945358-BIGPICTURE
Central Repository for Digital Pathology

WP3 - Data collection & management

D3.03 - Report on defined and first collected datasets

Lead contributor	20 - RO
	anna.c.boden@regionostergotland.se
Other contributors	21-MUW
	4-UMCU
	29-SECTRA AB

Document History

Version	**Date**	**Description**
V1.0	15 Jun 2022	First Draft specific input MUW, UMCU, SECTRA
V2.0	03 Aug 2022	Second draft complete, on review by WP3 on sharepoint
V3.0	31 Aug 2022	Third draft complete, reviewed by MB
V4.0	14 Sep 2022	Final Version

Abstract

In this deliverable report (D3.03) we describe the first versions of the defined and collected datasets for the Bigpicture repository. Also, it describes the underlying work procedures that have been performed in the processes to define datasets in terms of cohorts, storage file format standards, metadata schema and standards and metadata file storage standard. The work procedures also include practical work schemas for extracting, converting and anonymization of original data to generate datasets to the Bigpicture submission portal.

The main objective of this report is to describe the activities and procedures performed for submission of the first defined datasets fulfilling the Bigpicture Metadata Model. These datasets fulfill the Bigpicture Metadata Model in terms of being structured as defined in the Common Mandatory Metadata Schema (CMMS v.0.1.0), contained metadata adhering to the Mandatory Submission Metadata for Directly Accessible Datasets (MSMdad v0.1.0) and were transmitted to the repository by means of the developed flexible metadata storage format. The work procedures also include practical work schemas for extracting, converting and anonymization of original data to generate datasets to the Bigpicture submission portal and are described in five steps; Policy compliance, Define selection and parameters, Extraction, Curation and Export/submit data. These steps serve as a model and give a guidance of actions needed for future datasets contributors. Explicitly we hope to give good guidance in the steps described as well as to define these as selection and parameters, extraction, curate data and export/submit data. Furthermore, we hope to promote implementation of Wsidicomizer and the Bigpicture metadata interface in order to fulfill curation of datasets to Bigpicture.
The result of this deliverable is based on the practice and experience of three representative beneficiaries of surgical pathology departments connected to the collector nodes namely Cancer Node (lead by 4-UMC -Utrecht), Renal Node (lead by 21-MUW) and Skin Node (lead by 20-RO). The submitted data consists of one dataset per site within the scope of an anonymized directly accessible dataset.

By performing the steps described below, we have been able to meet Milestone 8, and submitted the first clinical data sets via the data submission portal to Bigpicture repository by June 2022.

Background

The first selected clinical datasets are defined as prototypes of datasets consisting of WSI and metadata from analyses in the pathology discipline on human tissue. They have been collected from clinical cases either from existing biobank information systems or laboratory information systems (LIS) and were anonymized. The data submission means Milestone 8 was met; and meeting this task (MS8) has rendered important experience on procedures and methods for extraction, conversion and submission as well as potential pitfalls and time required for data upload which aids in implementing specific measures to mitigate risks.
The work relates to D3.01 report on node network organisation model, D3.02 Minimal dataset deliverable, D4.03 Report on unified open digital slide and annotation format specification, D5.03 Report describing the Bigpicture pseudonymization strategy as well as D1.04 Data Management Plan and D2.02 Report on the pilot system establishment.

The main objective of this report is to frame a description of the activities and procedures performed for submission of the first defined and submitted datasets for the Bigpicture repository. By doing so, we provide guidance to future datasets contributors.

Scope

The report will mainly focus on the results from meeting MS8 in an outlined workflow which describes how to extract and submit data to Bigpicture as performed in the first collected datasets. In the first section there will be description of the background and relating activities combined with a general report structured in five main sections of data set creation. This general structure is followed by a more detailed report from the three data providers adding local actions and experiences based on the documentation by the three sites: MUW, RO and UMCU. The five main sections are:

- Policy compliance
- Define selection criteria and parameters
- Extraction
- Curate data
- Export/submit data

