Thesis

METADATA-CATALOG OF HISTOPATHOLOGICAL WHOLE SLIDE IMAGES

submitted by

Carmen Zerner BSc.

in partial fulfillment of the requirements for the degree of

Doktorin der gesamten Heilkunde

(Dr^{in}. med. univ.)

at the

Medical University of Graz

executed at the

Diagnostic & Research (D&F) Institute of Pathology

under supervision of

Dr. scient. med. Dipl.-Ing. Ing. Markus Plass BSc.

Ao. Univ.-Prof. Dr. med. univ. Peter Regitnig

Graz, 18 June 2026

Declaration of Academic Integrity

I hereby confirm that the present diploma thesis is the result of my own independent scholarly work. I also confirm that in all cases, where material from the work of others (in books, articles, essays, dissertations, and on the internet) is acknowledged, quotations and paraphrases are clearly indicated. No material other than that cited in the reference list has been used. I have read and understood the Medical University's regulations and procedures concerning plagiarism.

Furthermore, I hereby declare that if artificial intelligence (AI) tools were used for the generation and/or correction of certain text passages in the creation of this work, such employment was conducted in compliance with ethical principles, academic integrity, and the regulations of my university. Additionally, it was ensured that this usage was transparently disclosed and appropriately attributed.

Graz, 18 June 2026 Carmen Zerner, m.p.

Acknowledgement

I would like to express my deepest gratitude to my two supervisors, Dr. scient. med. Dipl.-Ing. Ing. Markus Plass BSc. and Ao. Univ.-Prof. Dr. med. univ. Peter Regitnig, for their patience and guidance. Their professional expertise and dedicated support contributed significantly to the successful completion of this thesis. Not only did their constructive feedback improve the quality of this work, but it also led to valuable new personal insights for me.

I would also like to extend my sincerest thanks to the members of the Müller and Plass research group for their professional support, critical discussions, and, above all, their constant encouragement of my academic development. It has been a pleasure to be part of this team for such a long time.

I would be remiss in not mentioning my parents, who have supported me not only while I was writing this thesis, but throughout my entire journey. Lastly, I'd like to thank my friends Anita, Birgit, Kathrin and Klara for their emotional support, their many words of encouragement, and the much-needed distraction during intense writing sessions. The support I received from them gave me the strength I needed to stay focused on my objectives.

Zusammenfassung

Einleitung

Die Digitalisierung der Pathologie schreitet global immer weiter voran. Durch die Implementierung von hochauflösenden Scannern in den histopathologischen Workflow werden große Mengen an Bilddaten mit dazugehörigen Metadaten erzeugt. Um diese Daten effizient zu verwalten, sie für die Forschung nutzbar zu machen und einzelne Fälle einfach aufzufinden, wird ein Metadatenkatalog für Whole Slide Images erstellt. Dieser Katalog soll von allen teilnehmenden österreichischen Pathologie-Instituten gleichermaßen befüllt werden, um so einen harmonisierten Datensatz zu erhalten, auf den alle beteiligten Institute Zugriff haben. Als Grundlage dienen das Laborinformationssystem des Diagnostik- und Forschungsinstituts für Pathologie in Graz, Kataloge, die im Rahmen von Forschungsprojekten entstanden und internationale Standards und Klassifizierungen. Ziel dieser Arbeit ist es, die Datenfelder für diesen Katalog zu ermitteln, die es ermöglichen, die vorhandenen Daten in strukturierter Form durchsuchbar zu machen und sie für die Forschung zur Verfügung zu stellen.

Methoden

Um zu eruieren, in welcher Form Daten aktuell am Diagnostik- und Forschungsinstitut für Pathologie in Graz gespeichert werden, werden die Struktur und die Datenfelder von PAS Xanthos, dem Laborinformationssystem, analysiert. Um den Bereich der Forschung miteinzubeziehen, werden Datenfelder untersucht, mit denen zwei Forschungskohorten, die ADOPT-Kohorte und die CRC-Kohorte, beschrieben werden. Auch Standards für Biobanking (MIABIS Core), Probenverwaltung (MIABIS Sample+Donor+Event) und digitale Pathologie (MIABIS Digital Pathology) werden untersucht. Auf Basis der Analyse dieser Datenfelder entsteht ein erster Entwurf von Datenfeldern für den Katalog. Dieser Entwurf wird in einem Workshop bearbeitet, an dem Patholog*innen, Forscher*innen, Techniker*innen von verschiedenen österreichischen Pathologie-Instituten teilnehmen. Nach der interdisziplinären Einigung auf ein finales Set an Datenfeldern im Rahmen des Workshops werden die Datenfelder in eine hierarchische Struktur gebracht und so miteinander verknüpft. Außerdem werden

die Inhalte spezifischer Datenfelder auf internationale Standards, wie SNOMED-CT gemappt.

Ergebnisse

Das Ergebnis dieser Arbeit ist die Struktur, in der die Felder in der Katalogisierungssoftware abgebildet werden, die Auswahl der Datenfelder und das Mapping spezifischer Datenfelder auf internationale Standards wie SNOMED-CT. Es wird eine patientenzentrierte hierarchische Struktur erstellt, bei der die Unterkategorien immer einem bestimmten Patienten zuordenbar ist. Die Datenfelder, die im Rahmen der Bestandsanalyse und des darauffolgenden Workshops entstehen, werden in diese Struktur eingebettet und somit miteinander verknüpft. Vordefinierte Werte, die zur Befüllung spezifischer Datenfelder verwendet werden, werden auf den internationalen Standard SNOMED-CT gemappt.

Konklusion

Im Rahmen dieser Arbeit entsteht ein interdisziplinär abgestimmter Metadatenkatalog, der durch die Integration von internationalen Standards wie SNOMED-CT und einer patientenzentrierten Struktur eine Grundlage für die österreichweite Harmonisierung von Daten der digitalen Pathologie schafft. Die erfolgreiche Implementierung von 70.000 Datensätzen in Graz belegt die Machbarkeit und das Potenzial für die multizentrische Forschung. Jedoch zeigen sich im Laufe der Arbeit auch deutliche Herausforderungen, wie das Finden der Balance zwischen diagnostischer Effizienz und Detailtiefe und der für die Forschung notwendigen Kodierung von Daten. Limitationen hinsichtlich der Repräsentativität der Bedürfnisse aller zukünftigen Nutzer*innen ergeben sich durch die kleine Gruppe der Teilnehmenden am Workshop. Auch der hohe Wartungsaufwand, der sich durch zukünftige Änderungen medizinischer Standards ergeben kann, muss berücksichtigt werden. Zukünftige Arbeiten sind erforderlich, um eine Automatisierung der Datenextraktion zu ermöglichen.

Abstract

Introduction

The digitization of pathology continues to advance globally. The implementation of high-performance scanners into the histopathological workflow generates large volumes of image data along with associated metadata. To manage this data efficiently, make it available for research, and easily locate individual cases, a metadata catalog for whole-slide images is being created. This catalog is to be populated equally by all participating Austrian pathology institutes in order to obtain a harmonized dataset accessible to all participating institutes. The basis for this are the laboratory information and management system of the Institute of Pathology in Graz, catalogs developed within the framework of research projects, and international standards and classifications. The aim of this work is to identify the data fields for this catalog that will make it possible to make the existing data searchable in a structured form and make it available for research.

Methods

To determine how data is currently stored at the Diagnostic and Research Institute of Pathology in Graz, the structure and data fields of PAS Xanthos, the laboratory information and management system, are being analyzed. To include the research sector, data fields used to describe two research cohorts - the ADOPT cohort and the CRC cohort - are being examined. Standards for biobanking (MIABIS Core), sample management (MIABIS Sample+Donor+Event), and digital pathology (MIABIS Digital Pathology) are also being examined. Based on the analysis of these data fields, an initial draft of data fields for the catalog is created. This draft is refined in a workshop attended by pathologists, researchers, and technicians from various Austrian pathology institutes. After interdisciplinary consensus on a final set of data fields is reached during the workshop, the data fields are incorporated into a hierarchical structure and linked together. In addition, the contents of specific data fields are mapped to international standards, such as SNOMED-CT.

Results

The result of this work is the structure in which the fields are represented in the cataloging software, the selection of data fields, and the mapping of specific data fields to international standards such as SNOMED-CT. A patient-centered hierarchical structure is created in which the subcategories can always be assigned to a specific patient. The data fields identified during the inventory analysis and the subsequent workshop are embedded within this structure and therefore linked to each other. Predefined values used to populate specific data fields are mapped to the international standard SNOMED-CT.

Conclusion

This thesis is resulting in an interdisciplinary metadata catalog that, through the integration of international standards such as SNOMED-CT and a patient-centered structure, establishes a foundation for the Austria-wide harmonization of digital pathology data. The successful implementation of 70,000 data records in Graz demonstrates the feasibility and potential for multicenter research. However, significant challenges have also emerged during the course of the work, such as finding the balance between diagnostic efficiency and depth of detail and the data coding required for research. Limitations regarding the representativeness of the needs of all future users arise from the small group of workshop participants. The high maintenance effort that may result from future changes in medical standards must also be considered. Further work is required to enable the automation of data extraction.

Table of Contents

List of abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CPath	Computational Pathology
CRC	Colorectal cancer
DICOM	Digital Imaging and Communications in Medicine
EHR	Electronic Health Record
EUCAIM	Cancer Image Europe
ICD	International Classification of Diseases
IHC	Immunohistochemistry
LIMS	Laboratory information and management system
MIABIS	Minimum Information About BIobank data Sharing
SNOMED-CT	Systematized Nomenclature of Medicine – Clinical Terms
WHO	World Health Organization
WSI	Whole Slide Image

List of figures

List of tables

1 Introduction

In modern pathology, tissue samples are processed, stained, and archived on glass slides. Thanks to ongoing digitization - as is currently being implemented at the Diagnostic and Research Institute of Pathology in Graz - these samples are digitized immediately after preparation using high-performance scanners. In addition to the actual image data, this process generates extensive metadata that provides a detailed description of each digital scan of a slide. In order to obtain an overview of all slides and their digital images, also known as Whole Slide Images (WSIs), throughout Austria, these should be mapped in a catalog. This catalog should be filled by all members of a national consortium consisting of several pathology institutes from Austria: The Johannes Kepler University Linz, the Medical University of Innsbruck, the Medical University of Vienna, the Veterinary University of Vienna and the Medical University of Graz. A metadata catalog for WSIs can be viewed as the backbone that transforms vast amounts of image data and metadata into a searchable database suitable for further scientific use. As part of this thesis, data fields will be selected and normalized to populate the catalog. Based on the selected data fields, a WSI will be described in adequate detail. The goal is to obtain an Austrian wide harmonized dataset. To this purpose, current systems will be analyzed, a workshop involving consortium members will be held, and findings from existing projects will be reviewed.

1.1 Relevance

The systematic cataloging of WSI data offers several key advantages. The goal is to create a harmonized dataset that all relevant stakeholders will have access to. This allows for an overview of what data is available throughout Austria and thus expands the possibilities for using this data for research purposes. The targeted selection of data fields and the storage of data in a structured format makes it easier to locate the desired data and provides a solid foundation for systematic queries from the catalog. This facilitates the creation of cohorts for research questions. To develop and apply algorithms and artificial intelligence (AI) on WSI, it is necessary to have consistent data structures in order to achieve valid results.

1.2 Background

Digital pathology marks a technological paradigm shift from conventional light microscopy to computer-assisted analysis of high-resolution digital tissue sections. Yet the digital image is just the end result of a process that still begins with the traditional preparation of the physical tissue sample in the laboratory.

1.2.1 Overview of digital pathology

The Diagnostic and Research Institute of Pathology at Medical University of Graz receives tissue samples from various sources. Biopsies from diagnostic procedures and surgical resections arrive at the sample reception unit of the pathology institute, where the medical case is registered in the pathology laboratory information and management system (LIMS).

After the registration the sample is transferred to the macropathology laboratory for the macroscopic assessment. Here a pathologist describes the origin, size, color, shape and consistency of the sample and any visible pathological condition as well as distance to surgical margins, if applicable. Larger pieces of tissue, for example surgically removed tumor tissue, are marked by the surgeons to understand the orientation of the tissue in the body. In order to make this orientation visible to the pathologist on the microscopic image, the edges of the sample are marked with ink in different colors. After examination and, if needed, cutting of the larger pieces, the samples are put into barcode-labelled plastic cassettes and kept in formalin until further processing. The description of the tissue and the number and contents of the plastic cassettes are dictated by the pathologist. The record is forwarded to the medical secretariat, which then prepares the macropathological report.

The formalin-fixed specimens are transferred to the histology laboratory for histoprocessing. Since the tissue needs to be cut into thin slices, it is necessary to support the tissue structure. Therefore, the plastic cassettes containing the specimens remain several hours in the tissue processor machines, where the specimens are dehydrated and completely infiltrated with paraffin wax. After this process, the specimens are manually embedded in rectangular blocks with paraffin and cooled to become solid. A microtome is used to cut 2-4 µm thin slices

from these blocks. The finished paraffin slices are applied to a warm water bath. From there they are mounted onto a glass slide. The tissue sections on the glass slides are now sufficiently prepared for the staining process.

In order to be able to examine the structures of the tissue under the microscope, they are made visible through staining. As standard, the overview staining is carried out with hematoxylin and eosin. This is applied to all cases and performed in an automated tissue staining machine. There are many other special stainings that can be performed automated and/or manually by the staff in the histology laboratory. If required, these are requested by the pathologist and subsequently carried out. This also applies to immunohistochemical (IHC) staining performed in the immunohistochemistry laboratory. After staining, a cover glass is placed over the tissue to protect the sample.

All finished slides that belong to a case are summarized in a folder and/or digital and then sent to the pathologist who is responsible for interpretation and diagnosis. The pathologist examines the stained tissue under a microscope to identify abnormal structures or disease processes. The histopathological findings and the diagnosis are dictated and sent to the medical secretariat. There, the record is written down and combined with the macropathological report to form the pathology report. The pathologist then receives the complete pathology report to validate it.

The following graphic (Figure 1) depicts the main steps of the histopathology process at the Diagnostic and Research Institute of Pathology in Graz: The blue arrows follow the process from specimen reception to the final pathology report, and the red arrows indicate at which steps along the histopathology workflow entries into the LIMS are made.

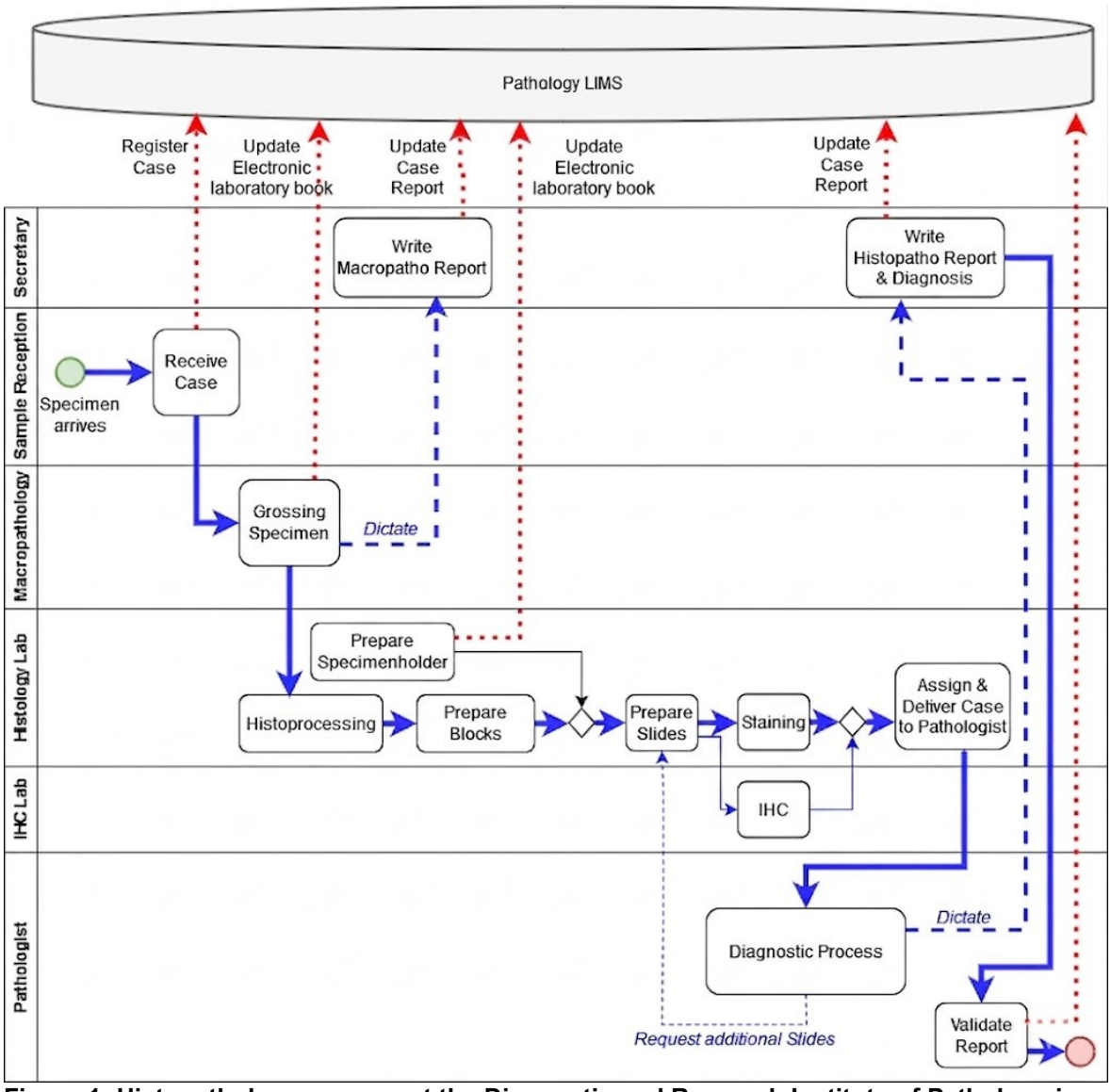


Figure 1: Histopathology process at the Diagnostic and Research Institute of Pathology in Graz

The histopathology process described above is suitable for the diagnostic pathway on site. However, the well-known shortage of doctors also affects the field of pathology. More pathologists are retiring than new pathologists are being recruited. (Ruiter et al., 2004) The population is also getting older, making it more likely that a person will need to use the services of a pathologist during their lifetime. (Warth et al., 2016) This is one field where digital pathology could lead to improvements. (Jeong et al., 2025)

The umbrella term "digital pathology" covers several areas of modern pathology. The most well-known difference to the traditional workflow in pathology is that

instead of viewing glass slides through the microscope, pathologists can view and process digital images of these slides on their computer screen. (Nam et al., 2020)
Although digital pathology is a phenomenon that has become more prominent in recent years, its origins go back much further. In 1986, Roland S. Weinstein, M.D. introduced the term "telepathology". The idea behind telepathology was to enable pathologists to work together over a long distance. (Weinstein et al., 2009)
There have been several approaches here over time. Initially, static images of individual slide sections were used, but later microscope stages were programmed to create a digital image of the entire slide. (Weinstein RS, 2001) These images can be regarded as precursors of the histopathological WSIs that are widely used today.

Like the development of staining techniques and microscopy, which were crucial for the transformation of pathology in the 19th century, the development and refinement of imaging techniques for the conversion of histopathological glass slides into high-resolution digital images initiated the currently ongoing transition to digital pathology. (Kiran et al., 2023)
Histopathological WSIs bring many benefits for pathology in diagnostics, education and research. These benefits include ease of access and sharing, integration with electronic workflows and LIMS, no risk of staining quality degradation or slide breakage, and the ability to view and examine multiple slides in parallel. (Jain et al., 2024)

The first WSI scanners were introduced in the 1990s. (Ferreira et al., 1997) A lot has happened in this sector since then. Various companies now offer fast, efficient and more cost-effective WSI scanners that can be used in routine diagnostics, teaching and research. The WSI therefore forms the base of the digitalized workflow in histopathology. (Jahn et al., 2020)
To create a WSI, the scanner captures the slide piece by piece. This is done in the form of tiles or lines. These tiles or lines are then assembled to create a complete image of the slide. (Indu et al., 2016)

Slides can be scanned at different magnifications. For most applications, such as ordinary image analysis and interpretation, a magnification of 20x is sufficient. If more information is required, a magnification of 40x may be necessary. 20x and 40x are the equivalent to 200x/400x magnification in the microscope. (Laurent et al., 2013)

The time it takes a scanner to generate a WSI depends on the device used, the size of the area being scanned and the magnification selected. The scanning time varies between 30 seconds and several minutes. (Pantanowitz et al., 2015)
In a Dutch study from 2009 the ScanScope XT scanner from Aperio, produced files with an average size of 373.1 MB, ranging from 5.0 MB to 3.9 GB. When the magnification of the scan is doubled, the scanning time and the size of the file will increase fourfold. (Huisman et al., 2010)
Another aspect that leads to a larger file size is the so-called z-stacking. In this process, several layers are scanned at the same spatial position along the z-axis of the slide. This allows the pathologists to switch between different focal planes on the WSI, which they are used to by adjusting the fine focus on the microscope. (Sturm et al., 2019)

WSIs are not just a digital reproduction of microscope slides, but function as a central element of modern pathology. (Zarella et al., 2019)
While glass slides run the risk of breaking, getting lost or the color of the tissue fading, WSIs can be digitally saved permanently without losing quality. This means that the original condition of the tissue can be viewed long after a slide has been produced and can then be used for research and further training. (Huisman et al., 2010)

WSIs also open up possibilities for telepathology or location-independent consultations. This is one of the most common and frequent use cases for WSIs. (Ghosh et al., 2017) Remote Consultation makes it possible to obtain a further expert opinion across national borders without the need to send the glass slide. The glass slide can remain at the original institute, saving the time, cost and risk of shipping. (Evans et al., 2015)

In a modern setting, the WSI Scanners are a fully inserted part of the workflow. Here, WSI scanners are integrated into the histopathology process. The glass slides are loaded directly from the histopathology laboratory into the scanner. This ensures optimal conditions, as the slides are fresh at this point and do not yet have any fingerprints, traces of markers or damage. (Hartman et al., 2017) This process was in development at the Diagnostic and Research Institute of Pathology in Graz at the time of writing this thesis. Six “Aperio GT 450 DX” WSI scanners from “Leica Biosystems” were purchased for this purpose and since January 2025 all histological glass slides are scanned and stored. In May 2026 already 550.000 WSIs have been scanned. Pathologists have been working digitally since May 2025. Most Pathologists of the institute switched from traditional microscopy to digital microscopy.

While static images of tissue samples were previously used in tumor boards, WSIs are now mainly used here as well. This means that during the patient presentation in the tumor board, you are no longer limited to the images previously taken. Several images from several cases can be presented in high quality and resolution. (Chen et al., 2014)

WSIs also play an important role in the training and further education of medical staff. The use of WSIs in courses on histology and histopathology continues to increase in training, whether in the field of human medicine, dentistry or veterinary medicine. (Saco et al., 2016) When it comes to standardizing the education of pathologists, WSIs are an important training tool. Carefully curated collections of appropriate cases can be made available in digitized form to all pathologists in training, regardless of where they complete their training. (Bruch et al., 2009)

Biotechnology and pharmaceutical companies use WSIs in their research on prognostic and predictive biomarkers. To this end, these companies have invested heavily in technologies for digital pathology. (Zarella et al., 2019)

While the use of WSIs modernizes pathology workflows and takes them into the digital age, it also opens up further possibilities. The availability of a large number of digital images provides an ideal working environment for AI. Digitalization and

the use of AI in the field of pathology is described under the term Computational Pathology (CPath). (Abels et al., 2019)
CPath benefits from the developments of recent years in the fields of histopathology, computer vision and bioinformatics. (Shmatko A, 2022) Relevant innovations in the field of histopathology include the aforementioned possibilities of in-line scanning, which is becoming an integral part of the histopathology process at more and more pathology institutes.
The development of digital image processing in histopathology began in the 1960s. Research was carried out into algorithms that would recognize predefined cell morphologies based on their geometric properties. (Prewitt and Mendelsohn, 1966)

This formed the basis for hand-crafted feature models. Hand-crafted features are features that can be linked to measurable attributes in the image. (Madabhushi and Lee, 2016)
These hand-crafted feature-based models thus enabled the first quantitative analysis of digital histology images. This includes algorithms for automatic detection of shapes and patterns that are typical for specific types of cancer or tissue. For example, mitoses, nuclei, lymphocytes or glandular structures are recognized and counted based on their architecture and arrangement. Accordingly hand-craftet feature-based models were an important step on the way to today's deep learning. (Andrew H. Song, 2023)

Subsequently, the classic machine learning algorithms were introduced, such as support vector machines and random forests. Support vector machines are used to separate two or more classes from each other. To do this, a line is drawn in a feature space that distinguishes two classes from each other. In the case of digital pathology, the support vector machine searches for a separating line that distinguishes two cell types from each other, for example. (Lecun et al., 1998)

Random forests are a collection of many individual decision trees that are trained on combinations of data and features. Each of these trees makes a prediction and the random forest then decides on the basis of the majority or average of the

individual trees. This enables random forests to make robust and reliable predictions. However, they reach their limit with large, high-dimensional data. (Becker et al., 2023)

From there, work on algorithms developed in the direction of so-called neural networks. The later emerging convolutional neural networks (CNNs) combine the extraction of features from images and the learning of associations. CNNs automatically recognize features in images without these features having to be defined manually beforehand. They learn directly from the raw data which features are relevant. A CNN consists of several layers that fulfill different tasks. In order to function precisely, CNNs need many images with which to be trained. Neural networks that have already been trained with a large amount of data can be fine-tuned with data from the area in which they are to be used. This method is called “transfer learning” and is used in fields where not enough data is available to initially train the neural network with this specific data. (Shmatko A, 2022)

While the algorithms and models described so far were aimed at recognizing the content of images and extracting features from images, the purpose of generative models is to generate new data. The generative models, such as the generative adversarial networks (GANs), are trained with real images and create new synthetic images based on this training data, which have the same properties as the images from the training data set. This synthetic data can be used to expand small or rather homogeneous data sets and thus make them usable as training data sets for machine learning algorithms. The conversation concerning data protection, which is particularly important for patient data, is also eliminated when using synthetic data. (Chen et al., 2021)

Digital pathology or CPath has the potential to enrich workflows in pathology in many ways. Diagnostic precision could be increased using machine learning algorithms. The combination of the expertise of experienced pathologists and these tools could reduce the error rate in diagnostics and thus improve the results for patients. (Ibrahim et al., 2020)

Another goal in the field of CPath is to digitize the diagnostic workflow and integrate supporting systems. These computer-aided diagnosis systems should not replace pathologists but support them on the path to diagnosis. Algorithms can also be helpful when making treatment decisions. Personalized medicine is on the rise and offers many new ways to fight different cancer cells. Deep learning models can help to identify patients who are likely to benefit from this treatment method. They can recognize certain structures in the tissue of the tumor that are associated with a positive outcome with certain treatments. (Hosseini et al., 2024)
AI methods can also be used to discover new prognostic and predictive biomarkers based on routine slides. These newly discovered biomarkers also offer approaches for research into cancer treatments. (Cifci et al., 2023)

There is therefore a great deal of hope for a significant improvement in pathological diagnostics and, accordingly, in the clinical treatment of cancer patients. However, the technical and legal limitations must also be kept in mind.

1.2.2 Software tools used in context of this thesis

Certain requirements are essential in order to successfully implement the planned metadata catalog. Firstly, a LIMS is required in which all diagnostic data is available. Furthermore, software is also needed in which the catalog is represented, and which makes it possible to manage the data records and create specific queries. The two systems intended for this purpose at the Diagnostic and Research Institute of Pathology in Graz are described in the following sections.

1.2.2.1 PAS Xanthos

PAS Xanthos, a LIMS, is the successor to PAS v2. It was introduced at the Diagnostic and Research Institute of Pathology in Graz in 2021 and covers the institute's processes - from sample reception and diagnosis to the distribution of cases, documentation of findings and all laboratory analysis. The workflow-controlled system accompanies the entire diagnostic process. This begins with sample acceptance. A unique accession number is assigned to an order, which is printed as text and barcode on the sample containers. The accompanying request form is digitized and also stored in the LIMS system. As described above, the

samples are sent to the macroscopic laboratory and transferred to plastic cassettes, which are also labeled and recorded in PAS Xanthos. The histological slides, which are obtained from these blanks from the macroscopic laboratory, are also labeled with barcodes and also mapped in LIMS. The slides are scanned in the digitization laboratory, and the scans are linked to the corresponding case in PAS Xanthos. Following this procedure, the slides and scans are sent to the pathologist for diagnosis. The macroscopic and histological description written from pathologist's dictations, information from the molecular diagnostics and the summarizing diagnosis and tumor classifications are also saved in PAS Xanthos.

1.2.2.2 OpenSpecimen

OpenSpecimen ("OpenSpecimen – Biobank LIMS Trusted by the World", n.d.) is the successor of caTissue Suite which was developed under the cancer Biomedical Informatics Grid (caBIG) program and funded from the National Cancer Center in 2004. (McIntosh et al., 2015)

OpenSpecimen is an open source biobanking informatics platform that is designed to be highly configurable in order to accommodate various specific researching and biobanking needs. The key feature of OpenSpecimen is the management of biospecimen data, however it also supports a wide range of other biobanking activities and offers several features which allow to improve efficiency and quality. OpenSpecimen is suitable for various purposes due to its adaptability and can therefore be used in biobanks, various research institutions and hospitals. It is possible to record the entire life cycle of a sample in OpenSpecimen - from collection, processing and storage to release or transfer to research projects. OpenSpecimen thus offers the possibility of complete traceability of samples. It is possible to log different sample types in OpenSpecimen. Various biological materials such as blood, urine, DNA or tissue can be systematically recorded, cataloged and managed. It is also possible to record aliquots and derivatives of individual samples. The modular structure enables a high degree of adaptability to individual requirements. Relevant data records, such as those required for research projects, can be collected via the user-friendly search function. OpenSpecimen is also very variable when it comes to managing user rights,

allowing fine-grained control over sample governance and who has access to which content. This is particularly important for medical data and ensures data security and compliance with regulatory requirements. The platform is web-based, open source and can be operated both locally and in the cloud. The architecture (RESTful APIs) enables uncomplicated integration into existing IT infrastructure. Interfaces to various LIMS, electronic health records (EHRs), Survey Tools such as REDCap, etc. are also available. (McIntosh et al., 2015)

1.2.3 Importance of metadata in medical imaging and diagnostics

In the digital era of pathology, large quantities of images are generated. It is essential to manage these images and handle them appropriately so that their full potential can be utilized for other non-diagnostic purposes, such as research or the development of AI algorithms. To achieve this, additional data associated with the images, so-called metadata, are essential. These metadata shall provide information about the provenance of the image as well as relevant clinical variables associated with the image. Metadata are of particular importance for the correct use and interpretation of medical images. (Kondylakis et al., 2022)

For metadata to be used effectively, it is required that the data is available in a machine-readable and standardized format. The most widely used standard worldwide for medical images and the associated information is Digital Imaging and Communications in Medicine (DICOM). (Larobina and Murino, 2014)
The origins of the DICOM standard date back to the 1980s. The American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) worked together on a way of transmitting data independently of manufacturer standards. They published DICOM in 1993. (Mildenberger et al., 2002)
DICOM files combine the pixel data, i.e. the images from X-ray, CT, MRI, ultrasound or histological images, with a standardized header containing essential metadata. This includes patient data, examination modalities, acquisition data and technical parameters. A universal standard such as DICOM makes it possible to work towards long-term interoperability and scalability. (Clunie, 2021)

Modern research inevitably involves large amounts of data. This poses a challenge for both humans and machines. Difficulties are caused, for example, by the availability of data in a wide variety of formats, inadequate documentation and difficult retrieval. Another challenge is scalability: humans cannot handle the mass of information efficiently, so machine processing is necessary. The FAIR principles were developed to remedy this situation. They are intended to increase the reusability of scientific data. The FAIR principles include Findability, Accessibility, Interoperability and Reusability. To be findable, data and metadata must have unique identifiers and be registered in searchable repositories. Data should also be sufficiently described with metadata. In order to be accessible, data must be retrievable via standardized communication protocols. These protocols are freely available and enable access controls where necessary. The metadata must be accessible even if the data is unavailable. In order to be interoperable, data must be available in a standardized format. This includes a vocabulary that complies with the FAIR principles. According to the FAIR principles, reusable data is characterized by being described in detail and accurately. For this purpose, a clearly defined and accessible data usage license and information on provenance must be available in advance. It is assumed that adherence to the FAIR principles will improve the reproducibility, transparency and long-term usability of data. (Wilkinson et al., 2016)

This indicates that data organized in datasets and listed in browsable, and searchable catalogs is significantly more useful when it is accompanied by relevant metadata. Metadata provides descriptive information about the available datasets.

1.2.4 Classification in pathology

Classifications play a central role in medicine, especially in pathology. They enable complex medical information to be presented in a structured and comparable manner, thus providing an important basis for communication, documentation, research and also for billing services. Numerous classifications and standards are used in medicine. Here, only five significant ones are described in more detail.

