Bringing Open Data to Whole Slide Imaging

Besson, Sebastien; Leigh, Roger; Linkert, Melissa; Allan, Chris; Burel, Jean-Marie; Carroll, Mark

Published in:
Digital Pathology

DOI:
10.1007/978-3-030-23937-4_1

Publication date:
2019

Document Version
Peer reviewed version

Bringing Open Data to Whole Slide Imaging

Sébastien Besson*[1][0000-0001-8783-1429], Roger Leigh*[1][0000-0003-2004-6965], Melissa Linkert*[2], Chris Allan[2], Jean-Marie Burel[1][0000-0002-1789-1861], Mark Carroll[1][0000-0001-7642-8749], David Gault[1], Riad Gozim[1], Simon Li[1][0000-0002-7929-2442], Dominik Lindner[1][0000-0001-8038-1250], Josh Moore[1][0000-0003-4028-811X], Will Moore[1][0000-0002-7264-8338], Petr Walczysko[1], Frances Wong[1][0000-0001-7397-8251] and Jason R. Swedlow[1,2(✉)] [0000-0002-2198-1958]

[1] Dept of Computational Biology, School of Life Sciences, University of Dundee, Dundee, DD1 5EH, United Kingdom
[2] Glencoe Software, Inc, 800 5th Ave, #101-259, Seattle, WA 98104, USA
* These authors contributed equally to this work
jrswedlow@dundee.ac.uk

Abstract. Faced with the need to support a growing number of whole slide imaging (WSI) file formats, our team has extended a long-standing community file format (OME-TIFF) for use in digital pathology. The format makes use of the core TIFF specification to store multi-resolution (or "pyramidal") representations of a single slide in a flexible, performant manner. Here we describe the structure of this format, its performance characteristics, as well as an open-source library support for reading and writing pyramidal OME-TIFFs.

Keywords: whole slide imaging, open file format, open data, OME-TIFF.

1 Introduction

Digital Pathology is a rapidly evolving field, with many new technologies being introduced for developing and using biomarkers [1, 2], imaging [3], and feature-based image analysis [4–7], most notably using various approaches to machine and deep learning [8, 9]. As is often the case in fields that cross research science and clinical practice, this transformation has been supported by rapid technology development driven both by academia and industry. A full ecosystem of open and commercial tools for preparing and scanning slides and analysing the resulting data is now evolving. These are starting to deliver advanced, innovative technologies that, at least in some cases, can evolve into defined products suitable for use in clinical laboratories.

During similar phases in the fields of radiology, genomics, structural biology, electron and light microscopy, and many others, one of key developments that helped accelerate development was the appearance of common, defined and open methods for writing, reading, and sharing data. Each of these fields has taken different approaches to

defining open data formats, and the approaches taken in different fields have had different levels of adoption. Digital Pathology, despite the rapid growth and potential of the field has not yet developed and adopted a mature open format that supports the wide range of data types that have emerged (with more on the horizon).

Since 2002, OME has built open software specifications and tools that accelerate and scale access to large, multi-dimensional datasets. OME's OME-TIFF [10], Bio-Formats [11] and OMERO [12] are used in 1000s of academic, industrial and clinical laboratories worldwide managing access to imaging data and also for publishing imaging data on-line [13, 14], In this report, we present an open, flexible data format based on accepted imaging community standards that supports all the whole slide imaging (WSI) modalities we are aware of today and can expand to support many of the emerging data types that are likely to appear in the near future. Critically, we provide open source, liberally licensed software for reading, writing and validating the format, freely available documentation and specifications, open build systems that anyone can monitor for development, and open, versioned example files for use in development and benchmarking experiments. Finally, we embed the format writer in a library that supports conversion from some of the dominant WSI proprietary file formats (PFFs).

2 State and support of WSI formats

The field of Digital Pathology has not yet adopted an open, supported, implemented data format for storing and exchanging WSI generated by acquisition scanners. The absence of such a format means that WSI in Digital Pathology uses PFFs, making the data fundamentally non-exchangeable, not available for long-term archiving, submission with regulatory filings or on-line publication. As more research funders and scientific journals adopt the principle that research data should be Findable, Accessible, Interoperable and Reproducible (FAIR) [15], this situation ultimately prevents the field of Digital Pathology from complying with emerging trends and regulations in research science and also inhibits further innovation as exemplar datasets are not available to technology developers. Technologies like deep learning require large, diverse datasets that realistically can only be assembled by combining datasets from multiple centres and/or clinics. Cohort datasets written in incompatible PFFs slow the development of new tools and waste precious resources (usually public funding) on converting incompatible data- a process that is error-prone and often leads to data loss.