Used nomenclature

AI, artificial intelligence. Machine intelligence, a branch of multiple computer science technologies that focuses on building and managing technology that can learn to autonomously make decisions and potentially carry out actions on behalf of a human minds
Bit, b, a binary data digit (0 or 1)
Byte, B, a group of 8 bits
Cytomine, a company and viewer platform (in Bigpicture community)
CMMS, Common Mandatory Metadata Structure
CPU, central processing unit, a component on your computer
D, Deliverable. Deliverables and milestones are the key elements of a Horizon Europe proposal workplan
DAC, data access committee
DICOM, Digital Imaging and Communications in Medicine (DICOM) an international standard to transmit, store, retrieve, print, process and display medical imaging information
DoA, Description of Action
GB, gigabyte a unit for computer information (byte) with SI unit prefix G for a billion, 10^9
Honest Broker, a function and service that will be developed in Bigpicture
IOPS, Input/output operations per second, performance measurement used to characterize computer storage devices
LIS, laboratory information system
LIMS, laboratory information and management system
MAB, multistakeholder advisory board
MB, megabyte a unit for computer information (byte) with SI unit prefix M for a million, 10^6
Mbps, Megabits per second are units of measurement for network bandwidth and throughput.
MS, Milestone. Deliverables and milestones are the key elements of a Horizon Europe proposal workplan
MSM, Mandatory Submission Metadata
MSMdad, Mandatory Submission Metadata for Directly Accessible Datasets
MUW, Medical University of Vienna
NBIS, National Bioinformatics Infrastructure Sweden
Ndpi, a proprietary file format of- a pathology slide specimen image created by any Hamamatsu slide scanner
PACS, picture and archive system
QC, quality control
RO, Region Ostergotland (Östergötland)
S3, a cloud object storage from Amazon
Sda-cli, Sensitive Data Archive (Sda) Command Line Interface (sda-cli)
SOP, Standard Operating Procedure
UMCU, University medical center Utrecht
WSI, whole slide image

Wsidicomizer, a tool developed as part of Task 4.5 in Bigpicture applicable for curation of selected data
XML, The Extensible Markup Language (XML) a text-based format to represent structured information

Method

This report is the result of intensive collaboration between multiple workpackages, namely WP2, WP3 & WP4. It is the result of work based on reference implementation work in Task 3.6 - Collection of clinical data sets which is depending on work from Task 3.1 - Establishing collector nodes, Task 3.2 - Prioritisation of datasets, Task 3.3 – Development of ontologies and standardized metadata sets, Task 3.4 - Defining and implementing datasets and metadata sets, and Task 3.11 - Data conversion, as well as tasks directly linked to other work packages but enables data submission e.g. Task 2.3 – Pilot system development, Task 2.5 – Pilot operation and Task 4.5– Develop a technical implementation of the open digital slide, annotation, and meta-data format and Task 5.1 – Ethical and legal compliance monitoring.

The metadata contained within the submitted datasets adhere to the established standardized metadata data model of Bigpicture as described in D3.02. The defined set of mandatory metadata fields, termed Mandatory Submission Metadata for Directly Accessible Datasets (MSMdad) is delineated and described in this deliverable (D3.02). The metadata model constitutes of an explicit description of the data bearing fields and entities behind these fields within the Bigpicture repository and can be reviewed in the Common Mandatory Metadata Structure (CMMS). A defined set of the metadata attributes of these entities have been coined as Mandatory Submission Metadata (MSM) for directly Accessible Data Sets (MSMdad). The metadata model has been developed by the Taskforce Metadata and Nomenclature, a contributing working group of members from WP3, WP4 and WP2. As reported in the MAB report by June 2022, the first versions of a flexible metadata storage format adhering to the CMMS and the MSMdad have accordingly been documented on Bigpicture SharePoint and specifications are published on the Bigpicture GitHub Repository.

The result of this deliverable is based on the practice and experience of three representative beneficiaries of surgical pathology departments connected to the collector nodes namely Cancer Node (lead by 4-UMC -Utrecht), Renal Node (lead by 21-MUW) and Skin Node (lead by 20-RO). Each submitting party has the dual role as data owner in the clinical aspect of pathology data and personal data relating to the pathology data/specimen but also as node leader for a Node. The submitted data consists of one dataset per site within the scope of an anonymized directly accessible dataset. The work has been executed independently per site, but Bigpicture project management and strategic activities took place through the Collector Node Network based on regularly bi-weekly meetings as well as WP3 bi-weekly meetings. Extensive discussions between members of different WPs representing different partners and different expertise, but most of all discussions between data contributing beneficiaries within WP3 and 29-SECTRA AB (Erik Gabrielsson), for application of digital slide format and extraction tools, have also been the foundation for these results. Additionally, a workshop on the theme data extraction was held during the Bigpicture annual meeting in 2-4 February 2022.

Results

Data submission architecture

Data submission service layered architecture view foreseen by BP Architecture definition 0.1 as shown in Figure 1 and 2. These diagrams describes different capabilities and functions and the structure of this application architecture area are outlined in the figures below; the data pre-processing layered view (Figure 1) and the data submission service layered architecture view

(Figure 2). Our report is based on results from implementation of the applicable components conversion tool and anonymization tool and the application service selecting criteria's as shown in Figure 1. The users as outlined in Figure 2 are dataset contributors utilizing business functions such as data conversion, selection of data for submission, anonymization as well as legal compliance documentation.