1.2.4.1 ICD-10

The International Classification of Diseases (ICD) is a globally recognized system for coding and classifying diseases, injuries and causes of death. It is published by the World Health Organization (WHO) and regularly revised to correspond to the current state of medical science. (Bundesministerium für Soziales, 2025)

The 10th revision (ICD-10) is currently in use in German-speaking countries. Although the successor version, ICD-11, officially came into effect in January 2022, its introduction is still in progress. The WHO is offering a flexible transition period of five years to enable smooth implementation in national health systems. (Bundesinstitut für Arzneimittel und Medizinprodukte, 2026)

The ICD-10 contains over 14,000 alphanumeric codes that precisely describe diseases, syndromes or specific health conditions. These codes are divided into chapters that are structured either according to the organ system affected or the etiology of the disease. For example, the chapter “Diseases of the respiratory system” contains codes for diagnoses such as bronchial asthma, chronic obstructive pulmonary disease or pneumonia. (World Health Organization, 2019b)

ICD-10 codes are alphanumeric and typically consist of one letter and one or more digits. The first three characters of the code represent the category of the disease or condition, while characters four to six provide additional, detailed information about the specific diagnosis. For example, the code for bronchial asthma is J45, where ‘J’ indicates that it is a disease of the respiratory system. The ‘45’ refers to bronchial asthma. Additional digits separated by a decimal point (.) can describe the diagnosis in even greater detail. For example, J45.0 stands for allergic bronchial asthma, while J45.1 refers to non-allergic bronchial asthma. Hierarchical coding allows for both a rough and a more detailed description of the diagnosis. (World Health Organization, 2019b)

In practice, the ICD serves several purposes, including clinical documentation. Doctors record diagnoses using ICD codes in electronic medical records. In addition, ICD codes form the basis for service documentation and billing of

services rendered to health insurance providers. Based on the documentation with ICD codes, epidemiological trends are observed, which form the basis for health policy planning by health authorities. (Bundesministerium für Soziales, 2025)

1.2.4.2 ICD-11

The eleventh revision of the ICD is the current global standard of the WHO for the systematic collection and analysis of data on causes of illness and death. Compared to its predecessor, the system has been fundamentally modernized to reflect the medical advances of the 21st century and the ongoing digitalization of healthcare. (World Health Organization, 2026)

Although ICD-11 officially took effect internationally in January 2022, its introduction in Austria, as well as in Germany and Switzerland, is currently still in the preparatory phase. The Federal Ministry of Social Affairs, Health, Care, and Consumer Protection is working on the national implementation and the German-language version. The WHO is granting a flexible transition period of at least five years, during which countries can continue to report data in ICD-10 format in parallel to enable a seamless technical and administrative transition of health systems. (Bundesinstitut für Arzneimittel und Medizinprodukte, 2026, World Health Organization, 2022a)

ICD-11 is significantly more comprehensive than its predecessor and comprises over 55,000 codes (compared to approximately 14,000 in ICD-10). It is designed as a digital database that allows for a more precise description of complex clinical pictures. The chapter structure has been expanded to 28 chapters (ICD-10 had 22). New features include separate chapters for diseases of the immune system, sleep-wake disorders, conditions related to sexual health, and a supplementary chapter on traditional medicine. (World Health Organization, 2022b)

The ICD-11 codes are alphanumeric and follow a new scheme. They consist of at least four characters. The first digit consists of a number or letter that describes the chapter number. The second digit is always a letter, except for “I” and “O,” which are not used to avoid confusion with ‘1’ and “0.” The letter in the second

position describes subchapters of the main chapters. The third digit is always a number and is determined by the hierarchical classification of a disease within its subchapter. The fourth digit can be a number or a letter and represents the final specification of the disease within its group. By using numbers and letters, it is possible to represent a large number of subtypes. These four digits form the basis of an ICD-11 code. This can be followed by a more detailed description of the disease, separated by a period (.), using additional numbers and letters. A major change to ICD-10 is the possibility of cluster coding. This allows additional information to be added to the master codes in the form of extension codes. These extensions allow the specification of location and side (right, left, both sides), severity (mild, moderate, severe, life-threatening), temporal course and status (acute, chronic, acute exacerbation, in remission), the cause (infectious agents, medications, substances), etc. These extension codes can be added to the root code following a slash (/) to provide further details about the condition. Multiple root codes can be linked together using an ampersand (&). This is used when one disease has caused another. For example, the root code for bronchial asthma is CA23. The C in the first position stands for chapter 12 – diseases of the respiratory system. The second position (A) describes the block of certain lower respiratory tract diseases in chapter 12. The number 2 in the third position stands for asthma. The number 3 in the fourth position further specifies this as bronchial asthma. There are further subcategories that can be indicated by the digit after the period (.). Thus, CA23.0 stands for allergic asthma, while CA23.1 describes non-allergic asthma. The severity can be added to the base code using an extension code from Chapter X with a slash (/). Here, CA23.02 / XS25 stands for allergic asthma, uncomplicated, with the extension code for “severe”. (World Health Organization, 2022b)

In practice, ICD-11 is used for modern clinical documentation in EHRs. It is “digitally ready” and can be integrated directly into practice software via interfaces. In addition to billing for health services, it forms the backbone of global health reporting. Thanks to its greater level of detail, epidemiological trends – such as antibiotic resistance or new disease phenomena can be observed and compared more uniformly worldwide. (World Health Organization, 2019a)

1.2.4.3 ICD-O-3

The ICD for Oncology, Third Edition (ICD-O-3) is a system developed by the WHO for the standardized coding of neoplasms (abnormal growth of cells). It is primarily used for cancer registration, epidemiology and oncological research. (World Health Organization, 2013)

In contrast to ICD-10, ICD-O-3 is a dual-axis classification system. It combines two classification systems: the topography code (ICD-O-T) and the morphology code (ICD-O-M). The topography code is also included in ICD-10, while the morphology code is absent. In the ICD-O-3, the topography code describes the location of the tumor tissue, while the morphology code refers to the histology, microscopic cell structure and biological behavior of the tumor. (World Health Organization, 2013)

The topography code in ICD-O-3 is based on the anatomical site of the tumor and is represented by a letter followed by a two-digit number. Another number can be added, separated by a decimal point (.). The topography section has been adapted from the malignant neoplasm section of the ICD-10. For example, "C50.3" represents the breast, lower inner quadrant. (World Health Organization, 2013)

The morphology code in ICD-O-3 is based on the histology or microscopic appearance of the tumor cells and is represented by a four-digit number followed by a forward slash and another number. The first four digits describe the histologic term, the fifth digit after the slash is called the behavior code. It indicates whether a tumor is malignant, benign, in situ or uncertain. For example, "8010/0" represents an epithelial tumor, benign. (World Health Organization, 2013)

By combining the topography and morphology codes, ICD-O-3 provides a detailed and standardized coding system for different types of cancer. Accordingly, it is used particularly in cancer registries and many national cancer statistics, where it forms the basis for clinical studies, care and survival analyses, and international comparability of incidence and mortality data. (World Health Organization, 2013)

1.2.4.4 SNOMED-CT

Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) is one of the world's most comprehensive, multilingual medical terminologies. With more than 360,000 logically defined and hierarchically organized terms, it offers the possibility to document detailed clinical information about patients in a standardized and structured manner. SNOMED-CT is used particularly in the context of EHRs, in research, and for storing and exchanging standardized clinical information between different healthcare institutions. (SNOMED International, 2024d)

The SNOMED-CT system is based on a coded vocabulary that includes medical concepts such as diagnoses, procedures, symptoms, medications, anatomical structures, organisms, and more. Each concept is represented by a unique numerical identifier. For example, the code 38341003 stands for “hypertensive disorder". (SNOMED International, 2024b)

A key feature of SNOMED-CT is its ability to enable semantic interoperability between different health information systems. While ICD-10 is more for statistical classification, SNOMED-CT allows for a more detailed, clinically oriented representation of medical facts. This not only improves the quality of patient care, but also makes health data more usable for research, AI applications, and public health. (SNOMED International, 2024c)

SNOMED-CT is maintained by SNOMED International (formerly: International Health Terminology Standards Development Organization (IHTSDO)). The terminology is used in healthcare systems worldwide and also serves as the standard for clinical terminology in EHRs in many countries. (SNOMED International, 2024a)

1.2.4.5 UICC TNM

The TNM classification is the most commonly used classification system internationally for describing the anatomical extent of tumor diseases. It was developed between 1943 and 1952 and is jointly maintained and regularly updated

by the UICC (Union for International Cancer Control) and the AJCC (American Joint Committee on Cancer). (Brierley et al., 2025)

The TNM code consists of three components, each describing a different aspect of tumor spread. The letter “T” stands for “tumor” and describes the local spread of the primary tumor. Depending on the type of tumor, the T category describes either the tumor size or the depth of tumor invasion. Typically, this category is described as T1–T4, with a higher number indicating an increasing size or deeper invasion into surrounding tissue. Additionally, the T category can be described as TX (primary tumor not assessable), T0 (no evidence of primary tumor), or Tis (carcinoma in situ). (Brierley et al., 2025)

The letter “N” stands for “node” and refers to the invasion of regional lymph nodes. N0 indicates no lymph node involvement, while N1–N3 describe increasing involvement in terms of number, size, or location. NX means that the lymph nodes could not be assessed. (Brierley et al., 2025)

The letter “M” represents “metastasis” and indicates the presence or absence of distant metastases. Here, M0 means that no distant metastases are present, while M1 indicates the presence of distant metastases. For some tumor entities, M1 is further subdivided (e.g. M1a, M1b, depending on the location of the metastases). (Brierley et al., 2025)

Prefixes are used to more precisely describe the timing and method of a tumor’s classification. For example, a classification with the prefix “c” (clinical) is based on clinical examination and imaging. The prefix “p” (pathological) indicates that the classification is based on histopathological examination. The suffix “m” refers to the presence of multiple primary tumors. The prefix “y” is placed before the prefixes “c” and “p” and indicates that tumor treatment has already been initiated prior to classification. The prefix “r” refers to recurrent tumors that have reappeared after a period during which the patient was tumor-free. The prefix “a” indicates that the classification was performed following an autopsy. (Brierley et al., 2025)

Based on the values of categories T, N, and M, tumors are classified into stages 0–IV. This classification is tumor-specific and varies depending on the tumor entity. (Brierley et al., 2025)

1.3 Problem Statement

While the digitization of pathology is paving the way for computer-assisted diagnoses, it also presents hospitals and research institutions with enormous logistical challenges. One of the major challenges is the large volume of data. In the field of diagnostics alone, around 60,000 medical orders of individual patients with approximately 120.000 specimens are processed at the Diagnostic and Research Institute of Pathology of the Medical University Graz every year. As part of the digitization of pathology approximately 400,000 whole slide images are generated annually at the institute.

1.3.1 Challenges faced by pathologists and researchers in managing histopathological images

The use of digitized images of slides not only increases the amount of data, but also the complexity of the data. These high-resolution images are a key resource for research, education, and the work on AI algorithms, but they also come with some significant challenges in terms of storage, structuring, and transmission. (Keighley et al., 2023)

It is becoming increasingly important for medical research and progress in the field of pathology to understand the complex data correlations. However, such correlations can only be identified by accessing large, well-structured data sets. In pathology in particular, where image data is supplemented by metadata containing clinical, molecular, or technical information, the effort required for quality-assured data preparation and storage increases exponentially. (Clunie, 2021, Wilkinson et al., 2016)

A key problem lies in the annotation of this data. High-quality annotation is essential for training and validating AI models. However, this requires a great deal of manual effort and time, which is hardly scalable. Prospective annotation by

pathologists is not feasible, either in terms of personnel or finances, even though it would significantly improve the findability, traceability, and reproducibility of the data in the long term. (Korzynska et al., 2021)

In the research work carried out at the Diagnostic and Research Institute of Pathology at the Medical University of Graz, researchers repeatedly encounter the problem that, although a great deal of data is available, these large data sets are currently not easily searchable. In addition, researchers are currently limited to their own data, as they are unaware of any useful data sets that may be available at other centers. The metadata catalog should remedy this situation and thus ensure better accessibility of existing data.

1.3.2 Need for a comprehensive metadata catalog

A metadata catalog or a standardized software interface ensures the standardization of data formats, enables seamless communication between different systems, and enhances the ability of pathologists and researchers to share data with each other. (Anagnostakis et al., 2014)

The need for a comprehensive metadata catalog in digital pathology is critical to ensuring interoperability, data management, and the effective search for requested data sets. By selecting the data fields and thereby standardizing the data formats, it is ensured that the data is comparable between institutions and can therefore be used interchangeably for research projects. Since the proposed metadata catalog is a project involving several Austrian institutes, each individual institute would benefit from a higher number of standardized data sets. The searchability and quick retrieval of data records that are of interest would also be improved by the metadata catalog.

1.4 Objectives

The metadata catalog is designed to bridge the existing gap between image creation and the efficient use of image data. The requirements of end users are the primary focus in this regard.

1.4.1 Aggregate the needs of pathologists and researchers

This work is intended to collect and present the requirements of the various user groups for a metadata catalog of histopathological WSIs. For this purpose, previous work is first searched for data fields that could be of interest to pathologists and researchers. These data fields from various sources are collected and evaluated. The aim is to determine whether the desired data fields can be filled from existing systems and how they can best be recorded in a structured form. This collection is presented to a national consortium of pathologists and researchers in a workshop. During this workshop, a standardized set of data fields for the metadata catalog of histopathological WSIs is defined in collaboration with the participating pathologists and researchers to meet the needs of future users.

1.4.2 Proposal of a minimal set of data fields for the catalog

The minimum set of data fields for the metadata catalog should be selected in such a way that the data fields can be filled in by all partner centers of the national consortium. The aim of the metadata catalog for histopathological WSIs is not to store all possible data relating to a WSI, but to identify the data fields necessary to make a WSI findable within a catalog. The aim of this work is therefore not to create the largest and most comprehensive metadata catalog, but to select a set of data fields that will make the catalog a valuable tool for users in their research work.

To achieve this, basic medical data for individual cases should be included in the catalog. This data can be obtained from EHRs. This basic data includes, among other things, the age of the patients, their gender, and the date of the sample acquisition. In order to protect patient privacy and comply with data protection guidelines, the data can be recorded in aggregated form. For example, the patient's data record is assigned to an age group (e.g., 45-50 years) instead of recording the patient's exact date of birth.

1.4.3 Coding of PAS Xanthos data fields

As already mentioned, approximately 120,000 samples are processed annually as part of diagnostics at the Diagnostic and Research Institute of Pathology in Graz. The corresponding data is currently collected, managed, and stored in PAS

Xanthos, the pathology LIMS. However, many data fields are currently not filled using standardized terminology such as SNOMED-CT, but rather as free text. Free texts are intuitive and easy for humans to write and foremost fulfill the medical purpose to submit the most important information to clinicians as well as to encompass biological diversity, but they are not machine-readable. Encoded data can be compared internationally and can be searched, sorted, or filtered in a targeted manner. These properties enable more efficient use of data in the fields of research. In an ideal setting, data coding would take place during the diagnostic process. This would require a change to the current system at the Diagnostic and Research Institute of Pathology in Graz as well as in most pathology departments worldwide. Therefore, in the course of this thesis, the keywords currently used to fill the data fields will be mapped to international standards. As part of this process, the values used are analyzed, grouped, and then assigned to internationally recognized codes. In cases where direct mapping is not possible, alternative coding proposals are provided.

2 Methods

In the process of this thesis, data fields from diagnostics and from research projects that use a catalog were collected and compared with each other. Data fields from PAS Xanthos, the LIMS at the Diagnostic and Research Institute of Pathology in Graz, were collected as a representative for diagnostics. As examples of the research projects the ADOPT cohort and the CRC cohort were examined. In order to address the field of biobanking, the MIABIS Core biobanking standard and its extensions MIABIS Sample+Donor+Event and MIABIS Digital Pathology were also considered in this thesis.

2.1 Process of designing the catalog structure

A list of proposed data fields will be created based on the data fields from the diagnostics and research projects, which will then be discussed and revised during the workshop. All the data fields compiled during the workshop will be integrated into the hierarchical structure of the cataloging software. This hierarchy is used to link the data fields together. This ensures that records are uniquely identified and can be easily located.

2.2 Ethics proposal

An ethics application was submitted in order to conduct a pilot run with diagnostic data. To this end, a study protocol was drafted that outlines the project's timeline and organizational structure. The pilot project was named "AUTO-IMD" and aimed to automatically index information in order to populate the catalog with structured data. The study protocol describes the inventory analysis of existing systems, coding according to international standards, and automatic mapping using a rule-based tool. The implementation of OpenSpecimen, which is intended to serve as cataloging software, is also described in detail in the study protocol. The entire study protocol can be viewed in the appendix.

2.3 Declaration on the Use of AI-Based Tools

In the process of writing this thesis, "DeepL Translator" and "DeepL Write" (DeepL SE, Cologne) were used on certain occasions for linguistic revision and stylistic

refinement. The purpose of using these tools was limited to improving grammar, style, and readability; the tools did not alter the substance of any scientific statements. Responsibility for the content, structure, interpretation of the results, and scientific statements rests entirely with the author.

3 Results

The result of this thesis is the structure in which the data fields are represented in the cataloging software, the data fields themselves, and the mapping of specific data to medical standards.

The first section of the results describes the existing source system in the diagnostic workflow of the Diagnostic and Research Institute of Pathology in Graz, as well as the data fields used in previous projects. These include PAS Xanthos, the ADOPT Cohort – a project in which 10,000 CRC cases were collected across Europe, the CRC Cohort – an industry collaboration with Google, in which 300,000 slides from the biobank were digitized, as well as the three MIABIS components: MIABIS Core, MIABIS Sample+Donor+Event and MIABIS Digital Pathology, which were developed with the support of the Medical University of Graz.

The second part of the results describes how the final proposed dataset is derived from the previously analyzed potential data fields. To this end, the previously identified fields are consolidated into an initial proposal. Based on this proposal, a workshop is being held to assess the importance and extractability of the data fields. The results of this workshop are then compiled into a final version of data fields.

The third section describes how these data fields can ultimately be populated. Among other things, this involved SNOMED coding of the medical terms from the fields transferred from PAS Xanthos.

3.1 Laboratory information management system for diagnostic workflow

PAS Xanthos is a workflow-controlled system that covers all processes at the Diagnostic and Research Institute of Pathology in Graz. An “order” is created in the sample reception area, which runs through the entire process. Samples from various senders are processed in the sample reception area. The institute receives the sample itself as well as the corresponding request form (see Figure 2).

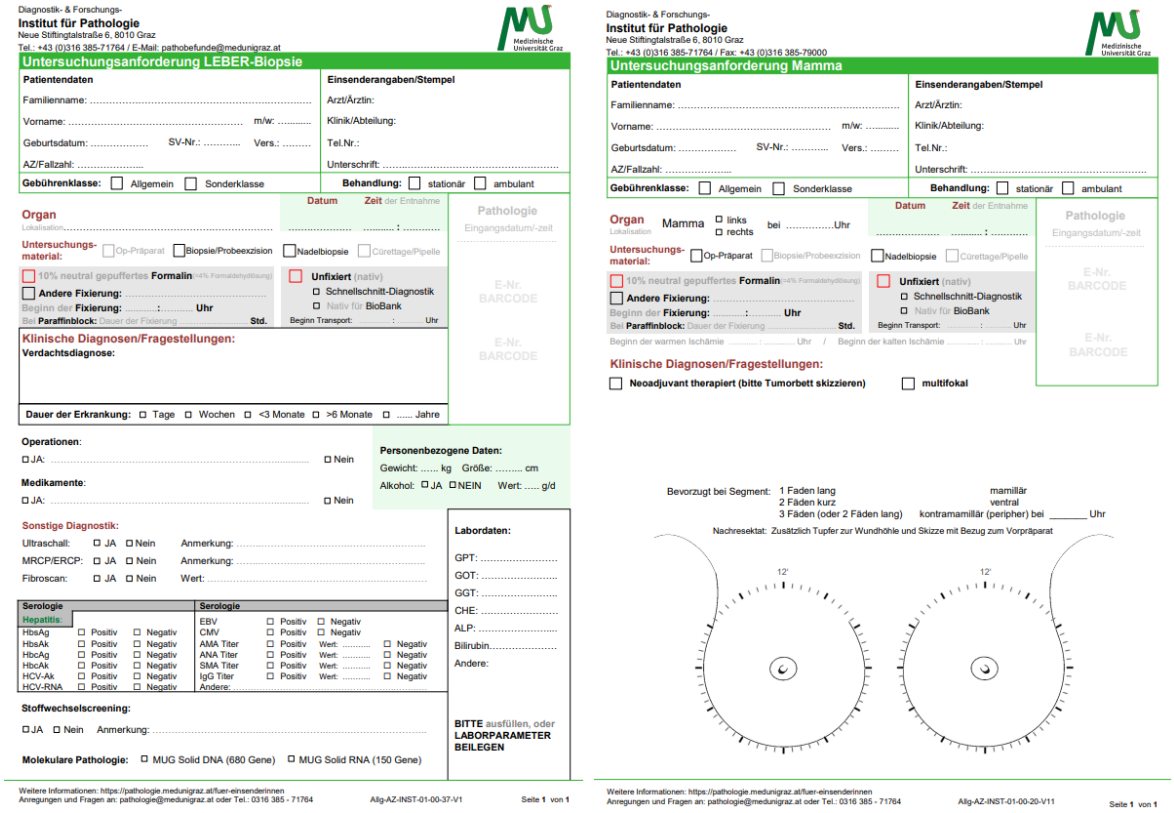


Figure 2: Form "Untersuchungsanforderung" for liver biopsy and mamma

The initial registration of the samples is carried out using an electronic stamp. Each submission is given a unique ID (E-number), which is a combination of the current year and a consecutive six-digit number. This E-number is printed on the sample containers as text and as a datamatrix code, which looks similar as a QR-code. The datamatrix code generated from this has the form: 2025123456, for example. A so called “electronic stamp” is used to record the following data for each entry: The number of containers, the categorized sample material received and the collection method of the sample material (e.g. biopsy or surgical resection). There is a predefined list for capturing the material, which can be seen as an example on Figure 3 (electronic stamp).

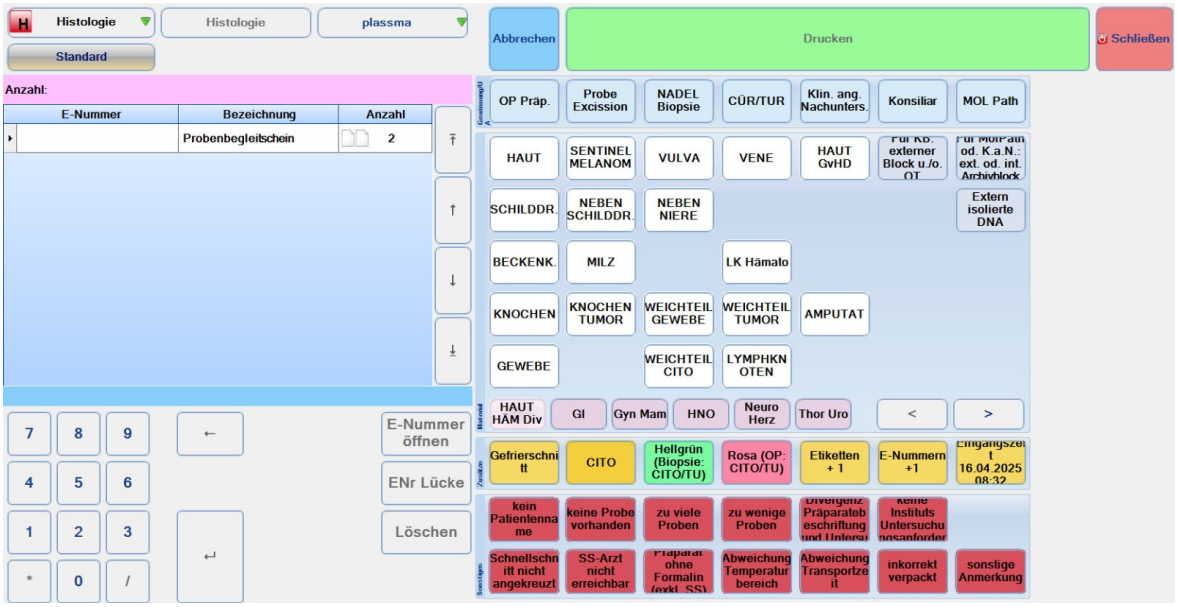


Figure 3: Electronic stamp in PAS Xanthos

A breakdown of the collection method is made between surgical specimen, specimen excision, needle biopsy, curettage, follow-up examination, consultation and molecular pathology request. After the electronic stamp, the request form (“Untersuchungsanforderung” Figure 2) is recorded. The patient data is compared with the data entered and additional information (e.g. further material details) is transcribed. The request form is stored as a digital image in the LIMS. The samples, which have now been recorded in the system and provided with a code, leave the sample reception area and are taken to the macroscopy laboratory. In the first step, the number of physically received sample containers is compared with the number of samples indicated on the accession form. In the next step, the samples are cut so that they fit into a plastic cassette, which is then used to create the block. In the macroscopic laboratory, the pathologist describes the sample in terms of origin, size, color, shape and consistency. Based on these observations, the pathologist cuts the whole sample into appropriate pieces and places representative samples or the whole tissue, depending on the medical need, in the labeled plastic cassettes. These are recorded as blocks in the so-called “Xanthos laboratory book”, which is a table with all blocks, slides, whole slide images and all further laboratory analyses. The barcode printed on these plastic cassettes is automatically assigned by the LIMS. The accession number, which the sample has already received in the sample reception area, is extended here by a further 2 digits. Separated by “/”, the block ID is added to the E-number as a two-digit

number, starting with 01. This results in a number in the format: 2025123456/01. These processes are documented using dictation and later appear as “Macroscopic description” in Xanthos. The next step is the histological processing of the blocks. The tissue in the plastic cassettes is dehydrated, then the blocks are filled with paraffin wax. At the same time, the labeling is printed on the slides using PAS Xanthos. Which slides belong to which block is also stored in PAS Xanthos. The block number already assigned is extended by 2 digits for this purpose. It is again separated by “/” and the slide ID is entered as a two-digit number. The format is therefore: 2025123456/01/05 (year 2025, consecutive number 123456, block ID 01, slide ID 05). The tissue sections are now mounted on the pre-printed slides in the histology laboratory. To do this, a thin section is cut from the block using a microtome and applied to the glass. This allows 1-n slices to be placed on one slide. For surgical specimens, two slices are usually placed on one slide, for biopsies up to four. The number of slices on a slide is not documented in PAS Xanthos. The slides with the applied pieces of tissue are now stained according to specified protocols. Various staining techniques and IHC methods are available for this purpose. The stained slides are sealed with a cover glass. This is necessary to protect the stained tissue pieces and to enable examination under the microscope and digitization in the slide scanner. The finished slides are now taken to the digitization laboratory. There they are digitized with a so-called WSI scanner. These digital images of the slides are stored in the PAS Xanthos lab book in a similar way to the blocks and slides. The ID of the slide is again extended by 2 digits. The two-digit scan ID is added, separated by “/”. This results in the following format: 2025123456/01/05/02 (year 2025, consecutive number 123456, block ID 01, slide ID 05, scan ID 02). After successful digitization, the corresponding case is assigned to a pathologist for diagnosis. The pathologist can request additional blocks, slides and scans if this is required for the diagnosis. These repeat orders are possible via the PAS Xanthos laboratory book (see Figure 4). The following data is recorded in PAS Xanthos as part of the diagnostic process: Histological description, molecular pathological description and diagnosis.

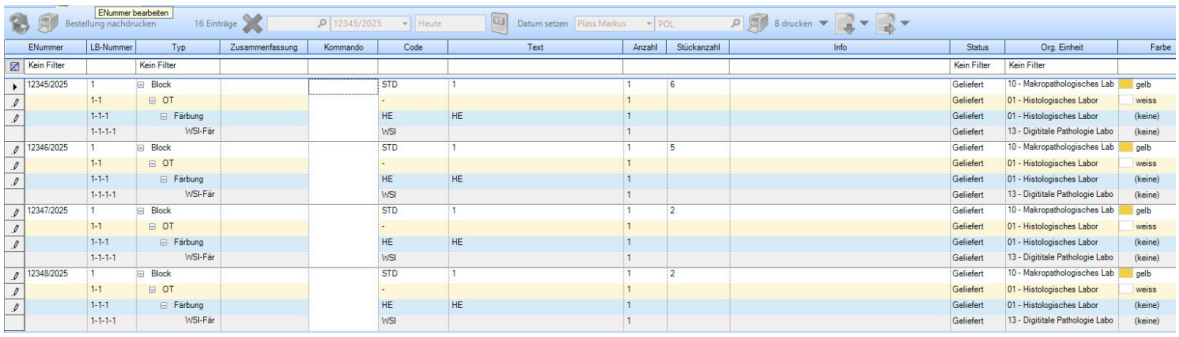


Figure 4: The laboratory book in PAS Xanthos

3.2 Data fields out of diagnostic system

PAS Xanthos covers all processes at the Diagnostic and Research Institute of Pathology in Graz, from the sample reception to the diagnosis, transmission of findings and invoicing. Accordingly, PAS Xanthos contains an enormous amount of data fields to ensure the management of these processes. A survey and listing of all these fields would go beyond the scope of this thesis. Therefore, only the relevant data fields from the field of diagnostics that could be of importance for the metadata catalog were collected here.

The data fields collected in PAS Xanthos provide information about the patient and the request. This includes identification numbers, date of birth, gender of the patient, as well as information about the request, such as the date of receipt and identification numbers of the request.

Sample and material management is also mapped in this system. It is possible to trace where the samples come from, how they were acquired and from which organ they were obtained.

The essential part of pathology, the diagnostic results, is also documented in PAS Xanthos. Findings from macroscopy and microscopy and diagnosis are unstructured texts, TNM staging, and ICD-O codes are examples of structured data fields used in this area.

The sample's journey through the laboratories is also recorded in the system. The tissue blocks and slides are described in the laboratory logbook, which also records the method used to fix the samples and the labeling of the containers.

The data fields collected from PAS Xanthos, including name, format, and information about each analysed data field, are shown in the appendix.

3.3 Data fields out of research projects

Additional to the data collected during diagnostics, large volumes of data are also generated in the context of research projects, each of which is defined according to project-specific requirements. In the course of this thesis, metadata fields from two large cohorts are considered, which were created in the scope of projects at the Diagnostic and Research Institute of Pathology in Graz: The ADOPT cohort and the CRC cohort.

In addition to these projects, the metadata fields that are intended to form a standard for biobanks are also described. The data fields described there were defined and established by the Medical University of Graz in cooperation with the “European Research Infrastructure for biobanking and biomolecular resources” (BBMRI-Eric), and other partner organizations. The projects considered here are “Minimum Information About BIobank data Sharing” (MIABIS) and “Cancer Image Europe” (EUCAIM).

As part of the EUCAIM project, the existing MIABIS data field set was expanded to include the category of digital pathology. To this end, the EUCAIM project team submitted an initial proposal for data fields, which was expanded and discussed in joint meetings. An EU-wide survey was then created and evaluated. The results of the survey were incorporated into the selection of data fields. This state of the data field selection was presented to the member countries of BBMRI-Eric and, following their feedback, MIABIS was expanded to include the data fields for digital pathology.

3.3.1 ADOPT Cohort

The ADOPT cohort consists of 10.000 colorectal cancer cases all around europe. The data fields used to describe this cohort were selected with a focus on clinical routine and to track the patient's history from diagnosis to survival status.

The catalog contains the sections “Diagnostic exam”, “Histopathology”, “Histopathology - TNM”, “Histopathology - UICC staging”, “Histopathology - WHO classification”, “Molecular markers”, “Molecular markers - KRAS mutation status”, “Patient Data”, “Pharmacotherapy”, “Radiation therapy”, “Response to therapy”, “Sample”, “Surgery”, “Targeted therapy” and “Vital status and survival information”.

The “Diagnostic exam” section documents which diagnostic procedures were performed and whether image data is available for the corresponding procedures. Computer tomography examinations, magnetic resonance imaging, imaging of the lungs and liver, and colonoscopies are displayed here.

The areas describing histopathology include information about tumors such as the location of the primary tumor and metastases, tumor morphology, and the availability of digital scans of the slides. Other histopathology categories describe the classification of tumors according to TNM, staging according to UICC, and grading according to WHO.

The sections on molecular markers describe the presence of various mutations, tests for microsatellite instability, defects in mismatch repair genes, and the risk of HNPCC (Lynch syndrome).

The “Patient Data” section includes the age at the time of diagnosis, the date of diagnosis, gender, participation in clinical trials, and the time between the initial diagnosis and potential recurrences.

The sections “Pharmacotherapy,” “Radiation therapy,” “Response to therapy,” “Surgery,” and “Targeted therapy” describe what types of therapies the patient

received, when these therapies took place, and how the patient responded to the therapies.

The “Sample” category provides information about the identification number, material type, and preservation method of the biological sample.

The section “Vital status and survival information” records whether the patient is still alive, when this information was last checked, and, in the event of death, the cause of death.

The data fields of the ADOPT cohort, including label, field name, data type, and description, are listed in the appendix.

3.3.2 CRC Cohort

The CRC cohort contains around 5.000 colorectal cancer cases from Graz. This set of data fields is designed to link pathological findings with clinical patient data.
In this cohort, the identification numbers of the individual hierarchy levels are mapped as UUIDs. These include the patientID, caseID, blockNumber, slideID, and scanID.

The biological sample is described in the “material” category. This includes information about the contents of the sample container and information about the block, such as the block ID, a description of the block contents, the embedded material, and the localization.

The “scan” category contains technical details about the scan of the slides. This includes the slideID, scanID, scannerID, time of scan, magnification, file format, and work package.

The “patient” section provides information about patient data. This includes, for example, the patient ID, gender, survival status, and the ICD-9 and ICD-10 codes for the cause of death.

The diagnosis is described in more detail in the “diagnosis” section. Here, the diagnosis is shown in coded form, such as ICD-10 code and ICD-O code. The classification of the tumor according to the TNM system is also shown here, as are the staging, grading, resection margin status, and vascular or lymphatic invasion. The age at diagnosis and survival in days and months are also described here.

In addition, system information is also recorded under the “general” category. Here, the version of the JSON file and the timestamp of the file creation of the JSON file are recorded.

The complete list of CRC Cohort data fields, including structure, description, and examples, can be viewed in the appendix.

3.3.3 MIABIS Core

MIABIS is a standardized framework designed to enhance the sharing and interoperability of biobank data. The first version was launched in 2012 by Biobanking and BioMolecular Resources Research Infrastructure of Sweden (BBMRI.se). (Norlin et al., 2012)
Since then, MIABIS has undergone several updates to accommodate the evolving needs of biobanks, particularly in harmonizing data across various platforms and enhancing semantic interoperability. The latest version, MIABIS Core v3, expands its scope to include data-driven biobanks and nonhuman sample collections, thereby improving usability and facilitating better data discovery for researchers. (Eklund N, 2024)

MIABIS provides a common language for biobanks, ensuring that data is described uniformly across different systems. It is designed to be extensible, allowing for the integration of additional data types and attributes as biobanking practices evolve. MIABIS facilitates the integration of biobanks on a global scale, promoting collaboration and resource sharing among researchers. (Merino-Martinez et al., 2016)

MIABIS Core v3 is the latest version of the standard and focuses on four main areas: “Biobank”, “Collection”, “Research resource”, and “Network”.

The “biobank” section describes the institute’s framework. It includes fields for identifying the biobank, name, acronym, website, legal basis, contact persons, capabilities, and the standards according to which the biobank is certified.

The “collection” category describes a specific group of samples within a biobank. It contains basic data about the collection, such as name, identification number, acronym, description, contact information, and URL. The samples in this collection are described in more detail, including information about the origin of the samples, gender of the individuals, age range of the donors, type of samples, diseases of the collective, and storage temperature of the samples. Information on study design, access authorization, number of individuals included, inclusion criteria, and publications based on the collection are also provided.

“Network” refers to associations of several biobanks or collections. This section explains the structure of the network, including its name, identification, acronym, URL, country, legal entity, contact information, description, network members, topics of collaboration, and network type.

The “research resource” category can combine material from different collections or biobanks for use in a specific research project. The data fields used correspond to those in the “collection” category. The “research resource” category can be used to flexibly describe compiled materials, even if they are not stored in the same collection.

The individual data fields of MIABIS Core v3, including attribute code, attribute name, allowed data type, description, constraints, and allowed values, can be found in the appendix.

3.3.4 MIABIS Sample+Donor+Event

In the field of clinical patient care and medical research, detailed descriptions of biological samples are important. To enable these biological samples to be represented effectively, the MIABIS standard has been expanded to include the add-on “Sample+Donor+Event”. (Eklund et al., 2020)

The standard defines the areas “Event”, “Sample donor”, and “Sample”. The “Event” category describes what happened and when. It refers to a sample or sample donor. A unique identification and the time of the event is provided. The time is described either as the exact date, including the donor's date of birth, or as the age at the time of the event.

The “Sample donor” section describes the individual from whom the sample was taken. Each donor is given a unique identification number. The donor's gender and date of birth are recorded. The type of data records that exist for the donor are also noted.

The “Sample” section provides information about the biological material or its digital images. A unique identification is assigned to the sample. The type of sample, storage temperature, date of creation, anatomical origin with various sub-attributes, diagnosis, access authorization, and origin of the sample are described here.

The complete list of data fields for MIABIS Sample+Donor+Event, including attribute code, attribute name, description, data type, allowed values, constraints, and cardinality, can be viewed in the appendix.

3.3.5 MIABIS Digital Pathology

MIABIS is therefore a standard that focuses on biobanks and their samples and associated data. However, due to the advancing digitization of medicine, there is now an increasing amount of digital data that needs to be stored, managed, and enhanced with additional data in the same way as physical samples. As part of the Cancer Image Europe (EUCAIM) project, the MIABIS standard was therefore

expanded to include attributes for digital pathology. For this purpose, a minimal set of essential attributes for the digital samples, the WSIs, was defined to ensure standardized data sharing and interoperability.

MIABI's digital pathology includes three categories “Assay”, “Scan”, and “File”. The “Assay” section documents what happens before the sample is scanned. The container is described by identification and type. The biological sample is also described, including what tissue it consists, what cell types are present, and what pathogens were tested for. Staining methods, details on antibodies, dyes, biomarkers, and their localization in the cell, as well as special procedures such as fluorescence and FISH analyses, are also documented.

The “Scan” category deals with the technical process of scanning and the data that is generated in the process. This category includes, among other things, scan identification, the timestamp of file creation, information about the scanner and its software, information about Z-stacking, the scan area, the magnification used during scanning, the resolution, the illuminating method, and error rates.

The “File” category is all about file attributes. In addition to the file identifier, it also records the format, file size, image dimensions, channel description, color space, compression method, availability of preview, label and annotations, as well as personal information.

All data fields available in MIABIS Digital Pathology, including attribute code, attribute name, description, data type, allowed values, constraints, and cardinality, can be found in the appendix.

3.4 Design of the Metadata Catalog

Within the framework of the workshop, it was agreed to adopt a participant-centered approach to sample management. Figure 5 illustrates the hierarchical data structure implemented in OpenSpecimen. It describes the logical path from a participant to a digital whole-slide image.

