Deliverable D2.1 Metadata Standardization Strategy and Database

Project	**CLARIFY – Cloud ARtificial Intelligence For pathologY**
Grant Agreement ID:	**860627**
Consortium coordinator:	**UNIVERSITAT POLITECNICA DE VALENCIA**
Start and end date:	**1 November 2019 - 31 October 2023**
Funded under:	**H2020-EU.1.3.1.**
Date of issue:	**30-Jun-2021**
Due date:	**28-Feb-2021**
Leader in charge of deliverable:	**University van Amsterdam**

Dissemination level	
X	PU = Public
	PP = Restricted to other programme participants (including the EC)
	RE = Restricted to a group specified by the consortium (including the EC)
	CO = Confidential, only for members of the consortium (including the EC)

European Commission

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 860627.

CHANGE REGISTER

Version	Date	Author	Organisation	Changes
A_DRAFT	5-Jan-2021	Na Li	UvA	Initialisation
	24-May-2021	Na Li	UvA	First draft. Contributions from all WP2 partners. Circulated for comments.
	25-May-2021	Saul Fuster	UiS	Comments to the previous version
	07-Jun-2021	Na Li	UvA	Version to review
	09-Jun-2021	Carlos Monteagudo	INCLIVA	Reviewed version
	10-Jun-2021	Kjersti Engan	UiS	Reviewed version
	11-Jun-2021	Sandra Morales	UPV	Reviewed version
	22-Jun-2021	Na Li	UvA	Reviews addressed. Final version.
A	30-Jun-2021	Valery Naranjo	UPV	Formatting and typos only

Statement of independence

The work described in this document is genuinely a result of efforts pertaining to the CLARIFY project: any external source is properly referenced.

Confirmation by Authors: Na Li, University of Amsterdam, NL
Zhiming Zhao, University of Amsterdam, NL

Abbreviations

HR-NMIBC	High-Risk Non-Muscle Invasive Bladder Cancer
SML	Spitzoid Melanocytic Lesions
TNBC	Triple Negative Breast Cancer
WSI	Whole Slide Image
WP	Work Package
AI	Artificial Intelligence
SUH	Helse Stavanger HF
INCLIVA	Fundación para La Investigación del Hospital Clínico de la Comunitat Valenciana
EMC	Erasmus Medisch Centrum Rotterdam
UiS	University of Stavanger
UPV	Universitat Politècnica de València
UGR	Universidad de Granada
TY	Tyris Software S.L.
RDA	Research Data Alliance
EHR	Electronic health record
ISO	International Organization for Standardization

DICOM	Digital Imaging and Communications in Medicine
IHE	Integrating the Healthcare Enterprise
HL7	Health Level Seven International
FHIR	Fast Healthcare Interoperability Resources
CDA	Clinical Document Architecture
PID	Persistent Identifier
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive

Table of Contents

1 Executive summary

This document aims to present the metadata standardization strategy based on the metadata requirements of the CLARIFY project and state-of-art metadata standards. More specifically, this document analyzes the possible research assets within the data flow and collects the metadata requirements for each type of research asset. Recommendations and a plan are given on the top of gap analysis between the requirements and the state of arts. Examples of annotated databases are also provided.

Deliverable 2.1 is under the Task T2.1 Metadata standarization, within the DoA of the CLARIFY project.

2 Introduction

This section will introduce the objectives and main actions of T2.1, answering how it is in line with the goal determined by WP2 and supports the overall goal of the CLARIFY project.

Metadata is necessary for data sharing and data reuse. Besides, it is the core element in FAIR principles[1]. The efficiency in data discovery largely depends on the quality of metadata. One big challenge involved in data sharing is the lack or the diversity of metadata standards, which poses great difficulty in data integration and data reuse. These issues can be addressed by adopting a unified metadata standard when generating data and metadata. In the context of digital pathology, annotation generated by pathologists is the crucial part of metadata.

The actions in Task 2.1 (T2.1) within Work Package2 (WP2) include:

1. Collection of a substantial reference data set of WSIs for a subset of cancer types and subsequent annotation with clinically relevant multi-disciplinary information.
2. Overview of which metadata types are necessary to be included as part of the reference dataset and review the available ontologies and common phenotype descriptors used for the cancer types addressed by CLARIFY.
3. Finally, a metadata standardization process will be performed.

Deliverable2.1 is the main output of T2.1, which will:

1. describe the requirements for metadata, metadata standardization strategies, annotated databases;
2. discuss the state of the arts;
3. identify the gaps between the available resources and the requirements;
4. propose the recommendations and work plans for next stages.

[1] https://www.go-fair.org/fair-principles/

3 Requirements and Current Status

3.1 Information Collection Methodology

The Research Data Alliance (RDA) has proposed five principles regarding metadata[2]:

1. The only difference between metadata and data is mode of use;
2. Metadata is not just for data, it is also for users, software services, computing resources;
3. Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions (rights, costs)) and for coupling users, software and computing resources to data (to provide a Virtual Research Environment)
4. Metadata must be machine-understandable as well as human understandable for autonomicity (formalism)
5. Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…)

According to these principles, it is of great importance to investigate the requirements for metadata from users before developing a metadata standard recommendation.

The requirements will be collected by following steps:

1. *Analyze CLARIFY assets in the workflow via use cases and questionnaire.* By considering possible assets produced within the CLARIFY project, the requirements for metadata and metadata standards are analyzed towards each type of asset. Examining the usage scenarios of data and metadata by partners in the CLARIFY project provides the users' point of views on metadata requirements. Questionnaire approach is leveraged to understand aforementioned workflow and usage scenarios.