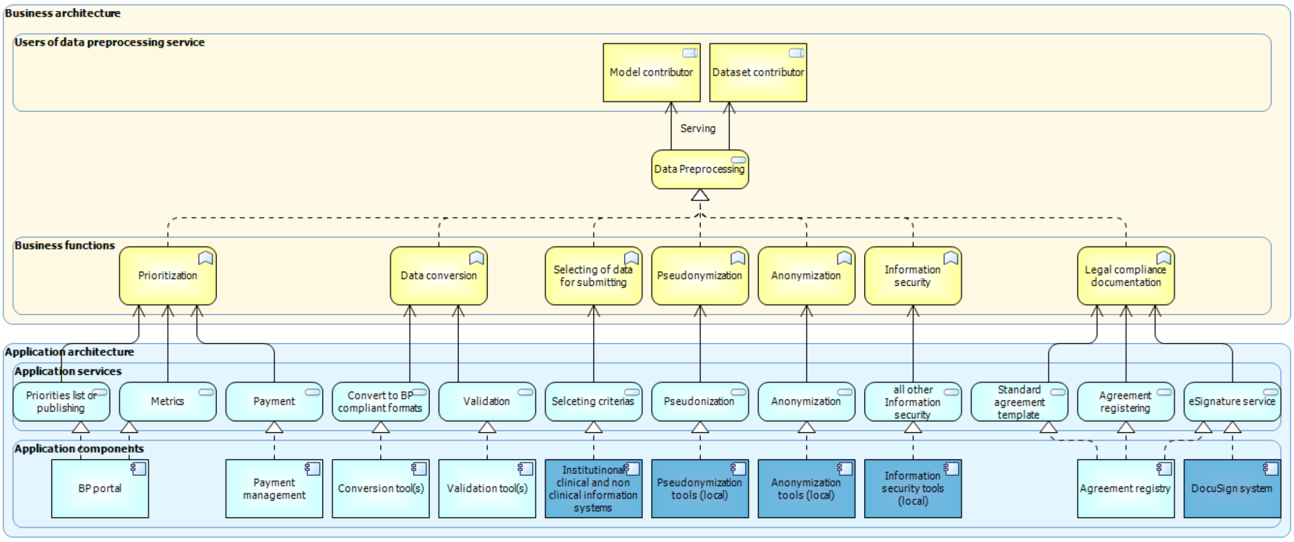

Figure 1. Data pre-processing application architecture view.

