

Landscape analysis of metadata usage in major bioimage repositories

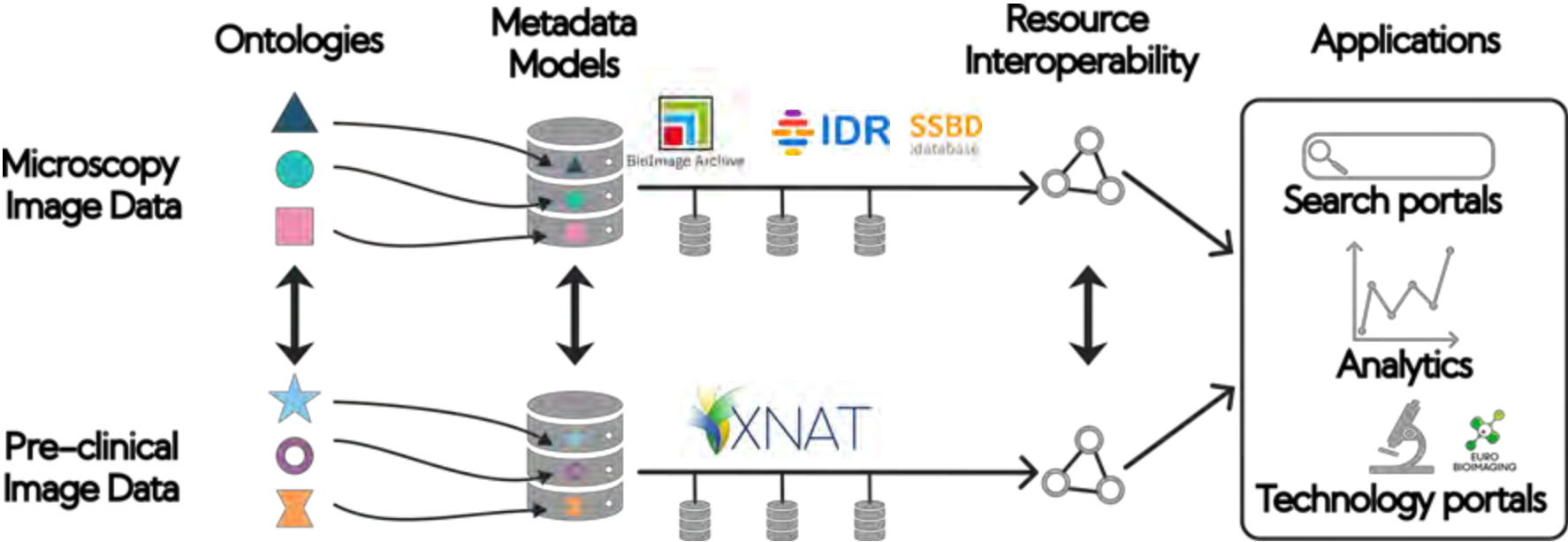
Shuichi Onami^{1,2}, Koji Kyoda¹, Yuki Yamagata^{2,3}

