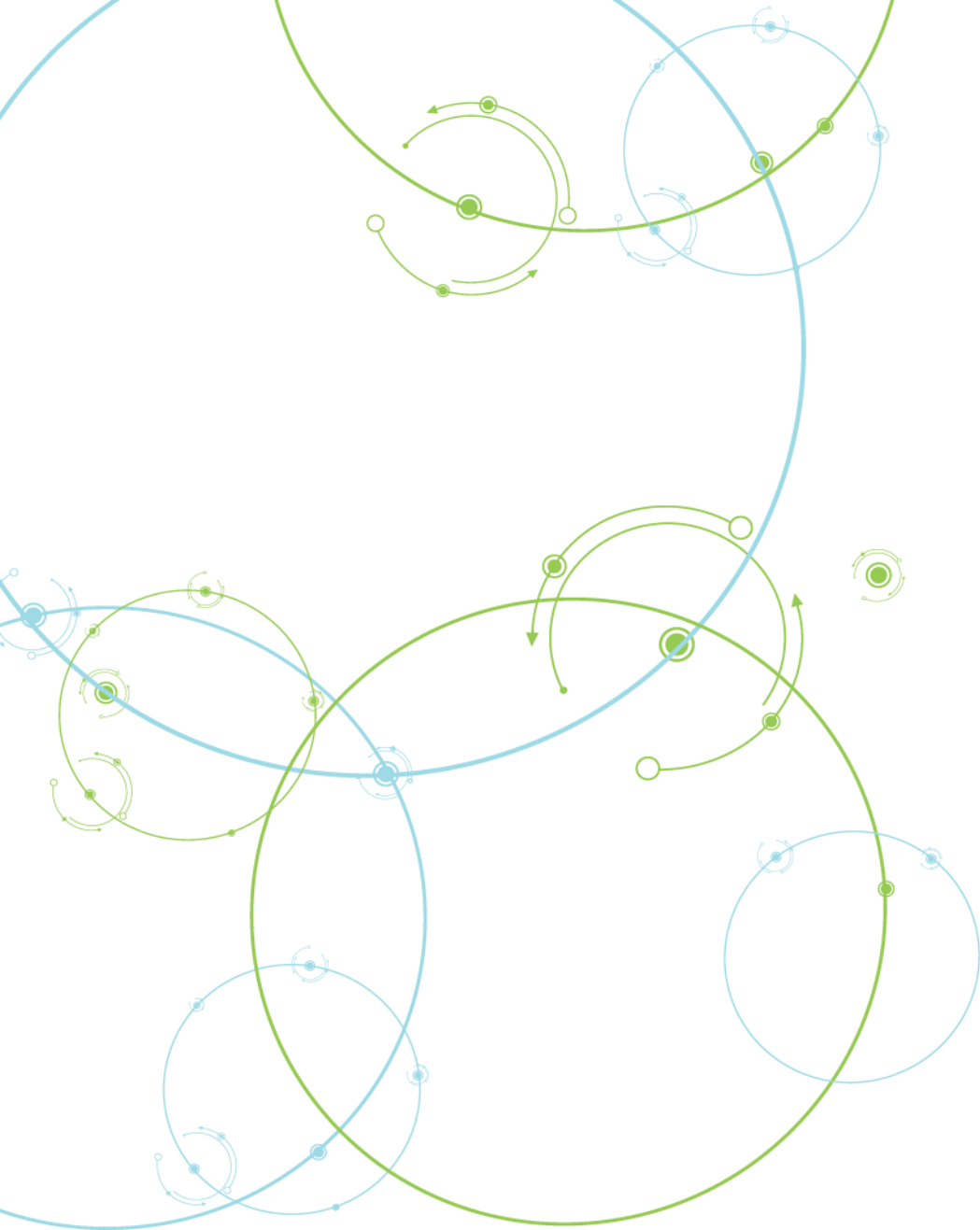


Whole Slide Image data sharing and de-identification, the NHLBI BioData Catalyst® Data Management Core experience

