Ing. Markus Plass, BSc

Large Scale Slide Digitalisation for Machine Learning in Computational Pathology

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieur

Master's degree programme:
Biomedical Engineering

submitted to

Graz University of Technology

Supervisor
Univ.-Prof. Dipl.-Ing. Dr.techn. Rudolf Stollberger
Institute of Medical Engineering
Univ.-Doz. Mag.phil. Mag.rer.nat. Dr.phil. Ing. Andreas Holzinger
Institute of Interactive Systems and Data Science

Graz, November 2020

This page intentionally left blank

Abstract

Computational pathology is a thriving research domain that uses large volumes of imaging data accompanied by sensitive clinical data collected to heterogenous data models to fundamentally improve how the histopathological and clinical diagnosis and oncological treatment of the patients is performed (1). The recent developments of high-throughput slide scanners offer a possibility for making the contained information of the glass slides stored in biobanks available for machine learning algorithms. Ensuring storage and access to digital slides, also called whole slide images (WSI), will overcome the current limitations to accessing and sharing pathology material together with the associated metadata.

This work describes the design and implementation of the digitization workflow, consisting of the following steps: (i) Selection and description of a scanning cohort by it's metadata, (ii) retrieving of the physical slides from the archive, (iii) cleaning and pre-processing, (iv) scanning with different scanners, (v) quality control, (vi) generation of technical scan metadata, (vii) linkage to phenotypical descriptions and (viii) cataloguing and long term storage.

The results were published in two papers and implemented as scanning infrastructure (software and hardware) consisting of a central database and several web interfaces to model the business logic of the scanning workflow. The solution is currently in production, and was already used for scanning and managing of more than 300.000 slides.

Keywords

Digital Pathology, Whole Slide Image, Digitisation, Machine Learing, Biobanking

This page intentionally left blank

Kurzfassung

Digitale Pathologie ist ein florierender Forschungsbereich, in dem große Mengen von Bilddaten zusammen mit sensiblen klinischen Daten, die in heterogenen Datenmodellen gesammelt wurden, verwendet werden, um die Durchführung der histopathologischen und klinischen Diagnose und onkologischen Behandlung der Patienten grundlegend zu verbessern (1). Die jüngsten Entwicklungen von Hochdurchsatz-Scannern bieten die Möglichkeit, die enthaltenen Informationen der in Biobanken gespeicherten Glasobjektträger für Algorithmen für maschinelles Lernen verfügbar zu machen. Durch die Sicherstellung der Speicherung und des Zugriffs auf digitale Objektträger, auch als WSI (Whole Slide Images) bezeichnet, werden die aktuellen Einschränkungen für den Zugriff auf und die gemeinsame Nutzung von Pathologiematerial zusammen mit den zugehörigen Metadaten überwunden.

Diese Arbeit befasst sich mit der Gestaltung und Anwendung digitaler Arbeitsprozesse, die sich aus folgenden Schritten zusammen setzen: (i) Auswahl und Beschreibung der gescannten Kohorte anhand der Metadaten, (ii) Aushebung der physischen Objektträger aus dem Archiv, (iii) die Reinigung und die Vorverarbeitung, (iv) das Digitalisieren mit unterschiedlichen Scannern, (v) Qualitätskontrolle, (vi) Erzeugung technischer Metadaten, (vii) die Verbindung zu phänotypischen Daten und (viii) die Katalogisierung und Langzeitlagerung.

Die wissenschaftlichen Ergebnisse wurden bereits in zwei Fachartikeln publiziert. Des Weiteren, werden die Resultate für die Scan Infrastruktur (Software und Hardware) genutzt, welche aus einer zentralen Datenbank und diversen Internetschnittstellen besteht, die den Scan-Workflow modellieren. Die Lösung ist derzeit in Produktion und wurde bereits zum Scannen und Verwalten von mehr als 300.000 Objektträger verwendet.

Schlüsselwörter

Digitale Pathologie, Digitalisierung, Machine Learing, Biobanking

This page intentionally left blank

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Graz, November 9, 2020

Markus Plass

This page intentionally left blank

Acknowledgements

I would like to express my gratitude to my supervisors Univ.-Prof. Dipl.-Ing. Dr.techn. Rudolf Stollberger and Univ.-Doz. Mag.phil. Mag.rer.nat. Dr.phil. Ing. Andreas Holzinger who gave me the opportunity to write this thesis.

I would like to thank Prof. Kurt Zatloukal and Dr. Heimo Müller from the Medical University of Graz, who placed the confidence in me by providing me with the resources to establish the Scanning-Lab.

A special thanks also to my Scanning-Team and my colleagues who accompanied me during the realization of the workflow.

Thanks to all of you!

Markus Plass
Graz, November 2020

This page intentionally left blank

Table of Contents

1. Introduction

Computational pathology is a thriving research domain that uses large volumes of imaging data accompanied by sensitive clinical data collected to heterogenous data models to fundamentally improve how the histopathological and clinical diagnosis and oncological treatment of the patients is performed.(1) The recent developments of high-throughput slide scanners offer a possibility for making the contained information of the glass slides stored in biobanks available for machine learning algorithms. Ensuring storage and access to digital slides, also called Whole Slide Images (WSI), will overcome the current limitations to accessing and sharing pathology material together with the associated metadata.

This work describes the design and implementation of the high throughput digitization workflow, consisting of the following steps: (i) Selection and description of a scanning cohort by it's metadata, (ii) retrieving of the physical slides from the archive, (iii) cleaning and pre-processing, (iv) scanning with different scanners, (v) quality control (vi) generation of technical scan metadata (vii) linkage to phenotypical descriptions and (viii) cataloguing and long term storage. The main three achievements of the master thesis are the definition of the MISS (Minimal Information about Slides and Scans) specification (2) algorithms for extraction and reconstruction of hierarchical information and relationships in Whole Slide Images (3) and the high throughput digitization workflow.

1.1 Digitization in Medicine

In recent years, the importance of digital medical imaging techniques such as CT, MRI, sonography or PET has steadily increased and other techniques such as projection radiography have become increasingly digital. This digitalization enabled the opportunity for the development of new diagnostic methods and also increased the efficiency of the health care system.

In pathology, imaging also plays a big role. Several years ago an analog photo-camera was mounted the microscope of the pathologist, so he was able to take pictures of regions of interest (ROI) and share them with others. The term telepathology was shaped in the 1980s starting with a remotely operated microscope. Over the last years, this has been replaced by a digital-photo-camera or a video camera to document the slide. With the increasing technical possibilities the first Whole-Slide-Image-Scanners (WSI-Scanners) were developed. This invention allowed to take pictures of the complete slide and the possibility to navigate through a digital version of the cut.

The digital images enable the possibility to share slides in real-time with the benefit of bridging physical distance (telepathology) between local hospitals and colleges for getting a second-opinion, and enabling home-office. New viewers allow alignment between different stains and across slides and measurements within the slides. These digital images lend themselves to computational pathology, for example basic tasks like measuring/counting and deep learning tasks. The evaluation with machine learning allows to look for features beyond the assessment of traditional histopathology (view through microscope), and allows direct links of the images to clinical data (e.g., prognosis, mutations). These emerging AI algorithms allow a certain amount of automatization and can decrease the workload of the medical staff. Beside these benefits, this digitization process also goes along with massive investments in IT-Infrastructure to get a reliably system, the need for new workflows for a safe implementation, regulatory requirements, artificial intelligence as an unaccountable "black-box", questions of cost-efficacy, and the transformation of the profession by automation.

1.2 Computational Pathology

High quality metadata and provenance information are essential to support product quality in almost all areas of computational pathology. We need the appropriate information to document the technical and medical validation and to support the regulatory approval process. There are several standards available covering dedicated parts, e.g. MIABIS for sample and donor metadata and DICOM or vendor specific attributes for file formats and scanning metadata. Our aim is not to propose yet another metadata standard, but to describe a small and minimal dataset across different standardization activities and initiate a community driven approach to collect and harmonize existing ontologies. In addition, MISS was defined within the use cases of a large scale digitization effort for machine learning. The minimal information about glass slides and their scanned representation is divided into three parts: Pre Scanning (Slide) Metadata: e.g. metadata from biobanks, glass slide labeling, cleaning; Scanning Metadata: e.g. technical parameters, resolutions and focus points and Post-Scanning (File) Metadata: e.g. image quality indicators.

Tissue areas on Whole Slide Images may be organized as single structures, in symmetric objects or spread over the carrier in a complex way. Since there is a huge amount of possible combinations of areas that may or may not form objects and objects that may be very similar there is a need of objective metrics to describe WSI training data sets.

1.2.1 Slide Digitisation

During the last decade, pathology has benefited from the rapid progress of image digitizing technologies, which led to the development of scanners capable to produce Whole Slide Images which can be explored by a pathologist on a computer screen (virtual microscope) comparable to the conventional microscope and can be used for education and training, diagnostics (clinico-pathological meetings, consultations, revisions, slide panels and upfront clinical diagnostics) and archiving.(4)
Compared to radiology, where the typical file sizes are in the range from 500 KB to 50 MB, a single WSI scan with 40x magnification consists of approximately 16 Gigapixels (Note: for the calculation of the WSI file size and comparison of different scanner manufacturers, the de-facto standard area of 15mm x 15mm, with an optical resolution of 0.12μm, which corresponds to a 80x magnification was used).

1.2.2 Whole Slide Images for Machine Learning and AI

Recent developments in high-throughput slide scanners offer the possibility for making the entire information contained in the millions of glass slides produced every year, available for machine learning applications. Access to whole slide images and related medical data will overcome the current limitations of accessing and sharing pathology material and will facilitate the development of new machine learning algorithms. In order to develop these algorithms, a large series of slides offering a broader coverage of tissues and cancer type / pathological deviations are required. To address this demand, samples and data from different biobanks in different countries must be suitable for integrated analyses. This is only possible if samples and data meet common quality criteria. Therefore, international standards (e.g. CEN Technical Specifications or ISO Standards) were implemented for sample pre-analytics, covering all steps from the sample collection of the patient to isolation of bio molecules (5), and (open-source) software for cataloguing and provenance management was developed, e.g. for rare diseases (6) and for biobanks in low and medium income countries.(7; 8)

Extraction and reconstruction of hierarchical information and relationships of objects in images are common problems, not only in medicine. The automated detection of certain objects and groups of objects is a highly specific task and depends on overall noise, background property, object sizes and shapes, colors and image qualities. With rising amounts of digital image data in the field of pathology, it is highly important to identify, label and structure those piles of information. Finding similarities between objects and groups of objects are intuitive tasks performed by human eyes and the brain but are far from trivial to be automated. Digital pathology generates data which is organised in different types and combined with lots of meta information .(1) Extracting, condensing and structuring these kinds of information are highly sensitive but crucial tasks and are mandatory parts to enable the possibilities of knowledge discovery processes in big data environments.

As described above, tissue areas on glass carrier plates can be organized as single structures, in symmetric objects or spread over the carrier in a complex way. Since there is a huge amount of possible combinations of areas that may or may not form objects and objects that may be very similar in size, shape and amount of areas they contain, there is a need of objective metrics that help to extract structural information and enhance the possibilities of methods for visualization.

It is easy for humans to identify tissue areas that "look similar" and to describe and pinpoint symmetry properties even if areas of interest are altered by air bubbles or folded tissues, for instance. Automatically extracting information containing the exact number and the size of tissue areas, identifying groups of tissues that share certain properties and calculate metrics of similarities between single tissues and whole constellations of various areas are problems that are not easy to solve.

1.3 Biobank Graz

Biobanks collect, preserve, and provide access to samples, e.g. from pathology in a transparent and quality controlled manner in compliance with ethical, legal, and regulatory requirements for research.(9) They require access to sufficient numbers of samples and data that properly cover the broad spectrum of disease sub-entities relevant for targeted therapies .(10) The Biobank Graz contains a collection of over 11 million paraffin-embedded slides, together with the information on the initial diagnosis, disease outcome and overall survival. Sub collections of the slide archive of the Biobank Graz are currently being digitized for several research projects and industry cooperations.

2. Background

Today (precision) medicine increasingly relies on rich data sets for better detection and treatment of diseases. This statement is particularly true for medical imaging and digital pathology. Digital pathology is more than just the transformation of the microscopic analysis of histological slides by pathologists to the digital domain (screen). Digital pathology and machine learning will change the education and training of pathologists (an urgently needed solution to address the global shortage of medical specialists) and it will generate new business models for diagnostic services (telepathology and AI assisted pathology). It is expected that several of the solutions developed in the context of digital pathology are also relevant to other fields of medical data analysis.

The development of machine learning algorithms in digital pathology requires access to large data sets that will cover the variety of human diseases in different organ systems.(1) Such data sets have to meet quality and regulatory criteria of medical devices (raw data) and be described by all necessary metadata, from patient to sample to the whole slide image. A good source for such data sets are biobanks, which provide samples and related data in a quality controlled matter. In the following section some basic definitions are given and relevant (data) standards from biobanking are presented.

2.1 Basic Definitions

Digitization and/or extraction of parameters is a frequently used entrance point for AI algorithm development. Biological samples are one of the key raw material for the generation of such data sets. The collection, preservation, and storage of biological samples, in addition to provision of access, are key activities of biobanks.

Therefore, it is essential that biobanks ensure proper quality of samples and data, ethical and legal compliance.(9)

A biobank is the legal entity or part of a legal entity that performs biobanking, which is the process of acquisitioning and storing, together with some or all of the activities related to collection, preparation, preservation, testing, analysing and distributing defined biological material as well as related information and data. The ISO standard 20387 defines biological samples / materials and any substance derived or part obtained from an organic entity such as a human, animal, plant, microorganism(s) or multicellular organism(s).(11)

In a biobank, endurant and perdurant entities can be distinguished, as described in (12):
Perdurant Entities denote temporal items such as events, periods, activities and processes. Perdurant entities happen over a limited continuous extent in time. If an event occurs later again at another point in time, we assume that the former event has ended and a new instance has come into existence, i.e. an event / activity cannot not happen twice. Quality management often defines templates for perdurant entities, usually called processes, as well as requirements for their documentation and assessment.

In contrast to perdurant entities *Endurant Entities* are documented as a single unit of discourse, also called persistent things. Such entities are either *Physical Entities*, *Conceptual Entities* or *Agents*. Physical entities / things in a biobank are e.g. samples, freezers, and containers. Physical entities can be moved and describe during their existence a trajectory in space and time, e.g. the transport, storage or disposal of a sample.

Conceptual Objects are also objects of a discourse, but they are in contrast to physical entities non-material products of our minds. A biobank, a collection, a study, an observation and a diagnosis are all conceptual objects. The production of such entities may originate by humans or by technical devices such as laboratory equipment. *Agents* are people (individually or in groups) or machines (algorithms) who have the potential to perform an intentional action. An agent is an entity that bears some form of responsibility for an activity or for the existence of an entity. A real person can be both a physical entity, e.g. as a donor of a sample, but also an agent as e.g. the responsible person for the scanning of slides stored in a biobank.

Agents can access and interact with entities. They can be on the one hand human beings, but also machines (scanners, robots) or even algorithms that perform actions on entities.

A conceptual object can exist on more than one particular *Carrier* at the same time. Example of carriers are the human memory, physical entities as paper or film and - most common today - electronic formats. It is important to note that conceptual objects cannot be destroyed, they exist as long as at least one carrier (even if it is only human memory) exists. Physical carriers for conceptual objects are known as *Data Objects*.[1] A carrier of a data object can be a digital object (stored as bitstream) or a physical entity, e.g. a sheet of paper.

A *Provenance Graph* is composed of endurant entities (nodes), activities (edges) and agents (attributes of edges) and describes the change of entities by activities. As entities in a provenance graph must be identified, *Appellations (signs)* are used to refer to and identify an entity within a certain context. Appellations do not identify things by their meaning, but through "pointing" to an entity by a technical or human agreement. An appellation is either an identifier, time appellation, or an agent appellation. Appellations are also known colloquially as "the name of a thing". A subclass of appellations are *Identifiers (codes)* assigned to entities in order to identify them uniquely and permanently within a specific (technical) context. In biobanking such codes are sample IDs, inventory numbers and registration codes. Identifiers are typically composed of alphanumeric sequences and are in most cases itself data objects defined in a specific namespace.

A special type of conceptual objects are *Metadata Objects*. Similar to appellations they describe other entities according to a specific convention, e.g. the Dublin Core schema for the basic description of data objects.

2.2 Data Quality

Data quality describes the degree to which a data object meets the expectations of data consumers (agents) based on their intended use, i.e. data quality can vary and can be difficult to measure, if the intended use is unknown. It can be distinguished

[1] In the course of this master thesis, we are looking at Whole Slide Images (WSI) and associated metadata as our main *data objects*.

between the following dimensions of data quality:

2.2.1 Congruence between Entities and Data Objects

The following attributes describe and measure the extend a data object, e.g. a scanned image or a temperature value, represents a real world (physical or conceptual) entity.

- *Accuracy* measures how similar a representation of an object is to the ground truth. Ground truth is always defined in a certain context by the acting agents, e.g. a group of scientists or a calibration device.
- *Currency* is defined by the time that is spend between an update of the conceptual entity and the data object (reaction time).
- *Completeness* and *Existence* define, how many of the parameters needed for the specific context are available to cover the intended use.
- *Reliability* defines if we can trust the agent who generated the data object.
- *Cost-effectiveness* is the cost of generation, storage and distribution of the data object reasonable to the intended usage.
- *Confidentiality* measures if data objects are only available to authorized persons for the intended use.
- *Granularity* and *Precision* define the level of detail captured in a representation, e.g number of significant digits to which a continuous value was measured for a continuous conceptual object as temperature, or for categorical variables, the resolution of the categories.

2.2.2 Congruence between Different Data Objects

The following attributes describe and measure the extend how several data objects representing the same (physical or conceptual) entity relate to each other [2].

[2] Future attributes can be found under `http://dimensionsofdataquality.com/alldimensions`

- *Consistency* measures if data objects are the same across systems or location of different storages. Data objects are consistent if objects representing the same conceptual object are not in conflict.

- *Volatility* measures information instability, the frequency of change of the value for an entity attribute.

2.2.3 Metadata Quality

A metadata object describes a data object within several dimensions, which can be grouped according to the FAIR data management principles[3]:

- *Findable Metadata Items* ensure that a data object has an appellation, so that it can be found either by humans or machines so it should be tagged with a persistent global unique identifier and a fixed linkage to the object.

- *Accessibility Metadata Items* ensure that the provenance of the data object is well documented and access information is well specified. Quality criteria cover usage of standards, access procedures, resolution of identifiers and availability of protocols.

- *Interoperability Metadata Items* ensure that the receiving agent (human or machine) has the necessary information to understand data objects at the syntactic level to be accessible (data syntax) and the semantic level to be understandable. Quality criteria cover the usage of standardised knowledge representations (ontologies), use of FAIR-compliant vocabularies and qualified references to other data.

- *Reusability Metadata Items* ensure that data access procedures are well documented, in order to allow the reuse of data objects. Quality criteria cover the existence of (machine readable) reuse licence(s), existence of provenance information and compliance with community standards.

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

2.3 Relevant Standards

In the following section relevant standards are described. The focus is on standards describing biological samples in a biobank (transport, storage, processing) and related metadata for the description of the sample provenance. Standards, which describe the medical history of the donor (openEHR, HL7, FHIR, etc) and/or phenotipical aspects and observational data (OMOP/OHDSI, PCORNet CDM, etc) and medical ontologies (ICD10, SNOMED, UBERON, HPO, etc) are not covered, as this would extend to scope of this thesis.

Logical Observation Identifiers Names and Codes (LOINC) Is the standard for identifying health measurements, observations, and documents.[4]

Unified Medical Language System (UMLS) Defines an ontology which unifies key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.[5]

ISO 20387 - General requirements for biobanking Containing the requirements to provide biological material and meta-data for research. Covering also the complete life cycle of biological materials and their meta data, see (11).

CEN/TC 140 - In vitro diagnostic and medical devices is a series of standards defining molecular in vitro diagnostic examinations for different samples, including the whole pre/examination process that is divided into two subgroups: "Outside the laboratory" facing the collection of the specimen (i.e. information about the patient, methods used,...) and the transport requirements (i.e. temperatures, humidity, time). And in the part "Inside the laboratory" including storage (ambient conditions, access,...), processing (fixation, cutting,...) and evaluation (devices used, selection of samples,...)

Standard PREanalytical Code (SPREC) is providing guidelines for documentation of pre/analytical processes as SPREC code, one for fluid specimen and

[4] https://loinc.org

[5] https://www.nlm.nih.gov/research/umls/

one for solid tissues[6]. The fluid specimen SPREC code consists of 7 elements: type of sample, type of primary container, pre/centrifugation, centrifugation, second centrifugation, post-centrifugation and storage condition. For solid tissues or tissue-derived cytological biospecimens the code elements are: type of sample, type of collection, warm ischemia time, cold ischemia time, fixation type, fixation time and storage condition.

The group also provides in addition to the SPREC definition the SPRECalc tool[7] allowing automatic SPREC generation. There is also a proposal for generating barcodes based on SPREC called SPRECWare (13; 14) and they have defined an interface to the Biospecimen Reporting for Improved Study Quality (BRISQ), see (15).

Minimum Information About Biobank data Sharing (MIABIS) was introduced by the BioMolecular Resources Research Infrastructure of Sweden in the year 2012 with the aim of providing a common biobank terminology to make samples and collections searchable (16). In 2016, an updated version the MIABIS 2.0 Core was released, defining basic attributes to describe a biobank, sample collection and study (17). An extension to define attributes for donors and samples and the QMS of a biobank is currently under development.

Minimum Information about Slides and Scans (MISS) is a small metadata set across different standardization activities and initiates a community driven approach to collect and harmonize existing ontologies to describe the provenance of a whole slide image. MISS was defined within the use cases of a large scale digitization effort for machine learning. Through several cycles with stakeholders from biobanking and machine learning a first draft was generated, which was proposed to the digital pathology community by the author, see (2). The minimal information about glass slides and their scanned representation is divided into three parts: (1) Pre Scanning (Slide) Metadata: e.g. metadata from biobanks, glass slide labeling, cleaning; (2) Scanning Metadata: e.g. technical parameters, resolutions and focus points; (3) Post-Scanning (File) Metadata: e.g. image quality indicators. A first version of MISS and exam-

[6] `hhttp://www.isber.org/?page=SPREC`

[7] `https://isber.site-ym.com/resource/resmgr/SPRECalc/SPRECalc.zip`

ples can be found in the MISS wiki page[8].

2.4 Whole Slide Image Formats

Due to the high demand on resources (very large gigapixel images, color space resolution) and specific medical metadata fields, general purpose image file-formats such as JPEG or PNG are not usable to store scanned slides. Therefore, several vendors implemented their own file-format, some based on the well known TIFF format with vendor specific extensions, others with completely new approaches. In all of vendor specific formats the image is stored as a pyramid together with other pictures like the label image. Vendor specific formats have the big drawback because they can only be accessed and processed with a proprietary software of one vendor. If a company gives access to the file-structure, third party solutions may be developed (e.g. the open-slide library[9]), however these often do not cover the full functionality (color correction, metadata fields). In addition to general purpose BIG TIFF format, the vendor-neutral whole slide image format was developed by the DICOM group[10]. At the time of writing (late 2020) there is still not a full implementation of the DICOM format across scanner vendors.

2.4.1 Vendor Specific Formats

Aperio (.svs) The base of the file-format from Aperio is TIFF format. This TIFF format was extended with an extra header containing information (Metadata) about the scanner (Serial Number, Magnification, Timestamp, Barcode,...). A .svs file can contain images with a different compression. The standard setting for the scanner is JPEG with a compression of 80. In a chain of single images the label image, thumbnail and the actual scan are stored, the latter in different pyramid layers, depending on the scan-resolution.

MIRAX (.mrxs) In comparison to most other file formats, the MIRAX file format is a mutli-file format (containing out of a .mrxs file, and a folder with serveral

[8] https://github.com/human-centered-ai-lab/MISS/wiki

[9] https://openslide.org/

[10] http://dicom.nema.org/Dicom/DICOMWSI/

.dat files and a slidedat.ini). The .mrxs file is the entry point for the user, consisting out of a simple JPG-image (low resolution layer of the scan). The viewer then opens the slidedat.ini (found in the folder with the same name as the .mrxs), in this file all the metadata is stored and also the references and indexes of the .dat files. For each layer of the pyramid, label and previewimage a new .dat file is created (for big layers multibe files), containing a data-stream of jpg tiles. taken by the scanner. The images taken by the camera were stored one to one without preprocessing. That means that the viewer has work with reconstructing and enhancing, making it difficult to implement open-source viewers.