Moreover, as each new WSI scanner arrives on the market, a new data format is introduced to the community. Manufacturers update their formats at arbitrary times, further expanding the number of versions of these proprietary file formats (PFFs).

To deal with this explosion of WSI PFFs, software translation libraries have emerged that read data stored across many formats into a common open representation using a unified application programming interface (API). As of today, the two most established libraries used in the WSI domain are OpenSlide, a C-based library developed at Carnegie Mellon University [16] and Bio-Formats, a Java-based library developed by the OME Consortium [11]. Both have been developed by academic groups as open

source projects. Many open-source and commercial tools in turn rely on the continued availability of these low-level libraries as a way to seamlessly access WSI data independently of its format. When reusing these libraries is not possible, commercial entities end up rewriting their own internal translational library allowing to achieve the same goal: reading WSI data independently of the format (e.g., https://free.pathomation.com/). Table 1 lists common types of WSI formats including their main manufacturer, their extension as well as their support in open-source libraries.

Table 1. List of common Proprietary File Formats (PFFs) used in the Whole Slide Imaging (WSI) domain alongside open-source libraries OpenSlide and Bio-Formats.

Manufacturer	File format extension	Support in open-source libraries
Aperio	.tiff	OpenSlide, Bio-Formats
Aperio	.svs, .afi	OpenSlide, Bio-Formats
Hamamatsu	.vms	OpenSlide, Bio-Formats
Hamamatsu	.ndpi,.ndpis	OpenSlide, Bio-Formats
Leica	.scn	OpenSlide, Bio-Formats
Mirax	.mrxs	OpenSlide
PerkinElmer	.qptiff	Bio-Formats
Philips	.tiff	OpenSlide
Sakura	.svslide	OpenSlide
Trestle	.tif	Bio-Formats, OpenSlide
Ventana	.bif, .tif	OpenSlide
Zeiss	.czi	Bio-Formats

It may appear that OpenSlide and Bio-Formats provide a convenient solution to the large and growing number of WSI PFFs. However, as shown in Table 1, no single implementation has a full coverage for the complete set of proprietary formats. Second, the burden of maintaining and expanding such libraries mainly remains the responsibility of the projects that build the libraries, as they reverse engineer each new PFF released by commercial manufacturers. The absence of prior discussion between manufacturers and community software developers involves constantly keeping up with the creation of new variants or new proprietary formats. Finally, data stored using these proprietary file formats remains fundamentally non-exchangeable between two researchers due to the absence of agreed-upon specification.

In response, we have embarked on a project to build a truly extensible, flexible, metadata-rich, cross-platform, open WSI data format for Digital Pathology.

3 Towards an Open WSI File Format

The Digital Imaging and Communications in Medicine (DICOM) working group published an official release (Supplement 145) in September 2010 specifically designed

to provide a standard specification for WSI data [17]. Conversion tools for generating DICOM-compliant files have been proposed [18], but community adoption of this format is limited. A key point is that the DICOM process only provides a data specification and leaves it to other entities to build reference implementations for the community. Delivering cross-platform, versioned, supported software that can be used across a broad community with many different use cases and applications is challenging and requires substantial dedicated resources. Moreover, DICOM supports private attributes and classes that can limit opportunities for implementing interoperability.

A separate issue with DICOM Suppl. 145 specification is the lack of software libraries for efficient reading and writing of the format for I/O intensive data processing, e.g., training of convolutional neural networks and other advanced learning applications. High performance software libraries that can contend with the large data volumes collected in WSI studies are essential for the routine use of large training sets and the development of new deep learning-based approaches in Digital Pathology.

