Anonymization of Whole Slide Images in Histopathology for Research and Education

Tom Bisson[1], Michael Franz[1], Isil Dogan O[1], Daniel Romberg[2], Christoph Jansen[1], Peter Hufnagl[1], Norman Zerbe[1]

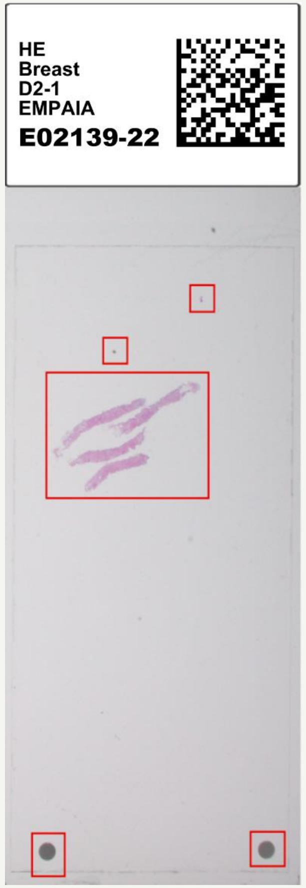
1 Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Pathology, Germany

2 Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

Federal Ministry
for Economic Affairs
and Climate Action

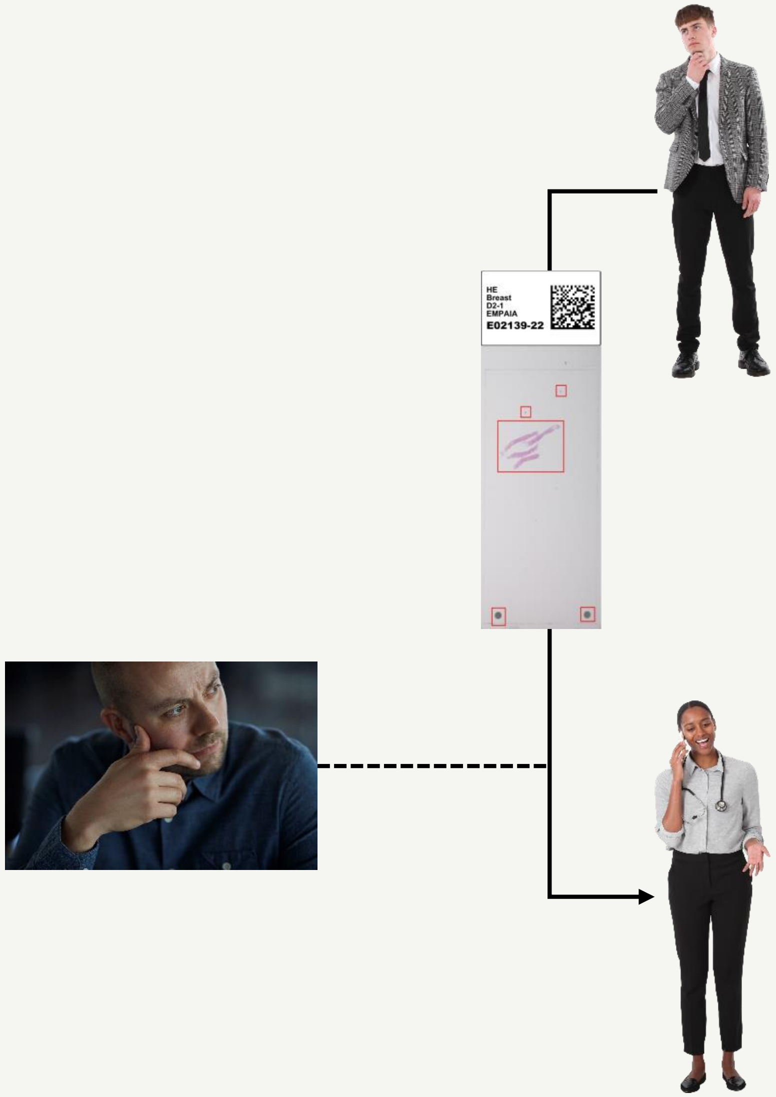
Motivation

- The fragmented scanner device market introduces a variety of proprietary WSI file formats, which differ significantly
- Depending on the specific format, sensitive information may be stored in different locations and extents
- Typically, glass slides contain a label with the case ID, in some cases additionally coded as a 1D or 2D barcode
- During scanning, the label is always captured and stored as an image file in the whole slide image (WSI)
- Additionally, some scanners digitize slide labels and store the contained data within the file, among other acquisition-related metadata



