Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2016 February 13; 9714: . doi:10.1117/12.2212085.

Photon-HDF5: Open Data Format and Computational Tools for Timestamp-based Single-Molecule Experiments

Antonino Ingargiola^a, Ted Laurence^b, Robert Boutelle^a, Shimon Weiss^a, and Xavier Michalet^a

^aDept. of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California, USA

^bPhysical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA

Abstract

Archival of experimental data in public databases has increasingly become a requirement for most funding agencies and journals. These data-sharing policies have the potential to maximize data reuse, and to enable confirmatory as well as novel studies. However, the lack of standard data formats can severely hinder data reuse.

In photon-counting-based single-molecule fluorescence experiments, data is stored in a variety of vendor-specific or even setup-specific (custom) file formats, making data interchange prohibitively laborious, unless the same hardware-software combination is used. Moreover, the number of available techniques and setup configurations make it difficult to find a common standard.

To address this problem, we developed Photon-HDF5 (www.photon-hdf5.org), an open data format for timestamp-based single-molecule fluorescence experiments. Building on the solid foundation of HDF5, Photon-HDF5 provides a platform- and language-independent, easy-to-use file format that is self-describing and supports rich metadata. Photon-HDF5 supports different types of measurements by separating raw data (e.g. photon-timestamps, detectors, etc) from measurement metadata. This approach allows representing several measurement types and setup configurations within the same core structure and makes possible extending the format in backward-compatible way.

Complementing the format specifications, we provide open source software to create and convert Photon-HDF5 files, together with code examples in multiple languages showing how to read Photon-HDF5 files. Photon-HDF5 allows sharing data in a format suitable for long term archival, avoiding the effort to document custom binary formats and increasing interoperability with different analysis software. We encourage participation of the single-molecule community to extend interoperability and to help defining future versions of Photon-HDF5.

Keywords

single-molecule; fluorescence; file format; FRET; FCS; HDF5; SPAD; open data

1. OPEN SCIENCE AND OPEN DATA

In 1660, a group of English polymaths and natural philosophers met in London for the first time, to create what would soon become the *The Royal Society of London for Improving Natural Knowledge*, also known as the *Royal Society*.¹ The very first scientific society in human history had a founding principle well summarized by its motto *nullius in verba*, which means “take nobody’s word for it”. By formalizing the commitment “to verify all statements by an appeal to facts determined by experiment”,² the Royal Society set the ground for modern science. A few years later, in 1676, a member of the Royal Society, Sir Isaac Newton, wrote “If I have seen further, it is by standing on the shoulders of giants.”

Fast-forward to present days, very little has changed in the principles inspiring science. However, one important change took place in the past decades: joining theory and experiments, computational science has become a fundamental pillar of the scientific endeavor. In parallel, the amount of data produced by scientific investigations has grown exponentially. While, on one hand, this explosive growth is contributing to impressive advances in science, on the other hand, it is making harder to validate each other’s work.³ Too often, data underpinning scientific conclusions is not publicly available, making impossible to validate or extend the results. Nowadays, by contrast to the early days of the digital era, publishing all data generated by a research project has become easier. General-purpose repositories such as Zenodo⁴ and Figshare⁵ can store, curate, and preserve datasets, providing permanent citable resource which are accessible to the public, free of charge. Moreover, best-practices for data curation are starting to gain widespread acceptance.^{6,7}

Data sharing empowers scientists to participate in the process of validating and extending the scientific knowledge, creating opportunities for collaborations, and thus, for faster scientific progress. Additionally, since a large fraction of scientific investigations world-wide are funded by governmental agencies, it is natural to expect that, in general, the produced data be preserved and made accessible to all members of society.

In recent years, reflecting this sensibility and the practical benefits of data sharing, both governmental and private funding agencies have started mandating increasing efforts to preserve, catalog and publish data. In this light, the availability of file formats that are well documented, reliable and easy to use (ideally on any computing platform) is instrumental for promoting data sharing and long term preservation.