Zixin Nie

RTI International



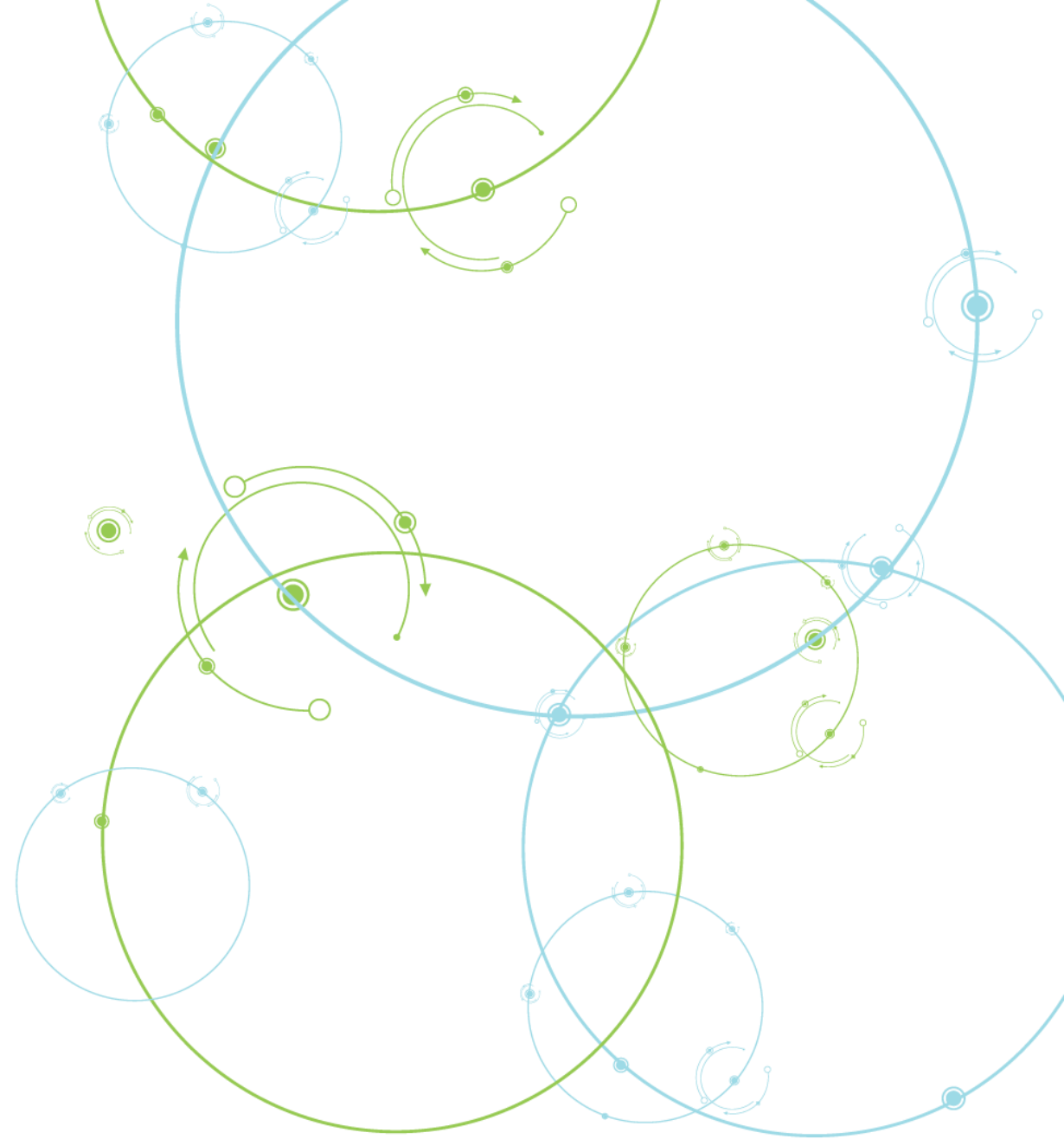


Disclaimer

The views expressed in this presentation do not necessarily reflect the official policies of the U.S. Department of Health and Human Services, the NIH or its components; nor does the inclusion of trade names/logos/trademarks/or references to outside entities constitute or imply an endorsement by any Federal entity.