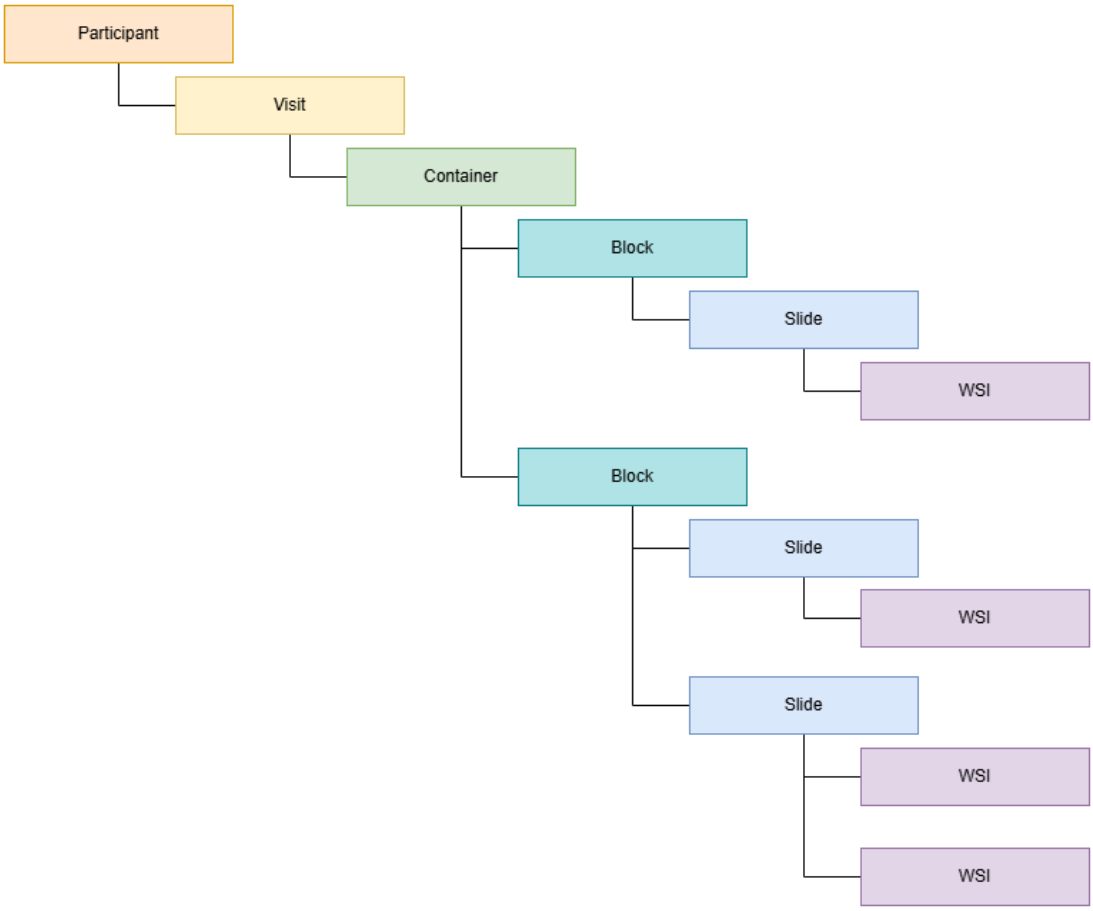


Figure 5: Data field structure

The process begins at the top level with the participant (patient), who is assigned a specific visit (transfer of biological material to the Department of Pathology). The collected medical material is placed in a container. From this source material in the container, one or more tissue blocks (usually embedded in paraffin) are prepared in the laboratory. These blocks then serve as the basis for preparing slides, with multiple thin tissue sections able to be cut from a single block. The final step in the process is digitization: high-resolution scans are created from the physical slides and stored in the system as whole-slide images. Sometimes WSIs have to be repeated due to quality issues, therefore WSI may appear more than once. This structure illustrates the increasing granularity of the process, in which a single sample can, through laboratory procedures and digitization, generate a multitude

of data sets and digital images - all of which remain fully traceable back to their origin.

3.5 Initial draft of data fields for the workshop

Based on the results of the analyses of the data fields from PAS Xanthos, the ADOPT cohort, the CRC cohort, and MIABIS (Core, Sample+Donor+Event, Digital Pathology), an initial set of data fields was developed. These are listed in Table 1.

Table 1: First draft of data fields

Branch	Feature
Governance	Location of stored slide
Governance	Access condition (e.g. non commercial, restricted till...date, personal restrition)
Technical WSI features	Type of Scanner (manufacturer, device type, serial number)
Technical WSI features	Resolution
Technical WSI features	Scan date and time
Technical WSI features	Brightfield, fluorescence
Technical WSI features	Compression quality level; JPEG, TIFF, etc.
Technical WSI features	Image format (NDPI, SVS, MRXS, DICOM, etc.)
Technical WSI features	Firmware version
Technical WSI features	Z-stacked
Slide features	Relation to block (slides from FFPE block)
Slide features	Relation to block (order of slides)
Slide features	Fixation type (formalin, frozen section)
Slide features	Block label (TU, 1, My, E, Po, Para, AR, OR, LY, etc.)
Slide features	Specimen type or shape of section (biopsy, resection, TMA, cytology)
Slide features	Size/area of tissue (excluding/including test tissue on IHC slides?)
Slide features	Number or presence of parallel sections
Slide features	Tissue type on specific WSI (fat, lymphatic tissue,

	fibrotic tissue, cartilage, bone, squamous epithelium, glandular epithelium (more subtypes?), smooth muscle, striated muscle, peripheral nerve, central nervous system)
Slide features	Slide annotation available (Y/N)
Slide features	Type of slide annotation
Slide features	Filetype of slide annotation
Stain features	Stain group (HE, special stain, IHC, fluorescent)
Stain features	Type of stain abbreviated including fluorescent (and long version); e.g. EvG (Elastica van Gieson), Ki-67 (Ki-67 clone MIB1)
Case features	Which slides belong together: i) to the same diagnostic event/case; ii) to the same patient over a long time period; iii) pre-defined research collection
Case features	Human / non-human (type of species animal)
Case features	Age in years at the diagnosis date
Case features	Sex (f/m/d/i/unknown/non applicable)
Case features	Oragan / material full text - on case level (example: right hemicolectomy)
Case features	Material abbreviated / coded (example: colon)
Case features	Year of diagnosis (grouped?)
Case features	Overall diagnosis full text
Case features	Coding T N M G, TNM ed., prefix
Case features	ICD-O-T, ICD-O-M code-value
Case features	Overall diagnosis abbreviated - disease concepts (neoplastic, metabolic, etc.)
Case features	Specific diagnosis on WSI
Case features	Biomarker (IHC, genetics): code - value
Follow-up data	Overall survival
Follow-up data	DFS (recurrence)
Follow-up data	Medication; CTX/RTX/immunotherapy (Y/N), neoadjuvant, adjuvant, second line, etc.

Follow-up data	Flag for other clinical data available on request (WSI usage fees for clinical annotation)
Follow-up data	Flag for other pathological data available (genetic data, IHC data)

3.6 Workshop

The outcomes of the analysis of the data fields from the diagnostics and the projects were compared with each other. The results of this comparison were summarized, processed and used as an initial proposal for the workshop “Digital Pathology: Interuniversity Infrastructure - Metadata Catalog”. This workshop took place on September 18 and 19, 2023 at the Medical University of Graz. Researchers, pathologists and technicians from the founder institutions, namely Johannes Kepler University Linz, the Medical University of Innsbruck, the Medical University of Vienna, the Veterinary University of Vienna and the Medical University of Graz took part in the workshop which was dedicated to planning the “Interuniversity Infrastructure - Metadata Catalog” project. The aim of the project is (i) to continuously create and store large volumes of sectional preparations as digital images, (ii) to collect and merge metadata from these images in a standardized way and (iii) to store them in a cross-location, anonymized data catalog. This catalog should be linked to the European research infrastructure (EOSC). The image data linked to the metadata should be available for AI research projects in an interuniversity infrastructure, while the governance rights remain with the respective contributors, an important point to all participants. The data fields proposed on the grounds of the analysis described above were discussed individually in the plenum. Data fields were removed from the original collection and new proposals were added, depending on their relevance to the most likely research questions and scenarios. Thoughts and comments on existing data fields were also incorporated. In addition to the selection of data fields, it was also noted how important the integration of the individual data fields in the catalog is for future users. Therefore, another objective of the workshop was to label the data fields depending on relevance. The categories for this were the letters A (very important), B (moderately important) and C (interesting, but not that important). The difficulty of collecting the contents of the individual data fields is also an

important point to consider. For this reason, the data fields were given numbers that indicate how easy or difficult it is to collect the data. The numbers 1 (easy to collect), 2 (moderately difficult to collect) and 3 (very difficult to collect) were assigned.

Table 2: Importance and difficulty of filling in the data field

Importance of the data field		Difficulty of filling in the data field	
A	Very important	1	Easy to collect
B	Moderatly important	2	Moderately difficult to collect
C	Not that important	3	Very difficult to collect

The insights gained regarding the importance and the difficulty of collecting the individual data fields were considered in the final selection of the data fields.

3.7 Final set of data fields

During the workshop, the results were discussed and the data fields were revised. It was agreed that the data fields would be embedded within a participant-centric structure. Within this structure, the data fields are embedded in OpenSpecimen and linked to one another.

The data fields for the metadata catalog are organized hierarchically. While the data fields are named and described in English, they are populated with content in German during the diagnostic process.

In the following section, references to the source systems refer to the original field names in the respective source. In the case of PAS Xanthos, they refer to the field names in the database. For example, the field “dbo”.”Patient”.”idPerson” in the PAS Xanthos database refers to the schema “dbo”, the table “Patient” and subsequently to the attribute “idPerson”.

3.7.1 Level 1 – Collection protocol

Level 1 represents the collection protocol, which describes the collection in OpenSpecimen and defines its content. The collection protocol defines the framework within which samples for a project or catalog are collected, stored, and

processed. It is essential for transforming the complex diagnostic process into a digital workflow. It includes key administrative details, such as the name of the collection, the people responsible, and the participating institutions. The collection protocol ensures that data can be collected consistently over the years, that standardized standards can be met, and that data quality is maintained.

3.7.2 Level 2 - Participant

Hierarchically viewed, “Level 1 - Collection protocol” is the parent category of “Level 2 – Participant”. The data fields at Level 2 describe the individual from whom the biological sample originates. The organism described here is referred to as the “participant”.

The **“Participant ID”** serves as a primary key to consistently merge all information, samples, and results for an individual. The participant must be uniquely identifiable, even when data from different sources is merged. The “Participant ID” data field represents the unique identifier of the individual from whom the biological sample originates. The data type is defined as text; an example of how this data field might be populated would be “115245”. The field is mapped differently in the source systems. In PAS Xanthos, the Participant ID is stored as an integer in the field "dbo"."Patient"."idPerson", in the CRC Cohort as a UUID under the field "patientID" and in MIABIS Sample+Donor+Event as text in the data field “MIABIS-SAMPLEDONOR-01”. The data field does not appear in the ADOPT Cohort.

The date of birth is primarily used to calculate age at the time of certain medical events. Age is a significant factor influencing biological processes and disease progression. In addition, the date of birth provides a further identity check, thereby ensuring data integrity within various systems, among other things. The **“Date of birth”** field records the participant's date of birth. It is stored in date format, for example “01.01.2000”. In PAS Xanthos, this information is stored as a datetime under the field "dbo"."Patient"."Geburtsdatum", and in MIABIS Sample+Donor+Event it is also stored as a datetime in the field “MIABIS-SAMPLEDONOR-04”. The date of birth is not displayed in the ADOPT Cohort and the CRC Cohort.

Recording the date of death is important for clinical research. This information can be used to perform survival analyses and investigate the long-term success of therapies. Without this date, it is impossible to determine whether a patient left the study because the therapy was successful or because they deceased. The data field **“Date of death”** is used to document the date of death of a participant. In the underlying data structure, this field is defined as data type “date.” An example of a date of death would be “31.01.2000”. PAS Xanthos displays this information in datetime format in the field "dbo"."Patient"."Sterbedatum". The date of death is not recorded in the ADOPT cohort, the CRC cohort, or in MIABIS.

The recording of the biological sex of the participant is an important variable in medical research. The onset of disease, symptoms, reactions to medication, biological markers, and much more differ greatly between the sexes. The **“Sex”** data field is used to record the biological sex of the participant. The field is filled in by selecting from a list of predefined values. Possible values are “f”; “m”; “i”; "unknown"; “non applicable”. The sex is described in all source systems. In PAS Xanthos, it is recorded as an integer and stored in the data field "dbo"."Patient"."Geschlecht". In the ADOPT cohort, the information is mapped as a fixed value set and stored in the “SEX” field; in the CRC cohort, the information can be found in the "gender" field as a fixed value set. MIABIS Sample+Donor+Event maps the information as a fixed value set in the “MIABIS-SAMPLEDONOR-02” field.

The social security number serves as a unique identifier in the national context. This means that every participant in Austria can be clearly identified. Hashing is used to convert the social security number into a format that complies with data protection regulations. The data field **“Social security number”** refers specifically to the participant's Austrian social security number. It is stored in the data type "Number"; an example would be “1234310100”. The social security number is only mapped in PAS Xanthos across all source systems. There, it is mapped as text in the field "dbo"."Patient"."SVNR". The ADOPT cohort, CRC cohort, and MIABIS do not cover the social security number.

The fields **“Origin”** and **“Origin detail”** are necessary to define the biological source of the data. The distinction between human and animal samples in the system is particularly important for database queries and the creation of research cohorts. The **“Origin”** data field describes whether a sample comes from a human or an animal. When filling in the field, it is possible to select from “human”; “non-human”. While this field does not appear in PAS Xanthos, the ADOPT cohort, or the CRC cohort, it is defined as a fixed value set in MIABIS Sample+Donor+Event in the field “MIABIS-SAMPLE-12”. If the source of the sample is an animal rather than a human, this is specified in more detail in the **“Origin detail”** field. This field does not exist in any of the source systems analyzed and is filled in with text, for example “Mouse”.

A table containing details about the individual data fields for this level and their availability in the source systems is included in the appendix.

When compiling the final dataset, care was taken to design it in such a way that no direct identifiers are required. The IDs used, such as patient IDs, can be pseudonymized via a mapping table when exported from PAS Xanthos.

The assignment of a non-traceable key would, in principle, be possible for potential dataset extensions, for example where a patient has multiple cases. Such a key could also be used in cases where access to the original data is required for a research project. However, this is not recommended in this instance. For all dates, the day and/or month may be omitted if necessary to make direct traceability more difficult. Age groups have not been used in order to maintain the ability to merge data and perform various search queries.

3.7.3 Level 3 – Visit

The hierarchical level “Level 2 - Participant” serves as the parent level for “Level 3 - Visit”. The data fields in this level describe a medical event, which is referred to here as a “visit”.

The **“Parent participant ID”** field is used to link the "visit" to the higher-level layer “participant”. This assigns a visit to a specific participant.

The **“Request ID”** is used for administrative traceability of requests or orders within the diagnostic process. It enables all materials and files belonging to a specific order to be clearly identified. This information is stored as text, for example “123456”. In PAS Xanthos, this is represented as an integer under "dbo"."Zuweisung"."Anforderungsnummer", while MIABIS Sample+Donor+Event lists it as text under "MIABIS-EVENT-01". This field does not exist in the ADOPT and CRC cohorts.

The **“Date of event”** field makes it possible to classify medical data chronologically and reconstruct the course of the disease precisely. It forms the basis for any longitudinal analysis of the patient's history. The field documents the date of a specific event in the date format. An example would be “31.01.2000”. In PAS Xanthos, it is listed as "dbo"."Zuweisung"."Eingangsdatum", in the ADOPT cohort as “DATE_DIAGNOSIS”, and in MIABIS Sample+Donor+Event as "MIABIS-EVENT-02", each in datetime format. It is not available in the CRC cohort.

The information about the **“Age in years at the diagnosis date”** is important for clinical research, as the age at diagnosis often has prognostic significance and is used for grouping patient populations. This is a calculated field defined as a natural number (example: “87”). While PAS Xanthos does not directly include this field, the information is available in the ADOPT cohort as a natural number in the field “AGE_AT_PRIMARY_DIAGNOSIS”, in the Gorilla Cohort as an integer in the field “ageAtDiagnose”, and in MIABIS Sample+Donor+Event as an integer in the field “MIABIS-EVENT-03”.

The appendix includes a table containing details about the individual data fields for this level and their availability in the source systems.

3.7.4 Level 4 – Container

The hierarchical level “Level 3 - Visit” serves as the parent level for “Level 4 - Container”. The information stored in this layer describes the biological sample, its origin, and the information that can be obtained from its analysis.

The **“Parent request ID”** field is used to link the "container" to the higher-level layer “visit”. This assigns a sample to a specific visit.

The unique identification of a sample is a basic requirement for full traceability in the diagnostic process. Without this identifier, it would no longer be possible to reliably assign analysis results to the original patient, which would jeopardize patient safety. The **"Container ID"** data field serves as a technical identifier for the primary sample and is defined as a text data type with the example value “2024000001”. In PAS Xanthos, the field "dbo"."MaterialEingang"."ENummer" is maintained in integer format, while CRC Cohort uses the name “containerNumber” and stores the identifier as a UUID. MIABIS digital pathology stores the identifier as text under “MIABIS-DIGITALPATHOLOGYASSAY-02”. This information is not available in the ADOPT cohort.

In addition to the technical ID, a human-readable label is necessary to minimize confusion during manual handling of sample containers. The **“Human readable container ID”** field serves this purpose and uses identical sample values as the technical ID. In PAS Xanthos, the CRC Cohort, and MIABIS digital pathology, this field is available in the same way as the container ID. It is not maintained in the ADOPT cohort.

To provide the biological context, it is necessary to specify which organ the majority of the sample comes from. To ensure that datasets are internationally comparable for research purposes, the material is recorded as a SNOMED-CT code in the **“Material SNOMED-CT ID”** field, for example “128169007”. While the field is not displayed in PAS Xanthos, the ADOPT cohort, or the CRC cohort, MIABIS Digital Pathology refers to it as a predefined list under the field “MIABIS-DIGITALPATHOLOGYASSAY-05”.

In addition to the material’s coding, the **“Material”** and **“Material description”** fields allow for a textual description of the tissue. The **“Material”** field provides a human-readable description of the material, which is significantly more convenient in everyday diagnostic practice than a coded format. This is a field that is filled with free text, such as “LIVER”. In PAS Xanthos, this information is stored as free text

in field "dbo"."Material"."Bezeichnung"; in the ADOPT cohort, a predefined list is available, and its values are stored in field “SAMPLE_MATERIAL_TYPE”. In the CRC cohort, the information is stored as free text in field “materialEmbeded”, and in MIABIS Sample+Donor+Event, it is stored as a predefined list in field “MIABIS-SAMPLE-05”. The **“Material description”** field allows for a more detailed description of the tissue using free text, for example, “2 liver punch biopsy cylinders”. In PAS Xanthos, this information is stored as free text in the field "dbo"."MaterialEingangBefundtext"."Text”; the ADOPT cohort displays it as a predefined list in the field “SAMPLE_MATERIAL_TYPE”, while the CRC cohort displays it as free text in the field “materialEmbeded”. This information is not available in MIABIS.

Documentation of the **“Extraction method”** is necessary to provide a more detailed description of the sample. The type of sample collection and the form of the sample determine what diagnostic conclusions can be drawn from it. The field is populated with predefined values: “Conventional smear”; “Conventional smear with a cytology brush”; “Fluid sample”; “Smear or impression (on a slide)”; “Surgical specimen”; “Needle biopsy”; “Excisional biopsy”; “Autopsy”; “Curettage, pipelle biopsy, or transurethral resection (TUR)”. In PAS Xanthos, this information is stored as free text in the "dbo"."Gewinnungsart"."Bezeichnung" field, whereas it is not available in the ADOPT cohort, the CRC cohort, or in MIABIS.

The **“Macroscopic description”**, **“Histological description”** and **“Diagnosis”** fields form the core of the pathological report. Their purpose is to provide a detailed description of the pathological condition and thus serve as the basis for clinical decisions. These fields are therefore filled in with free-text entries.
An example for the **“Macroscopic description”** field is “The specimen consists of a segment of tissue measuring 4.5 × 3.0 × 2.0 cm. On the cut section, a firm, whitish, ill-defined tumor measuring 1.8 cm in greatest diameter is identified.”, while the **“Histological description”** field could, for example, be filled with “Sections show an invasive epithelial neoplasm composed of irregular glandular structures infiltrating the surrounding stroma. The tumor cells display moderate nuclear atypia with enlarged, hyperchromatic nuclei and prominent nucleoli. Mitotic figures are readily identified. The surrounding stroma shows desmoplastic

reaction. No definite lymphovascular invasion is seen in the examined sections.”. The contents of both fields are stored in PAS Xanthos in the "dbo"."MaterialEingangBefundtext".”Text” field; they do not appear in the ADOPT cohort, the CRC cohort, or MIABIS.
The **“Diagnosis”** field is also filled with free text; an example would be: “Invasive moderately differentiated adenocarcinoma”. This field is also stored in PAS Xanthos in the "dbo"."MaterialEingangBefundtext".”Text” field; in MIABIS Sample+Donor+Event, this information is stored as a predefined list in the “MIABIS-SAMPLE-10” field. This information is not available in the ADOPT cohort and the CRC cohort.

The recording of the fields **“Molecular pathological diagnosis”** and **“Molecular pathological description”** is becoming increasingly important due to the growing use of targeted therapies, as patients' genetic profiles determine the choice of treatment. These fields document the results of molecular biological analyses as free text. An example of how to enter data in the **“Molecularpathological diagnosis”** field would be: "Molecular analysis demonstrates the presence of a pathogenic KRAS mutation (c.35G>A, p.G12D). No mutations were detected in NRAS or BRAF. The findings are consistent with a RAS-mutated tumor profile.“, while the content of the **”Molecularpathological description“** field could be ”DNA extracted from tumor tissue was analyzed using targeted next-generation sequencing. A pathogenic mutation in the TP53 gene was detected.". The contents of both fields are recorded in PAS Xanthos in the "dbo"."MaterialEingangBefundtext".”Text” field; they do not occur in the ADOPT cohort, the CRC cohort, or MIABIS.

Some fields are only required in certain cases. The **“Neuropathological report”** field contains the results of the neuropathological examination. It is populated with text, such as: “Sections show a diffusely infiltrating glial neoplasm with increased cellularity and nuclear atypia. Mitotic figures are identified. No definite microvascular proliferation or necrosis is observed.”. In general, this description corresponds to the Histologic Description in the pathology report and is therefore added to this text field.

If a rapid section is required, the results of the rapid section are entered as text in the **“Rapid section examination”** field, for example: “Sections show a cellular neoplasm composed of atypical epithelial cells arranged in irregular glandular structures. Nuclear pleomorphism and mitotic activity are present.”.

In the event of the patient's death, a postmortem examination is performed. The results are saved in the **“Postmortem examination”** field and entered there as descriptive text, for example: “The lungs are heavy and congested. The heart is enlarged with left ventricular hypertrophy. The liver shows moderate steatosis. No acute hemorrhage or traumatic injury is identified.”.

The cause of death is also determined during the postmortem examination and recorded in the **“Cause of death”** field, for example: "Immediate cause of death: Acute respiratory failure; Underlying cause: Metastatic adenocarcinoma of the lung".

If a test of a cytological sample has been requested, the results are recorded in the **“Cytological description”** field, for example, “Smears show a moderately cellular specimen composed of cohesive clusters and scattered atypical epithelial cells.”.

To track the timeline of the report generation and the status of the case, the **“Report type”** field is used to specify the type of report. This field is populated with predefined values, which include “Initial report” (the orignal report), “Addendum” (extention of an existing report) and “Amended report” (change of exisiting report).

The fields **“Neuropathological report”, “Rapid section examination”, “Cause of death”, “Postmortem examination”, “Cytological description”** and **“Report type”** are each stored as text in field "dbo"."MaterialEingangBefundtext"."Text” in PAS Xanthos. This information is not included in the ADOPT Cohort, the CRC Cohort, or MIABIS.

The classification of tumors according to the TNM system is represented in several fields in the catalog. Each main category is described by three fields: The prefix, the main field, and a field for additions. The **“TNM T prefix”** field describes the prefix for the T category of the TNM classification. This prefix indicates whether the diagnosis was clinically or pathologically confirmed, whether treatment has already been performed, whether the case involves a recurrence, or whether the classification was made during a postmortem examination following the patient’s death. The field can be populated with predefined values. The following are possible: “c”; “p”; “yc”; “yp”; “a”; “r”; ‘rc’; “rp”. In PAS Xanthos, the information is stored as text in field "dbo"."MaterialEingang_Codes_TNM"."TPrefix"; the ADOPT cohort maps all information regarding the T category of the TNM system to data field “TNM_PRIMARY_TUMOR”. This field contains the prefix, the main information, and all additional details as a fixed value set. In the CRC cohort as well, all information regarding the T category is stored in a single data field. The “t” field contains all the information in text format. MIABIS does not include this information.

The **“TNM T”** field describes the T category of the TNM classification, i.e., the extent of the primary tumor. This field is populated with predefined values: "TX"; "T0"; "Ta"; "Tis"; "T1"; "T1mi"; "T1a"; "T1a1"; "T1a2"; "T1b"; "T1b1"; "T1b2"; "T1b3"; "T1c"; "T1c1"; "T1c2"; "T1c3"; "T1d"; "T2"; "T2a"; "T2a1"; "T2a2"; "T2b"; "T2c"; "T2d"; "T3"; "T3a"; "T3b"; "T3c"; "T3d"; "T3e"; "T4"; "T4a"; "T4b"; "T4c"; "T4d"; "T4e". PAS Xanthos stores this value as text in field "dbo"."MaterialEingang_Codes_TNM"."TPostfix"; for the ADOPT cohort and the CRC cohort, the value of this field, as described above, is part of a shared field that describes category T. This information is not available in MIABIS.

The **“TNM T addition”** field allows for additional information and specifications and is filled in as free text; for example, “m” indicates the presence of multiple primary tumors. In PAS Xanthos, this information is stored as text in the "dbo"."MaterialEingang_Codes_TNM"."TErgaenzung" field; the ADOPT cohort and the CRC cohort use a shared data field for the entire T category of the TNM system, as described above. MIABIS does not map this information.

The **“TNM N prefix”** field describes the prefix for the N category of the TNM classification. This prefix indicates whether the diagnosis was clinically or pathologically verified, whether treatment has already been done, whether it is a relapse, or whether the classification was done as part of a postmortem examination following the patient’s death. The field can be populated with predefined values; the following are possible: “c”; “p”; “yc”; “yp”; “a”; “r”; “rc”; “rp”. In PAS Xanthos, this information is stored as text in the "dbo"."MaterialEingang_Codes_TNM"."NPrefix" field. The ADOPT cohort maps all information regarding the N category to the data field “TNM_REGIONAL_LYMPH_NODES,” which contains the prefix, the main information, and all additional details as a fixed value set. In the CRC Cohort as well, all information regarding the N category is stored in a single data field named “n” in text format. MIABIS does not contain this information.

The **“TNM N”** field describes the N category of the TNM classification, i.e., the extent of regional lymph node involvement. This field is predefined with the following values: “NX”; “N0”; “N1”; “N1mi”; “N1a”; ‘N1b’; “N1c”; “N2”; “N2a”; “N2b”; “N2c”; “N3”; “N3a”; “N3b”; “N3c”. PAS Xanthos stores this value as text in the field "dbo"."MaterialEingang_Codes_TNM"."NPostfix". For the ADOPT cohort and the CRC Cohort, the value of this field, as described above, is part of a shared field that describes the N category. This information is not available in MIABIS.

The **“TNM N addition”** field allows for additional information and specifications and is filled in as free text; for example, “(i+)” indicates the detection of isolated tumor cells. In PAS Xanthos, this information is stored as text in the field "dbo"."MaterialEingang_Codes_TNM"."NErgaenzung". As previously described, the ADOPT cohort and the CRC Cohort use a common data field for the entire N category of the TNM system. MIABIS does not map this information.

The **"TNM M prefix"** field describes the prefix for the M category of the TNM classification. This prefix documents the status of the metastasis diagnosis, such as whether it was confirmed clinically, pathologically, or following prior treatment. The field can be populated with the predefined values “c”; “p”; “yc”; “yp”; “a”; “r”;

‘rc’; “rp”. In PAS Xanthos, the information is stored as text in the field "dbo"."MaterialEingang_Codes_TNM"."MPrefix". The ADOPT cohort maps all details regarding the M category to the data field “TNM_DISTANT_METASTASIS” as a fixed value set. In the CRC Cohort, this information is uniformly stored in the text field “m.” In MIABIS, this prefix is not provided.

The **"TNM M"** field describes the M category of the TNM classification and indicates whether distant metastases are present. The field is filled with the following predefined values: “M0”; “M1”; “M1a”; “M1a1”; “M1a2”; “M1b”; “M1b1”; ‘M1b2’; “M1c”; “M1c1”; “M1c2”; “M1d”. PAS Xanthos stores this value as text in the field "dbo"."MaterialEingang_Codes_TNM"."MPrefix". In the ADOPT cohort and the CRC Cohort, this value is part of the respective shared data field for the M category mentioned above. This information is not available in MIABIS.

The **“TNM M addition”** field allows for further specifications regarding the M category in free-text format, such as the notation “(mol+)” for purely molecular biological evidence. This information is stored in PAS Xanthos in the text field "dbo"."MaterialEingang_Codes_TNM"."MErgaenzung". Both the ADOPT cohort and the CRC Cohort use their respective collection fields for the M classification for these additions. MIABIS does not map this additional information.

The **"TNM stage"** field is used to provide a summary classification of a tumor’s anatomical spread, based on a combination of the T, N, and M values. This classification is of utmost importance in clinical practice, as it enables a rapid assessment of the prognosis and serves as the primary guideline for selecting a treatment strategy (e.g., surgery, radiation therapy, or chemotherapy). The field is populated with predefined values, such as: "0"; "0a"; "0is"; "I"; "IA"; "IA1"; "IA2"; "IA3"; "IB"; "IB1"; "IB2"; "IC"; "IS"; "II"; "IIA"; "IIA1"; "IIA2"; "IIB"; "IIC"; "III"; "IIIA"; "IIIB"; "IIIC"; "IIIC1"; "IIIC2"; "IIID"; "IV"; "IVA"; "IVB"; "IVC". In PAS Xanthos, this value is stored as a text in the field "dbo"."MaterialEingang_Codes_TNM"."Stadium". The ADOPT cohort uses the field "UICC_STAGE" with a predefined value set for this purpose, while the CRC

Cohort lists the value under “staging” as a text. This information is not available in MIABIS.

The **"TNM UICC"** field documents the specific version of the UICC classification used to generate the findings. Since the criteria for staging may vary between different editions (e.g., Version 8 versus Version 9), this information is essential for comparing study results over time and for meta-analyses. The field is defined as the “Number” data type, with the value “9” provided as an example. In PAS Xanthos, the version number is stored in the field "dbo"."MaterialEingang_Codes_TNM"."UICC". The ADOPT cohort uses the “UICC_VERSION” field with a fixed value set, and in the CRC Cohort, this information is stored as text in the “stagingDefinition” field. This information is also not mapped separately in MIABIS.

The **"ICD-O code"** field is used for the specific oncological classification of tumors according to the “ICD for Oncology”. This coding is of crucial importance for cancer research, as it precisely describes not only the location (ICD-O-T, topography) but also the biological behavior and cell type (ICD-O-M, morphology) of the tumor (example: “8500/3”). To populate this field, a list of all available ICD-O codes is used. In PAS Xanthos, this value is stored via a link between the tables "dbo"."MaterialEingang_Codes_Sonstige"."idCode" and "dbo"."Code"."Kuerzel". The CRC Cohort stores this value as text in the “icd0[]” field. This specific field is not available in the ADOPT Cohort or in MIABIS.

The **“ICD-10 code”** field enables the general coding of diagnoses according to the globally established standard of the WHO. This data is essential for the statistical analysis of clinical presentations and for billing in the healthcare sector. It is filled in using a list of all available ICD-10 codes (example: “J45.0” for asthma). It has to be noted that the ICD-O-T Code corresponds to malignant tumors in the ICD-10. While this field is not listed in PAS Xanthos and the ADOPT cohort in this specific mapping, it is listed as a text in the CRC Cohort under “icd10[]” and as a fixed value set in the MIABIS Sample+Donor+Event standard under the identifier “MIABIS-SAMPLE-10”.

The **“ICD-11 code”** field represents the latest generation of the ICD and ensures that the data structure is future-proofed to meet the most advanced diagnostic requirements. Here, as well, a list of all available ICD-11 codes is used to populate the field. (Example: “CA23.0”). The inclusion of this field allows for an even more detailed digital representation of medical conditions compared to previous versions. This field is not described in the existing systems and cohorts, and represents a new addition to the catalog.

The appendix contains a table with details of the individual data fields available at this level and their availability in the source systems.

3.7.5 Level 5 – Block

The “Level 4 - Container” level serves as the parent level for “Level 5 - Block”. The information in this layer focuses on the block in which biological samples are preserved and prepared for further processing into slides and for storage.

The **“Parent container ID”** field is used to link the "block" to the higher-level layer “container”. This assigns a block to a specific container.

The **“Block ID”** field serves as a technical identifier for a specific tissue block and is essential for traceability within the biobank and the laboratory. Without this unique ID, the sections obtained from the block could no longer be reliably traced back to the original sample. The field is defined as text (example: “2024000001/01”). In PAS Xanthos, the information is stored as an integer in field "dbo"."Laborbuch"."Nummer", while in the CRC cohort it is stored as a UUID in field “blockNumber”. This field is not available in the ADOPT cohort or in the MIABIS standard.

The **“Human readable block ID”** displays the block identifier in a format that is easy to read for humans, to facilitate manual handling and archiving. This minimizes the risk of mix-ups during physical sorting in the archive. Here, the information is also stored as an integer in the "dbo"."Laborbuch"."Nummer" field in

PAS Xanthos and as a UUID in the “blockNumber”field in the CRC cohort. This field is neither available in the ADOPT cohort nor in the MIABIS standard.

The **“Location of stored block”** field records the physical storage location of the block (e.g., “Biobank Graz”). This information is useful for quickly locating samples for future research purposes. It is defined as a text field and is not available in the current systems and cohorts. This field therefore represents a new feature.

The fields **“Access condition block”** and **“Access condition block date”** govern the access rights and their duration for the respective tissue block. The **“Access condition block”** field is populated with predefined values, which include: “Open access”; “Restricted access”; “Controlled access”; “Non-commercial use only”; “Academic use only”; “Personal use only”; "Institutional access only"; "Embargoed". The **“Access condition block date”** field is filled in date format, for example, “31.01.2000”. Both fields are relevant for compliance with ethical guidelines and data protection regulations in research. They are new data fields and are not available in PAS Xanthos, ADOPT, CRC, or MIABIS.

The **“Fixation type”** field describes the method used to preserve the tissue. Since the fixation method has a significant impact on the quality of subsequent analyses (such as IHC or DNA extraction), this information is important for the scientific usability of the sample. There are predefined values for this field: “Formalin fixation”; “Alcohol fixation”; “Frozen section”; “Glutaraldehyde fixation”; “Michel’s medium”; “Bouin’s solution”; “Cytology preservative solutions”; “Unfixed tissue”. In the ADOPT cohort, this is stored as a fixed set of values in the “SAMPLE_PRESERVATION_MODE” field, whereas this information is not available in this form in PAS Xanthos, the CRC cohort, and MIABIS.

The **“Block label”** field provides space for a short, descriptive label for the block (e.g., ‘TU’ for tumor or “1,” “My”). This makes it easier for laboratory staff to quickly identify the sample type at a glance. The field is populated with free-text data in the catalog. In PAS Xanthos, this information is displayed as text in table

"dbo"."Laborbuch"."Text" in combination with the laboratory book, whereas the CRC cohort uses a UUID in the “description” field for this purpose.

The appendix comprises a table detailing the individual data fields for this level and their availability in the source systems.

3.7.6 Level 6 – Slide

The “Level 5 - Block” level serves as the parent level for “Level 6 - Slide”. This level describes the glass slide on which the sections cut from the corresponding block are mounted.

The **“Parent block ID”** field is used to link the "slide" to the higher-level layer “block”. This assigns a slide to a specific block.

The **“Slide ID”** field serves as a technical identifier for a specific slide and is essential for diagnostic accuracy and scientific analysis, as it is the only way to directly link microscopic findings to the corresponding tissue block. Without this unique ID, seamless traceability from the digital image or slide back to the original sample cannot be guaranteed. The field is defined as text (example: “2024000001/01/01”). In PAS Xanthos, this information is stored as an integer in field "dbo"."Laborbuch"."Nummer"; in the ADOPT cohort, this ID is represented as text in field “SAMPLE_ID”; in the CRC cohort, the ID is stored as a UUID in field “slideID”; and in MIABIS Digital Pathology, it appears as text in field “MIABIS-DIGITALPATHOLOGYASSAY-03”.

The **“Human-readable slide ID”** presents the slide identifier in a format that is easily readable by humans, to support manual handling, sorting, and microscopy in everyday laboratory work. This minimizes the risk of errors when physically handling the slides. The field is populated with text. Like the “Slide ID,” this field is stored as an integer in field "dbo"."Laborbuch"."Nummer" in PAS Xanthos; in the ADOPT cohort, the information is stored as text in field “SAMPLE_ID”. In the CRC cohort, the ID is represented as a UUID in field “slideID”, and in MIABIS Digital

Pathology, it is represented as text in field “MIABIS-DIGITALPATHOLOGYASSAY-03”.

The **“Location of stored slide”** field records the physical storage location of a specific glass slide, for example at a facility such as the “Biobank Graz”. This information is essential for efficiently locating samples for subsequent diagnostic follow-ups or scientific research projects, as it enables rapid retrieval from the archive. The field is populated with data of the text data type, for example, “Biobank Graz”. This field was added as an extension; there is no information related to it in PAS Xanthos, the ADOPT cohort, the CRC cohort, or MIABIS.