¹RIKEN Center for Biosystems Dynamics Research

²RIKEN Information R&D and Strategy Headquarters

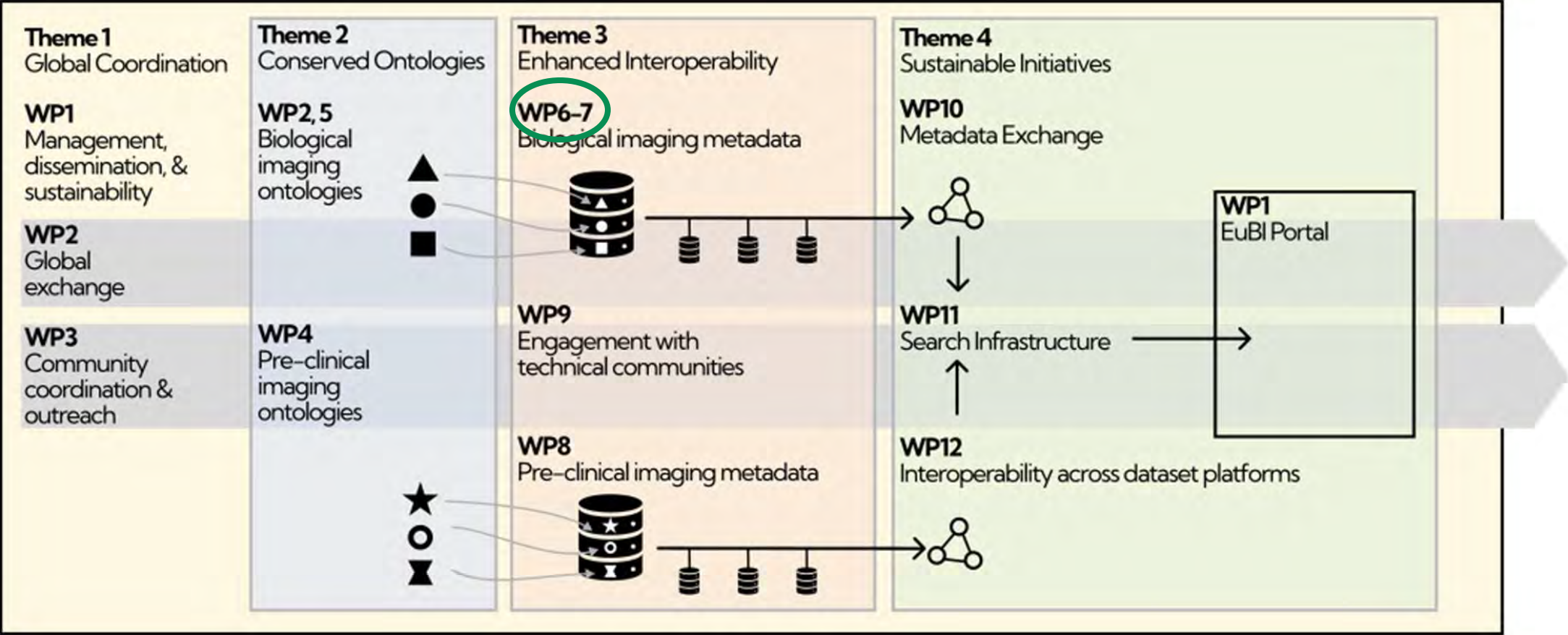
³RIKEN BioResource Research Center

foundingGIDE



foundingGIDE Project

Groundwork for harmonization of BioimageArchive, IDR, and SSBD metadata models



Motivation

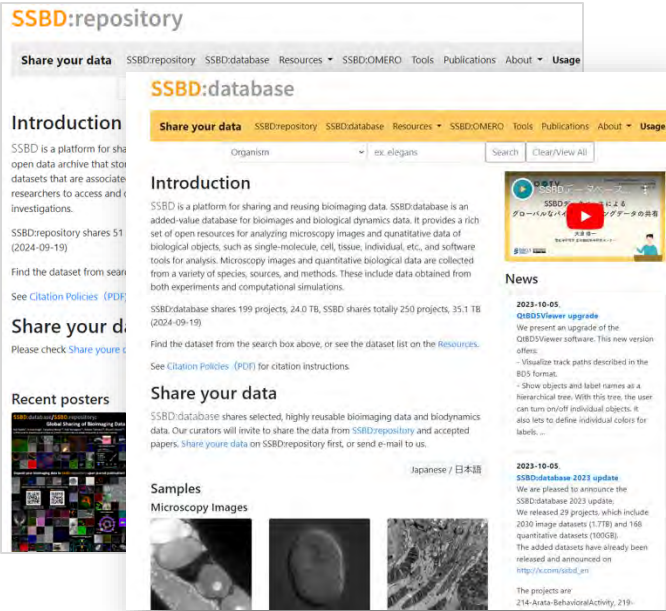
To harmonize the metadata models among resources in the Global Image Data Ecosystem, it is necessary

- **to analyze and compare the requirements of the existing metadata models** used by resources in the bioimaging field,
- **elucidate the overlaps and differences** among the models, and
- **map the model components.**

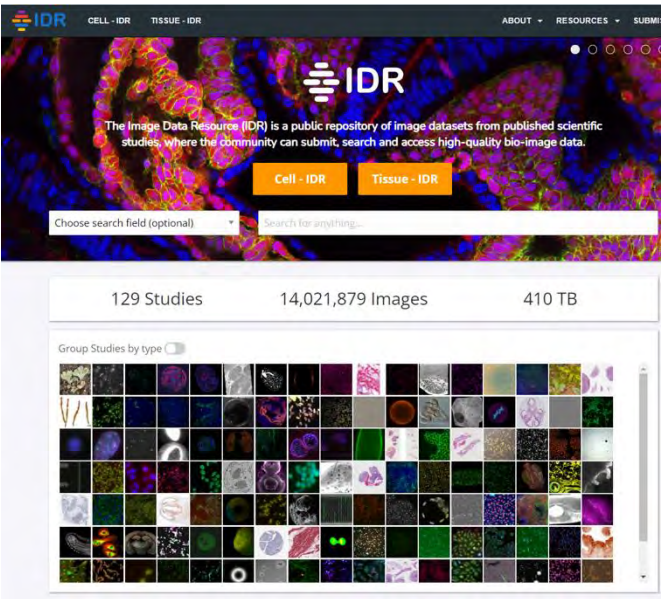
The target existing bioimage data resources