Philips (.tiff), Hamamatsu (.vms, .vmu, .ndpi), Leica (.scn), Trestle (.tif) Similar to the .svs file, all of them are based TIFF format. Philips, for example, has in the beginning of the file a XML-Header containing the metadata and also the encoded label and the preview-image.

Sakura (.svslide) Sakura has a database approach for the scans. All the information is stored in a SQL-Lite database including the metadata and the image tiles. Through different tables the indexes are picked and the image is reconstructed.

2.4.2 Vendor Neutral Formats

Beside the proprietary file formats, some vendor neutral formats have emerged allowing to build a setup that isn't locked to a specific vendor (on the scanner side and also on the viewer side). This also allows to use scanners from different vendors in one lab and to integrate them into a pathology information system.

Generic tiled TIFF The Generic tiled TIFF is the basic tiff file without future metadata, only containing the pyramid image. Other images such as the label or the preview image do not fit into the file and have to be stored separately.

OME-TIFF The Open Microscopy Environment Consortium[11] introduced the file format OME-TIFF. OME-TIFF is extending the generic tiled TIFF with the

[11] `https://docs.openmicroscopy.org`

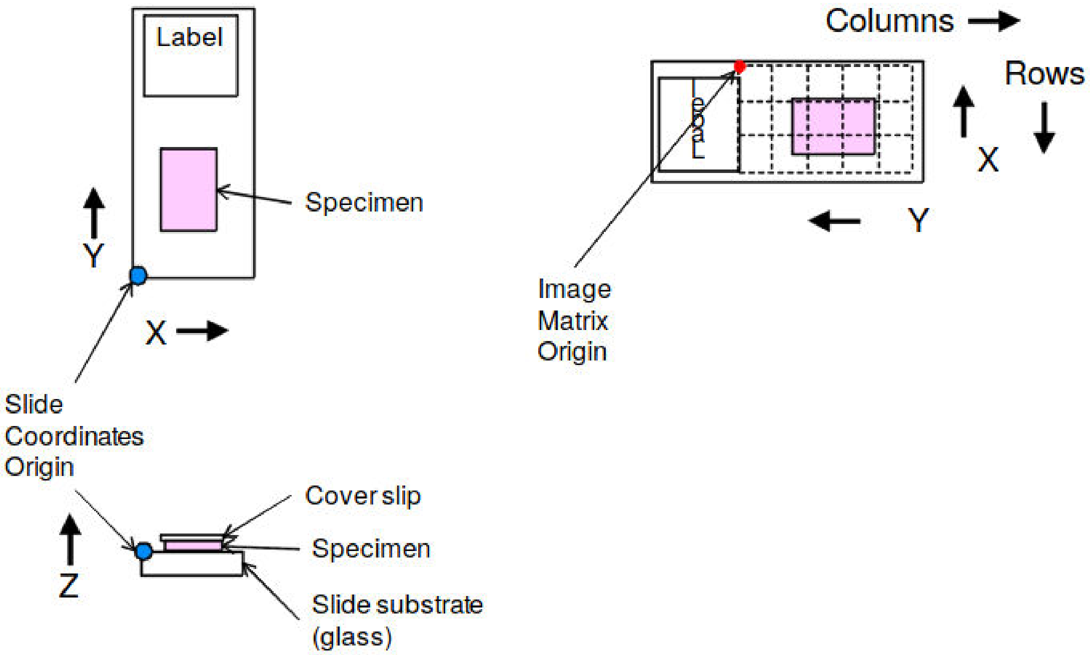


Figure 2.1: Slide Coordinates Origin and (X,Y,Z)(19)

OME-XML header allows storing metadata information about the scanner directly in the image-file. The flexibility of the OME-XML also allows encode images within the header (such as label and preview).

DICOM With the 'Supplement 145' of the DICOM standards a new standard for Whole Slide Images was created. (18) With the big goal to have one standardized format for all images within the clinical domain. This enables an integration into different viewers and allowing a combination of different domains (i.e. pathology and radiology). The 'Supplement 145' is not only defining how the metadata is organized, it also describes the coordinate systems of the scanner (see figure 2.1).

Unfortunately, the different scanner-vendors are not a big determinant factor for the propagation of an open standard for Whole Slide Images and open their viewers for other vendors.

2.5 Provenance

The exchange of samples, metadata and derived data (Whole Slide Images) has become a fundamental exigency in the development of AI algorithms and consequently interoperability and quality measures of data have become imperative. The generation of such data sets has undergone significant changes during recent years, evolving away from individual (small) scanning projects to transnational consortia covering a wide range of data formats, techniques and expertise.

To ensure the required data quality, there is an urgent need for standardized and comprehensive documentation of the whole workflow from the collection, generation, processing and analysis of the biological material (pre-scanning) the scanning process itself and to data analysis and statistics afterwards. Such a provenance information serves as a quality indicator and provides information on the reliability thus enabling transparency and comparability of derived results.

This demand was the starting point for the ISO TC 276, WG5 to define a standard for *Provenance information model for biological specimen and data.* The aim of this effort is to extend well-established approaches to provenance information management generally available in information technology (e.g., OPM1 or W3C PROV2) towards the requirements of the biotechnology domain.(9)

As biotechnology involves data generation from biological material, the provenance information needs to start with the source of the biological material, through its processing and all the steps of data generation and processing to final data analysis. High-throughput microscopy is developing scanning pipelines to extract biologically relevant information. Such automated system have a) to rely on a well documented pre-analytical quality of samples and b) support machine readable provenance documentation. With the help of provenance information of Whole Slide Images the following aims can be met:

- Continuous quality monitoring of samples and data entered in a scanning workflow.

- Retrospective analysis of results, pointing to the used Standard Operating Procedures (SOPs) and workflow parameters for a single scan.

- Retrospective analysis and quality control comparisons between scanning pipelines analysing a series of scans.

- Assessment of samples and data fitness for purpose in relation to the intended use. Provenance information can add valuable metadata to Whole Slide Images.

- Profiling of sample and data analysis pipelines in order to optimize scanning workflows and identify bottlenecks.

2.5.1 Provenance Standards

Dublin Core is the most used general purpose metadata standard to document resources, often applied in library purposes. DC has become endorsed by multiple standardization organizations as IETF RFC 5013 (20), ISO 15836-2009 (21), and NISO Standard Z39.85.

CIDOC/CRM is a reference ontology for the interchange of cultural heritage information. At first sight, this maybe a very different domain, but at a closer look there are also overlaps to provenance models in the biotechnology domain. Both approaches define an ontology for the exchange and integration of heterogeneous scientific documentation of "things" in scientific collections. According to the CIDOC/CRM ontology we have in a biobank a series of physical objects (e.g. scientist, subject, specimen and derivatives), conceptual objects (e.g. study, collection, diagnosis), digital born data objects (e.g. surveys, experiment outcomes) and digital samples (scanned slides). The CIDOC/CRM was developed by the ICOM's International Committee for Documentation and provides scientific collections with advice on good practice . The development started 1994 and is based on an object-oriented knowledge modeling approach. Since 2014, CIDOC/CRM has been the official standard ISO 21127:2014[12] (22).

An entity in the CIDOC/CRM[13] belongs either to the class Temporal Entity (E2) or to the class Persistent Item (E77). There is also a distinction

[12] https://www.iso.org/standard/34424.html

[13] http://www.cidoc-crm.org/Version/version-6.2

between Physical Things (E18) and Conceptual Objects (E28), where conceptual objects are being exclusively man-made. The CIDOC/CRM defines 168 properties between entities, e.g. identification, type assignment, occurrence at a specific event, usage of entities and ownership. The above example illustrates that the CIDOC/CRM goes much beyond a general purpose provenance model by defining domain semantics of "collecting and curating things".

OPM Open Provenance Model[14] is one of the early attempts to standardize provenance in the academic and computer science communities (23). OPM is based on the assumption that the provenance of an "object" is represented by an annotated causality graph expressing dependencies among things. The graphs, intended as records of past executions, are based on a set of syntactic rules and consist mainly of nodes (artifacts, processes and agents) and case-effects dependencies between sources and destinations.

HL7 FHIR Provenance is part of the FHIR Specification (v4.0.1)[15] and based on the W3C PROV model. The HL7 FHIR Provenance model is based on events triggered by HL7 FHIR and also allows links to objects that are not using HL7 FHIR. For security digital signatures can be added to verify the transactions taken within the model.

W3C PROV data model[16] consists out of 13 documents defining a model for provenance on the Web.

Historically, there were lots of different ontologies for provenance in use (including the Dublin Core Metadata Terms) and some of them were also an input of the process itself, e.g.:

- PREservation Metadata: Implementation Strategies (PREMIS)[17], a data dictionary focused on the preservation aspects of digital objects

[14] `http://openprovenance.org/`

[15] `https://www.hl7.org/fhir/provenance.html`

[16] `https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/` and `https://www.w3.org/standards/techs/provenance`

[17] `http://www.oclc.org/content/dam/research/activities/pmwg/premis-final.pdf` for version 1.0 (2005)
`http://www.loc.gov/standards/premis/v2/` for version 2.0 (2008–2012)
`http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf` for version 3.0 (2015)

- Provenir[18] (24), an ontology describing of e-Science applications
- Provenance Markup Language (PML)[19] (25)
- SWAN Provenance Ontology[20] (26) is a biomedical ontology including mechanisms to describe authorship and attribution lifecycles

All the above efforts on provenance have strongly influenced and led to the final W3C PROV recommendations. An in-depth comparison of provenance data models can be found on W3C Provenance Working Group Wiki[21]. The general idea of activities (processes) producing and consuming entities, is a common concept in many workflow engines, which is modelled by the PROV standard with a simple Entity–Agent–Activity model, see figure 2.2.

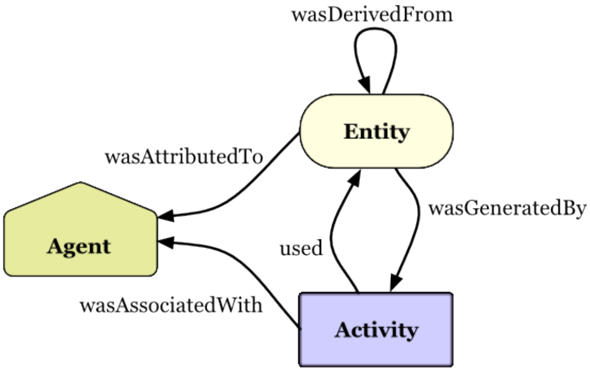


Figure 2.2: Key concepts of W3C PROV model. Image by W3C

A provenance graph can be used to visualize provenance information, as e.g. defined by the PROV standard family. Each node in the provenance graph is either an entity, an activity or an agent. A workflow is a data processing/analysis task represented by a directed graph detailing a sequence of operations that transform an input data set to an output. Each operations could be itself, another workflow or a "well defined" black-box. In this case, called nested provenance graphs.

The computational action within a workflow can be captured at a high level of detail including user interactions, quality inspections and other activities. In figure 2.3 an example provenance graph is shown, describing the provenance of a Whole

[18] http://wiki.knoesis.org/index.php/Provenir_Ontology

[19] https://tw.rpi.edu//portal/PML_Provenance_OWL_Ontology

[20] https://code.google.com/archive/p/swan-ontology/

[21] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings
https://www.w3.org/2011/prov/wiki/Interoperability

Slide Image starting at the sample acquisition (e.g. a surgery), the sample transport, sample pre-analytic, storage and finally, the scanning process.

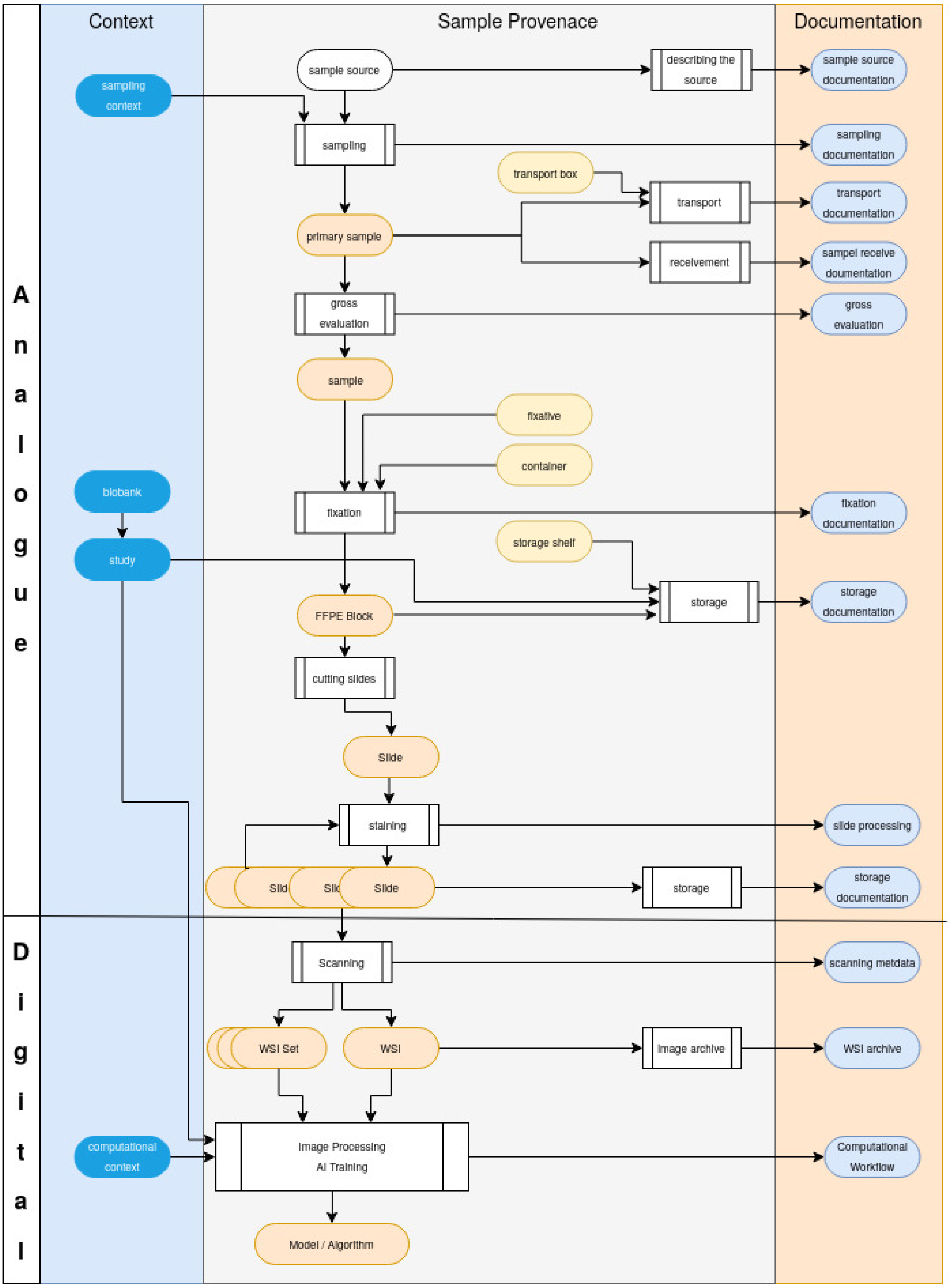


Figure 2.3: Example of a provenance graph covering the provenance of a (physical) biological sample, the scanning workflow, and the analysis of the Whole Slide Image

3. Methods

In this chapter, all steps of the high-throughput digitization workflow are described, see figure 3.1. They can be grouped into 1) case selection and data pre-processing, all processes in the first two rows of figure 3.1 2) scanning and scan quality control, third row in figure 3.1 and 3) post processing and cataloguing, the last two steps in the workflow.

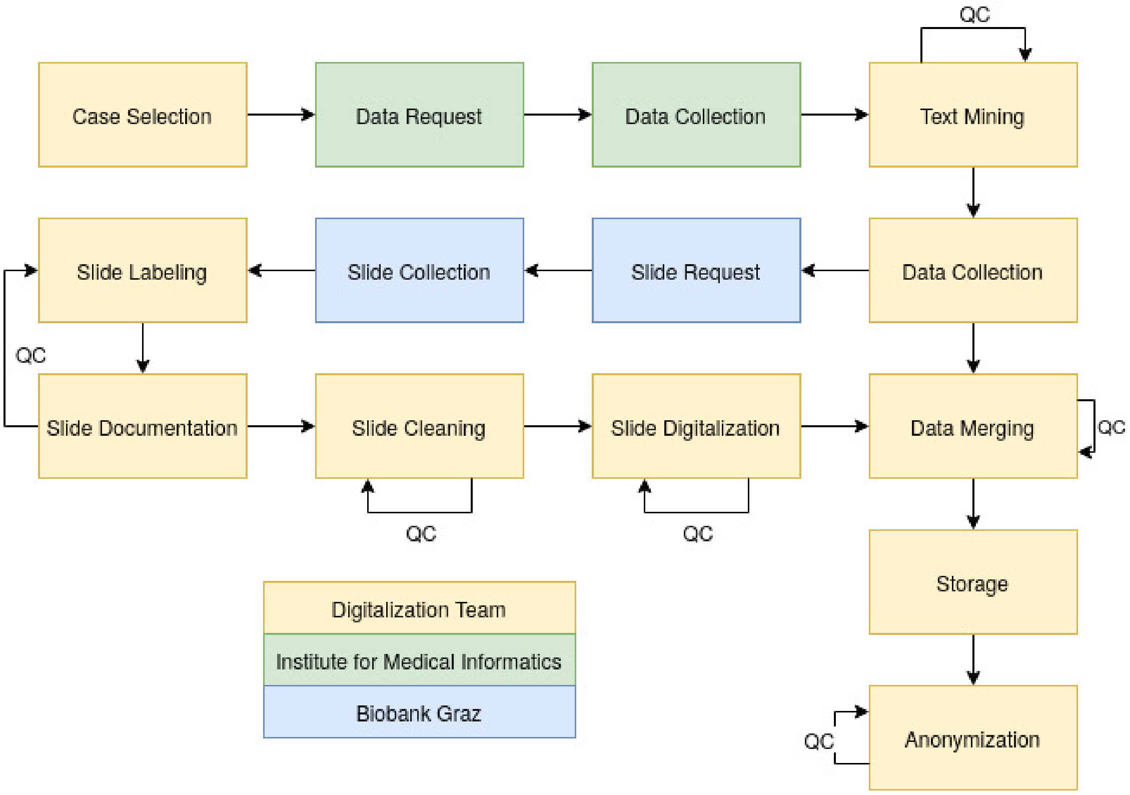


Figure 3.1: Overview Digitalization Workflow

3.1 Case Selection and Data Pre-processing

In the first step all medical cases and their related (physical) slides that should be scanned are selected. This is done according to the research question and the study protocol defined in the corresponding and approved ethics application. The following parameters are specified:

- Diagnoses inclusion and exclusion criteria
- Type of samples needed (surgery, biopsy....)
- Time period of initial diagnoses
- Donor information requested (gender, age, diagnoses, survival, . . .)

With the help of several databases, e.g. the *Pathology information system of the Medical University Graz* (PAS System), the *Hospital Information System* (Medocs) and the *Death Registry* of Statistic Austria the cases are selected and structured metadata descriptions (disease code, TNM staging, ICD-O classification) are extracted from their medical records, see (27). In this step for every case and all the related slides an unique research code is generated, which is later on used for pseudonymisation of the cases and tracking of the slides during the scanning workflow. All identifiable information is removed in this step, and the physical slides are "re-labeled" with this research code before shipping from the biobank to the scanning laboratory.

For the delivery of the glass slides, special drawers are used, were the glass slides are sorted by the year of examination and the research code, which replaces the old, in some cases even handwritten label in the glass slide. A typical (handwritten) label contains the following information:

- ***Histonumber*** Consecutively numbered restarting each year, one number per sample
- ***Year*** Year of the sample creation
- $\boldsymbol{Sample - Type}$ (Optional) Type of tissue

- ***Staining*** (Optional) Staining of the slide
- ***Position*** (Optional) Cut position on the sample

This label is replaced for the automated digitalization process by a new datamatrix (2D barcode) to ensure that each slide has a unique ID and can be identified by a machine (and human) during the scanning workflow. The datamatrix follows the structure: <PA-H YYYY ZZZZZZ \BB \SS> where:

- $\boldsymbol{PA-H}$ Stands for: Pathology Histology
- $\boldsymbol{YYYY}$ Year of the sample creation
- $\boldsymbol{ZZZZZZ}$ Research code
- $\boldsymbol{BB}$ Block-number
- $\boldsymbol{SS}$ Cut-number

Before the original label is replaced it is essential that a scan of the original label is archived. With this a quality control of the label transcription process is implemented. Furthermore, the archived original labels, together with the manual transcription of the label area may be used for training of AI-Algorithms to detect handwritten labels.

3.2 Scanning and Scan Quality Control

In this part of the workflow the glass slides received from the biobank are registered in the scanning laboratory and in the first step a low resolution "preview image" is taken. The preview images are used to document the original condition (before cleaning and marker[1] removal) and as a reference in the final quality control where the completeness of the scan is checked.

Before the scanning, the slides have to be cleaned as even a small part of dust and dirt can interrupt the automatic focusing process and have thus a major impact

[1] To speed up future diagnoses and reviews, the pathologist often marks the region of interest with a pencil. This marker is applied on the cover-slip, which creates major problems with the focus point setting during the scanning-process and has to be removed for the scanning

on the image quality. Most critical here are markers and dust on top of the cover-slip and air bubbles between the tissue and the cover-slip. The main artifacts on glass slides are fingerprints, glue, tilted cover-slips, markers and dust.

After the cleaning process is completed the slides are bulk loaded into the WSI-scanner. For the digitization of archive slides, usually a manual scanning routine is selected, were the focus points are set manually for each slide. In addition to the manually specification of the focus point (approx. 20 per slide), all other scanning attributes (magnification, file format, compression parameters) are set in an automatic way for all slides of a scanning job.

The resulting gigapixel images require a huge amount of data storage, one slide resulting in an average file size of 8GB, which is produced within 30 seconds. Storing this on a network storage within this short time is not possible. Therefore, a multi-tier storage solution was implemented. In the first step, the scans are stored on a local drive. In a highly efficient workflow, even a local spinning HDD(SATA/SAS) can not fulfill the requirements of the needed speed, so the integrated HDD are removed from the scanners and replaced by SSDs[2]. After the scan is completed, the file is transferred to the next tier (storage system), a local HDD RAID within 2 minutes, In this spinning disc setup, the scans are stored for up to two weeks, until all quality analysis and metadata generation is finished. Finally, the WSI are transferred from the local scan laboratory storage to a cheaper and more scale-able storage system (CEPH-Cluster).

3.3 Post Processing and Cataloguing

In the post processing step the data files are anonymized. Here it is checked if no identifiable information is present at the slide or label area and all the clinical parameters are generalized to achieve a k-anonymity (28). This process is done manually and is combined with the quality control of the Whole Slide Images. In the quality control, each slide is checked towards completeness, scanning artifacts as stitching errors, color fadings and out of focus areas.

For cataloging the Whole Slide Images a rich set of metadata attributes describ-

[2] Here enterprise sector SSD has to be used, as consumer sector SSDs have an average 0.5 Drive Writes Per Day (DWPD) resulting in a lifetime of only 5 month in continues scanning operation

ing the structure of the scanned tissue is necessary. Tissue areas on glass carrier plates may be organized as single structures, in symmetric objects or spread over the carrier in a complex way. Since there is a huge amount of possible combinations of areas that may or may not form objects and objects that may be very similar in size, shape and amount of areas they contain, there is a need of objective metrics that help to extract structural information and enhance the possibilities of methods for visualization. It is easy for humans to identify tissue areas that "look similar" and to describe and pinpoint symmetry properties even if areas of interest are altered by air bubbles or folded tissues. However, automatically extracting this information containing the exact number and the size of tissue areas, identifying groups of tissues that share certain properties and calculate metrics of similarities between single tissues and whole constellations of various areas is necessary in a high throughput scanning setup. The applied automatic method is illustrated in the next section.

3.3.1 Example Lymph Node Metastasis

In the detection of lymph node metastasis one block is sliced into 10 to 15 levels with a distance between the levels of 200μm each consisting out of 2 slides with each 2 sections. In the case of removed and stained breast lymph nodes three different slide categories were defined as shown in Figure 3.2 and Figure 3.3.

These specified sets of possibilities can occur once or several times per slide. For instance, there might be three small lymph node slices on one side of a slide and three lymph node slices of the next or previous level on the other side of the slide, so it can happen that one single case contains 20 up to 80 slides (40 to 160 single cuts) with a well defined hierarchical structure.

