On bringing bioimaging data into the open (World)

Moore, Josh; Kobayashi, Norio; Kunis, Susanne; Onami, Shuichi; Swedlow, Jason R.

Published in:
CEUR Workshop Proceedings

Publication date:
2019

On Bringing Bioimaging Data into the Open (World)

Josh Moore[1][0000-0003-4028-811X], Norio Kobayashi[2][0000-0002-2962-0073], Susanne Kunis[3], Shuichi Onami[4][0000-0002-8255-1724], Jason R. Swedlow[1][0000-0002-2198-1958], and the OME Consortium

[1] Division of Computational Biology School of Life Sciences, University of Dundee, Nethergate, Dundee, DD1 4HN, UK
{j.a.moore, j.r.swedlow}@dundee.ac.uk
[2] Head Office for Inrormation Systems and Cybersecurity (ISC), RIKEN, 2-1 Hirosawa, Wako, Saitama, 351-0198 Japan
norio.kobayashi@riken.jp
[3] Department of Biology/Chemistry and Center for Cellular Nanoanalytics, Osnabrück University, 49076 Osnabrück, Germany.
susanne.kunis@biologie.uni-osnabrueck.de
[4] RIKEN Center for Biosystems Dynamics Research (BDR), 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo, 650-0047 Japan
sonami@riken.jp

Abstract. For over 15 years, the Open Microscopy Environment (OME) Data Model has provided a basis for the storage, exchange and re-use of bioimaging data. During that time, XML Schema and XSL Transformations have provided a reliable mechanism to support the yearly updates to the model, keeping valuable data accessible by the research community. However, the acceleration of developments in the bioimaging domain now demand a more flexible, collaborative representation without the loss of versioning control. The OME Consortium proposes to adopt the semantic web stack for a next generation of data formats.

Keywords: Bioimaging, Domain Model, FAIR Data Principles, Linked Open Data, Semantic Framework.

1 Introduction

1.1 Bioimaging Advances

The potential applications for imaging in the life and biomedical sciences have driven an explosion in new modalities and technologies. Super-resolution, light sheet, and wave front correction methods have revolutionized light microscopy, providing previously impossible improvements to spatial and temporal resolution, and thus enabling experiments that were previously impossible [1–7]. Alongside these revolutions in imaging devices and systems, a parallel revolution in image processing has occurred with tools for deep learning-based segmentation [8,

9], restoration [10], imputing molecular localization from unlabeled images [11, 12] and advanced protocols for mapping effects of drugs or gene perturbations on cells and tissues [13].

1.2 A FAIR and Open World

The Open Microscopy Environment (OME) has played a key role in developing technologies that enable sharing and publishing of bioimage data [14]. The OME Data Model [15] contains fundamental concepts (Illumination, Detector, etc.) but these largely represent bioimaging as it was known circa 2000-2010. With the establishment of public image data resources like the Image Data Resource (IDR) [16], it is essential that OME's specifications for bioimaging metadata evolve to better support the wide range of technologies that are now routinely used in the life and biomedical sciences and finally make biomaging data available in a form that meets the FAIR data principles.

OME aims to address this fundamental block by providing mechanisms for others to properly annotate and describe these new methodologies. We propose to bring semantic modelling technologies developed for the modern Web to bioimaging. This will be achieved by expressing OME metadata in Resource Description Framework (RDF) triples conforming to an OWL ontology.

As a result, bioimaging datasets will be accessible beyond the closed-world of databases and schemas and become consistent with the open-world assumption (OWA). Input from diverse sources will be on equal footing to unite into a single, complete record of modern imaging systems and analysis workflows, improving the ability of users to find and access data for re-analysis and integration. *Our overall goal in joining the SWAT4HCLS community is to deliver more bioimage data into the public domain with more value than ever before.*

2 State of the art

2.1 OME-XML and OME-TIFF

OME's OME-TIFF format is used by several commercial companies and independent technology developers to write bioimage data in an open, easily accessible format. OME-TIFF incorporates the OME Data Model in the header of a TIFF file [17] as the specification for storing critical image acquisition, experimental and analytic metadata. This model was developed from 2001 to 2004 and published in 2005 [14] and as a result now only expresses a small portion of metadata concepts relevant to modern bioimaging. Most fundamentally, the idea that a single, static data model can capture the diversity of modern imaging modalities is no longer correct and must be transformed.