- BioImage Archive and IDR in Europe, and SSBD in Japan are the key existing repositories and added-value databases.

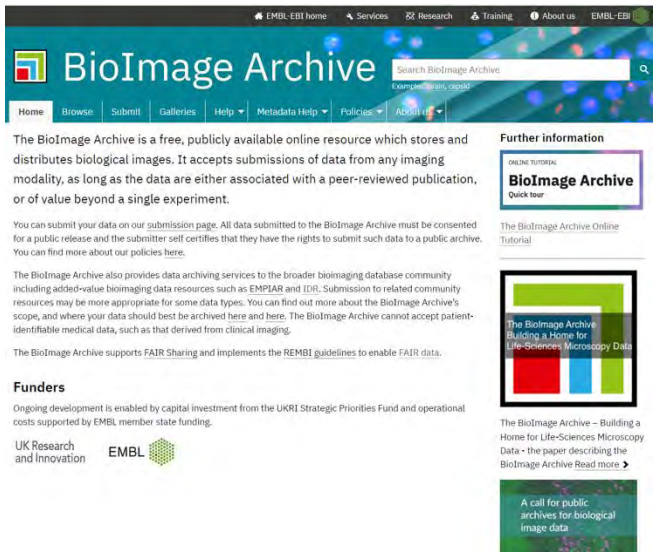
SSBD



IDR



BioImage Archive



Approaches

- We conducted
 1. Analyses the models of BIA, IDR, and SSBD, identify the overlaps and differences among the metadata models and the correspondence among model components and ontology, and determine any mismatches between them.
 2. Use these results to examine the requirements for achieving metadata harmonization among data resources in the GIDE.
 3. Propose criteria for harmonizing metadata for the BIA, IDR, and SSBD and make recommendations that further harmonization (or standardization) of metadata models.

Metadata models of BIA, IDR, and SSBD

- BIA and IDR kindly provided us with their metadata models.
- Each metadata component of the models is mapped onto the attributes of the REMBI model.