Contents

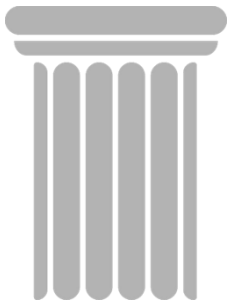
- What is NHLBI BioData Catalyst®
 - Who is the Data Management Core
- Histopathology whole slide images
 - Contents of SVS files
 - Where can PII occur
- Using open-source tools to review SVS files
- Practical tips for reviewing files



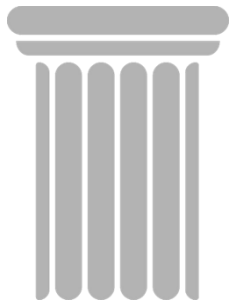
NHLBI BioData Catalyst®

- NHLBI BioData Catalyst® (BDC) is a cloud-based ecosystem funded by the National Heart, Lung, and Blood Institute (NHLBI) that offers researchers data, analytical tools, applications, and workflows in secure workspaces.
 - BDC is a community where researchers can find, access, share, store, and analyze heart, lung, blood, and sleep data.
 - BDC is also an NHLBI data repository where researchers share scientific data from NHLBI-funded research, so they and others can reproduce findings and reuse data to advance science.
- By increasing access to NHLBI data and innovative analytic capabilities, BDC accelerates reproducible biomedical research to drive scientific advances that can help prevent, diagnose, and treat heart, lung, blood, and sleep disorders.

Mission



Vision



- The *mission* is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.
- The *vision* is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.

BioData Catalyst Data Management Core

Enable data generators to ingest their datasets into BDC systems such that high value data is available for research

- ❖ Help ensure the efficient and effective management of data before and during ingest into BDC
- ❖ Identify and prioritize data for onboarding into BDC
- ❖ Standardize data and streamline the data onboarding process

Histopathology Whole Slide Images

- Digitized microscope slides of tissue sections, scanned at high magnification (typically 20× or 40×) to produce multi-gigabyte, multi-resolution images used for pathology research and AI/ML model development.
- WSIs are increasingly contributed alongside clinical, genomic, and phenotypic data to enable secondary research and large-scale analytics in the NHLBI ecosystem.
- Unlike DICOM, WSI vendor formats (SVS, NDPI, MRXS, BIF, etc.) are proprietary, non-standardized, and scanner vendors generally do not ship anonymization functionality — so identifiers land in files that most tools can't safely modify.
- We focus on Aperio SVS, a common vendor format in digital pathology and the one currently in scope for BDC ingest.

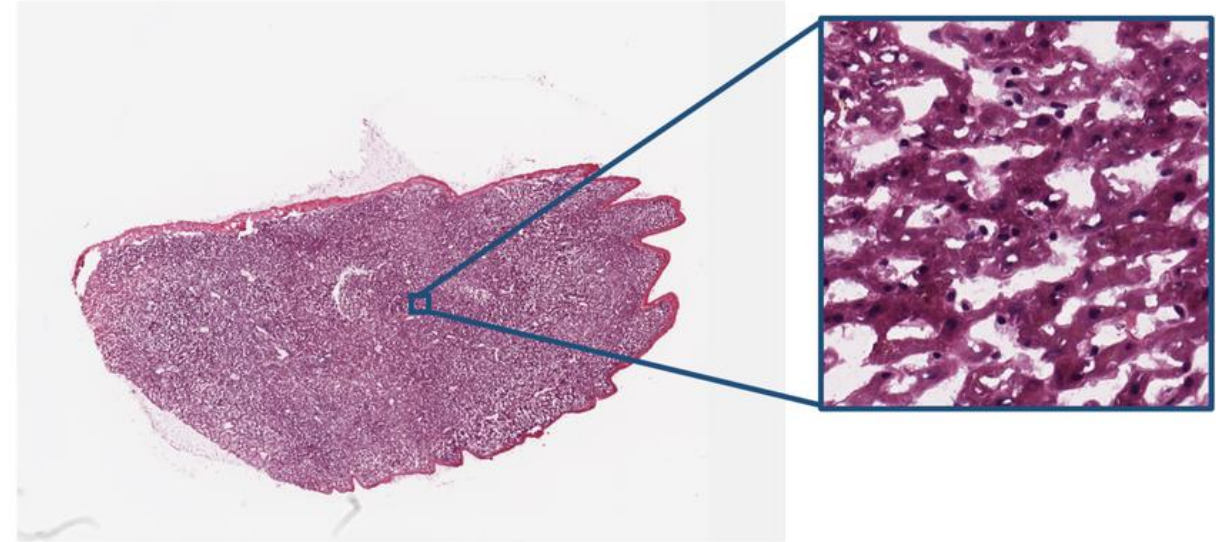


Image Source: The Current Role of Image Compression Standards in Medical Imaging - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/A-digital-pathology-whole-slide-image-The-20-000-14-000-whole-slide-image-is-shown-on_fig3_320511493 [accessed 1 Jul 2026]