The **“Access condition slide”** field governs the specific terms of use and access rights for a physical glass slide, which is essential for ensuring compliance with ethical guidelines and data protection regulations in research. It defines whether a sample is freely available for scientific use or whether restrictive conditions apply to ensure the protection of donor data and compliance with informed consent. The field is populated with predefined values, which may include: “Open access”; “Restricted access”; “Controlled access”; “Non-commercial use only”; “Academic use only”; “Personal use only”; “Institutional access only”; “Embargoed”. This field is not described in PAS Xanthos, the ADOPT Cohort, the CRC Cohort, or MIABIS, and is therefore a new addition to the catalog.

The **“Access condition slide date”** field records the date until which the specified access conditions for the respective slide are valid. This time limit is useful for technically mapping legal retention periods or time-limited patient consent forms, thereby ensuring legally compliant sample management. The field is defined as a Date data type with the example value “31.01.2000”. Like the field for the access type itself, this date field is not described in the existing systems, cohorts, and standards.

The **“Slide label”** field provides a brief, descriptive label - displayed as free text (e.g., “/2”) - that is also visible directly on the glass slide, facilitating quick orientation when reviewing multiple slices of a block. This information is particularly useful for visually distinguishing between different staining levels or

special examinations within the same case. In PAS Xanthos, this label is stored as text in the "dbo"."Laborbuch"."Text" field; however, it is not available in this form as a separate data field in the ADOPT and CRC cohorts or in the MIABIS standard.

The **“Assay type”** field documents the testing procedure performed in the laboratory on the biological specimen mounted on the glass slide, which is crucial for the correct interpretation of the microscopic results. It allows for the distinction between standard procedures, such as hematoxylin-eosin staining, and specialized IHC examinations. This field uses the following predefined values: “Hematoxylin and Eosin Stain”; “Special Stains”; “Immunohistochemistry”; “Immunocytochemistry”; “Immunofluorescence”; “In Situ Hybridization”; “Fluorescence In Situ Hybridization”; “Fluorescent Multiplex Immunohistochemistry”; “Multiplex Immunohistochemistry”; “other (the value is known but cannot be classified into any of the predefined categories)”; “NULL (the value is unknown)”. It is recorded as text in the "dbo"."Laborbuch"."Code"."Text” field in PAS Xanthos and as a fixed value set in the “MIABIS-DIGITALPATHOLOGYASSAY-11” field in MIABIS Digital Pathology, while it is not directly mapped in the ADOPT and CRC cohorts.

The **“Special stain technique”** field documents the specific method of a special staining procedure that goes beyond standard hematoxylin-eosin staining to make certain tissue structures, microorganisms, or deposits visible under a microscope. This information is of crucial importance for pathological diagnosis, as it enables the identification of specific disease features that might remain hidden in routine staining. The field is defined as “Enumerated Values”, with the permissible values to be derived from a standardized catalog - the "CID 8112 Specimen Stain" (https://dicom.nema.org/medical/dicom/current/output/chtml/part16/sect_cid_8112.html). The values are to be encoded as SNOMED-CT ID. While this field is not directly mapped in PAS Xanthos, the ADOPT cohort, and the Gooogle Cohort, it is stored in the MIABIS Digital Pathology standard under the identifier “MIABIS-DIGITALPATHOLOGYASSAY-12” as a fixed value set.

The **“Dye name”** field specifies the exact chemical name of the neutral or synthetic dye used for staining specific tissues or cellular components. Accurate documentation of the dye is essential for scientific reproducibility and internal quality control in the laboratory to ensure consistent staining intensity and result quality. As with the staining technique, this is an “Enumerated Values” data type, whose values also originate from the "CID 8112 Specimen Stain". The values also are to be encoded as SNOMED-CT ID. This field is not available in PAS Xanthos, the ADOPT cohort, and the Gorilla cohort, but is listed as a fixed value set in the MIABIS Digital Pathology standard under “MIABIS-DIGITALPATHOLOGYASSAY-13”.

The **“Antibody name”** field records the name of the antibody used as specified by the manufacturer, which is essential for the diagnostic classification of IHC antibodies. This information enables the pathologist to identify specific protein expressions in the tissue and thus make precise determinations regarding the tumor type or therapeutic response. The field is populated with text, for example, “Recombinant Monoclonal Anti-Ki-67 Antibody, Rabbit (7A4)”. In PAS Xanthos, this information is stored as text in field "dbo"."Laborbuch"."Code"; in MIABIS Digital Pathology, it is stored as text in field “MIABIS-DIGITALPATHOLOGYASSAY-14”. This information is not available in the ADOPT cohort or the CRC cohort.

The **“Antibody identifier”** field is used to uniquely identify an antibody using a standardized ID, which is useful for scientific reproducibility and cross-system data exchange. Since different manufacturers may offer antibodies with similar names, this identifier ensures that exactly the same reagent is referenced in research databases. The field is populated with values from a list as specified on the page “https://www.antibodyregistry.org/”, for example “AB_3741092”. If the antibody is not included in the list, it is described as free text. This identifier is not mapped in PAS Xanthos, the ADOPT cohort, or the CRC cohort. In MIABIS Digital Pathology, it is stored as a fixed value set in the “MIABIS-DIGITALPATHOLOGYASSAY-15” field.

The **“Antibody clone freetext”** field specifies the name of the antibody clone used, which is of great importance for validating staining results and for quality control in the laboratory. Since different clones of the same antibody may exhibit varying binding properties or sensitivities, this level of detail is necessary for comparing findings across different laboratories. The field is populated with text, for example, “Recombinant Monoclonal Anti-Ki-67 Antibody”. This information is displayed as text in field "dbo"."Laborbuch"."Code" in PAS Xanthos; it does not appear in the ADOPT cohort, the CRC cohort, or the MIABIS standard.

The appendix section contains a table that provides an overview of the individual data fields for this level and their availability in the source systems.

3.7.7 Level 7 – WSI

The “Level 6 - Slide” level serves as the parent level for “Level 7 - WSI”. This level describes the glass slide on which the sections cut from the corresponding block are mounted.

The **“Parent slide ID”** field is used to link the "WSI" to the higher-level layer “slide”. This assigns a WSI to a specific slide.

The **“WSI ID”** field serves as a unique technical identifier for the digital image of a glass slide and is the fundamental reference for linking image data to clinical metadata. Without this ID, automated image analyses or AI algorithms could not be correctly assigned to the corresponding patient case. The field is populated with text, for example “d7880b0a-b2a2-488e-a4ab-67103fd225e8”. PAS Xanthos uses the field "dbo"."Zusatzfeld"."Laborbuch" for this purpose, in which the ID is stored as text. In the CRC Cohort, the ID is stored as a UUID in the “scanID” field; in MIABIS Digital Pathology, it is stored as text in the “MIABIS-DIGITALPATHOLOGYSCAN-01” field. This information is not included in the ADOPT cohort.

The **“Human-readable WSI ID”** field displays the identifier of the digital scan in a human-readable format, allowing lab staff to quickly and visually match the image

to the physical slide. This is particularly useful when manually reviewing scan results in the archive. The field is populated with text, for example, “2024000001/01/01/01”. PAS Xanthos stores this ID as an integer in the "dbo"."Laborbuch"."Nummer" field. The CRC Cohort also maps this ID to a UUID in the “scanID” field, and MIABIS Digital Pathology also stores this ID as text in the “MIABIS-DIGITALPATHOLOGYSCAN-01” field. This information is not represented in the ADOPT cohort.

The **“Location of stored WSI”** field records the digital storage location or path of the image file, which is essential for quick access to the data within the hospital network or biobank. A correct path specification, such as “\\DigitalArchive4245\WSI\d7880b0a-b2a2-488e-a4ab-67103fd225e8,” ensures that viewer software can load the image without delay. This is stored as text in the field. This field is a new addition to the catalog; this information was not included in the existing systems, cohorts, and standards.

The **“Access condition WSI”** field governs the specific terms of use and digital access rights for a WSI to ensure data protection and compliance with ethical guidelines. It defines, for example, whether an image is freely available for research purposes or may only be used under restricted conditions. This field is populated with values from a predefined list, which may include “Open access”; “Restricted access”; “Controlled access”; “Non-commercial use only”; “Academic use only”; “Personal use only”; “Institutional access only”; “Embargoed”. This field has not yet been described in PAS Xanthos, the ADOPT cohort, the CRC cohort, or MIABIS, and is therefore considered a new field.

The **“Access condition WSI date”** field records the date until which the specified access conditions for the digital scan remain valid. This time limit is useful for technically mapping any patient withdrawal periods or statutory retention periods. The field is populated in the date format, for example “31.01.2000”. Like the access type itself, this date field is also marked as “New” and is not yet available in the current source systems and cohorts.

The **“Manufacturer of scanner”** field specifies the manufacturer of the scanning system used, which is important for technical standardization and the scientific reproducibility of the image data. Different manufacturers often use proprietary optics that can affect the image. This field is populated with text, such as “Aperio”. This information can be found in MIABIS Digital Pathology in the “MIABIS-DIGITALPATHOLOGYSCAN-05” field as text. In PAS Xanthos, the ADOPT cohort, and the CRC cohort, this information is not available.

The **“Type of scanner”** field specifies the exact model of the scanner in order to accurately document the technical parameters of image acquisition. Since different models produce varying scan times and image qualities, this information is useful for process optimization in the laboratory. This field is also populated with text, for example “GT 450 DX”. In MIABIS Digital Pathology, this information appears as text in field “MIABIS-DIGITALPATHOLOGYSCAN-05”; in PAS Xanthos, the ADOPT cohort, and the CRC cohort, it does not occur.

The **“Serial number of scanner”** field is used to identify the specific device used to create the scan, which is useful for the laboratory’s internal quality assurance and maintenance records. In the event of image errors, the device responsible can be precisely identified. The information is recorded as text, for example “SS45789”. In the CRC cohort, the information can be found as text in the “ScannerID” field; in PAS Xanthos, the ADOPT cohort, and MIABIS, it is not present.

The **“Resolution field”** records the scan resolution in micrometers per pixel (µm/pixel), which forms the basis for all quantitative image analysis and morphometric measurements. An accurate resolution is crucial to ensure that fine cellular details are sufficiently visible for diagnostic purposes. The resolution is stored as a decimal value, for example, “0.25”. In the CRC cohort, this information is stored as text in the “Magnification” field, while in MIABIS Digital Pathology it is stored as an integer in the “MIABIS-DIGITALPATHOLOGYFILE-08” field. PAS Xanthos and the ADOPT cohort do not contain this information.

The **“Scan date and time”** field records the exact time of digitization, which is useful for reconstructing the sample workflow and conducting temporal analyses. The field uses the datetime format, for example “31.01.2025 06:00:00”. This information is available in PAS Xanthos in datetime format in field "dbo"."Laborbuch"."Datum", as well as in the CRC Cohort as text in field “Timestamp” and in MIABIS Digital Pathology in datetime format in field “MIABIS-DIGITALPATHOLOGYSCAN-02”.

The **“Imaging technique”** field describes the imaging method used, such as brightfield microscopy, which is necessary for the correct visualization and interpretation of tissue structures. Different techniques require different digital processing algorithms. The field is populated with predefined values, including “Transmission”; “Reflection”; “Epifluorescence”; “Brightfield”; “Darkfield”; “Oblique”; “Phase Contrast”; “Differential Interference Contrast”; “Total Internal Reflection Fluorescence”. In MIABIS Digital Pathology, this information is represented as a fixed value set in the “MIABIS-DIGITALPATHOLOGYSCAN-14” field. In PAS Xanthos, the ADOPT cohort, and the CRC cohort, this information is not available.

The **“Compression method”** field specifies the data compression method used to manage the file size of high-resolution scans. Since compression methods can affect image quality, this information is critical for the validation of AI models. The compression method is specified using predefined values, which include “JPEG”; “JPEG 2000”; “JPEG-LS”; “JPEG XR”; “PNG (DEFLATE)”; “TIFF (LZW)”; “TIFF (PackBits)”; “TIFF (Deflate / ZIP)”; “WebP”; “HEIF / HEIC”; “AVIF”; “GIF (LZW)”; “BMP (RLE)”; “BPG”; “FLIF”; “JPEG XL”; “LERC (Limited Error Raster Compression)”; “Zstandard (Zstd)”; “LZMA”. In MIABIS Digital Pathology, this information is represented in the “MIABIS-DIGITALPATHOLOGYFILE-11” field as a fixed value set. In PAS Xanthos, the ADOPT cohort, and the CRC cohort, this information is not provided.

The **“Compression quality level”** field specifies the degree of compression, with a higher value indicating better image quality but requiring more storage space. This information is useful for IT storage management and is stored as text, for

example, "80". In MIABIS Digital Pathology, this information is stored as a decimal value in the “MIABIS-DIGITALPATHOLOGYFILE-12” field. In PAS Xanthos, the ADOPT cohort, and the CRC cohort, this information is not available.

The **“Image format”** field specifies the digital file format of the scan, which is crucial for compatibility with various image viewers. Different formats require specific software decoders. The field is populated with predefined values; possible values include “svs”; “ndpi”; “scn”; “mrxs”; “czi”; “vms”; “vmu”; “bif”; “tif”; “tiff”; “ometiff”; “isyntax”; “kfb”; “svslide”; “dcm”; “zvi”; “lif”. The CRC cohort represents this information as text in the “ScanFileType” field, while MIABIS Digital Pathology represents it in the “MIABIS-DIGITALPATHOLOGYFILE-02” field as a fixed value set. PAS Xanthos and the ADOPT cohort do not include this information.

The **“Firmware version”** field records the scanner's firmware version at the time the image was captured, which is useful for troubleshooting and standardizing image acquisition. Software updates can affect image processing and thus the final result. The field is populated with text, for example “1.2.5”. This field is available exclusively in MIABIS Digital Pathology, where the information is stored as text in field “MIABIS-DIGITALPATHOLOGYSCAN-06”. In PAS Xanthos, the ADOPT cohort, and the CRC cohort, this information is not provided.

The **“Z-stacked”** field indicates whether the scan contains multiple focal planes in the depth direction (Z-planes), which is particularly necessary for diagnosis in the case of thick tissue sections or cytological specimens. The field uses the values "single"; "multiple"; "extended". MIABIS Digital Pathology provides this information in the “MIABIS-DIGITALPATHOLOGYSCAN-08” field in the form of a fixed value set. This data is not available in PAS Xanthos, the ADOPT cohort, or the CRC cohort.

The **“Scan area”** field describes the digitized portion of the slide to indicate whether the entire section or only a portion of it was captured. This is useful for assessing the completeness of the digital report. This field can be populated with the options “full slide” or “partial.” This information is not recorded in the source

systems, cohorts, or standards. Accordingly, this field is also one of the new additions.

The **“WSI annotation available”** field indicates whether a pathologist has already added markings or comments to the digital scan, which is extremely valuable for subsequent scientific analyses or for training algorithms. The field uses a fixed value set with the values “Yes” and “No.” This information is described in MIABIS Digital Pathology in the xxx field as a fixed value set. In PAS Xanthos, the ADOPT cohort, and the CRC cohort, this data is not included. In PAS Xanthos, a proprietary list of possible sources for sample material has been used to this point to describe the composition of the sample.

The appendix contains a table that provides an overview of the individual data fields for this level and their availability in the source systems.

3.8 Alignment with existing metadata standards

In PAS Xanthos, a proprietary list of possible sources for sample material has been used to this point to describe the composition of the sample. To make the query options more precise and to convert the sample data into a standardized, machine-readable format, the contents of the currently used structured list were mapped to SNOMED-CT codes as part of this thesis.

For example, the histology material label **“Milz”** was mapped to “Specimen from spleen” with the SNOMED-CT code “433308004”. Semantically, this code is classified as “specimen”.

The description currently in use in gynecological cytology, **“Vagina”**, is represented by the SNOMED-CT code “258520000”, which stands for “Vaginal swab”. This code also falls under the semantic category “specimen”.

While mapping the original list to SNOMED-CT codes, it became apparent that it wasn't possible to assign a SNOMED-CT code to every value that precisely describes the type of sample material. Therefore, for the affected values, higher-level categories were used to at least broadly categorize the sample.

Table 3: SNOMED-CT mapping of materials

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Histology	STUHL	119339001	Stool specimen	specimen
Histology	MAMMA	127456000	Specimen from breast	specimen
Histology	WEICHTEILGEWEBE	309072003	Soft tissue specimen	specimen
Histology	MAMMA TE	127456000	Specimen from breast	specimen
Histology	VAGINA	119394009	Specimen from vagina	specimen
Histology	MILZ	433308004	Specimen from spleen	spceimen
Histology	MUNDHOEHLE	309185002	Oral cavity specimen	specimen
Histology	MUNDHÖHLEN CA	309185002	Oral cavity specimen	specimen
Histology	NECKDISSECTION	309476009	Neck block dissection specimen	specimen
Histology	NETZ (OMENTUM MAJUS)	433309007	Specimen from omentum	specimen
Histology	NIERE CA	127474009	Tissue specimen from kidney	specimen
Histology	PHARYNX	309168004	Upper respiratory tissue specimen	specimen
Histology	TONSILLE	430222001	Specimen from tonsil	spceimen
Histology	MAMMA ABLATIO	127456000	Specimen from breast	specimen
Extragenital cytology	Gastrointestinale Bürstenzyto	127465007	Specimen from gastrointestinal tract	specimen
Extragenital cytology	Knochenmark/Blut	119359002	Bone marrow specimen	specimen
Extragenital cytology	Gastrointestinale FNA	127465007	Specimen from gastrointestinal tract	specimen
Extragenital cytology	Leber	119383005	Specimen from liver	specimen
Extragenital cytology	Harn - Lavage	122575003	Urine specimen	specimen
Extragenital cytology	Harn - Spontanurin	122575003	Urine specimen	specimen
Extragenital cytology	Harn ohne Gewinnungsangaben	122575003	Urine specimen	specimen
Extragenital cytology	Endozervix (CK Bürstenabstriich)	258524009	Cervical swab	specimen
Extragenital cytology	Abstrich zur Erregerdiagn.	257261003	Swab	specimen
Extragenital cytology	ISOLIERTE DNA	258566005	Deoxyribonucleic acid specimen	specimen

Table 3 shows a selection of materials from PAS Xanthos along with the SNOMED-CT codes and SNOMED-CT terms they correspond to. The appendix section contains a comprehensive list of the materials used and their associated SNOMED-CT codes and terms.

With regard to staining methods and IHC techniques, existing standards or databases were investigated in order to enable standardized coding in this area as well. However, since no standards or databases covering the commonly used stains could be found in this field, the values from CID 8112 Specimen Stain – a DICOM Content Mapping Resource - are used and coded in SNOMED-CT.

Since the Content Mapping Resource CID 8112 Specimen Stain described here only maps stains and does not include IHC methods, another database has to be consulted for mapping the IHC methods. The Antibody Registry, created at the University of California (https://www.antibodyregistry.org/), was used for this purpose. The ID from this database is stored in the “Antibody identifier” field. If the antibody used is not found in this registry, the name of the antibody is recorded as free text in the “Antibody name” field.

4 Discussion

This discussion examines the process of selecting data fields and establishing technical standards for the metadata catalog, and analyzes its potential as an essential interface for the structured storage and exchange of digital image data in pathology.

4.1 Key Findings

The main result of this work is the list of data fields that form the basis for the metadata catalog. In order to select a set of data fields, it was necessary to gain an overview of the existing datasets and their structure and format. After the initial draft was prepared, the data fields were revised during the workshop. An important aspect of this process was reaching an interdisciplinary consensus. The workshop was attended by pathologists, data scientists, and researchers, each of whom shared insights on how the catalog could make their daily work more efficient and what requirements it would need to meet to do so. Pathologists, who were increasingly involved in diagnostic work, noted that the structured data entry should not significantly prolong or complicate the diagnostic process. Participants focused on research emphasized the advantage of coded and machine-readable data for cohort creation. Another point of discussion was the availability of the data. Data fields that were deemed non-essential and, moreover, difficult to populate were removed from the list of data fields.

Following this interdisciplinary discussion and agreement on a list of data fields, these fields were supplemented with descriptions, examples, and the data type they are intended to contain. When selecting data types, care was taken to ensure that filling out the data fields was not overly burdensome, while still achieving the highest possible rate of structured or coded data. The challenge of balancing practicality and efficiency in the diagnostic process with the need to generate a dataset that is as machine-readable as possible and offers benefits for research was successfully met through collaboration with the various professional fields involved. As part of the standardization of selected data fields, some of the fields that were free-text fields in the source systems were converted into structured or coded fields. A crucial part of the thesis was mapping the existing list of materials

to SNOMED-CT codes. Recording the origin of the sample using codes enables a machine-readable form of documentation that is internationally understandable and comparable. In the area of staining and IHC techniques, it was not possible to map all stains to a SNOMED-CT code; however, a list of codes was utilized for the most commonly used stains. For any stains not included in this list and any IHC methods not included in the Antibody Registry, there is an option to record them in a designated free-text field.

Another key achievement of this work is the creation of a structure in which the data fields are embedded. The data fields are assigned to different hierarchical levels and thus linked to one another. This ensures that it is always clear how the data within a record are related to one another. For example, a scan can be assigned to a specific patient visit. During the workshop, participants agreed to adopt a patient-centered structure rather than a specimen-centered one. The advantage of using this approach is the possibility to clearly identify the clinical context at any time, which is ideal for evaluating the patient's clinical progress or specimens which belong together, but were sent in different containers within a short time period (e.g. several specimens during a longer surgical procedure). For example, one can see at a first glance whether there are already findings from previous visits for the patient whose sample is currently being examined. This allows one to retrieve old information as needed and use it to assess the current situation. However, this type of structure can become less easy to navigate when there is a large number of samples—and, consequently, WSIs—to manage. A sample-centric structure would focus on the glass slide or the WSI. In this case, the overview at the metadata level would simply be preserved, while the clinical context would be sacrificed.

4.2 Recap of the thesis objectives and findings

The goal of this thesis was to create a metadata catalog that contains a harmonized dataset usable throughout Austria, which is easy to navigate and in which cases can be easily located. Currently as of May 2026, the catalog has already been technically implemented at the Diagnostic and Research Institute of Pathology in Graz: At this point, data from 70,000 patients has already been

integrated into the structure and is now searchable. Newly generated data is already being entered into this structure, and already existing data is being exported and added gradually. The partner institutions in Linz, Innsbruck and Vienna are also already in the process of implementing a catalog modeled after this work. This has laid the foundation for Austria-wide collaboration across various institutions.

To select the data fields for this catalog, various source systems were analyzed. The data fields from the diagnostic LIMS (PAS Xanthos) were compared with the data fields from research projects (ADOPT cohort, CRC cohort) and the standards (MIABIS Core, MIABIS Sample+Donor+Event and MIABIS Digital Pathology). This revealed that most of the fields in PAS Xanthos are filled in as free text. This makes perfect sense in diagnostics, as it ensures a high degree of accuracy and flexibility. In fact, structured data is not necessary for diagnostics; it would be more of an advantage in research, as structured data can be compared more easily. However, this increases the difficulty of extracting useful data for research from PAS Xanthos. The projects and standards differ in terms of the available data fields not only from the diagnostics but also from each other. The data fields in each project are specifically geared to the respective question. The ADOPT cohort focuses more on the clinical course of the patient (information on imaging procedures, molecular markers, pharmacotherapy, radiotherapy, surgical therapy, survival status, etc.). The CRC cohort, on the other hand, relates more to pathological diagnostics (sample material, localization, slide ID, block ID, etc.). However, both the ADOPT cohort and the CRC cohort describe the histopathological diagnosis, information on the TNM classification and the age at diagnosis. The clinical parameters, which are mainly described in the ADOPT cohort, are not so easy to collect from our side, as a lot of information is not available in the systems used. All findings and examinations carried out in private practice or in other hospitals that are not part of KAGES or Medical University of Graz are not accessible at the Diagnostic and Research Institute of Pathology in Graz. While the data fields from diagnostics and from the projects relate more to patient data and the associated examinations, processes and therapies, the focus of MIABIS Core, MIABIS Sample+Donor+Event and MIABIS Digital Pathology is

more on technical aspects. In particular, the MIABIS Digital Pathology extension provides a solid foundation for fields related to the description of WSIs. At the time of the analysis, no information regarding this area was stored in PAS Xanthos. The fields available in MIABIS Core are less specific, which makes them difficult to search for. Since the aim of this work is to collect data fields for an easily searchable catalog, the technical parameters of MIABIS Core are probably less suitable for achieving this goal.

4.3 Challenges and Limitations

The development of the metadata catalog was shaped by a constant balancing act between scientific precision, technical standardization, and practical everyday clinical applicability, which resulted in specific methodological and structural challenges.

4.3.1 Discussion of challenges faced during the research

One of the challenges was reaching a consensus among all the stakeholders involved regarding the selection of data fields. In an interdisciplinary team, there are inevitably differing priorities when it comes to expectations and requirements for creating a metadata catalog.

Pathologists working in diagnostics place a high value on practicality. For them, the metadata catalog must under no circumstances make it more time-consuming or complicated. They place great importance on diagnostic accuracy. To this end, they prefer fields that help clearly describe tumors, such as fields for grading or TNM staging. Many pathologists prefer free-text fields or dropdown lists with a limited number of options, provided this is feasible given the wide range of diagnoses and biological variability. Additional required fields can be perceived as administrative hurdles that unnecessarily bloat the diagnostic workflow, which would further exacerbate the staffing situation. Medical researchers, on the other hand, tend to prefer structured information that is not particularly relevant in routine diagnostic practice. There is significant interest here in linking data with molecular pathology, genetic data, or specific biomarkers. Longitudinal data is also of interest to this group, as it is well-suited for identifying correlations. Among data

scientists, the primary focus is on data quality. Machine readability - achieved through coded fields - is a key priority. Technical metadata and information about the files generated during the scanning process are also relevant here. Data scientists want a data collection that is as complete as possible - they prefer a few fields that are fully populated over many fields that are only partially filled in. Only when these requirements are met, meaningful research can be conducted using AI models or algorithms.

The common goal of the workshop was therefore to meet the needs of the various professional groups as effectively as possible without unnecessarily bloating the metadata catalog. The aim here was to structure and standardize the data, but only to the extent that the diagnostic process is not adversely affected. Information that is irrelevant to routine diagnosis and would unnecessarily prolong the process has been excluded as far as possible. Discussions took place regarding which fields should be filled in as free text, which in a structured format, and which in a coded format.

Free text reporting offers several advantages in clinical practice. It enables a high degree of flexibility and allows detailed descriptions to be recorded for unique cases. It also supports expressiveness. Users can record clinical interpretations and observations in various nuances. Free text reports are easy to use. Virtually no training is required, and it is not necessary to stick to predefined formats. However, these advantages are accompanied by certain challenges. Due to the high variability of the inputs, difficulties arise if standardization is desired. It is difficult to compare findings from different cases or different institutions. Extracting data is also difficult with free text reporting if structured data is required for research or analysis. The variability in the language and terminology used between individual users can lead to inconsistencies. Various abbreviations in the medical field in particular can lead to problems. Free text reporting therefore scores with a high degree of flexibility and the ability to describe the situation in detail. However, the biggest disadvantage is the lack of structure and therefore also the comparability of the data.

Structured reporting, however, makes it easier to use data for subsequent analyses and research. Thanks to the predefined input options, the extracted data is available in standardized form. This ensures uniformity and comparability of reports. Standardization increases the quality of the data extracted and simplifies the integration of the data into databases. Structured data facilitates the integration of guidelines and automated alerts when clinical decision support systems are used. The predefined templates also speed up the entry and thus make the documentation process more efficient. However, structured reporting also brings limitations, such as a reduction in flexibility. Complex cases may not be described in sufficient detail and therefore cannot be presented in their full range. The input also requires more training and familiarity with the available templates. The learning curve is significantly flatter here compared to free text input. Predefined templates inevitably lead to a simplification of the presentation. There is a risk here that important details cannot be recorded and are therefore lost in the course of documentation. Structured reporting therefore sacrifices details for structure. The Diagnostic and Research Institute of Pathology in Graz uses pre-formatted templates for reporting on biomarkers.

Coded reporting refers to the structured and standardized documentation of medical information using defined code systems such as ICD or SNOMED-CT. Instead of free text, medical findings, diagnoses, procedures or laboratory results are recorded in coded form, making it possible to make the content machine-readable, comparable and analyzable - both within a system and between different systems and institutions. A major advantage of coded reporting is its interoperability. Different institutions worldwide can exchange data and use it in different systems if standardized codes are used. The standardized form of the data also enables automated analysis of the data. This is helpful in research, quality control and also for trend analysis. In addition, entries in coded form increase precision and reduce the risk of misinterpretation. Of course, coded reporting also has its drawbacks. The complexity of the systems used for coding requires detailed knowledge of the respective systems. You need to be familiar with both the individual diseases and the coding in the used code system in order to be able to document correct diagnoses in coded form. As with structured

reporting, the nuances are also lost here. By being limited to one coding, you may lose the ability to fully cover more complex cases and provide context for the report. The costs and resources involved in setting up such coding systems should also not be neglected. The same applies to coded reporting as to structured reporting, only here again more clearly: The details are compromised for the structure. However, the standardization here is so high that you can work with this data practically anywhere in the world. When creating reports for tumor diseases, the Diagnostic and Research Institute of Pathology in Graz uses the UICC-TNM coding system in specific data fields within templates.

In diagnostics, a lot of data is only available in unstructured form. This is because a structured form of the data is not necessary to create a clinically applicable diagnosis. Since the main focus in diagnostics is of course on the rapid creation of a correct diagnosis, which should also be described in detail, it stands to reason that free text reporting is predominant here. Structured data collection would have no added value for the establishment of a diagnosis; it would possibly even make it more difficult because it could restrict the clinician in the detailed presentation of the case. Furthermore, structured reporting does not offer any added value for the patient or the clinician, unless the clinician is conducting research in the field and would benefit from the data in structured form. Structured reporting is also very time-consuming when compared to free text reporting. And even though SNOMED-CT codes offer a wide variety of clinical terms, integrating such a standard - which lacks human-readable coding - poses certain risks in terms of diagnostic accuracy.

In the field of research, however, structured or even coded data offers a significant advantage. For example, when creating cohorts and working with AI models or algorithms, researchers benefit greatly from high-quality, complete datasets in structured or coded form. This also facilitates collaboration across different institutions.

The difficulty in this phase of the work, therefore, was to harmonize the flexibility of free-text fields - which is essential for diagnostic nuance - with the consistency of standardized coding required for automated research within a combined model.

Challenges also arise when mapping the existing values from the material list in PAS Xanthos to SNOMED-CT codes. It is important to note that, despite the large number of existing SNOMED-CT IDs, there are entities that cannot be adequately described using this nomenclature. In such cases, it is often necessary to select a higher-level category of the specimen material in order to code the field. However, the specimens are not described with consistent granularity and precision. For example, no SNOMED-CT code can be found that adequately describes a specimen from the pharynx. As a compromised solution, the code for “Upper respiratory tissue specimen” is assigned to the field, even though this is not the area where tissue from the pharynx would typically be classified.

Another challenge that occurred in the course of this work was the coding of the staining and IHC procedures. While the origin of the biological samples could be coded using SNOMED-CT codes, obstacles were encountered when attempting to do the same with the staining and IHC methods. After extensive research, no database or standard was found that includes the commonly used stainings and antibody clones for IHC analyses. To overcome this obstacle, inspiration was drawn from the MIABIS Digital Pathology Standard. Here, the staining methods are listed under “CID 8112 Specimen Stain” and mapped to SNOMED-CT codes. A free-text field (“Antibody name”) is available to accurately record the vendor-stated name of the antibodies used for IHC examinations. Another field (“Antibody identifier”) allows the antibody used in an assay to be coded using the ID from the https://www.antibodyregistry.org/, assuming the antibody is available in the registry. If the antibody is not listed in the registry, the name of the antibody clone can be entered as free text in an additional field (“Antibody clone freetext”).

4.3.2 Limitations of the study

Like any piece of work, this one also comes with limitations. One of these is the representativeness and bias of the workshop. Given the small number of

participants and the restriction to institutes affiliated with universities, it cannot be assumed that the requirements and wishes identified here reflect those of all pathologists, researchers, and data scientists across Austria. Employees in smaller, more peripheral hospitals would likely express different requirements than those identified in this work. Even within the field itself, it is not unlikely that certain subcategories of pathology - such as cytology - may be underrepresented, while the focus may lie on the pathology of solid-organ tumors.

Furthermore, this thesis merely describes the structure and data fields of the metadata catalog on paper. The technical implementation and integration into routine diagnostic procedures are important practical steps necessary to verify the practicality of the metadata catalog in real-world settings. Since the technical implementation is not part of this study, there is consequently no practical testing. Therefore, there is no information available on how long it actually takes a pathologist in their daily work to fill out the fields in the proposed format, or on how useful a dataset generated by completing this catalog actually is for research.

Another issue that could cause difficulties in the future is the evolving nature of medical standards. Medical classifications such as TNM or standards such as the ICD are subject to constant change and further development. This catalog represents a momentary snapshot and currently reflects the latest standards. In the event of changes to medical classifications and standards, either manual post-processing or an automated process is required to incorporate these updates into the catalog. Neither of these is covered within the scope of this work.

4.4 Implications for Pathologists and Researchers

One change that pathologists and researchers can expect in association with this metadata catalog is a shift in the way information is entered as part of the routine diagnostic process. While structured data entry initially requires additional effort, it leads to significant overall time savings in clinical and scientific practice by eliminating manual search processes and reducing redundancies.

Furthermore, the catalog serves as a technical enabler for the FAIR principles. By using unique identifiers and standardized codes, such as SNOMED-CT codes, samples can be easily located across databases. This fulfills the “Findable” aspect of the FAIR principles. The clear structure specifies which metadata is accessible to which group of people, thereby upholding the principle of accessibility. The aspect of interoperability is the most noticeable feature in this catalog. By mapping to international standards such as SNOMED-CT, this dataset is comparable to other datasets that use the same standards worldwide. The issue of reusability is also addressed. The data can be used for research for years to come, as the metadata provides detailed information about how the data was obtained.

However, this catalog is probably most useful for research. By completely filling out the catalog and structuring and coding selected data fields, it is possible to collect high-quality data. This high-quality data is essential for advancing research using AI models and algorithms. The structured format of the data is a prerequisite for this, as it must be machine-readable. Having this structured and coded data eliminates the need to “clean up” free-text data, a process that consumes a lot of resources.

Since this structure is used by several Austrian institutions to create a local metadata catalog, there is a high degree of interoperability among the participating institutions. Data exchange between institutions for the purpose of conducting multicenter studies, as well as the creation of large-scale research cohorts across institutional boundaries, is enabled by the compatible data structures of the individual institutions. This allows for studies with larger sample sizes than a single institution could manage on its own, thereby improving the quality of the studies and the significance of the results.

4.5 Future Work

An aspect that was not included in this work is that existing data in the diagnostic system must be organized into the structure required for integration into the metadata catalog. Future projects could aim to modify the input fields and structure of the PAS Xanthos LIMS so that data is captured directly in the format

required for transfer to the catalog. This would eliminate the need for time-consuming and error-prone manual post-processing and mapping to the new structure.

Another potential development based on this work would be a tool that enables the automatic export of data from the PAS Xanthos diagnostic system and its import into the OpenSpecimen cataloging software. This would require a tool that extracts the data from the diagnostic system, converts it into the appropriate format, and then loads it into the target system. The development of this ETL (Extract, Transform, Load) tool is necessary for populating the metadata catalog as part of daily diagnostic procedures.

To further structure and code the data, it is necessary to find a solution for coding staining methods and IHC techniques. Since this work has not identified a suitable standard that would allow for effective coding, developing a standard for staining methods and IHC antibodies would be a potential next step that could lead to a higher density of coded data.

To improve the searchability of the proposed datasets, three complementary approaches could be established in future: searching within already structured datasets, developing retrieval-augmented generation systems based on diagnostic reports, and the direct application of large language models to free-text fields. The latter two approaches benefit from upstream structured pre-filtering, which reduces the number of patients or cases to be analyzed. This allows the context to be processed to be limited, which improves both the efficiency and the reliability of the search.

4.6 Final thoughts on the importance and potential of the metadata catalog

The metadata catalog for whole-slide images developed as part of this work was not intended merely to be a list of medical and technical parameters, but rather to serve as the backbone of digital pathology. In modern pathology, image data is an essential resource in medical research. Its quality is a key determinant of the value

of this resource. Without an appropriate structure and hierarchical organization, it is not possible to fully realize the potential of WSIs. It is only through the relations and codings defined in the catalog that a searchable dataset is created that can provide scientific value. This catalog provides a common language, particularly within the Austrian research community. It enables Austrian pathologists and researchers to efficiently identify cases across institutional boundaries and assemble cohorts for multicenter studies in a significantly reduced amount of time compared to the past. The catalog also provides a foundation for future research in the field of AI. The structure it establishes ensures that the data required for training is not only available but also contextually accurate from a medical perspective. In summary, it can be stated that this catalog establishes the infrastructural foundation necessary to complete the digital transformation of pathology - moving beyond the mere digitization of glass slides toward data-driven precision medicine.