2. *Collect information about requirements and current status* via online discussion and expert interview. More detailed and specific requirements are directly acquired from domain experts by the means of interviews and online discussions.

3. *Analyze detailed requirements* to derive the minimal requirements for all involved research assets in the context of CLARIFY use case scenarios.

In the rest of this chapter, we will explain these steps and the output in more details.

3.2 CLARIFY Asset Analysis

It is of great importance to know the research assets (images, clinical data, models, etc.) in the CLARIFY project prior to understanding metadata requirements. A questionnaire[3] that collects information about the management and usage of research assets is devised and circulated among WP2 partners of the CLARIFY project. It contains 3 sections and 57 questions. The first section asks for basic information of the participants. The second section is specifically for

[2] https://www.rd-alliance.org/metadata-principles
[3] https://forms.gle/i7GeomJszfJTUJEEA

asset producers on the asset generating process, storage technologies, etc. And the last section is for asset users on their targeting tasks, asset access methods, metadata requirements, etc. 9 responses of the questionnaire are received from the following institutions:

- University van Amsterdam;
- Stavanger University Hospital (SUH);
- Tyris;
- INCLIVA;
- University of Stavanger (UiS);
- Universitat Politècnica de València (UPV);
- Erasmus Medical Center (EMC).

Based on the responses, we determined four research assets produced by the CLARIFY project, which are: (1) Whole slide images (WSIs); (2) WSI annotations; (3) Clinical data; (4) AI models. The analysis on the responses also helps us understand, from the perspective of data, the different roles played by partners of the CLARIFY project, as well as the processes involved. As a result, a dataflow that describes the generation and usage of research assets is used to illustrate how metadata is intertwined with research assets (see Figure 3.1).

Within the CLARIFY project, WSI images and clinical data are the most essential data which are primarily provided by medical partners, namely SUH, INCLIVA and EMC. Figure 3.1 shows the generic dataflow of whole slide images and clinical data in WSI-based research. The upper part shows the clinical dataflow and the bottom part shows the WSI dataflow. In the clinical dataflow, doctors register patient information into clinical records. Clinical data is then preprocessed and utilized by AI experts to build models. The WSI dataflow starts from whole slide image scanning, followed by WSI annotation, and then WSI preprocessing. AI experts exploit WSIs (with or without annotation) together with clinical data to train and test AI models. Note that the metadata describing the process of glass slide preparation, such as dehydration, fixation, slicing, staining, etc., is treated as internal information within the pathology department, and thus omitted from later processes. The workflow here is not always entirely implemented in practice, with some steps being skipped in some cases. For instance, AI experts also utilize images that are not annotated and maybe not preprocessed, since everything is not annotated completely.

We identify 2 important roles in the dataflow: (1) Medical experts (SUH, INCLIVA, EMC); (2) AI experts (UiS, UPV, UGR, TY). Medical experts are usually from hospitals and medical centers that possess the first-hand information of the patients. They prepare WSI and clinical data for downstream users and annotate WSI as well. AI experts mainly focus on developing AI models by leveraging the data to finish medically related tasks, such as detection, classification, segmentation, content based retrieval, etc. To simplify the process, we did not take the software/service providers into consideration so far, such as annotation tool provider, data manager, metadata standardization developer, etc.

With the dataflow, it is easy to distinguish between providers and users with respect to each type of asset. WSIs, annotations over WSIs and clinical data are provided by medical experts and used by AI experts. AI models are provided by the AI experts (Uis, UPV, TY and UGR in the context of CLARIFY project).

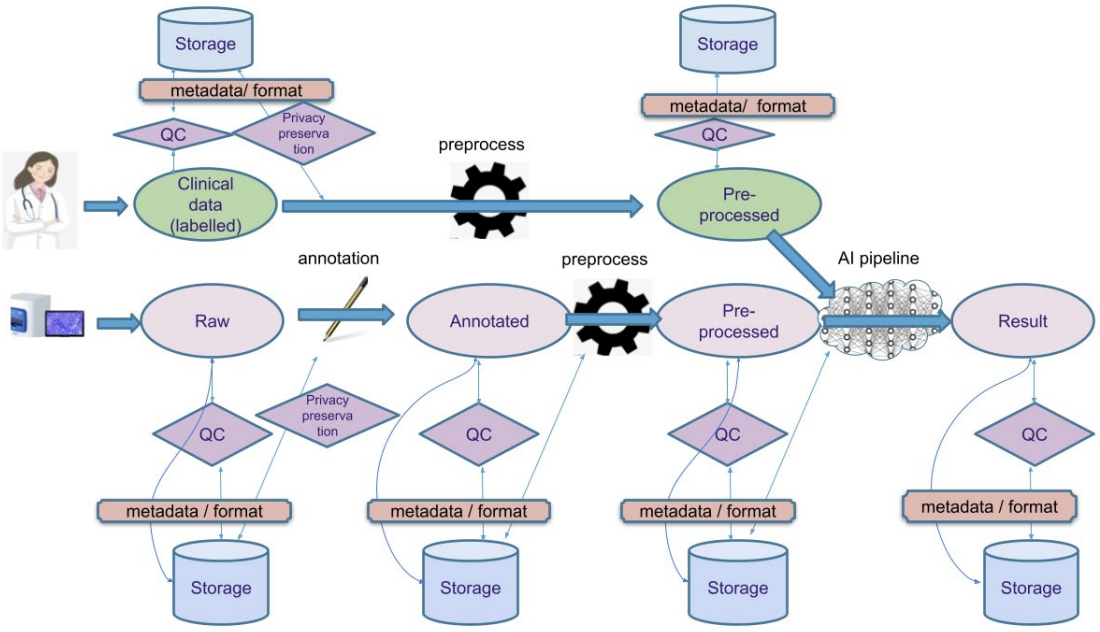

Figure 3.1 Generic dataflows of whole slide images and clinical data in WSI-based research. Here "QC" means quality control.