BIA



IDR



SSBD



Module	Attribute	Comments	Data entry method	Relevant existing standards and ontologies
Study	Study type	Type of the overall study, which may include other imaging and/or non-imaging data	text, ontology	EDAM-BIOIMAGING, FBbi, EFO, IDR
	Study description	Study description, e.g., title of published paper	text	IDR
	General dataset info	Authors, publications, licenses etc	misc.	Dublin Core, DataCite Metadata, schema.org, IDR
Study component	Imaging method	Technique used to acquire image data	ontology	EDAM-BIOIMAGING, FBbi, OME
	Study component description	Description specific to this image dataset component	text	IDR
Biosample	Identity	Internal unique ID		
	Biological entity	What is being imaged	text and/or ontology entry (multiple possible)	EFO
	Organism	Species (multiple possible)	taxonomy	NCBI Taxonomy
	Intrinsic variable	Intrinsic (e.g. genetic) alteration if applicable	text and/or ontology entry (multiple possible)	EFO
	Extrinsic variable	External biosample treatment (e.g. reagent) if applicable	text and/or ontology entry (multiple possible) or associated file	EFO, IDR
	Experimental variables	What is intentionally varied (e.g. time) between multiple entries in this study component	text and/or ontology entry (multiple possible)	EFO
Specimen	Experimental status	Test/ control		
	Location within Biosample	Plate/dish coordinate or tissue location	text or associated file	OME
	Preparation method	Sample preparation protocol	text, file, ontology, or widget for specific method types	EDAM-BIOIMAGING, FBbi
	Signal/contrast mechanism	How is the signal generated by this sample	text, ontology	EDAM-BIOIMAGING, FBbi
	Channel - content	Specific specimen staining (e.g. IEM, DAB)	text	
	Channel - biological entity	What molecule is stained	text, ontology entries	EFO
Image acquisition	Instrument attributes	Details about instruments used	text, file, ontology, or widget for specific instrument types	EDAM-BIOIMAGING, FBbi, OME, 4DN-BINA-OME
	Image acquisition parameters	Image acquisition details	text, file, ontology, or widget for specific acquisition method types	EDAM-BIOIMAGING, OME, 4DN-BINA-OME
Image data	Type	Primary image/processed image/segmentation	pull-down	EDAM-BIOIMAGING
	Format & compression	File type	extract from data if possible	EDAM-BIOIMAGING, OME
	Dimension extents	Volume in pixels: x, y, z, tilts	extract from data if possible	OME
	Size description	Physical size of image volume in x,y,z & units (pull-down), OR magnification	extract from data if possible	OME
	Pixel/voxel size description	Physical size of pixels in x, y, z & units (pull-down)	extract from data if possible	OME
	Channel information	How are individual channels represented in the image	extract from data if possible	OME
	Image processing method	Image registration, other processing applied to this dataset	text, file, ontology, or widget for specific method types	EDAM-BIOIMAGING, FBbi
	Contrast inversion to TEM	Y/N; N if stained features result in brighter (whiter) signal; Y if it looks like a TEM image	pull-down	
Image Correlation	QC info	QC score for uploaded image quality if applicable	text or controlled vocabulary	
	Spatial and temporal alignment	Method used to correlate images from different modalities (e.g. manual overlay, alignment algorithm etc)	text, ontology	EDAM-BIOIMAGING
	Fiducials used	Features from correlated datasets used for colocalization	text	
	Transformation matrix/ other info	Correlation transformations	text, or related project files (e.g. .h5 Amira files)	
Analyzed data	Related images and relationship	Correlated dataset or images	link	
	Analysis result type	Numerical analyses, segmentation (non-image), categorical features/phenotypes	text, ontology	EDAM-BIOIMAGING, OME
	Data used for analysis	Specific feature set used for analysis (e.g. volume measurements, locations of features)	text or file(s)	
	Analysis method and details	Analysis method	text, file, ontology, or pointer to Methods section	EDAM-BIOIMAGING

Mapping and gap analysis of the metadata models

16 items are shared among the three models

Module	Attribute	Details	BIA	IDR	SSBD	Standards
Study	Identity	unique ID	uuid, accession_id	Id	SiD	
	Study type	Study type		Study Type	type	
	Study description	Study description	description	Experiment/Screen description	description	
	General dataset info	authors	author:	Publication Authors	person:	
		publication	title, doi, pubmed_id	Publication Title/DOI	paper info, DOI, PubMed ID	DOI
		license	license	License	license	Creative Commons
		release date	release_date	Release date	date of release	date
Study Component	Imaging Method	Imaging Method	fbbi_method_type	Imaging Method	imaging method:	FBbi
	Study component description	Study component description	annotation_component			
Biosample	Identity	internal unique ID	uuid		(strain)	
	Biological entity	biological entity	biological_entity			
		Cell Line		cell_lines	cell line	CLO
		Cell Type			cell	CO
		Pathology/Disease		Pathology		SNO MED-CT
	Organism	organism	organism_classification	organism	organism	NCBITaxon
	Intrinsic variable	Intrinsic variable	intrinsic_variables		intrinsic variable	
		Gene		Gene	gene	Ensembl ID/NCBI Gene
	Extrinsic variable	Extrinsic variable	extrinsic_variables		extrinsic variable	
		Compound		Compound	reagent/compound	PubChem/ChEBI
		Antibody		Antibody	reagent/compound	ChEBI, FBbi
		siRNA		siRNA	oligoprimer	
Specimen	Experimental variables	experimental variables	experimental_variables		experimental variables	experimental variables
	Experimental status	Experimental status				
	Location within Biosample	Location within Biosample				
	Preparation method	specimen_preparation_method	{method, growth_control} description		preparation method	FBbi (obo)
	Signal/contrast mechanism	Signal/contract mechanism	signal contrast mechanism description			
	Channel - content	Channel - content	channel_content_description	ChannelInfo:	tag	
	Channel - biological entity	channel - biological entity	channel biological entity	ChannelInfo:?	(protein)	
Image acquisition	Instrument attributes	Instrument	imaging_instrument_description	Instrument:	instruments:	
	Image acquisition parameters	acquisition method	acquisition_method			
Image data	Identity	unique ID	uuid	id	ID	
	Type	Type				
	Format & compression	format	image_format		file format	
	Dimension extents	dimension	size {x, y, z, c, t}	Size {X,Y,Z,C,T}	dimensions	
	Size description	data size	total size in bytes		file size	
	Pixel/voxel size description	pixel/voxel size	physical_size {x, y, z}	PixelSize {X,Y,Z}, TimeIncrement	{X, Y, Z, T} scale	
	Channel information	channel	channel_content_description, channel biological entity	ChannelInfo:	tag, protein	
	Image processing method	Image processing method	annotation_method			
	Contrast inversion to TEM	Contrast inversion to TEM				
	QC info	QC info				
Image Correlation	Spatial and temporal alignment	Spatial and temporal alignment				
	Fiducials used	Fiducials used	fiducials used			
	Transformation matrix/other info	Transformation matrix/other info	transformation_matrix			
	Related images and relationship	Related images and relationship				
Analysed data	Analysis result type	type Phenotype	method_type	Phenotype	type	CMPO
	Data used for analysis	Data used for analysis	annotation_file:		Source	
	Analysis method and details	Analysis method and details	annotation_method:		Workflow	