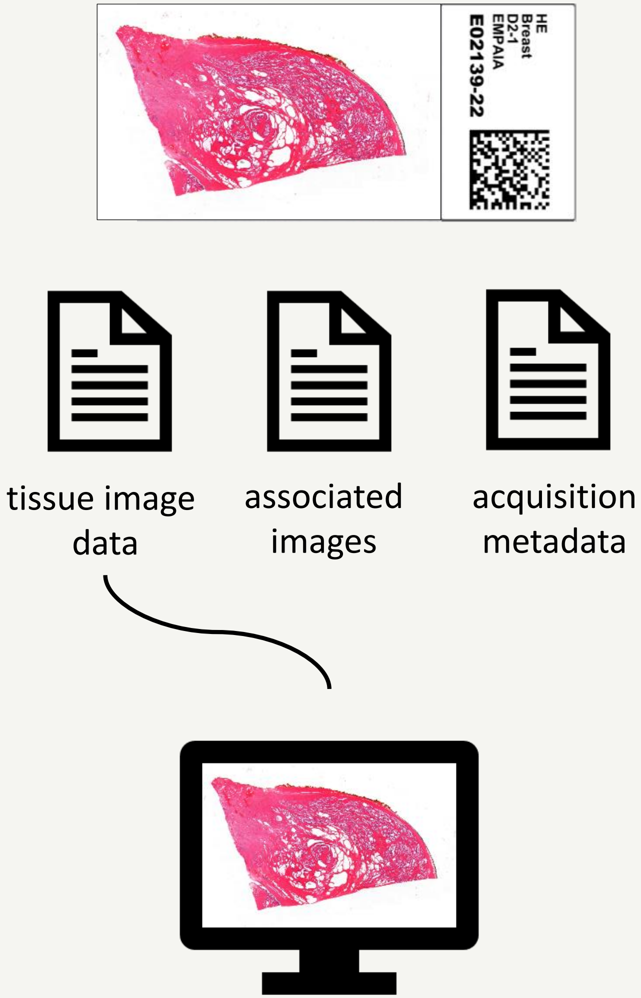
Motivation

- when exchanging WSIs, is the partner receiving the data permitted to access this sensitive information?
  - billing for second opinions requires this data
  - little need for the sensitive information stored in WSIs for educational purposes or research contributions
- transmission channel and networking security also have a major impact on data protection