Each step involves both data storage and data retrieval (except for the whole slide image scanning step). Metadata plays an important role in accuracy and efficiency of data retrieval. It is worthy to emphasize that the clinical data is data for its own sake, but also metadata for the corresponding WSIs, providing necessary information for AI experts to analyze the images.

3.3 Requirement Collection

To understand the metadata requirements for different research assets, a survey is conducted among CLARIFY partners by the means of online discussions and expert interviews. Online discussions have been held every two weeks ever since February 25th 2021 among the Work Package 2 (WP2) working group members. The discussions are aimed at updating work progresses in WP2 and reaching a common understanding for the work undertaken by different partners. During the meetings, the comments towards the metadata requirements are collected and summarized. Expert interview is the final step to acquire specific metadata requirements from the CLARIFY partners.

Metadata requirements vary for different cancer types because pathologists utilize different factors and metrics when diagnosing different types of cancer. More specifically, for Spitzoid Melanocytic Lesions (SML), they aim to distinguish between different types of lesions. For Triple Negative Breast Cancer (TNBC), they distinguish between different histopathological subtypes such as infiltrating duct carcinoma, lobular carcinoma, mixed ductal-lobular carcinoma, etc. And for High-Risk Non-Muscle Invasive Bladder Cancer (HR-NMIBC), they aim to classify cancer grade and stage, and to predict risk of recurrence and progression.

The rest of this section follows the "Field-Type-Value" template to organize the metadata requirements with respect to different cancer types, with an additional "explanation" column when necessary. "Field" refers to the name of the information. "Type" denotes the valid type

of content for each “Field”. “Value” suggests the acceptable contents when the “Type” is specified as controlled vocabulary. When the “Type” is not specified as controlled vocabulary, it means the data type for the “Value”, such as string, number, date, boolean, etc.

3.3.1 Metadata requirements for WSIs, annotations and clinical data

The requirements for metadata of WSIs are collected from AI experts, who are the end users of WSIs and the producers of the AI models. Table 3.1 lists all the requirements for metadata. Since the contents in WSI annotations and the clinical data are metadata of WSI from the viewpoint of AI experts, all they need are identifiers to the corresponding annotations and clinical data.

Table 3.1 Requirements for metadata of WSIs, annotations and clinical data.

Asset type	Field	Type	Value	Description
WSI	name	string		Name of the WSI, usually anonymised
	identifier	string		Identifier of the WSI
	data format	string		File format of the WSI
	resolution	number		Resolution of the WSI
	identifier to clinical record	string		Identifier connecting to the corresponding clinical record
	identifier to WSI annotation	string		Identifier connecting to the corresponding annotations
	size	number		Storage size of each WSI
	magnification levels	string		Magnification levels that the WSI contains
WSI annotation	Identifier to the WSI	string		The identifier connecting to the WSI
Clinical data	Identifier to the WSI	string		The identifier connecting to the WSI

3.3.2 Metadata for WSI annotations

Annotations over whole slide images provided by pathologists usually serve as the ground truths in AI model training and testing. In the CLARIFY project, all the annotations over the WSIs w.r.t a specific cancer type follow a specific annotation protocol established by the relevant partners. Table 3.2 aggregates the required annotating information for SML, TNBC, HR-NMIBC, respectively. According to the AI experts, the specified annotations cover all the essential information needed to develop the AI models.

Table 3.2 Metadata requirements for WSI annotations

Cancer type	Field	Type	Value
SML (Spitzoid Melanocytic Lesions)	Global regions	controlled vocabulary	1. Spitzoid nevus (benign) 2. Spitzoid tumor of uncertain malignant potential (STUMP) 3. Spitzoid melanoma (malignant)
	Small patterns	region and controlled vocabulary	Typical mitosis Atypical mitosis Ulcers Necrosis
TNBC (Triple Negative Breast Cancer)	Histopathological subtype labels	controlled vocabulary	1. infiltrating duct carcinoma 2. lobular carcinoma 3. mixed ductal-lobular carcinoma 4. medullary carcinoma 5. metaplastic carcinoma 6. apocrine adenocarcinoma 7. adenoid cystic carcinoma
	Interesting factor labels	controlled vocabulary	1. Tumor infiltrating lymphocytes (TILs) 2. Typical mitosis 3. Atypical mitosis 4. Fibrotic focus 5. Necrosis 6. Adipocytes
HR-NMIBC (High-Risk Non-Muscle Invasive Bladder Cancer)	Tissue types	controlled vocabulary	1. Urothelium tissue: (1) without immune cell infiltration (2) with immune cell infiltration (3) normal (4) cancerous (5) variant histology (6) flat lesion 2. Stroma tissue: (1) without immune cell infiltration (2) with immune cell infiltration 3. Muscle 4. Blood 5. Damaged tissue (1) Cauterized (2) Blurry (3) Folded 6. Cancerous invasive areas 7.Mitosis

	Grading labels	controlled vocabulary	1. WHO04/16 (1) High grade (2) Low grade 2. WHO73 (1) Grade 1 (2) Grade 2 (3) Grade 3 3. Flat lesions (1) Dysplasia (2) CIS
	Staging labels	controlled vocabulary	1. pTa 2. pT1 3. pTis

3.3.3 Metadata for clinical data

Clinical records register patient's information related to the pathological images. They provide crucial information for diagnosis and prognosis in clinics and are also important resources alongside whole slide images for AI algorithm development. INCLIVA provides a table of variables used in clinical data for SML (see Table 3.3). SUH provides a table of variables used in clinical data for TNBC (see Table 3.4) and HR-NMIBC (see Table 3.5).