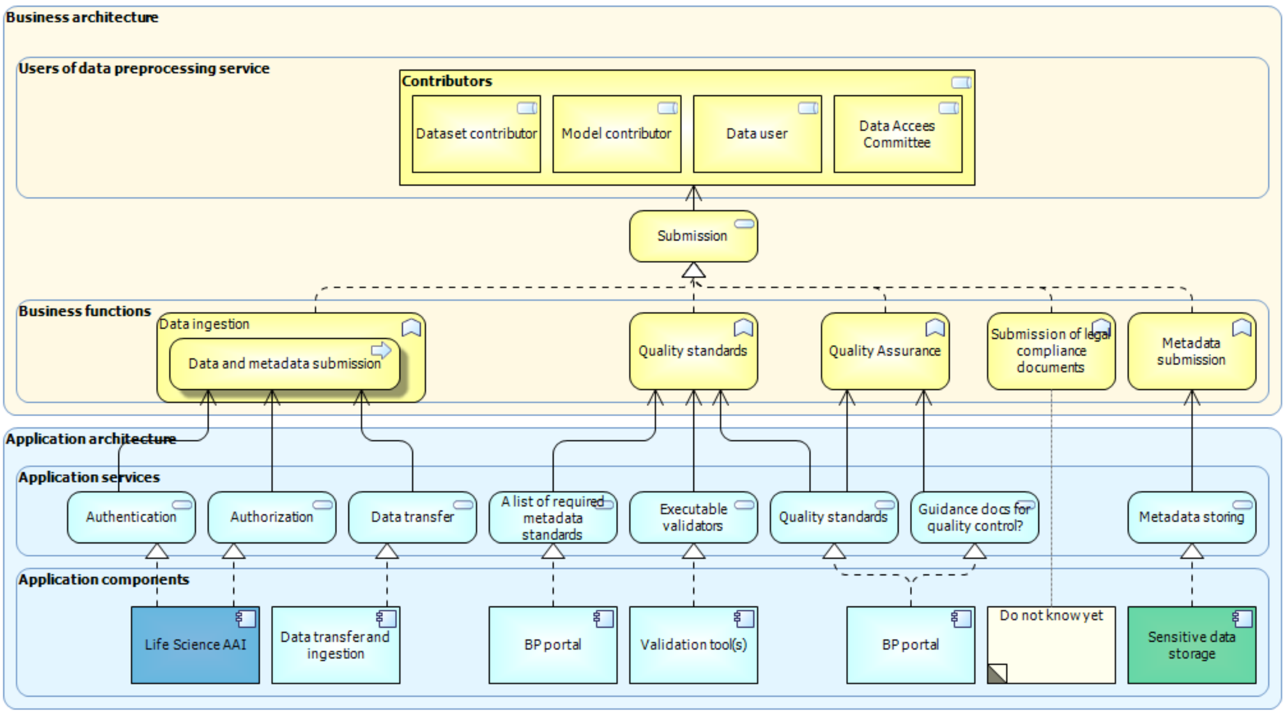

Figure 2. Data Submission application architecture view.

As stated in the conclusion of D4.03 the Bigpicture platform is based on existing technology that specifies a submission as metadata in a specific schema that is combined with data files in terms of images. The schema is initiated as XML files with filename references to the data files. Data submitters package their data in a submission format and submit through a data submission portal. Once data is submitted, the data is stored as is and is used by a number of services. It is possible to access the submitted data. A submission consists of metadata, digital slides, observations and

annotations as a complete package.
In the future data submission will constitute a wider range of different capabilities and requirements compared to the description of the prototypes in this report. One reason for this uncertainty is caused by the differences in technologies of data systems and organization of data for WSI and metadata among different data contributors connected to the clinical nodes. In addition, there will be data contributors who have pre-existing datasets primarily for another purpose but willing to contribute to the Bigpicture repository. The datasets/collection with already defined metadata might need to be customized individually for integration. The results in this report are based on the first submission from three data contributors (MUW, RO and UMC) who have applied the selection criteria retrospectively on reported and coded cases from the LIS system and extracted data (both WSI and metadata).
Overall, the logical sequence of steps taken in preparation of a dataset for submission is described as follows:

- Policy compliance
- Define selection and parameters
- Extraction
- Curate data
    - Implementing Wsidicomizer and the Bigpicture metadata interface
    - Quality control
- Export/submit data
    - Encryption and Data transfer

The outlined workflow, as presented by RO, for the first clinical dataset contribution can be reviewed in Figure 3.

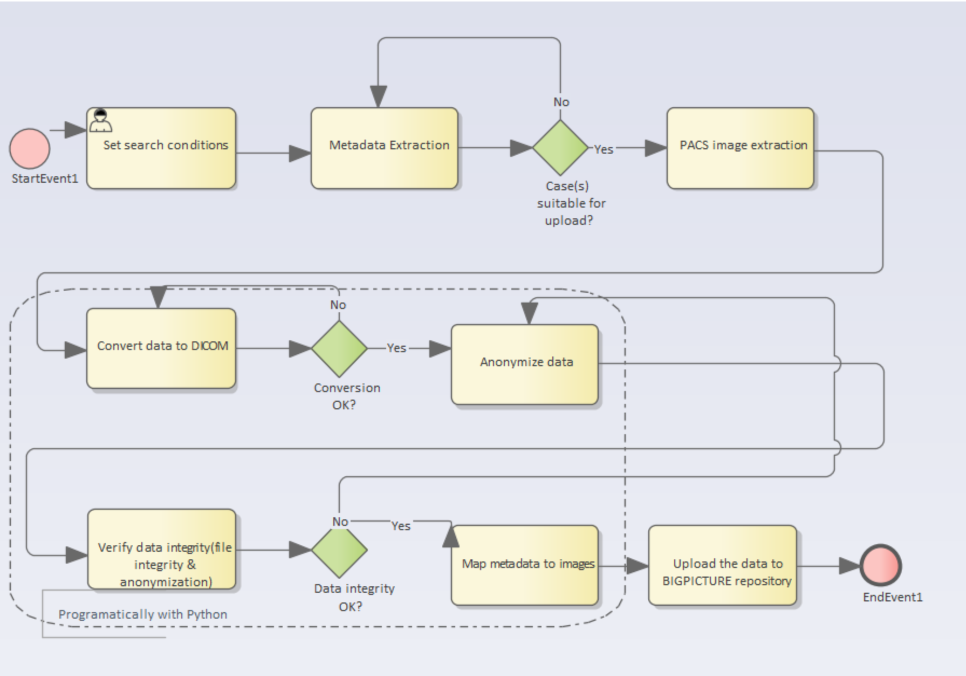

Figure 3. The workflow of RO to pre-process data before sending it to the Bigpicture data repository.

The boxed activities, being quite similar for all contributors, are as follows; set search conditions, extract data (metadata and WSI), conversion of data to DICOM format and XML file structure,

anonymization of data, verify data integrity, map metadata to images and upload of data to repository.
The protocols describing how the specific activity for each step will be performed must be defined and controlled locally since systems, data information structure and personal resources differs, hence a detailed description from the three different data contributors can be found in the appendix.