Controlled vocabularies used in the models

- In all models, the *imaging method* and *organism* were described in the controlled vocabulary.
- Several topics (*compound* and *gene*) were described in different controlled vocabularies.

Topic	Controlled vocabulary/ Ontology	BIA	IDR	SSBD
Biological process	Gene Ontology (Biological Process)	N	N	Y
Cellular component	Gene Ontology (Cellular Component)	N	N	Y
Molecular function, Molecular activity	Gene Ontology (Molecular Function)	N	N	Y
Cell	Cell Ontology	N	N	Y
Cell line	Cell Line Ontology	N	N	Y
Anatomy	Uberon multi-species anatomy ontology	N	N	Y
Imaging method	Biological Imaging Methods Ontology	Y	Y	Y
Disease	Human Disease Ontology	N	N	N
Disease	Mondo Disease Ontology	N	N	N
Disease, Pathology	SNOMED CT	N	Y	N
Chemical compound	Chemical Entities of Biological Interest Ontology	N	N	Y
	PubChem	N	Y	N
Taxonomic classification, Organisms	NCBI organismal classification (NCBI Taxonomy)	Y	Y	Y
Experimental conditions	Ontology for Biomedical Investigations	N	N	Y
Experimental conditions	Experimental Factor Ontology	N	Y	Y
Phenotype	Human Phenotype Ontology	N	N	N
Phenotype	Mammalian Phenotype Ontology	N	N	N
Unit	Units of measurement ontology	N	N	Y
Cellular phenotype	Cellular microscopy phenotype ontology	N	N	N
Gene	Ensembl gene	N	Y	Y
	NCBI Gene	N	Y	N
Protein	UniProt	N	Y	Y
Controlled vocabulary	Medical Subject Headings (MeSH)	N	N	Y

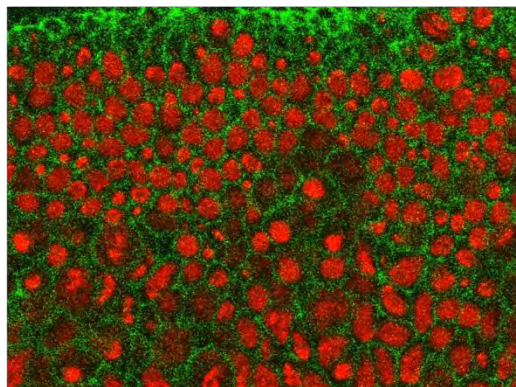
Use cases of bioimaging metadata

- Bioimaging data can be used for
 - Generation of new hypotheses by data observation and comparison
 - Development and validation of new image analysis methods and tools
 - Exploration and evaluation of chemical compounds' functions
 - Comparative phenotype analysis with preclinical and clinical data
 - Integrated analysis (Meta-analysis) of bioimaging data
 - Multi-modal analysis with transcriptome, proteome, and the other data