In order to detect and describe such a complex structure, the following computing steps are done: Extract and analyse areas from images, find clusters of tissues, find measures of similarities (between single areas and between groups of areas) and store all this information in an adequate data structure. Areas of interest (possible tissue areas) are extracted in the first step with methods described in (29). Metrics on out-of-focus errors, regions of air bubbles and removed objects are calculated and stored as quality control measures. More recent work discusses methods of calculating

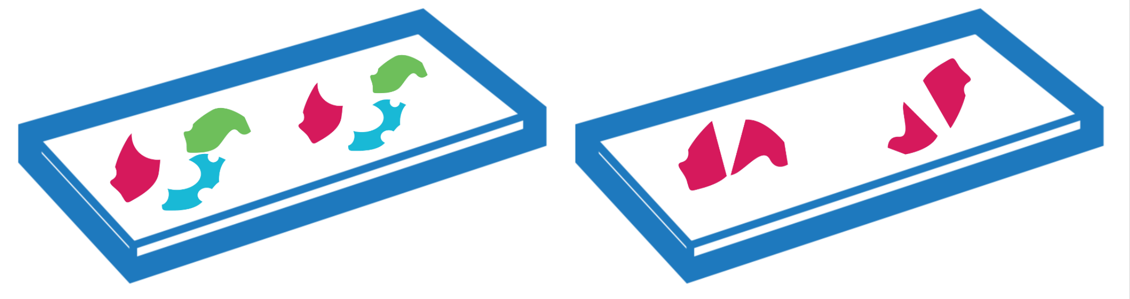


Figure 3.2: A (Left): One or more lymph nodes on one slide (usually: lymph nodes smaller, <5mm major dimension) B (Right): One lymph node split on one slide (usually: lymph nodes larger than case A, >10mm major dimension)

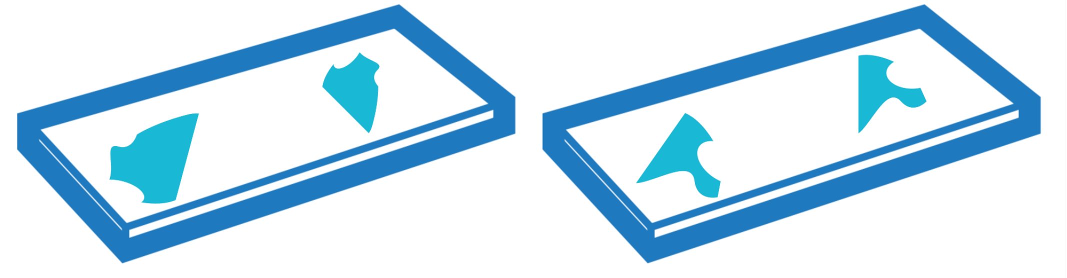


Figure 3.3: C: One lymph node split on two slides (usually: very large lymph nodes: >20mm major dimension)

tissue masks on preview images for automated scanning procedures (30). For the classification of leaf shapes, HU-Moments are used to train a support vector machine by (31). The creation of a binary mask, finding local clusters of tissue areas and calculating shape properties is done in the following steps.

- Create the binary mask parts with the HistoQC pipeline. First, the image is thresholded at the standard deviation over all channels. This is then transformed to a grayscale image keeping all values that are below a certain grayscale-threshold, resulting in a binary mask that represents areas of interest as well as other objects that passed this simple filter.

- Areas smaller than a certain value are removed and small are were filled so that fine connections between departments of several parts of objects could be kept intact to not adulterate the final image mask in terms of object quantity, size and other properties. In the end, all remaining areas are labelled: background area = 0, object areas = [1...N].

To characterize these labelled areas, the following set of different parameters are calculated according to (32) :

Perimeter using the boundary list: $[X_1...X_N]$

$$perimeter = \sum_{1=1}^{N-1} d_i = \sum_{1=1}^{N-1} |X_i - X_{i+1}| \tag{3.1}$$

Major Axis using the end points: Two pixels of boundary that are farthest away

Major axis angle:

$$angle = tan^{-1}(\frac{Y_2 - Y_1}{X_2 - X_1}) \tag{3.2}$$

Minor axis endpoints: Points farthest away on a line that is perpendicular to the major axis.

$$axislength = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{3.3}$$

Compactness :

$$compactness = \frac{4\pi * area}{perimeter^2} \tag{3.4}$$

which equals 1 for a circle or pi/4 for a square.

Roundness is s

$$roundness = \frac{4\pi * area}{convexperimeter^2} \tag{3.5}$$

Elongation using the bounding box (bbox) of the object with dimensions of major and minor axis for width and length:

$$elongation = \frac{width_{bbox}}{length_{bbox}} \tag{3.6}$$

Eccentricity:

$$eccentricity = \frac{minoraxislength}{majoraxislength} \tag{3.7}$$

Spatial moments were defined by:

$$m_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q \quad \text{for} \quad p, q = 0, 1, 2... \tag{3.8}$$

This yields the area for zero order were p = q = 0.Further, this formula results in centres of gravity, or first order moments:

$$centroid = [\bar{x}, \bar{y}] = [\frac{m_{10}}{m_{00}} \frac{m_{01}}{m_{00}}] \tag{3.9}$$

Central Moments:

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^p (y - \bar{y})^q \quad \text{for} \quad p + q > 1 \tag{3.10}$$

Normalising those central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad \text{with:} \quad \gamma = (p + q)/2 + 1 \tag{3.11}$$

From those normalized central moments, HU moments were calculated (scaling, translation and rotation invariant parameters) as described in Formula (3.12). These were introduced by (33).

$$\begin{aligned}
\phi_1 =& \eta_{20} + \eta_{02} \\
\phi_2 =& \eta(\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 =& (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \mu_{03})^2 \\
\phi_4 =& (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \mu_{03})^2 \\
\phi_5 =& (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
& + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30}) + \eta_{12}^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 =& (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21}) + \eta_{03})^2] \\
& + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 =& (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
& - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{3.12}$$

The k-Means algorithm was used on the calculated centroids of the areas. Since the amount of clusters on each Whole Slide Image was not known beforehand, several clusterings between 3 and C_{max} clusters were calculated (where C_{max} depends on images size and expected amount of clusters). If a slide only held one or two objects it wasn't looked for local groupings. In each run, k-Means tried to minimize a potential function using a number of cluster centers set in a specific smart way (34) (introduced by the authors as k-Means++) and given data points (area centroids in this case).

For evaluating the performance of each clustering process the Silhouette Score according to (35), shown in Formula (3.13) was used.

$$Sscore = mean(\sum_{n=1}^{N} \frac{b_n - a}{max(a,b)}) \tag{3.13}$$

In this formula "a" is the mean intra cluster distance and "b" the distance to the nearest cluster that the current sample is not part of. Values are generally between -1 and 1. Higher scores represent small clusters with a larger distance between each other (good separability) compared to smaller numbers.
After the clustering for each group, the minimum bounding box was calculated from the individual areas belonging to the group containing x and y - coordinates and rotational parameter. These boxes are the fundamental single objects of the sample space that are later used for visualisation.

A convex hull around the areas that belong to one group that yielded the highest silhouette score was used as a representative group area. Finally, those convex hull areas or representative group areas were compared calculating the euclidean distance of their HU-moments. The closer the distance between two areas in this seven dimensional space the higher is their similarity. As another independent measurement also the structural similiarties (SSIM) of the objects between groups, according to (36) were calculated. Since SSIM yield scores between -1 and 1 the result was shifted as follows:

$$SSIM_{new} = \frac{SSIM + 1}{2} \tag{3.14}$$

This transforms the interval [-1;1] to [0;1] where 0 reflects the lowest and 1 the highest similarity. Next, it was important to find normalized values for the distances in HU-Space in order to be able to combine them with SSIM results (see Formula 3.15).

$$HU_{similarity} = \frac{1}{7} \sum_{i=1}^{7} \frac{|min(H_1[i], H_2[i])|}{|max(H_1[i], H_2[i])|} \tag{3.15}$$

Here the areas 1 and 2 were compared. For every HU-Moment, the division of the smaller value and the greater value was calculated. The result for each division is 1 or smaller. The mean of those values was added to the SSIM score as shown in Formula 3.16:

$$SIMSCORE_{total} = \frac{SSIM_{new} + HU_{similarity}}{2} \tag{3.16}$$

This combination of the transformed SSIM and the normalized HU-Distances is used as final similarity score. It is calculated for every pair of objects in the sample space and the final similarity measures are stored in a NxN matrix.

Finally, the similarity matrix that was calculated in the previous subsection is evaluated. A specific similarity threshold is set to find the N-best matches of each object. Since the evaluation is done for every object, the resulting list for each virtual group may contain the same entries multiple times as shown in Table 3.1.

	A	B	C	D
A	1	0.97	0.15	0.98
B	0.97	1	0.2	0.95
C	0.15	0.2	1	0.2
D	0.98	0.95	0.2	1

Table 3.1: Different similarities for Objects A-D

Multiple entries of groups that already existed are ignored (in this case three equal groups existed for Members A, B and C) which would lead to the following list of virtual groups (compare Table 3.1 and Table 3.2):

Group ID		1
Member	A	1
Member	B	0.97
Member	D	0.98
Group ID		2
Member	C	1

Table 3.2: Objects sorted in groups

This representation is stored in a JSON file format with additional information such as: CollectionID, CollectionName and VirtualGroups that further contain: VgroupID and Members with: ObjectID, Similarity and Rotation. This information on relationships was then combined with information from the sample space as mentioned in previous section.

For the visualization of similar structures each group is cut out using the information of the bounding box stored in the sample space and rotated in order to make alignments possible. The angle from structural similarities is used for the rotation parameters. The basic translation as shown in (3.17) and the basic rotation shown in (3.18) are combined to a transformation-matrix T(x) (3.19).

Translation Matrix:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.17}$$

Rotation Matrix:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} cos(\theta) & -sin(\theta) & 0 \\ sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.18}$$

Combined Matrix T(x):

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = (\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(\theta) & -sin(\theta) & 0 \\ sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}) \begin{bmatrix} x \\ y \\ w \end{bmatrix} \tag{3.19}$$

4. Results

In this chapter the implementation and quality-control of the digitization process for archived glass-slides are described. In order to collect and digitalise archive slides in a high throughput way, a complete new workflow has been developed. An overview of this workflow is given in figure 3.1.

4.1 Web-service Implementation

This section shows how the workflow described above was implemented at the Medical University of Graz. Handling up to 1600 different slides per day and managing the data behind them requires a tool that is supporting the scan-team. Therefore, a Python Web-service was developed. The implemented Web-service has the following main tasks:

- ***Data − Handling*** With approximately 500.000 slides from 40.000 different patients per year an efficient data management behind the workflow is one of the key elements.

- ***Data − Movement and Data − Storage*** Another important task of the web-service is the movement and the control of the files generated by the scanners. This part has to ensure an efficient way to access the scans and the files. Each year, 12.000.000 new files are generated with a total size of approximate +5PB needed storage volume per year.

- ***Anonymization*** The web-service is also in charge of the anonymization, the the key handling for the pseudonymization, and allows data access according to the requirements of the General Data Protection Regulation (GDPR)

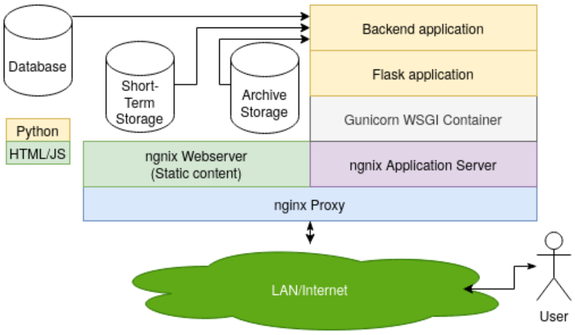


Figure 4.1: Overview Web-service Setup

- ***Data − Hosting*** To provide an human accessible view on the data stored and managed by the web-service a slide-viewer and a case-viewer was added to the web-service. Also a fully automated upload of the anonymized cases to the project partners is integrated into the system.

- ***Qualitycontrol*** Also a part of the web-service is a semi-automated quality control of the scanned images, and consistency-checks of the imported data.

4.1.1 Web-service Setup

As mentioned above the software implemented is a python web-service with several different background tasks for handling the data. The GUI is implemented in Javascript and HTML, this static content is handeled by a nginx-web-server, the dynamic content of the web-service is provided by the WSGI-Container Gunicorn that is running a Flask application see Figure 4.1.

The background tasks (that are running as thread on top of the Flask application) are managing the movement and the handling of the items (scans, slides,...) without user interaction.

- ***Import − Thread*** Each scanner has a specific folder to store the finished scans. This "landing zones" are managed and observed by the "Import Thread". If a new file is added by one of the scanners and the number of maximum simultaneous imports is not exceeded, a creation of a new "Import Inbox Thread" is triggered. This allows a coordinated pick-up process from all the different scanners and avoids an overload of the system.

- ***Import − Inbox − Thread*** As mentioned above, this thread is importing a single slide from the scanner into the Long-Term Storage. In a first step a consistency check of the slide is done to avoid broken files to be imported After that the SHA1 checksum is calculated and the scan is moved to the Storage. In a future step, the label-image and a small layer of the pyramid is extracted to allow a fast access to that information, without the need of handling the whole scan-file and to extract the barcode. After that, the SHA1 checksum is calculated again to detect possible file corruptions during the transfer and finally, the scan is added to the database.

- ***Merge − Data − Thread*** In the thread all the scans imported by the "Import Inbox Thread" are linked with the data that is already present of the scan in the database, this is for example the slide, project, case, patient, preview-image and the other scans of the same case. After this step, all the available data is linked together.

- ***Export − Thread*** On the other side of the import threads are the export threads. The "Export Thread" is in charge of handling and queuing all exports from the database. This follows a similar setup like the "Import Thread" and is checking constantly the maximum number of uploads, what data is needed to be exported and if the meta-data needed is available in the database.

- ***Export − Item − Thread*** After a new "Export Item Thread" is triggered by the "Export Thread" the scan is transferred to the export storage. After checking the SHA1 hash again, the anonymization of the scan is done (Removal of the label, and barcode deletion if present). Then all the needed meta-data is collected and converted into a pre-defined structure (see section "External Data Access"). Now, the data can be automatically stored on a FTP-Server or uploaded to a S3 Bucket.

- ***Map − Thread*** This thread is in charge of extracting the information out of the preview images and to map the label of the slide with the tissue type, section order and the staining. This is also the first point where the physical glas-slide is added with an UUID to the database.

- ***Log − and − Status − Thread*** The "Log and Status Thread" is one of the key elements within the web-service.

The first task of this thread is to handle all the logs from the different thread within the web-service.

The second task it to provide the data for both status panels (Scanner Status and Web-Service Status).

4.1.2 Anonymization

All the data processed within the workflow are pseudonymized by the Institute for Medical Informatics, meaning that all the identifying data like names were already removed and replaced by pseudonyms. During the import of the data another second pseudonymization step is applied, for each existing pseudonym an unique ID (UUID) is given.

In the exporting step all the UUIDs are changed again to a project specific UUID, this allows to keep track of the IDs in the scope of a project but it does not allow to link different projects with each other, also the other data fields are reduced to reach a k-anonymity of three as described in (28). In a final step, the link between the internal and the external UUID is deleted.

4.1.3 Data Access

The access to the processed data is granted in four gradations, all with different privileges and different levels of information details.

Researcher Access To provide access to all the researches of the Medical University of Graz an open research access has been implemented. By entering the patient identifier from the pathology information system the researchers can have a view on the scan and the corresponding metadata and preview image. This allows to reduce the number of slides that must be picked by the Biobank for other projects, and also allows to browse across the scanned slides .

Team Access In difference to access for researchers, the scan-team members have the possibility to change the status of the scans, and to do quality checks. The patient identifier is not accessible by the scan-team.

Admin Access The admin access allows full access to the database, within this also the import and export module is implemented, allowing to define what kind of information can be seen by whom.

External Data Access The external data access was implemented by exporting all the data according to MISS in an anonymized way. All the external users have no direct access to the system, instead the metadata is exported as a JSON-File, the labels of the scan are removed and uploaded to a cloud all together (or stored on a provided device).

4.1.4 Database Design

The data hosted by the web-service is coming from different resources. A centralized database is collecting them all together. As mentioned above the entries in the database only contain pseudonymous data, all the direct identifiers are replaced during the import with a virtual ID, the mapping back to the original value is stored outside the database so that it can't be accessed on a potential attack. These virtual IDs are containing a prefix that is explaining the type of the ID in example 'MUGGRZ-PATIENT' for a Patient or 'MUGGRZ-CASE' for a case, followed by an 128-Bit Universally Unique Identifier (UUID). This enables a unique ID within the whole name-space.

Project table and related map-tables

On top of the database structure the project table is containing the basic information about all the projects handled by the web-service. All the cases, patients, scans, slides, storages, inboxes have referees to these tables allowing a direct assignment of each item to one or more of the projects.
This results in the following map-tables:

- ***Project − Storage − Map*** N to N link between the Project Table and the Storage Table, which enables to assign a dedicated storage system to a project, for Machine-Learning projects for example a fast SSD storage, and for archive projects a Tape Storage can be assigned.

- $\boldsymbol{Project - Case - Map}$ N to N link between the cases and the different projects, with an implicit mapping of the project to the patients through the Case-Patient-Map. With this table it is secured that a project has only access to the assigned cases and patients.

- $\boldsymbol{Project - Scan - Map}$ N to N link between the Scan-Table and the Project Table, this assigns a specific scan of a slide to one ore more projects. Implicitly, this is also containing the N to N Slide-Project Map.

- $\boldsymbol{Project - Workpackage - Map}$ N to 1 link between the work-packages and the project. Each project is divided into different work-packages, each of these containing an assigned 'inbox'. This assigned inbox allows a dropping stage for the created scans and enables a mapping to a distinct project.

Case Table, Patient Table and related map-tables

The case table is containing all the case-related information of a patient. Besides its UUID of the case, it contains the age of the patient, when the case took place, a free text for the diagnoses (extracted from the PAS-System of the University), the examination type (if it is an operation sample or a biopsy, the date of the examination, the TNM-Classification, the Dukes-Classification and the staging.

- $\boldsymbol{ICD10 - Case - Map}$ N to N map link between a case and the extracted ICD10 Code table. One case can have multiple ICD10 codes and vice versa.

- $\boldsymbol{ICDO - Case - Map}$ N to N map link between a case and the extracted ICDO Code table. One case can have multiple ICDO codes and vice versa.

- $\boldsymbol{Stagingreference - Map}$ 1 to N map link between a case and the stating reference. As described above, the rules for the Tumor-Staging are changing over the years. The stage is based on the TNM-Staging with this map it is possible to have a common definition of the stage for all years despite the long time period included into the different projects.

The patient table is collecting all the information belonging to the patient, starting with the birth date, gender, survival status (if alive or dead), cutoff day (cutoff date from the death registry), ICDN (if available) to track the cause of death clinical or ICDE (if available) for the cause of death external with the associated date of death.

Scanner Table

The scanner table is storing the status of each scanner containing information about the fill-status, the current job the scanner is doing, and an error log to forward faults and down-times to the web-service and for the scanner-status-dashboard. For each scanner, the racks loaded into the scanner are also tracked, this enables a tracking of the slides and if a quality issue occurs, also the neighbouring scans can be especially checked for errors.

Slide Tables

An important differentiation was done with splitting up slides and scans (see below) into two different tables. In the Slide Table all information according to the physical glass slide is stored. This is containing the link to the Case-Table, to the label transcribed handwritten label of the slide (containing information like the tissue type, region, staining), to the UUID of the slide and to the two storage paths of the 'Preview Images' (one with barcode and one without barcode) taken by the 'Preview Station'

- $\boldsymbol{Slide - Type - Map}$ 1 to N map, link between the slide and the associated slide type. Slide type is the information extracted from the handwritten label. Due to the historical changes over the years and the amount of disease specific labels this was implemented as 1 to N and not N to N to ensure that the right label is chosen.

Scan and Scan-Quality Table

In this table, all the scan related information is stored and additional meta-data to the actual scan-file. This is including the file type, scanner ID (UUID of the scanner), position and rack the slide was in, timestamp when the slide was created and the magnification of the scan, also file-based parameters such as the storage path, the file size, the MD5 Hash-Value and last but not least the quality status.

In case of the quality status is 'FALSE', a link into the $\boldsymbol{Quality - Scan - Map}$ is added. This is a N to N map between a scan and a quality issue. These quality issues are subdivided as mentioned above into groups scan-based (i.e. scanning errors) and slide-based (i.e. air bubbles).

Log Tables

The Log Tables are containing all the information about the data access and the data provision. This allows a fully re-traceable overview about all data that is being accessed, changed, or viewed by the web-service itself and the users that have access to the web-service. Also the whole meta-data transferred to third parties is documented in the log files. This enables to have an overview over the data transmitted without breaking the anonymization. Logs and messages created by the web-service itself (Program logs) are not stored in a database, they a separated in log files.

4.2 Quality Control

Due to the age of the processed slides and the therefore corresponding error rates of the scanners, a quality control of the scanned images was unavoidable. Hence, a process of quality control was added to the workflow in the start-up phase (Phase 1) of the scanning operations, completely manually done by the staff, later in a semi-automated way (Phase 2). In the future, it is planned to implement Phase 3, an automated quality check with request for feedback to humans if obscurities occur.

Phase 1 - Manual Quality Check In the initial phase of the project each slide was checked by 2 employees. If they came up with a different result a third one checked the slide. This manual check was split into 2 parts. First, the check if all the regions with tissue were scanned by checking the Preview Image (see section Preview Station). If this check were positive the out of focus regions were checked. Therefore, the operators had to zoom in (to maximum magnification) at least five points of the scan and evaluate the sharpness, then the staff decided if the scan is good or a re-scan is needed.

Phase 2 - Semi-automated quality check and Phase 3 - Automated

quality check

In order to reduce the workload and also to have a more reproducible quality check a new method was introduced. The idea for an Automated Quality Check was born. With the fact that in-focus images have strong gradients and edges the plan was to apply a Gaussian Laplace filter on the scanned images (37), drawing a histogram and evaluating the distribution (high distribution -> out of focus; low distribution -> in focus). After trying this on several Whole-Slide-Images the results were very poor, also the inclusion of several pre-processing steps didn't improve the results. Through the creation process of the slide, each image has its 'natural' unsharpness caused by multiple layers of cells on one cut. This type of defocusing shouldn't cause a quality issue, because you also have this imprecise areas on the microscope. A filter like the Gaussian Laplace filter can not differentiate between this 'natural' blurriness and the scanner caused unsharpness. Caglar Senaras (38) introduced a new method for finding out of focus (OOF) regions in slides. With the help of deep learning they developed the DeepFocus algorithm done by training a five convolution layers neuronal network, resulting in a heat-map with a tile-size of 64x64 pixels. This algorithm (available on GitHub) performs much better the previous approach with the Gaussian Laplace filters.

Choping up a gigapixel image into small parts with 64x64 pixels will result in more than 2 millions tiles which the algorithm needs to analyze. On the GPU cluster of the scanning lab (4x Nvidia V100), the run-time of the algorithm for one slide is in average more than 7 minutes (without the time needed for loading the slide into the memory of the cluster). In the scanning lab in average there were 2.5 slides scanned per minute to run the algorithm in our lab. This means that the computing power needs to be increased by a factor of 17.5 to check all the images generated - an unfeasible extension for the lab (for cost, it-administrative and rack-space reasons). So a new method was introduced, the Semi-automated Quality Check. This software was put on top of the existing DeepFocus software. The idea behind it was instead of checking the 2 millions tiles with the algorithm only pick some random tiles and check them.

For a good coverage of the whole image the user can specify the amount of parts the WSI should be divided into. From each part one random selected tile is chosen for the Quality Check (as described in the Manual Quality Check as described above).

To avoid that the algorithm is picking empty space (height amount of white pixels) it creates a histogram of the pixel values. If it is empty, another random tile is chosen (after N-times (i.e 100) he cancels the search, because it could be that there is no tissue at all). The in-this-way-chosen tiles are now presented to the user in a grid with the highest magnification, a red circle indicates the results of the OOF-Algorithm, shown on the left slide in Figure 4.2. On the right side in Figure 4.2, an open-slide viewer with an overlay of the grid and the areas chosen by the circle is shown. The user now can zoom into the image and decide if the quality is good enough or a re-scan must be done. With this method it can be ensured that most areas of the image are checked and also it can be reproduced at what parts of the image the user has been looking.