Policy compliance

The very first activity before export and anonymization of clinical data is to comply with policies. The policies refer firstly to the local/national policy that needs to be applied in regard to consent and/or ethical approval for research purposes and for mining personal data in the different data systems. A review of the ethical and legal compliance is described in reports from Task 5.1 – Ethical and legal compliance monitoring, namely D1.17 Ethics HCT-Requirements No.1, D1.18 Ethics-HCT-Requirements No.2, D1.19 Ethics-HCT-Requirements No.4, D1.20 Ethics-HCT-Requirements No.5 and D5.1 Overview of the legal framework of all countries involved in Bigpicture. Therefore, this report will not describe how each data contributor met these local and national legal and ethical policies.

Secondly, data transfer approval from the main authority is advisable. In addition, an established data management plan applicable for the dataset and Bigpicture repository should be in place, (see D1.4 Initial Data Management Plan). Besides ethical approval, additional local approval might be necessary to obtain, e.g., a dialogue and risk analysis with legal officers on personal data assessment and data protection was executed by one submitting organization as well as the director from the pathology department permitted an approval for the data extraction. These additional activities differ among organizations as well as countries and each submitter need to have a dialogue with the appropriate local bodies to get advice on what is needed. When requesting access to data, the local requirements need to be met and as a practical example from one of the submitters this was met by two connected activities. A general description of image data withdrawal was published, and a specific approval and documentation for the intended data extraction was applied based on this description.

Thirdly, protocols and procedures for anonymization need to be adapted by each data contributor as described in D5.03, Report describing the Bigpicture pseudonymisation strategy. Figure 5 gives an example of how such anonymization or pseudonymisation strategy was defined referring to each metadata field.

Lastly, procedures for quality control were executed. This was performed through a combination of automatic script and manual (visual) inspection of some of the data (visual inspection of approximately 5% of the data). There was no detailed data storage kept at the site after the submission of the dataset. Only logs of data extraction related to search conditions (diagnosis, specimen area and time period) were saved to avoid repeated data extraction for Bigpicture purpose.

Data parameter	Source	Value and format	Anonymize by
Laboratory ID	LIS	NN12345-12	Delete
Age	LIS	years, 12	Stratify 0-4, 5-9, 10-14…
Date of request	LIS	yyyy-mm-dd, 2022-02-20	Include years only. Must have >10 instance of the specific diagnosis and anatomical localisation

Figure 5 Example of structure of an anonymization strategy

Defined selection and parameters

The datasets need to meet applicable local search criteria and to be designed for the specific Node area. In addition, datasets had to adhere to the developed Common Mandatory Metadata Structure (CMMS) and contain metadata as defined within the Mandatory Submission Metadata (MSM) for direct accessible data sets (MSMdad) within Bigpicture. The submitted datasets in MS8 were based on the approved MSMdad v0.1.0 and CMMS v0.1.0.

All nodes have in different ways been engaged in clinical node coordinator meetings. Three nodes (skin, renal and cancer) had the opportunity to apply the first versions of extraction and conversion tools to their systems with the aim of producing anonymized datasets. Due to the complexity and diversity of information relating to a disease (pathological findings, diagnoses, patterns, classifications, ontologies and standards) disease information was at this point excluded in these datasets. This exclusion encompassed metadata references on all levels (case, slide and image). Nevertheless, the selection criteria for the specific cohorts that defined a dataset were selected from each LIS/LIMS on a combination of at least: 1) a selected coded diagnosis/disease (observation) by a pathologist per case, and 2) a defined time period. The work on observational metadata is continuing in order to establish a data and metadata format for observational information within the Bigpicture project led by the TF metadata and WP3.

Extraction

Extraction was done either manually or semiautomatically, and for each submitter there is a description in the appendix. Since manual extraction is tedious and resource-intensive, innovation on different automatic search tools was explored. In relation to the information structure of the data in the LIS/LIMS or PACS systems, one of the databases, RO's LIS, was queried directly in the test system. Exploration was performed manually, and tracing profiles and results from queries were explored to identify the required data fields according to MSMdad. Subsequently, data were mapped in the local database to attributes from the MSMdad. Depending on the local data system, different tracing server profiles were used to trace user transactions in the database to find the required data in the local database. A verification process was performed between the personnel who created the annotation (the metadata file) and a domain expert. Lessons learned was that RO's LIS could serve as a model for direct queries by external python application. This was tested on the test system. For the submitted dataset (using the production system) the queries were run 'manually'. The search queries developed by RO were later used in a Python application developed by SECTRA, which will be further fitted for local application by other users who can adapt such search queries.