Table 3.3 Required variables used in clinical data for SML

Field	Type	Value	Explanation
Age (years)	controlled vocabulary	1. 0-11 2. 12-18 3. 19-29 4. 30-49 5. >50	Age at the time of diagnosis
Gender	controlled vocabulary	1. Female 2. Male	Type of gender
Tumor location	controlled vocabulary	1. Head and neck 2. Trunk 3. Upper limb and shoulder 4. Lower limb and hip 5. Not recorded or specified	Location of the primary tumor
Date of Diagnosis	date (year/month)		Year and month of the histopathological diagnosis
Local recurrence	boolean	1. No 2. Yes	New tumor growth at the primary site after the surgical excision
Date of local recurrence	date (year/month)		Year/month of recurrence diagnosis

Regional relapse	controlled vocabulary	1. No 2. Satellite/in transit metastasis 3. Lymph node metastasis 4. Lymph node AND satellite/in transit metastasis	
Date of regional relapse	date (year/month)		Year/month of regional relapse diagnosis
Distant metastasis	controlled vocabulary	1. No 2. Distant metastasis to skin, soft tissue and/or non-regional lymph node 3. Distant metastasis to lung 4. Distant metastasis to non-Central Nervous System visceral sites 5. Distant metastasis to Central Nervous System	Tumor dissemination to distant sites
Date of distant metastasis	date (year/month)		Year/month when the distant metastasis was detected
Follow up	controlled vocabulary	1: Alive and well 2: Alive with local recurrence 3: Alive with regional relapse 4. Alive with distant metastasis 5: Dead of Disease	
Last date of follow up	date (year/month)		Last control

Table 3.4 Required variables used in clinical data for TNBC

Field	Type	Value	Explanation
SUS-number	string		ID_number
eofus	controlled vocabulary	1: Alive and well 2: Alive with Distant Metastasis 3: Alive with local recurrences 4: Dead of other causes 5: Dead of Disease with local recurrences 6: Dead of Disease with Distant Metastasis 7: Lost to follow-up	End Of Follow Up Status
FUT	date (days)	Date last control – date breast cancer diagnosis in days /30.5	Follow Up Time in Months
cur_rela	controlled vocabulary	1: no metastasis 2:locoregional recurrences 3:distant metastasis 4: Lost from follow-up 9:Unknown	Current relapse status, at date of last control
metalife	controlled vocabulary	1:No meta 2:Meta 9:Unknown or LFTU	Any Meta or not in a lifetime, including locoregional recurrences
meta1e	controlled vocabulary	1: local regional 2: other breast 3: bone 4: liver or other organ 5: brain 6: multiple 8: not applicable 9: unknown	First metastasis
MAI	number		Mitotic Activity Index (mitosis/1.59 mm2)
PPH3	number		PPH3 evaluation (positive cells/1.59mm2)
Ki67	number		Ki67 percentage hotspot/coldspot in 500 tumor cells
MAI_10	controlled vocabulary	1: MAI<10 2: MAI>9 9: unknown	MAI 9 or lower vs MAI 10 and higher
PPH3_13	controlled vocabulary	0: PPH#<13 1: PPH3>=13 9: unknown	H3 with threshold 13

Age	number (days/365)	Time of first diagnosis – birthday in days /365	Age at time of diagnosis
origrade	controlled vocabulary	1: grade 1 2: grade 2 3: grade 3	Original grade
tub_form	controlled vocabulary	1: >75% 2: 10-75% 3: >10% 9: unknown	Tubular formation
nuc_atyp	controlled vocabulary	1: mild 2: moderate 3: marked 9: unknown	Nuclear Atypia
mit_imp	controlled vocabulary	1: 0-5 2: 5-10 3: >10 9: unknown	Mitotic Impression
sumgrad	number		Sum of tub, Nuc At, Mit Imp.
Nottgrade	controlled vocabulary	1: 3-5 2: 6-7 3: 8-9 99: unknown	Nottingham grade
tum_size	number (cm)	From the original pathology report	Tumor size [cm]
tsize2	controlled vocabulary	1: Tsize<= 2.0 cm 2: Tsize>2.0 cm 9: unknown	
oestr_re	controlled vocabulary	1: positive 2: dubious (1-10%) 3: negative 9: unknown	Estrogen Receptor based upon IHC >1%
prog_rec	controlled vocabulary	1: positive 2: dubious (1-10%) 3: negative 9: unknown	Progesteron Receptor based upon IHC >10%
htype	controlled vocabulary	5: Ductal 7: other	Histologic type Tumor
l_status	controlled vocabulary	1: positive 2: negative 3: no lymph nodes found 9: unknown/not performed	Lymph Node Status
n_nodes	number		Number of nodes

np_nodes	number		Number positive nodes
HER2	controlled vocabulary	0: negative =score 0 1: negative = score 1 2: positive = score 2 3: positive = score 3	HER2 based upon IHC, 10% tumor cells
FUT_recur	number		Time to first recurrence
LeucoInfiltration	boolean	1: No 2: Yes	Leukocyte infiltration= TILs
Siteleuco	controlled vocabulary	1: infiltrating 2: tumor border 3: spot like 4: everywhere 5: focally without a typical pattern	Where are the leucocytes located
FF	controlled vocabulary	0: absent 1: present	Fibrotic focus
FFsize	controlled vocabulary	1: < 1/3 of the tumor 2: > 1/3 of the tumor	Fibrotic focus size
Necrosis	controlled vocabulary	0: absent 1: present	Necrosis in fibrotic focus