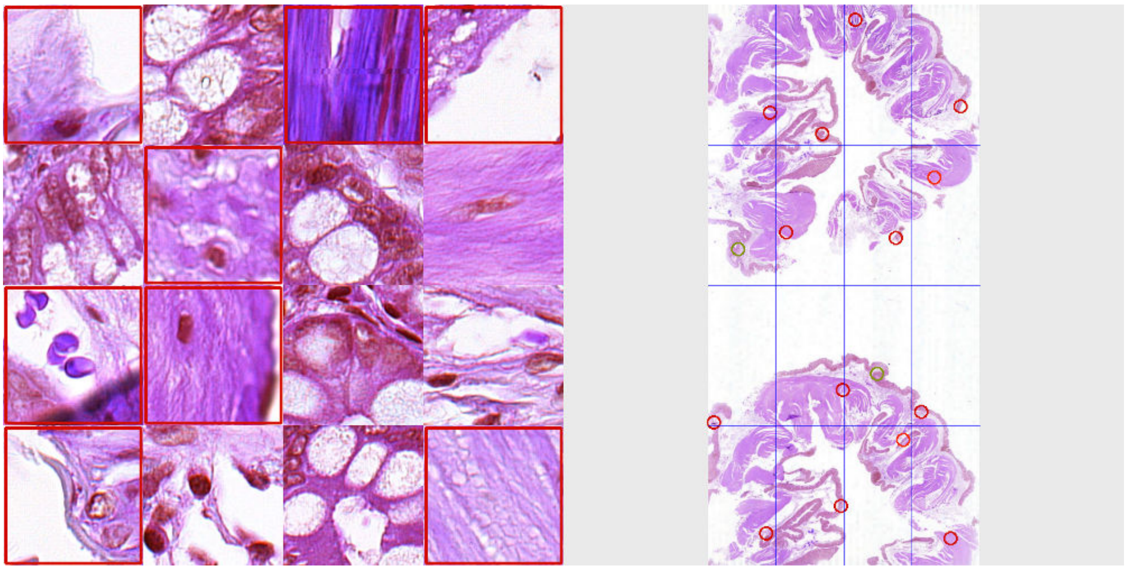


Figure 4.2: Phase 2 - Semi-Automated Quality Check

4.3 Tissue Feature Extraction Example Data

4.3.1 Analyse Areas and Find Clusters

First, the Whole Slide Image images were scaled and centered onto a white background, shown in Figure 4.3. Next, the binary mask was calculated, small holes were filled and small objects removed (see Figures 4.4 and 4.5).

Characteristic values were calculated as presented above and stored to be able

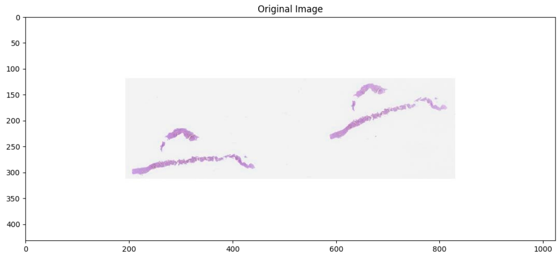


Figure 4.3: Original Whole Slide Image on white background

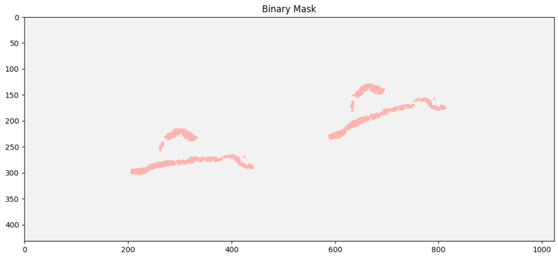


Figure 4.4: Binary mask of original image

to compare areas as shown in Table 4.1. These values serve as similarity measures between single areas and support the grouping done by the k-Means algorithm.

Parameter	Area1	Area2
Perimeter	150	145
Angle	7	78
Axislength	54	51
Compactness	0.63	0.67
Roundness	0.44	0.46
Eccentricity	0.33	0.31
HU-Moment1	0.87	0.86

Table 4.1: Similar parameters of 2 areas

After the different k-Means calculations, clusters that share similar locations are identified and evaluated using the silhouette score demonstrated in Figure 4.6. Centroids that belong to a certain cluster are marked with different colors.

After grouping areas and identifying objects, the characteristic values (HU-Moments) were calculated using the convex hull representation of those objects (see Figure 4.7).

Then, for these hull representations a minimum bounding box was calculated. ObjectID, SlideID, the coordinates of the bounding box and its rotation are stored

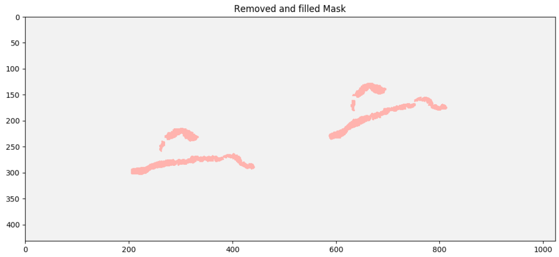


Figure 4.5: Binary mask with filled holes and removed small objects

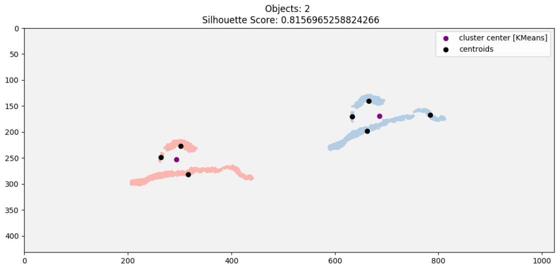


Figure 4.6: k-Means clusters and silhouette score for K=2 clusters

in the "Sample Space" as objects (see Figure 4.8).

4.3.2 Compare Groups

For every found object on all slides in a given sample space, SSIM and HU-Distances are combined (as shown above) and the final similarity score is stored in a NxN matrix (see Table 4.2).

	OB5751	OB12	OB96972	OB68775	OB4575
OB5751	1	0.375	0.157	0.482	0.947
OB12	0.375	1	0.951	0.991	0.288
OB96972	0.157	0.951	1	0.269	0.181
OB68775	0.482	0.991	0.269	1	0.334
OB4575	0.947	0.288	0.181	0.334	1

Table 4.2: Similarity matrix for 5 different objects

4.3.3 Similarity JSON

Each object holds its ID, the slide number it belongs to, the similarity score to the first object and the rotation in degrees. The sample space stores other information on

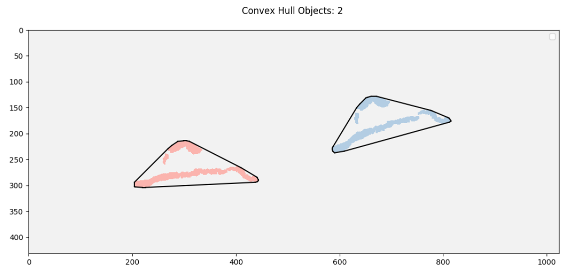


Figure 4.7: Hull representation of two objects

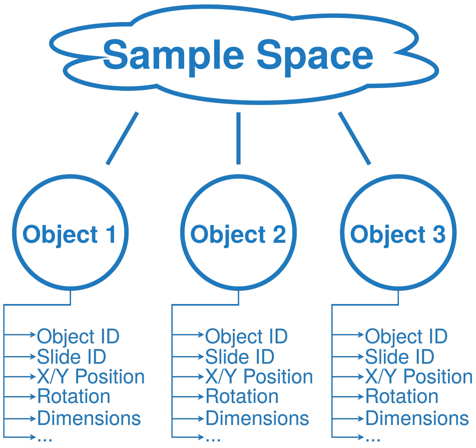


Figure 4.8: Sample Space and Objects

parameters of each object as mentioned above. These results were merged together in the Similarity JSON, see Listing 4.1.

```
{
  "CollectionID": "COL12754272",
  "CollectionName": "Sample Collection 1",
  "VirtualGroups": [
    {
      "VgroupID": "VG052754",
      "Members": [
        {
          "ObjectID": "OB5751",
          "Similarity": 1,
          "Rotation": 0
        },{
          "ObjectID": "OB68775",
          "Similarity": 0.974,
          "Rotation": 25
        }
      ]
    },{
      "VgroupID": "VG127",
      "Members": [
        {
          "ObjectID": "OB12",
          "Similarity": 1,
          "Rotation": 0
        },{
          "ObjectID": "OB96972",
          "Similarity": 0.951,
          "Rotation": 169
        },{
          "ObjectID": "OB4575",
          "Similarity": 0.991,
          "Rotation": 52
        }
      ]
    }
  ]
}
```

Listing 4.1: Similarity JSON

4.3.4 Visualised Similarities

With the Sample Space, the Similarity Json from the previous steps and the application of the transformation T(x) (3.19), now a new way of representing WSI is possible. In the example in Figure 4.9, the different detected objects on the slides are re-aligned and transformed to a virtual group. This enables a new way for the pathologist to view the scanned image, browse through the entire sample and survey whole collections.

4.4 Preview Station

The Preview Station is a small device developed for the "Slide Documentation" part of the workflow. One of the main needs for such a device was that the pathologists

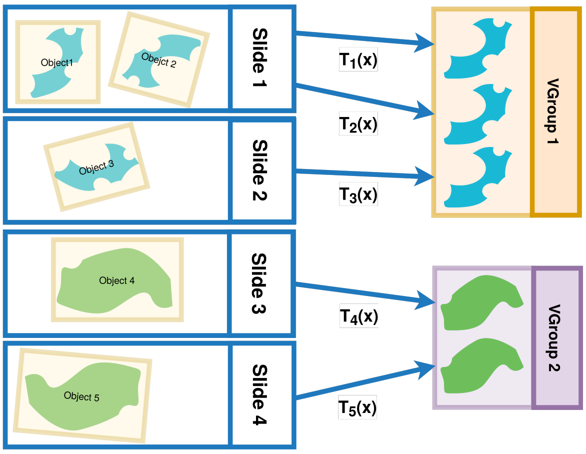


Figure 4.9: Transformation and Visualisation

often apply marker with a pencil on the glass-slide, with this the region of interest is marked. As part of the digitalization this marker is removed to get the best possible image quality of the scan.

The Preview Station concerns a device and a method for generating an image of a biological sample on a glass slide for generating an inventory in an image database. The device comprises a receiving unit configured for receiving the biological sample on the glass slide, a camera configured for generating an image of the biological sample while being received by the receiving unit, a releasing unit configured for automatically releasing the glass slide from the receiving unit after generating the image by the camera. The camera is configured to generate the image in a way that the image of the biological sample on the glass slide comprises an overall image of the glass slide and the biological sample on the glass slide, wherein the releasing unit is configured for applying a mechanical force onto the glass slide received by the receiving unit for releasing the glass slide from the receiving unit.

This invention was also put into a patient application (see patent application in the attachment).

4.4.1 Hardware-Implementation

In a first round experiments with a simple Smart-Phone-camera were made: The results of this were not very promising, because of the lack of setting the focus-distance and the brightness.

After this, a standard web-cam was used. This was not the best way to take pictures of the tissue, because fatty tissue is to bright to be seen with the normal ambient light.

So the first prototype of the Preview-Station was developed. This was a simple setup made out of a Raspberry Pi3b+ with an attached touch screen, and a "Logitech Brio" webcam. The parts were covered by a box made out of laser-chopped medium-density fibreboard. To solve the problem of the initial setup with the missing light, a LED-backlight covered by acrylic glass was integrated into the case. The Raspberry Pi had a small web-service running to control the monitor and the webcam.

Soon two new problems came up the way: The speed of Raspberry Pi3b+ was to slow to allow efficient work and the ambient light in the laboratory has caused reflections on the slide. As a fast workaround a box out of cardboard was put over the webcam-area and the Raspberry Pi3b+ was replaced by an external PC and a monitor. In addition to the issues above, the acrylic glass was destroyed by the sharp edges of the glass-slide. Resulting picture: see Figure 4.10, left picture.

During the further procedure a complete redesign of the setup was made. The result was a closed box with a new loading area for the slide (staging area) and with an external computer. Also an additional light source from the top was added, to highlight the handwritten label. To avoid the reflections of the top-LED in the tissue area, a separator was integrated into the box. Results: see Figure 4.10, middle picture. With a new staging area the problem with the sharp glass seems to be solved, but after 50.000 slides the staging area out of medium-density fibreboard was destroyed again. Concerning the image quality aspects the limits of the used webcam were reached. The main problem was the small distance between the object (slide) and the camera, webcams are usually designed for a minimum distance of 50cm.

Back to the drafting table the third and final big re-design phase started.This should solve the following problems:

- ***Wear of the stage*** To create a Preview Station that has a long lifetime a new concept for the stage is needed.

- ***Distance between camera and stage*** Due do the lens and the focus distance of the webcam it is not possible to get a sharp image.

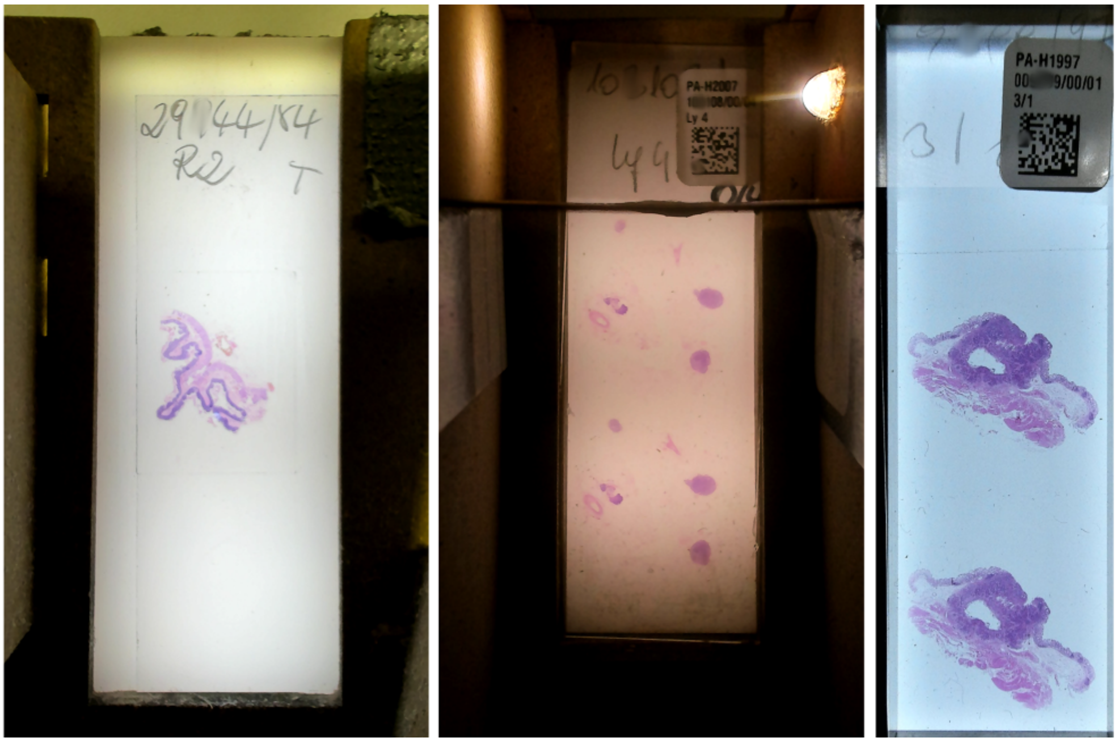


Figure 4.10: Preview Station Images

- ***Get rid of the external computer*** The initial idea of the Preview Station was to have a simple stand-alone device. Due the lack of the performance of the Raspberry PI3b+, it became necessary to have the external computer for the image processing.

- ***Allow different label sizes*** In the last version of the Preview Station a separator between the label area and the tissue area was added made out of cardboard. This creates the problem that slides with a different label size won't work with this setup.

- ***Speed*** To take a preview picture that needs in average 15s, the goal is to speed up this process to 10s. Thus, an average 1 hour work per day can be saved.

Camera

The camera of the preview station is the most important part, and was also one of the most challenging parts in the whole system. An usual webcam that was used in the first drafts of the device was not the best choice because it had the following problems: No fixed focus point; large minimal focus distance; requirement of bright ambient light; not adjustable settings like contrast, gamma,... but at least the chosen webcam had three big benefits: size, price and resolution. Finally, the E-con See3CAM-CU135 seemed to be the right USB-Cam for this device. With its 4k resolution, its adjustable fixed-focus lens and its internal flash buffer this cam meets all the requirements. After the additional adjusting of the timings of the driver of the USB-Cam it was also possible to use the full resolution of the webcam without overloading the Odroid XU4.

Housing

The initial versions of the Preview Stations were made out of medium-density fibreboard. This material was initially chosen because it is one of the best materials for laser-cutting. One of its biggest drawbacks is its porous nature, and that it can't be cleaned with wet tissues. To keep the simplicity of having a laser-cut case the decision was made to make the complete box out of a non transparent acrylic glass, with the drawback that the production tolerance of such a material is quite big (± 0.05mm). This material tolerance was the main reason for moving from a puzzle-part construction method to a screwed based fixture. In Figure 4.11 the final case of the Preview Station can be seen.

Light Sources

In the Preview Station two different types of light sources were added, both LEDs are controlled by PWM-Moudle of the device.

- $\boldsymbol{Bottom - Light}$ With the Bottom-Light the tissue section of the slide is highlighted. This light has to be as constant illuminated as possible, otherwise parts of the tissue get lost. In the previous prototypes simple LEDs with an acrylic glass on top for the dispersion were used. But as seen in Figure 4.10 there are dark parts between the LEDs. Experiments with reflectors under the staging area and also with an indirect light source were not really promising. An old Cell-Phone display brought really good results, so the idea was born to use back-lights for LCD displays (they also have the perfect size for this task). The back-light is usually an LED that is encased in acrylic, this is producing

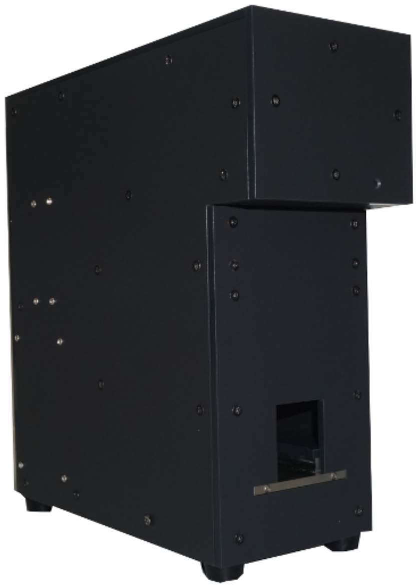


Figure 4.11: Preview Station

an even light, on the back side there is a foil reflector to maximise the light emitted to the front.

- ***Top – Light*** The Medical University of Graz has collect over the years several different types of slides with different labels (some of them are with a semi-transparent label area, some with non transparent area. To illuminate also the non transparent labels a top-light, made out of standard LEDs, was added; Expanded with a reflector made out of acrylic glass to avoid reflection in the label area.

Staging Area

In this version the old stage made out of medium-density fibreboard was replaced by a stage out of stainless steel. This ensures that there is no abrasion from the slide and also allows an easy glide of the slide. The experiences made with the first versions of the Preview Station showed that one of the most time consuming steps was the removal of the slide from the stage, caused by the small opening in the case, which can't be opened more because otherwise to much ambient light will

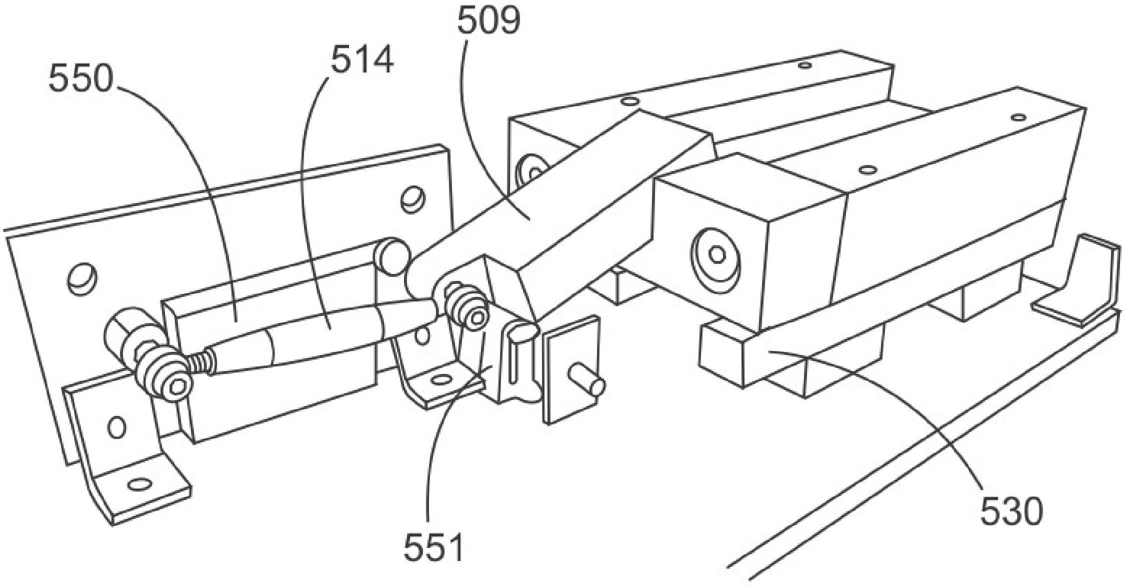


Figure 4.12: Stage of the Preview Station (excerpt of the pend. patent)

influence the picture. With the new material of the stage, it is now possible to add a mechanism for an automated slide remover. This was done with a pestle mounted over a turnbuckle with a step motor 4.12.

Electrical Components

Computer

As mentioned above the Raspberry PI3B+ has been removed from the last version because of the lacking power. On a research for other devices the Odroid XU4 came up, this single-board computer has some big benefits compared to the Raspberry:

- USB3.1 connector for higher webcam speed
- CPU with 8 cores
- Optimal alignment of the connectors to save space in the box

This single-board computer has Ubuntu 18.04 in an ARM-Version running that offers a great repository on libraries and also the big benefit of an operating system that can be handled by the average user.

$\boldsymbol{PWM - ModulandStep - Motor}$

Attached to Odroid XU4 over an I2C an external PWM-Modul is connected to the system. On two of the channels, the LED top and bottom light is connected. This allows a smooth brightness adjustment of the LEDs. On the third channel the step-motor is connected that controls the slide remover.

4.4.2 Software-Implementation

The task of the software is on the one hand to guide the user through the process of taking preview pictures of the slides and on the other hand to enhance the images, manage and sort them. The backend is based on a Python Flask webserver and the frontend is HTML and Java-Script.

Backend

In this part of the software the communication with the components takes place (Camera, Motors, LEDs) and also the processing of all the images is happening here. If the user takes a new picture a series of events are triggered and clocked:

- Turn on top LED
- Take picture (to capture the label area of the slide)
- Turn on bottom LED and the top LED off
- Take picture (to capture the tissue area of the slide)
- Eject the slide
- Turn off LED
- Merge the two pictures together and colour calibration
- Extract barcode
- Store and sort the image

The images are by default stored in an advanced folder structure that is handling the huge amount of different slides, also an CSV-Map for the location of the slides is built so that the images can be searched. The images can be stored on a local USB-Drive, the local disk or an attached NAS-Storage-an optional cloud sync allows to load the images directly in a cloud system (private could or if GDPR-confirm to a public hoster like Dropbox).

Frontend

The user-interface of the device allows the operators to interact with the device. This was implemented as a "hands-free" device for the main functions, allowing the user to trigger the camera by a feet-switch to have the hands available for handling and changing the slides. Only for changing the configuration of the device a keyboard and/or a mouse is needed. After the image is taken the user can verify the picture and he/she has the chance to do a retake if there was an error or a finger on the slide.

Through the configuration page of the Preview Station the settings of the Preview Station can be changed such as the colour correction, image naming, barcode parsing, cropping... Also the user has the chance to configure shot-cuts to often used settings and slide-namings, with the help of pre-defined buildings blocks.

4.5 Scanning Lab Infrastructure and Resources

In this chapter the main components of the Scanning Lab are described including the basic infrastructure for the slide cleaning as well as the IT-Components needed for the digitization. These components allow to increase the efficiency and the utilization of the scanners.

The whole lab is organized into 4 main parts:

- Cleaning Area: In this part of the lab the slide preparation is done.
- Slide-Storage: Storage Area for the slides waiting to be cleaned and for the scanned slides (before final quality check is done and the slides are returned to the BioBank).
- Scanning Area: Separated room for the slide-scanners. In this separated room the slide-scanners are positioned to keep the noise pollution for the employees as low as possible.
- Control Area: This is the controlling area for the scanners to select the region of interest to be scanned.

The infrastructure of the lab is designed to host up to 8 staff members at once, divided into 2 groups, one for cleaning and one region of interest selection.

4.5.1 Basic-Infrastructure

Cleaning Area:

Each employee that is in charge of the cleaning needs their own cleaning equipment, consisting of the following components:

- Scoring tool: Diamond dust loaded pen-like tool to remove the overhanging glass edges of the cover-slip.