Curate data

There are two tools in Bigpicture applicable for curation of selected data, Wsidicomizer and the Bigpicture metadata interface. The background for Wsidicomizer as described in report D4.03, on unified open digital slide and annotation format specification, is that the python tool Wsidicom was developed as part of Task 4.5 to serve as a reference implementation of a DICOM WSI reader. Wsidicom builds upon the more general python tool Pydicom, with added functionality to enable random frame access, to treat a set of DICOM WSI files as a pyramid, and to interface with WSI metadata. Related to Wsidicom, WP4 developed two tools, Opentile and Wsidicomizer, for reading other WSI formats and convert them to DICOM. Opentile is a high-level interface for reading WSI tiles from tiff-based WSI formats with well-defined tile organization. Wsidicomizer can use a tile-producing object, for example a file opened with Opentile, and save the content as DICOM. The Wsidicomizer tool had not been fully developed at this stage of the project though it will be finalized in D4.4. The format for digital slides and annotation is combined with the format for metadata. The extracted WSI data should therefore be converted into the unified open digital slide and annotation format specification applying the Wsidicomizer leading to anonymization and DICOM conversion. Metadata mapping to Bigpicture format and anonymization was handled by the Bigpicture metadata interface, as the interface defines required and optional metadata and replaces all identifiers with randomized identifiers and no other personal data information was introduced into the metadata. Image data was anonymized by the DICOM conversion process, as no personal data information present in the original image files were carried over into the new files.

Image quality control reference was applied from a standpoint that approved quality for diagnostic should be acceptable for Bigpicture repository. For Utrecht and RÖ who have implemented digital pathology and have a digital archive of WSIs that have been the background for primary diagnostic, we anticipated that the image quality of these WSI would be acceptable for a pathologist. Some variation of image quality is permitted and part of the clinical reality that comes along when implementing a large-scale digitization process of glass slides. Despite the variation in image quality, there is a broad acceptance in the overall quality of digitized glass slides, where multiple validation studies have shown that there is a high concordance in primary digital diagnostics compared to glass slide diagnostic.
Quality control of the result of the automatic extraction process was executed before submission in collaboration between a system developer and a pathologist verifying that only the defined metadata fields and images were extracted according to the selection criteria, and that no strings of personal data information were present in the Bigpicture metadata files nor WSI.

Export/submit data

Encryption and data transfer
The data was encrypted and submitted to the Bigpicture portal using the sda_cli-tool developed by NBIS (WP2). This tool is referenced in D2.04 and the command line user interface for the Sensitive Data Archive is found in https://github.com/NBISweden/sda-cli/.

Discussion

According to the description of the milestone in the DoA this deliverable has been fulfilled, as there are now submitted prototypes of clinical pathology WSI datasets fulfilling the Bigpicture metadata criteria in the Bigpicture repository. However, as described in the DoA, the metadata criteria should be part of the honest broker mechanism (HB), and according to the time plan for the HB mechanism pilot testing and operation such a pilot has not yet been established. Although metadata criteria are to be aligned to support HB functions, a great majority of the metadata criteria are independent of HB. In MS8, the goal was to submit datasets to the BP repository and HB functions are not required at this point. A consequence is that the submitted datasets (the prototypes) contain anonymized data, with an addendum of scrambled metadata fields to simulate a realistic dataset for testing technical conditions of submission and further on access but also to protect backtracking on anonymized data to personal data. These precautionary measures were seen as necessary and the action was decided by each data owner.

Furthermore, the report D3.03, “Report on defined and first collected datasets” is described in the DoA relating to the results from the tasks 3.2, 3.6 and 3.7. 3.2 (Prioritization of datasets) and 3.6 (Collection of clinical data sets) are fulfilled since the three currently submitted data sets are good demonstrator cases from three different nodes. These demonstrate a proposed submission workflow including the implementation of the developed tools for adapting to the developed Bigpicture open format. Task 3.7 however is probably an incorrect typing in the DoA since task 3.7, Collection of nonclinical data sets, is planned in the period of month 18 - 72 which is a time period outside the activity of MS8 and the due date of this deliverable (month18).

Lessons learned from the creation of these datasets are that work on defining data search criteria and curation of data needs resources from system developers as well as pathologists to meet semi-automation and scalability of data detection. The development of the open source search queries are promising including the Wsidicomizer and the Bigpicture interface. Another lesson learned is that slow upload speeds during the upload process are a risk, as a speed of 1.5-2 Mbps hinders data transfer of digital pathology image files: one image can easily consist of 1-2 GB, which will require about 15 minutes in time for upload. This will not be acceptable for submitters who intend to thousands of images. Also, the longer upload time the bigger the risk for disconnection. Submitters need to investigate the limitations of the technical infrastructure on site. Data traffic from a local server to the cloud is a complex process and further work resources on this aspect need to be discussed. On the other hand, developers in WP2 experienced no problems at their end and their service should easily allow for upload speeds > 100 Mbps (provided there is no bottleneck on the users’ end).