Table 3.5 Required variables used in clinical data for HR-NMIBC

Field	Type	Value	Explanation
WHO 73	controlled vocabulary	1= grade 1 2= grade 2 3= grade 3	Grade of primary tumor after WHO -73. This is done by an experienced uropathologist when going through all cases for validation.
WHO 04	controlled vocabulary	0= PUNLMP 1=low grade 2=high grade	Grade of primary tumor after WHO -04. This is done by an experienced uropathologist when going through all cases for validation.
Stage	controlled vocabulary	1=Ta 2=T1 3=T2 4=T3 5=Tis	T-stage for primary tumor
CK20	controlled vocabulary	1= negative (≤3) 2=positive (>3)	CK20 Immunohistochemistry ImmunoReactiveScore, threshold 3
MAI SUS protocol	controlled vocabulary	0=Low ≤15 1=High >15	Mitotic activity index, number of mitosis in tumor cells at 1,59 mm2. Threshold 15

Ki67	controlled vocabulary	0=Low ≤39 1=High >39	Ki67 measured by image analysis. Threshold 39%
Recurrence	boolean	0=No 1=Yes	Recurrence of urothelial carcinoma/ carcinoma in situ locally in the bladder (not registered after cystectomy).
Time to recurrence	number (months)		Time from date at diagnosis to local recurrence in the bladder (in months). Those without recurrence are given time from diagnosis to last known control with cystoscopy, or cystectomy.
Stage Progression	boolean	0=No 1=Yes	Any progression in TNM (also after cystectomy).
Time to progression	number (months)		Time from date at diagnosis to any progression in TNM (in months). Those without progression are given time from diagnosis to death, to last known contact if they moved, or until 30.6.2016 for the rest.
Metastasis	boolean	0=No 1=Yes	Metastasis to nodes (N) or distant metastasis (M) from urothelial carcinoma/ carcinoma in situ in the bladder. The evaluation is based on histology, radiology and/ or clinical information in the patient journal.
Follow-up recurrence	number (months)		Time from diagnosis to cystectomy or last control with cystoscopy (might be most relevant for recurrences).
Follow-up_progression	number (months)		Time from diagnosis to death or until 30.6.2016 (might be most relevant for progression).
Sex	controlled vocabulary	1=male 2=female	Sex
Age	number		Age at time of diagnosis in 5 year cohorts
Multifoc	boolean	0=No 1=Yes	Multifocal tumor
Size	number (cm)		Tumor size in cm
CIS	boolean	0=No 1=Yes	Carcinoma in situ
PPH3	controlled vocabulary	1= low (≤38) 2=high (>38)	Phospho HistoneH3 Immunohistochemistry in 1.59 mm2. Threshold 38
CD25	controlled vocabulary	1= low (<1) 2=high (≥1)	CD25 Immunohistochemistry. Threshold 1%.
P53	controlled vocabulary	1= low (<15) 2=high (≥15)	P53 immunohistochemistry. Threshold 15%
BCG	boolean	0=No 1=Yes	BCG treatment

Chemo	boolean	0=No 1=Yes	Received chemotherapy
Immunotherapy	boolean	0=No 1=Yes	Received immunotherapy
NGS	boolean	0=No 1=Yes	Next Generation sequencing performed Oncomine Focus Panel (DNA only)
FGFR3-alteration	boolean	0=No 1=Yes	FGFR3 hotspot mutation or amp?

3.3.4 Metadata for AI models

Compared to other research assets in CLARIFY (WSI, WSI annotations and clinical data), AI models are the products of the research which can be reproducible and reused by clinical practitioners and researchers. For AI models, UPV and UiS provide the metadata that is important for reusing AI models (see Table 3.6). Considering the complexity of AI models, description of model architectures, input-output constraints as well as implementation details are usually beyond simple controlled vocabularies. Thus the "Type" for each "Field" in Table 3.6 is specified as "text", which means natural language description is allowed.

Table 3.6 The metadata requirements for AI models.

Field	Type	Value	Description
task	text		Targeted tasks of the models, such as classification, segmentation, etc.
cancer type	text		Cancer type that is dealt with by the models
model type	text		Type of the models, such CNN, autoencoder, etc.
model architecture	text		Architecture of the models, including layers, classes
training dataset	text		Training datasets being used, including the size of the dataset and the numbers of images
test dataset	text		Test datasets being used, including the size of the dataset and the numbers of images
performance	text		performance metrics given for the model used on the test dataset
hardware	text		Hardware being used during experiments
patch size	text		Patch size of the WSI , or patch sizes in case of multiscale models
resolution levels	text		if the models are multiscale or single scale models, and which WSI resolution level(s) that are required as input
preprocessing requirement	text		Preprocessing that has been performed on the training and test set prior to learning and testing (that has to be done

			before feeding WSI patches to the model)
programming language	text		Programming language for implementing the models
library used	text		Code library used for implementing the models

4 State of the Art

A survey on state-of-art metadata standards and metadata standardization strategies is also conducted outside the CLARIFY consortium.

The survey methodology is as below:

- Web search using Google search engine. Documents associated with the metadata standards w.r.t, SML, TNBC, HR-NMIBC are searched through Google search engine.
- Output review for leading standard development organizations. By checking the output of the leading standard development organizations, we collect the recommendations and standards relevant to the targeted research assets.
- Literature review. By reviewing the recent scientific papers related to histopathological image analysis on top venues, the data usage information can be extracted, following which we analyzed the data management situation in the academic world and selected the best practices.