- Scalpel/Craft knife (including spare blades): Needed to scratch of the markers of the glass slide. The back side can also be used to remove the overhanging cover slip and to remove crooked applied labels.

- Window scraper (and blades): Used to remove bigger parts of glue and used to distribute the cleaners over the slide

- Task light and magnifier: Magnification light to have a better view on the slide, also needed to over-light the ambient light (reflections from the room light)

- Automated soap dispenser: Needed for an automated dosage of the cleaners. This allows the employee to keep his hands free.

- Blade disposal container: Trashcan for the used blades of the scalpel and the window scraper.

- Cleaners (Glass cleaner and Ethanol 75%): Ethanol used to remove the glue and and glass cleaner used to remove the stripes.

- Sponge-daubers: Used to apply the cleaner on the slide.

- Microfiber cloth (fine and abrasive): Cloth to remove the glue and cleaner form the slide

- Adhesive tape: Used to glue broken slides.

- Slide-measuring tool: Needed to check if the slide fits into the scanner (Some slides are outside of the standard and can't be scanned, these slides will cause an error during the scanning process.

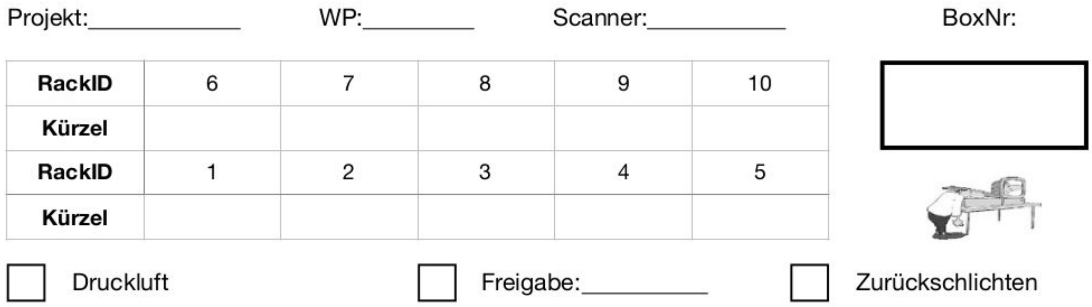
- Gloves: For protecting the staff from getting a) cut by the sharp edges and b) protecting the hand from the ethanol.
- Lab Coat: Used to avoid getting paraffin on the clothes of the employees.
- Oil-free Compressed Air: This is needed to remove the last dust particles of the glass slides. During the wipe with the microfiber cloth and the time between the cleaning and the loading into the scanner some dust particles remain/are added on the slide. These can be removed easily with compressed air (important is not to exceed a pressure of 2.5 bar, otherwise it can happen that some cover slips are loosened).

Racks and Transport boxes:

For an efficient workflow more additional racks than the standard amount (1 set) are needed. In the best case, 4 sets of racks are available, this allows that one set is in the scanner, one is in the cleaning step and one is in the unload phase. The fourth set allows to have an additional rack cleaned during the week to have one ready after the weekend. The amount needed is also depending on the magnification/speed of the scanner (the faster the scanner is, the more racks are needed for the employees to handle the demand). In addition to the racks, transport and storage boxes are needed to organize them; on the market there were no such rack available. The result were laser cut (done at the TU-Graz in the FAB-Lab), plastic inserts for the standard Euro-Boxes allowing a cost effective solution for the storage and transportation, besides this ending up with three European Design Patents on them.

Another important tool added to the workflow was the so-called "tracer". This sheet of paper keeps trace of the current state of slides that are momentarily in progress. Answering questions such as: Who cleaned what number of rack? To which project and work-package does the slide belong? Are the slides compressed air cleaned? Are they double checked? In what scanner were they? Are they ready to be sorted back? This process also allows, in cases of quality issue, going back to the person who cleaned the slide and doing a re-training to avoid future errors like this 4.13.

For a good utilisation of the scanners the coordination and planning of the staff is one of the key elements, this enables a seven days a week and 24 hours per day up



Projekt:__________ WP:________ Scanner:__________ BoxNr:

RackID	6	7	8	9	10
Kürzel					
RackID	1	2	3	4	5
Kürzel					

□ Druckluft □ Freigabe:__________ □ Zurückschlichten

Figure 4.13: Tracer - for tracking the status of the slides

time of the scanners. The scanning lab was man-powered by a total number of 14 student assistants with a weekly workload of 186 hours per week. During the project it turned out that the employment of student assistants offers a lot of benefits to both sides (the student side and the employer side).

- Flexible working hours: This offers the students to work beside the university up to 20 hours a week. The lab does not have fixed operation hours, so they were very free to organize the attendance on there own. Some guiding rules were given, like the maximum number of employees in the lab at once (4 people, except for an overlap of 15 minutes) and the attendance of at least one person in the lab during the day-time, to react to faults of the scanners.

- Long operation times: Due to the fact that the students have lectures in the morning and others on the evening, the time when operations were available in the lab was very long (usually starting from 6 in the morning till 6 in the evening.

- Breaks: The possibility for the students to make breaks and go visit lectures during the day.

- Networking: The social-networks of the students offered the great possibility to hire new crew-members very easy.

The flexible working hours were realized with a common team-calendar, in which the students have to assign their working hours for the next week until Wednesday of the current week.

5. Discussion and Future Work

Within the last two years more than 300.000 different slides have been scanned and processed within the workflow described above, resulting in 2PB of data and metadata. With the presented metadata and properties, similarity scores and data structure, a pipeline was provided to describe and organize Whole Slide Images.
A scanning lab with eleven WSI-Scanners (7x Leica Aperio AT2; 1x 3DHistech P1000; 1x DHistech P250; 1x Ventana DP 200; 1x Grundium Ocus) was setup during that time, providing a capacity of up to 3.000 archive slides (registration, labeling, cleaning, scanning, quality check) per day.

In comparison to the digitization of fresh cut slides, the scanning of archive slides comes up with great challenges, such as the unstructured metadata and the partly worse quality of the slides (caused by the age of the samples). The methods, described in more detail in chapter 3, show how to handle these problems and what solutions have been developed.

Whole slide scanning of archive tissue slides as data source for the development of AI algorithms will introduce further requirements. For future AI supported workflows in digital pathology, we need a clear understanding of how pathologists make diagnoses (39), and also a good insight into possible visualization methods (40). This is essential for the design and development of new human-AI interfaces, which enable pathologists to ask questions (41; 42) interactively via dialog systems (43). This would help to increase the quality of making diagnoses and is the ultimate goal of an AI assisted digital pathology pipeline.

The parameters and similarity scores need further investigation in terms of error rates and stability over a variety of different test images and data sets. While some parameters are interesting candidates for further classification tasks, future investigation is needed for the possibilities of 3D reconstructions of Whole Slide Images based on this work.

Besides the scanning and structuring of archive slides, the next step at the Medical University of Graz is to transfer the lessons learned within this project to a process for routine diagnoses. The main challenges are, in this context, the integration into the existing workflow and the highlighting of the benefits of such a transformation to justify the massive effort needed to do the transformation. The real benefit is not achieved by moving from the microscope to a screen, it is achieved by introducing new methods and AI-Algorithms in the daily work of the pathologists.

Transferring the daily routine workflow from its current analogue way to a completely digital workflow is coming along with a lot of advantages and disadvantages.

Telepathology allows easy access to get a second opinion from an expert (this can be in-house or at another institute) and small institutes are able to get expert knowledge without having them employed. The drawback of this is that this is creating a global competition on the histopathology market and second opinions are may interrupting the workflow of the pathologist. Telepathology will also reduce the amount of slides that are shipped around the world for diagnoses reducing the risk of damage and loss, with the disadvantage that no re-stains or molecular assays can be done.

Whole Slide Images have the big advantage that they are available without any delay on the pathologist desk, so physical slide distribution is needed, this will decrease the sign-out time. An integration into the laboratory information system helps to reduce the error rates. For the scans a reliable quality check is mandatory to ensure that no parts of the slides is lost during the scanning process and no artifacts are present. Several changes are needed in the slide-creation process, like the positioning of the tissue areas (most scanner are not able to scan the whole area of a slide). Air bubbles and tissue folds must also be reduced to a minimum to avoid out of focus areas. If only one focus plane is scanned critical parts can be lost.

Reporting, this new technology enables new ways a medical report can be generated. For example, the pathologist can add directly some pictures into the report, make annotations on the scans and do measurements. With this a diagnoses can be reproduced very fast. Also historical cases can be accessed immediately without the time delay for picking them out of the biobank again. The combination of different image sources (WSI, MRI, CT) allows a completely new view on the case, and will improve the quality.

Teaching, the education of pathologists is based on theoretical knowledge and practical skills. The latter is without the help of digital pathology achieved by a direct training from an expert. New teaching methods like described in (40) can help to increase quality of teaching and allow transferring explicit knowledge. Also remote classes and courses for experts are emerging with that technology.

Costs and efficiency, with the implementation of a digital workflow the work time on one case can be reduced (faster turnaround times), algorithms will help to reduce the amount of immunohistochemistry stains needed, and the slide handling staff can be reduced.The digitization costs for a single slide (including scanning, storage, viewer, staff,...) can be estimated at around 60-70 Cent. A huge initial investment for the infrastructure is needed (scanner, storage, clients and network) to implement digital pathology. An open question is if the pathologists still need a microscope on their desk or if it is possible to implement a 100 percent reliable digital workflow.

In order to fulfill the regulatory requirements the complete process must be tracked, including the pre-analytics, the digitization process, the AI-algorithms (if used) and the viewer. The scanning workflow must be combined in the future with devices from different vendors, allowing institutes to be independent of distinct vendors. Nowadays, for example, algorithms only get a FDA approval if the slides are stained with a certain antibody, scanned on a specified scanner and viewed in a specific viewer. This results currently in a closed system without the chance to implement a vendor neutral digital pathology. The big future challenge is to create an environment that allows access to pathology data in a structured way to exploit the full potential of computational pathology.

List of Figures

List of Tables

Glossary

CEN European Committee for Standardization (Comité Européen de Normalisation), `http://www.cen.eu/`. 85

CEN/TC Technical Committee of CEN. 28

CIDOC/CRM A reference ontology for the interchange of cultural heritage information developed by the International Council of Museums (ICOM)'s International Committee for Documentation. 34, 35

HL7 Health Level Seven International (HL7) organization focusing on developing data standards for electronic health information. 85

HL7 FHIR Fast Healthcare Interoperability Resources (FHIR, pronounced "Fire"), upcoming healthcare data exchange standard developed by HL7. `https://www.hl7.org/fhir/`. 35

ICOM Global organization of museums and museum professionals. 85

ISO International Organization for Standardization, `http://www.iso.org/`. 28, 34

LOINC A standard providing universal identifiers for laboratory and clinical measurements, observations, and documents, `https://loinc.org`. 28

MIABIS Minimum Information About Biobank data Sharing. 29

MISS Minimum Information about Slides and Scans. 29, 55

OPM Open Provenance Model. 35

SPREC Standard PREanalytical Code v. 2.0, `http://www.isber.org/?page=SPREC`. 28

UMLS An ontology which unifies key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. See `https://www.nlm.nih.gov/research/umls/`. 28

W3C World Wide Web Consortium. 86

W3C PROV Provenance information standard created by W3C. 35, 36, 81

References

[1] A. Holzinger, B. Malle, P. Kieseberg, P. M. Roth, H. Müller, R. Reihs, and K. Zatloukal, “Towards the augmented pathologist: Challenges of explainable-ai in digital pathology,” *CoRR*, vol. abs/1712.06657, 2017.

[2] M. Plass, R. Reihs, R. Merino Martinez, and H. Müller, “Miss - minimal information about slides and scans,” *European Congress on Digital Pathology*, 2020.

[3] M. Plass, P. Faulhammer, R. Reihs, A. Holzinger, K. Zatloukal, and H. Müller, “Reconstruct and visualise hierarchical relationships in whole slide images,” in *24rd International Conference Information Visualisation (IV)*, IEEE, 2020.

[4] S. Al-Janabi, A. Huisman, and P. J. Van Diest, “Digital pathology: current status and future perspectives,” *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

[5] H. Müller, N. Malservet, P. Quinlan, R. Reihs, M. Penicaud, A. Chami, K. Zatloukal, and G. Dagher, “From the evaluation of existing solutions to an all-inclusive package for biobanks,” *Health and technology*, vol. 7, no. 1, pp. 89–95, 2017.

[6] S. Gainotti, P. Torreri, C. M. Wang, R. Reihs, H. Müller, E. Heslop, M. Roos, D. M. Badowska, F. De Paulis, Y. Kodra, *et al.*, “The rd-connect registry & biobank finder: a tool for sharing aggregated data and metadata among rare disease researchers,” *European Journal of Human Genetics*, vol. 26, no. 5, pp. 631–643, 2018.

[7] T. Klingstrom, M. Mendy, D. Meunier, A. Berger, J. Reichel, A. Christoffels, H. Bendou, C. Swanepoel, L. Smit, C. Mckellar-Basset, E. Bongcam-Rudloff, J. Soderberg, R. Merino-Martinez, S. Amatya, A. Kihara, S. Kemp, R. Reihs,

and H. Müller, "Supporting the development of biobanks in low and medium income countries," in *2016 IST-Africa Week Conference*, IEEE, May 2016.

[8] H. Bendou, L. Sizani, T. Reid, C. Swanepoel, T. Ademuyiwa, R. Merino-Martinez, H. Müller, A. Abayomi, and A. Christoffels, "Baobab laboratory information management system: development of an open-source laboratory information management system for biobanking," *Biopreservation and biobanking*, vol. 15, no. 2, pp. 116–120, 2017.

[9] H. Müller, G. Dagher, M. Loibner, C. Stumptner, P. Kungl, and K. Zatloukal, "Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management," *Current Opinion in Biotechnology*, vol. 65, pp. 45–51, 2020.

[10] H. Müller, R. Reihs, K. Zatloukal, F. Jeanquartier, R. Merino-Martinez, D. van Enckevort, M. A. Swertz, and A. Holzinger, "State-of-the-art and future challenges in the integration of biobank catalogues," in *Smart Health*, pp. 261–273, Springer International Publishing, 2015.

[11] T. ISO, "Biotechnology — biobanking — general requirements for biobanking iso 20387: 2018," 2018.

[12] M. Doerr, F. Murano, and A. Felicetti, "Definition of the crmtex," 2017.

[13] U. Nanni, F. Betsou, S. Riondino, L. Rossetti, A. Spila, M. G. Valente, D. Della-Morte, R. Palmirotta, M. Roselli, P. Ferroni, *et al.*, "Sprecware: software tools for standard preanalytical code (sprec) labeling–effective exchange and search of stored biospecimens," *The International journal of biological markers*, vol. 27, no. 3, pp. 272–279, 2012.

[14] S. Riondino, "Sample preanalytical code for labeling of biospecimens: an analysis of specimen labeling protocols," *J Biorepository Sci Appl Med*, vol. 3, pp. 15–21, 2015.

[15] H. M. Moore, A. B. Kelly, S. D. Jewell, L. M. McShane, D. P. Clark, R. Greenspan, D. F. Hayes, P. Hainaut, P. Kim, E. Mansfield, *et al.*, "Biospecimen reporting for improved study quality (brisq)," *Journal of proteome research*, vol. 10, no. 8, pp. 3429–3438, 2011.

[16] L. Norlin, M. N. Fransson, M. Eriksson, R. Merino-Martinez, M. Anderberg, S. Kurtovic, and J.-E. Litton, "A minimum data set for sharing biobank samples, information, and data: Miabis," *Biopreservation and biobanking*, vol. 10, no. 4, pp. 343–348, 2012.

[17] R. Merino-Martinez, L. Norlin, D. van Enckevort, G. Anton, S. Schuffenhauer, K. Silander, L. Mook, P. Holub, R. Bild, M. Swertz, *et al.*, "Toward global biobank integration by implementation of the minimum information about biobank data sharing (miabis 2.0 core)," *Biopreservation and biobanking*, vol. 14, no. 4, pp. 298–306, 2016.

[18] R. Singh, L. Chubb, L. Pantanowitz, and A. Parwani, "Standardization in digital pathology: Supplement 145 of the dicom standards," *Journal of pathology informatics*, vol. 2, 2011.

[19] DICOM-Standard-Committee, "Working groups 26, pathology,"digital imaging and communications in medicine (dicom) supplement 145: Whole slide microscopic image iod and sop classes,"."

[20] J. Kunze and T. Baker, "The dublin core metadata element set: Rfc 5013," *California: IETF*, 2007.

[21] I. Iso, "15836: 2003-information and documentation-the dublin core metadata element set," 2003.

[22] T. ISO, "46. cidoc coneptual reference model (crm)—iso 21127: 2006," *International Standardizaton Organization (ISO)*, p. 44, 2006.

[23] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, *et al.*, "The open provenance model core specification (v1. 1)," *Future generation computer systems*, vol. 27, no. 6, pp. 743–756, 2011.

[24] S. S. Sahoo and A. P. Sheth, "Provenir ontology: Towards a framework for escience provenance management," 2009.

[25] P. P. Da Silva, D. L. McGuinness, and R. Fikes, "A proof markup language for semantic web services," *Information Systems*, vol. 31, no. 4-5, pp. 381–395, 2006.

[26] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, “The swan biomedical discourse ontology,” *Journal of biomedical informatics*, vol. 41, no. 5, pp. 739–751, 2008.

[27] R. Reihs, H. Müller, and K. Zatloukal, “Ontology-based text mining for large scale digital slide annotation.,” *Journal of Pathology Informatics*, vol. 10, 2019.

[28] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[29] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, “Histoqc: An open-source quality control tool for digital pathology slides,” *JCO Clinical Cancer Informatics*, vol. 3, pp. 1–7, 04 2019.

[30] M. Hosseini, D. Lee, D. Gershanik, D. Lee, and S. Damaskinos, “Whole slide preview image segmentation and setup for digital pathology scanners,” 02 2020.

[31] M. Lukic, E. Tuba, and M. Tuba, “Leaf recognition algorithm using support vector machine with hu moments and local binary patterns,” pp. 000485–000490, 01 2017.

[32] L. Kopanja, D. Zunic, B. Loncar, S. Gyergyek, and M. Tadic, “Quantifying shapes of nanoparticles using modified circularity and ellipticity measures,” *Measurement*, vol. 92, 06 2016.

[33] Ming-Kuei Hu, “Visual pattern recognition by moment invariants,” *IRE Transactions on Information Theory*, vol. 8, pp. 179–187, February 1962.

[34] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’07, (USA), p. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[35] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.

[36] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

[37] A. Huertas and G. Medioni, “Detection of intensity changes with subpixel accuracy using laplacian-gaussian masks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5, pp. 651–664, 1986.

[38] C. Senaras, M. K. K. Niazi, G. Lozanski, and M. N. Gurcan, “Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning,” *PloS one*, vol. 13, no. 10, p. e0205387, 2018.

[39] B. Pohn, M. Kargl, R. Reihs, A. Holzinger, K. Zatloukal, and H. Müller, “Towards a deeper understanding of how a pathologist makes a diagnosis: Visualization of the diagnostic process in histopathology,” in *IEEE Symposium on Computers and Communications (ISCC 2019)*, IEEE, 2019.

[40] B. Pohn, M.-C. Mayer, R. Reihs, A. Holzinger, K. Zatloukal, and H. Müller, “Visualization of histopathological decision making using a roadbook metaphor,” in *23rd International Conference Information Visualisation (IV)*, IEEE, 2019.

[41] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, 2019.

[42] A. Holzinger, A. Carrington, and H. Müller, “Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations,” *KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt*, vol. 34, no. 2, 2020.

[43] E. Merdivan, D. Singh, S. Hanke, and A. Holzinger, “Dialogue systems for intelligent human computer interactions,” *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 57–71, 2019.

A. Appendix

Reconstruct and Visualise Hierarchical Relationships in Whole Slide Images

Markus Plass
Institute of Pathology
Medical University Graz
Graz, Austria
markus.plass@medunigraz.at

Philipp Faulhammer
Institute of Pathology
Medical University Graz
Graz, Austria
philipp.faulhammer@medunigraz.at

Robert Reihs
Institute of Pathology
Medical University Graz
BBMRI-ERIC
Graz, Austria
robert.reihs@medunigraz.at

Andreas Holzinger
Institute for Medical Informatics,
Statistics and Documentation
Medical University Graz
Graz, Austria
andreas.holzinger@medunigraz.at

Kurt Zatloukal
Institute of Pathology
Medical University Graz
Graz, Austria
kurt.zatloukal@medunigraz.at

Heimo Müller
Institute of Pathology
Medical University Graz
BBMRI-ERIC
Graz, Austria
heimo.mueller@medunigraz.at

***Abstract*—Extracting hierarchical properties from Whole Slide Images automatically and expanding the possibilities of visualising data from digital pathology will not only drastically improve speed and accuracy of the pathologists' work but also simplify the necessary pre-processing steps for machine learning tool-chains. The introduced pipeline identifies and converts areas of interest into binary masks and finds groups of areas that share similar locations using k-Means. This grouping is evaluated by the Silhouette Score which serves as a measure of confidence for the separability of clusters. Found objects are compared using structural similarities and HU-Moments. These results are then stored as measures of similarities creating virtual groups of the most similar objects. Finally, the information on similarities combined with further structural parameters are visualised.**

I. Introduction

Extraction and reconstruction of hierarchical information and relationships of objects in images are common problems not only in medicine. The automated detection of certain objects and groups of objects is a highly specific task and depends on overall noise, background property, object sizes and shapes, colors and image qualities. With rising amounts of digital image data in the field of pathology, it is highly important to identify, label and structure those piles of information. Finding similarities between objects and groups of objects are intuitive tasks performed by human eyes and the brain but are far from trivial to be automated. Digital pathology generates data which is organised in different types and combined with lots of meta information [1]. Extracting, condensing and structuring these kinds of information are highly sensitive but crucial tasks and are mandatory parts to enable the possibilities of knowledge discovery processes in big data environments.

In general, tissue areas on glass carrier plates may be organized as single structures, in symmetric objects or spread over the carrier in a complex way. Since there is a huge amount of possible combinations of areas that may or may not form objects and objects that may be very similar in size, shape and amount of areas they contain, there is a need of objective metrics that help to extract structural information and enhance the possibilities of methods for visualization. It is easy for humans to identify tissue areas that "look similar" and to describe and pinpoint symmetry properties even if areas of interest are altered by air bubbles or folded tissues, for instance. Automatically extracting information containing the exact number and the size of tissue areas, identifying groups of tissues that share certain properties and calculate metrics of similarities between single tissues and whole constellations of various areas are problems that are not easy to solve.

II. Background

A. Slide Digitisation

During the last decade, pathology has benefited from the rapid progress of image digitizing technologies, which led to the development of scanners capable to produce so called Whole Slide Images (WSI) which can

be explored by a pathologist on a computer screen (virtual microscope) comparable to the conventional microscope and can be used for education and training, diagnostics (clinico-pathological meetings, consultations, revisions, slide panels and upfront clinical diagnostics) and archiving [2].
Compared to radiology, where the typical file sizes are in the range from 500 KB to 50 MB, a single WSI scan with 40x magnification consists of approximately 16 Gigapixels (Note: for the calculation of the WSI file size and comparison of different scanner manufacturers, we use the de-facto standard area of 15mm x 15mm, with an optical resolution of 0.12µm, which corresponds to a 40x magnification).

B. Whole Slide Images for Machine Learning and AI

Recent developments in high-throughput slide scanners offer the possibility for making the entire information contained in the millions of glass slides produced every year, available for machine learning applications. Access to whole slide images and related medical data will overcome the current limitations of accessing and sharing pathology material and will facilitate the development of new machine learning algorithms. In order to develop these algorithms, a large series of slides offering a broader coverage of tissues and cancer type / pathological deviations are required. Biobanks collect, preserve, and provide access to samples, e.g. from pathology in a transparent and quality controlled manner in compliance with ethical, legal, and regulatory requirements for research [3]. They require access to sufficient numbers of samples and data that properly cover the broad spectrum of disease subentities relevant for targeted therapies [4]. To address this demand, samples and data from different biobanks in different countries must be suitable for integrated analyses. This is only possible if samples and data meet common quality criteria. Therefore, international standards (e.g. CEN Technical Specifications or ISO Standards) were implemented for sample pre-analytics, covering all steps from the sample collection of the patient to isolation of bio molecules [5], and (open-source) software for cataloging and provenance management was developed, e.g. for rare diseases [6] and for biobanks in low and medium income countries [7], [8].

C. Example Lymph Node Metastasis

Concerning the detection of lymph node metastasis, for example, one block is sliced into 10 to 15 levels with a distance between the levels of 200µm each consisting out of 2 slides with each 2 sections.

In the case of removed and stained breast lymph nodes we define three different slide categories as shown in Figure 1 and Figure 2.