Contents of SVS Files

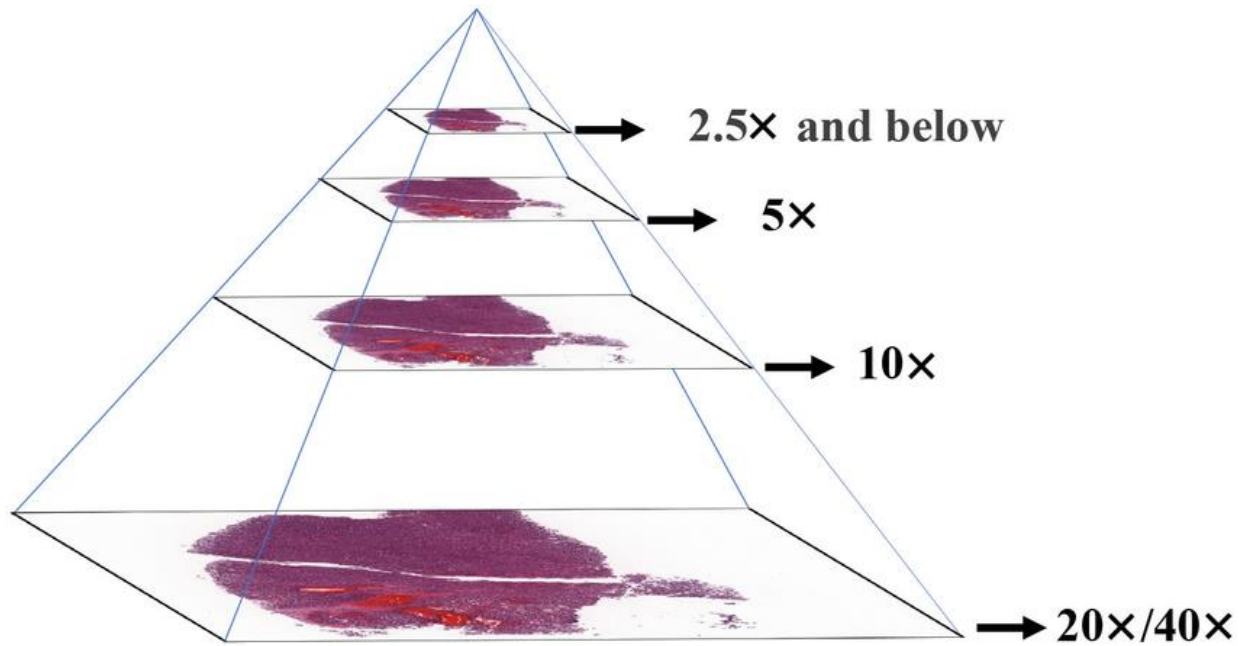


Image Source: Multi-Magnification Image Search in Digital Pathology - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Pyramidal-structures-of-a-whole-slide-image-WSI-in-digital-pathology-The-base_fig1_361233038 [accessed 1 Jul 2026]

- SVS is a multi-page TIFF file with pyramidal, tiled images plus vendor-specific metadata in the ImageDescription tag.
- Typical page layout:
 - Page 0 — Baseline full-resolution image, tiled (usually 240×240 px, JPEG-compressed)
 - Page 1 — Thumbnail (stripped, ~1024×768)
 - Pages 2–N — Downsampled pyramid levels (tiled, same compression as baseline)
 - Optional trailing pages — Label image (NewSubfileType = 1) and Macro image (NewSubfileType = 9)
- Metadata: Stored in the ImageDescription TIFF tag as pipe-delimited key/value pairs. Typical fields include AppMag, MPP (microns per pixel), Date, Time, ScanScope ID, and often free-text comments.