4.1 Metadata Standards for Medical Data

Digital Imaging and Communications in Medicine (DICOM)[4] is the international standard for medical images and related information. It defines the formats for medical images that can be exchanged with the data and quality necessary for clinical use. DICOM Working Group 26 (DICOM WG-26: Pathology)[5] specifically aims at developing the DICOM Standard in Pathology Domain so that the whole slide images and also the macros can be handled when it comes to produce, store and communicate. The DICOM WG-26 provides a document to describe the characteristics of whole slide images and the DICOM Whole Slide Image Storage IOD[6].

The Pathology and Laboratory Medicine (PaLM) domain of Integrating the Healthcare Enterprise (IHE) issues the “Anatomic Pathology Workflow in an Era of Digital Imaging” (APW-EDM) White Paper[7] to describe use cases, data elements, actors, and transactions necessary to support anatomic pathology workflows that leverage digital technologies.

The International Organization for Standardization (ISO)[8] is a worldwide federation of national standards bodies (ISO member bodies). There are some standards developed by ISO that are related to health data exchange and pathology.

ISO 15189:2012 Medical laboratories — Requirements for quality and competence[9] specifies requirements for competence and quality that are particular to medical laboratories. Medical laboratory services are essential to patient care and therefore have to be available to meet the needs of all patients and the clinical personnel responsible for the care of those patients. Such

[4] https://www.dicomstandard.org/current
[5] https://www.dicomstandard.org/activity/wgs/wg-26
[6] http://dicom.nema.org/Dicom/DICOMWSI/
[7] https://wiki.ihe.net/index.php/APW-EDM_White_Paper
[8] https://www.iso.org/home.html
[9] https://www.iso.org/standard/56115.html

services include arrangements for examination requests, patient preparation, patient identification, collection of samples, transportation, storage, processing and examination of clinical samples, together with subsequent interpretation, reporting and advice, in addition to the considerations of safety and ethics in medical laboratory work.

ISO 22857:2013 Health informatics — Guidelines on data protection to facilitate trans-border flows of personal health data[10] aims to facilitate international and trans-jurisdictional health-related applications involving the transfer of personal health data. It seeks to provide the means by which health data relating to data subjects, such as patients, will be adequately protected when sent to, and processed in, another country/jurisdiction.

ISO 13606:2019 Health informatics — Electronic health record communication is to define a rigorous and stable information architecture for communicating part or all of the electronic health record (EHR) of a single subject of care (patient) between EHR systems, or between EHR systems and a centralized EHR data repository. It consists of 5 parts, which are: (1) Reference model[11]; (2) Archetype interchange specification[12]; (3) Reference archetypes and term lists[13]; (4) Security[14]; (5) Interface specification[15].

Health Level Seven International (HL7)[16] is a not-for-profit, ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.

Fast Healthcare Interoperability Resources (FHIR)[17] is a next generation standards framework created by HL7. It is designed to enable information exchange to support the provision of healthcare in a wide variety of settings. The specification builds on and adapts modern, widely used RESTful practices to enable the provision of integrated healthcare across a wide range of teams and organizations.

Clinical Document Architecture (CDA)[18] developed by HL7 is a document markup standard that specifies the structure and semantics of "clinical documents" for the purpose of exchange between healthcare providers and patients. It defines a clinical document as having the following six characteristics: (1) Persistence, (2) Stewardship, (3) Potential for authentication, (4) Context, (5) Wholeness and (6) Human readability.

This Note “Dataset Descriptions: HCLS Community Profile”[19] was produced by the Semantic Web in Health Care and Life Sciences (HCLS) Interest Group is a specification for the description of datasets that meets key functional requirements, uses existing vocabularies, and is expressed using the Resource Description Framework (RDF). It discusses elements of data

[10] https://www.iso.org/standard/52955.html
[11] https://www.iso.org/standard/67868.html
[12] https://www.iso.org/standard/62305.html
[13] https://www.iso.org/standard/62303.html
[14] https://www.iso.org/standard/62306.html
[15] https://www.iso.org/standard/62304.html
[16] https://www.hl7.org/index.cfm
[17] http://hl7.org/fhir/summary.html
[18] https://www.hl7.org/implement/standards/product_brief.cfm?product_id=7
[19] https://www.w3.org/TR/hcls-dataset/

description including provenance and versioning, and describes how these can be used for data discovery, exchange, and query (with SPARQL), which then enables the retrieval and reuse of data to encourage reproducible science.

The Research Data Alliance (RDA)[20] is a research community organization with the goal of building the social and technical infrastructure to enable open sharing and re-use of data. RDA's Metadata Interest Group[21] concerns itself with all aspects of metadata for research data. RDA's Health Data Interest Group[22] focuses on the intricacies of Health Data, especially as it relates to privacy and security issues in Healthcare. However, no metadata standards related to digital pathology are found from above Interest Groups.

4.2 Annotated Databases

Literature review is to collect the annotated databases in the academic area, as well as to obtain a clearer picture of how the histopathological data are stored and managed in practice. The methodology is as following:

1. Select papers from three top venues in medical image analysis domain using the keywords "patholog*", "WSI", "whole slide image";
2. Extract information of data sets from selected papers;
3. Collect extra information from data portals, data publication websites, etc.