2.2 Version maintenance

To evolve OME's XML Schema (XSD), new versions are released along with XML Stylesheet Language (XSL) templates for both down-grading and upgrading documents. Software written with support for only a single version applies

multiple templates to convert documents. Data written with any version is still supported by OME-compatible software. Eleven updates of the model have taken place since the original release in 2003 to handle, e.g., High Content Screening data [18] as well as fluorescence lifetime, optical tomography and other advanced modalities (Table 1).

Table 1. Upgrade/downgrade compatibility afforded by the XSL transforms

Schema Version	**Upgrade** (excellent)	**Downgrade** (good)	**Downgrade** (fair)	**Downgrade** (poor)
2003-FC	2016-06	–	–	–
2007-06	2016-06	–	–	2003-FC
2008-02	2016-06	–	–	2003-FC
2010-04	2016-06	–	2009-09	2003-FC
2010-06	2016-06	–	2008-02	2003-FC
2011-06	2016-06	2010-04	2008-02	2003-FC
2012-06	2016-06	2010-04	2008-02	2003-FC
2013-06	2016-06	2010-04	2008-02	2003-FC
2015-01	2016-06	2010-04	2008-02	2003-FC
2016-06	–	2010-04	2008-02	2003-FC

The nearly yearly release schedule of the OME Data Model has been successful in maintaining the utility of the model for more established or complete technologies, but users must wait until the next release to express any newly emerging concepts.

2.3 Insufficiently Open-World

Though XSD and XSL have enabled the OME Consortium to provide a stable platform for the community, they neither allow for the storage of arbitrary statements nor provide a mechanism for the collaborative development of vocabularies. Proposals for new attributes to the model must be accepted by core developers, and due to the overhead of releases, conservative choices are made.

A result of the open-world assumption is that no single source of truth exists for what statements can be made. In an Open-World setting, individual documents can contain arbitrary statements which need not adhere to a single, closed-world schema. Of course, commonly used concepts can be incorporated over time into community-accepted standards, but no one group should have a monopoly on expressing biological truth.

2.4 Informal flexibility

To enable the storage of third-party metadata without modifying the central model, an extension to the data model allowed for free-form annotation using key-value pairs [19]. This mechanism flexibly stores spreadsheet-like metadata

to objects in the OME Data Model, where the column keys function roughly like RDF predicates (Fig. 1). This successfully captured the emerging metadata that needed to be expressed but did not provide the community a clear method for structuring these fledgling vocabularies. For example, even where **terms for a key are chosen from an existing vocabulary, discovering which vocabulary is in use is not possible.**

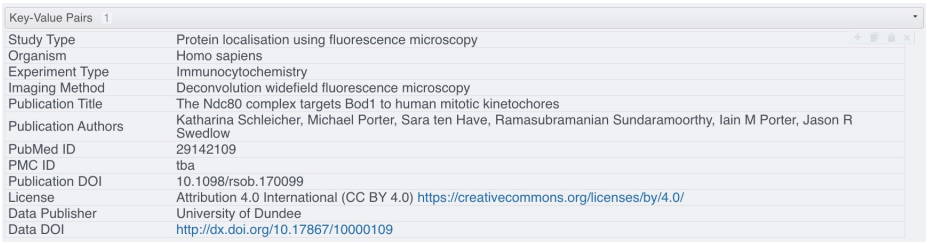

Fig. 1. Example of key-value metadata in OMERO [20]

2.5 Limited Alternatives

In their pursuit of flexibility, a frequent suggestion from users is that OME could replace the older XML technology stack with a more modern, web-based JSON one. This transition would provide a more comprehensive extension point than the key-value pairs, a simpler syntax, and a good deal of user-friendliness. The OME Consortium however has been hesitant to give up the versioning semantics and standards-track framework which are so key to the management of an evolving data model. JSON Schema could provide a partial solution, however, (a) it lacks the open-world assumption capability of OWL, (b) documents are not inherently linked to their schema, and (c) no standard tool is available for migrating documents between schema versions, as listed in Table 2.