Where can PII Occur in an SVS File

- Metadata (ImageDescription tag):
 - Original scanner IDs, filenames, and file paths
 - Scan Date/Time — a HIPAA identifier when tied to a patient event
 - Free-text User, Comments, or study fields sometimes contain names or MRNs
- Label image: in clinical scans the label commonly shows patient name, MRN, accession number, DOB, and a 1D/2D barcode encoding the same.
- Macro image: Depending on scanner illumination and framing, the macro view can capture the label edge-on and render its text legible.
- Baseline / pyramid tiles: Occasionally handwritten annotations, ink marks with initials/dates, or accession stickers sit on the coverslip and are baked into the tissue pyramid itself. These cannot be scrubbed without destroying pixels.

```
tiff.ImageDescription : Aperio Image Library v10.0.51
46920x33014 [0,100 46000x32914] (256x256) JPEG/RGB
Q=30|AppMag = 20|StripeWidth = 2040|ScanScope ID =
CPAPERIOCS|Filename = CMU-1|Date = 12/29/09|Time =
09:59:15|User = b414003d-95c6-48b0-9369-
8010ed517ba7|Parmset = USM Filter|MPP = 0.4990|Left =
25.691574|Top = 23.449873|LineCameraSkew = -
0.000424|LineAreaXOffset = 0.019265|LineAreaYOffset =
-0.000313|Focus Offset = 0.000000|ImageID =
1004486|OriginalWidth = 46920|Originalheight =
33014|Filtered = 5|ICC Profile = ScanScope v1
```

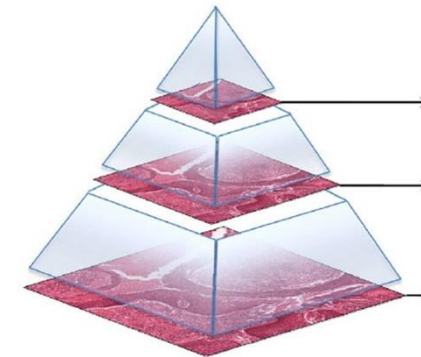
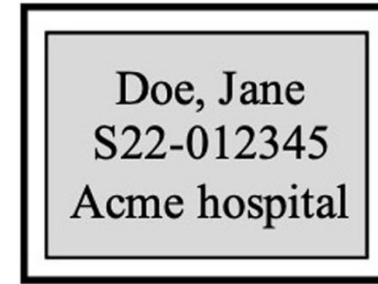
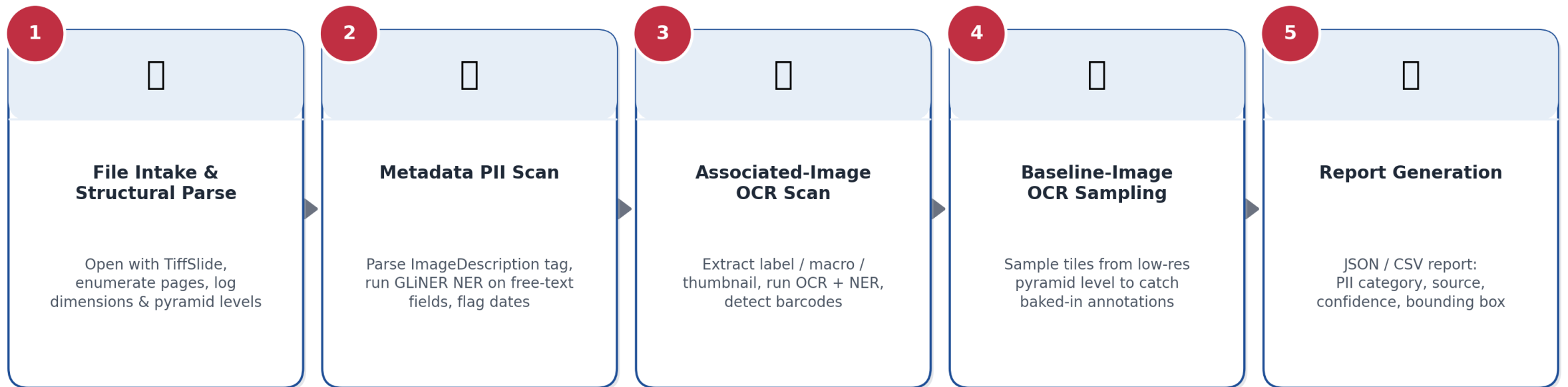


Image Source: Simonson R, Zarella M. 1384 PHI Can Be Present in Whole-slide Images After Conventional Deidentification Methods are Applied. Laboratory Investigation, 105

Using Open-Source Tools for Reviewing SVS Files

- Aperio ImageScope: the vendor viewer; useful as a ground-truth visual reference during QC.
- OpenSlide: C library with Python bindings; the de facto standard for reading Aperio SVS and 12+ other WSI formats. Requires a separate binary install alongside openslide-python.
- TiffSlide: pure-Python, tiff file-based drop-in replacement for openslide-python. Pip-installable with no external binary, works with fsspec for direct reads from S3/GCS.