As a result of the literature collection, totally 51 papers published on three top venues are sorted out for review. The number of papers from each venue are shown in Table 4.1. Among all the datasets, 58% of them are open and publicly accessible. Unfortunately, the papers that mention open datasets seldom provide the URL to the datasets. Thus the datasets are searched with Google search engine by their names and verified by the descriptions of the corresponding papers. Open datasets as well as the hosting repositories are listed in Table 4.2. As the information of the datasets usually leads to the hosting repositories and data portals, it is more reasonable to investigate the repositories than a single dataset. The information on persistent identifier (PID) and metadata schema of data repositories are provided by Table 4.3.

Table 4.1 Number of papers selected from three top venues.

Publication Venues	Number of papers
MICCAI (International conference on Medical Image Computing & Computer Assisted Intervention)	21
TMI (IEEE Transactions on Medical Imaging)	14
MIA (Medical Image Analysis)	16
Total	51

[20] https://rd-alliance.org/
[21] https://www.rd-alliance.org/groups/metadata-ig.html
[22] https://www.rd-alliance.org/groups/health-data.html

Table 4.2 Open datasets as well as the hosting repositories.

Dataset	Repository	URL	Conferences/Journals
PCam database	Github, Google drive	https://github.com/basveeling/pcam	MICCAI
TUmor Proliferation Assessment Challenge 2016 (TUPAC16)	Google drive	http://tupac.tue-image.nl/	TMI
BreAst Cancer Histology images (BACH)	Grand challenge	https://iciar2018-challenge.grand-challenge.org/Dataset/	MIA
Camelyon16 dataset	Grand challenge	https://camelyon17.grand-challenge.org/	MICCAI
MICCAI 2018 Monuseg challenge	Grand challenge	https://monuseg.grand-challenge.org/	MICCAI
2014 MITOSIS dataset	Grand challenge	https://mitos-atypia-14.grand-challenge.org/dataset/	MIA
133 whole slide H&E tissue sections from 133 different patients	Internal server	http://www.sfu.ca/~abentaie/LSVM_CTXT/LSVM_CTXT.html	MIA
Colorectal nuclear segmentation and phenotypes (CoNSeP) dataset	Warwick TIA Lab	https://warwick.ac.uk/fac/sci/dcs/research/tia/data/	MIA
GlaS (Gland Segmentation dataset)	Warwick TIA Lab	https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/	MIA
2018 Data Science Bowl (DSB2018)	Kaggle	https://www.kaggle.com/c/data-science-bowl-2018	MICCAI
BreaKHis	Kaggle	https://www.kaggle.com/ambarish/breakhis	TMI

Table 4.3 Information of PID and metadata standard of open data repositories.

Name	Type	PID	Metadata schema/standard
The Cancer Genome Atlas (TCGA)	platform	UUID	GDC data dictionary
IEEE Dataport	platform	DOI	some required fields
The Cancer Imaging Archive (TCIA)	platform	DOI	managed by users
Grand Challenge	platform	Zenodo	managed by users
Kaggle	platform	managed by users	Frictionless data specification
Broad Bioimage Benchmark Collection (BBBC)	collection	No PID	5 required fields
Warwick TIA Lab: Datasets for Sharing	collection	No PID	No metadata schema

Apart from the datasets mentioned in the selected papers, the Digital Pathology Association (DPA)[23] provides a Whole Slide Image Repository[24] that aggregates more than 30 collections of pathological images from academic, independent and industry sources. Among them, the Cancer Imaging Archive (TCIA)[25] is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download. It contains multiple modalities of images, such as MRI, CT, Pathology, etc.

The Cancer Genome Atlas (TCGA)[26] is dedicated to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from Clinical, genetic, and pathological data resides in the Genomic Data Commons (GDC) Data Portal[27] while the radiological data is stored on The Cancer Imaging Archive (TCIA). Matched TCGA patient identifiers allow researchers to explore the TCGA/TCIA databases for correlations between tissue genotype, radiological phenotype and patient outcomes.

TCGA provides Breast Invasive Carcinoma (TCGA-BRCA)[28], Skin Cutaneous Melanoma (TCGA-SKCM)[29], Bladder Urothelial Carcinoma (TCGA-BLCA)[30] datasets for breast cancer, skin cancer and bladder cancer, respectively. TCGA-BRCA contains 1098 cases of Adenomas and Adenocarcinomas; Adnexal and Skin Appendage Neoplasms; Basal Cell Neoplasms; Complex Epithelial Neoplasms Cystic; Mucinous and Serous Neoplasms; Ductal and Lobular

[23] https://digitalpathologyassociation.org/
[24] https://digitalpathologyassociation.org/whole-slide-imaging-repository
[25] https://www.cancerimagingarchive.net/collections/
[26] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[27] https://portal.gdc.cancer.gov/
[28] https://portal.gdc.cancer.gov/projects/TCGA-BRCA
[29] https://portal.gdc.cancer.gov/projects/TCGA-SKCM
[30] https://portal.gdc.cancer.gov/projects/TCGA-BLCA

Neoplasms; Epithelial Neoplasms, NOS; Fibroepithelial Neoplasms; Squamous Cell Neoplasms. TCGA-SKCM contains 470 cases of Nevi and Melanomas. TCGA-BLCA contains 412 cases of Adenomas and Adenocarcinomas; Epithelial Neoplasms, NOS; Squamous Cell Neoplasms; Transitional Cell Papillomas and Carcinomas.