Bibliography

ABELS, E., PANTANOWITZ, L., AEFFNER, F., ZARELLA, M. D., VAN DER LAAK, J., BUI, M. M., VEMURI, V. N., PARWANI, A. V., GIBBS, J., AGOSTO-ARROYO, E., BECK, A. H. & KOZLOWSKI, C. 2019. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *The Journal of Pathology,* 249**,** 286-294.

ANAGNOSTAKIS, A., PAPPAS, A., SUCAET, Y. & WAELPUT, W. 2014. Digital Pathology Data Brokerage: A Standard Recommendation for Complex Digital Pathology Information Web-Services. *Analytical Cellular Pathology,* 2014**,** 1-2.

ANDREW H. SONG, G. J., DREW F.K. WILLIAMSON, MING Y. LU, ANURAG VAIDYA, TIFFANY R. MILLER, FAISAL MAHMOOD 2023. Artificial Intelligence for Digital and Computational Pathology. *Nature Reviews Bioengineering,* 1**,** 930–949.

BECKER, T., ROUSSEAU, A.-J., GEUBBELMANS, M., BURZYKOWSKI, T. & VALKENBORG, D. 2023. Decision trees and random forests. *American Journal of Orthodontics and Dentofacial Orthopedics,* 164**,** 894-897.

BRIERLEY, J. D., GIULIANI, M., O'SULLIVAN, B., ROUS, B. & VAN EYCKEN, E. (eds.) 2025. *TNM Classification of Malignant Tumours,* Oxford: Wiley-Blackwell.

BRUCH, L. A., DE YOUNG, B. R., KREITER, C. D., HAUGEN, T. H., LEAVEN, T. C. & DEE, F. R. 2009. Competency assessment of residents in surgical pathology using virtual microscopy. *Human Pathology,* 40**,** 1122-1128.

BUNDESINSTITUT FÜR ARZNEIMITTEL UND MEDIZINPRODUKTE. 2026. *ICD-11 – Die 11. Revision der ICD* [Online]. Available: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-11/_node.html [Accessed 28 April 2026].

BUNDESMINISTERIUM FÜR SOZIALES, G., PFLEGE UND KONSUMENTENSCHUTZ (BMSGPK) 2025. Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision – BMSGPK-Version 2025+.

CHEN, R. J., LU, M. Y., CHEN, T. Y., WILLIAMSON, D. F. K. & MAHMOOD, F. 2021. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng,* 5**,** 493-497.

CHEN, Z. W., KOHAN, J., PERKINS, S. L., HUSSONG, J. W. & SALAMA, M. E. 2014. Web-based oil immersion whole slide imaging increases efficiency and clinical team satisfaction in hematopathology tumor board. *Journal of Pathology Informatics,* 5**,** 41.

CIFCI, D., VELDHUIZEN, G. P., FOERSCH, S. & KATHER, J. N. 2023. AI in Computational Pathology of Cancer: Improving Diagnostic Workflows and Clinical Outcomes? *Annual Review of Cancer Biology,* 7**,** 57-71.

CLUNIE, D. A. 2021. DICOM Format and Protocol Standardization-A Core Requirement for Digital Pathology Success. *Toxicol Pathol,* 49**,** 738-749.

EKLUND, N., ANDRIANARISOA, N. H., VAN ENCKEVORT, E., ANTON, G., DEBUCQUOY, A., MÜLLER, H., ZAHARENKO, L., ENGELS, C., EBERT, L., NEUMANN, M., GEERAERT, J., T'JOEN, V., DEMSKI, H., CABOUX, É.,

PROYNOVA, R., PARODI, B., MATE, S., VAN IPEREN, E., MERINO-MARTINEZ, R., QUINLAN, P. R., HOLUB, P. & SILANDER, K. 2020. Extending the Minimum Information About BIobank Data Sharing Terminology to Describe Samples, Sample Donors, and Events. *Biopreservation and Biobanking,* 18**,** 155-164.

EKLUND N, E. C., NEUMANN M, STRUG A, VAN ENCKEVORT E, BABER R, BLOEMERS M, DEBUCQUOY A, VAN DER LUGT A, MÜLLER H, PARKKONEN L, QUINLAN PR, URWIN E, HOLUB P, SILANDER K, ANTON G. 2024. Update of the Minimum Information About BIobank Data Sharing (MIABIS) Core Terminology to the 3rd Version. *Biopreservation and Biobanking***,** 346-362.

EVANS, A. J., KRUPINSKI, E. A., WEINSTEIN, R. S. & PANTANOWITZ, L. 2015. 2014 American Telemedicine Association clinical guidelines for telepathology: Another important step in support of increased adoption of telepathology for patient care. *J Pathol Inform,* 6**,** 13.

FERREIRA, R., MOON, B., HUMPHRIES, J., SUSSMAN, A., SALTZ, J., MILLER, R. & DEMARZO, A. 1997. The Virtual Microscope. *Proc AMIA Annu Fall Symp***,** 449-53.

GHOSH, A., BROWN, G. T. & FONTELO, P. 2017. Telepathology at the Armed Forces Institute of Pathology: A Retrospective Review of Consultations From 1996 to 1997. *Archives of Pathology & Laboratory Medicine,* 142**,** 248-252.

HARTMAN, D. J., PANTANOWITZ, L., MCHUGH, J. S., PICCOLI, A. L., OLEARY, M. J. & LAURO, G. R. 2017. Enterprise Implementation of Digital Pathology: Feasibility, Challenges, and Opportunities. *Journal of Digital Imaging,* 30**,** 555-560.

HOSSEINI, M. S., BEJNORDI, B. E., TRINH, V. Q.-H., CHAN, L., HASAN, D., LI, X., YANG, S., KIM, T., ZHANG, H., WU, T., CHINNIAH, K., MAGHSOUDLOU, S., ZHANG, R., ZHU, J., KHAKI, S., BUIN, A., CHAJI, F., SALEHI, A., NGUYEN, B. N., SAMARAS, D. & PLATANIOTIS, K. N. 2024. Computational pathology: A survey review and the way forward. *Journal of Pathology Informatics,* 15**,** 100357.

HUISMAN, A., LOOIJEN, A., VAN DEN BRINK, S. M. & VAN DIEST, P. J. 2010. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. *Hum Pathol,* 41**,** 751-7.

IBRAHIM, A., GAMBLE, P., JAROENSRI, R., ABDELSAMEA, M. M., MERMEL, C. H., CHEN, P.-H. C. & RAKHA, E. A. 2020. Artificial intelligence in digital breast pathology: Techniques and applications. *The Breast,* 49**,** 267-273.

INDU, M., RATHY, R. & BINU, M. P. 2016. "Slide less pathology": Fairy tale or reality? *J Oral Maxillofac Pathol,* 20**,** 284-8.

JAHN, S. W., PLASS, M. & MOINFAR, F. 2020. Digital Pathology: Advantages, Limitations and Emerging Perspectives. *Journal of Clinical Medicine,* 9**,** 3697.

JAIN, E., PATEL, A., PARWANI, A. V., SHAFI, S., BRAR, Z., SHARMA, S. & MOHANTY, S. K. 2024. Whole Slide Imaging Technology and Its Applications: Current and Emerging Perspectives. *International Journal of Surgical Pathology,* 32**,** 433-448.

JEONG, J., KIM, S., PAN, L., HWANG, D., KIM, D., CHOI, J., KWON, Y., YI, P., JEONG, J. & YOO, S. J. 2025. Reducing the workload of medical diagnosis

through artificial intelligence: A narrative review. *Medicine (Baltimore),* 104**,** e41470.
KEIGHLEY, J., DE KAMPS, M., WRIGHT, A. & TREANOR, D. Digital pathology whole slide image compression with vector quantized variational autoencoders. 2023 2023. SPIE.
KIRAN, N., SAPNA, F., KIRAN, F., KUMAR, D., RAJA, F., SHIWLANI, S., PALADINI, A., SONAM, F., BENDARI, A., PERKASH, R. S., ANJALI, F. & VARRASSI, G. 2023. Digital Pathology: Transforming Diagnosis in the Digital Age. *Cureus*.
KONDYLAKIS, H., CIARROCCHI, E., CERDA-ALBERICH, L., CHOUVARDA, I., FROMONT, L. A., GARCIA-AZNAR, J. M., KALOKYRI, V., KOSVYRA, A., WALKER, D., YANG, G. & NERI, E. 2022. Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks. *European Radiology Experimental,* 6.
KORZYNSKA, A., ROSZKOWIAK, L., ZAK, J. & SIEMION, K. 2021. A review of current systems for annotation of cell and tissue images in digital pathology. *Biocybernetics and Biomedical Engineering,* 41**,** 1436-1453.
LAROBINA, M. & MURINO, L. 2014. Medical image file formats. *J Digit Imaging,* 27**,** 200-6.
LAURENT, C., GUÉRIN, M., FRENOIS, F. X., THURIES, V., JALABERT, L., BROUSSET, P. & VALMARY-DEGANO, S. 2013. Whole-slide imaging is a robust alternative to traditional fluorescent microscopy for fluorescence in situ hybridization imaging using break-apart DNA probes. *Hum Pathol,* 44**,** 1544-55.
LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE,* 86**,** 2278-2324.
MADABHUSHI, A. & LEE, G. 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis,* 33**,** 170-175.
MCINTOSH, L. D., SHARMA, M. K., MULVIHILL, D., GUPTA, S., JUEHNE, A., GEORGE, B., KHOT, S. B., KAUSHAL, A., WATSON, M. A. & NAGARAJAN, R. 2015. caTissue Suite to OpenSpecimen: Developing an extensible, open source, web-based biobanking management system. *Journal of Biomedical Informatics,* 57**,** 456-464.
MERINO-MARTINEZ, R., NORLIN, L., VAN ENCKEVORT, D., ANTON, G., SCHUFFENHAUER, S., SILANDER, K., MOOK, L., HOLUB, P., BILD, R., SWERTZ, M. & LITTON, J.-E. 2016. Toward Global Biobank Integration by Implementation of the Minimum Information About Biobank Data Sharing (MIABIS 2.0 Core). *Biopreservation and Biobanking,* 14**,** 298-306.
MILDENBERGER, P., EICHELBERG, M. & MARTIN, E. 2002. Introduction to the DICOM standard. *Eur Radiol,* 12**,** 920-7.
NAM, S., CHONG, Y., JUNG, C. K., KWAK, T.-Y., LEE, J. Y., PARK, J., RHO, M. J. & GO, H. 2020. Introduction to digital pathology and computer-aided pathology. *Journal of Pathology and Translational Medicine,* 54**,** 125-134.
NORLIN, L., FRANSSON, M. N., ERIKSSON, M., MERINO-MARTINEZ, R., ANDERBERG, M., KURTOVIC, S. & LITTON, J. E. 2012. A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. *Biopreserv Biobank,* 10**,** 343-8.

PANTANOWITZ, L., FARAHANI, N. & PARWANI, A. 2015. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 23.
PREWITT, J. M. S. & MENDELSOHN, M. L. 1966. THE ANALYSIS OF CELL IMAGES. *Annals of the New York Academy of Sciences,* 128, 1035-1053.
RUITER, D. J., ROALD, B., UNDERWOOD, J. & PRAT, J. 2004. Histopathology training in Europe: a lesson for other specialties? *Virchows Arch,* 444, 278-82.
SACO, A., BOMBI, J. A., GARCIA, A., RAMÍREZ, J. & ORDI, J. 2016. Current Status of Whole-Slide Imaging in Education. *Pathobiology,* 83, 79-88.
SHMATKO A, G. L. N., GERSTUNG M, KATHER JN. 2022. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer,* 3, 1026-1038.
SNOMED INTERNATIONAL. 2024a. *About SNOMED International* [Online]. Available: https://www.snomed.org/about-us [Accessed 28 April 2026].
SNOMED INTERNATIONAL. 2024b. *SNOMED CT Editorial Guide* [Online]. Available: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-editorial-guide [Accessed 28 April 2026].
SNOMED INTERNATIONAL. 2024c. *SNOMED CT Value Proposition* [Online]. Available: https://www.snomed.org/value-proposition [Accessed 28 April 2026].
SNOMED INTERNATIONAL. 2024d. *What is SNOMED CT?* [Online]. Available: https://www.snomed.org/what-is-snomed-ct [Accessed 28 April 2026].
STURM, B., CREYTENS, D., COOK, M. G., SMITS, J., VAN DIJK, M., EIJKEN, E., KURPERSHOEK, E., KÜSTERS-VANDEVELDE, H. V. N., OOMS, A., WAUTERS, C., BLOKX, W. A. M. & VAN DER LAAK, J. 2019. Validation of Whole-slide Digitally Imaged Melanocytic Lesions: Does Z-Stack Scanning Improve Diagnostic Accuracy? *J Pathol Inform,* 10, 6.
WARTH, A., STENZINGER, A., ANDRULIS, M., SCHLAKE, W., KEMPNY, G., SCHIRMACHER, P. & WEICHERT, W. 2016. Individualized medicine and demographic change as determining workload factors in pathology: quo vadis? *Virchows Arch,* 468, 101-8.
WEINSTEIN RS, D. M., LIANG C, BHATTACHARYYA AK, GRAHAM AR, DAVIS JR, SCOTT KM, RICHTER L, KRUPINSKI EA, SZYMUS J, KAYSER K, DUNN BE. 2001. Telepathology overview: From concept to implementation. *Human Pathology,* 32, 1283-1299.
WEINSTEIN, R. S., GRAHAM, A. R., RICHTER, L. C., BARKER, G. P., KRUPINSKI, E. A., LOPEZ, A. M., ERPS, K. A., BHATTACHARYYA, A. K., YAGI, Y. & GILBERTSON, J. R. 2009. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. *Hum Pathol,* 40, 1057-69.
WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., BOITEN, J. W., DA SILVA SANTOS, L. B., BOURNE, P. E., BOUWMAN, J., BROOKES, A. J., CLARK, T., CROSAS, M., DILLO, I., DUMON, O., EDMUNDS, S., EVELO, C. T., FINKERS, R., GONZALEZ-BELTRAN, A., GRAY, A. J., GROTH, P., GOBLE, C., GRETHE, J. S., HERINGA, J., T HOEN, P. A., HOOFT, R., KUHN, T., KOK, R., KOK, J., LUSHER, S. J., MARTONE, M. E., MONS, A., PACKER, A. L., PERSSON, B., ROCCA-SERRA, P., ROOS, M., VAN SCHAIK, R., SANSONE, S. A., SCHULTES, E., SENGSTAG, T., SLATER,

T., STRAWN, G., SWERTZ, M. A., THOMPSON, M., VAN DER LEI, J., VAN MULLIGEN, E., VELTEROP, J., WAAGMEESTER, A., WITTENBURG, P., WOLSTENCROFT, K., ZHAO, J. & MONS, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data,* 3**,** 160018.
WORLD HEALTH ORGANIZATION 2013. International Classification of Diseases for Oncology. 3rd ed. ed.
WORLD HEALTH ORGANIZATION 2019a. ICD-11 Implementation or Transition Guide.
WORLD HEALTH ORGANIZATION 2019b. International Statistical Classification of Diseases and Related Health Problems: Instruction Manual. *Tenth Revision.* World Health Organization.
WORLD HEALTH ORGANIZATION. 2022a. *ICD-11 2022 release* [Online]. Available: https://www.who.int/news/item/11-02-2022-icd-11-2022-release [Accessed 28 April 2026].
WORLD HEALTH ORGANIZATION 2022b. ICD-11 Reference Guide.
WORLD HEALTH ORGANIZATION. 2026. *International Statistical Classification of Diseases and Related Health Problems (ICD)* [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/international-classification-of-diseases [Accessed 28 April 2026].
ZARELLA, M. D., BOWMAN, DOUGLAS, AEFFNER, F., FARAHANI, N., XTHONA, ALBERT, ABSAR, S. F., PARWANI, A., BUI, M. & HARTMAN, D. J. 2019. A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association. *Archives of Pathology & Laboratory Medicine,* 143**,** 222-234.

The following tools were used to translate and optimize the language of the text:

• DeepL Translator (Version 23.5 – 26.5.1)

• Publisher/Provider: DeepL SE

• Date the content was generated: May 2023 – June 2026

• Uniform resource locator (URL): https://www.deepl.com/de/translator

• DeepL Write (Version 26.4)

• Publisher/Provider: DeepL SE

• Date the content was generated: 15 April 2026

• Uniform resource locator (URL): https://www.deepl.com/de/write

Appendix

Data fields PAS Xanthos

The date of birth is represented by **"dbo"."Patient"."Geburtsdatum"** with a data type of datetime. The field **"dbo"."Patient"."Geschlecht"** represents the sex of the patient and is stored as an integer value. The possible entries include: *female (f), male (m), diverse (d), intersex (i), unknown, or not applicable*. The field **"dbo"."Patient"."idPerson"** corresponds to the patient's unique identifier, also referred to as the P-number. It is stored as a bigint data type. The field **"dbo"."Zuweisung"."Eingangsdatum"** represents the date and time of the event or admission. It is stored in the datetime format. The field **"dbo"."Zuweisung"."Anforderungsnummer"** represents the request ID. It is stored as a nvarchar(50) data type. The field **"dbo"."MaterialEingang"."ENummer"** represents the container ID. It is stored as an int data type and is used to uniquely identify a specific container or sample unit received in the system. The field **"dbo"."Material"."Kuerzel"** represents the type of material or specimen, indicated by an abbreviated code. It is stored as a nvarchar(20) data type and is used to specify the material associated with a particular sample or entry. The field **"dbo"."Material"."Bezeichnung"** represents a free-text description of the material or specimen. It is stored as a nvarchar(100) data type and allows for a more detailed or customized designation of the material beyond the standard abbreviated code. The field **"dbo"."Gewinnungsart"."Bezeichnung"** describes the method of collection, such as biopsy, resection, tissue microarray (TMA), or cytology. It is stored as a nvarchar(100) data type.

The field **[Xanthos_Mug].[dbo].[MaterialEingang_Codes_TNM] -> Textart** refers to the macroscopic description of the specimen. It provides a textual summary of the visible characteristics of the sample as observed without a microscope, such as size, shape, color, or notable features. This field is typically used for documentation purposes during the gross examination of the material. This field also represents the histological description. Text blocks in the form of a gap text are provided for this purpose, which can be filled in by the pathologist. There are also fields in this gap text where you can choose from a preselection.

These text blocks are used by pathologists to make quick and comparable diagnoses. However, they can always be overwritten with free text if more detailed information is required. The molecular pathological description is also covered in this field.

These fields from the table **"dbo"."MaterialEingang_Codes_TNM"** are related to the TNM (Tumor, Node, Metastasis) staging system, which is used for cancer staging. **"TPrefix"** represents the prefix for the "T" classification (tumor size/extent), stored as nvarchar(20). **"TPostfix"** represents the postfix for the "T" classification, stored as nvarchar(20). **"TErgaenzung"** represents additional information or details for the "T" classification, stored as nvarchar(50). **"NPrefix"** represents the prefix for the "N" classification (node involvement), stored as nvarchar(20). **"NPostfix"** represents the postfix for the "N" classification, stored as nvarchar(20). **"NErgaenzung"** represents additional information or details for the "N" classification, stored as nvarchar(50). **"MPrefix"** represents the prefix for the "M" classification (metastasis), stored as nvarchar(20). **"MPostfix"** represents the postfix for the "M" classification, stored as nvarchar(20). **"MErgaenzung"** represents additional information or details for the "M" classification, stored as nvarchar(50). **"Stadium"** represents the overall stage of the cancer (based on the TNM system), stored as nvarchar(20). **"UICC"** represents the UICC (Union for International Cancer Control) staging code, stored as int. This code corresponds to the standardized cancer stage according to the UICC guidelines.

The ICD-O-T and ICD-O-M code-value is represented as **"dbo"."MaterialEingang_Codes_Sonstige"."idCode"** with the data type nvarchar(20). The "Block ID" is represented by **"dbo"."Laborbuch"."Nummer"** with the data type int. The terms “Block”, “OT”, and “Färbung” refer to various rows within the "Laborbuch" (laboratory book). The fixation type, such as formalin or frozen section, is represented by **"dbo"."Zusatzfeld"."id"**. The block label, which can include values like *TU, 1, My, E, Po, Para, AR, OR, LY*, *etc.*, is represented by **"dbo"."Laborbuch"."Text"**. It is combined with the **"Laborbuch Type.Bezeichnung B"** and has a data type of nvarchar(200). The slide label, which can include values like *\2, \7, etc.*, is represented by **"dbo"."Laborbuch"."Text"** with a data type of nvarchar(200). The Slide ID is represented by **"dbo"."Laborbuch"."Nummer"** with a data type of int.

Data fields ADOPT Cohort

The section “Diagnostic exam” includes a required field labeled **DIAG_CT_DONE**, which captures information about whether a CT (Computed tomography) scan has been performed. The data type is a list of predefined values: *CT - Unknown*, *CT - Done, data not available*, *CT - Done, data available*, and *CT - Not done*. The required field **DIAG_MRI_DONE** records whether an MRI (Magnetic Resonance Imaging) scan has been conducted. The data type is a list of predefined values: *MRI - Unknown*, *MRI - Done, data not available*, *MRI - Done, data available*, and *MRI - Not done*. The required field **DIAG_X_DONE** pertains to lung imaging procedures, such as chest X-rays. The field uses a predefined list of values: *Lung imaging - Unknown*, *Lung imaging - Done, data not available*, *Lung imaging - Done, data available*, and *Lung imaging - Not done*. The field **DIAG_Liver_imaging_DONE** is required and relates to liver imaging diagnostic exams. It captures whether liver imaging has been performed and whether corresponding data is available. The data type is a predefined list of values: *Liver imaging - Unknown*, *Liver imaging - Done, data not available*, *Liver imaging - Done, data available*, and *Liver imaging - Unknown Not done*. The field **DIAG_COLONOSCOPY** is required and pertains to the colonoscopy diagnostic exam. It indicates whether the procedure was performed and, if so, whether the result was positive or negative. The field uses a predefined list of values: *Colonoscopy diagnostic exam - Not done*, *Colonoscopy diagnostic exam - Negative*, and *Colonoscopy diagnostic exam - Positive*.

The section “Histopathology” includes a required field labeled **HIST_METASTASIS,** which refers to the localization of metastasis as part of the histopathological assessment. It captures specific sites where metastases may be present, using a predefined list of values: *Localization of metastasis – Others*, *Skin*, *Adrenals*, *Peritoneum*, *Pleura*, *Bone marrow*, *Lymph nodes*, *Brain*, *Hepatic*, *Osseous*, *Pulmonary*, and *None*. The field **HIST_LOCALIZATION** is required and relates to the localization of the primary tumor within the histopathology section. It uses a predefined list of values corresponding to specific ICD-10 codes: *C20*, *C19*, *C18.7*, *C18.6*, *C18.5*, *C18.4*, *C18.3*, *C18.2*, and *C18.1*. The field **HIST_MORPHOLOGY** is required and pertains to the morphological classification of the tumor in the histopathology section. It includes a predefined list of values

that represent various tumor types, such as *Adenocarcinoma*, *Mucinous carcinoma*, *Signet-ring cell carcinoma*, *Medullary carcinoma*, *High-grade neuroendocrine carcinoma*, *Large cell neuroendocrine carcinoma*, *Small cell neuroendocrine carcinoma*, *Squamous cell carcinoma*, *Adenosquamous carcinoma*, *Micropapillary carcinoma*, *Serrated adenocarcinoma*, *Spindle cell carcinoma*, *Mixed adenoneuroendocrine carcinoma*, *Undifferentiated carcinoma*, and *Other*. The field **DIGITAL_IMAGING_AVAILABILITY** is required and addresses the availability of high-resolution digital imaging from histopathology, specifically at a magnification of 40x. The predefined list of values includes: *No*, *Can be generated*, and *Readily available*. The field **DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY** is optional and refers to the availability of high-resolution digital imaging (at 40x magnification) specifically capturing the invasion front in histopathology samples. The predefined list of values includes: *Invasion front not included*, *Readily available*, *Can be generated*, and *No*.

In the section “Histopathology - TNM”, the field **TNM_DISTANT_METASTASIS** is required and pertains to the classification of distant metastasis in the TNM staging system. It uses a predefined list of values: *Distant metastasis - M1b*, *Distant metastasis - M1a*, *Distant metastasis - M1*, and *Distant metastasis - M0*. The field **TNM_REGIONAL_LYMPH_NODES** is required and relates to the classification of regional lymph nodes in the TNM staging system. It uses a predefined list of values: *Regional lymph nodes - N2b*, *Regional lymph nodes - N2a*, *Regional lymph nodes - N2*, *Regional lymph nodes - N1c*, *Regional lymph nodes - N1b*, *Regional lymph nodes - N1a*, *Regional lymph nodes - N1*, *Regional lymph nodes - N0*, and *Regional lymph nodes - NX*. The field **TNM_PRIMARY_TUMOR** is required and refers to the classification of the primary tumor in the TNM staging system. It uses a predefined list of values: *Primary Tumor - T4b*, *Primary Tumor - T4a*, *Primary Tumor - T3*, *Primary Tumor - T2*, *Primary Tumor - T1*, *Primary Tumor - Tis*, *Primary Tumor - T0*, and *Primary Tumor - TX*.

The field **UICC_STAGE** in the section “Histopathology - UICC staging” is required and refers to the staging of cancer according to the UICC (Union for International Cancer Control) system. It uses a predefined list of values: *Stage - IVB*, *Stage - IVA*, *Stage - IIIC*, *Stage - IIIB*, *Stage - IIIA*, *Stage - IIB*, *Stage - IIA*, *Stage - I*, and

Stage - 0. The field **UICC_VERSION** is required and refers to the version of the UICC (Union for International Cancer Control) staging system used for cancer staging. It uses a predefined list of values: *4th edition or earlier*, *5th edition*, *6th edition*, *7th edition*, and *Not known*.

In the section “Histopathology - WHO classification”, the field **WHO_GRADE** is required and refers to the grade of the tumor according to the World Health Organization (WHO) grading system. It uses a predefined list of values: *WHO Grading - Grade - G4*, *WHO Grading - Grade - G3*, *WHO Grading - Grade - G2*, and *WHO Grading - Grade - G1*. The field **WHO_GRADE_VERSION** is required and refers to the version of the WHO (World Health Organization) classification system used for tumor grading. It uses a predefined list of values: *Not known*, *1st edition*, *2nd edition*, *3rd edition*, and *4th edition*.

The field **MM_MICROSAT_INSTABILITY** in the section “Molecular markers” is required and refers to the presence of microsatellite instability (MSI) in the tumor. It uses a predefined list of values: *YES*, *NO*, and *NOT_DONE*. This field indicates whether microsatellites (such as BAT26, D17S250, D5S346, BAT40, D2S123, and BAT25) were analyzed and whether MSI was detected. The field **BRAF_PIC3CA_HER_MUTATION_STATUS** is optional and refers to the mutation status of the BRAF, PIC3CA, and HER2 genes. It uses a predefined list of values: *BRAF, PIC3CA, HER2 mutation status - not mutated*, *BRAF, PIC3CA, HER2 mutation status - mutated*, and *BRAF, PIC3CA, HER2 mutation status - not done*. This field captures whether mutations in these specific genes were tested and, if tested, whether they were found to be mutated or not. The field **MM_MISMATCH_REPAIR_GE** is required and pertains to the expression status of mismatch repair (MMR) genes, specifically MLH1, MSH2, PMS2, and MSH6. It uses a predefined list of values: *LOSS_OF_EXPRESSION*, *EXPRESSION*, and *NOT_DONE*. This field indicates whether the expression of these MMR genes was analyzed, and if so, whether the expression was present or lost. The field **MM_RISK_SITUATION_HNPCC** is optional and refers to the risk situation for Hereditary Non-Polyposis Colorectal Cancer (HNPCC), based on the Amsterdam criteria. It uses a predefined list of values: *true*, *false*, *yes*, *no*, *f*, and *t*. This field is used to indicate whether the individual meets the criteria for HNPCC risk. The

ontology referenced in the provided source is discussed in detail in the article https://pmc.ncbi.nlm.nih.gov/articles/PMC2933058/.

In the section “Molecular markers - KRAS mutation status”, the field **MM_KRAS_MUTATION_NRAS_EX4** is required and refers to the mutation status of NRAS exon 4, specifically codons 117 or 146. It uses a predefined list of values: *Not mutated*, *Mutated*, and *Not done*. This field is important for documenting whether mutations in these specific codons of the NRAS gene have been detected. The field **MM_KRAS_MUTATION_NRAS_EX3** is required and refers to the mutation status of NRAS exon 3, specifically codons 59 or 61. It uses a predefined list of values: *Not mutated*, *Mutated*, and *Not done*. This field is important for documenting whether mutations in these specific codons of the NRAS gene have been detected. The field **MM_KRAS_MUTATION_NRAS_EX2** is required and refers to the mutation status of NRAS exon 2, specifically codons 12 or 13. It uses a predefined list of values: *Not mutated*, *Mutated*, and *Not done*. This field is important for documenting whether mutations in these specific codons of the NRAS gene have been detected. The field **MM_KRAS_MUTATION_KRAS_EX4** is required and refers to the mutation status of KRAS exon 4, specifically codons 117 or 146. It uses a predefined list of values: *Not mutated*, *Mutated*, and *Not done*. This field is important for documenting whether mutations in these specific codons of the KRAS gene have been detected. The field **MM_KRAS_MUTATION_KRAS_EX3** is required and refers to the mutation status of KRAS exon 3, specifically codons 59 or 61. It uses a predefined list of values: *Not mutated*, *Mutated*, and *Not done*. This field is important for documenting whether mutations in these specific codons of the KRAS gene have been detected. The field **MM_KRAS_MUTATION_KRAS_EX2** is required and refers to the mutation status of KRAS exon 2, specifically codons 12 or 13. It uses a predefined list of values: *Not mutated*, *Mutated*, and *Not done*. This field is important for documenting whether mutations in these specific codons of the KRAS gene have been detected.

In the section “Patient Data”, the field **AGE_AT_PRIMARY_DIAGNOSIS** is required and refers to the patient's age at the time of the initial histopathological diagnosis, based on a biopsy or surgical specimen of the primary tumor. The age is rounded to the nearest year and is represented as a natural number, where *0* <=

x. The ontology referenced in the source provided is described in detail at http://purl.bioontology.org/ontology/SNOMEDCT/423493009. The field **DATE_DIAGNOSIS** is optional and refers to the specific date on which colon cancer was diagnosed for the first time. It follows the ISO 8601 date format (with days), ensuring a standardized representation of the diagnosis date. The field **SEX** is required and refers to the biological sex of the person, defined by chromosomes. It uses a predefined list of values: *male* and *female*. The ontology referenced in the provided source is discussed in detail at https://ontobee.org/ontology/PATO?iri=http://purl.obolibrary.org/obo/PATO_0020000. The field **CLINICAL_STUDY_PARTICIPANT** is optional and refers to whether the person is participating in a clinical study. It uses a predefined list of values: *true*, *false*, *yes*, *no*, *f*, and *t*. The field **TIME_OF_RECURRENCE_RELATIVE** is optional and refers to the time in weeks between the primary diagnosis and the diagnosis of recurrence (or metastasis). The value is represented as a natural number of weeks, where *0* <= *x*. If only months are available, the conversion to weeks is done by multiplying the number of months by 4.

The field **PHARMACOTHERAPY_END_RELATIVE** in the section “Pharmacotherapy” is required and refers to the time, in weeks, between the initial diagnosis and the end of pharmacotherapy (drug intake). The value is represented as a natural number of weeks, where *0* <= *x*. The field **PHARMACOTHERAPY_START_RELATIVE** is required and refers to the time, in weeks, between the initial diagnosis and the start of pharmacotherapy (drug intake). The value is represented as a natural number of weeks, where *0* <= *x*. The field **PHARMACOTHERAPY_SCHEME** is required and refers to the specific scheme of pharmacotherapy used in the treatment. It uses a predefined list of values, which include various pharmacotherapy schemes such as:

- *Other*
- *Scheme of pharmacotherapy - 5-FU 325-350 mg/m2 + LV 20 mg/m2 i.v. bolus, day 1-5, weeks 1 and 5*
- *Scheme of pharmacotherapy - 5-FU 400 mg/m2 + 100 mg i.v. bolus, day 1, 2, 11, 12, 21, 22*
- *Scheme of pharmacotherapy - 5-FU 225 mg/m2 i.v. continuous infusion, 5 days per week*

- *Scheme of pharmacotherapy - 5-FU 1000 mg/m2 i.v. continuous infusion, day 1-5, weeks 1 and 5*
- *Scheme of pharmacotherapy - Capecitabine 800-825 mg/m2 bid po, day 1-5, together with radiation or continuously until end of radiation*
- *Scheme of pharmacotherapy - UFT (300-340 mg/m2/day) and LV (22.5-90 mg/day) po continuously, 5(-7) days per week, together with radiotherapy*
- *Scheme of pharmacotherapy - Only preoperatively (no standard): 5-FU 250 mg/m2 i.v. continuous infusion on days 1-13 and 22-35 and oxaliplatin 50 mg/m2 i.v. day 1, 8, 22, and 29*

In the section “Radiation therapy ”, the field **RADIATION_THERAPY_END_RELATIVE** is required and refers to the time, in weeks, between the initial diagnosis and the end of radiation therapy. The value is represented as a natural number of weeks, where *0* <= *x*. The field **RADIATION_THERAPY_START_RELATIVE** is required and refers to the time, in weeks, between the initial diagnosis and the start of radiation therapy. The value is represented as a natural number of weeks, where *0* <= *x*.

The field **THERAPY_RESPONSE_TIMESTAMP_RELATIVE** in the section “Response to thearapy” is required and refers to the time, in weeks, between the initial diagnosis and the date when a response to therapy was obtained. The value is represented as a natural number of weeks, where *0* <= *x*. The field **THERAPY_RESPONSE** is required and refers to the specific response to therapy. It uses a predefined list of values, which include:

- *Specific response - Complete response*
- *Specific response - Partial response*
- *Specific response - Stable disease*
- *Specific response - Progressive disease*

In the section “Sample”, the field **SAMPLE_ID** is required and refers to a unique identifier for the sample, which is distinct within the biobank. This field is filled with free text. The field **SAMPLE_MATERIAL_TYPE** is required and refers to the type of specimen collected. It uses a predefined list of values, which includes: *Other*, *Healthy colon tissue* and *Tumor.* The field **SAMPLE_PRESERVATION_MODE** is required and refers to the method used to preserve the specimen. It uses a

predefined list of values, which includes: *Other*, *Cryopreservation* and *FFPE* (Formalin-Fixed, Paraffin-Embedded).

The field **SURGERY_RADICALITY** in the section “Surgery” is required and indicates the radicality of the surgery, specifically whether the entire tumor was removed. It includes predefined values such as *R0*, *R1*, *R2*, and *RX*, which reflect the completeness of the tumor resection. The field **SURGERY_START_RELATIVE** is required and captures the time difference, in weeks, between the initial diagnosis and the surgical intervention. The value is recorded as a natural number, where *0 <= x.* The field **SURGERY_TYPE** is required and specifies the type of surgery performed. It includes predefined values such as *Endo-rectal tumor resection*, *Abdomino-perineal resection*, *Anterior resection of rectum*, *Low anterior colon resection*, *Pan-procto colectomy*, *Total colectomy*, *Sigmoid colectomy*, *Transverse colectomy*, *Left hemicolectomy*, *Right hemicolectomy*, or *Other*. The field **SURGERY_TYPE_OTHER** is optional and allows for the specification of a surgery type if it is not included in the predefined list. It is a free-text field.

In the section “Targeted therapy”, the field **TARGETED_THERAPY_END_RELATIVE** is optional and refers to the date when targeted therapy ended, expressed as the number of weeks since the initial diagnosis. The value is a natural number where *0 <= x.* The field **TARGETED_THERAPY_START_RELATIVE** is required and indicates the start date of targeted therapy, recorded as the number of weeks since the initial diagnosis. The value must be a natural number where $0 \leq x$.

The field **OVERALL_SURVIVAL_STATUS** in the section “Vital status and survival information” is required and records the overall survival time in weeks following the start of the first colon cancer therapy. The value is a natural number where $0 \leq x$. If survival time is originally documented in months, it should be converted using the formula $weeks = months \times 4$. The field **VITAL_STATUS** is required and specifies the current vital status of the individual. It includes predefined values such as *ALIVE*, *DEATH_COLON_CANCER*, *DEATH_OTHER*, *DEATH_UNKNOWN_REASON*, and *UNKNOWN*. The field **VITAL_STATUS_TIMESTAMP** is required if the **VITAL_STATUS** is not marked

as "UNKNOWN" and optional otherwise. It refers to the timestamp of the last update regarding the vital status of the individual, formatted in ISO 8601 with days.