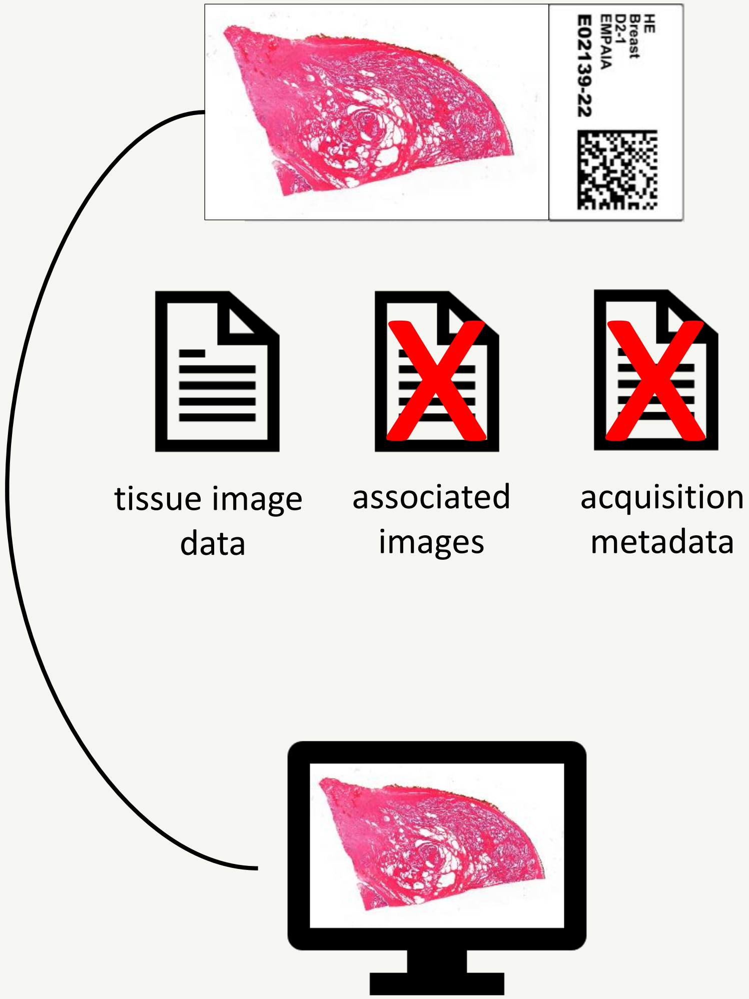
Technical Approaches

- **Separate Tissue Image Data from identifying information**
  - Needs an interface for accessing the data
- Remove all identifying information from the Whole Slide Image
  - Case ID cannot be retrieved
  - No access control needs to be implemented
- Both are only effective when there is no additional information in the scanned region (text, pen markings or even parts of the label)



Technical Approaches

- Separate Tissue Image Data from identifying information
  - Needs an interface for accessing the data
- **Remove all identifying information from the Whole Slide Image**
  - Case ID cannot be retrieved
  - No access control needs to be implemented
- Both are only effective when there is no additional information in the scanned region (text, pen markings or even parts of the label)