Automated Pipeline for PII Reviews



Challenges with this Format

- Baseline images routinely exceed $40,000 \times 30,000$ pixels.
 - Many Python image libraries (e.g., OpenCV's `cv2.imread`, PIL's default `MAX_IMAGE_PIXELS`) refuse to load images past a safety threshold.
 - Full decompression can easily exceed 10 GB of RAM.
- Because the baseline is stored as $\sim 240 \times 240$ px tiles, any PII text baked into the tissue image can be split across tiles
 - Prevents OCR unless you overlap or reassemble regions.
- SVS uses non-standard TIFF tags. Naively rewriting tags with generic TIFF libraries frequently produces files that Aperio ImageScope or OpenSlide can no longer open.
- The full SVS specification is not publicly redistributable, and Leica/Aperio scanners don't ship a de-identification workflow.
- DICOM WSI is coming but not here yet: Industry adoption of DICOM for pathology is still in a tentative state.

Tools for Performing SVS File De-Identification

- Proprietary
 - **John Snow Labs — Medical LLM / Visual NLP de-ID**
 - Commercial pipeline that combines OCR + medical NER with format-preserving output
- Open Source
 - **svs-deidentifier (pearcetm)**
 - GUI application (Chrome/QWebEngine) that strips label and macro images from Aperio SVS files.
 - Supports "Copy" mode (safe, produces a new file) and "Modify" mode (in-place, faster but destructive).
 - Last meaningful update ~2 years ago but still functional and MIT-licensed.
 - **wsi-anon / WSI-anon (Fraunhofer MEVIS / Charité)**
 - C library with CLI + Python wrappers;
 - multi-vendor (Aperio, Hamamatsu, MIRAX, Ventana, Philips) with format-preserving GDPR-compliant anonymization.
 - Published in Digital Health (2023)
 - **anonymize-slide (bgilbert) / AnonymizeSlide fork (DeepMicroscopy)**
 - the original "Anonymyslide"-lineage script.
 - Deletes labels from SVS/NDPI/MRXS
 - Last updated in 2014, still requires Python 2.6/2.7, and is unmaintained.