Table 2. Comparison of schema language features

Schema language	XSD	JSON Schema	OWL
JSON representation	No	Yes	Yes
Open-World Assumption	No	No	Yes
Versioned documents	Yes	No	Yes
Declarative transformation	Yes (XSLT)	No	Yes (Inference)

3 Next Generation

The turning point in the search for an alternative technology was the previously reported work from RIKEN group [21]. As a collaboration between the University of Dundee and RIKEN, an implementation of the OME Data Model has been built in RDF/OWL and applied to the modeling of complex imaging workflows, demonstrating the possibility and utility of this approach. After subsequent community discussions, this work has been chosen for adoption as a path towards modernizing and extending OME's bioimaging data specifications. We aim to develop this specification as a candidate standard for the bioimaging community. The OME Consortium will maintain both the XML- and RDF-based representations. RDF-based bioimaging metadata will be equally supported in OME-TIFF as well as all future OME file formats. All existing elements of the XSD model will bidirectionally map into their OWL counterparts, while the RDF/OWL elements can more accurately express concepts from external semi-structured and structured metadata, like key-value pairs and other OWL domains, respectively.

3.1 Other features

Transformable and Versionable. Most critically, the semantic web stack provides the infrastructure for automatic upgrades and downgrade needed for the long-term support of existing documents. Each RDF/OWL version of the data model will be accompanied by the necessary descriptions needed to perform version maintenance. Initially, inference will be used to replace XSLT for migrating RDF documents from one OWL ontology to the next.

Submittable and Validatable. At the same time, the openness of the metadata approach should not place an undue burden on submitters or curators. Transformations from user-friendly formats like XLS simplify the submission, while transformations back to the closed-world OME-XML provide a first, strict validation for checking that constraints are still met. Longer-term, newer semantic validation technologies will need to be evaluated, such as ShEX.

Searchable and Integrative. Once curated, the metadata graph should also lead to an increase in reuse and integration. Though the existing database for OME data (OMERO) provides a query language, an API, and a web service, none of these are known outside of the bioimaging domain. For cross-domain searching, support for several FAIR principles can be improved by having a more widely known protocol (e.g. A1, I1) and resolvable URL-based identifiers (e.g. F1). [22]. This fosters collaboration and integration into existing RDF platforms like EBI's [23] or RIKEN's described below.

Extensible. Finally, the primary driver for this adoption is access to open-world extensibility. The most direct method is via the use of existing ontologies

like EDAM-imaging [24], CMPO [25], and EFO [26]. These are already in use in the IDR (below) but there is no method for discovering which key-values map to which ontology.

More substantially, however, the OME Data Model in OWL provides for the development by third parties of new models along with the necessary assertions for mapping between the models. This is precisely what the 4D-Nucleome project [27] has done and described in a SWAT4HCLS poster .

3.2 Applications

There are applications of similar mechanisms across the scientific domains [28, 29], but few widely known implementations for bioimaging. Below the authors' efforts are described as representative examples of what is currently underway.

IDR In 2016 OME began a collaboration with EMBL-EBI to build the Image Data Resource (IDR; `https://idr.openmicroscopy.org`), an added value, journal-independent database publishing reference bioimage datasets associated with peer-reviewed publications [16]. A critical aspect of IDR is its focus on curation, annotation, and publication of reference images: those likely to be heavily reused by the community and that integrate with other studies available in IDR. Metadata is currently collected from authors in a tabular format (e.g. CSV or XLS). The next IDR metadata version will be based on the OME Data Model in OWL. Tabular submissions can continue but will gain an enhanced semantic interpretation. An example scenario making use of these metadata is shown in Fig. 2.