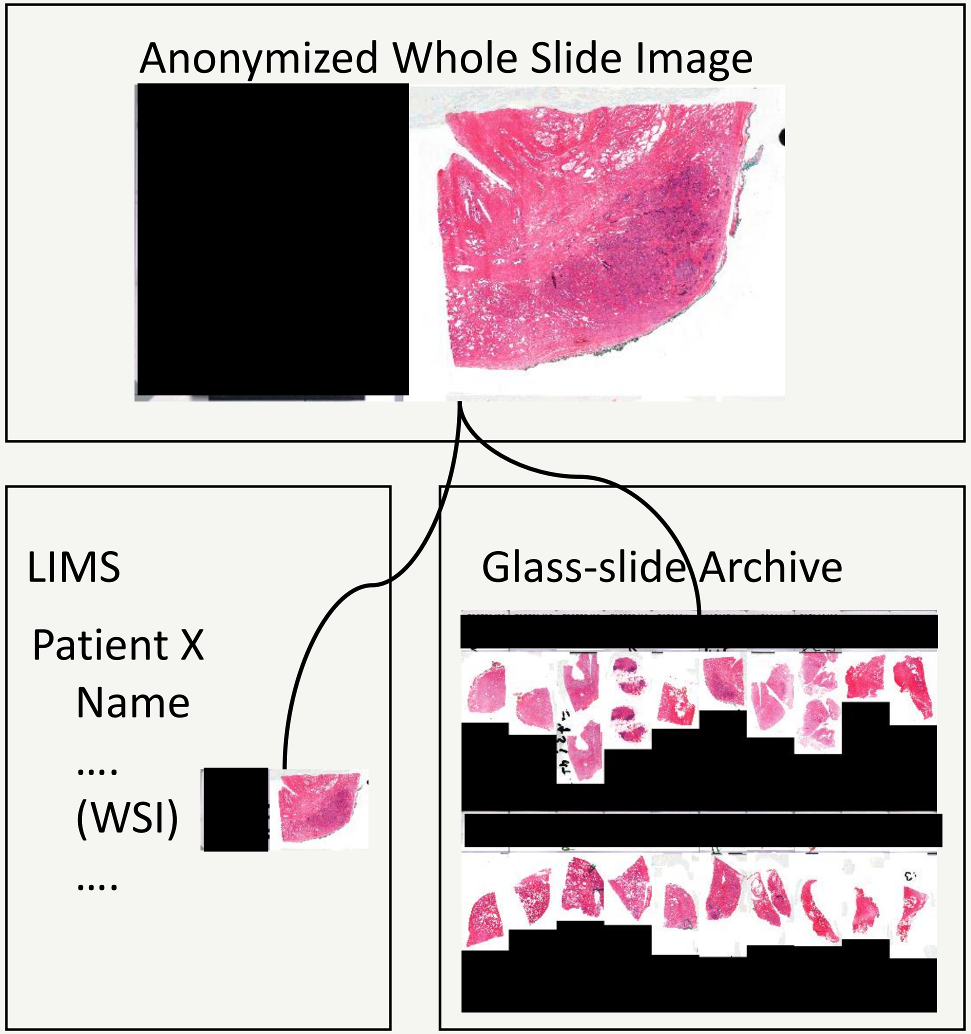
Regulatory

- In Europe: GDPR as the legal framework
- No specific guidelines
- Instead: protection of personal data must ensure a „reasonable“ level (open to interpretation)
- rules for the processing, storing, and sharing of any data identifying a natural person
- Health related and medical data are not specifically addressed

commons.wikimedia.org

„Anonymization“

- digitized tissue itself can serve as an identifier if both the original tissue and its association with the patient are accessible
- GDPR already considers data anonymous if it can only be traced back to specific individuals with a disproportionate effort
- From a technical POV, we can only achieve a “de facto” anonymization (Holub et al., 2023)
- Extensive manual effort would theoretically allow the particular case or patient to be identified, if the clinical system or the original slides are accessible
- But: if such access exists, matching does not add any value, since it is already possible to view the electronic patient record
- Thus, if the case ID can no longer be extracted from the WSI, the effort of retrieving patient information may be considered disproportionately large and thus the use of the data may be considered to be covered by the GDPR



Anonymized Whole Slide Image

LIMS

Patient X

Name

....

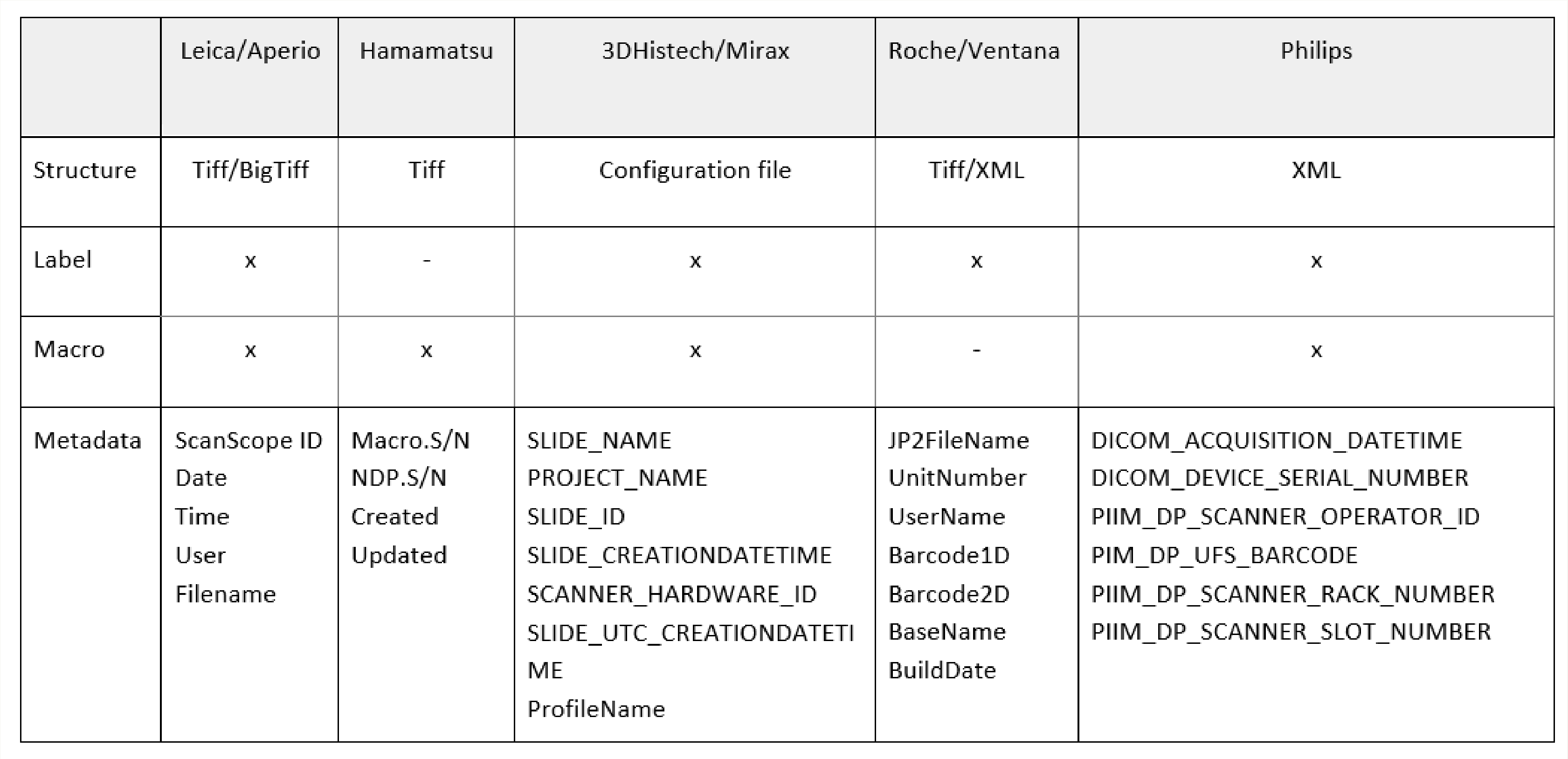
(WSI)