Data fields CRC Cohort

In the CRC cohort, the filename equals the **scanID**. An example is *MUGGRZ-PATH-SCAN-SS74561018723.json*. The **caseID** is a unique identifier for the case, represented as a UUID (Universally Unique Identifier) for each case. For example, *MUGGRZ-CASE-30e7e984-a3c1-4d5f-b013-2595be28e27c*.

The section “material” is related to the biological material of the sample and contains various fields. The field **containerNumber** is the identifier for the container holding the sample, for example *1*. The field **block** contains information about the block and is described in multiple subfields: The subfield **blockNumber** is an UUID per block, used to uniquely identify the block. For example, *MUGGRZ-PATH-BLOCK-dcf667df-5fd4-401d-beb4-6278dac20c1b*. The subfield **description** describes what is inside of the block, for example *Tumor*. The subfield **materialEmbeded** describes if any material is embedded within the block, such as *None*. The subfield **localisisation** relates to the location of material in the block, such as *Tumor*. The subfield **blockType** describes the type of block, which in this case is labeled as *virtual* due to the absence of block information. The subfield **slides** indicates whether slides have been created from this block or not.

The section “scan” contains fields like the **slideID**, which is a UUID assigned to each slide, such as *MUGGRZ-PATH-SLIDE-6e6935e7-8cf6-4b53-9529-18d1bd4e91e9*. The **scanID** is another UUID, unique to each scan, for example, *MUGGRZ-PATH-SCAN-SS74561018723*. The **ScannerID** refers to the identifier of the scanner used, such as *SS7456*. The **Timestamp** represents the Unix timestamp when the scan was performed, for instance, *1531415534*. The **Magnification** indicates the magnification setting of the scanner, such as *40* (40x magnification). The **ScanFileType** specifies the type of file generated by the scanner, which could be *aperio* or *mirax*. Finally, the **workpackage** refers to the workpackage number such as *1.2*. This structure provides detailed documentation and tracking of the scan and related metadata.

The “patient” section includes fields like the **patientID**, which is a UUID assigned to each patient, ensuring their unique identification, such as *MUGGRZ-PATIENT-*

d8f46ce9-3451-427f-baf9-c4034aada1ac. The **ICDN** is the ICD 9/10 code that represents the cause of death, for example, *A047*. The **ICDE** represents the ICD 9/10 code for the cause of death from external causes, such as *Y834*. The **gender** field is used to specify the patient's gender, marked as *W* for female or *M* for male. The **survivalStatus** indicates whether the patient is alive or dead, with values like *DEAD* or *ALIVE*. The **cutOffDay** specifies the date at which the data was cut off, such as *29.06.2018*. The **ageAtCutoff** shows the patient's age at the cutoff date, and for a patient marked as *ALIVE*, this could be an age like *70*.

The section “diagnosis” contains fields like the **icd10** field, which contains a list of all ICD 10 codes used for coding the diagnosis, for example, *"C18.7"*, *"C21.0"*, *"D12.5"*. Similarly, the **icd0** field lists all ICD O-3 codes for coding the diagnosis, such as *"8140/3"* and *"8480/3"*. The **t** field represents the T staging, such as *3A*, indicating the size and extent of the primary tumor. The **n** field is for the N staging, such as *0*, showing the regional lymph node involvement. The **m** field for M staging, indicated by *x*, represents the presence or absence of distant metastasis. The **g** field represents the G grade, such as *3*, indicating the tumor’s differentiation grade. The **r** field refers to the R staging, which assesses resection margin status, such as *0*, meaning the margin is clear. The **l** and **v** fields are for L and V staging, represented by *x* and *1*, respectively, providing additional information about local invasion and vascular invasion. The **ageAtDiagnose** indicates the patient's age at the time of diagnosis, such as *60*. The **survivalInDays** field gives the number of days the patient survived from the first diagnosis until death, such as *4171 days*. The **survivalInMonth** field indicates the number of months the patient survived from the diagnosis until the cutoff date, provided the patient is alive. The **staging** field specifies the tumor stage, such as *I*, *II*, or *III*. The **stagingDefinition** refers to the classification system used, for example, *AJCC 6th classification 2004*. The **estimatedNumberOfSlides** field provides an estimate of the number of slides according to the PAS, such as *10 slides*, and the **numberOfSlidesReceivedFromBB** indicates the actual number of slides received from the biobank for the diagnosis, like *8 slides*. The **yearOfDiagnose** specifies the year the diagnosis was made, such as *1996*.

The fields **radiationCalculatedDays** and **therapyOnkologyDepartmentCalculatedDays** both contain the subfield

daysFromDiagnosisToStart which describes the days from the diagnosis to the start of the therapy, such as *99*. The subfield **daysFromDiagnosisToEnd** describes the days from the diagnosis to the end of the therapy, such as *146*.
The "general" section contains the field **jsonVersion** which describes the version of the Json-file, like MUG_V10. The field **timestampJsonCreated** describes the unix timestamp, such as 1531415534.

Data fields MIABIS Core v3

The entity Biobank is described in MIABIS as one that stores samples and data from these samples. The attribute **MIABIS-BIOBANK-01 (ID)** is a technical identifier for the biobank that is assigned by the implementation. This ID is created in accordance with the specifications of the implementing instance. The ID of the biobank is stored as text. The field **MIABIS-BIOBANK-02 (Acronym)** refers to any abbreviations or acronyms used for the biobank and is also stored as text. **MIABIS-BIOBANK-03 (Name)** specifies the name of the biobank, preferably in English. This data is also saved as text. The attribute **MIABIS-BIOBANK-04 (URL)** contains the complete Internet address (http format) of the biobank as text. **MIABIS-BIOBANK-05 (Juristic Person)** describes the legal entity, such as a university, a company, or a local authority, that is the responsible party for the biobank in text form. For the attribute **MIABIS-BIOBANK-06 (Country)**, the two-letter country code must be specified as text, based on the ISO standard 3166 alpha2. **MIABIS-BIOBANK-07 (Contact Information)** contains structured data on the contact person or persons responsible for the biobank. **MIABIS-BIOBANK-08 (Description)** allows a free-text description of the biobank in English, with a recommended maximum length of 2000 characters. **MIABIS-BIOBANK-09 (Infrastructural Capabilities)** describes the technical infrastructure capabilities of the biobank, e.g., the availability of sample storage, data storage, or biosafety measures. The field is filled with a value from a predefined list. **MIABIS-BIOBANK-10 (Organizational Capabilities)** refers to the organizational services that the biobank can offer to support users, including, for example, the ability to recontact donors, support for clinical studies, access to omics data, laboratory analysis data, clinical data or archives (pathological, radiological), and access to national medical registries. Predefined values can be selected here. **MIABIS-**

BIOBANK-11 (Bioprocessing and Analytical Capabilities) describes the bioprocessing and analytical services offered by the biobank. These include, for example, biochemical analyses, genomics, nucleic acid extraction, proteomics, metabolomics, histology, cell line processing, virology, sample processing, shipping, and quality control of samples. Values from predefined lists are also used here. Finally, **MIABIS-BIOBANK-12 (Quality Management Standard)** records the quality management standards according to which the biobank is certified or accredited. Permissible values include ISO 20387, ISO 9001, or "Other" for other standards. The information is verified by the implementing organization, and the list of permissible values can be expanded as needed.

The MIABIS Collection category represents structured information on collections of biological samples and the associated data. The attributes in this category are used for the standardized recording and description of collections within a biobank. The attribute **MIABIS-COLLECTION-01 (ID)** is a technical identifier for the collection that is assigned by the implementation. This ID is created in accordance with the specifications of the implementing instance. The ID of the collection is stored as text. **MIABIS-COLLECTION-02 (acronym)** refers to a short abbreviation that may be used for the collection and is recorded as text. The **name of the collection (MIABIS-COLLECTION-03)** must be specified as text and should preferably be in English. Under **MIABIS-COLLECTION-04 (URL)**, a complete Internet address for the collection can be entered as free text. The **contact information (MIABIS-COLLECTION-05)** is a structured attribute that identifies the contact person responsible for the collection and must be provided. A **description of the collection (MIABIS-COLLECTION-06)** should be written in English as free text and ideally should not exceed 2000 characters. The attribute **MIABIS-COLLECTION-07 (Sample source)** specifies the source of the samples, with *Human, Animal, or Environment* as permissible values. **MIABIS-COLLECTION-08 (Sex)** describes the biological sex of the donors in the collection and can contain multiple entries (*Male, Female, Unknown, Undifferentiated, or Not applicable*). For age information, **MIABIS-COLLECTION-09 (Age Low)** and **MIABIS-COLLECTION-10 (Age High)** fields are available, which describe the age of the youngest and oldest sample donors at the time of sample collection. In addition, **MIABIS-COLLECTION-11 (Age Low Unit)** and **MIABIS-COLLECTION-12 (Age**

High Unit) specify the respective time units, e.g., *Years, Months, Weeks, Days, or Gestational weeks*. The attribute **MIABIS-COLLECTION-13 (dataset type)** describes the type of data sets associated with the samples. Permitted values include *Lifestyle dataset, Environmental dataset, Physiological dataset, Biochemical dataset, Clinical dataset, Psychological dataset, Genomic dataset, Proteomic dataset, Metabolomic dataset, Body (Radiological) image, Whole slide image, Photo image, Genealogical records, Other*. **MIABIS-COLLECTION-14 (sample type)** describes the sample type. Permitted values are *Blood, Buffy coat, Cancer cell lines, Digital sample, DNA, Entire body organ, Faeces, Embryo or fetal tissue, Immortalized cell lines, Isolated microbes, Other body fluid, Plasma, Primary cells, Post-mortem tissue, RNA, Saliva, Serum, Specimen from environment or food, Swab, Tissue (Frozen), Tissue (FFPE), Urine, Other.* **MIABIS-COLLECTION-15 (Storage temperature)** specifies the temperature at which the samples are stored long-term. Possible values include *RT (Room temperature), 2 °C to 10°C, -18 °C to -35 °C, -60 °C to -85 °C, <-135 °C, Liquid nitrogen vapor-phase, Liquid nitrogen liquid-phase, Other.* The structured attribute **MIABIS-COLLECTION-16 (Disease)** describes information about diseases or disease groups that the collection focuses on. **MIABIS-COLLECTION-17 (Sample collection setting)** describes the context in which the sample collection was conducted. Possible values are *Routine health care setting, Clinical trial, Research study, Public health/population-based study, Museum and/or archeological collection, Environment, Unknown, Other.* The attribute **MIABIS-COLLECTION-18 (Collection design)** describes the design of the collection. Permitted values are *Case-control, Cross-sectional, Longitudinal cohort, Twin-study, Quality control study, Population-based cohort, Disease-specific cohort, Birth cohort, Microbial collection (if applicable with resistance data), Reference collection, Rare disease collection, Other.* **MIABIS-COLLECTION-19 (Use & Access conditions)** includes conditions for the use and access of the samples/data in the collection. Possible values are *Commercial use, Collaboration, Specific research use, Genetic data use, Outside EU access, Xenograft, Other animal work, Other.* The **status of the collection (MIABIS-COLLECTION-20)** indicates whether the collection is still active or has already been terminated. Permitted values are *Active, Ended, Other*. The attribute **MIABIS-COLLECTION-21 (Total number of subjects)** documents

the total number of individuals included in the collection and is recorded as an integer. The **inclusion criteria (MIABIS-COLLECTION-22)** define the parameters used to select participants for the collection. The possible parameters are *Health status, Hospital patient, Use of medication, Gravidity, Age group, Familial status, Sex, Country of residence, Ethnic origin, Population representative sampling, Lifestyle/Exposure, Other.* Finally, **MIABIS-COLLECTION-23 (Publications)** provides space for listing scientific publications that have emerged from the collection as free text – with DOI if possible.

The attributes of the MIABIS entity Network describe a group of interconnected biobanks, collections, and/or research resources. The attribute **MIABIS-NETWORK-01 (ID)** is a technical identifier for the network that is assigned by the implementation. This ID is created in accordance with the specifications of the implementing instance. The ID of the network is stored as text. **MIABIS-NETWORK-02 (Acronym)** refers to a short abbreviation for the network, if used, as text. The **name of the network (MIABIS-NETWORK-03)** should be specified as text, preferably in English. The full Internet address of the network can be specified as text under **MIABIS-NETWORK-04 (URL)**. The attribute **MIABIS-NETWORK-05 (Juristic Person)** describes the legal entity, i.e. a university, a company, or a county council that is responsible for the network. The countries to which the network partners are assigned are specified in the **MIABIS-NETWORK-06 (Country)** field. The information is provided as text in ISO 3166 alpha-2 format whereby multiple entries are permitted and necessary if the network is international. The **contact information (MIABIS-NETWORK-07)** includes structured details about the network's responsible contact person. This attribute allows a free text description (recommended maximum 2000 characters) but should correspond to the structured definition of the attribute. **MIABIS-NETWORK-08 (Description)** allows a description of the network as text in English. The attribute **MIABIS-NETWORK-09 (Network status)** describes the current functional status of the network. Permitted values are *Active, Ended, Other*. **MIABIS-NETWORK-10 (Network members)** lists the names of the entities (organizations, biobanks, collections, or research institutions) participating in the network as text. The attribute **MIABIS-NETWORK-11 (Common collaboration topics)** describes the common topics or agreements on which the collaboration

within the network is based. Possible values are *Common charter, Common SOPs, Common data access policy, Common sample access policy, Common MTA, Common image access policy, Common image MTA, Common representation, Common URL, Other*. The final attribute **MIABIS-NETWORK-12 (Network type)** describes the type or content focus of the network. Permitted values are *BBMRI-ERIC National Node network, Biobank network, Collection network, Disease-specific network, Project network, Rare Disease network, Other.*
The attributes in the Research Resource category are used to describe research resources that contain biological samples and/or associated data. A research resource can combine material from different collections and different biobanks. The attribute **MIABIS- RESEARCHRESOURCE-01 (ID)** is a technical identifier for the research resource that is assigned by the implementation. This ID is created in accordance with the specifications of the implementing instance. The ID of the research resource is stored as text. **MIABIS-RESEARCHRESOURCE-02 (acronym)** refers to a short abbreviation used in text form for the research resource, if applicable. The **name of the research resource (MIABIS-RESEARCHRESOURCE-03)** should be specified as text and preferably in English. **MIABIS-RESEARCHRESOURCE-04 (URL)** provides the complete Internet address of the research resource. **MIABIS-RESEARCHRESOURCE-05 (Contact Information)** provides structured contact details for the person responsible. **MIABIS-RESEARCHRESOURCE-06 (Description)** allows a free-text description of the resource in English. **MIABIS-RESEARCHRESOURCE-07 (Sample source)** describes the origin of the samples. Permitted values are *Human, Animal, or Environment.* MIABIS-RESEARCHRESOURCE-08 (Sex) indicates the biological sex of the donors. Permitted values are *Male, Female, Unknown, Undifferentiated, or Not applicable.* The attributes **MIABIS-RESEARCHRESOURCE-09 (Age Low)** and **MIABIS-RESEARCHRESOURCE-10 (Age High)** indicate the age of the youngest and oldest donor at the time of sample collection. The age data is specified in units using the attributes **MIABIS-RESEARCHRESOURCE-11 (Age Low Unit)** and **MIABIS-RESEARCHRESOURCE-12 (Age High Unit)**, which allow the following to be specified: *Years, Months, Weeks, Days, or Gestational weeks.* **MIABIS-RESEARCHRESOURCE-13 (Dataset type)** describes the types of datasets

derived from donor information or samples. Possible values are *Lifestyle dataset, Environmental dataset, Physiological dataset, Biochemical dataset, Clinical dataset, Psychological dataset, Genomic dataset, Proteomic dataset, Metabolomic dataset, Body (Radiological) image, Whole slide image, Photo image, Genealogical records, Other.* The attribute **MIABIS-RESEARCHRESOURCE-14 (Sample type)** describes the type of samples, e.g. *Blood, Buffy coat, Cancer cell lines, Digital sample, DNA, Entire body organ, Faeces, Embryo or fetal tissue, Immortalized cell lines, Isolated microbes, Other body fluid, Plasma, Primary cells, Post-mortem tissue, RNA, Saliva, Serum, Specimen from environment or food, Swab, Tissue (Frozen), Tissue (FFPE), Urine, Other.* The **storage conditions (MIABIS-RESEARCHRESOURCE-15)** specify the temperatures at which the samples are stored long-term, in accordance with SPREC v3. Permitted values are *RT (Room temperature), 2 °C to 10°C, -18 °C to -35 °C, -60 °C to -85 °C, <-135 °C, Liquid nitrogen vapor-phase, Liquid nitrogen liquid-phase, Other*. **MIABIS-RESEARCHRESOURCE-16 (Disease)** contains structured information on the diseases or disease categories that are the focus of the research resource – multiple entries are possible. The **sample collection setting (MIABIS-RESEARCHRESOURCE-17)** describes the context in which the sample collection was conducted. Possible values are *Routine health care setting, Clinical trial, Research study, Public health/population-based study, Museum and/or archeological collection, Environment, Unknown, Other.* The attribute **MIABIS-RESEARCHRESOURCE-18 (Research Resource design)** describes the overall study design. Possible values are *Case-control, Cross-sectional, Longitudinal cohort, Twin-study, Quality control study, Population-based cohort, Disease-specific cohort, Birth cohort, Microbial collection (if applicable with resistance data), Reference collection, Rare disease collection, Other*. **MIABIS-RESEARCHRESOURCE-19 (Use & Access conditions)** describes special terms of use and access. These may include *Commercial use, Collaboration, Specific research use, Genetic data use, Outside EU access, Xenograft, Other animal work, Other.* **MIABIS-RESEARCHRESOURCE-20 (Research Resource status)** describes the status of the resource with the following possible values: *Active, Ended, Other.* **MIABIS-RESEARCHRESOURCE-21 (Total number of subjects)** indicates the total number of donors or participants in the research resource and is

recorded as an integer. The attribute **MIABIS-RESEARCHRESOURCE-22 (Inclusion criteria)** contains information on the criteria used to determine whether individuals are included in the resource. Possible criteria are *Health status, Hospital patient, Use of medication, Gravidity, Age group, Familial status, Sex, Country of residence, Ethnic origin, Population representative sampling, Lifestyle/Exposure, Other*. Finally, **MIABIS-RESEARCHRESOURCE-23 (Publications)** lists the most important publications that have emerged from the research resource as free text. Where possible, DOIs should be provided.

Data fields MIABIS Sample+Donor+Event

The “Event” section describes something that happens in relation to the sample and/or the sample donor at a certain time and place. The attribute **MIABIS-EVENT-01 (Event ID)** is a technical identifier for each event that is assigned by the database implementation. This ID is created as a coded string and serves as a random, unique reference. It is mandatory to provide this ID if an event is recorded. The attribute **MIABIS-EVENT-02 (Event date and time)** specifies the exact moment the event occurred, stored in the ISO8601 format (yyyy-mm-ddThh:mm:ss) or as a partial date. You must use either this date or the age at the event, but not both; if this field is used, the date of birth of the individual is also required. The attribute **MIABIS-EVENT-03 (Age at event)** is a decimal value representing how old the individual was when the event took place. This field is used as an alternative to the specific event date and time, and it must not be filled if a date is already provided. The attribute **MIABIS-EVENT-04 (Age at event unit)** defines the scale used for the age, chosen from a specific list (years, months, weeks, days, or gestational weeks). This information is mandatory whenever the age at the event (MIABIS-EVENT-03) is provided to ensure the numerical value is interpreted correctly.

The category “Sample donor” describes a person who serves as a source for biological samples or the digital image of a biological entity. The attribute **MIABIS-SAMPLEDONOR-01 (Sample donor ID)** is a technical identifier for the sample donor, stored as a coded string. This ID must be a unique, pseudonymised alphanumeric code within the sample collection and is a mandatory field for each donor. The attribute **MIABIS-SAMPLEDONOR-02 (Sex)** describes the biological

sex of the sample donor using enumerated values. The allowed entries are restricted to Male, Female, Unknown, Undifferentiated, or Not applicable, and it is a required field for the dataset. The attribute **MIABIS-SAMPLEDONOR-03 (Dataset type)** identifies the specific data categories available for or linked to the sample donor, such as clinical, genomic, or lifestyle datasets. This attribute allows for multiple values to represent all relevant types of information associated with the donor. The attribute **MIABIS-SAMPLEDONOR-04 (Birth date)** records the donor's date of birth in accordance with the ISO-standard 8601 (yyyy-mm-ddThh:mm:ss). While partial dates like the birth year are permitted for general use, the full date of birth is strictly required if the event date (MIABIS-EVENT-02) is utilized.

The term “sample” refers to biological material obtained from a donor or a digital image of a biological specimen. The attribute **MIABIS-SAMPLE-01 (Sample ID)** is a technical identifier for the sample within a collection, often represented by a barcode. This ID is stored as a pseudonymized, alphanumeric string and is intended for sharing; it is recommended that these IDs remain persistent within a given biobank. The attribute **MIABIS-SAMPLE-02 (Sample type)** categorizes the biological specimen intended for diagnostics, research, or treatment. It uses enumerated values ranging from fluids like plasma and serum to various tissues and isolated components like DNA or RNA, and it is a mandatory field for each sample record. The attribute **MIABIS-SAMPLE-03 (Sample storage temperature)** identifies the long-term temperature at which the sample is stored after preparation. The values are selected from a defined list (e.g., RT, -60 °C to -85 °C, or Liquid nitrogen) based on SPREC v4 and MIABIS standards. The attribute **MIABIS-SAMPLE-04 (Sample creation date and time)** records when the sample was established in its specific form. The data must follow the ISO 8601 format (yyyy-mm-ddThh:mm:ss), though partial values such as the year are permitted if detailed information is unavailable. The attribute **MIABIS-SAMPLE-05 (Anatomical site)** is a structured attribute used to describe the anatomical source of the sample material. It serves as a container for more specific sub-attributes regarding ontologies and topographical codes. The attribute **MIABIS-SAMPLE-05-01 (Anatomical site ontology)** specifies the name of the ontology used to describe the sample source, such as ICD-O-3. This field, along with the version (05-02), is required if any anatomical site information is provided. The attribute

MIABIS-SAMPLE-05-02 (Anatomical site ontology version) identifies the specific version of the selected anatomical ontology. It must be provided whenever anatomical site information is being documented. The attribute **MIABIS-SAMPLE-05-03 (Anatomical site ontology code)** stores the specific code representing the anatomical location, retrieved from the chosen ontology version. The attribute **MIABIS-SAMPLE-05-04 (Anatomical site ontology description)** provides the textual description associated with the specific anatomical site ontology code. The attribute **MIABIS-SAMPLE-05-05 (Anatomical site free text)** allows for an explanation or manual entry in cases where the anatomical site is unknown or the available ontology codes provide insufficient information. The attribute **MIABIS-SAMPLE-10 (Sample content diagnosis)** uses ICD-10 or ORPHA codes to describe the diagnosis of the sample content, such as whether it contains carcinogenic material. Multiple codes can be assigned to a single sample. The attribute **MIABIS-SAMPLE-11 (Use & Access conditions)** details the restrictions or permissions that affect sample availability, such as commercial use, genetic data use, or animal work. This attribute allows for multiple enumerated values. The attribute **MIABIS-SAMPLE-12 (Sample source)** indicates the origin from which the samples were collected or isolated. The allowed values are restricted to Human, Animal, or Environment.

Data fields MIABIS DigitalPathology

The attributes of the Assay category contain information about the containers (glass slides) and the preparation steps in the laboratory. The attribute **MIABIS-DIGITALPATHOLOGYASSAY-01 (Container Configuration)** classifies a container based on the organization of the samples it contains. Possible values are *TMA, Multiple patients, Multiple blocks, Single block, Single section, Cytology sample, Other (the value is known but does not fit into the specified categories), and NULL (the value is unknown)*. **MIABIS-DIGITALPATHOLOGYASSAY-02 (Container Identifier)** is a text field that contains a unique identifier for a container in accordance with DICOM. **MIABIS-DIGITALPATHOLOGYASSAY-03 (Specimens)** contains one or more unique identifiers of specimens contained in a container. It is a mandatory field with free text format. **MIABIS-DIGITALPATHOLOGYASSAY-04 (Container Type)** describes the type of non-

biological object used and can take the following values: *Tissue Cassette, Tissue Microarray Cassette, Microscope Slide, Specimen Container, Specimen Vial, Specimen Well, Electron Microscopy Grid*. **MIABIS-DIGITALPATHOLOGYASSAY-05 (Tissue Type)** specifies which tissue types are marked and examined. The values are taken from CID 7166 Common Tissue Segmentation Type and coded with SNOMED-RD ID. **MIABIS-DIGITALPATHOLOGYASSAY-06 (Cell Type)** describes specific cell types that have been examined. The values are taken from the Cell Annotation Platform and the Data Coordination Platform. **MIABIS-DIGITALPATHOLOGYASSAY-07 (Pathogen Type)** names the pathogenic microorganisms tested. Possible values are *viruses, bacteria, fungi, parasites, and prions.* **MIABIS-DIGITALPATHOLOGYASSAY-08 (Pathogen Identifier)** is an attribute that contains the name or ID of a tested pathogen. The values come from the National Healthcare Safety Network. **MIABIS-DIGITALPATHOLOGYASSAY-09 (Biomarker Type)** describes the tested cell component, e.g., *Protein, DNA, RNA, Carbohydrate, Lipid, Inorganic Ions, other (the value is known but cannot be classified into any of the predefined categories), NULL (the value is unknown).* **MIABIS-DIGITALPATHOLOGYASSAY-10 (Biomarker Spatial Localization)** specifies the subcellular localization of the biomarker. Possible values are *extracellularMatrix, cellMembrane, cytoplasm, nucleus, microorganism*. **MIABIS-DIGITALPATHOLOGYASSAY-11 (Assay type)** describes the laboratory test used on specimens mounted on a glass slide. Permitted values are *Hematoxylin And Eosin Stain, Special Stains, Immunohistochemistry, Immunocytochemistry, Immunofluorescence, Insitu Hybridization, Fluorescence Insitu Hybridization, Fluorescent Multiplex Immunohistochemistry, Multiplex Immunohistochemistry, other (the value is known but cannot be classified into any of the predefined categories), NULL (the value is unknown).* **MIABIS-DIGITALPATHOLOGYASSAY-12 (Special Stain Technique)** specifies the technique of special staining methods, except hematoxylin and eosin stain. The values are taken from CID 8112 Specimen Stain and are coded with SNOMED-RD ID. **MIABIS-DIGITALPATHOLOGYASSAY-13 (Dye Name)** specifies the dyes used (e.g., for special staining), based on CID 8112 Specimen Stain and SNOMED-RD ID for encoding. **MIABIS-DIGITALPATHOLOGYASSAY-14**

(Antibody name) specifies the name assigned by the manufacturer to the antibody used as a text string. **MIABIS-DIGITALPATHOLOGYASSAY-15 (Antibody identifier**) is a structured value from the antibody registry that specifies a unique ID for the antibody used in an assay. **MIABIS-DIGITALPATHOLOGYASSAY-16 (Chromogenic Substrate Name**) names the chromogenic substrate used in an assay. Permitted values are *3,3'-Diaminobenzidine, 3-Amino-9-Ethylcarbazole, 3,3´,5,5´ tetramethylbenzidine, fastRed, permanentRed, emerald, BCIP/NBT.* **MIABIS-DIGITALPATHOLOGYASSAY-17 (Fluorophore Substance Name)** names the fluorophore used. The values are taken from https://fluorophores.tugraz.at. **MIABIS-DIGITALPATHOLOGYASSAY-18 (FISH Signal Pattern**) describes the pattern of fluorescence signals in FISH analyses. Possible values are *Deletion Probes, Amplification Probes, Dual-Colour Breakapart Probes, Deletion-Fusion Probes, Tri-Colour Breakapart Probes, Translocation, Dual Fusion Probes.* **MIABIS-DIGITALPATHOLOGYASSAY-19 (FISH Probe Identifier)** contains the name of the FISH probe used, as specified by the manufacturer and saved as a string. **MIABIS-DIGITALPATHOLOGYASSAY-20 (Counter Labeling)** describes the counterstaining method used to enhance contrast. The value can be provided as a free text (string) or by using values from a predefined list like *Hematoxylin, Eosin, Nuclear Fast Red, Methyl Green, 4',6-diamidino-2-phenylindole, Hoechst 33342, PropidiumIodide, Phalloidin, other (the value is known but cannot be classified into any of the predefined categories), NULL (the value is unknown).*

The Scan category of the MIABIS Digital Pathology standard describes attributes related to scanning and information about scanners. The attribute **MIABIS-DIGITALPATHOLOGYSCAN-01 (Scan Identifier)** contains a unique identifier for the scan as free text. **MIABIS-DIGITALPATHOLOGYSCAN-02 (Acquisition Date)** describes the date and time of the digital image capture or scanning of a slide. The value must be encoded in ISO 8601 format as YYYY-MM-DD hh:mm:ss. **MIABIS-DIGITALPATHOLOGYSCAN-03 (Imaging Protocol)** describes the protocol used for image acquisition. Permitted values are *whole Slide Imaging, slide Microscopy, WSI40XRGB, WSI20XRGB, gross Specimen Imaging, peri-operative Photographic Imaging, other (the value is known but cannot be classified into any of the predefined categories) NULL (the value is unknown).* **MIABIS-**

DIGITALPATHOLOGYSCAN-04 (Container Condition) evaluates the condition of the scanned slide and cover glass. Possible values are *Good, Acceptable, Unacceptable, Subject To Interpretation (The data provider is uncertain whether the container condition is acceptable or unacceptable)*. **MIABIS-DIGITALPATHOLOGYSCAN-05 (Device)** contains the name of the scanner or device used as specified by the manufacturer (including manufacturer name and model designation) in text format. **MIABIS-DIGITALPATHOLOGYSCAN-06 (Device Software Version)** describes the software and firmware versions of the device used as free text. **MIABIS-DIGITALPATHOLOGYSCAN-07 (Scanning Operation Mode)** describes the type of scanning operation, not the capabilities of the scanner. Possible values are *Automatic (digitization without any interference from the personnel responsible), Semi-automatic (digitization with minimal interference from the personnel responsible), Manual (the personnel responsible adjust several parameters during the digitization procedure).* **MIABIS-DIGITALPATHOLOGYSCAN-08 (Z-stacking)** describes the applied method of Z-stacking (focus stacking). Permitted values *are single (one image layer generated by one focal plane in scanning), multiple (n image layers generated by n focal planes in scanning), extended (one image layer generated by the combination of n focal planes in scanning).* **MIABIS-DIGITALPATHOLOGYSCAN-09 (Number of Focal Planes)** specifies the number of focal planes along the vertical z-axis, used for multiple and extended z-stacking and is stored as integer. **MIABIS-DIGITALPATHOLOGYSCAN-10 (Distance between Focal Planes)** describes the distance between the focal planes in micrometers (µm). This field must be filled with a decimal number. **MIABIS-DIGITALPATHOLOGYSCAN-11 (Tissue Scan Area)** is the scanned tissue area, specified in mm². A decimal number is also required here. **MIABIS-DIGITALPATHOLOGYSCAN-12 (Objective Lens Magnification)** describes the objective magnification used to capture the whole slide images. Permitted values are *5x, 10x, 20x, 40x, other, NULL*. **MIABIS-DIGITALPATHOLOGYSCAN-13 (Spatial Resolution)** describes the physical size of a pixel in the image in micrometers (µm) as a decimal number. **MIABIS-DIGITALPATHOLOGYSCAN-14 (Illumination Method)** specifies the illumination method for the imaging process. Possible values are *Transmission Illumination, Reflection Illumination, Epifluorescence Illumination, Brightfield Illumination,*

Darkfield Illumination, Oblique Illumination, Phase Contrast Illumination, Differential Interference Contrast, Total Internal Reflection Fluorescence. **MIABIS-DIGITALPATHOLOGYSCAN-15 (Illumination WaveLength)** describes the nominal wavelength of the illumination in nanometers (nm) as a decimal number. **MIABIS-DIGITALPATHOLOGYSCAN-16 (Out-of-Focus Error Rate)** indicates the percentage of the area that was captured out of focus. Ideally, the value should be determined by a validated algorithm and is expressed as a decimal number between 0 and 1. **MIABIS-DIGITALPATHOLOGYSCAN-17 (Stitching and other Artifact Error Rate)** describes the percentage of the scanned area affected by stitching errors or other artifacts. This value should also be determined by a validated algorithm and expressed as a decimal number between 0 and 1.

The File entity of MIABIS Digital Pathology contains information about the digital files generated during digitization. The attribute **MIABIS-DIGITALPATHOLOGYFILE-01 (File Identifier)** contains a unique identifier for an image file as free text (string). **MIABIS-DIGITALPATHOLOGYFILE-02 (File Format)** describes the format of the image file. Permitted values are *svs, tif, tiff, dcm, vms, vmu, ndpi, scn, mrxs, svslide, bif,other (the value is known but cannot be classified into any of the predefined categories), NULL (the value is unknown).* **MIABIS-DIGITALPATHOLOGYFILE-03 (File Size)** specifies the size of the image file in bytes and is stored as an integer. **MIABIS-DIGITALPATHOLOGYFILE-04 (Image Width)** describes the number of pixels per row of the image as an integer. **MIABIS-DIGITALPATHOLOGYFILE-05 (Image Height)** refers to the number of rows in the image and is also recorded as an integer. **MIABIS-DIGITALPATHOLOGYFILE-06 (Image Depth)** specifies the number of image planes, which is particularly relevant for Z-stacking. For single or extended Z-stacking, the image depth is 1; for multiple > 1. The values are entered as integers. **MIABIS-DIGITALPATHOLOGYFILE-07 (Number of Channels)** specifies the depth of data contained within a pixel as an integer. MIABIS-**DIGITALPATHOLOGYFILE-08 (Channel Resolution)** describes the image channel resolution in bits. The data type is an array of numbers as integer. **MIABIS-DIGITALPATHOLOGYFILE-09 (Color Space)** defines the color space model that describes the image channels. Valid values are *grey, RGB, lab, HSV, multispectral (the channels can be mapped to a specific light spectrum), multiplex*

(the channels are abstract information layers), other, NULL. **MIABIS-DIGITALPATHOLOGYFILE-10 (Color (ICC) Profile)** specifies the color profile of the image capture. Values from a controlled set of color profiles are possible. It is recommended to specify the color profile, such as *sRGB or Gray Gamma 1.8/2.2, GSDF, etc*. **MIABIS-DIGITALPATHOLOGYFILE-11 (Compression Method)** describes the compression method used. Where possible, the value should be taken from the Lossy Image Compression Method (0028,2114). **MIABIS-DIGITALPATHOLOGYFILE-12 (Compression Quality Level)** describes the quality of the compression as a decimal value. **MIABIS-DIGITALPATHOLOGYFILE-13 (Pyramid Levels)** specifies the number of resolution levels of an image pyramid as an integer. **MIABIS-DIGITALPATHOLOGYFILE-14 (Preview Available)** indicates whether a preview image version is available in the image file, typically in low resolution. Possible values are *yes, no, NULL (the value is unknown)*. **MIABIS-DIGITALPATHOLOGYFILE-15 (Label Available)** indicates whether a label image version is contained in the file. Permitted values are *yes, no, NULL (the value is unknown)*. **MIABIS-DIGITALPATHOLOGYFILE-16 (Annotations Available)** indicates whether the image file contains digital annotation layers. Markers or handwritten annotations on the slide are excluded from this. Possible values are *yes, no, NULL (the value is unknown)*. **MIABIS-DIGITALPATHOLOGYFILE-17 (Personal Information)** describes whether the image file contains personal information – either in the metadata or embedded in the image (e.g., burned-in labels). What constitutes identifiable information is at the discretion of the metadata creator. Permitted values are *yes, no, NULL (the value is unknown)*.

Med Uni Graz		Version: v01
		Datum: 10.12.2023
		Seite: 1

Studienprotokoll zum Projekt:

“AUTO-IMD”

Verantwortlichkeiten und Anschriften

Leitung:
Ao.Univ.-Prof. Dr.med.univ. Peter Regitnig
Diagnostik und Forschungsinstitut für Pathologie
Medizinische Universität Graz
Neue Stiftingtalstraße 2
8010 Graz

Weitere verantwortliche Personen:
Dipl. Ing., Ing. Markus Plass
Dr. Heimo Müller

Bezeichnung des Vorhabens

“Pilotprojekt: Automatisierte Indexierung von Bilddaten und Verknüpfung mit Metadaten zur Bereitstellung für Diagnostik und Forschung”

Einleitung

Im Zuge der Forschung und Diagnostik am Institut für Pathologie fallen große Datenmengen an. Diese bestehen einerseits aus Bilddaten, die generiert werden, und andererseits aus den zugehörigen Metadaten, die im PAS-Xanthos und verschiedenen Projektdatenbanken gespeichert sind. Leider liegen diese Informationen derzeit nur in semi-strukturierter und verteilter Form vor.