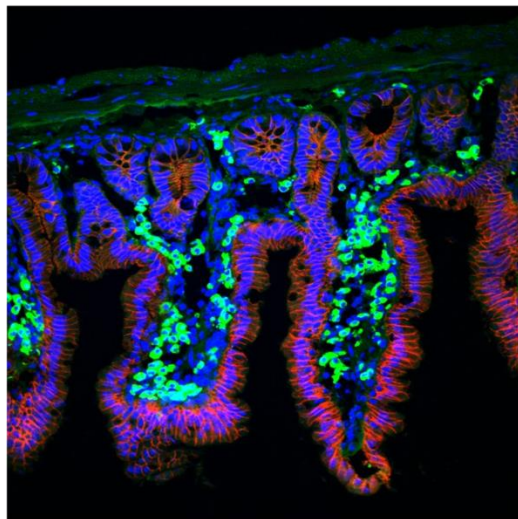
Example) Genetic research and disease study/drug discovery

- Genetic research
 - Expression patterns for the gene of interest across different species
 - Cellular phenotypes for genes of interest
- Disease study/Drug discovery
 - Search for data related to specific cellular phenotype and disease
 - Effect of compounds

Cadherin localization

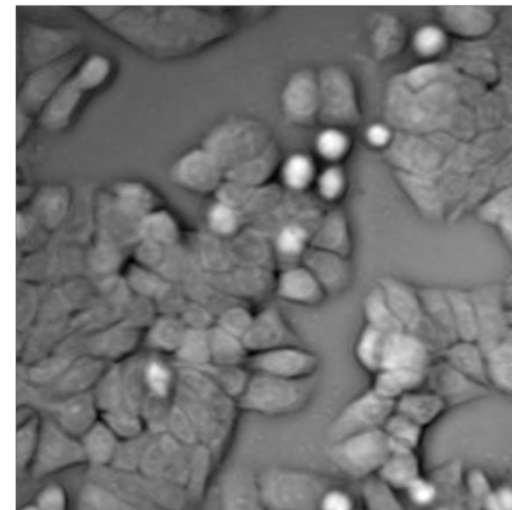


D. melanogaster
tracheal invagination
(Kondo and Hayashi (2013))



M. musculus
intestine
(Muta et al. (2018))

Drug treatment



H. sapiens
nocodazole
(Ito et al. (2017))

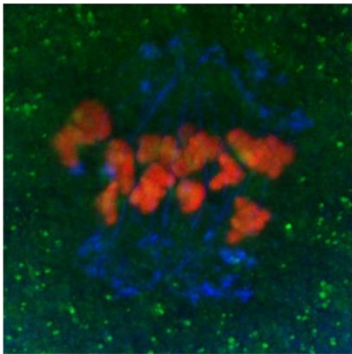
Required metadata: **Gene, Protein, Organism, Phenotype**

Required metadata: **Compound, Phenotype**

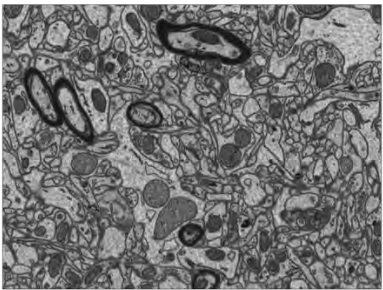
Example) Tool development and validation

- Development and validation of image segmentation, tracking, and error correlation tool

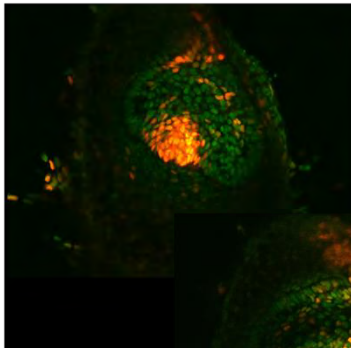
Imaging modality



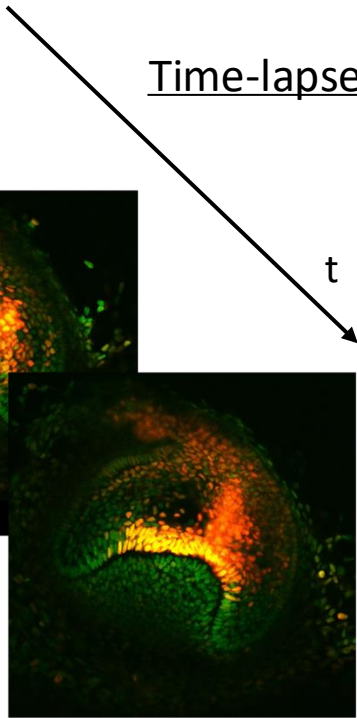
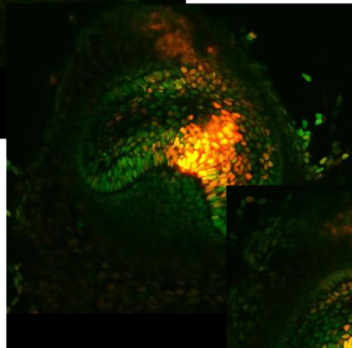
fluorescence microscopy
chromosome, microtubule
(Yoshida et al. (2015))