These specified sets of possibilities can occur once or several times per slide. For instance, there might be three small lymph node slices on one side of a slide and three lymph node slices of the next or previous level on the other side of the slide, so it can happen that one single case contains 20 up to 80 slides (40 to 160 single cuts) with a well defined hierarchical structure.

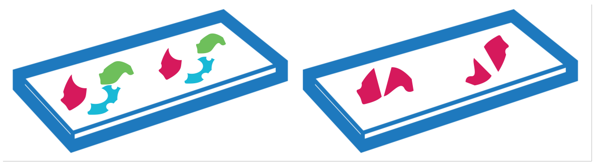


Fig. 1. A (Left): One or more lymph nodes on one slide (usually: lymph nodes smaller, <5mm major dimension) B (Right): One lymph node split on one slide (usually: lymph nodes larger than case A, >10mm major dimension)

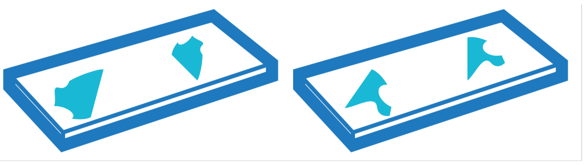


Fig. 2. C: One lymph node split on two slides (usually: very large lymph nodes: >20mm major dimension)

In order to handle these amounts of complexly generated visual data we need to consider the following problems: Extract and analyse areas from images, find clusters of tissues, find measures of similarities (between single areas and between groups of areas) and store all this information in an adequate data structure. Previous work on the extraction of tissue areas from WSI was done by [9]. Areas of interest (possible tissue areas) are extracted step by step by their "HistoQC" - pipeline. Metrics on out-of-focus errors, regions of air bubbles and removed objects are calculated and stored as quality control measures. More recent work discusses methods of calculating tissue masks on preview images for automated scanning procedures [10]. For the classification of leaf shapes, HU-Moments are used to train a support vector machine by [11]. The creation of a binary mask, finding local clusters of tissue areas and calculating shape properties will be discussed in the following section.

III. Methods

A. Analyse Areas and Find Clusters

1) Binary Mask: To create the binary mask we used parts of the HistoQC pipeline: First, the image is thresholded at the standard deviation over all channels. This is then transformed to a grayscale keeping all values that are below a certain grayscale-threshold, resulting in a binary mask that represents areas of interest as well as other objects that passed this simple filter.

Next, areas smaller than a certain value were removed and small holes were filled so that fine connections between departments of several parts of objects could be kept intact to not adulterate the final image mask in terms of object quantity, size and other properties. In the end, all remaining areas are labelled: background area = 0, object areas = [1...N].

2) Shape Properties: To characterize these labelled areas, the following set of different parameters were calculated according to [12]:
Perimeter using the boundary list: $[X_1...X_N]$

$$perimeter = \sum_{1=1}^{N-1} d_i = \sum_{1=1}^{N-1} |X_i - X_{i+1}| \tag{1}$$

Major Axis using the end points: Two pixels of boundary that are farthest away
Major axis angle:

$$angle = tan^{-1}(\frac{Y_2 - Y_1}{X_2 - X_1}) \tag{2}$$

Minor axis endpoints: Points farthest away on a line that is perpendicular to the major axis.

$$axislength = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{3}$$

Compactness :

$$compactness = \frac{4\pi * area}{perimeter^2} \tag{4}$$

which equals 1 for a circle or pi/4 for a square.
Roundness is s

$$roundness = \frac{4\pi * area}{convexperimeter^2} \tag{5}$$

Elongation using the bounding box (bbox) of the object with dimensions of major and minor axis for width and length:

$$elongation = \frac{width_{bbox}}{length_{bbox}} \tag{6}$$

Eccentricity:

$$eccentricity = \frac{minoraxislength}{majoraxislength} \tag{7}$$

Spatial moments were defined by:

$$m_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q \quad \text{for} \quad p, q = 0, 1, 2... \tag{8}$$

This yields the area for zero order were p = q = 0. Further, this formula results in centres of gravity, or first order moments:

$$centroid = [\bar{x}, \bar{y}] = [\frac{m_{10}}{m_{00}} \frac{m_{01}}{m_{00}}] \tag{9}$$

Central Moments:

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^p (y - \bar{y})^q \quad \text{for} \quad p + q > 1 \tag{10}$$

Normalising those central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad \text{with:} \quad \gamma = (p + q)/2 + 1 \tag{11}$$

From those normalized central moments, HU moments were calculated (scaling, translation and rotation invariant parameters) as described in Formula (12). These were introduced by [13].

$$\begin{aligned}
\phi_1 =& \eta_{20} + \eta_{02} \\
\phi_2 =& \eta(\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 =& (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \mu_{03})^2 \\
\phi_4 =& (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \mu_{03})^2 \\
\phi_5 =& (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
& + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30}) + \eta_{12}^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 =& (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21}) + \eta_{03})^2] \\
& + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 =& (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
& - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{12}$$

3) k-Means and Silhouette Score: The k-Means algorithm was used on the calculated centroids of the areas. Since the amount of clusters on each Whole Slide Image was not known beforehand, several clusterings between 3 and C_{max} clusters were calculated (where C_{max} depends on images size and expected amount of clusters). If a slide only held one or two objects we did not look for local groupings. In each run, k-Means tried to minimize a potential function using a number of cluster centers set in a specific smart way

[14] (introduced by the authors as k-Means++) and given data points (area centroids in this case).
For evaluating the performance of each clustering process we used the Silhouette Score according to [15], shown in Formula (13).

$$Sscore = mean(\sum_{n=1}^{N} \frac{b_n - a}{max(a,b)}) \tag{13}$$

In this formula "a" is the mean intra cluster distance and "b" the distance to the nearest cluster that the current sample is not part of. Values are generally between -1 and 1. Higher scores represent small clusters with a larger distance between each other (good separability) compared to smaller numbers.
After the clustering for each group, the minimum bounding box was calculated from the individual areas belonging to the group containing x and y - coordinates and rotational parameter. These boxes are the fundamental single objects of the sample space that are later used for visualisation.

B. Compare Groups

A convex hull around the areas that belong to one group that yielded the highest silhouette score was used as a representative group area. Finally, those convex hull areas or representative group areas were compared calculating the euclidean distance of their HU-moments. The closer the distance between two areas in this seven dimensional space the higher is their similarity. As another independent measurement we also calculated the structural similiarties (SSIM) of the objects between groups, according to [16]. Since SSIM yield scores between -1 and 1 we shifted the result as follows:

$$SSIM_{new} = \frac{SSIM + 1}{2} \tag{14}$$

This transforms the interval [-1;1] to [0;1] where 0 reflects the lowest and 1 the highest similarity. Next, it was important to find normalized values for the distances in HU-Space in order to be able to combine them with SSIM results (see Formula 15).

$$HU_{similarity} = \frac{1}{7} \sum_{i=1}^{7} \frac{|min(H_1[i], H_2[i])|}{|max(H_1[i], H_2[i])|} \tag{15}$$

Here, we compared the areas 1 and 2. For every HU-Moment, the division of the smaller value and the greater value was calculated. The result for each division is 1 or smaller. The mean of those values was added to the SSIM score as shown in Formula 16:

$$SIMSCORE_{total} = \frac{SSIM_{new} + HU_{similarity}}{2} \tag{16}$$

This combination of the transformed SSIM and the normalized HU-Distances was used as final similarity score. It was calculated for every pair of objects in the sample space and the final similarity measures were stored in a NxN matrix.

C. Similarity JSON

Afterwards, the similarity matrix that was calculated in the previous subsection was evaluated. A certain similarity threshold was set to find the N-best matches of each object. Since the evaluation was done for every object, the resulting list for each virtual group may contain the same entries multiple times as shown in Table 1.

	A	B	C	D
A	1	0.97	0.15	0.98
B	0.97	1	0.2	0.95
C	0.15	0.2	1	0.2
D	0.98	0.95	0.2	1

TABLE I
Different similarities for Objects A-D

Multiple entries of groups that already existed were ignored (in this case three equal groups existed for Members A, B and C) which would lead to the following list of virtual groups (compare Table I and Table II):

Group ID		1
Member	A	1
Member	B	0.97
Member	D	0.98
Group ID		2
Member	C	1

TABLE II
Objects sorted in groups

This representation was stored in a JSON file format with additional information such as: CollectionID, CollectionName and VirtualGroups that further contain: VgroupID and Members with: ObjectID, Similarity and Rotation. This information on relationships was then combined with information from the sample space as mentioned in section A.

D. Visualised Similarities

For the visualization of similar structures each group was cut out using the information of the bounding box stored in the sample space and rotated in order to make alignments possible. The angle from structural similarities was used for rotation. The basic translation as shown in (17) and the basic rotation shown in (18) are combined to a transformation-matrix T(x) (19).

Translation Matrix:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{17}$$

Rotation Matrix:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} cos(\theta) & -sin(\theta) & 0 \\ sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{18}$$

Combined Matrix T(x):

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = (\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(\theta) & -sin(\theta) & 0 \\ sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}) \begin{bmatrix} x \\ y \\ w \end{bmatrix} \tag{19}$$

IV. Results

A. Analyse Areas and Find Clusters

First, we scaled our images and centered the Whole Slide Image onto a white background, shown in Figure 3. Next, the binary mask was calculated, small holes were filled and small objects removed (see Figures 4 and 5).

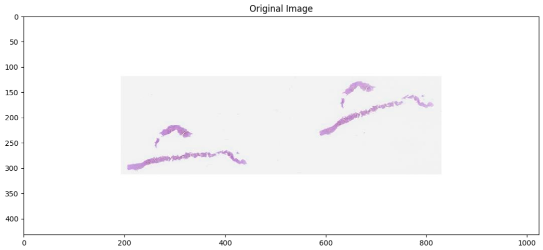


Fig. 3. Original Whole Slide Image on white background

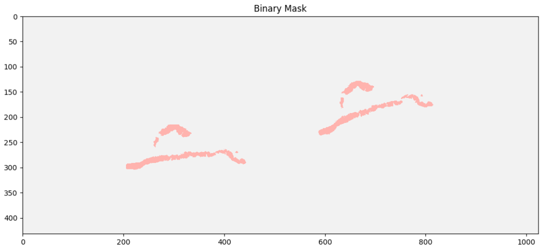


Fig. 4. Binary mask of original image

Characteristic values were calculated as presented above and stored to be able to compare areas as shown in Table 3. These values serve as similarity measures between single areas and support the grouping done by the k-Means algorithm.

After the different k-Means calculations, clusters that share similar locations are identified and evaluated using the silhouette score demonstrated in Figure 6. Centroids that belong to a certain cluster are marked with different colors.

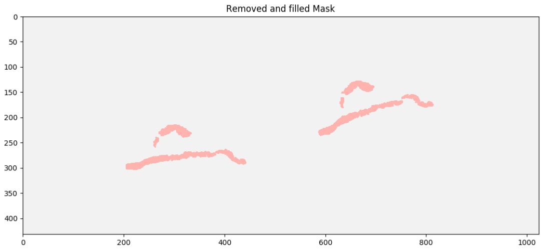


Fig. 5. Binary mask with filled wholes and removed small objects

Parameter	Area1	Area2
Perimeter	150	145
Angle	7	78
Axislength	54	51
Compactness	0.63	0.67
Roundness	0.44	0.46
Eccentricity	0.33	0.31
HU-Moment1	0.87	0.86

TABLE III
Similar parameters of 2 areas

After grouping areas and identifying objects, the characteristic values (HU-Moments) were calculated using the convex hull representation of those objects (see Figure 7).

Then, for these hull representations a minimum bounding box was calculated. ObjectID, SlideID, the coordinates of the bounding box and its rotation are stored in the "Sample Space" as objects (see Figure 8).

B. Compare Groups

For every found object on all slides in a given sample space, SSIM and HU-Distances are combined (as shown above) and the final similarity score is stored in a NxN matrix (see Table IV).

	OB5751	OB12	OB96972	OB68775	OB4575
OB5751	1	0.375	0.157	0.482	0.947
OB12	0.375	1	0.951	0.991	0.288
OB96972	0.157	0.951	1	0.269	0.181
OB68775	0.482	0.991	0.269	1	0.334
OB4575	0.947	0.288	0.181	0.334	1

TABLE IV
Similarity matrix for 5 different objects

C. Similarity JSON

Each object holds its ID, the slide number it belongs to, the similarity score to the first object and the rotation in degrees. The sample space stores other information on parameters of each object as mentioned above.

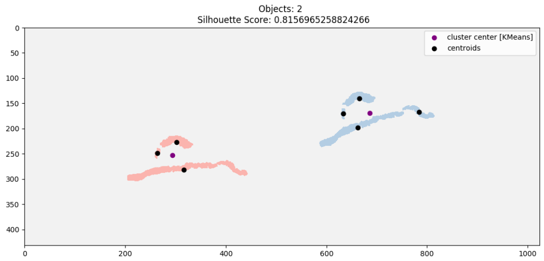


Fig. 6. k-Means clusters and silhouette score for K=2 clusters

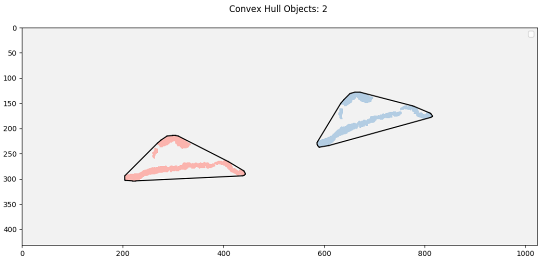


Fig. 7. Hull representation of two objects

```
{
  "CollectionID": "COL12754272",
  "CollectionName": "Sample Collection 1",
  "VirtualGroups": [
    {
      "VgroupID": "VG052754",
      "Members": [
        {
          "ObjectID": "OB5751",
          "Similarity": 1,
          "Rotation": 0
        },{
          "ObjectID": "OB68775",
          "Similarity": 0.974,
          "Rotation": 25
        }
      ]
    },{
      "VgroupID": "VG127",
      "Members": [
        {
          "ObjectID": "OB12",
          "Similarity": 1,
          "Rotation": 0
        },{
          "ObjectID": "OB96972",
          "Similarity": 0.951,
          "Rotation": 169
        },{
          "ObjectID": "OB4575",
          "Similarity": 0.991,
          "Rotation": 52
        }
      ]
    }
  ]
}
```

Listing 1. Similarity JSON

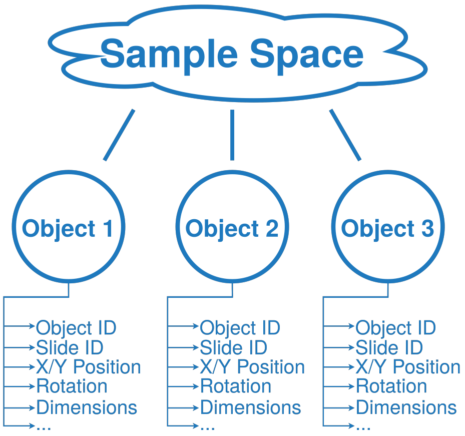


Fig. 8. Sample Space and Objects

D. Visualised Similarities

With the Sample Space, the Similarity Json from the previous steps and the application of the transformation T(x) (19), now a new way of representing WSI is possible. In the example in Figure 9, the different detected objects on the slides are re-aligned and transformed to a virtual group. This enables a new way for the pathologist to view the scanned image, browse through the entire sample and survey whole collections.

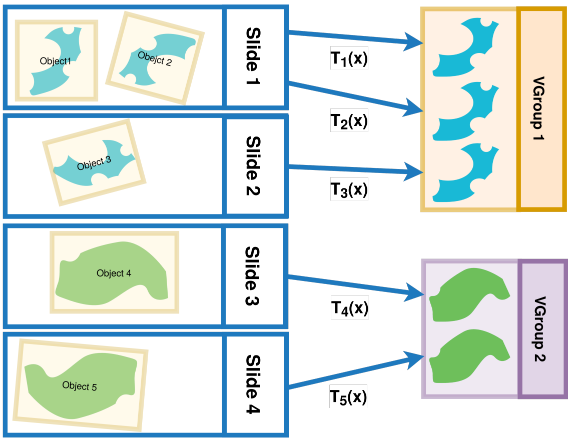


Fig. 9. Transformation and Visualisation

V. Discussion

With the presented specific parameters and properties of areas, similarity scores and data structure, we provide a pipeline to describe and organize tissue areas on Whole Slide Images. The parameters and similarity scores need further investigation in terms of error rates and stability over a variety of different test images and data sets. While some parameters are interesting candidates for further classification tasks, we will also investigate the possibilities of 3D reconstructions of Whole Slide Images based on this work in the future.

Whole slide scanning of archive tissue slides as data source for the development of AI algorithms will introduce further requirements. For future AI supported workflows in digital pathology we need a clear understanding of how pathologists make diagnoses [17], and also a good insight into possible visualization methods [18]. This is essential for the design and development of new human-AI interfaces, which enable pathologists to ask questions [19], [20] interactively via dialog systems [21]. This would help to increase the quality of making diagnoses and is the ultimate goal of an AI assisted digital pathology pipeline.

VI. Acknowledgements

The authors declare that there are no conflicts of interests and the work does not raise any ethical issues. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 “A reference model of explainable Artificial Intelligence for the Medical Domain”, and by the European Union’s Horizon 2020 research and innovation program under grant agreements No824087 "EOSC-Life".

References

[1] A. Holzinger, B. Malle, P. Kieseberg, P. M. Roth, H. Müller, R. Reihs, and K. Zatloukal, “Towards the augmented pathologist: Challenges of explainable-ai in digital pathology,” *CoRR*, vol. abs/1712.06657, 2017. [Online]. Available: http://arxiv.org/abs/1712.06657

[2] S. Al-Janabi, A. Huisman, and P. J. Van Diest, “Digital pathology: current status and future perspectives,” *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

[3] H. Müller, G. Dagher, M. Loibner, C. Stumptner, P. Kungl, and K. Zatloukal, “Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management,” *Current Opinion in Biotechnology*, vol. 65, pp. 45–51, 2020.

[4] H. Müller, R. Reihs, K. Zatloukal, F. Jeanquartier, R. Merino-Martinez, D. van Enckevort, M. A. Swertz, and A. Holzinger, “State-of-the-art and future challenges in the integration of biobank catalogues,” in *Smart Health*. Springer International Publishing, 2015, pp. 261–273.

[5] H. Müller, N. Malservet, P. Quinlan, R. Reihs, M. Penicaud, A. Chami, K. Zatloukal, and G. Dagher, “From the evaluation of existing solutions to an all-inclusive package for biobanks,” *Health and technology*, vol. 7, no. 1, pp. 89–95, 2017.

[6] S. Gainotti, P. Torreri, C. M. Wang, R. Reihs, H. Müller, E. Heslop, M. Roos, D. M. Badowska, F. De Paulis, Y. Kodra *et al.*, “The rd-connect registry & biobank finder: a tool for sharing aggregated data and metadata among rare disease researchers,” *European Journal of Human Genetics*, vol. 26, no. 5, pp. 631–643, 2018.

[7] T. Klingstrom, M. Mendy, D. Meunier, A. Berger, J. Reichel, A. Christoffels, H. Bendou, C. Swanepoel, L. Smit, C. Mckellar-Basset, E. Bongcam-Rudloff, J. Soderberg, R. Merino-Martinez, S. Amatya, A. Kihara, S. Kemp, R. Reihs, and H. Müller, “Supporting the development of biobanks in low and medium income countries,” in *2016 IST-Africa Week Conference*. IEEE, May 2016. [Online]. Available: https://doi.org/10.1109/istafrica.2016.7530672

[8] H. Bendou, L. Sizani, T. Reid, C. Swanepoel, T. Ademuyiwa, R. Merino-Martinez, H. Müller, A. Abayomi, and A. Christoffels, “Baobab laboratory information management system: development of an open-source laboratory information management system for biobanking,” *Biopreservation and biobanking*, vol. 15, no. 2, pp. 116–120, 2017.

[9] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, “Histoqc: An open-source quality control tool for digital pathology slides,” *JCO Clinical Cancer Informatics*, vol. 3, pp. 1–7, 04 2019.

[10] M. Hosseini, D. Lee, D. Gershanik, D. Lee, and S. Damaskinos, “Whole slide preview image segmentation and setup for digital pathology scanners,” 02 2020.

[11] M. Lukic, E. Tuba, and M. Tuba, “Leaf recognition algorithm using support vector machine with hu moments and local binary patterns,” 01 2017, pp. 000 485–000 490.

[12] L. Kopanja, D. Žunić, B. Lončar, S. Gyergyek, and M. Tadic, “Quantifying shapes of nanoparticles using modified circularity and ellipticity measures,” *Measurement*, vol. 92, 06 2016.

[13] Ming-Kuei Hu, “Visual pattern recognition by moment invariants,” *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, February 1962.

[14] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA ’07. USA: Society for Industrial and Applied Mathematics, 2007, p. 1027–1035.

[15] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.

[16] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[17] B. Pohn, M. Kargl, R. Reihs, A. Holzinger, K. Zatloukal, and H. Müller, “Towards a deeper understanding of how a pathologist makes a diagnosis: Visualization of the diagnostic process in histopathology,” in *IEEE Symposium on Computers and Communications (ISCC 2019)*. IEEE, 2019.

[18] B. Pohn, M.-C. Mayer, R. Reihs, A. Holzinger, K. Zatloukal, and H. Müller, “Visualization of histopathological decision making using a roadbook metaphor,” in *23rd International Conference Information Visualisation (IV)*. IEEE, 2019.

[19] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, 2019.

[20] A. Holzinger, A. Carrington, and H. Müller, “Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations,” *KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt*, vol. 34, no. 2, 2020.

[21] E. Merdivan, D. Singh, S. Hanke, and A. Holzinger, “Dialogue systems for intelligent human computer interactions,” *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 57–71, 2019.

MISS - Minimal Information about Slides and Scans

M Plass[1], R Reihs[1,2], RM Martinez[3], H Müller[1,2]

[1]Medical University Graz - Institute of Pathology
[2]BBMRI-ERIC - Biobanking and BioMolecular resources Research Infrastructure - European Research Infrastructure Consortium
[3]KI Karolinska Institutet - Department of Laboratory Medicine

(Corresponding author markus.plass@medunigraz.at)

Introduction - High quality metadata and provenance information are essential to support product quality in almost all areas of digital pathology. Especially when datasets are used in computational pathology, we need the appropriate information to document the technical and medical validation and to support the regulatory approval process. There are several standards available covering dedicated parts, e.g. MIABIS for sample and donor metadata and DICOM or vendor specific attributes for file formats and scanning metadata. Our aim is not to propose yet another metadata standard, but to describe a small and minimal dataset across different standardization activities and initiate a community driven approach to collect and harmonize existing ontologies.

Materials and Methods - MISS was defined within the use cases of a large scale digitization effort for machine learning. Through several cycles with stakeholders from biobanking and machine learning we generated a first proposal.

Results - The minimal information about glass slides and their scanned representation is divided into three parts: Pre Scanning (Slide) Metadata: e.g. metadata from biobanks, glass slide labeling, cleaning; Scanning Metadata: e.g. technical parameters, resolutions and focus points; Post-Scanning (File) Metadata: e.g. image quality indicators. A first version of MISS and examples can be found at https://github.com/human-centered-ai-lab/MISS/wiki.

Conclusions – We invite the digital pathology community to comment on and contribute to the MISS github repository and to provide examples of their scanning metadata in specific application scenarios.

MEDIZINISCHE UNIVERSITÄT GRAZ   Munich, 25 March 2020
Our reference: M11572EP / AO

Medizinische Universität Graz
Technologieverwertung
Auenbruggerplatz 2./4. Stock, 8036 Graz, Austria

Device for generating an image of a biological sample on a glass slide for generating an inventory in an image database

---

FIELD OF THE INVENTION

The invention relates to the field of image generating of biological samples, in particular of biological samples on a glass slide, wherein the images are generated for generating an inventory in an image database.

BACKGROUND OF THE INVENTION

For performing a diagnosis on the basis of a sample on a glass slide, the sample, i.e. the biological sample, is nowadays typically marked with a patient number and type of the sample by hand and if appropriate also provided with markers for indicating an interesting region of the sample. These handwritten markings have the drawback that they are not machine-readable.