....

Glass-slide Archive

Metadata and associated images

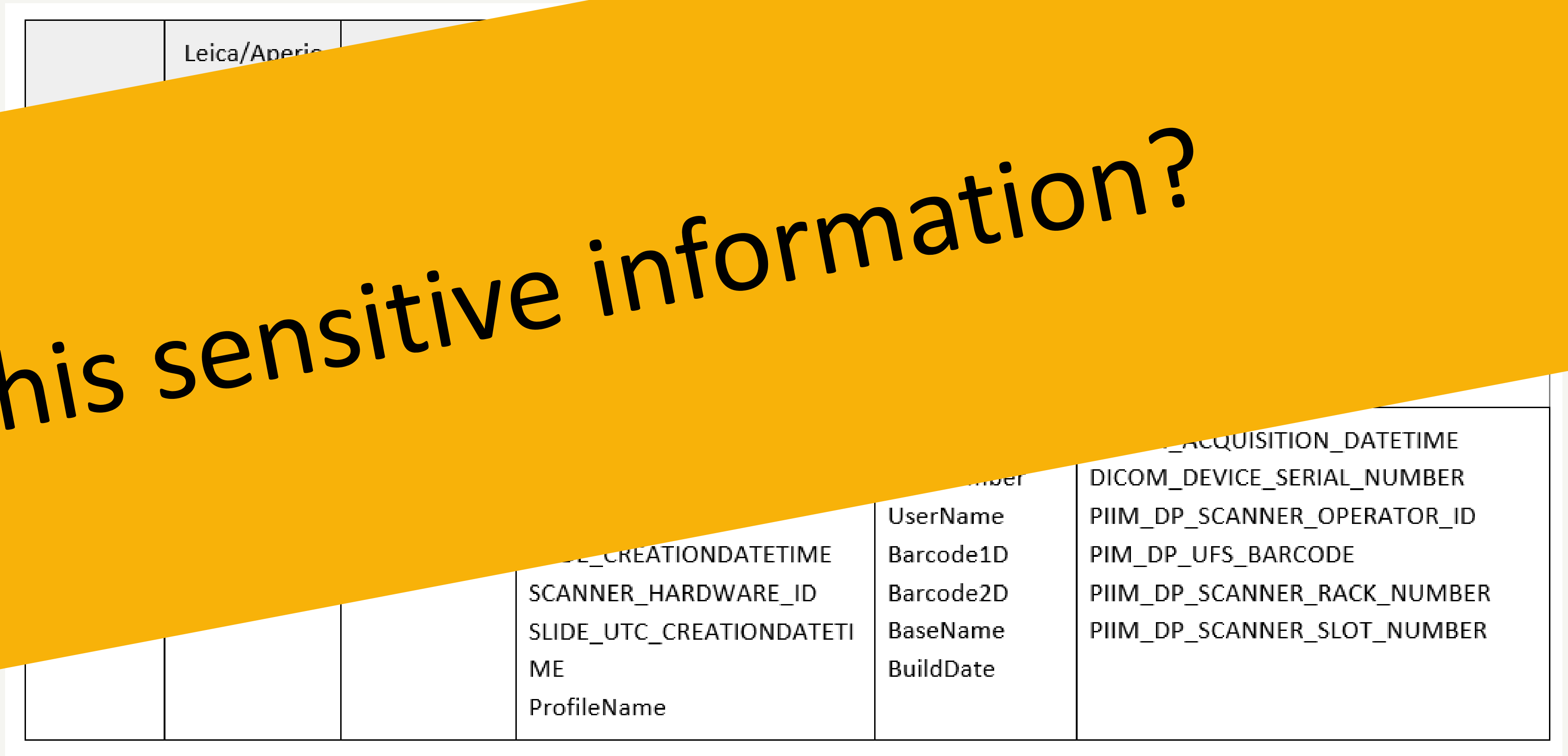
- Associated Images
  - Macro image (sometimes including the label)
  - Label image
- Barcodes and text can be digitized and stored inside the WSI
- **Acquisition-related metadata**