For breast cancer, TCIA hosts relevant datasets such as Assessment of Residual Breast Cancer Cellularity after Neoadjuvant Chemotherapy using Digital Pathology (Post-NAT-BRCA)[31] and Breast Metastases to Axillary Lymph Nodes (Breast-Mets-Lymph-Nodes)[32]. The Post-NAT-BRCA dataset is a collection of representative sections from breast resections in patients with residual invasive BC following NAT. Histologic sections were prepared and digitized to produce high resolution, microscopic images of treated BC tumors. Also included, are clinical features and expert pathology annotations of tumor cellularity and cell types. The Breast-Mets-Lymph-Nodes dataset consists of 130 de-identified whole slide images of H&E stained axillary lymph node specimens from 78 patients. The slides were scanned at Memorial Sloan Kettering Cancer Center (MSKCC) with Leica Aperio AT2 scanners at 20x equivalent magnification (0.5 microns per pixel). Together with the slides, the class label of each slide, either positive or negative for breast carcinoma, is given. The slide class label was obtained from the pathology report of the respective case.

Meanwhile, CLARIFY provides a list of public WSI datasets, shown in Table 4.4.

Table 4.4 Additional public WSI datasets collected by CLARIFY.

Dataset	URL
CAMELYON17 challenge	https://camelyon17.grand-challenge.org/
SPIE-AAPM-NCI BreastPathQ: Cancer cellularity challenge 2019	https://breastpathq.grand-challenge.org/
DigestPath 2019: Digestive-system pathological detection and segmentation challenge 2019	https://digestpath2019.grand-challenge.org/
Prostate cANcer graDe Assessment (PANDA) challenge	https://panda.grand-challenge.org/
MoNuSAC 2020: Multi-organ nuclei segmentation and classi_cation challenge	https://nucls.grand-challenge.org/NuCLS/
Herohe: ECDP2020	https://ecdp2020.grand-challenge.org/
NuCLS datasets	https://nucls.grand-challenge.org/NuCLS/
Breast cancer semantic segmentation	https://bcsegmentation.grand-challenge.org/
Prostate fused-MRI-pathology	https://wiki.cancerimagingarchive.net/display/Public/Prostate+Fused-MRI-Pathology
SICAPv2 - Prostate Whole Slide Images with Gleason Grades Annotations	https://data.mendeley.com/datasets/9xxm58dvs3/1

[31] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52758117
[32] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52763339

5 Summary

5.1 Gap Analysis

5.1.1 WSI

The general requirement for the whole slide images is a common image format across different types of pathological images. Currently the image formats used in CLARIFY are Deep Zoom Image (.dzi), Hamamatsu NanoZoomer Digital Pathology Image (.ndpi) , Tagged Image File Format (.tiff) and file format from Leica scanners (.scn) . The most promising work is the implementation and the adoption of DICOM standard for whole slide image generation and storage. However, in reality not all medical partners have scanners that support DICOM format. Therefore, in current status, it is not compulsory to adopt DICOM for WSIs, but the image format should be specified in the metadata.

5.1.2 WSI annotation

Annotation standards are elusive for the specific cancer types, hence it is CLARIFY's responsibility to develop annotation protocols. The metadata requirements are derived from the annotation protocols from the CLAIRY project. It is obvious that different annotation strategies are adopted for different cancer types, and thus produce varying labels as metadata.

5.1.3 Clinical data

According to the survey results, although a lot of standards exist for health data exchanging and sharing, there are no commonly adopted metadata standards that target clinical data with respect to the selected cancer types, namely SML, TNBC and HR-NMIBC. As pointed out by the FAIR principle F2: Data are described with rich metadata[33], the metadata should be generous and extensive without presuming the intended data users and the purposes. Therefore, the principle for generating clinical data is to provide as much information as possible while complying with the privacy-protection regulations.

5.1.4 AI models

Compared to WSIs, annotations and clinical data that usually have confined values for a certain field, AI models encompass a large variety of models that may have thousands of different architectures and settings. Meanwhile, the AI experts have been exploring new models leveraging new information. For instance, the classification models over SML whole slide images now use the global label (benign or malignant), but the regional information can also be exploited to increase classification accuracy and robustness. Thus, it is not feasible to design controlled vocabulary for the metadata.

[33] https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/

5.2 Recommendation and Plan

Based on the requirements, the state of arts, and current status of the CLARIFY project, we recommend a minimum set of metadata being adopted for WSI, identifiers being used by WSI annotations and clinical data, as shown in Table 3.1. The contents of the annotations should follow the annotation protocols specified for each cancer type and the metadata should be provided according to Table 3.2. For contents in clinical data, it is recommended that variables specified by Table 3.3, Table 3.4, and Table 3.5 are filled with corresponding clinical information with respect to each cancer type. For AI models, a minimum set of metadata is also recommended as in Table 3.6 to publish the models.

While the recommendations above are made based on current understanding between AI researchers, pathologists and computer scientists within the CLARIFY project, further plans will be dedicated to improve the interoperability of research assets beyond the scope of CLARIFY. Considering that the CLARIFY project is still at its early phase and meanwhile there are on-going movements and activities in the domain of digital pathology, better standards will be achieved.

6 References

1. NEMA PS3 / ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard, National Electrical Manufacturers Association, Rosslyn, VA, USA (available free at http://www.dicomstandard.org/)

2. Singh R, Chubb L, Pantanowitz L, Parwani A. Standardization in digital pathology: Supplement 145 of the DICOM standards. J Pathol Inform. 2011 Jan 1;2(1):23–23. doi: 10.4103/2153-3539.80719

3. Herrmann, Markus D et al. "Implementing the DICOM Standard for Digital Pathology." Journal of pathology informatics vol. 9 37. 2 Nov. 2018, doi:10.4103/jpi.jpi_42_18

4. Berman, F. (2019): The Research Data Alliance --The First Five Years, Supplement to: Berman, F., & Crosas, M. (2020). The Research Data Alliance: Benefits and Challenges of Building a Community Organization. Harvard Data Science Review, 2(1). doi: 10.1162/99608f92.5e126552