1.1 Examples of Other Formats

The need for standard file formats for domain-specific data is a common recurrence in many scientific communities. For examples, formats such as FITS⁸ and the recently proposed ASDF⁹ for astronomy, netCDF¹⁰ for geosciences and NeXus¹¹ for particle physics, have been developed by their respective communities and have helped catalyzing research and collaboration. Remarkably, except for FITS and ASDF, all other formats mentioned here are HDF5-based. HDF5 has also seen wide adoption in the industry, as exemplified by Math Works choosing HDF5 as the default data format in MATLAB (.mat files are in fact HDF5 files since version 7.3).

In the microscopy community, it is customary to save image and video data in TIFF format or as plain binary files. In the single-molecule imaging community, a new file format for camera-based experiments called SMD has been recently proposed.¹² Camera-based experiments typically employ surface-immobilized samples. In these experiments, time-binned intensity time traces are extracted for each emitter detected in the recorded video file. Due to fluorophore photobleaching, time traces have relatively short duration and therefore space requirements are much less of a problem than for freely-diffusing time stamp-based measurements. SMD files are based on the widespread JSON format, which, since it is text-based, ensures long term accessibility to the data. Unfortunately, SMD is unsuitable to store large time stamp-based datasets such as those produced by freely-diffusing single-molecule fluorescence experiments.

2. INTRODUCING PHOTON-HDF5

Photon-HDF5 is a file format for storing single-molecule fluorescence data acquired using single-photon detectors (typically single-photon avalanche diodes, or SPADs). In this class of experiments, encompassing a great variety of experimental techniques, photon time stamps and possibly other per-photon data (e.g. detector numbers, TCSPC nanotimes, etc...) are saved without binning. Photon-HDF5 defines a conventional layout that is suitable for this kind of time stamp-based data. By defining a common format, we aim to facilitate data sharing and interoperability between different scientific software. Additionally, by employing a well documented structure and by including a rich set of metadata, we aim to preserve the accessibility of experimental data in the foreseeable future.

In this article, which closely follows a previous publication of ours, we briefly illustrate design principles, general structure and resources related to Photon-HDF5. For a more complete description we point the reader to our previous publications^{13,14} and to the Photon-HDF5 website.¹⁵

2.1 Design and Features

Photon-HDF5 is based on HDF5 and therefore inherits all its advantages. In particular, HDF5 is an open standard which is platform- and programming language-agnostic. HDF5 is also an efficient binary format, which transparently supports compression and partial loading of datasets. Open source HDF5 implementations for the majority of programming languages are widely available (in most high level languages, HDF5 libraries bind to the official HDF5 C library).

In designing Photon-HDF5, we followed a set of general principles which are also useful for other scientific formats:

- **self-describing:** each data field embeds a plain text description explaining the purpose of the field;
- **self-contained:** it contains all the information necessary to analyze the data;
- suitable for **long-term archival:** rich metadata record experimental details, provenance, author and software version;

- supports arbitrary **user data**;
- support software is **open source** (under MIT license).

Additionally, in order to support a wide range of single-molecule fluorescence data, Photon-HDF5 has the following features:

- supports any number of spectral, polarization or “beam-split” channels (i.e. channels split with non-polarizing beam splitters);
- supports single- and multi-spot data;¹⁶
- extensible: the bulk “photon-data” (present in all types of measurements) is logically separated from data specific of a single measurement type (figure 1).

Photon-HDF5 currently supports 2-color smFRET, μ s-ALEX with 2 or 3 colors and ns-ALEX (also known as PIE) measurement types. Furthermore, thanks to its extensible structure, new measurement-types can be defined in backward-compatible manner.¹⁷ We encourage users to join the mailing list¹⁸ to discuss the definition of new measurement types (see section 2.3).