Die Hauptziele dieses Pilotprojektes sind:

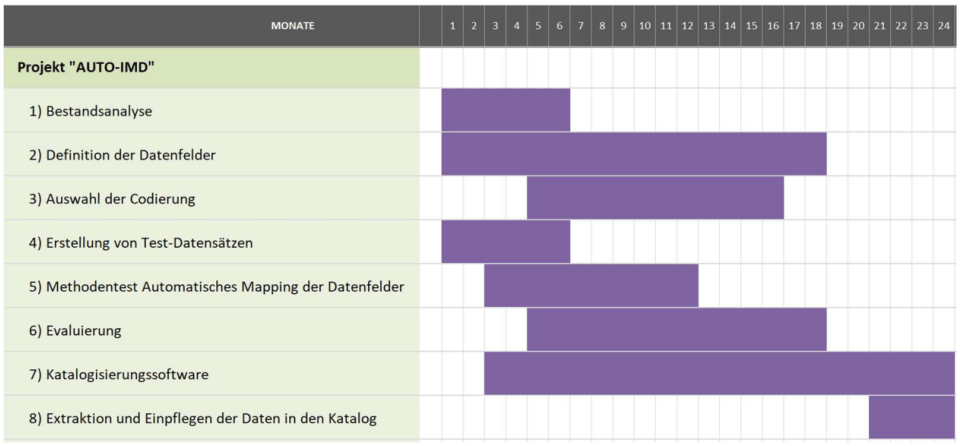
a) Durchführung von Methodentests/Evaluierung, um herauszufinden, wie diese semi-strukturierten Daten strukturiert werden können.

b) Reduktion der Subsysteme, um die gesammelten Informationen übersichtlich darzustellen und sie anschließend für Folgeforschungszwecke bereitzustellen.

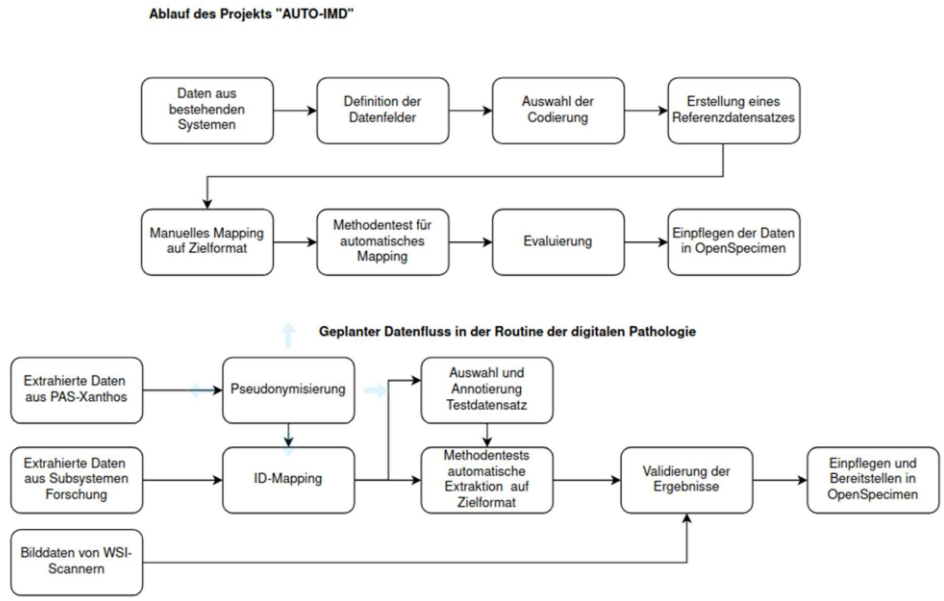
Um diese Ziele zu erreichen, soll aus den bestehenden Datenbanken ein repräsentativer Testdatensatz ausgewählt und annotiert werden. Auf diesem Datensatz sollen dann Methoden getestet werden, um die Strukturierung in Zukunft automatisiert durchführen zu können. Nachfolgend wird ein Katalog (OpenSpecimen) mit den extrahierten Daten gefüllt, und die Nutzbarkeit dessen wird evaluiert.

Vorgehensweise

Das Projekt “AUTO-IMD” teilt sich in acht Phasen auf und ist für einen Zeitraum von zwei Jahren geplant.



Gantt-Chart des Projekts “AUTO-IMD”



Datenflussdiagramm

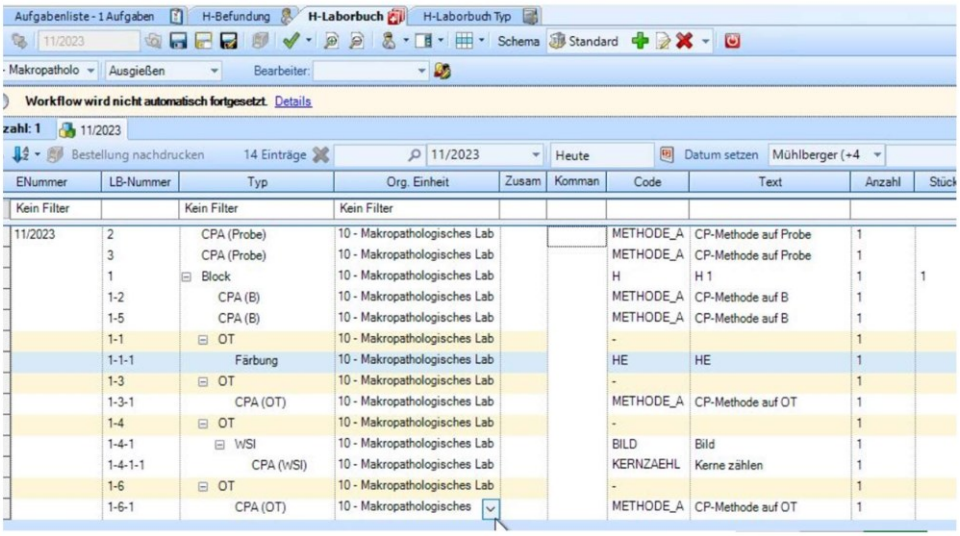
1) Bestandsanalyse

Zu Beginn des Projektes werden die am Diagnostik und Forschungsinstitut für Pathologie der Medizinischen Universität Graz etablierten Systeme zur Verarbeitung der Diagnosedaten analysiert und deren Datenstrukturen/Datenfelder erfasst. Zu diesen Systemen zählen das am Pathologie-Institut eingesetzte Labor-Informations-Management System *PAS Xanthos*, Softwarelösungen zum Management der Whole Slide Images (WSIs) in der digitalen Pathologie und das Krankenhausinformationssystem (KIS) *openMEDOCS*.

Systeme zur Verarbeitung der Daten:

Labor-Informations-Management System *PAS Xanthos*

Die speziell auf die Anforderungen im Bereich der Pathologie zugeschnittene Softwarelösung PAS Xanthos (Hersteller: Programmierfabrik Hagenberg) wird am Diagnostik und Forschungsinstitut für Pathologie der Medizinischen Universität Graz als Labor-Informations-Management System (LIMS) verwendet, das die prozessorientierte Abarbeitung der Untersuchungen und qualitätsstandard-konforme Dokumentation sämtlicher Laborleistungen des Instituts beginnend bei der Probenannahme bis zur Befunderstellung unterstützt.



Aktuelle Ansicht des Laborbuchs in *PAS Xanthos*

Softwarelösungen zum WSI-Management

WSI-Webservice

Mit Hilfe der Softwarelösung *WSI-Webservice* werden alle digitalen histopathologischen Bilddaten von Forschungsprojekten verwaltet. Dazu enthält *WSI-Webservice* verknüpfte Metadaten zu den Bildern und extrahierte, pseudonymisierte Patienten-Informationen von Forschungsprojekten.

Case-Center

Das serverbasierte WSI-Managementsystem *CaseCenter* (Hersteller: 3DHISTECH) dient den Patholog*innen als Plattform, um WSIs für Lehre und Forschung zu verwalten. Auch dieses System enthält Metadaten zu den gespeicherten WSIs.

WSI-Management für die Diagnostik

Eine WSI-Viewer Softwarelösung für die (zukünftige) digitale Pathologie Diagnostik ist derzeit im Ausschreibungs- bzw. Anschaffungsprozess.

Krankenhausinformationssystem *openMEDOCS*

Über eine Schnittstelle zum Krankenhausinformationssystem *openMEDOCS* haben Patholog*innen Zugriff auf die Grunddaten der Patient*innen. Dort befinden sich Befunde, Informationen zu weiterführenden Behandlungen sowie Bilddaten (z.B. radiologische Bilder).

Die Analyse der bestehenden Systeme zur Verarbeitung der Diagnosedaten (wie *PAS-Xanthos* und *openMEDOCS*) und zum Bildmanagement ist ein wichtiger Schritt, um den Überblick über vorhandene Daten und Systeme zu erhalten. Hierbei werden alle relevanten Informationen über diese Systeme gesammelt und ermittelt, welche Arten von Daten in jedem System gespeichert sind. Auch die Quelle dieser Daten wird identifiziert, um zu wissen, ob diese Daten direkt von medizinischen Geräten, durch manuelle Eingabe oder durch eine Schnittstelle zu einem anderen System bezogen werden. Des Weiteren wird untersucht, in welchen Formaten und Strukturen die Daten in den bestehenden Systemen vorliegen. Man unterscheidet hier zwischen strukturierten Daten, unstrukturierten Daten und einer Kombination aus beidem. Es wird überprüft, ob es Standards für die Datendarstellung gibt.
Es gilt herauszufinden, wer auf welche Daten in den bestehenden Systemen zugreifen kann und welche Berechtigungen und Zugriffsbeschränkungen gelten, um zu gewährleisten, dass auch durch die im weiteren Projektverlauf geplante automatisierte Indexierung von Bilddaten und Verknüpfung mit Metadaten zur Bereitstellung für Diagnostik und Forschung weiterhin Datenschutz- und Sicherungsrichtlinien eingehalten werden.
Es wird überprüft, wie die Daten in den bestehenden Systemen verwaltet und gesichert werden, um sicherzustellen, dass die Integrität der Daten gewahrt bleibt. Die Möglichkeiten, Daten aus den bestehenden Systemen zu exportieren, werden getestet.
Die Ergebnisse dieser Analyse der bestehenden Systeme werden umfassend dokumentiert. Diese Dokumentation wird alle gefundenen Informationen zu den bestehenden Systemen, Datenquellen, Datenformaten und anderen Aspekten, die für die weitere Projektarbeit relevant sind, enthalten.

2) Definition der Datenfelder

Um bestehende und auch zukünftige Daten strukturiert und kategorisiert zur Verfügung zu haben, ist die Erstellung eines Katalogs geplant. Um sicherzustellen, dass der Katalog effektiv erstellt, verwaltet und genutzt werden kann, ist eine Definition der Datenstruktur entscheidend. Es ist wichtig, dass die Datenstruktur den Anforderungen der Benutzer*innen gerecht wird und Flexibilität für zukünftige Änderungen und Erweiterungen bietet.
Zunächst wird festgelegt, welche Informationen und Daten im Katalog enthalten sein sollen. Dies geschieht auf Basis einer Kategorisierung und Priorisierung der möglichen Datenfelder. Hauptaugenmerk liegt hier auf der Bedeutung und Schwierigkeit der Erhebung der Daten. Alle relevanten Datenkategorien werden identifiziert.
Für jede Kategorie von Informationen im Katalog wird ein Datentyp definiert. Hierarchien und Beziehungen zwischen den einzelnen Elementen im Katalog werden definiert. Um sicherzustellen, dass die Daten im Katalog konsistent und korrekt sind, werden Regeln für die Validierung und Formatierung der Daten festgelegt. In der Dokumentation sind die Definitionen der Datentypen, Attribute, Pflichtfelder, Hierarchie und alle relevanten Informationen enthalten.

Für die Auswahl der Datenfelder sind folgende Schritte geplant:

Analyse der Datenfelder aus bisherigen Forschungsprojekten

Datenfelder aus bisherigen Forschungsprojekte, die in ähnlichen Kontexten durchgeführt wurden, werden untersucht. Die Datenfelder, die als wichtig erachtet werden, werden identifiziert.

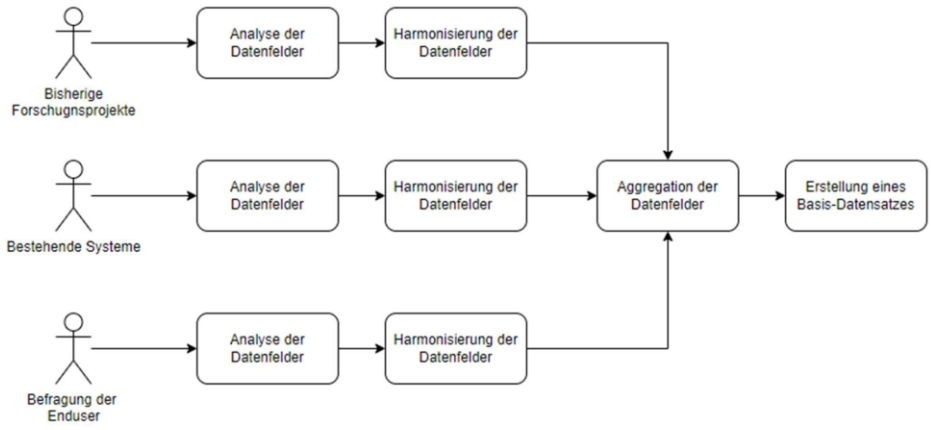
Analyse der Datenfelder der bestehenden Systeme

Die Datenfelder, die in den oben beschriebenen bestehenden Systemen vorhanden sind, werden analysiert und die, die für sinnvoll erachtet werden, identifiziert.

Analyse der Datenfelder, die von den Endbenutzer*innen gewünscht werden

In Form von Interviews wird eine Gruppe von Endbenutzer*innen nach ihren Bedürfnissen und Wünschen hinsichtlich der Datenfelder befragt. Die auf diese Weise erhobenen Datenfelder werden zusammengefasst und in Form eines Fragebogens aufbereitet. Dieser Fragebogen wird wieder an die Endbenutzer*innen verteilt mit dem Ziel der Priorisierung der einzelnen Datenfelder.

Die durch diese Methoden erhobenen Datenfelder werden in weiterer Folge harmonisiert und aggregiert. Die aggregierten Datenfelder werden benutzt, um einen Basis-Datensatz zu erstellen, der im weiteren Verlauf des Projekts als Referenzdatensatz verwendet wird.



Nötige Schritte zur Erstellung des Basis-Datensatzes

3) Auswahl der Codierung

Für die strukturierte Speicherung der Daten ist es notwendig, ein geeignetes Datenformat bzw. eine geeignete Codierung auszuwählen. Im Bereich der digitalen Pathologie bieten sich diverse Standards an:

ICD (International Classification of Diseases)

Der ICD ist ein weltweit anerkanntes Klassifikationssystem für medizinische Diagnosen, entwickelt und herausgegeben von der Weltgesundheitsorganisation (WHO). Der ICD dient dazu, Krankheiten und Gesundheitsprobleme zu klassifizieren und zu kodieren. Er bietet eine standardisierte Methode zur Erfassung, Darstellung und Analyse von Gesundheitsdaten und Statistiken auf globaler Ebene.
Der ICD ist hierarchisch strukturiert und besteht aus verschiedenen Kategorien, Gruppen und Unterkategorien. Jeder Zustand oder jede Krankheit wird durch einen alphanumerischen Code dargestellt. Der Code setzt sich aus Buchstaben und Ziffern zusammen und bietet eine eindeutige Identifikation für jeden Zustand.
Die zehnte Revision des ICD, bekannt als ICD-10, ist die derzeit am weitesten verbreitete Version. Es gibt auch eine fortgeschrittenere Version namens ICD-11, die veröffentlicht wurde und nach und nach weltweit eingeführt wird.

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms)

SNOMED CT ist eine umfassende und international anerkannte klinische Terminologie und Ontologie, die speziell für den Gebrauch im Gesundheitswesen entwickelt wurde. SNOMED CT bietet eine standardisierte Möglichkeit, klinische Informationen und medizinische Begriffe zu repräsentieren, um eine einheitliche Kommunikation und Interoperabilität im Gesundheitswesen zu fördern. SNOMED CT dient dazu, eine eindeutige und standardisierte Sprache für die Darstellung von klinischen Begriffen, Diagnosen, Prozeduren, Befunden und anderen medizinischen Konzepten bereitzustellen.
SNOMED CT verwendet eine hierarchische Struktur, bei der medizinische Begriffe in einem Baumdiagramm organisiert sind. Dies ermöglicht eine präzise Klassifikation und Repräsentation von medizinischen Konzepten. Jedes Konzept in SNOMED CT wird durch eine eindeutige Konzept-ID identifiziert. Die Terminologie enthält auch Beziehungen zwischen den Konzepten, um die semantischen Verbindungen zwischen den medizinischen Begriffen zu repräsentieren.
SNOMED CT kann mit anderen medizinischen Klassifikationssystemen und Terminologien, wie ICD (International Classification of Diseases), LOINC (Logical Observation Identifiers Names and Codes) und anderen, gemappt werden.

OMOP (Observational Medical Outcomes Partnership)

OMOP ist ein Konsortium und eine Initiative, die darauf abzielt, Daten aus der realen Welt (Real-World Data, RWD) für die Forschung im Gesundheitswesen nutzbar zu machen. Das Hauptziel von OMOP ist es, einen gemeinsamen Datenstandard zu schaffen, der es

Forschern ermöglicht, strukturierte Daten aus verschiedenen Quellen zu integrieren und zu analysieren, um Erkenntnisse über die Sicherheit und Wirksamkeit von medizinischen Interventionen zu gewinnen. OMOP fördert die Nutzung von strukturierten Daten, um Fragen zur Sicherheit, Wirksamkeit und Effizienz von medizinischen Behandlungen zu beantworten.
Das OMOP CDM (Common Data Model) ist ein standardisiertes Datenmodell, das Tabellen und Felder für verschiedene Arten von Gesundheitsdaten definiert. Es umfasst Informationen zu Patienten, Diagnosen, Medikamenten, Labortests und anderen relevanten Aspekten der Gesundheitsversorgung.

FHIR (Fast Healthcare Interoperability Resources)
FHIR ist ein international anerkannter Standard für den Austausch von Gesundheitsinformationen und die Interoperabilität im Gesundheitswesen. FHIR wurde vom Health Level Seven International (HL7) entwickelt, einer Organisation, die sich auf die Entwicklung von Standards für den Austausch von Gesundheitsinformationen spezialisiert hat. Der Hauptzweck von FHIR besteht darin, den Datenaustausch zwischen verschiedenen Gesundheitsinformationssystemen zu erleichtern, insbesondere im Kontext der elektronischen Gesundheitsakte (EHR) und der modernen Gesundheitsinformatik. FHIR wird sowohl in der Patientenversorgung als auch in der Forschung eingesetzt. Im klinischen Kontext ermöglicht es die Interaktion zwischen elektronischen Patientenakten, Gesundheitsportalen und anderen Systemen. In der Forschung erleichtert es den Austausch von klinischen Studiendaten.

Die Auswahl des geeigneten Formats hängt von mehreren Faktoren ab, wie zum Beispiel der Interoperabilität (Daten sollten zwischen verschiedenen Systemen ausgetauscht werden können). Der Standard sollte auch von bestehenden Systemen unterstützt und im Zuge der Diagnostik erhoben werden. Es werden Kombinationen verschiedener Standards in Erwägung gezogen. Für die Auswahl werden Expert*innen aus den Bereichen der Pathologie und der Technik zur Rate gezogen.

4) Erstellung von Test-Datensätzen

Das Erstellen eines geeigneten Referenzdatensatzes ist ein wichtiger Schritt, um die automatisierte Indexierung von Bilddaten und Verknüpfung mit Metadaten zu realisieren.
Es werden relevante Daten in den bestehenden Systemen identifiziert. Dabei wird auf eine ausreichende Vielfalt und Repräsentativität geachtet, um die Anforderungen an das Katalog-System angemessen abzudecken. Es wird ein repräsentativer Basis-Datensatz erstellt, der ~10.000 Fälle beinhalten wird.
Die ausgewählten Demo-Daten werden aus den bestehenden Systemen (beschrieben unter Bestandsanalyse) extrahiert. Dies kann bedeuten, Daten aus Datenbanken zu exportieren, Dateien herunterzuladen oder Schnittstellen zu verwenden, um die Daten abzurufen. Die extrahierten Daten werden auf Inkonsistenzen, Fehler oder inkorrekte Informationen überprüft

und bereinigt, indem fehlende Werte aufgefüllt, Duplikate entfernt und Datenfehler korrigiert werden.
Die Daten werden für die weitere Verwendung pseudonymisiert und in weiterer Folge manuell annotiert. Im Zuge dieses Prozesses werden die Daten mit den erforderlichen Metadaten versehen. Hierfür wird eine klare Annotations-Richtlinie erstellt, um konsistente und zuverlässige Ergebnisse zu erzielen. Das manuelle Zuordnen ("Mappen") von Datensätzen auf ein Zielformat ist ein wichtiger Schritt, um sicherzustellen, dass die Daten in einer strukturierten Form vorliegen. Für diesen Schritt werden klare Mapping-Regeln definiert, die beschreiben, wie die Daten aus dem Quelldatensatz in das Zielformat überführt werden. Die Quelldaten werden in weiterer Folge händisch in das Zielformat übertragen. Nach dem Mapping werden die Daten im Zielformat auf Genauigkeit und Übereinstimmung mit den Anforderungen überprüft. Dabei wird sichergestellt, dass keine Daten verloren gegangen sind und dass sie korrekt und vollständig im Zielformat vorliegen.
Die Ergebnisse der Annotation unterlaufen einer Qualitätssicherung, um sicherzustellen, dass der Referenzdatensatz eine hohe Qualität aufweist. Durch Validierung wird sichergestellt, dass der Referenzdatensatz für den beabsichtigten Zweck geeignet ist.

5) Methodentest für automatisches Mapping der Datenfelder

Nach Erstellung, händischer Annotierung und Validierung des Referenzdatensatzes werden mehrere Methoden getestet, um das Mapping auf das Zielformat zu automatisieren. Hierfür kommen unter anderem Tools zum Einsatz, welche in Forschungsprojekten bisher verwendet wurden. Zudem werden neue Ansätze getestet.

SAAT (Semi-Automated-Annotation-Tool)

SAAT ist ein System, das zur Extraktion strukturierter Informationen aus unstrukturierten pathologischen Diagnosedaten dient. Es verwendet einen Regelsatz-basierten Algorithmus und reguläre Ausdrücke zur Strukturierung der Daten.

AIDAVA

Im Zuge des Projekts "AIDAVA" werden Daten aus Befunden und anderen Gesundheitsdokumenten extrahiert und in einen Personal Health Knowledge Graph (PHKG) übertragen.

Large Language Models (LLMs)

Die Entwicklung im Bereich der Large Language Models schreitet rasant voran.
Für das Verarbeiten und Verstehen von klinischen Texten gibt es spezialisierte Sprachmodelle und NLP-Modelle, die darauf abzielen, die spezifischen Anforderungen und Nuancen von medizinischem Fachjargon und klinischer Terminologie zu bewältigen:

BioBERT ist ein auf BERT (Bidirectional Encoder Representations from Transformers) basierendes Sprachmodell, das speziell für die Verarbeitung von biomedizinischen und

klinischen Texten entwickelt wurde. Es wurde von Forschern der University of Washington und des National Institutes of Health (NIH) erstellt. BioBERT wurde darauf trainiert, das Verständnis von spezifischem medizinischen Fachjargon, klinischer Terminologie und biomedizinischem Kontext zu verbessern.

ClinicalBERT ist ein auf BERT (Bidirectional Encoder Representations from Transformers) basierendes Sprachmodell, das speziell für die Verarbeitung von klinischen Texten und medizinischem Fachjargon entwickelt wurde. Es wurde von Forschern der University of Massachusetts Medical School erstellt, um die spezifischen Anforderungen des medizinischen Kontextes besser zu adressieren.

GPT-3 bzw. GPT-4 ist ein leistungsstarkes, generatives Sprachmodell von OpenAI, das auf vielfältige Weise eingesetzt werden kann, darunter auch in medizinischen Anwendungen. Es kann für die automatische Generierung von medizinischen Texten, Berichten oder Patienteninformationen verwendet werden und bei der automatischen Erfassung und Dokumentation von klinischen Informationen in Patientenakten oder anderen medizinischen Aufzeichnungen unterstützen.

Nach der Testung der oben beschriebenen Tools werden jeweils deren Vor- und Nachteile analysiert. Es wird getestet, inwiefern diese Tools lokal ausführbar sind und ob man diese Tools spezifisch mit vorhandenen Datensätzen trainieren kann, um die Performance für diesen speziellen Anwendungszweck zu verbessern.

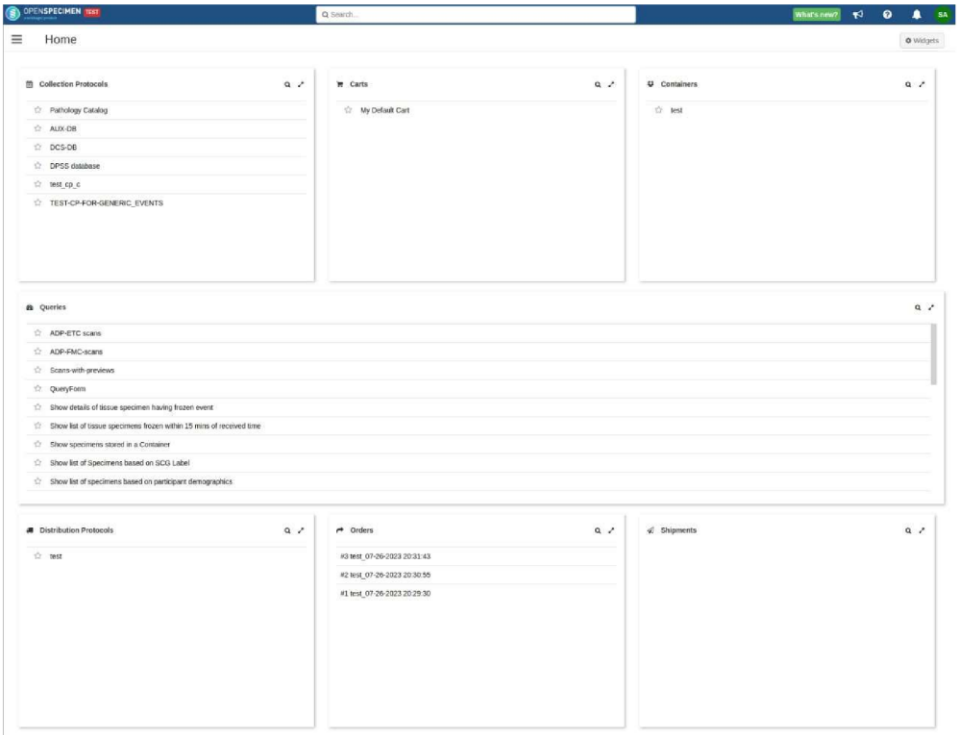
Es wird auch recherchiert, welche neuen Tools für das Mapping auf das Zielformat verwendet werden können. Diese Tools werden genau wie die bereits bekannten Tools getestet und die Ergebnisse mit dem manuell erstellten Referenzdatensatz abgeglichen.

6) Evaluierung

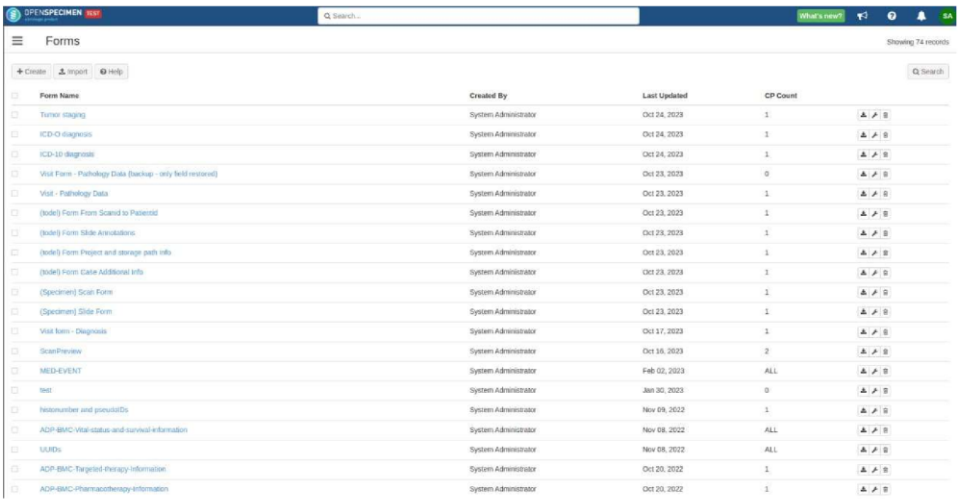
Die Evaluierung der getesteten Methoden und der Vergleich der verschiedenen Methoden miteinander ist nötig, um die bestmögliche Methode für das Mapping auf das Zielformat auswählen zu können. Hierfür werden die Genauigkeit des Mappings und die Geschwindigkeit der Ausführung berücksichtigt. Nach der Testung aller Methoden bzw. Tools werden die Vorteile und Nachteile der einzelnen Möglichkeiten gegeneinander abgewogen. Ein wichtiger Aspekt ist hierbei die Skalierbarkeit, da die Methode sich für große Datenmengen eignen muss. Auch auf die Flexibilität wird Wert gelegt, da die Methode sich an ändernde Anforderungen und zukünftige Entwicklung anpassen lassen muss. Ziel ist es, eine fehlerresistente Methode auszuwählen, die in der Lage ist, auch mit unerwarteten Daten oder Änderungen in der Quelldatei umzugehen.

7) Katalogisierungssoftware

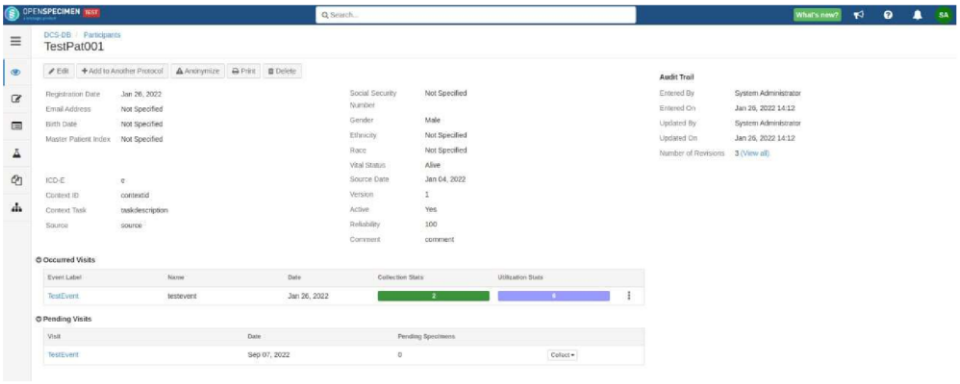
Als Katalogisierungssoftware wird OpenSpecimen (https://www.openspecimen.org/) verwendet. OpenSpecimen (früher bekannt als caTissue Plus) ist eine freie und quelloffene Software zur Verwaltung von Biobanken und Bioproben. Der Grundgedanke von OpenSpecimen ist, dass "Bioproben ohne qualitativ hochwertige Daten keinen Wert haben". OpenSpecimen wird weltweit in einigen Biobanken unterschiedlicher Größe und Entitäten eingesetzt. OpenSpecimen rationalisiert die Verwaltung von Bioproben in den Bereichen Entnahme, Zustimmung, Qualitätskontrolle, Anforderung und Verteilung. Die Software ist in hohem Maße konfigurierbar und anpassbar.



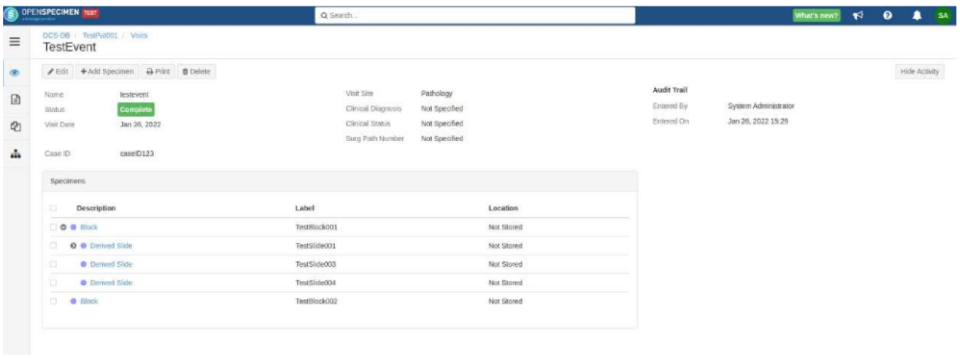
OpenSpecimen - Home/Dashboard



OpenSpecimen - Forms

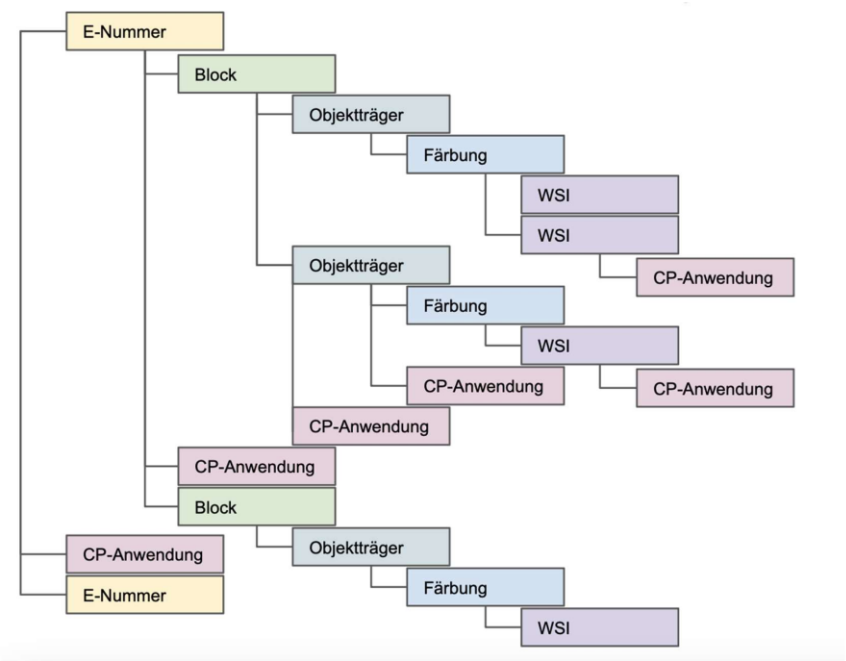


OpenSpecimen - Test Patient



OpenSpecimen - Test Event

Für die Installation von OpenSpecimen wird zuerst ein Datenbanksytem installiert, das von OpenSpecimen unterstützt wird (PostgreSQL oder MySQL). Das Datenbanksystem wird eingerichtet und die erforderlichen Datenbanken und Benutzerkonten erstellt. Da OpenSpecimen eine Java-Anwendung ist, wird sichergestellt, dass die erforderliche Version des Java Development Kit (JDK) installiert ist. Da OpenSpecimen in der Regel auf einem Tomcat-Webserver ausgeführt wird, wird Tomcat installiert und der Webserver entsprechend konfiguriert. Die aktuelle Version von OpenSpecimen wird von https://github.com/krishagni/openspecimen bezogen. OpenSpecimen wird installiert und die Datenbank-Initialisierung wird durchgeführt, um die erforderlichen Tabellen und Daten in der Datenbank zu erstellen. OpenSpecimen wird konfiguriert, um die gewünschte Datenstruktur (siehe Abbildung) darstellen zu können.



Abzubildende Struktur in OpenSpecimen

Die Sicherheitseinstellungen und die Zugriffsbeschränkungen werden verwaltet, um die in OpenSpecimen abgebildeten Daten zu schützen. Dafür werden Rollen mit den benötigten Rechten festgelegt, denen die erstellten Accounts der Benutzer*innen zugewiesen werden. Es werden eindeutige Regeln zur Nutzung der Datensätze festgelegt, um eine sichere Umgebung für die Daten zu schaffen.
Die ordnungsgemäße Installation und Konfiguration wird im Zuge von Tests und Validierungsverfahren geprüft. Dies umfasst den Zugriff auf die Anwendung und die Verbindung zur Datenbank. Benutzerkonten werden eingerichtet und Berechtigungen vergeben, um den Zugriff auf Daten zu kontrollieren. Die Nutzer der Software werden in der Verwendung von OpenSpecimen geschult. Die Software wird regelmäßig gewartet und aktualisiert.