Practical Tips

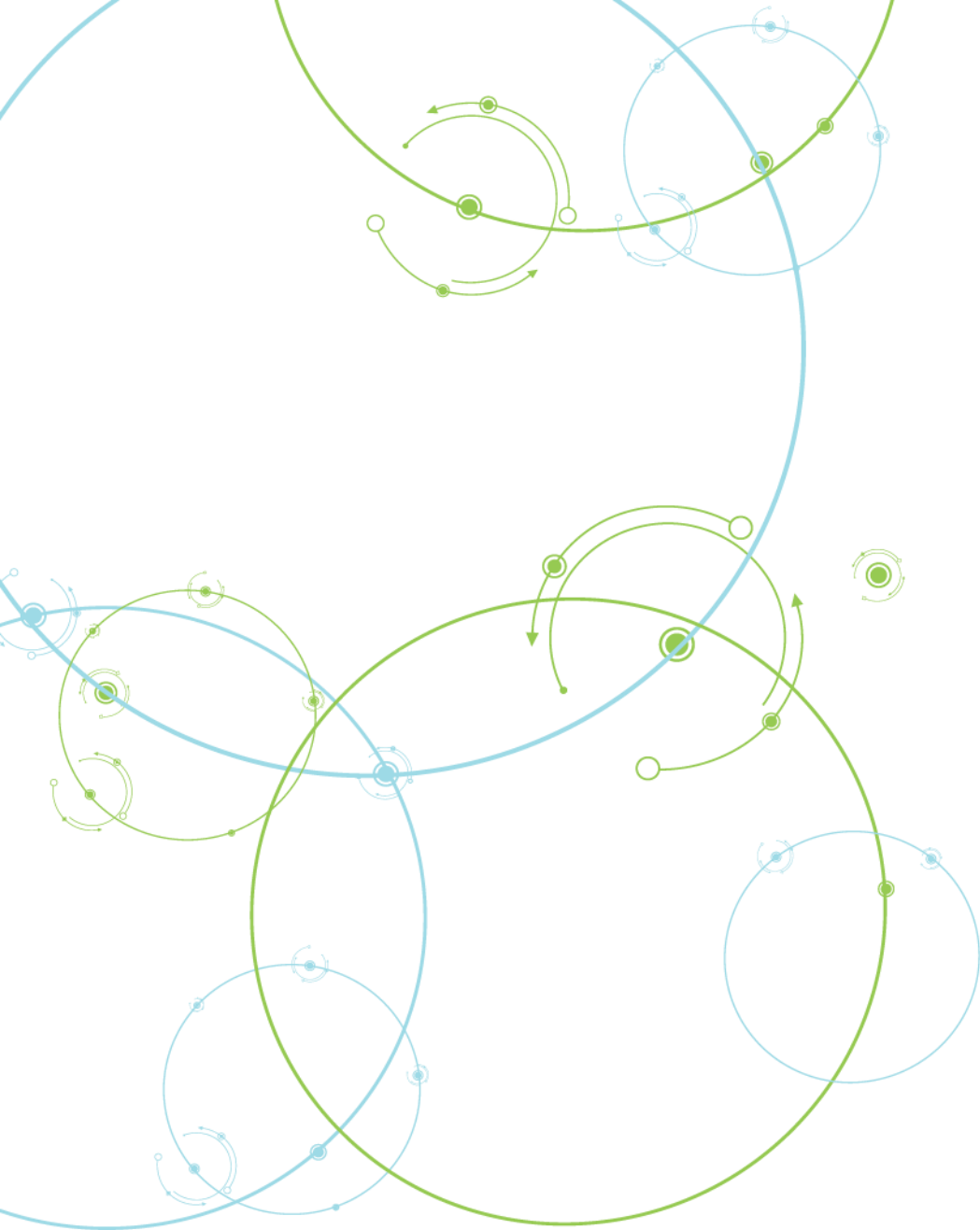
- **De-identify at the source.** The most effective control is stopping identifiers from ever entering the scanner — e.g., use de-identified accession labels on the physical slide before scanning.
- **Always strip label + macro before sharing.** This addresses the vast majority of PII risk with minimal file impact.
- **Scrub the metadata explicitly.** After label/macro removal, inspect ImageDescription for original IDs, scanner serials, dates, and free-text Comments fields.
- **Preserve, don't drop, useful acquisition metadata.** AppMag, MPP, ICC profile, and scanner model are important for downstream analysis and are not PII.
- **Verify with a viewer after de-ID.** Round-trip every de-identified file through OpenSlide/TiffSlide and ImageScope to confirm (a) the file still opens, (b) label/macro no longer render, and (c) metadata scrub took effect.
- **Keep the original offline.** In-place ("Modify") tooling is fast but irreversible; mirror originals to a controlled-access location before any batch run.

Conclusion

- SVS is a BigTIFF-based, proprietary, pyramidal format with three well-defined PII surfaces: metadata, label/macro, and (rarely) tissue tiles.
- Reading and inspecting SVS is well-supported by OpenSlide and TiffSlide.
- Automated PII detection is feasible by combining metadata parsing + GLiNER NER + OCR on associated images, but is bounded by tile size, memory limits, and OCR recall on baked-in text.
- Automated modification remains the harder problem. Only a handful of maintained tools (svs-deidentifier, wsi-anon) safely rewrite SVS in place, and none cover every edge case.

Example Datasets

- OpenSlide public test data — <https://openslide.cs.cmu.edu/download/openslide-testdata/> — canonical set of Aperio, Hamamatsu, Leica, MIRAX, and Ventana samples used across the community



Acknowledgements

- The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst® (BDC) ecosystem.
- I would like to thank the members of the BDC DMC from RTI International for their support for this presentation:
Christopher Siege and Oswaldo Loyola
- I would also like to thank the NHLBI BDC leadership and Communications Team for helping craft the content of this presentation.



Thank you

Contact: Zixin Nie | email: zixinnie@rti.org