Conclusion

The main objective of this report is to describe the activities and procedures performed for submission of the first defined datasets fulfilling the Bigpicture Metadata Model. These datasets fulfill the Bigpicture Metadata Model in terms of being structured as defined in the Common Mandatory Metadata Schema (CMMS v.0.1.0), contained metadata adhering to the Mandatory Submission Metadata for Directly Accessible Datasets (MSMdad v0.1.0) and were transmitted to the repository by means of the developed flexible metadata storage format. The aim of the description of the five steps is to give a guidance of actions needed for future dataset contributors. Explicitly we hope to give guidance in the steps described as well as to define these as selection and parameters, extraction, curate data and export/submit data. Furthermore, we hope to promote implementation of Wsidicomizer and the Bigpicture metadata interface in order to fulfill curation of datasets to Bigpicture.
By performing the steps described above, we have been able to meet Milestone 8, and submitted the first clinical data sets via the data submission portal to the Bigpicture repository by June 2022. Future steps that will be taken by following submitters are under development to learn from experiences from new data submitters as well as to meet expectations of more efficient and simplified extraction, curation and export processes.

That being said, the success of creation and submission of the three described prototype datasets will be of great value for different outlined tasks. One task is the need of prototyped datasets by

WP2 for reiteration of design and function of a secure processor environment. Another task is for WP3 to further develop the honest broker mechanism for more efficient submission and access. WP3 also needs these prototypes to collaborate with WP4 and evaluate access and reader friendly environment in Cytomine of datasets. Also, the prototypes can be used to further develop, modify and give experiences to guide future decisions on metadata format within TF metadata to meet the need of AI developers and pathology users to enhance clinical knowledge of AI in digital pathology and to meet the vision how Bigpicture will become the catalyst in digital transformation in pathology.

Appendix

Specific results from MUW: Define selection and parameters

WSIs of biopsies of native and transplanted kidneys as well as surgically resected tissue are generated for research purposes (i.e., telemedicine conferences, morphometry studies, Bigpicture) and stored in a redundant, access-controlled local storage. Cases with insufficient material for diagnosis (<5 glomeruli, no arteries) were excluded. Two random cases were selected for submission to Bigpicture.

Extraction

At MUW, pathology reports are stored in a dedicated LIS (PAS Xanthos, Programmierfabrik GmbH, Linz, Austria), which does not support direct access to its database. Using an export feature, data were extracted as lists and re-combined into a database, ready for querying and assembly of a Bigpicture dataset.

Patient-related metadata were taken from internal databases, technical metadata (i.e., scan resolution) were read from WSI file. Both source data were converted to XML representations following the Bigpicture schema using Bigpicture_metadata_interface. A considerable amount of time (approx. 40 hours) was spent on understanding and optimizing this process.

Curate data

A dataset of kidney biopsies (2 cases, each containing approx. 10 slides of different stainings) was assembled. Patient identifiers in images and metadata were removed. WSIs were converted to DICOM using Pydicom.

All images and metadata were encrypted via Crypt4GH tool and Bigpicture public key.
Final quality control was done before submission (number of files match expected count, checking for non-encrypted data).

Extract data according to Bigpicture MSMdad format

Encrypted WSIs (in DICOM format) and metadata (in Bigpicture xml format) were uploaded using sda-cli to S3 storage provided by Bigpicture.

Specific results from UMCU: Define selection and parameters

We searched in our archive LIS database for cases that arrived after 2018, with ‘invasive carcinoma’ as a diagnosis term and type of material as ‘breast’. That resulted in a large amount of cases (>500). This dataset was then checked for against our database for patients that have opted out of research and those cases were removed.

Since the environment that would host the data is still in test/beta phase and this particular set would not be used by the project, we opted for an extra step in the process by scrambling the complete dataset. Every column was scrambled/randomized which would generate a dataset that resembled the structure of the actual exported cases without actually having to submit real patient data. After the scrambling, we exported a small subset of cases (N=24) for submitting to the project. We then exported the lab metadata for that subset (slide staining, specimen and block information).

Curation

We followed a similar approach as described in the general section concerning the metadata conversion, but we did not map the stains to individual constituent stains, we opted to submit the name of the stain as is. That would require to make a list of stains and identify all of the individual stains that comprise a staining. That information does not exist in our LIMS and is only described in laboratory SOPs.
That is only readily available for fluorescent staining. If the individual stains are required to describe a slide staining, then that would make the threshold of submission higher.