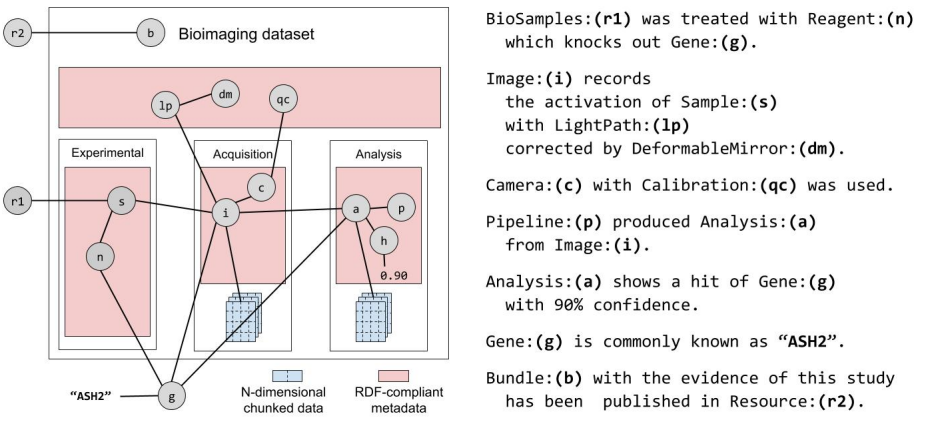

Fig. 2. Example of metadata modeling with LOD describing bioimaging metadata

The open-source IDR technology stack has inspired a number of independent IDRs in the same mold, related to marine biology, digital pathology as well

as efforts in several European and Asian countries to build national bioimage databases. RDF and OWL provide an ideal basis for expressing the breadth of bioimaging metadata and sharing them within and across communities.

RIKEN RIKEN is a comprehensive natural science research institute in Japan with a focus on the development of data-driven biomedical and open life sciences. Initial interest in the OME Data Model stems from work on the Systems Science of Biological Dynamics database (SSBD:database; `http://ssbd.qbic.riken.jp`) [30], which uses the OMERO platform [31] and provides quantitative resources for spatiotemporal dynamics of biological objects of various scales from single molecules to organisms, and peer-reviewed microscopy images obtained by using a variety of state-of-the-art technologies. Sample and experimental metadata were obtained from the authors and are provided in RDF. Originally, the OME Data Model in OWL was developed in order to extend that ontology for the description of electron microscopy (EM), X ray CT and MRI experimental conditions and samples. Future tasks include integration of imaging and other omics datasets using RIKEN MetaDatabase [32], an RDF-based data integration and publication platform.

German Bioimaging As a part of the Image Data Analysis and Management work group (`https://www.gerbi-gmb.de/WG6`), the Center of Cellular Nanoanalytics Osnabrück (CellNanOs) is focused on increasing the usability of bioimaging metadata capture early in the acquisition process and has built a specialized user-interface to that end. It is possible to extend the OME XSD model locally with any objects and define one's own reference points to OME elements. In progress is an OWL and RDF-based update of the interface to make it easier for the user to use one or more ontologies.

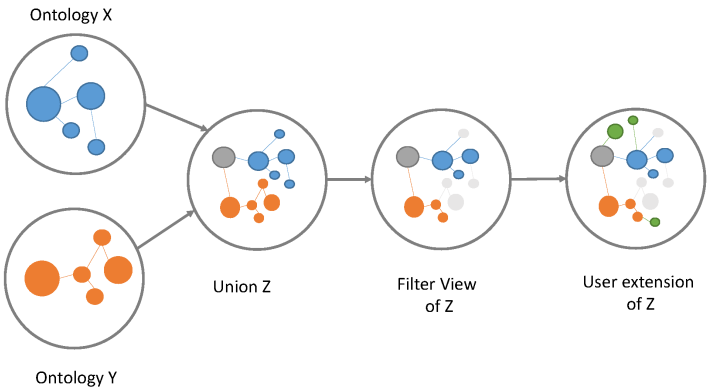

Fig. 3. User view of combined ontologies

In particular, the user can restrict ontologies to the areas they require ("filter view") for better clarity. For more flexibility the user is given the opportunity to integrate her "own world" into existing ontologies (Fig. 3).