For digitalizing objects on glass slides mostly whole slide image scanners are used, so called WSI-scanner. For this digitalization, the purity of the glass slide is a crucial factor and the glass slides need an unambiguous machine-readable code, for example, a QR-Code or a barcode, but the original (handwritten) code or number may disappear below the new machine-readable code. Further, the handwritten markings, which should remain for documentation and possible follow-up diagnosis, disturb or harm the digitalization process, because biological sample can get indistinct below. The sample being analyzed with the slide image scanner has to be cleaned before the slide image scanner can start to analyze it. This cleaning also includes the removement of the marker indicating a relevant region on the sample for a specific diagnosis. During the removal of the markers for the digitalization and/or the cleaning of the sample, the sample may be destroyed. Further, due to the elaborate sample preparation process a sample quality control has to be conducted by hand before the sample is prepared to be used by the slide image scanner of the prior art. The currently used slide image scanner are expensive in purchasing and further has a low throughput of glass slides, such that it is not possible to easily and cheaply catalogue or inventory biological samples.

SUMMARY OF THE INVENTION

Therefore, there exists a need for a device, which allows for a simple, cheap and easy digitalization of a biological sample on a glass slide such that an overview of the whole sample on the glass slide including all markers and relevant handwritten patient codes can be provided, wherein the image is usable for generating an inventory in an image database. Further, there exists a need that failures during the creation of the machine-readable codes are prevented and that the original codes have to remain and must not disappear under the machine-readable code.

The object of the present invention is solved with the subject matter of the independent claims, wherein further embodiments are incorporated in the dependent claims. An object of the invention is to provide a device which is able to create a fast and easy to use overview image of the whole biological sample on a glass slide including markers and codes on the sample (and on the glass slide). Thereby, an inventory in an image database has access to the image of original non-processed, not changed, biological samples on the glass slide. This will be explained in more detail hereinafter.

According to a first aspect of the invention, a device for generating an image of a biological sample on a glass slide for generating an inventory in an image database comprises a receiving unit which is configured for receiving the biological sample on the glass slide and a camera configured for generating an image of the biological sample while being received by the receiving unit. The device further comprises a releasing unit configured for automatically releasing the glass slide from the receiving unit after generating the image by the camera, wherein the camera is configured to generate the image such that the image of the biological sample on the glass slide comprises an overall image of the glass slide and the biological sample on the glass slide. Further, the releasing unit is configured for applying a mechanical force onto the glass slide received by the receiving unit for releasing the glass slide from the receiving unit.

In particular, the device is configured for generating an image of a biological sample on a glass slide for inventorizing an image of a biological sample in an image database. The device is able to automatically release the glass slide after the generation of the image, this leads to a fast throughput of glass slides and enhances the workflow for digitalization and for creating an inventory of the images of biological samples. The generated image can be used for digitalization of the biological sample and an inventory for an image database can be build up after generating a plurality of images.

In the context of the present invention, the term “inventory” shall be understood to describe a collection and/ or a library of images of biological samples, in particular digitalized images of samples on a glass slide. The inventory is further generated for providing an archive of biological samples. Further, with the inventory of the images an archive back up relating to the patients and/or to specific diagnosis or pathological findings can be provided.

In the context of the present invention, the term “automatically” shall be understood to describe that the releasing of the glass slide from the receiving unit after the generating of the image by the camera can be carried out without interaction of a user. The device itself automatically releases the glass slide. For example in other available scanners of the prior art, the glass slide has to be removed manually by the user.

In the context of the present invention, the term “overall image” shall be understood to describe an entire image of the biological sample on the glass slide, wherein the image can comprises a picture of the entire size of the glass slide and not only the area on which the biological sample is positioned on the glass slide. Hence, the overall image comprises the biological sample and additional the markers indicating a region being relevant for diagnosis and additional codes and/or numbers of the patient for identifying the patient, the tissue from which the sample has been taken, and/or the organ from which the sample has been taken, or other clinical relevant codes. On the other hand, the overall image shall be understood to describe an entire image of the biological sample excluding non-relevant parts of the glass slide. In particular, the overall image comprises the biological sample and these parts of the glass slides onto which the additional markers are provided, wherein parts of the glass slide which do not comprise any information relating to the biological sample are not comprised in the overall image.

The receiving unit is further configured for holding the glass slide when received by the receiving unit. The camera may be arranged above the receiving unit and also above the releasing unit such that the camera may generate an image of the biological sample on the glass slide from above. A conventional camera may be used for generating the image of the sample. When using a conventional camera the use of expensive optical microscope, digital microscope, may be avoided. It may also be possible to use more than one camera for generating the image of the sample, wherein one or more cameras are generating the same overall image or each camera is generating an image of different parts of the sample and the glass slide, which plurality of images may afterwards be processed to one overview image.

The releasing unit applies a mechanical force onto the glass slide for releasing the glass slide from the receiving unit, in particular for releasing the glass slide from the device itself. The mechanical force may be applied onto at least one side of the glass slide, wherein the at least one side of the glass slide may be a side surface of the glass slide on which no sample is attached. The mechanical force may directly interact at the glass slide. Further, the mechanical force may be a pushing force and/or pulling force directly interacting on the glass slide for moving the glass slide out of the device.

Further, the releasing unit may be configured and arranged in such a manner that it forms part of the receiving unit. Hence, the receiving and the releasing unit may be embodied in one spatial unit, wherein different elements of said spatial unit are configured and arranged for the receiving function and the releasing function, as was described herein before.

According to an exemplary embodiment of the invention, the releasing unit comprises a release element and a piston configured for releasing the glass slide, when the mechanical force is applied, wherein the mechanical force is generated by the piston for moving the release element.

The piston is configured and arranged to move the release element, wherein the release element may only be moved when the piston will be moved. In particular, the piston is automatically moved using a pushing force for pushing the release element towards the glass slide, such that the glass slide may be moved out of the releasing unit and out of the device.

An exemplary embodiment of such a release element and piston can be seen Figure 5 and Figure 6. For instance, the piston may be movable from a rest position into a release position by the mechanical force, wherein the release position is a position in which the piston causes the release element to move the glass slide out of the releasing unit. In particular, in the release position the release element is in contact with the glass slide, wherein the contact is established by the movement of the piston into the release position. On the other side, the rest position is a position of the piston wherein the piston does not cause the release element to move the glass slide. In particular, the rest position may be a position wherein the release element is not in contact with the glass slide. For instance, the piston is moving the release element away from the glass slide and the piston is positioned in the rest position. Hence, when a new glass slide has to be inserted into the receiving unit, the piston has to be moved into the rest position different to the release position. In particular, before the insertion of a new glass slide the piston is moved by a pulling force into the rest position. By moving the piston back into the rest position, the piston, in particular the release element will not disturb the receipt of the glass slide. The glass slide is moved by the release element and the piston in a predetermined direction and further along a predetermined direction. The predetermined direction may be any direction, which leads the glass slide out of the releasing unit, in particular out of the device. The configuration of the release element and the piston realizes an easy constructive way for an automatical release mechanism.

The release element may be a simple mechanical component, having a cuboid or cylindrical shape, wherein one end of the release element is attached at the piston and the opposite end may form a contact with the glass slide for moving the glass slide. The release element may be in direct contact with at least one side of the glass slide when moving the glass slide out of the releasing unit.

According to an exemplary embodiment of the invention, the piston is a turnbuckle configured for adjusting the position of the release element.

The turnbuckle may be used to adjust the tension and/or the length of the distance between the turnbuckle and the release element, in particular for increasing or decreasing the distance between the release element and the turnbuckle. Insofar, the release element can be positioned closer or further away from the glass slide to be moved by the release element, this may also serve for adjusting the strength of the mechanical force, which will be applied to the glass slide by the release element. Further, by adjusting the distance between the glass slide and the release element by the turnbuckle, glass slides with different lengths can be inserted into the releasing unit. Therefore, the device may be used for glass slides having different lengths. The elements of this embodiment can be seen from the very detailed embodiment shown in Figure 5 and Figure 6.

According to an exemplary embodiment of the invention, the device further comprises a receiving opening configured for receiving the glass slide in the receiving unit, and a support unit for holding the glass slide in a predefined position in the receiving unit for generating the image by the camera.

For example, the receive opening is an opening in a housing of the device, i.e. an opening in a wall of the device through which the glass slide has to be inserted into the device, in particular into the receiving unit. Further, the receiving opening may be an opening in the receiving unit, wherein the receiving unit is configured and formed in such a manner that one glass slide fits through the opening. Moreover, the receiving opening is formed such that a size of the receiving opening corresponds to a size of the glass slide. When the glass slide is received through the receiving opening in the receiving unit the glass slide will afterwards rest on the support unit in the receiving unit. In other words, the support unit is arranged in the receiving unit, forming a base onto which the glass slide may be stored or held for generating the image of the biological sample. The support unit may be a simple plate or table onto which the glass slide may be laid after insertion into the device. For instance, the support unit may be configured and formed in such a manner that the surface onto which the glass slide rests on the support unit may be minimal. For this reason, an illuminating area of a light source used for image generation may be maximal, wherein the light source and the image generation will subsequently be elucidated in detail. The tolerances of the support unit may be chosen such that also non-processed glass slides, which may have an inclined cover slip, may be inserted into the receiving unit. Further, the support unit may be configured and formed in such a manner that the glass slide is supported by the support unit along its entire length.

According to an exemplary embodiment of the invention, the release element is configured for moving the glass slide out of the device along the support unit.

In particular, the support unit may be arranged below the glass slide, such that the glass slide is inserted into the receiving unit laying above the support unit, wherein the support unit may be formed along the entire length below the glass slide. Hence, the release element may move and/or push the glass slide out of the receiving unit along the entire support unit, such that the support unit serves as a support below the glass slide during the whole releasing movement of the glass slide.

According to an exemplary embodiment of the invention, the device further comprises a plate configured for guiding the release element towards the glass slide when releasing the glass slide, wherein a release opening formed in the plate is surrounded by the plate on three sides, wherein the glass slide is released by the release element through the release opening.

For instance, the plate may be arranged above the release element, such that the plate may form a barrier or a support above the release element, which causes the release element not to incline into a direction above the glass slide. The glass slide may be inserted from above through the plate, in particular through the release opening, wherein the release opening may have a size corresponding to the glass slide. The glass slide may rest on the support unit below the plate, or also in the plate, wherein the plate may form walls surrounding the glass slide on at least three sides and the support unit may form the bottom onto which the glass slide rests during the image generation. The releasing opening may be formed on a fourth wall, wherein the opening is formed in the entire fourth wall, such that the glass slide is not covered by the fourth wall for being removable through the fourth wall, i.e. the release opening.

According to an exemplary embodiment of the invention, the device further comprises a motor for moving the piston. Preferably, the motor is an electromotor, in particular a servomotor, or a step motor. The motor is attached to the piston in such manner that the motor moves the piston forward and backward. In particular, the piston is moved by the motor towards the release element. Further, the motor moves the piston from the rest position into the release position and backwards into the rest position. The motor may be electrically coupled to a power supply that may be part of the device of the present invention. Furthermore, the motor may be electrically coupled to a control unit configured for controlling the motor, and hence the removing of the glass slide. Moreover, the control unit may also control the camera of generating the image of the sample.

According to an exemplary embodiment of the invention, the device further comprises at least a first light source for illuminating the glass slide, and a second light source for illuminating the glass slide, wherein the first light source is arranged and configured to illuminate a first main surface of the received glass slide, wherein the second light source is arranged and configured to illuminate a second main surface of the received glass slide.

A main surface of the glass slide may be a surface on which the biological sample is attached to the glass slide; this means on the main surface on which the biological sample is visible. The first main surface may lay opposite to the second main surface of the glass slide. In other words, both light sources are illuminating the glass slide from different opposing sides, wherein the glass slide is illuminated by the first light source from above and illuminated by the second light source from below. The illumination at two sides of the glass slide may ensure that for instance by illuminating the first main surface a detailed image of the biological sample may be generated and by illuminating the second main surface a detailed image of the markers and codes on the glass slide for the biological sample may be generated, or vice versa. It may of course also be possible to use more than one light source for illuminating one surface of the glass slide. For instance, different spectral light sources may be used for illuminating the glass slide from one side and/or from both sides.

According to an exemplary embodiment of the invention, the first light source and/or the second light source is an LED. Further, the light sources may be an OLED or an LED-Panel. For instance, the glass slide is illuminated from above and from below by one LED covering the spectrum of visible light. Other light emitting sources may also be possible for being used in the device, wherein when using more than one light source different light sources for different light spectrums may be used.

According to an exemplary embodiment of the invention, the camera is configured for generating a first image from the glass slide when illuminated by the first light source and a second image from the glass slide when illuminated by the second light source.

According to this embodiment of the invention, the camera generates the first image from above, wherein the first main surface of the glass slide is illuminated. Further, the camera generates the second image from above the glass slide, wherein the second main surface of the glass slide is illuminated. In other words, the camera generates the first and the second image from above the glass slide, wherein at least the illumination of the glass slide is changed. The device generates two images from two illuminated main surfaces of the glass slide for ensuring that in the overall image all information of the sample on the glass slide will be included. During the image generation of one main surface typically only the respective light source is switched on, the other light source, which would illuminate the main surface from which no image presently is generated, would be switched off.

On the other hand, the camera may be configured to generate an image of the first main surface when the first light source is switched on, and may be further configured to generate an image of the second main surface when the second light source is switched on. According to this embodiment of the invention, the camera is arranged and configured to generate an image of each main surface of the glass slide. This means that the camera, in particular the image generating position of the camera, may be switchable between different positions for being able to generate an image of different surfaces of the glass slide.

Furthermore, more than one camera may be used for generating the image of the glass slide, wherein a first camera may be configured and arranged for generating an image of the first main surface of the glass slide and a second camera may be configured and arranged for generating an image of the second main surface of the glass slide. This may avoid the need to switch camera positions.

According to an exemplary embodiment of the invention, wherein the device further comprises an image processing unit configured for generating a combined image comprising the first image and the second image. In particular, the image-processing unit is configured for joining the first image and the second image together to one single image, one joint image. The combined image will comprise all information, which can be seen on the biological sample and on the glass slide. For instance, the generated first image may include the markers and the codes on the glass slide and the generated second image includes a whole view of the sample, or vice versa. Hence, the combined image comprises all information of both images describing the front- and the backside of the main plain of the glass slide. Therefore, also information other than relating to the biological sample itself may be comprised in the image.

According to an exemplary embodiment of the invention, the device is configured to control the at least two light sources by a pulse width modulation.

In particular, at least two LEDs are controllable by pulse width modulation. For instance, a pulse width modulation module is used for generating pulse width modulation, wherein the light sources are in electrically contact with the pulse width modulation module for being controlled by the pulse width modulation module. The pulse width modulation may be used for controlling the order in which the light sources should be switched on and off for the image generation. In particular, the pulse width modulation may control whether the glass slide is illuminated from above or below. Moreover, the pulse width modulation may be used to control the intensity, brightness, of the light source. Furthermore, the pulse width modulation may be configured to control the releasing unit, in particular to control the release element and the piston, wherein the pulse width modulation may control the motor, which moves the piston of the releasing unit. For controlling the motor, the pulse width modulation module may be in electrically contact with the motor.

According to an exemplary embodiment of the invention, the device further comprises a storage unit for storing the generated image in the device. Further, the storage unit may be used for storing the generated image in the image database and/or for storing the generated image for further processing. The further processing may use a diagnosis tool, or displaying for diagnosis on a display, a documentation of the images, and/or a quality control of the sample (for instance, whether the sample is air bubble free or not).

According to an exemplary embodiment of the invention, the camera creates the image with a resolution in a range of three-times to seven-times resolution, preferably a five-times resolution.

The device uses a lower resolution than the conventional slide image scanner, this has the advantage that a lower data capacity/volume may be needed for storing the image and in particular for processing the image. The lower resolution of the image is sufficient for generating an inventory of an image database, wherein a higher resolution is compulsory for an image generation by current WSI scanners

According to a second aspect of the invention, a method for generating an image of a biological sample on a glass slide and for generating an inventory in an image database comprises the steps of receiving the biological sample on the glass slide by a receiving unit of a device, generating an image of the biological sample on the glass slide by a camera of the device, while being received by the receiving unit, automatically releasing the glass slide by a releasing unit of the device after the generating of the image by the camera. The generation of the image of the biological sample on the glass slide comprises an overall image of the glass slide and the biological sample on the glass slide, wherein the releasing comprises applying a mechanical force onto the glass slide received by the receiving unit for releasing the glass slide from the receiving unit. In particular, the glass slide may firstly be received by the receiving unit and may be held by the receiving unit for image generation. Secondly, the camera generates the overall image and thirdly the glass slide may be removed from the receiving unit, and/or from the device automatically by the releasing unit.

According to an exemplary embodiment of the invention, the method further comprises the steps of illuminating a first main surface of the received glass slide by a first light source of the device, and illuminating a second main surface of the received glass slide by a second light source of the device. In particular, the light sources may be arranged and configured for illuminating the respective main surface, wherein they may not be able to illuminate the other main surface.

According to an exemplary embodiment of the invention, the method further comprises the steps of generating a first image of the glass slide when illuminated by the first light source, and generating a second image of the glass slide when illuminated by the second light source. In particular, the image is generated by the camera of the device from above the glass slide such that only the illumination direction changes and the image generation is conducted from the same side of the glass slide. On the other hand, it may also be possible to generate a first image of the first main surface illuminated by the first light source and generating a second image of the second main surface illuminated by the second light source. Hence, the image is generated from the respective illuminated glass slide and not only from one side of the glass slide.

According to an exemplary embodiment of the invention, the method further comprises the step of generating a combined image comprising the first image and the second image.

According to an exemplary embodiment of the invention, the method further comprises the step of using the generated image for generating the inventory of the image database.

Further, the device for generating an image of the biological sample on a glass slide may be configured to use the generated image for generating the inventory of the image database. The use of the generated image may be understood as separately using the generated first image and the generated second image for generating the inventory. On the other hand the use of the generated image may be understood as using the combined image, which comprises the first and the second image, for generating the inventory. In particular a plurality of images are generated by the device (and the method respectively) wherein the inventory is generated by using the plurality of generated images. The plurality of images may be the first generated image, the second generated image and/or the combined image. The images are used for being transferred into the image database and/or the images are stored in the image database, wherein as explained hereinbefore the device may comprise a storage which may be used as the image database. In particular, the use of the generated image may comprise the digitalization of the generated images, wherein each generated images may be automatically digitalized after the generation of the image by the camera. Afterwards, the digitalized image may be automatically stored in the inventory. The inventory may be used for automatically performing a diagnosis on the basis of the images stored in the inventory, wherein different images of the inventory may be compared with each other for diagnosis comparison.

BRIEF DESCRIPTION OF THE DRAWINGS

The aspects defined above and further aspects of the present invention are apparent from the examples of embodiment to be described hereinafter and are explained with reference to the examples of embodiments. The invention will be described in more detail hereinafter with reference to examples of embodiments, but to which the invention is not limited.

Fig. 1 illustrates a device according to an exemplary embodiment of the invention.

Fig. 2 illustrates a sidewall of the device according to an exemplary embodiment of the invention.

Fig. 3 illustrates a releasing unit used in an exemplary embodiment of the invention.

Fig. 4 illustrates a releasing unit used in an exemplary embodiment of the invention.

Fig. 5 illustrates a side view of a releasing unit used in an exemplary embodiment of the invention.

Fig. 6 illustrates a top view of a releasing unit used in an exemplary embodiment of the invention.

Fig. 7 illustrates a schematic block diagram of a method according to an exemplary embodiment of the invention.

DETAILED DESCRIPTION OF EMBODIMENTS

Figure 1 and Figure 2 illustrate a device (100) for generating an image of a biological sample on a glass slide (115) for generating an inventory in an image database. The device (100) comprises a receiving unit (110) configured for receiving the biological sample on the glass slide (115), a camera (102) configured for generating an image of the biological sample while being received by the receiving unit (110), a releasing unit (101) configured for automatically releasing the glass slide (115) from the receiving unit (110) after generating the image by the camera (102). The camera (102) is configured to generate the image such that the image of the biological sample on the glass slide (115) comprises an overall image of the glass slide (115) and the biological sample on the glass slide (115), wherein the releasing unit (101) is configured for applying a mechanical force onto the glass slide (115) received by the receiving unit (110) for releasing the glass slide (115) from the receiving unit (110). The device (100) comprises a housing, which in Figure 1 and 2 is indicated with a sidewall (111, 211). Inside the housing the receiving unit (110), the releasing unit (101) and the camera (102) are arranged. As can be seen in Figure 1 the device (100) comprises a receiving opening for the receiving unit (110), wherein the receiving opening is configured for receiving the glass slide (115) in the receiving unit (110), wherein in this embodiment the receiving opening is arranged at the housing of the device (100). The device (100) may be able to perform the method for generating an image of a biological sample on a glass slide (115) and for generating an inventory in an image database. The method may comprise the steps of receiving the biological sample on the glass slide (115) by the receiving unit (110) of the device (100). The method and the respective step will be explained in detail with Figure 7.

As schematically illustrated in Figure 2, the device (100) further comprises at least a first light source (203) for illuminating the glass slide (115) and a second light source (204) for illuminating the glass slide (115). The first light source (203) is arranged and configured to illuminate the first main surface (116) of the received glass slide (115), wherein the second light source (204) is arranged and configured to illuminate the second main surface (117) of the received glass slide (115). In particular, the first light source (203) is arranged above the glass slide (115) and the second light source (204) is arranged below the glass slide (115). The first light source (203) and/or the second light source (204) is an LED. The device may further comprise a scattering unit (118) for reducing the reflection of the light source. The scattering unit (118) may be arranged between the first light source (203) and the glass slide (115). The scattering unit may also be arranged between the second light source (204) and the glass slide (115). It may also be possible to use two scattering units on each side of the glass slide (115), such that reflections of the light source on both sides of the glass slide (115) may be reduced. The scattering unit may be arranged in each embodiment of the present invention. The camera (102) is configured for generating a first image from the glass slide (115) when illuminated by the first light source (203) and a second image from the glass slide (115) when illuminated by the second light source (204). In particular, the camera (102) generates an image of the glass slide (115) from above, which means from the first main surface (116) when illuminated by the first light source (203). Further, the camera (102) generates an image of the glass slide (115) from above, from the first main surface (116) when illuminating the second main surface (117), i.e. the glass slide (115) from below. On the other hand, it may also be possible that the camera (102) may be movably arranged at the device (100), such that the camera may generate an image of the first main surface (116) of the glass slide, when the first main surface (116) is illuminated by the first light source (203), and the camera may generate an image of the second main surface (117) of the glass slide (115), when the second main surface (117) is illuminated by the second light source (204), or vice versa. The device (100) may further comprise a printed circuit board (213) configured for further processing the generated images and/or for analysis of the generated images. Therefore, the device (100) further comprises an image processing unit arranged e.g. at the printed circuit board (213), which image processing unit is configured for generating a combined image comprising the first image and the second image. The printed circuit board (213) may therefore be electrically connected with the camera (102) and the light sources (203, 204). Further, the device (100) is configured to control the at least two light sources (203, 204) by a pulse width modulation (212), in particular by a pulse width modulation module (212). Hence, the pulse width modulation module (212) may be electrically connected with the camera (102), the two light sources (203 and 204) and also with the printed circuit board (213). For example, the printed circuit board (213) may further comprise a storage or a memory for storing the generated images. Further, the printed circuit board (213) may comprise a data connection for providing the generated images to a database for creating an image database. On the other hand, the memory, the storage, of the printed circuit board (213) may be configured to form or create an image database in the device (100) itself. Moreover, the device (100) may comprise an interface configured to read out data, in particular, to read out image data. The interface may be electrically connected to the printed circuit board (213) or it may be arranged at the printed circuit board (213). In the device (100), the camera (102) creates the image with a resolution in a range of three-times to seven-times resolution, preferably five-times resolution. Hence, a simple conventional camera may be applicable for the image generation in the device (100).

Figure 3 illustrates the releasing unit (101) according to an exemplary embodiment of the invention. The releasing unit (101) comprises a release element (309) and a piston (not visible in Figure 3, the piston can be seen in Figure 5 and 6) configured for releasing the glass slide (115), when the mechanical force is applied, wherein the mechanical force is generated by the piston moving the release element (309). The device (100) further comprises a support unit (307) for holding the glass slide (115) in a predefined position in the receiving unit (110) for generating the image by the camera (102). In particular, the support unit (307) may be formed as a plate or a table onto which the glass slide (115) may be laid. The release element (309) is configured for moving the glass slide (115) out of the device along the support unit (307). Further, the support unit (307) may be formed in such a manner that it extends along the release direction of the glass slide (115) such that the glass slide (115) may be supported along the entire movement by the support unit (307). The device (100) further comprises a plate (305) configured for guiding the release element (309) towards the glass slide (115) when releasing the glass slide (115), wherein a release opening (340) is formed in the plate (305). The release opening (340) is surrounded by the plate (305) on three sides, wherein the glass slide (115) is released by the release element

(309) through the release opening (340). Further, the support unit (307) may extend in the release direction beyond the plate (305) and the releasing opening (340). In other words, the support unit (307) extends further than the plate (305), i.e. is longer than the plate (305) in the direction of the release direction. As can be seen in Figure 3, the release opening (340) is formed on one side of the plate (305), for example the front side, and the release element (309) is getting in contact with the glass slide (115) on the opposite side. In particular, the plate (305) is formed above the received glass slide (115), such that the glass slide (115) is spatially moved above the support unit (307) and below the plate (305). The plate (305) may be formed of acrylic material, wherein the material is chosen to be less reflective, such that less reflection on the glass slide (115) may be generated during the illumination by the light sources (203, 304). Further, the plate (305) may guide the release element (309) and preventing it from being inclined into the above direction. Hence, the release element (309) is guided by the plate (305) during it's towards movement into the direction of the glass slide (115). The glass slide (115) may have sharp edges, therefore the receiving unit (110), in particular the support unit (307), may be made of stainless steel for providing a high stability and a long lifetime. This arrangement maximizes the support of the glass slide, the insertion and the release of the glass slide (115).

