support/ome-model/ome-tiff/
[3] http://www.openmicroscopy.org/site/products/bio-formats

these are not easy to discover (their content is not referenced by web search engines) and there is no standard access mode to these databases. An additional obstacle is that these repositories usually cannot be searched or browsed in a meaningful way outside of the scope of the original project.

A solution to these problems will come from Euro-BioImaging in the coming years. Euro-BioImaging will, together with ELIXIR, provide access to standardized, annotated image data repositories of general relevance for the research community in a common European repository of standardized and quality controlled tools for image data analysis.

3.4 Image annotation

Image annotation associates information about the sample in the image and the content of the image with the image itself. Systematic descriptions of images are key to their reuse; making full use of image data requires machine-readable annotations. Ontologies are a means to formally express knowledge in a machine-readable form. By defining controlled vocabularies and relationships between vocabulary terms, ontologies enable consistent descriptions of images across data sets. Ontology-based annotations are computable such that algorithms can be applied to the automatic retrieval or classification of images.

Ontologies cover many different domains such as biological entities, medical conditions or experiment descriptions. In Table 2 we provide a list of ontologies that are relevant for cellular imaging annotation. A full list of available ontologies related to the biomedical field can be found on http://bioportal.bioontology.org.

Although this list of ontologies covers many different cellular aspects, some aspects have not yet been captured and suitable ontologies are missing. One of the most important aspects for the description of cellular images which is poorly covered by ontologies is a cellular phenotype. The lack of this particular ontology is directly linked to the activities of the WP6, which is addressing interoperability of the image data sets and relies heavily on the description of cellular phenotypes. To be interoperable, cellular images have to be annotated with a suitable ontology. In particular, description of phenotypes at the cellular level generally cannot be made using ontologies aimed at describing 'normal'

cells or 'normal' cellular processes, such as those captured by the Gene Ontology biological process domain or using ontologies describing phenotypes at a different scale (e.g. tissue or organ). To bridge the gap between existing ontologies, the cellular microscopy phenotype ontology (CMPO) suitable for use in WP6 is currently under development by the WP6 partners. This ontology will be the best practice example and solution for the annotation of cellular image data sets in a way that allows interoperability with image data sets, including those coming from mouse and human tissue samples.

Full interoperability of data sets on multiple levels depends also on the existing links between ontologies. For example, a full description of cellular phenotypes is likely to also require information about the experimental context as captured by other ontologies such as the experimental factors or bioassays ontologies or as provided in reporting guidelines such as MIACA (Minimum Information About a Cellular Assay), MIARE (Minimum Information about an RNAi experiment) or the ISA metadata tracking tools[4]. In addition, the full description of mouse tissue phenotypes will require the genotype and linked mutation as metadata. Unfortunately, many related ontologies are not linked so that additional effort is required to map terms across ontologies.

Even with existing ontologies, capturing information relevant for all possible use of the images is not practical. As a result, most images are only annotated with basic information relevant to the project for which they have been produced. Therefore, reusing image data often involves re-annotation through a manual curation process, although for some data sets automated annotation is becoming possible.

Table 2 *Ontologies relevant to tissue imaging*

Gene Ontology: Biological process
Cell ontology
Phenotypic Quality Ontology
Mammalian Phenotype Ontology
Mammalian pathology ontology (MPATH-Pathbase)
Adult Mouse Anatomy Dictionary

[4] http://isa-tools.org/

3.5 Conclusion

Analysing the pool of used file formats and annotations, we can see several challenges that stand in the way of easy interoperability of imaging data and propose the following solutions:

- promote standard representation of the numerical matrices representing digital images and their metadata, which will allow for easy processing and visualization of images
- apply consistent annotation of cellular and tissue images, particularly of cellular phenotype, allowing sharing and analysis of diverse image data
- improve access to image repositories, increasing their visibility and promoting standard querying modes

Progress is being made by the gradually wider adoption of the OME-TIFF and HDF5 file formats in the cell biology community. Euro-BioImaging will provide central access to image data repositories in the future. The consistent annotation of cellular microscopy phenotypes still has to be addressed and WP6 is currently putting efforts in this direction, developing a standard Cellular Microscopy Phenotype Ontology.

4 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

5 Adjustments made

No adjustments were made.

6 Background information

This deliverable relates to WP 6; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 6 Title: Interoperability of large scale image data sets from different biological scales

Lead: Jan Ellenberg (EMBL)
Participants: EMBL, HMGU, CIRMMP

Work package number	WP6	**Start date or starting event:**	month 13
Work package title	PhenoBridge-crossing the species bridge between mouse and human		
Activity Type	RTD		

Participant number	1:EMBL	11:HMGU	20: CRIMMP
Person-months per participant	37	27	16

Objectives

This work package will demonstrate the utility of the interoperability of large scale image data sets from different biological scales (cell – tissue – organism) to enable drug target and biomarker discovery for human disease with cancer as an example. Based on the standards and services developed in WP2 and 3, we will use integrated access to systematic imaging data of disease gene function in cultured human cells (EMBL-Heidelberg) and systematic imaging data available from tissue microarrays of diseased tissue from both human patients (FIMM, U Helsinki) and mouse models (HMGU).

This use case will thus link the four BMS ESFRI infrastructures Euro-BioImaging (EMBL-Heidelberg), BBMRI (U Helsinki), EATRIS (U Helsinki-FIMM) and Infrafrontier (HMGU) with the standards and services provided by ELIXIR (EMBL-EBI) and require strong links to ELIXIR's molecular data resources. The comparison of morphological image data on cellular phenotypes of individual genes, with morphological image data of the diseased tissues in mouse models and human patients could create a powerful predictor of optimized biomarkers as well as drug targets in cancer. Linking these imaging data with molecular data including the cancer genome sequence and cancer expression data, will allow in silico validation of the predictions and prioritization of biomarkers for validation in clinical research.

Description of work and role of participants

Task 1.1. Implementation of interoperability standards and ontologies for reference image data sets
(Leader: U. Helsinki-FIMM, Participants: EMBL-Heidelberg, U Helsinki, EMBL-EBI, HMGU)

The first task to make integrated image data access usable is to map the (meta)data standards and ontologies present within each image data domain (cell, human and mouse tumor tissue) onto each other to enable correlative analysis. In line with the standards and services developed in WP2, 3 and where applicable respecting the secure access to medical data developed in WP4, we will implement unambiguous maps between the respective metadata. For this we will select high throughput imaging reference data sets with cancer related assays (e.g. www.mitocheck.org, www.cellmigration.org/resource/discovery/#genes) as well as tumor tissue and clinical data (we will make these available at: fimm.webmicroscope.net/)

Task 1.2. Prediction of novel cancer biomarkers (e.g. breast and prostate cancer) (Leader: EMBL-Heidelberg, Participants: FIMM, U Helsinki, HMGU, EMBL-EBI).

Correlative analysis of interoperable cell and tissue image datasets with their associated annotation and metadata will be mined with state of the art bioinformatic tools to predict novel biomarker candidates. In particular we will focus on genes with function in cell cycle and cell division control as well as invasive behavior, for which comprehensive molecular and cellular datasets are available. An initial set of predicted biomarkers will then be further cross-validated against the biomolecular databases hosted by ELIXIR, drawing on cancer genome and expression data, as well as general sequence and structural properties of the identified genes to also explore their potential as drug targets.