2.2 Photon-HDF5 Layout

Data in Photon-HDF5 files is organized in 5 main groups (see figure 2 and 3). The `photon_data` and `setup` groups (figure 2) contain all the raw data, measurement type specifications and parameters and hardware description. The remaining groups (figure 3), `sample`, `identity` and `provenance`, contain metadata about the measured sample, identification information (author, file creation date, software used, etc...) and provenance (original data file before conversion, original acquisition software, etc...). A more detailed overview can be found in previous papers,^{13,14} while the full specifications are part of the official Photon-HDF5 documentation.¹⁹

2.3 Development and Feedback

All Photon-HDF5 development (both specification documents and software) take place publicly on GitHub²⁰ and follows a participative open source model. The Photon-HDF5 homepage¹⁵ lists all the project resources and announces updates. Discussions on Photon-HDF5 usage, or on development of new features (e.g. new measurement specs) take place on the Photon-HDF5 forum.¹⁸

Users can contribute feedback and corrections to software or documentation using GitHub “Issues” or “Pull Requests”. Furthermore, users are welcome to propose extensions to the file format to accommodate or facilitate new use cases. We provide contributor’s guidelines²¹ and publicly acknowledge all contributions.

We encourage all interested parties to join this effort and participate in shaping the future of the Photon-HDF5 format.

2.4 Supporting Software

Photon-HDF5 files can be read in the HDF5 graphical viewer HDFView²² which also allows exporting arrays for use in other software. Photon-HDF5 files can be accessed by the majority of programming languages used for scientific computing, including C, Java, Python, R, Perl, Julia, MATLAB and LabVIEW. Specific procedures vary for each programming language. To help new users, we provide examples of how to read Photon-HDF5 files in Python, MATLAB and LabVIEW.²³

To write valid Photon-HDF5 files from scratch, we developed an open source Python library called phconvert.²⁴ Phconvert simplifies creating or converting Photon-HDF5 files, automating many of the requirements for creating valid files (e.g. it checks data type, inserts descriptions and automatically generates metadata fields). In addition to the library, phconvert also includes a set of Jupyter notebooks²⁵ (a web-based interface) which can be used to convert common file formats (PicoQuant's HT3 or Becker&Hickl's SPC) to Photon-HDF5. These notebooks serve also as examples of how to use the phconvert library. Users can run phconvert's notebooks locally or online using the Photon-HDF5 Converter Service described in section 2.5.

For all other programming languages, we provide an easy-to-install script called phforge²⁶ which has the same benefits as the phconvert library (which is used under the hood).

2.5 Online Converter

To allow users to try generating Photon-HDF5 from existing data files without requiring any software installation, we provide a demonstration online converter.²⁷ With this service, users can convert existing data files from PicoQuant's HT3 or Becker&Hickl's SPC formats to Photon-HDF5. The converter leverages the [MyBinder.org](https://mybinder.org) service from Freeman Lab to provide remote execution of the file-conversion notebooks.²⁵ The same Jupyter notebooks are included in the phconvert repository and can be executed locally.

3. SOFTWARE USING PHOTON-HDF5

In this section we list software supporting Photon-HDF5 either as input format (read support) or as storage (write support). All software listed here is open source.

3.1 FRETbursts

FRETbursts²⁸ is an open source Python package for single-molecule FRET burst analysis. It supports a wide range of standard analysis techniques such as single- and dual-channel burst search,^{29,30} background-dependent burst search threshold, burst selection, population fitting. FRETbursts is a Python library which can be conveniently used inside the Jupyter Notebook environment.²⁵

FRETbursts can load any smFRET data stored in Photon-HDF5 files. Supported measurement types include 2-color smFRET, μ S-ALEX, ns-ALEX or multi-spot smFRET data. When loading a Photon-HDF5 file, FRETbursts recognizes the measurement type by reading the relevant Photon-HDF5 fields.

FRETbursts can be used to perform standard state-of-the-art burst analysis or as a toolkit for implementing novel analysis techniques.

3.2 PyBroMo

PyBroMo is a 3D Brownian motion simulator for one or multiple populations of freely-diffusing singly- or doubly-labeled particles, excited by a diffraction limited laser. PyBroMo saves fully-compatible Photon-HDF5 files containing smFRET data.