The following functions of the interface supports the collection of this data: user-supported input by predefined objects (e.g. microscope hardware settings to compensate missing metadata in the image containers) and automatically recognizing missing data as well as the creation of a reusable template to enable fully automatic annotation of similar data.

The simple collection of metadata for data submission is critical for all scales of a bioimaging ecosystem. Institutional and national endeavors can flexibly capture relevant metadata. On submission to an international resource like the IDR, vocabularies can be normalized for maximizing the FAIR-ness and therefore value of all integrated datasets.

4 Conclusion

With this work, the OME Data Model gains a flexible representation with an open semantic framework as the basis for this FAIR- and open-world. The OME Consortium is excited to be joining the SWAT4HCLS community for this next phase of metadata development. We envision that this new semantic capability will facilitate the integration and analysis of bioimaging resources with a wide range of existing bioinformatics resources, multiplying the value of SSBD, IDR, and other participating endeavors.

Acknowledgements

NK was supported by JSPS KAKENHI (17K00434, 17K00424, 18K19766). SK was supported by the Deutsche Forschungsgemeinschaft (SFB944-INF). SO was supported by Core Research for Evolutionary Science and Technology (JST-CREST) (JPMJCR1511). Work on OME by JM and JRS is supported by awards from the BBSRC (BB/R015384/1) and the Wellcome Trust (212962/Z/18/Z).

References

1. Sahl, S.J., Hell, S.W., Jakobs, S.: Fluorescence nanoscopy in cell biology. Nature Reviews Molecular Cell Biology. 18, 685–701 (2017). doi: 10.1038/nrm.2017.71
2. Schermelleh, L., Ferrand, A., Huser, T., et al: Super-resolution microscopy demystified. Nature Cell Biology. 21, 72–84 (2019). doi: 10.1038/s41556-018-0251-8
3. Balzarotti, F., Eilers, Y., Gwosch, K.C., et al: Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. Science. 355, 606–612 (2016). doi: 10.1126/science.aak9913
4. Abrahamsson, S., Blom, H., Agostinho, A., et al: Multifocus structured illumination microscopy for fast volumetric super-resolution imaging. Biomedical Optics Express. 8, 4135 (2017). doi: 10.1364/BOE.8.004135