An alternative approach is to build an open format based on known, established standards that are widely supported by communities and both open and commercial software and is proven to be useful for computational workflows. For example, the Tagged Image File Format (TIFF) specification is widely used as a binary vessel for image data storage (https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml). Since 2005, the OME Consortium has released OME-TIFF, a variant that complies with the TIFF specification, but adds OME's flexible imaging metadata model to the TIFF header [10]. As the OME metadata model includes support for imaging metadata, region of interest annotations, and a flexible key-value store [19], the format has been used to support many different imaging modalities in research, industrial and commercial settings (https://docs.openmicroscopy.org/latest/ome-model/ome-tiff/). Open source reader and writer implementations in Java and C++ are available [11, 20], along with a large number of example files (https://downloads.openmicroscopy.org/images/OME-TIFF/).

Given the interoperability of TIFF, it is no surprise that many PFFs have adopted the TIFF layout as a convenient way to store WSI data. Some libraries (OpenSlide, VIPS) use a so-called tiled multi-resolution TIFF format where each resolution is stored as a separate layer within a multi-page TIFF. A direct advantage of this approach is its great simplicity. However, while it applies well to single-plane RGB pyramidal images, this approach does not immediately support multi-channel data from fluorescence WSI, multiplexed data from cyclic immunofluorescence [2] and mass spectrometry-based CODEX data [3] or a through-focus series ("Z-stack"). Finally, each of these approaches, while TIFF-based is yet another PFF.

An alternative layout is to extend the TIFF specification to store reduced resolutions internally and refer to them from each layer using a specific tag SubIFD. This approach is also compatible with standard TIFF tools like libtiff (http://www.libtiff.org/) and commercial tools like Adobe Photoshop. It also allows flexibility to store new multiplexed data, or any other extensions available in the TIFF specification. In 2018,

OME proposed the usage of this strategy as an extension of its OME-TIFF specification to be able to generate exchangeable pyramidal images (https://openmicroscopy.github.io/design/OME005/). In addition to the interoperability with other tools, this updated OME-TIFF format makes it possible to store and exchange multi-dimensional pyramidal images, so multiplexed data, through-focus Z-series and several others are supported [11]. Finally, OME's flexible metadata schemes support multiple WSI pyramidal images as well as typical ancillary images generated by WSI scanners, e.g., barcodes, macro images of the full slide, all as part of an OME-TIFF file.

A key design requirement is that this updated form of OME-TIFF is backwards compatible with existing software that reads OME-TIFF. Following discussion and feedback on the proposed approach, an update to OME-TIFF readers and writers was released that and fulfilled these requirements and several others.

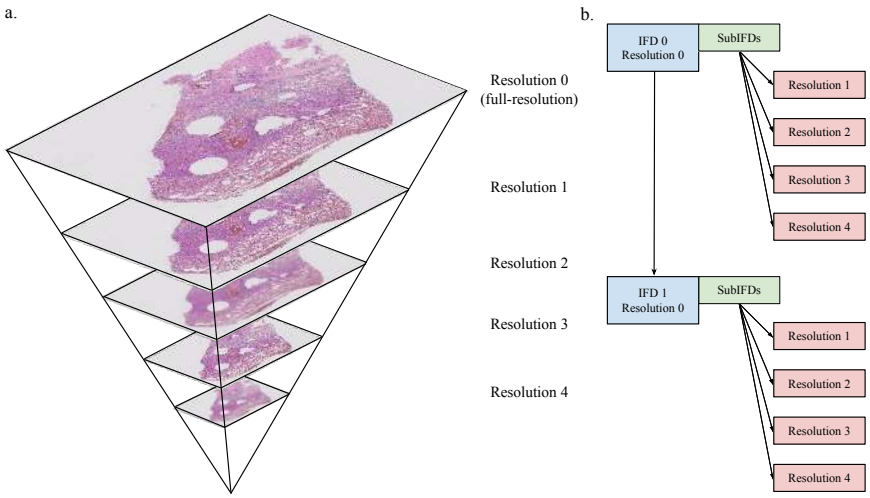

Fig. 1. a. Pyramidal image with five levels of resolution. Resolution 0 is the full-resolution plane while resolutions 1 to 4 are reduced along the X and Y dimensions using a consistent downsampling factor. b. In the updated OME-TIFF specification, this data is supported by storing metadata for sub-resolutions using the TIFF SubIFDs extension tag.