PyBroMo, by default, uses a numerical PSF computed via rigorous vectorial electromagnetic simulation of a 3-layer system (objective medium, cover-glass and sample) and an Abbe-corrected objective lens computed with the PSFLab software.³¹ Users can compute their own PSF and use it with PyBroMo. PyBroMo simulates diffusion of N particles in a simulation box of user-defined size (typically, 6 – 12 μm side length). Number of particles and volume determine the simulated particle concentration. In a first step, the particle spatial trajectories and emission rates are simulated. The same trajectories can be reused to simulate discrete photon emission events (timestamps) for a single population or for a mixture of populations characterized by different FRET efficiencies and/or brightness.

Among other things, PyBroMo can be used to test the performance of smFRET burst analysis methods in identifying and characterizing different smFRET populations. Files created by PyBroMo can be directly analyzed with FRETbursts.

4. CONCLUSIONS

Within a single research group, Photon-HDF5 will allow saving time and efforts to comply with data sharing policies by using a well documented format, instead of maintaining and documenting internally used formats. The open development of Photon-HDF5, guarantees that individuals or groups can contribute and propose new features to fit their own particular needs, while immediately making the new files intelligible to and shareable with other users.

Additionally, by facilitating data sharing an interoperability between software, Photon-HDF5 can improve reproducibility and promote new collaborations between research groups. We encourage interested parties in the industry and in academia to participate in making Photon-HDF5 a valuable tools for the single-molecule community.

Acknowledgments

We thank Dr. Sangyoon Chung and Dr. Eitan Lerner for providing experimental data files. This work was supported in part by NIH Grant R01-GM95904 and by DOE Grant DE-FC02-02ER63421-00. Dr. Weiss discloses equity in Neshor Technologies and intellectual property used in the research reported here. The work at UCLA was conducted in Dr. Weiss's Laboratory. Dr. Laurence's work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

1. Royal Society on Wikipedia. https://en.wikipedia.org/wiki/Royal_Society. Accessed: 2016-01-30
2. The Royal Society History. <https://royalsociety.org/about-us/history/>. Accessed: 2016-01-30
3. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green

DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon TA, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T. SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* (New York, NY). Jun. 2015 348:1422–5.