Furthermore, the release element (309) may be moved through an opening (308) arranged on a backside of the receiving unit (110). The opening (308) for the release element (309) may be formed below the plate (305), wherein the plate (305) may form an upper wall of the opening (308). The opening (308) for the release element (309) may have a shape corresponding to the shape of the release element (309), such that the guiding of the release element (309) may be further improved.

Figure 4 illustrates the releasing unit (101) according to an exemplary embodiment of the present invention. The releasing unit (101) may further comprise a guide rail (330) which may be configured for receiving a light source (204), wherein the light source (204) may be in a form of a LED plate. The material of the guide rail (330) may be duroplastic plastics for being light in weight. The LED plate can be easy pulled out through the guide rail (330), such that the LED plate may be easily cleaned.

Further, the plate (405) for guiding the release element (409) and which plate (405) comprises the release opening (440) may comprise inlet guide elements (441) arranged on each side of the release opening (440). The inlet guide elements (441) may narrow the release opening (440) such that the glass slide (115) may be prevented from being from being tilted during the insertion of the glass slide (115) onto the support (407). Therefore, these inlet guide elements (441) serve for a secure and easy insertion of the glass slide (115) onto the support unit (407).

Figure 5 illustrates the releasing unit (101) from a side view according to an exemplary embodiment of the invention, wherein in this view the plate (305) is removed from the releasing unit (101). As can be seen from Figure 5, when the plate (305) is removed, the release element (509) would not be guided and may be moved inclined. Further, the piston (514) is illustrated in this Figure, wherein the piston (514) may be a turnbuckle configured for adjusting the position of the release element (509). The piston (514) is connected with the motor (550), which moves the piston (514) and the connected release element (509) from a rest position to a release position. In this view of the releasing unit (101) the electrical connection of the LED panel representing one light source (204) is illustrated, wherein the LED panel may be electrically connected by a plug (551) to the printed circuit board (213) and/or to the pulse width modulation module (212).

Figure 6 illustrates the releasing unit (101) from a top view according to an exemplary embodiment of the invention. In particular, the electrical connection (660) of the second light source (204) and the electrical connection (660) of the motor (550) is illustrated. The second light source (204) and the motor (550) may be electrically connected to the printed circuit board (213) of the device (100) and may also be electrically connected to the pulse width modulation module (212) of the device (100).

Figure 7 illustrates a schematically block diagram of a method according to an exemplary embodiment of the invention. The device (100) may be able to perform the method for generating an image of a biological sample on a glass slide (115) and for generating an inventory in an image database. The method may comprise the steps of receiving the biological sample on the glass slide (115) by the receiving unit (110) of the device (100). This step is indicated in Figure 7 using block S1, wherein step S1 may be the first step carried out by the method. Further, the method may comprise generating an image of the biological sample on the glass slide (115) by a camera (102 displayed in Figure 2) of the device (100), while being received by the receiving unit (110). This step is indicated in Figure 7 using block S2, wherein this step S2 may be the second step carried out by the method. The method may further comprise the step S3 of automatically releasing the glass slide (115) by a releasing unit (101) of the device (100) after the generating of the image by the camera (102). This step may be a third step carried out by the method. The generation of the image of the biological sample on the glass slide (115) comprises an overall image of the glass slide (115) and the biological sample on the glass slide (115). The release comprises the application of a mechanical force onto the glass slide (115) received by the receiving unit (110) for releasing the glass slide (115) from the receiving unit (110).

While the invention has been illustrated and described in detail in the drawings and foregoing description, such illustration and description are to be considered illustrative or exemplary and not restrictive; the invention is not limited to the disclosed embodiments.

Other variations to the disclosed embodiments can be understood and effected by those skilled in the art in practicing the claimed invention, from a study of the drawings, the disclosure, and the appended claims. In the claims, the word "comprising" does not exclude other elements or steps and the indefinite article "a" or "an" does not exclude a plurality. A single processor or other unit may fulfil the functions of several items recited in the claims. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measured cannot be used to advantage. Any reference signs in the claims should not be construed as limiting the scope.

List of reference signs:

100	device for generating an image of a biological sample on a glass slide
101	releasing unit
102	camera
110	receiving unit
111, 211	sidewall
115	glass slide
116	first main surface of the glass slide
117	second main surface of the glass slide
118	scattering unit
203	first light source
204, 404	second light source
212	pulse width modulation
213	printed circuit board
305, 405	plate
306	receiving opening
307, 407	support unit
308	opening for the release element
309, 409, 509	release element
330, 430, 530	guide rail
340, 440	release opening
441	inlet guide element
514	piston
550	motor
551	light source connector
660	electrical connection

Claims:

1. Device for generating an image of a biological sample on a glass slide for generating an inventory in an image database, the device comprising

a receiving unit (110) configured for receiving the biological sample on the glass slide (115),

a camera (102) configured for generating an image of the biological sample while being received by the receiving unit (110),

a releasing unit (101) configured for automatically releasing the glass slide (115) from the receiving unit (110) after generating the image by the camera,

wherein the camera (102) is configured to generate the image such that the image of the biological sample on the glass slide comprises an overall image of the glass slide and the biological sample on the glass slide,

wherein the releasing unit (101) is configured for applying a mechanical force onto the glass slide (115) received by the receiving unit (110) for releasing the glass slide from the receiving unit (110).

2. Device according to claim 1, wherein the releasing unit comprises

a release element (309) and a piston (514) configured for releasing the glass slide (115), when the mechanical force is applied, wherein the mechanical force is generated by the piston moving the release element (309).

3. Device according to claim 2,
wherein the piston is a turnbuckle configured for adjusting the position of the release element (309).

4. Device according to any of the preceding claims, wherein the device further comprises

a receiving opening (306) configured for receiving the glass slide (115) in the receiving unit (110), and

a support unit (307) for holding the glass slide in a predefined position in the receiving unit for generating the image by the camera (102).

5. Device according to claim 4, wherein the release element (309) is configured for moving the glass slide out of the device along the support unit (307).

6. Device according to any of the claims 2 to 5, wherein the device further comprises

a plate (305) configured for guiding the release element (309) towards the glass slide (115) when releasing the glass slide,

wherein a release opening formed in the plate is surrounded by the plate on three sides,

wherein the glass slide is released by the release element through the release opening

7. Device according to any of the preceding claims, wherein the device further comprises

at least a first light source (203) for illuminating the glass slide, and

a second light source (204) for illuminating the glass slide,

wherein the first light source (203) is arranged and configured to illuminate a first main surface of the received glass slide (115),

wherein the second light source (204) is arranged and configured to illuminate a second main surface of the received glass slide (115).

8. Device according to claim 7, wherein the first light source and/or the second light source is a LED.

9. Device according to claim 7 or 8, wherein the camera (102) is configured for generating a first image from the glass slide when illuminated by the first light source (203) and a second image from the glass slide when illuminated by the second light source (204).

10. Device according to claim 9, wherein the device further comprises

an image processing unit configured for generating a combined image comprising the first image and the second image.

11. Device according to any of the preceding claims 7 to 10, wherein the device is configured to control the at least two light sources by a pulse width modulation.

12. Device according to any of the preceding claims wherein the camera (102) creates the image with a resolution in a range of three-times to seven-times resolution, preferably a five-times resolution.

13. Method for generating an image of a biological sample on a glass slide and generating an inventory in an image database, the method comprising the steps of

receiving the biological sample on the glass slide (115) by a receiving unit (110) of a device (S1),

generating an image of the biological sample on the glass slide by a camera (102) of the device, while being received by the receiving unit (110) (S2),

automatically releasing the glass slide by a releasing unit (101) of the device after the generating of the image by the camera (S3),

wherein the generating of the image of the biological sample on the glass slide comprises an overall image of the glass slide and the biological sample on the glass slide,

wherein the releasing comprises applying a mechanical force onto the glass slide received by the receiving unit (110) for releasing the glass slide from the receiving unit (110).

14. Method according to clam 13, wherein the method further comprises the steps of

illuminating a first main surface of the received glass slide by a first light source (203) of the device,

illuminating a second main surface of the received glass slide by a second light source (204) of the device.

generating a first image of the glass slide when illuminated by the first light source, and

generating a second image of the glass slide when illuminated by the second light source.

15. Method according to claim 14, further comprising the step of

generating a combined image comprising the first image and the second image.

16. Method according to any of the previous claims 13 to 15, further comprising the step of

using the generated image for generating the inventory of the image database.

Abstract:

The invention concerns a device and a method for generating an image of a biological sample on a glass slide for generating an inventory in an image database. The device comprises a receiving unit configured for receiving the biological sample on the glass slide, a camera configured for generating an image of the biological sample while being received by the receiving unit, a releasing unit configured for automatically releasing the glass slide from the receiving unit after generating the image by the camera. The camera is configured to generate the image such that the image of the biological sample on the glass slide comprises an overall image of the glass slide and the biological sample on the glass slide, wherein the releasing unit is configured for applying a mechanical force onto the glass slide received by the receiving unit for releasing the glass slide from the receiving unit.

(Fig. 3)

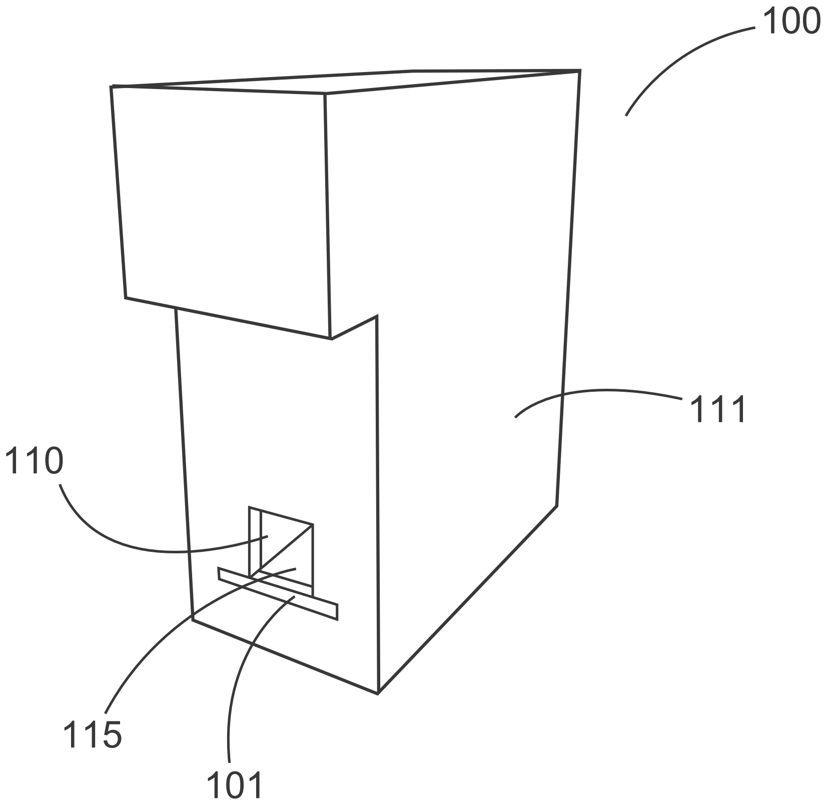

100
111
110
115
101

Fig. 1

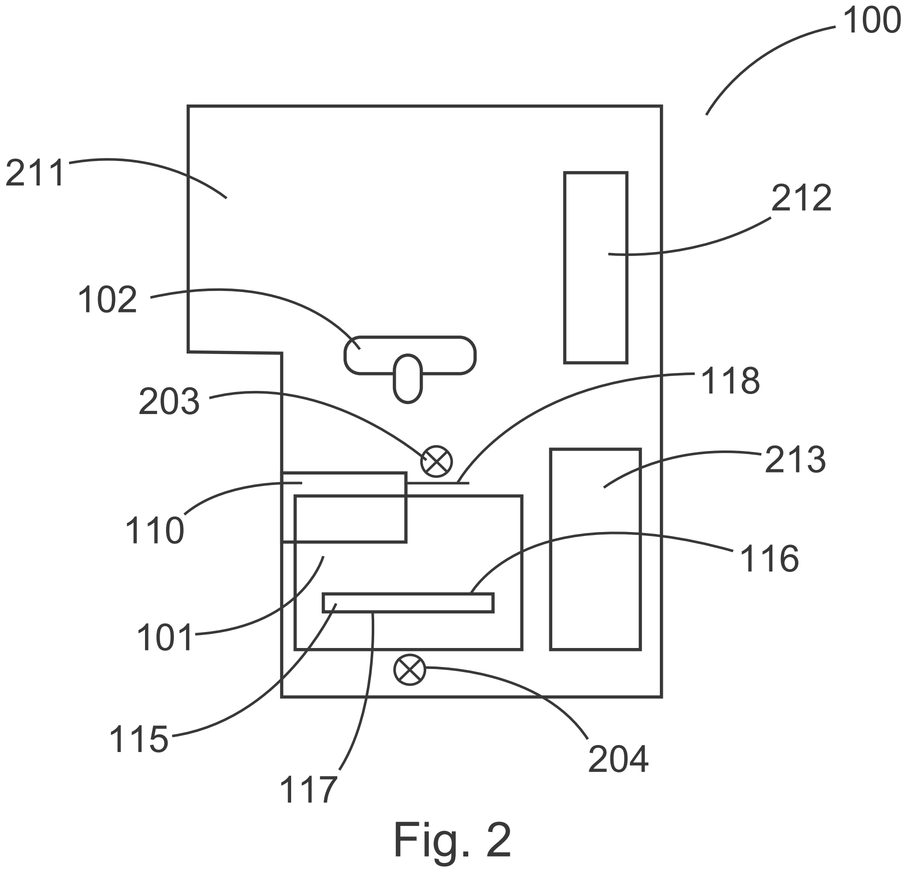

100
211
212
102
203
118
213
110
116
101
115
117
204

Fig. 2

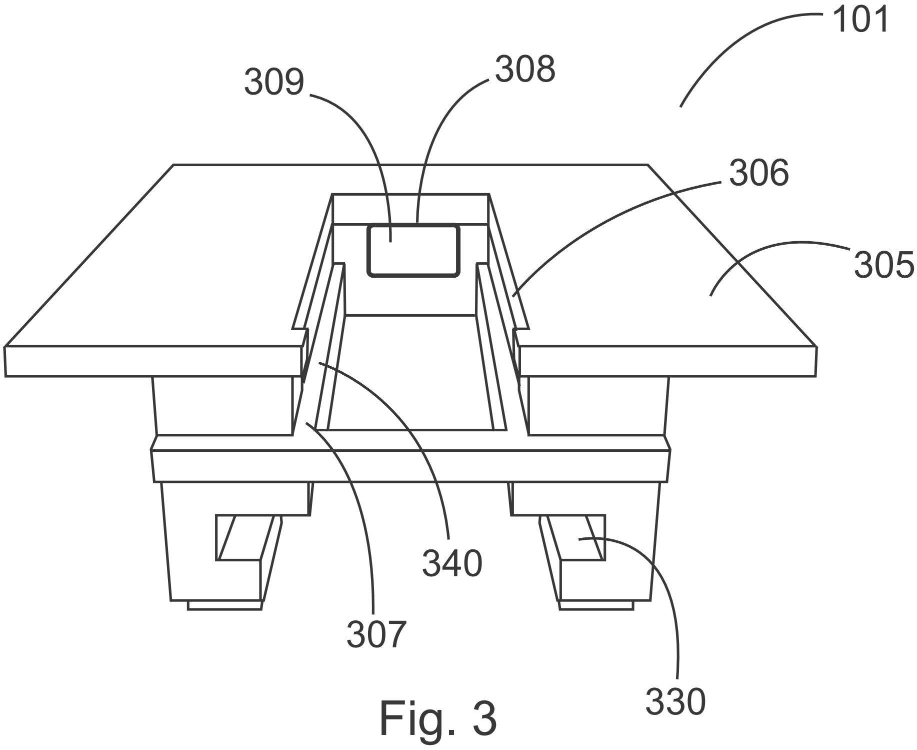

101
309
308
306
305
340
307
330

Fig. 3

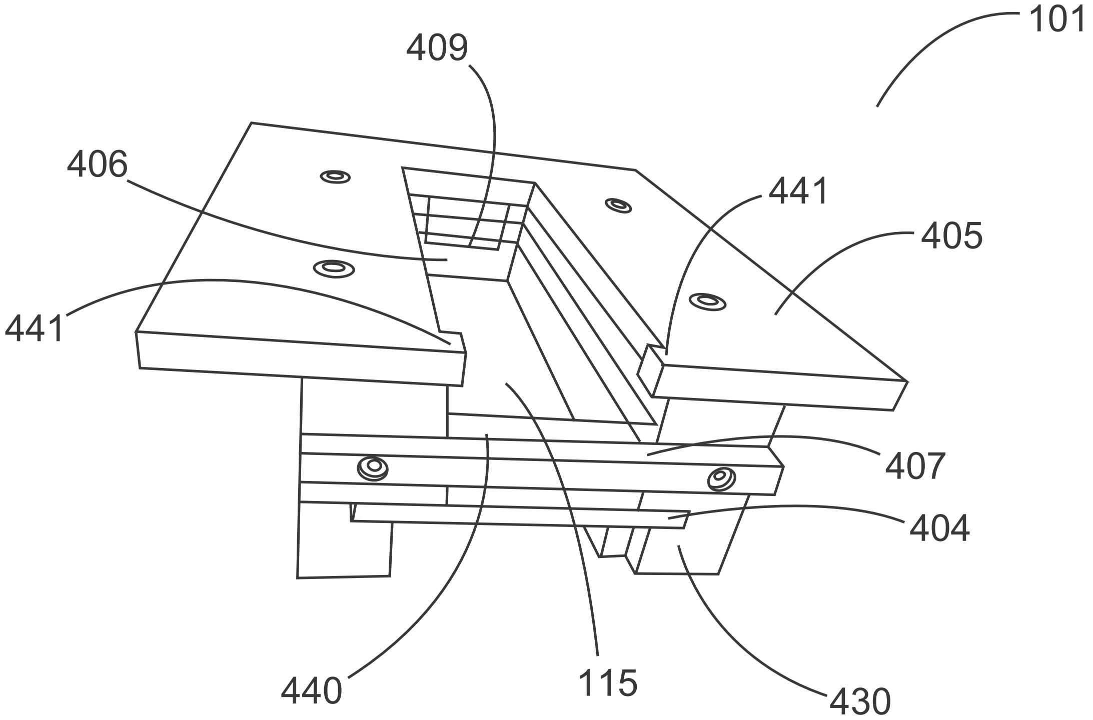

101
409
406
441
441
405
407
404
440
115
430

Fig. 4

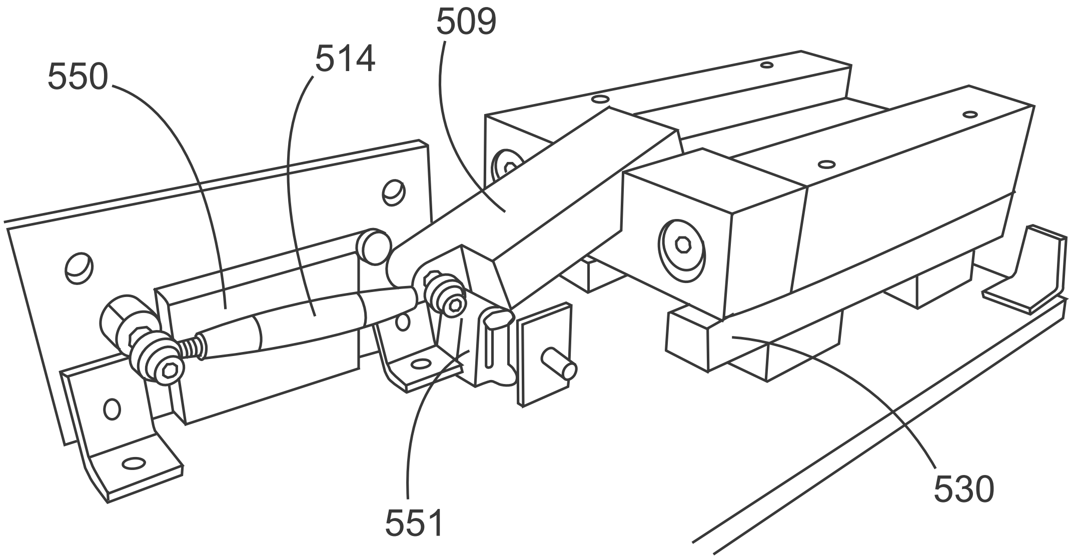


Fig. 5

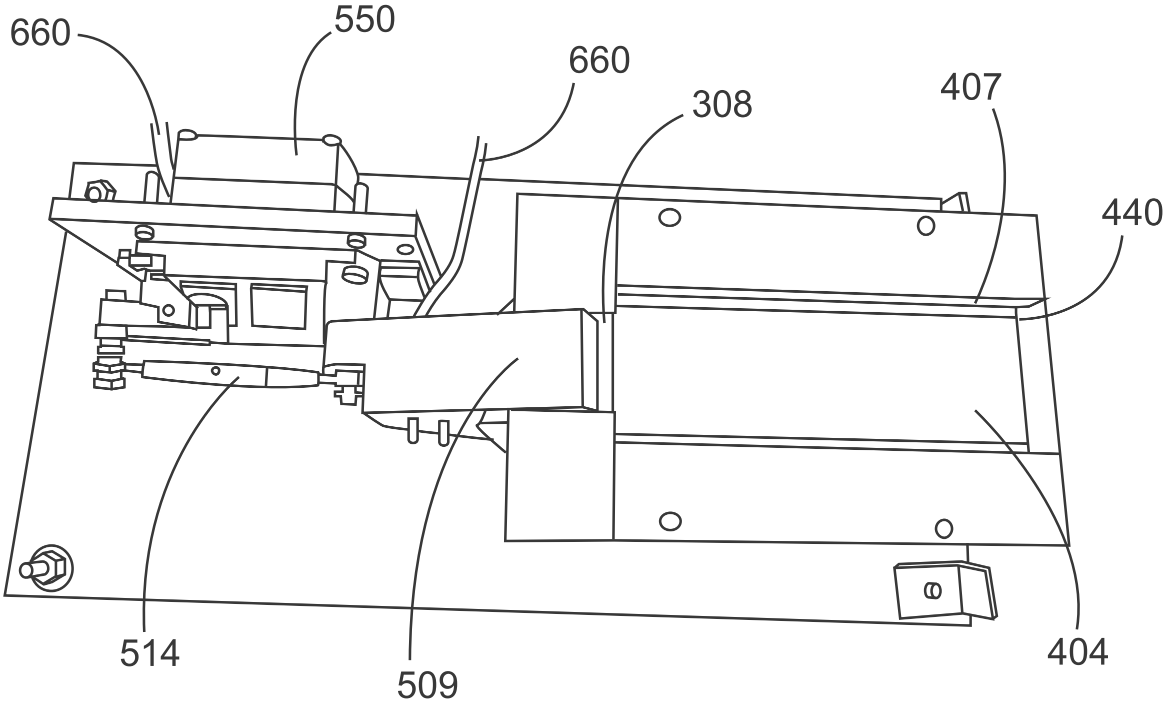


Fig. 6

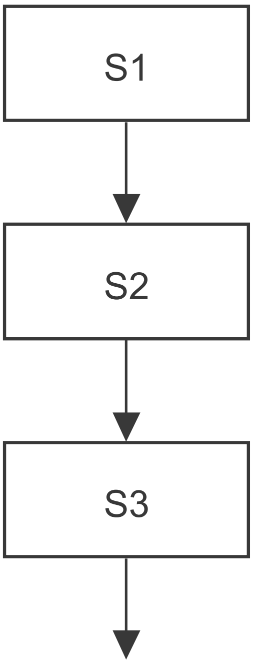


Fig. 7