Images (in ndpi-format) were converted to DICOM using the open-source Python application Wsidicomizer developed by SECTRA. The conversion was integrated into the data extraction tool. Only pyramid levels and no label or overview images were converted, as the later image types can contain personal data information.

The extracted metadata and images were matched and parsed together into Bigpicture format using a Python application developed by SECTRA. The SECTRA 'Dicomizer' application was adapted to the metadata format of our lab. Specifically, we changed the naming convention used by Dicomizer to import metadata of blocks and specimens from our lab since the Dicomizer was configured to parse that information from the slide id whereas our lab does not store that information in the slide id, it exists in our LIMS.

Export/submit data

The data was encrypted and submitted to the Bigpicture portal using the sda_cli-tool developed by NBIS (WP2). The encryption and upload process took quite a long time to complete. The encryption process relies on the local infrastructure to run so running the process on an average workstation resulted in the process taking way longer than it needed to. A lesson learned is that the data encryption should run on workstations that have access to storage systems with high IOPS and read/write speeds as well as ample CPU resources. Further, the upload process took also quite long which means that the network connection to the NBIS was somehow limited.

Specific results from RO:

Define selection and parameters

We started by searching retrospectively for suitable skin tumour diagnosis, initially we started with cases coded as invasive melanoma. The selection criteria were defined by a limited time interval in combination with the local diagnosis (M type SNOMED code) in our Laboratory Information System (LIS).
Since we wanted to create a melanoma dataset, we started developing search queries that searched for types of referral (in our setting this is named as "PL"), the time interval was set as year (2020) and the local diagnose code was (M87203 with definition "melanoma"). If a case was deemed not suitable for upload (e.g., a unique case and hence risk of recognition by persons related to the specific observation) another case was selected instead. Exclusion criteria for avoiding unique case were e.g., a second opinion had been performed on that specific case. The result rendered more than one hundred cases, and as we applied an explorative working strategy for application of different tools for each subsequent task, we narrowed the search result even further to 8 cases containing ~80 slides in total.

Extraction

Region Ostergotlands (RO) first challenge was to find an efficient method for extracting metadata from the LIS. As the LIS only allowed limited functionality for exporting metadata, effort was put into deciphering where the actual metadata resides in LIS database in order to enable direct extraction of metadata from the database. Once we knew where the data could be found and their linkage, we developed search queries that would extract relevant metadata given an input case number. For this limited extraction the queries were run manually against the production LIS database, but an automatic extraction method was also developed and tested against the test LIS database. The search queries developed in the project can be of great use for future research projects since we have acquired a greater understanding of the pathological clinical data in RO LIS.

Curate data

A major step in the curation process was how to map coded and non-coded attributes found in the extracted LIS metadata to coded attributes in the Bigpicture format. To efficiently do this, a mapper concept was introduced into the Bigpicture metadata interface, which allows mappable attributes to be assigned to metadata objects (for example staining for a slide). Mappers would then be created by adding expressions and corresponding results, for example 'H&E' -> 'Stain('haematoxylin'), Stain('eosin')'. When the mapper was applied to the objects, the mappable attributes would be converted into their target attributes, i.e., the string 'H&E' expanded to a Staining-attribute containing the haematoxylin and eosin stains.

Images (in ndpi-format) were converted lossless to DICOM using the open-source python application Wsidicomizer developed by SECTRA. The conversion was integrated into the data extraction tool. Only pyramid levels and no label or overview images were converted, as the later image types can contain personal data information. Thumbnail images were created for each converted image to enable fast and convenient checking of the conversion result.

Metadata anonymization was handled by the Bigpicture metadata interface, as the interface replaces all identifiers with randomized identifiers and no other personal data information was introduced into the metadata. Image data was anonymized by the DICOM conversion process, as any personal data information present in the original image files were not carried over into the new files.

The extracted metadata and images were matched and parsed together into Bigpicture format using the Python application developed by SECTRA as described above. The metadata and image interfaces for the application was later abstracted and shared internally within Bigpicture to allow the tool to be usable at other nodes.

Export/submit data

The data was encrypted and submitted to the Bigpicture portal using the sda_cli-tool developed by NBIS (WP2). Since we were two developers working with encryption of the data, we encountered a minor communication error; one of us used the wrong encryption key on a subset of the data that was uploaded to the repository which led to us having to upload the data again. This error could easily have been avoided if one person assumes full responsibility from data encryption to upload.

RO experienced slow upload speeds (1.5-2 Mbps) during the upload process, it is unclear if this was due to limitations of the infrastructure on site or a temporary problem somewhere else. Developers in WP2 experienced no problems at their end and their service should easily allow for upload speeds > 100 Mbps (provided there is no bottlenecks on the users end).