4. Zenodo Homepage. <https://zenodo.org/>. Accessed: 2016-01-30
5. Figshare Homepage. <https://figshare.com/>. Accessed: 2016-01-30
6. Gewin V. Data sharing: An open mind on open data. *Nature*. Jan.2016 529:117–119. [PubMed: 26744755]
7. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, Gil Y, Groth P, Hedstrom M, Hogg DW, Kashyap V, Mahabal A, Siemiginowska A, Slavkovic A. Ten simple rules for the care and feeding of scientific data. *PLoS computational biology*. Apr.2014 10:e1003542. [PubMed: 24763340]
8. FITS, The Astronomical Image Table Format. <http://fits.gsfc.nasa.gov/>. Accessed: 2016-01-30
9. Greenfield P, Droettboom M, Bray E. ASDF: A new data format for astronomy. *Astronomy and Computing*. Sep.2015 12:240–251.
10. Network Common Data Form (NetCDF). <http://www.unidata.ucar.edu/software/netcdf/>. Accessed: 2016-01-30
11. NeXus, A common data format for neutron, x-ray and muon science. <http://www.nexusformat.org/>. Accessed: 2016-01-30
12. Greenfield M, van de Meent J-W, Pavlichin DS, Mabuchi H, Wiggins CH, Gonzalez RL, Herschlag D. Single-molecule dataset (SMD): a generalized storage format for raw and processed single-molecule data. *BMC bioinformatics*. Jan.2015 16(3)
13. Ingargiola A, Laurence T, Boutelle R, Weiss S, Michalet X. Photon-HDF5: an open file format for single-molecule fluorescence experiments using photon-counting detectors. *bioRxiv*. Sep.2015 doi: 10.1101/026484
14. Ingargiola A, Laurence T, Boutelle R, Weiss S, Michalet X. Photon-HDF5: An Open File Format for Timestamp-Based Single-Molecule Fluorescence Experiments. *Biophysical Journal*. Jan.2016 110:26–33. [PubMed: 26745406]
15. Photon-HDF5 Home Page. <http://photon-hdf5.org>
16. Ingargiola, A., Panzeri, F., Sarkhosh, N., Gulinatti, A., Rech, I., Ghioni, M., Weiss, S., Michalet, X. 8-spot smFRET analysis using two 8-pixel SPAD arrays. In: Enderlein, J.Gregor, I.Gryczynski, ZK.Erdmann, R., Koberling, F., editors. *Proceedings of SPIE*. Vol. 8590. Feb. 2013 p. 85900E-85900E-11.
17. Photon-HDF5: Defining new measurement types. http://photon-hdf5.readthedocs.org/en/latest/new_measurement_specs.html
18. Photon-HDF5 Public Forum. <https://groups.google.com/forum/#!forum/photon-hdf5>
19. Photon-HDF5 format definition. <http://photon-hdf5.readthedocs.org/en/latest/phdata.html>
20. Photon-HDF5 GitHub Project Page. <https://github.com/Photon-HDF5>
21. Photon-HDF5: Guidelines for contributors. <http://photon-hdf5.readthedocs.org/en/latest/contributing.html>
22. HDFView Homepage. <https://www.hdfgroup.org/products/java/hdfview/>. Accessed: 2016-01-30
23. Reading Photon-HDF5 in multiple languages. http://photon-hdf5.github.io/photon_hdf5_reading_examples/
24. Ingargiola, A. phconvert - Library and converter for Photon-HDF5 files. <http://photon-hdf5.github.io/phconvert>
25. Shen H. Interactive notebooks: Sharing the code. *Nature*. Nov.2014 515:151–2. [PubMed: 25373681]
26. phforge: a script for creating Photon-HDF5 files. <http://photon-hdf5.github.io/phforge/>
27. Photon-HDF5 Online Converter. <http://photon-hdf5.github.io/Photon-HDF5-Converter/>
28. Ingargiola A, Lerner E, Chung S, Weiss S, Michalet X. FRETbursts: Open Source Burst Analysis Toolkit for Confocal Single-Molecule FRET. *bioRxiv*. Feb.2016 doi: 10.1101/039198

29. Fries JR, Brand L, Eggeling C, Köllner M, Seidel CAM. Quantitative Identification of Different Single Molecules by Selective Time-Resolved Confocal Fluorescence Spectroscopy. *The Journal of Physical Chemistry A*. Aug.1998 102:6601–6613.
30. Nir E, Michalet X, Hamadani KM, Laurence TA, Neuhauser D, Kovchegov Y, Weiss S. Shot-noise limited single-molecule FRET histograms: comparison between theory and experiments. *Journal of Physical Chemistry B*. Nov.2006 110:22103–24.
31. Nasse MJ, Woehl JC. Realistic modeling of the illumination point spread function in confocal scanning optical microscopy. *Journal of the Optical Society of America A, Optics, image science, and vision*. Mar.2010 27:295–302.

Figure 1.

Logical sections of Photon-HDF5 files. We highlight the separation between experimental data and metadata, as well as the modularity of defining measurement type. New measurement types can be independently defined by users and submitted the Photon-HDF5 community for inclusion in the official specifications. This approach assures the flexibility needed to cover all the experimental niches, while promoting transparency and a common schema to describe different experimental techniques.

Figure 2.

Photon-HDF5 groups containing the data and hardware information. The *photon data* group contains raw data and measurement specifications. The *setup* group contains hardware and setup configurations.

Figure 3. Photon-HDF5 metadata groups containing sample and author information, provenance and software name and version used to save or convert the file.