4 Implementations and Results

Figure 1 presents a graphical representation on how WSI data is stored in an OME-TIFF file. SubIFDs are used to indicate the location of sub-resolution tiles. Any software that implements the TIFF specification can be updated to read and write the file format. To demonstrate this, we modified the Bio-Formats library to read sub-resolutions from OME-TIFF files containing sub-resolution tiles. Test files were manually generated from public domain TIFF-based WSI PFFs to comply with the specification described above. These sample files were validated under two separate separate libraries that use Bio-Formats as a plug-in library, OMERO, a client-server data management application and QuPath, a desktop WSI data analysis application [12, 21]. In

both cases, updating the version of Bio-Formats enabled the software to read and display the updated OME-TIFF files. We validated the number of detected images including WSI, macro and label images and the number of sub-resolutions for each image, the metadata associated with each image and finally the pixel values for regions of each sub-resolution. The updated Bio-Formats library correctly passed all image parameters via metadata requests to the Bio-Formats API and properly delivered all tiles for rendering and display (Fig. 2).

a. b.

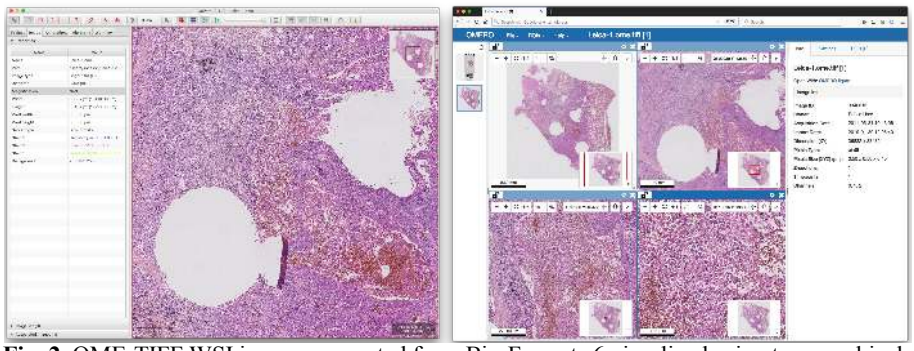

Fig. 2. OME-TIFF WSI images generated from Bio-Formats 6 visualized using two graphical clients a. QuPath and b. OMERO.iviewer [12, 21].

We also modified the Bio-Formats library to include support for writing pyramidal OME-TIFFs. The writer API was modified to allow setting the number of resolutions of an image and changing between sub-resolutions while writing data to disk. In addition, the OME-TIFF writer was updated to write sub-resolutions as described above. Finally, we implemented simple options to generate downsampled images from very large planes in the Bio-Formats conversion tools. These updates were tested using five different datasets: a selection of brightfield and fluorescent pyramidal images expressed in the main WSI PFFs supported by Bio-Formats (see Table 1), a collection of large single-plane TIFF files from the Human Protein Atlas project published in the Image Data Resource [14, 22], a synthetic image with 1400 Z-stacks, a multi-channel fluorescence image and a large electron-microscopy published in EMPIAR [23]. We converted all these datasets into OME-TIFFs using the command-line Bio-Formats tools and validated them as described above.

Table 2. List of resources publicly available for testing and validating the open OME-TIFF file format with support for multi-resolution.

Name	Description	URL
OME-TIFF	Format specification	https://docs.openmicroscopy.org/latest/ome-model/ome-tiff/specification.html
OME-TIFF	Public WSI samples	https://docs.openmicroscopy.org/latest/ome-model/ome-tiff/data.html#sub-resolutions
Bio-Formats 6	Binaries and API documentation	https://www.openmicroscopy.org/bio-formats/downloads/
Bio-Formats 6	Technical documentation	https://docs.openmicroscopy.org/latest/bio-formats6/

All of these functions have been built into and released as reference implementations that support the updated OME-TIFF formats (see Table 2) that include OME-TIFF samples for all the modalities described above, software libraries and documentation. The source code allowing to reproduce the data generation and validation is available at https://doi.org/10.5281/zenodo.2595928.

5 Discussion

We have developed an updated specification and implementation for OME-TIFF, an open image data format by adding support for multi-resolution tiles alongside existing capability for multiplexed, multi-focus and multi-timepoint images. Multi-resolution capability is important as it makes OME-TIFF usable as an exchange and/or transport format for WSI data. We have built and released example files, documentation and open source reference software implementations to ease OME-TIFF adoption by software developers and also research and clinical users.