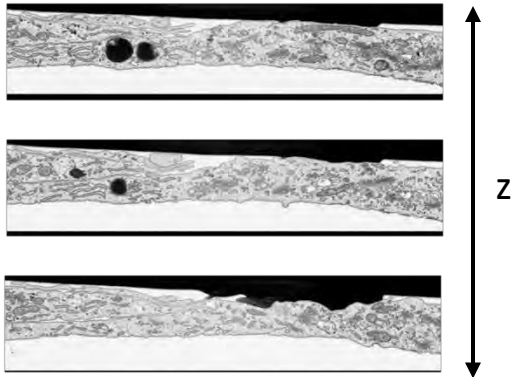
FIB-SEM
synapse morphology
(Sato et al. (2019))



862x855x43x2x271
(z: 43, c: 2, t: 271)
(Morita et al. (2016))



Time-lapse and Z-stack images



3438x760x121x1x1
(z: 121, c: 1, t: 1)
(Miyamoto et al. (2020))

Required metadata: **Imaging method, Instrument, Dimension, Pixel/Voxel size**

Ontology Assessment for harmonization

- **Coverage:**
 - How well does the ontology cover the domain?
 - ✓ PubChem: Extensive coverage with >320 million compounds
 - ✓ Experimental Factor Ontology (EFO) : used for experimental conditions in IDR and SSBD. OBI can complement terms.
- **User:**
 - How widely is the ontology used?
 - ✓ NCBI Taxon and GO: widely used across species.
 - ✓ FBbi : Used for imaging methods across BIA, IDR and SSBD.
- **Maintenance:**
 - Is the ontology actively maintained and updated?
 - FBbi:** Last updated three years ago, which may limit its ability to incorporate the latest imaging technologies in the future.
- **Integration:**
 - How well can the ontology be integrated with other data systems?
 - ✓ OBO Foundry ontologies (GO, Cell Ontology, ChEBI, MONDO/Human Disease Ontology, UBERON) prioritize interoperability
- **Open Access:**
 - Is the ontology available for open use?
 - SNOMEDCT: Not openly accessible in Japan; requires mapping to MONDO for disease terms
- **Consistency and Uniformity:**
 - Does the ontology provide consistent and uniform terms and definitions?
 - ✓ OWL DL: Built on Description Logic to ensure logical consistency

Future Directions from the Ontology Point of view

- Interoperability for Preclinical Data
 - Interoperability with DICOM and SNOMED CT will become essential, though challenges exist:
 - DICOM: Primarily targets clinical images and is centered on human patients.
 - SNOMED CT: Not freely accessible in Japan, limiting its applicability.
- Harmonizing Pathology and Disease Data
 - To ensure compatibility with SNOMED CT terms used in IDR, mapping to ICD-11, MONDO, and DOID is recommended.

Proposal of the requirements for metadata harmonization

- In addition to the 16 components, 4 additional components were identified (two accession IDs and two components for connecting with preclinical data).

Proposal of formats and query for data access for 20 metadata components

Component	Description	Format	Query for data access
Study Description	Description of the study	Text	Free text
Authors	Information about the authors	First name and Last name; ORCID	Author name, ORCID
Publication	Information on the related publication	Text (Title), Publication year, DOI	DOI
License	Information about the dataset's license	Creative Commons with the version and URL	CC version, URL
Release Date	The release date of the data	YYYY-MM-DD format	Release date, range search
Imaging Method	The method used to obtain image data	FBbi term and ID	FBbi ID, term
Cell Line	Information about the cell line	Cell Line Ontology term and ID	CLO ID, term
Organism	Information about the organism studied	NCBI Taxonomy term and ID	NCBI Taxonomy ID, term
Gene	Information about related genes	Ensembl Gene or NCBI Gene name or ID	Gene ID, name
Compound	Information about the compound	ChEBI or PubChem term and ID	Compound ID, term
Antibody	Information about the antibody used	FBbi or ChEBI term and ID	Antibody ID, term
Channel – Content	Content related to the channel	FBbi term and ID	Channel content ID, term
Channel – Biological Entity	Biological entities related to the channel	Experimental Factor Ontology term and UniProt ID	Biological entity ID, term
Instrument	Information about the instrument used	Compliant with OME or 4DN-BINA-OME standards	Instrument name, ID
Dimension	Information about data dimensions	Stored in X, Y, Z, T, C format	Dimension specification
Pixel/Voxel Size	Size of the pixels or voxels	Stored with units; recommended for microscopy data	Pixel/voxel size
Study Unique ID	Unique ID for the study	Accession ID, DOI	Accession ID, DOI
Dataset Unique ID	Unique ID for the dataset	Accession ID	Accession ID
Pathology/Disease (Biological Entity)	Pathology/Disease related to the biological entity	SNOMED-CT, ICD-11 or MONDO	Pathology/Disease ID, term
Phenotype (Analysis Data)	Phenotypic data related to the analysis	Cell Morphology Phenotype Ontology, MPO, or HPO	Phenotype ID, term