5. Trinh, L.A., Fraser, S.E.: Imaging the Cell and Molecular Dynamics of Craniofacial Development. Craniofacial Development. 599–629 (2015). doi: 10.1016/bs.ctdb.2015.09.002
6. Power, R.M., Huisken, J.: A guide to light-sheet fluorescence microscopy for multiscale imaging. Nature Methods. 14, 360–373 (2017). doi: 10.1038/nmeth.4224
7. McDole, K., Guignard, L., Amat, F., et al: In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. Cell. 175, 859–876.e33 (2018). doi: 10.1016/j.cell.2018.09.031
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR. `arXiv:1910.11370 [cs.CV]` (2015).
9. Hollandi, R., Szkalisity, A., Toth, T., et al: A deep learning framework for nucleus segmentation using image style transfer. bioRxiv. p. 580605 (2019). doi: 10.1101/580605
10. Weigert, M., Schmidt, U., Boothe, T., et al: Content-aware image restoration: pushing the limits of fluorescence microscopy. Nature Methods. 15, 1090–1097 (2018). doi: 10.1038/s41592-018-0216-7
11. Christiansen, E.M., Yang, S.J., Ando, D.M., et al: In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. Cell. 173, 792–803.e19 (2018). doi: 10.1016/j.cell.2018.03.040
12. Ounkomol, C., Seshamani, S., Maleckar, M.M., et al: Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nature Methods. 15, 917–920 (2018). doi: 10.1038/s41592-018-0111-2
13. Bray, M.-A., Singh, S., Han, H., et al: Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature Protocols. 11, 1757–1774 (2016). doi: 10.1038/nprot.2016.105
14. Swedlow, J.R., Goldberg, I.G., Brauner, E., et al: Informatics and Quantitative Analysis in Biological Imaging. Science. 300, 100–102 (2003). doi: 10.1126/science.1082602
15. Goldberg, I.G., Allan, C., Burel, J.-M., et al: The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biology. 6, R47 (2005). doi: 10.1186/gb-2005-6-5-r47
16. Williams, E., Moore, J., Li, S.W., et al: Image Data Resource: a bioimage data integration and publication platform. Nature Methods. 14, 775–781 (2017). doi: 10.1038/nmeth.4326
17. Linkert, M., Rueden, C.T., Allan, C., et al: Metadata matters: access to image data in the real world. The Journal of Cell Biology. 189, 777–782 (2010). doi: 10.1083/jcb.201004104
18. Li, S., Besson, S., Blackburn, C., et al: Metadata management for high content screening in OMERO. Methods. 96, 27–32 (2016). doi: 10.1016/j.ymeth.2015.10.006
19. Burel, J.-M., Besson, S., Blackburn, C., et al: Publishing and sharing multi-dimensional image data with OMERO. Mammalian Genome. 26, 441–447 (2015). doi: 10.1007/s00335-015-9587-6
20. Schleicher, K., Porter, M., ten Have, S., et al: The Ndc80 complex targets Bod1 to human mitotic kinetochores. 2017. data doi: 10.17867/10000109
21. Kume, S., Masuya, H., Kataoka, Y., et al: Development of an Ontology for an Integrated Image Analysis Platform to enable Global Sharing of Microscopy Imaging Data. International Semantic Web Conference (2016)
22. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., et al: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3, (2016). doi: 10.1038/sdata.2016.18

23. RDF platform: Linked Open Data platform for EBI data. https://ebi.ac.uk/rdf. Last accessed 22 Nov 2019
24. Kalaš, M., Plantard, L., Sladoje, N., et al.: EDAM-bioimaging: the ontology of bioimage informatics operations, topics, data, and formats (2019 update). F1000Research Feb 6 (2019). doi: 10.7490/f1000research.1116432.1
25. Jupp, S., Malone, J., Burdett, T., et al: The cellular microscopy phenotype ontology. Journal of Biomedical Semantics. 7, 28 (2016). doi: 10.1186/s13326-016-0074-0
26. Malone, J., Holloway, E., Adamusiak, T., et al: Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 26(8), 1112–1118 (2010). doi: 10.1093/bioinformatics/btq099
27. Huisman, M., Hammer, M., Rigano, A., et al: Minimum Information guidelines for fluorescence microscopy: increasing the value, quality, and fidelity of image data. arXiv:1910.11370 [q-bio.QM] 2019.
28. Faulconbridge, A., Burdett, T., Brandizi, M., et al: Updates to BioSamples database at European Bioinformatics Institute. Nucleic Acids Research. 42, D50–D52 (2013). doi: 10.1093/nar/gkt1081
29. Salvadores, M., Alexander, P. R., Musen, M. A., et al: BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. Sem. Web. 4, 277-284 (2013).
30. Tohsato, Y., Ho, K.H.L., Kyoda, K., et al: SSBD: a database of quantitative data of spatiotemporal dynamics of biological phenomena. Bioinformatics. 32, 3471-3479 (2016). doi: 10.1093/bioinformatics/btw417
31. Allan, C., Burel, J.-M., Moore, J., et al: OMERO: flexible, model-driven data management for experimental biology. Nature Methods. 9, 245–253 (2012). doi: 10.1038/nmeth.1896
32. Kobayashi, N., Kume, S., Lenz, K., et al: RIKEN MetaDatabase: A Database Platform for Health Care and Life Sciences as a Microcosm of Linked Open Data Cloud. Int J Semant Web Inf Syst. 14(1), 140-164 (2018).