

# Open-Source, Cross-Platform Drag and Drop Pipeline for Real-Time Whole-Slide Image Anonymization

André Lametti<sup>1</sup>

<sup>1</sup>Department of Pathology, McGill University. Disclosures: none.

## Background and problem description

The need to remove identifying information from medical images for non-clinical uses is well-established. Digital pathology slides are used extensively for research and teaching, but contain slide labels and other sensitive data. Our patients expect us to treat personally identifying information very carefully; this is especially true for research use in the era of artificial intelligence, where research ethics boards require robust and detailed data stewardship practices.

In practice, there are several problems to large-scale or continuous digital slide anonymization (fig. 1). Proprietary solutions provided by digital pathology systems are usually not flexible, conversion to other file formats is compute-intensive, and most other solutions require use of a command-line interface.

One excellent available solution is the wsi-anon library from the EMPAIA consortium, which is fast, has few dependencies, and is extensively compatible with slide scanner vendors [1]. However, additional features are required for routine use by non-expert end users. We propose a solution to some of these challenges.

Figure 1: Problems faced by end users requiring mass anonymization of digital pathology slides for non-clinical purposes

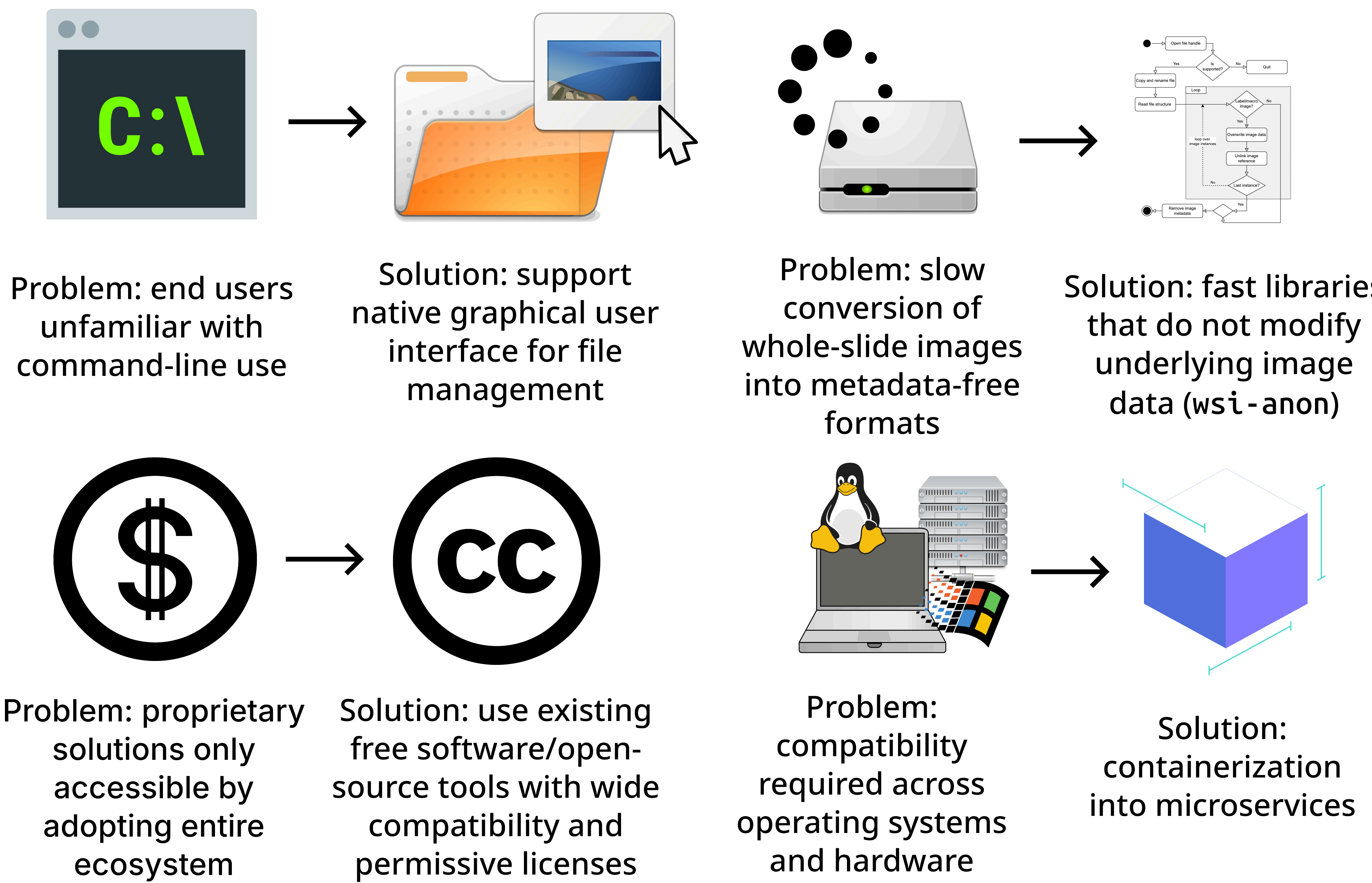
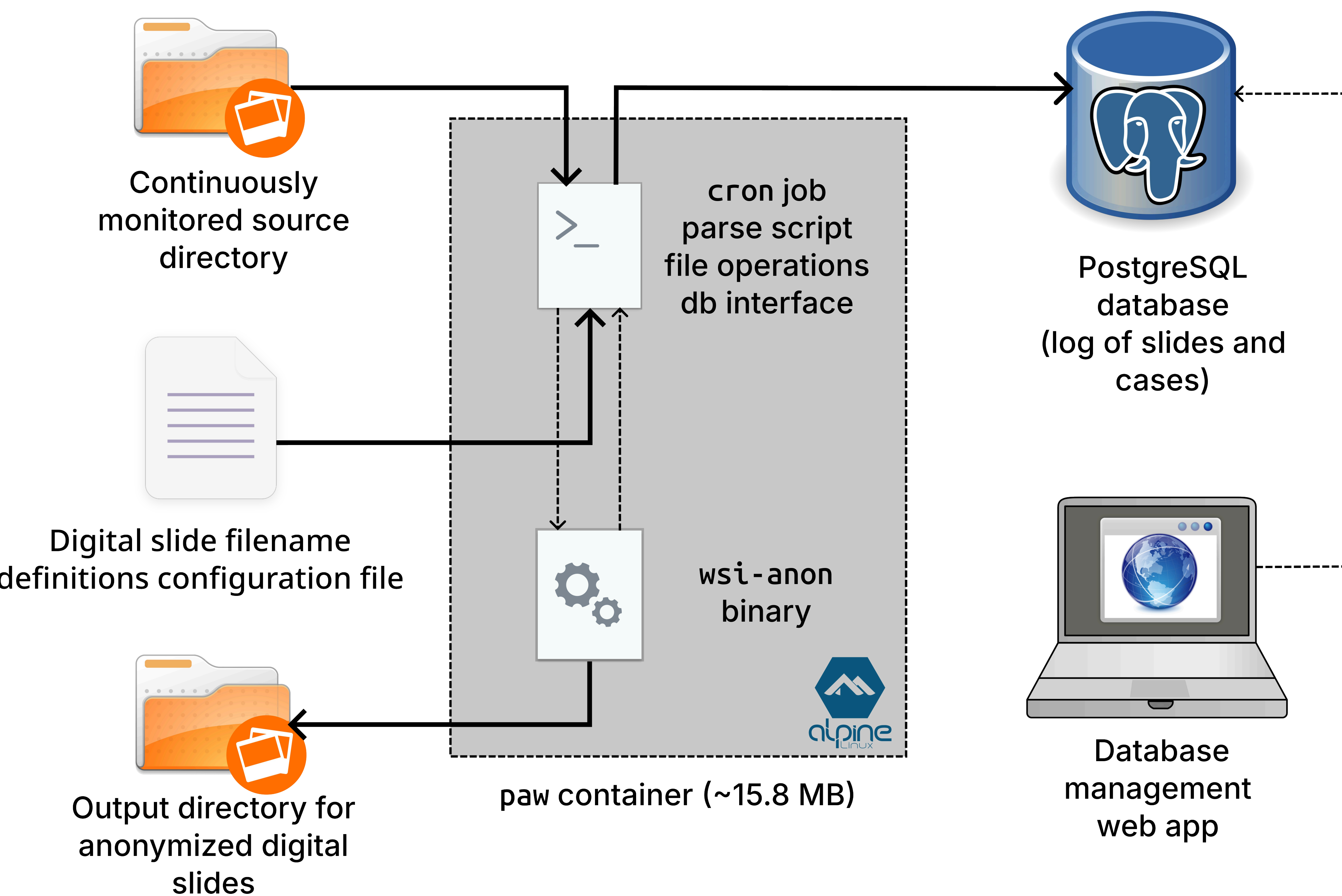