Metadata annotation with Large Language Models (LLMs)

- Current landscape of metadata annotation
 - Metadata is traditionally annotated by human curators.
 - Recent advances in LLMs have made automatic metadata annotation a realistic alternative.
- Flagging metadata as LLM or Curator-generated
 - Metadata origin supports scalable annotation by leveraging both LLMs and curators.
- Metadata extraction from free-text descriptions
 - Free-text description allows for future LLM-driven metadata enhancements as these models evolve.

Towards easier data access

- It is desirable to assign a unique and standardized accession ID.
 - **Study**: DOI of the datasets published in journal paper
 - **Image Data**: an identifier provided by GIDE consortium
- How to achieve global sharing of bioimage data:
 - Development of a **universal identifier** for bioimaging study and data
 - Establishment of a **(metadata-based) data portal** for bioimaging study and data
- Versioning of metadata is necessary

Draft report

FoundingGIDE WP6

Mapping and gap analysis of the metadata components

The goal is to define the requirements for harmonizing the metadata of the BIA, IDR, and SSBD metadata through a comparative analysis of the metadata components that comprise these models.

Each metadata model consists of multiple classes, such as Study and Biosample, with each class containing various metadata components. In this comparative analysis of metadata components, we used Modules and Attributes from REMBI as a foundation to establish correspondences between components. As a result, we identified 16 items that should describe the same metadata in the three models (Appendix 1), although there are differences in class names and component names.

Next, we explored the requirements (in terms of format and access methods) needed to harmonize these metadata based on the above table and use cases of metadata. In addition to the 16 components mentioned above, we defined requirements for a total of 20 components, which include four additional components that consider essential IDs and connections to preclinical data for data sharing in public repositories.

Controlled vocabularies and ontologies are essential for data consistency and reusability in metadata harmonization. A comparative analysis of the ontologies used across the BIA, IDR, and SSBD metadata models demonstrates that interoperability is significantly enhanced through shared ontologies. Across the metadata models for BIA, IDR, and SSBD, the Biological Imaging Methods Ontology (FBbi) is commonly used for imaging methods, and NCBI taxonomy provides standardization for species classification. For experimental conditions, the Experimental Factor Ontology (EFO) is shared between IDR and SSBD, with SSBD further utilizing the Ontology for Biomedical Investigations (OBI) as a complementary resource. Furthermore, SSBD uses the Cell Ontology (CL) for cell types, the Cell Line Ontology (CLO) for cell lines, Uberon for anatomical terms, and the Gene Ontology for biological processes. These ontologies enhance data interoperability, allowing seamless integration and analysis across diverse datasets and species.

For compound data, IDR employs PubChem, one of the largest compound databases globally, encompassing over 320 million substances and more than 100 million structures, offering extensive data coverage. In contrast, SSBD utilizes ChEBI, which covers approximately 60,000 compounds and effectively systematizes compound knowledge through a hierarchical structure. As a member of the Open Biological and Biomedical Ontology Foundry (OBO Foundry), ChEBI will facilitate high interoperability with other ontologies, such as GO, CL, and CLO, supporting cross-ontology integration.

Share your metadata model!

Acknowledgement

- BioImage Archive
 - Matthew Hartley (EMBL-EBI)
 - Francois Sherwood (EMBL-EBI)
- IDR
 - Josh Moore (German BioImaging)
 - Khaled Mohamed (University of Dundee)
- FoundingGIDE
 - Members of Work Package 6
 - Dario Longo (Italian National Research Council)
- RIKEN
 - Members of SSBD team



Koji Kyoda



Yuki Yamagata