	Leica/Aperio	Hamamatsu	3DHistech/Mirax	Roche/Ventana	Philips
Structure	Tiff/BigTiff	Tiff	Configuration file	Tiff/XML	XML
Label	x	-	x	x	x
Macro	x	x	x	-	x
Metadata	ScanScope ID Date Time User Filename	Macro.S/N NDP.S/N Created Updated	SLIDE_NAME PROJECT_NAME SLIDE_ID SLIDE_CREATIONDATETIME SCANNER_HARDWARE_ID SLIDE_UTC_CREATIONDATETIME ProfileName	JP2FileName UnitNumber UserName Barcode1D Barcode2D BaseName BuildDate	DICOM_ACQUISITION_DATETIME DICOM_DEVICE_SERIAL_NUMBER PIIM_DP_SCANNER_OPERATOR_ID PIM_DP_UFS_BARCODE PIIM_DP_SCANNER_RACK_NUMBER PIIM_DP_SCANNER_SLOT_NUMBER

Metadata and associated images

- Associated Images
  - Macro image (sometimes in th



	Leica/Aperic				
			_CREATIONDATETIME SCANNER_HARDWARE_ID SLIDE_UTC_CREATIONDATETIME ProfileName	ber UserName Barcode1D Barcode2D BaseName BuildDate	ACQUISITION_DATETIME DICOM_DEVICE_SERIAL_NUMBER PIIM_DP_SCANNER_OPERATOR_ID PIM_DP_UFS_BARCODE PIIM_DP_SCANNER_RACK_NUMBER PIIM_DP_SCANNER_SLOT_NUMBER