Figure 2: Design of a microservice to continuously anonymize digital pathology slides dragged and dropped into a folder



## Design considerations

We prioritized the following characteristics for our solution: (1) once installed, end users should not need to use any command-line tools to perform digital slide anonymization; (2) when pseudonymous identifiers have been pre-assigned to cases (for example, biobank IDs), anonymized file names should use such identifiers; (3) file names for anonymized slides should follow an informative, user-defined pattern, and allow for preservation of non-identifying metadata such as specimen, block and slide information; (4) the solution should be highly portable and lightweight.

To achieve this, we implemented paw, a Docker container based on Alpine Linux that includes the compiled wsi-anon binary and a compatibility layer. In addition, the container includes continuously operating scripts that perform the following jobs (fig. 2):

(1) Monitor the root of a user-defined source directory, whether local or remote, and execute wsi-anon on any new compatible digital slide; (2) query a local comma-separated value file or a database to match and retrieve any previously attributed anonymous identifier; (3) rename anonymized files using universally unique identifiers

(UUIDs) or optionally a user-defined pattern, and move them to a local or remote output directory; (4) log all anonymized digital slides to a local comma-separated value file or a database.

Use of a PostgreSQL database is optional, but our quick-start configuration file uses existing tools to create a database and launch a web-based interface to access it.

## Practical use case description

As part of an ongoing research ethics board-approved project, our team wanted to digitally scan several thousand archival glass slides associated with the surgical pathology cases of an existing patient cohort. As the expected file size of the scanned slides was large (several terabytes), we selected a commercial network-attached storage device to host them (fig. 3).

As this device supports Docker containers, we started up our microservices stack using the default paw configuration file. Using the provided web-based database interface, we uploaded the list of surgical accession numbers and corresponding identifiers to the database. Since our archival slides had used several distinct types of barcodes and different case naming conventions, we mapped each barcode type to a filename pattern allowing paw to decode them.

We then simply requested that when digitizing slides destined for this research project, the operator of the slide scanner set its output to the source directory monitored by paw. This is a routine operation that does not require any knowledge of our anonymization workflow, and that can be performed with the slide scanner control software.

As a result, without requiring additional user intervention, we obtained a dataset of anonymized digital slides, stripped of identifying metadata, named using a consistent pattern, as well as a complete table of all the slides and cases in the dataset.

Figure 3: Storage device hosting paw directly reading from a slide scanner



## Code repository

<<https://codeberg.org/bertogatti/paw>>  
DOI:10.5281/zenodo.14994532

## Contact

André Lametti <[andre.lametti@mail.mcgill.ca](mailto:andre.lametti@mail.mcgill.ca)>

## Reference

[1] T. Bisson et al., "Anonymization of whole slide images in histopathology for research and education," Digital Health, vol. 9, p. 20552076231171475, Jan. 2023, doi: 10.1177/20552076231171475.