Our goal in this work is not to declare a single data standard, but rather to build an open, supported WSI data format that is as flexible as possible, supports a wide range of metadata and binary data from many different applications, and can support the range of current and emerging domains using whole slide imaging. We have successfully tested the format across several different applications. We expect that the release of the updated OME-TIFF specification and open source software will enable the community to test the use of the format in many other domains and evaluate the utility of the specification and software. This will likely lead to several updates that steadily improve the utility and performance of OME-TIFF.

The reference implementation of the updated OME-TIFF has been developed in Java and integrated into the open-source Bio-Formats library [11]. For manufacturers, C++ and C# are usually the language of choice for writing software that drives commercial software for WSI acquisition. In addition to the Java-based library, the OME Consortium has built and released OME Files, a C++ reference implementation for reading and writing open OME formats [20] which we aim to update in the near future.

6 Acknowledgements

This work was funded by grants from the BBSRC (Ref: BB/P027032/1, BB/R015384/1) and the Wellcome Trust (Ref: 202908/Z/16/Z).

References

1. M. Udall et al.: PD-L1 diagnostic tests: a systematic literature review of scoring algorithms and test-validation metrics. Diagn. Pathol. 13, 12 (2018).
2. J.-R. Lin et al.: Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. Elife. 7, 31657 (2018).

3. Y. Goltsev et al.: Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Cell. 174, 968–981.e15 (2018).
4. P. Leo et al.: Stable and discriminating features are predictive of cancer presence and Gleason grade in radical prostatectomy specimens: a multi-site study. Sci. Rep. 8, 14918 (2018).
5. N. Beig et al.: Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas. Radiology, 180910 (2018).
6. R. Awan et al.: Glandular Morphometrics for Objective Grading of Colorectal Adenocarcinoma Histology Images. Sci. Rep. 7, 16852 (2017).
7. K. Sirinukunwattana et al.: Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Sci. Rep. 8, 13692 (2018).
8. A. Janowczyk, A. Madabhushi: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J. Pathol. Inform. 7, 29 (2016).
9. B. Ehteshami Bejnordi et al.: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 318, 2199–2210 (2017).
10. I. G. Goldberg et al.: The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol. 6, R47 (2005).
11. M. Linkert et al.: Metadata matters: access to image data in the real world. J. Cell Biol. 189, 777–782 (2010).
12. C. Allan et al.: OMERO: flexible, model-driven data management for experimental biology. Nat. Methods. 9, 245–253 (2012).
13. J.-M. Burel et al.: Publishing and sharing multi-dimensional image data with OMERO. Mamm. Genome. 26, 441–447 (2015).
14. E. Williams et al.: The Image Data Resource: A Bioimage Data Integration and Publication Platform. Nat. Methods. 14, 775–781 (2017).
15. M. D. Wilkinson et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3, 160018 (2016).
16. A. Goode, B. Gilbert, J. Harkes, D. Jukic, M. Satyanarayanan: OpenSlide: A vendor-neutral software foundation for digital pathology. J. Pathol. Inform. 4, 27 (2013).
17. R. Singh, L. Chubb, L. Pantanowitz, A. Parwani: Standardization in digital pathology: Supplement 145 of the DICOM standards. J. Pathol. Inform. 2, 23 (2011).
18. T. Marques Godinho, R. Lebre, L. B. Silva, C. Costa: An efficient architecture to support digital pathology in standard medical imaging repositories. J. Biomed. Inform. 71, 190–197 (2017).
19. S. Li et al.: Metadata management for high content screening in OMERO. Methods. 96, 27–32 (2016).
20. R. Leigh et al.: OME Files-An open source reference library for the OME-XML metadata model and the OME-TIFF file format. bioRxiv, 088740 (2016).
21. P. Bankhead et al.: QuPath: Open source software for digital pathology image analysis. Sci. Rep. 7, 16878 (2017).
22. M. Uhlén et al.: Proteomics. Tissue-based map of the human proteome. Science. 347, 1260419 (2015).
23. A. Iudin, P. K. Korir, J. Salavert-Torres, G. J. Kleywegt, A. Patwardhan: EMPIAR: a public archive for raw electron microscopy image data. Nat. Methods. 13, 387–388 (2016).