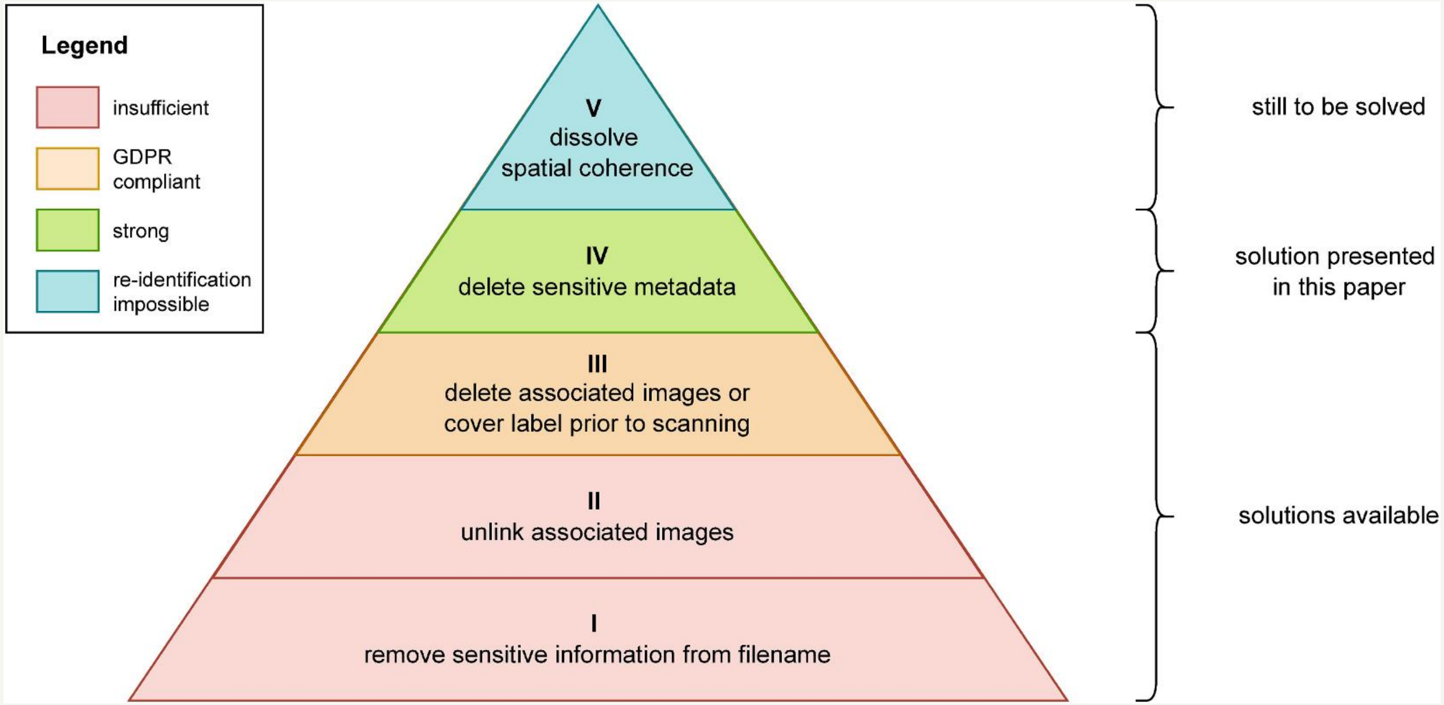
Why is this sensitive information?

Metadata and associated images

- This data is critical on its own, but in combination with other non-critical data may allow the patient to be identified
- A Scenario:
  - WSIs are published as part of an investigation of a rare disease
  - The label images are removed, but associated metadata is still present within the file
  - The same scanner serial number of these WSIs can be found in other available WSIs (e.g. from other studies)
  - If those are single-center studies or teaching materials, the WSI may already be assigned to the institute where the patient was admitted
  - The acquisition date combined with various background information such as study descriptions or appendices can then be used to narrow down the treatment period in more detail
  - Although access to the information system would still be necessary here, making use of the previously mentioned information in combination with the rarity of cases of that disease could reduce the set of eligible patients

Anonymization Policy

- Level IV requires that all sensitive metadata be deleted in addition to the label image
- Levels III and IV can theoretically be achieved with conversion to DICOM or OMERO but require conversion to another file format, which in turn requires support for that additional file format.



Legend

- 🟥 insufficient
- 🟧 GDPR compliant
- 🟩 strong
- 🟦 re-identification impossible

V
dissolve
spatial coherence

IV
delete sensitive metadata

III
delete associated images or
cover label prior to scanning

II
unlink associated images

I
remove sensitive information from filename

still to be solved

solution presented in this paper

solutions available

Solution

- De facto Anonymization can be achieved by converting the WSI to DICOM WSI and subsequent removal of all sensitive information
- But: DICOM is yet only tentatively adopted in digital pathology and not all slide scanner vendors provide anonymization or DICOM conversion
- To close this gap we developed a software library that enables a level IV anonymization of WSIs within the proprietary file formats
- In addition to the C library, we provide wrappers for Python and JavaScript as well as a command line interface tool and show how it can be used in the browser via WebAssembly
- The library is designed to be extended easily to new formats or updated versions of already existing formats

Thank you very much!

tom.bisson@charite.de

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

empaia

Fraunhofer
MEVIS

Fraunhofer Institute for Digital Medicine MEVIS

Federal Ministry
for Economic Affairs
and Climate Action