Nach der Installation und Konfiguration von OpenSpecimen werden die Testdaten in OpenSpecimen eingepflegt. Dafür wird eine CSV-Datei erstellt, die die Testdaten enthält. Jede Zeile der CSV-Datei entspricht hierbei einem Datensatz, und die Spalten entsprechen den Attributen in OpenSpecimen. Die CSV-Datei wird in OpenSpecimen hochgeladen. In OpenSpecimen werden die Spalten in der CSV-Datei den entsprechenden Datenfeldern in der Anwendung zugeordnet, damit die Daten korrekt interpretiert werden können. Durch Start

des Datenimportprozesses wird die CSV-Datei von OpenSpecimen verarbeitet und die Testdaten werden in die Datenbank eingefügt. Nach dem Datenimport wird überprüft, ob die Test-Daten korrekt und vollständig sind. Auch die Beziehungen zwischen den Datensätzen müssen ordnungsgemäß sein. Durch Funktionstests wird sichergestellt, dass die Software wie erwartet funktioniert. Dies geschieht durch Suchen nach Daten, Anzeige von Datensätzen und Testen anderer Funktionen.

Wenn die Testdaten vollständig und korrekt in OpenSpecimen eingepflegt sind, wird die Nutzbarkeit des Katalogs mit den Testdaten durch Patholog*innen evaluiert. Es wird eine repräsentative Gruppe von Nutzer*innen ausgewählt, die die tatsächlichen Anwender*innen der Software darstellen. Klare Kriterien und Metriken, anhand derer die Nutzbarkeit des Katalogs bewertet wird, werden definiert. Dies kann die Effizienz bei der Dateneingabe, die Genauigkeit der Suche, die Benutzerfreundlichkeit der Benutzeroberfläche und andere relevante Aspekte umfassen. Es wird sichergestellt, dass die Testumgebung den Bedingungen in der tatsächlichen Produktionsumgebung entspricht. Dies beinhaltet die Verwendung der Testdaten, die zuvor in OpenSpecimen eingepflegt wurden. Die Patholog*innen erhalten Schulung und Anleitung zur Verwendung von OpenSpecimen. Sie sollten mit den Funktionen und Werkzeugen der Software vertraut sein. Die Patholog*innen werden die Software in einer realistischen Umgebung verwenden. Es wird beobachtet, wie sie mit OpenSpecimen interagieren. Es werden Aufgaben durchgeführt, die den täglichen Arbeitsabläufen entsprechen. Feedback von den Patholog*innen wird in Form eines strukturierten Fragebogens gesammelt. Die Patholog*innen werden auch bei der Arbeit mit der Software beobachtet, um mögliche Probleme oder Herausforderungen zu identifizieren. Mithilfe der gesammelten Daten wird ein Bericht erstellt, der die Ergebnisse der Evaluierung zusammenfasst. Hier werden die Stärken und Schwächen der Software aus Sicht der Patholog*innen identifiziert. Basierend auf diesem Feedback und den Ergebnissen der Evaluierung können Iterationen und Verbesserungen in Betracht gezogen werden. Dies kann die Anpassung der Benutzeroberfläche, die Implementierung neuer Funktionen oder die Korrektur von Fehlern umfassen.

8) Extraktion und Einpflegen der Daten in den Katalog

In Zukunft sollen Daten automatisiert aus Subsystemen extrahiert und in den Katalog eingespielt werden. Hierfür werden Schnittstellen zu den bestehenden Systemen (beschrieben in Bestandsanalyse) erstellt. Wenn die Systeme APIs unterstützen, wäre dies eine Möglichkeit, die Daten zwischen den Systemen auszutauschen. Alternativ können die Daten in einem standardisierten Format (z.B. CSV, JSON) exportiert werden. Auch der WSI-Scanner wird in diesen Prozess eingebunden, um Daten direkt vom Scanner in den Katalog zu übertragen.
Es muss sichergestellt werden, dass die Daten aus den bestehenden Systemen so transformiert werden, dass sie dem gewünschten Format im Katalog entsprechen. Dafür

werden klare Mappings zwischen den Datenfeldern in den bestehenden Systemen und den Datenfeldern im Katalog definiert.
Es wird ein Sicherheitsmechanismus implementiert, um sicherzustellen, dass der Zugriff auf die Daten und die Übertragung zwischen den Systemen sicher und geschützt ist. Zugriffskontrollen und Authentifizierungsmethoden sorgen dafür, dass nur autorisierte Benutzer auf die Schnittstellen zugreifen können.
Der Prozess der Datenextraktion und -übertragung wird automatisiert und erfolgt regelmäßig ohne manuelle Eingriffe. Dazu wird der Zeitpunkt für die Datenextraktion gewählt, um die Aktualität der Daten im Katalog sicherzustellen. Ein Protokollierungsmechanismus wird implementiert, um den Status und Verlauf der Datenübertragung zu überwachen. Mechanismen zur Fehlerbehandlung werden hinzugefügt, um Probleme während der Datenübertragung zu erkennen und zu beheben. Die Schnittstellen werden umfassend getestet und validiert.

Data fields Level 2 – Participant

Field name	Description	Example	Data type
Parent collection protocol ID	Identifier of the Collection protocol		Impicit link to layer above
Participant ID	Identifier of the source Participant	115245	Text
Date of birth	Date of Birth	31.01.2000	Date
Date of death	Date of Death	31.01.2000	Date
Sex	Sex	f	'f'; 'm'; 'i'; 'unknown'; 'non applicable'
Social security number	Austrian social security number	1234310100	Number
Origin	Origin of sample	human	'human';'non human'
Origin detail	Detailed type of species if not human	Mouse	Text

	Availability			
Field name	**PAS Xanthos**	**ADOPT cohort**	**CRC Cohort**	**MIABIS**
Parent collection protocol ID				
Participant ID	Int	Not available	UUID	Text
Date of birth	Datetime	Not available	Not available	Datetime
Date of death	Datetime	Not available	Not available	Not available
Sex	Int	Fixed value set	Fixed value set	Fixed value set
Social security number	Text	Not available	Not available	Not available
Origin	Not available	Not available	Not available	Fixed value set
Origin detail	Not available	Not available	Not available	Not available

Data fields Level 3 – Visit

Field name	Description	Example	Data type
Parent participant ID	Identifier of the Participant		Impicit link to layer above
Request ID	Identifier of the Request	123456	Text
Date of event	Date of Event	31.01.2000	Date
Age in years at the diagnosis date	Age in years at the diagnosis date	87	Integer

Field name	Availability			
	PAS Xanthos	ADOPT cohort	CRC Cohort	MIABIS
Parent participant ID				
Request ID	Int	Not available	Not available	Text
Date of event	Datetime	Datetime	Not available	Datetime
Age in years at the diagnosis date	Not available	Natural number	Int	Int

Data fields Level 4 – Container

Field name	Description	Example	Data type
Parent request ID	Identifier of the Request		Impicit link to layer above
Container ID	Identifier of the primary sample	2024000001	Text
Human readable container ID	Identifier of the primary sample in human readable format	2024000001	Text
Material SNOMED-CT ID	SNOMED-CT coded material	128169007	List according to https://www.snomed.org/
Material	Human readable description of material	LIVER	Text
Material description	Detailed description of the material	2 Liver punch biopsy cylinders	Text
Extraction method	Specimen type or shape of section	Needle biopsy	‘Conventional smear'; 'Conventional smear using a cytobrush'; 'Fluid specimen'; 'Smear or imprint (on glass slide)'; 'Surgical specimen'; 'Needle biopsy'; 'Excisional biopsy'; 'Autopsy / Post-mortem examination'; 'Curettage, Pipelle biopsy, or transurethral resection (TUR)'
Macroscopic description	Description of the results of the macroscopic examination	The specimen consists of a segment of tissue measuring 4.5 × 3.0 × 2.0 cm. On cut section, a firm, whitish, ill-defined tumor measuring 1.8 cm in greatest diameter is identified.	Text

Field name	Description	Example	Data type
Histological description	Description of the results of the histological examination	Sections show an invasive epithelial neoplasm composed of irregular glandular structures infiltrating the surrounding stroma. The tumor cells display moderate nuclear atypia with enlarged, hyperchromatic nuclei and prominent nucleoli. Mitotic figures are readily identified. The surrounding stroma shows desmoplastic reaction. No definite lymphovascular invasion is seen in the examined sections.	Text
Molecularpathological diagnosis	Results of the molecular pathological examination	Molecular analysis demonstrates the presence of a pathogenic KRAS mutation (c.35G>A, p.G12D). No mutations were detected in NRAS or BRAF. The findings are consistent with a RAS-mutated tumor profile.	Text
Molecularpathological description	Description of the results of the molecular pathological examination	DNA extracted from tumor tissue was analyzed using targeted next-generation sequencing. A pathogenic mutation in the TP53 gene was detected.	Text
Diagnosis	Result of the diagnostic procedure	Invasive moderately differentiated adenocarcinoma	Text
Neuropathological report	Result of the neuropathological examination	Sections show a diffusely infiltrating glial neoplasm with increased cellularity and nuclear atypia. Mitotic figures are identified. No definite microvascular proliferation or necrosis is observed.	Text
Rapid section examination	Result of the rapid section examination	Sections show a cellular neoplasm composed of atypical epithelial cells arranged in irregular glandular structures. Nuclear pleomorphism and mitotic activity are present.	Text
Cause of death	Cause of death	Immediate cause of death: Acute respiratory failure Underlying cause: Metastatic adenocarcinoma of the lung	Text

Field name	Description	Example	Data type
Postmortem examination	Result of the postmortem examination	The lungs are heavy and congested. The heart is enlarged with left ventricular hypertrophy. The liver shows moderate steatosis. No acute hemorrhage or traumatic injury is identified.	Text
Cytological description	Description of the results of the cytological examination	Smears show a moderately cellular specimen composed of cohesive clusters and scattered atypical epithelial cells.	Text
Report type	Type of the report	Initial report	'Initial report'; 'Addendum'; 'Amended report'
TNM T prefix	Prefix of the T category of the TNM classification	p	'c'; 'p'; 'yc'; 'yp'; 'a'; 'r'; 'rc'; 'rp'
TNM T	T category of the TNM classification	1	'TX'; 'T0'; 'Ta'; 'Tis'; 'T1'; 'T1mi'; 'T1a'; 'T1a1'; 'T1a2'; 'T1b'; 'T1b1'; 'T1b2'; 'T1b3'; 'T1c'; 'T1c1'; 'T1c2'; 'T1c3'; 'T1d'; 'T2'; 'T2a'; 'T2a1'; 'T2a2'; 'T2b'; 'T2c'; 'T2d'; 'T3'; 'T3a'; 'T3b'; 'T3c'; 'T3d'; 'T3e'; 'T4'; 'T4a'; 'T4b'; 'T4c'; 'T4d'; 'T4e'
TNM T addition	Addition to the T category of the TNM classification	m	Text
TNM N prefix	Prefix of the N category of the TNM classification	p	'c'; 'p'; 'yc'; 'yp'; 'a'; 'r'; 'rc'; 'rp'
TNM N	N category of the TNM classification	1	'NX'; 'N0'; 'N1'; 'N1mi'; 'N1a'; 'N1b'; 'N1c'; 'N2'; 'N2a'; 'N2b'; 'N2c'; 'N3'; 'N3a'; 'N3b'; 'N3c'
TNM N addition	Addition to the N category of the TNM classification	(i+)	Text

Field name	Description	Example	Data type
TNM M prefix	Prefix of the M category of the TNM classification	p	'c'; 'p'; 'yc'; 'yp'; 'a'; 'r'; 'rc'; 'rp'
TNM M	M category of the TNM classification	1	'M0'; 'M1'; 'M1a'; 'M1a1'; 'M1a2'; 'M1b'; 'M1b1'; 'M1b2'; 'M1c'; 'M1c1'; 'M1c2'; 'M1d'
TNM M addition	Addition to the M category of the TNM classification	(mol+)	Text
TNM stage	UICC classification for tumor staging	IVA	'0'; '0a'; '0is'; 'I'; 'IA'; 'IA1'; 'IA2'; 'IA3'; 'IB'; 'IB1'; 'IB2'; 'IC'; 'IS'; 'II'; 'IIA'; 'IIA1'; 'IIA2'; 'IIB'; 'IIC'; 'III'; 'IIIA'; 'IIIB'; 'IIIC'; 'IIIC1'; 'IIIC2'; 'IIID'; 'IV'; 'IVA'; 'IVB'; 'IVC'
TNM UICC	UICC version used	9	Number
ICD-O code	ICD-O coding of the diagnosis	8010/0	List according to https://seer.cancer.gov/icd-o-3/sitetype.icdo3.20220429.pdf
ICD-10 code	ICD-10 coding of the diagnosis	J45.0	List according to https://icd.who.int/browse10/2019/en
ICD-11 code	ICD-11 coding of the diagnosis	CA23.0	List according to https://icd.who.int/en/

Field name	Availability			
	PAS Xanthos	ADOPT cohort	CRC Cohort	MIABIS
Parent request ID				
Container ID	Int	Not available	UUID	Text
Human readable container ID	Int	Not available	UUID	Text
Material SNOMED-CT ID	Not available	Not available	Not available	Fixed value set
Material	Text	Fixed value set	Text	Fixed value set
Material description	Text	Fixed value set	Text	Not available
Extraction method	Text	Not available	Not available	Not available
Macroscopic description	Text	Not available	Not available	Not available
Histological description	Text	Not available	Not available	Not available
Molecularpathological diagnosis	Text	Not available	Not available	Not available
Molecularpathological description	Text	Not available	Not available	Not available
Diagnosis	Text	Not available	Not available	Fixed value set
Neuropathological report	Text	Not available	Not available	Not available
Rapid section examination	Text	Not available	Not available	Not available
Cause of death	Text	Not available	Not available	Not available
Postmortem examination	Text	Not available	Not available	Not available
Cytological description	Text	Not available	Not available	Not available
Report type	Text	Not available	Not available	Not available
TNM T prefix	Text	Fixed value set	Text	Not available
TNM T	Text	Fixed value set	Text	Not available
TNM T addition	Text	Fixed value set	Text	Not available
TNM N prefix	Text	Fixed value set	Text	Not available
TNM N	Text	Fixed value set	Text	Not available
TNM N addition	Text	Fixed value set	Text	Not available
TNM M prefix	Text	Fixed value set	Text	Not available
TNM M	Text	Fixed value set	Text	Not available
TNM M addition	Text	Fixed value set	Text	Not available
TNM stage	Text	Fixed value set	Text	Not available
TNM UICC	Text	Fixed value set	Text	Not available
ICD-O code	Text	Not available	Text	Not available
ICD-10 code	Not available	Not available	Text	Fixed value set
ICD-11 code	Not available	Not available	Not available	Not available

Data fields Level 5 – Block

Field name	Description	Example	Data type
Parent container ID	Identifier of the Container		Impicit link to layer above
Block ID	Block Identifier	2024000001/01	Text
Human readable block ID	Block Identifier in human readable format	2024000001/01	Text
Location of stored block	Location of stored block (e.g. Biobank Graz, Archive)	Biobank Graz	Text
Access condition block	Access condition of the block	Open access	'Open access'; 'Restricted access'; 'Controlled access'; 'Non-commercial use only'; 'Academic use only'; 'Personal use only'; 'Institutional access only'; 'Embargoed'
Access condition block date	Date until access condition of the block is valid	31.01.2000	Date
Fixation type	Fixation type	Frozen section	'Formalin fixation'; 'Alcohol fixation'; 'Frozen section'; 'Glutaraldehyde fixation'; 'Michel's medium'; 'Bouin's solution'; 'Cytology preservative solutions'; 'Unfixed tissue'
Block label	Block label (e.g. TU, 1, My, E, Po, Para, AR, OR, LY)	TU	Text

	Availability			
Field name	PAS Xanthos	ADOPT cohort	CRC Cohort	MIABIS
Parent container ID				
Block ID	Int	Not available	UUID	Not available
Human readable block ID	Int	Not available	UUID	Not available
Location of stored block	Not available	Not available	Not available	Not available
Access condition block	Not available	Not available	Not available	Not available
Access condition block date	Not available	Not available	Not available	Not available
Fixation type	Not available	Fixed value set	Not available	Not available
Block label	Text	Not available	UUID	Not available

Data fields Level 6 – Slide

Field name	Description	Example	Data type
Parent block ID	Identifier of the Block		Impicit link to layer above
Slide ID	Slide Identifier	2024000001/01/01	Text
Human readable slide ID	Slide Identifier in human readable format	2024000001/01/01	Text
Location of stored slide	Location of stored Slide (e.g. Biobank Graz, Archive)	Biobank Graz	Text
Access condition slide	Access condition slide	Open access	'Open access'; 'Restricted access'; 'Controlled access'; 'Non-commercial use only'; 'Academic use only'; 'Personal use only'; 'Institutional access only'; 'Embargoed'
Access condition slide date	Date until access condition of the slide is valid	31.01.2000	Date
Slide label	Label of slide (e.g. /2, /7)	/2	Text
Assay type	The laboratory test performed on a bio-specimen mounted on a glass slide	Enumerated Values	'Hematoxylin And Eosin Stain'; 'Special Stains'; 'Immunohistochemistry'; 'Immunocytochemistry'; 'Immunofluorescence'; 'Insitu Hybridization'; 'Fluorescence Insitu Hybridization'; 'Fluorescent Multiplex Immunohistochemistry'; 'Multiplex Immunohistochemistry'; 'other (the value is known but cannot be classified into any of the predefined categories)'; 'NULL (the value is unknown)'

Field name	Description	Example	Data type
Special stain technique	The method of a special staining technique performed to visualize a selected tissue type or tissue element, a cell type, or a pathogen, except for the H&E method	Enumerated Values	Values are to be taken from CID 8112 Specimen Stain and be encoded as SNOMED-CT ID
Dye name	The neutral or synthetic colored substance used for staining specific tissues or cellular components	Enumerated Values	Values are to be taken from CID 8112 Specimen Stain and be encoded as SNOMED-CT ID
Antibody name	The vendor-stated name of an antibody used to detect and visualize the presence of an antigen	Recombinant Monoclonal Anti-Ki-67 Antibody, Rabbit (7A4)	Text
Antibody identifier	The distinct ID for an antibody used in an assay	AB_3741092	List according to https://www.antibodyregistry.org/; else use Freetext
Antibody clone freetext	Name of the used antybody	Recombinant Monoclonal Anti-Ki-67 Antibody	Text

Field name	Availability			
	PAS Xanthos	ADOPT cohort	CRC Cohort	MIABIS
Parent block ID				
Slide ID	Int	Text	UUID	Text
Human readable slide ID	Int	Text	UUID	Text
Location of stored slide	Not available	Not available	Not available	Not available
Access condition slide	Not available	Not available	Not available	Not available
Access condition slide date	Not available	Not available	Not available	Not available
Slide label	Text	Not available	Not available	Not available
Assay type	Text	Not available	Not available	Fixed value set
Special stain technique	Not available	Not available	Not available	Fixed value set
Dye name	Not available	Not available	Not available	Fixed value set
Antibody name	Text	Not available	Not available	Text
Antibody identifier	Not available	Not available	Not available	Fixed value set
Antibody clone freetext	Text	Not available	Not available	Not available

Data fields Level 7 – WSI

Field name	Description	Example	Data type
Parent slide ID	Identifier of the Slide		Impicit link to layer above
WSI ID	WSI Identifier	d7880b0a-b2a2-488e-a4ab-67103fd225e8	Text
Human readable WSI ID	WSI Identifier in human readable format	2024000001/01/01/01	Text
Location of stored WSI	Location of stored WSI	\\DigitalArchive4245\WSI\d7880b0a-b2a2-488e-a4ab-67103fd225e8	Text
Access condition WSI	Access condition WSI	Open access	'Open access'; 'Restricted access'; 'Controlled access'; 'Non-commercial use only'; 'Academic use only'; 'Personal use only'; 'Institutional access only'; 'Embargoed'
Access condition WSI date	Date until access condition of the WSI is valid	31.01.2000	Date
Manufacturer of scanner	Manufacturer of Scanner	Aperio	Text
Type of scanner	Type of Scanner	GT 450 DX	Text
Serial number of scanner	Serial number of Scanner	SS45789	Text
Resolution	Resolution of Scan in µm/Pixel	0,25	Decimal
Scan date and time	Scan date and time	31.01.2025 06:00:00	Datetime
Imaging technique	Imaging technique	Brightfield	'Transmission'; 'Reflection'; 'Epifluorescence'; 'Brightfield'; 'Darkfield'; 'Oblique'; 'Phase Contrast'; 'Differential Interference Contrast'; 'Total Internal Reflection Fluorescence'

Field name	Description	Example	Data type
Compression method	Compression method applied	JPEG	'JPEG'; 'JPEG 2000'; 'JPEG-LS'; 'JPEG XR'; 'PNG (DEFLATE)'; 'TIFF (LZW)'; 'TIFF (PackBits)'; 'TIFF (Deflate / ZIP)'; 'WebP'; 'HEIF / HEIC'; 'AVIF'; 'GIF (LZW)'; 'BMP (RLE)'; 'BPG'; 'FLIF'; 'JPEG XL'; 'LERC (Limited Error Raster Compression)'; 'Zstandard (Zstd)'; 'LZMA'
Compression quality level	Compression quality level	80	Text
Image format	Image format	dcm	'svs'; 'ndpi'; 'scn'; 'mrxs'; 'czi'; 'vms'; 'vmu'; 'bif'; 'tif'; 'tiff'; 'ometiff'; 'isyntax'; 'kfb'; 'svslide'; 'dcm'; 'zvi'; 'lif'
Firmware version	Firmware version	1.2.5	Text
Z-stacked	Z-stacked	single	'single'; 'multiple'; 'extended'
Scan area	Scan Area	full slide	'full slide'; 'partial'
WSI annotation available	WSI annotation available	Yes	'Yes'; 'No'

	Availability			
Field name	PAS Xanthos	ADOPT cohort	CRC Cohort	MIABIS
Parent slide ID				
WSI ID	Text	Not available	UUID	Text
Human readable WSI ID	Int	Not available	UUID	Text
Location of stored WSI	Not available	Not available	Not available	Not available
Access condition WSI	Not available	Not available	Not available	Not available
Access condition WSI date	Not available	Not available	Not available	Not available
Manufacturer of scanner	Not available	Not available	Not available	Text
Type of scanner	Not available	Not available	Not available	Text
Serial number of scanner	Not available	Not available	Text	Not available
Resolution	Not available	Not available	Text	Integer
Scan date and time	Datetime	Not available	Text	Datetime
Imaging technique	Not available	Not available	Not available	Fixed value set
Compression method	Not available	Not available	Not available	Fixed value set
Compression quality level	Not available	Not available	Not available	Decimal
Image format	Not available	Not available	Text	Fixed value set
Firmware version	Not available	Not available	Not available	Text
Z-stacked	Not available	Not available	Not available	Fixed value set
Scan area	Not available	Not available	Not available	Not available
WSI annotation available	Not available	Not available	Not available	Fixed value set

List of Material mapped to SNOMED-CT

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Histology	STUHL	119339001	Stool specimen	specimen
Histology	MAMMA	127456000	Specimen from breast	specimen
Histology	WEICHTEILGEWEBE	309072003	Soft tissue specimen	specimen
Histology	MAMMA TE	127456000	Specimen from breast	specimen
Histology	VAGINA	119394009	Specimen from vagina	specimen
Histology	MILZ	433308004	Specimen from spleen	spceimen
Histology	MUNDHOEHLE	309185002	Oral cavity specimen	specimen
Histology	MUNDHÖHLEN CA	309185002	Oral cavity specimen	specimen
Histology	NECKDISSECTION	309476009	Neck block dissection specimen	specimen
Histology	NETZ (OMENTUM MAJUS)	433309007	Specimen from omentum	specimen
Histology	NIERE CA	127474009	Tissue specimen from kidney	specimen
Histology	PHARYNX	309168004	Upper respiratory tissue specimen	specimen
Histology	TONSILLE	430222001	Specimen from tonsil	spceimen
Histology	MAMMA ABLATIO	127456000	Specimen from breast	specimen
Histology	HODEN	128154006	Specimen from testis	spceimen
Histology	NIERE	127474009	Tissue specimen from kidney	specimen
Histology	HYPOPHYSE	309154006	Pituitary specimen	specimen
Histology	MUSKELBIOPSIE	309507001	Muscle biopsy specimen	specimen
Histology	PORTIOKONUS	127481002	Tissue specimen from uterine cervix	specimen
Histology	LARYNX	430144001	Specimen from larynx	specimen
Histology	LEBER	119383005	Specimen from liver	specimen
Histology	KNOCHEN	258417006	Bone tissue specimen	specimen
Histology	LYMPHKNOTEN	258589002	Lymph node specimen	specimen
Histology	LUNGE	127458004	Specimen from lung	specimen
Histology	NEBENNIERE	309141004	Adrenal gland specimen	specimen
Histology	NASENNEBENHÖHLE	430238007	Specimen from nasal sinus	specimen
Histology	PE DES OBEREN GI-TRAKTES	608842007	Specimen from upper gastrointestinal tract	specimen
Histology	TUBE	127476006	Specimen from fallopian tube	specimen
Histology	HEMICOLEKTOMIE RECHTS	122648004	Specimen from colon obtained by right hemicolectomy	specimen
Histology	HEMICOLEKTOMIE LINKS	122650007	Specimen from colon obtained by left hemicolectomy	specimen
Histology	UTERUS	127479004	Specimen from uterus	specimen

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Histology	GELENK	309125000	Joint specimen	specimen
Histology	VULVA	397129002	Specimen from vulva	specimen
Histology	CYSTEKTOMIE + UTERUS	309275006	Cystectomy specimen	specimen
Histology	WERTHEIM OPERATIONSPRÄPARAT	309501000	Hysterectomy specimen	specimen
Histology	CYSTEKTOMIE + PROSTATA	309275006	Cystectomy specimen	specimen
Histology	WHIPPLE OPERATIONSPRÄPARAT	122665008	Specimen from pancreas obtained by pancreaticoduodenectomy (Whipple resection)	specimen
Histology	OESOPHAGUS	127463000	Specimen from oesophagus	specimen
Histology	RECTUM SAUGBIOPSIE	309262006	Rectal biopsy specimen	specimen
Histology	PARAFFINBLOCK	3040001000004107	Paraffin embedded tissue block specimen	specimen
Histology	OVAR	128155007	Specimen from ovary	specimen
Histology	PANKREAS	127469001	Specimen from pancreas	specimen
Histology	PENIS	128169007	Tissue specimen from penis	specimen
Histology	AMPUTAT	408654003	Specimen obtained by amputation	specimen
Histology	PERITONEUM	430250001	Specimen from peritoneum	specimen
Histology	PLAZENTA	122736005	Tissue specimen from placenta	specimen
Histology	ARTERIE	309479002	Artery specimen	specimen
Histology	PLEURA	127459007	Specimen from pleura	specimen
Histology	PORTIO	127481002	Tissue specimen from uterine cervix	specimen
Histology	COLON POLYP (PO)	309226005	Colonic polyp specimen	specimen
Histology	PORTIO UTERI	127481002	Tissue specimen from uterine cervix	specimen
Histology	PROSTATA	119386002	Specimen from prostate	specimen
Histology	NERVENBIOPSIE	309130001	Nerve biopsy specimen	specimen
Histology	RECTUM CA / OP-PRÄP.	447955000	Specimen from rectum	specimen
Histology	SCHILDDRÜSE	399680007	Specimen from thyroid	specimen
Histology	SPEICHELDRÜSE	399747006	Specimen from salivary gland	specimen
Histology	HIRN	128157004	Tissue specimen from brain	specimen
Histology	MAMMA TE IN SERIE (MIKROKALK)	127456000	Specimen from breast	specimen
Histology	HERZ	128166000	Tissue specimen from heart	specimen
Histology	BECKENKAMM	271515009	Iliac crest bone marrow specimen	specimen

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Histology	SENTINEL MALIGNES MELANOM	441709004	Specimen from sentinel lymph node	specimen
Histology	BLUT	119297000	Blood specimen	specimen
Histology	BRONCHUS	128158009	Tissue specimen from bronchus	specimen
Histology	MAMMAREDUKTION	127457009	Tissue specimen from breast	specimen
Histology	MAGEN	128171007	Tissue specimen from stomach	specimen
Histology	COLON	128159001	Tissue specimen from colon	specimen
Histology	MEDIASTINUM	433326001	Specimen from mediastinum	specimen
Histology	CYSTEKTOMIE (HARNBLASE)	309275006	Cystectomy specimen	specimen
Histology	URETHRA	128174004	Tissue specimen from urethra	specimen
Histology	AXILLA LYMPHKNOTEN	258589002	Lymph node specimen	specimen
Histology	NEBENSCHILDDRÜSE	309149002	Parathyroid specimen	specimen
Histology	ABORTUSMATERIAL	119396006	Specimen from endometrium	specimen
Histology	MUNDHÖHLE	309185002	Oral cavity specimen	specimen
Histology	ADENOIDE	309168004	Upper respiratory tissue specimen	specimen
Histology	ADNEXE	128155007	Specimen from ovary	specimen
Histology	DUCTUS CHOLEDOCHUS	309491006	Biliry tract tissue specimen	specimen
Histology	ANUS	128156008	Tissue specimen from anus	specimen
Histology	MYOKARDBIOPSIEN	119377007	Myocardial specimen	specimen
Histology	AORTA	725372003	Tissue specimen from aorta	specimen
Histology	APPENDIX VERMIFORMIS	309222007	Appendix specimen	specimen
Histology	AUGE	128164002	Tissue specimen from eye	specimen
Histology	ANASTOMOSE	127465007	Specimen from gastrointestinal tract	specimen
Histology	URETER	399624001	Tissue specimen from ureter	specimen
Histology	COLON DIVERTIKULOSE	122651006	Specimen from colon obtained by sigmoidectomy	specimen
Histology	DUCTUS DEFERENS	309139000	Vas deferens specimen	specimen
Histology	DÜNNDARM	122638001	Tissue specimen from small intestine	specimen
Histology	GALLENBLASE	122656001	Tissue specimen from gallbladder	specimen
Histology	GEWEBE (ohne nähere Zuordnung)	119376003	Tissue specimen	specimen
Histology	VENE	309480004	Vein specimen	specimen

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Histology	DISCUS	438960000	Specimen from intervertebral disc	specimen
Histology	HARNBLASE	309272009	Urinary bladder tissue specimen	specimen
Histology	HAUT	119325001	Tissue specimen from skin	specimen
Histology	HERZKLAPPE	258424007	Heart valve tissue	specimen
Histology	OVARIAL TUMOR OP	128155007	Specimen from ovary	specimen
Histology	SENTINEL LYMPHKNOTEN MAMMA	16235301000119101	Specimen from sentinel lymph node of breast	specimen
Histology	NASENHÖHLE	119388001	Specimen from internal nose	specimen
Histology	ZNS	399542003	Tissue specimen from central nervous system	specimen
Histology	SENTINEL VULVA; ENDO; OV	441709004	Specimen from sentinel lymph node	specimen
Histology	DÜNNDARM TUMOR	119380008	Specimen from small intestine	specimen
Histology	WEICHTEIL TUMOR	309072003	Soft tissue specimen	specimen
Histology	KNOCHEN TUMOR	258417006	Bone tissue specimen	specimen
Histology	UTERUS ADNEXE	309287004	Hysterectomy and bilateral salpingo-oophorectomy specimen	specimen
Histology	SENTINEL LYMPHKNOTEN	441709004	Specimen from sentinel lymph node	specimen
Histology	Hämato Lymphknoten	309078004	Lymph node tissue specimen	specimen
Histology	ZAHN	430319000	Specimen from tooth	specimen
Histology	NEBENHODEN	309140003	Epididymis specimen	specimen
Histology	PERICARD	699874006	Tissue specimen from pericardium	specimen
Histology	MUNDHÖHLEN CA mit NECK DISSECTION	309185002	Oral cavity specimen	specimen
Histology	ISOLIERTE DNA	258566005	Deoxyribonucleic acid specimen	specimen
Histology	DUODENUM	122638001	Tissue specimen from small intestine	specimen
Gynecological Cytology	Cervikalkanal	258524009	Cervical swab	specimen
Gynecological Cytology	Scheidenblindsack	258520000	Vaginal swab	specimen
Gynecological Cytology	Portio	258524009	Cervical swab	specimen
Gynecological Cytology	Vulva	258523003	Vulva swab	specimen
Gynecological Cytology	Vagina	258520000	Vaginal swab	specimen
Extragenital cytology	Stuhl	119339001	Stool specimen	specimen

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Extragenital cytology	MEDIASTINUM	433326001	Specimen from mediastinum	specimen
Extragenital cytology	Nebenschilddrüse	309149002	Parathyroid specimen	specimen
Extragenital cytology	Vorderkammer-Punktat	258444001	Aqueous humor specimen	specimen
Extragenital cytology	Niere	127474009	Tissue specimen from kidney	specimen
Extragenital cytology	Ovar	128155007	Specimen from ovary	specimen
Extragenital cytology	Knochenmarksaspirat	119359002	Bone marrow specimen	specimen
Extragenital cytology	Pankreas	127469001	Specimen from pancreas	specimen
Extragenital cytology	PERICARD	430244006	Specimen from pericardium	specimen
Extragenital cytology	Gelenk	309125000	Joint specimen	specimen
Extragenital cytology	Peritoneum	430250001	Specimen from peritoneum	specimen
Extragenital cytology	Pleura	127459007	Specimen from pleura	specimen
Extragenital cytology	Pharynx	472881004	Swab from pharynx	specimen
Extragenital cytology	Prostata	119386002	Specimen from prostate	specimen
Extragenital cytology	Retroperitoneum	433323009	Specimen from retroperitoneum	specimen
Extragenital cytology	Schilddrüse	399680007	Specimen from thyroid	specimen
Extragenital cytology	Neuro Zytologie (Punktion)	258450006	Cerebrospinal fluid specimen	specimen
Extragenital cytology	Mamma FNA (cystische Läsion)	110894008	Mammary cytologic material	specimen
Extragenital cytology	Mamma FNA (solide Läsion)	110894008	Mammary cytologic material	specimen
Extragenital cytology	Mamma FNA	110894008	Mammary cytologic material	specimen
Extragenital cytology	Larynx	430144001	Specimen from larynx	specimen
Extragenital cytology	Mamillensekret	110894008	Mammary cytologic material	specimen
Extragenital cytology	Transthorakale Punktionszyt.	127458004	Specimen from lung	specimen
Extragenital cytology	Lymphknoten	258589002	Lymph node specimen	specimen
Extragenital cytology	Blut	119297000	Blood specimen	specimen
Extragenital cytology	Magensaft	258459007	Gastric fluid sample	specimen
Extragenital cytology	Plasma	119361006	Plasma specimen	specimen

Section	Label	SNOMED-CT ID	SNOMED-CT Term	Semantic
Extragenital cytology	Speicheldrüsen	399747006	Specimen from salivary gland	specimen
Extragenital cytology	Vitrektomiematerial	258438000	Vitreous humor specimen	specimen
Extragenital cytology	Sputum	119334006	Sputum specimen	specimen
Extragenital cytology	Liquor	258450006	Cerebrospinal fluid specimen	specimen
Extragenital cytology	Vorderkammer-Punktat	258444001	Aqueous humor specimen	specimen
Extragenital cytology	Weichgewebe	309072003	Soft tissue specimen	specimen
Extragenital cytology	Zellen	258591005	White blood cell specimen	specimen
Extragenital cytology	BAL	258607008	Bronchoalveolar lavage fluid specimen	specimen
Extragenital cytology	Lunge,Bronchien Exfoliativmat.	127458004	Specimen from lung	specimen
Extragenital cytology	Gallenblase und Gallegänge	449446003	Specimen from gallbladder	specimen
Extragenital cytology	Gastrointestinale Bürstenzyto	127465007	Specimen from gastrointestinal tract	specimen
Extragenital cytology	Knochenmark/Blut	119359002	Bone marrow specimen	specimen
Extragenital cytology	Gastrointestinale FNA	127465007	Specimen from gastrointestinal tract	specimen
Extragenital cytology	Leber	119383005	Specimen from liver	specimen
Extragenital cytology	Harn - Lavage	122575003	Urine specimen	specimen
Extragenital cytology	Harn - Spontanurin	122575003	Urine specimen	specimen
Extragenital cytology	Harn ohne Gewinnungsangaben	122575003	Urine specimen	specimen
Extragenital cytology	Endozervix (CK Bürstenabstriich)	258524009	Cervical swab	specimen
Extragenital cytology	Abstrich zur Erregerdiagn.	257261003	Swab	specimen
Extragenital cytology	ISOLIERTE DNA	258566005	Deoxyribonucleic acid specimen	specimen