

REVIEW

Community standards for open cell migration data



Alejandra N. Gonzalez-Beltran ^{1,2,†}, Paola Masuzzo ^{3,4,5,†},
 Christophe Ampe ⁴, Gert-Jan Bakker ⁶, Sébastien Besson ⁷, Robert
 H. Eibl ⁸, Peter Friedl ^{6,9,10}, Matthias Gunzer ^{11,12},
 Mark Kittisopikul ^{13,14}, Sylvia E. Le Dévédec ¹⁵, Simone Leo ^{7,16},
 Josh Moore ⁷, Yael Paran ¹⁷, Jaime Prilusky ¹⁸, Philippe Rocca-Serra ¹,
 Philippe Roudot ¹⁹, Marc Schuster ¹¹, Gwendolien Sergeant ⁴,
 Staffan Strömblad ²⁰, Jason R. Swedlow ⁷, Merijn van Erp ⁶,
 Marleen Van Troys ⁴, Assaf Zaritsky ²¹, Susanna-Assunta Sansone ^{1,*}
 and Lennart Martens ^{3,4,*}

¹Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, Oxford OX1 3QG, Oxford, UK; ²Present address: Scientific Computing Department, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell Campus OX11 0QX, Didcot, UK; ³VIB-UGent Center for Medical Biotechnology, VIB, A. Baertsoenkaai 3, B-9000, Ghent, Belgium; ⁴Department of Biomolecular Medicine, Ghent University, A. Baertsoenkaai 3, B-9000, Ghent, Belgium; ⁵Institute for Globally Distributed Open Research and Education (IGDORE), Kabupaten Gianyar, Bali 80571, Indonesia; ⁶Department of Cell Biology, Radboud Institute for Molecular Life Sciences, Geert Grooteplein 28 6525 GA Nijmegen, The Netherlands; ⁷Centre for Gene Regulation & Expression & Division of Computational Biology, University of Dundee, School of Life Sciences, Dow St Dundee DD1 5EH, Scotland, UK; ⁸German Cancer Research Center, DKFZ Alumni Association, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany; ⁹David H. Koch Center for Applied Genitourinary Medicine, UT MD Anderson Cancer Center, 6767 Bertner Ave, Mitchell Basic Science Research Building, 77030 Houston, TX, USA; ¹⁰Cancer Genomics Center, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands; ¹¹Institute for Experimental Immunology and Imaging, University Hospital, University Duisburg-Essen, Universitätsstr. 2, 45141 Essen, Germany; ¹²Leibniz Institute for Analytical Sciences, ISAS, Bunsen-Kirchhoff-Straße 11, 44139 Dortmund, Germany; ¹³Department of Biophysics, UT Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, TX 75390, USA; ¹⁴Department of Cell and Developmental Biology, Feinberg School of Medicine, Northwestern University, 303 E. Chicago Ave, Chicago, IL 60611, USA; ¹⁵Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, PO box 9502 2300 RA Leiden, The Netherlands; ¹⁶Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Loc. Piscina Manna, Edificio 1, 09050 Pula (CA), Italy; ¹⁷IDEA Bio-Medical Ltd, 2 Prof. Bergman St., Rehovot 76705, Israel; ¹⁸Life Science Core Facilities, Weizmann Institute of Science, P.O. Box 26 Rehovot 76100, Israel; ¹⁹Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, 5323 Harry Hines Blvd.

Received: 26 October 2019; Revised: 2 April 2020; Accepted: 2 April 2020

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Dallas, TX 75390, USA; ²⁰Department of Biosciences and Nutrition, Karolinska Institutet, Neo, SE-141 83 Huddinge, Sweden and ²¹Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, P.O.B. 653, 8410501 Beer-Sheva, Israel

*Correspondence address. Susanna-Assunta Sansone, Oxford e-Research Centre, Department of Engineering Science, University of Oxford, Oxford e-Research Centre, 7 Keble Road, Oxford OX1 3QGOxford, UK. E-mail: susanna-assunta.sansone@oerc.ox.ac.uk  <http://orcid.org/0000-0001-5306-5690>; Lennart Martens, VIB and Ghent University, A. Baertsoenkaai 3, B-9000, Ghent, Belgium. E-mail: lennart.martens@vib-ugent.be  <http://orcid.org/0000-0003-4277-658X>

[†]Contributed equally to this work.

Abstract

Cell migration research has become a high-content field. However, the quantitative information encapsulated in these complex and high-dimensional datasets is not fully exploited owing to the diversity of experimental protocols and non-standardized output formats. In addition, typically the datasets are not open for reuse. Making the data open and Findable, Accessible, Interoperable, and Reusable (FAIR) will enable meta-analysis, data integration, and data mining. Standardized data formats and controlled vocabularies are essential for building a suitable infrastructure for that purpose but are not available in the cell migration domain. We here present standardization efforts by the Cell Migration Standardisation Organisation (CMSO), an open community-driven organization to facilitate the development of standards for cell migration data. This work will foster the development of improved algorithms and tools and enable secondary analysis of public datasets, ultimately unlocking new knowledge of the complex biological process of cell migration.

Keywords: cell migration; data standards; metadata; CMSO; MIACME; biotricks; frictionless data package; FAIR data

Introduction: Towards FAIR and Open Cell Migration Data

Owing to advances in molecular biology, microscopy technologies, and automated image analysis, cell migration research currently produces spatially and temporally resolved, complex, and large datasets. Consequently, experimental imaging techniques have de facto entered the “big data” era [1, 2]. This creates, on the one hand, challenges [3] for standardizing and maintaining data-driven cell migration research in public repositories while, on the other hand, offering unprecedented opportunities for data integration, data mining, and meta-analyses.

This situation resembles the progress that has been made in the omics fields integrating standardized data generation, sharing, and analysis over the past 2 decades [4, 5]. The ultimate goal for cell migration data processing is to follow a similar route to progress and become more quantitative, interdisciplinary, and collaborative.

To enable cell migration data integration, mining, and meta-analysis, we initiated an open data exchange ecosystem for cell migration research [6]. The aim was to overcome the current fragmentation of cell migration research and facilitate data exchange, dissemination, verification, interoperability, and reuse, as well as to encourage data sharing [7]. This should also increase the reproducibility of experiments, enable data mining and meta-analyses, and thus satisfy the FAIR principles for Findable, Accessible, Interoperable, and Reusable data [8]. Public availability of both cell migration data and metadata, and designated tools to mine these data, will facilitate the understanding of complex cell functions and their relevance for clinical use in health and disease. In addition, it will attract computational scientists to the field, producing *in silico* models allowing numerical hypotheses to be tested experimentally [9, 10].

Establishing such an open cell migration data ecosystem necessitates community consensus on what content to report, what terminologies to use, and what structured machine-readable formats to use in order to represent the experimental details, workflows, and analysis results. A significant chal-

lenge is the inherent heterogeneity of experimental data: experiments are performed in a wide array of assays, at all levels of throughput, in diverse cellular models, maintained in various microenvironments, using multiple microscopy techniques and analysis methods. For example, a common readout in a cell migration experiment is the measurement of the movement over time of cells and/or subcellular compartments. Other quantitative readouts include cellular morphology and its temporal dynamics [11]. However, there is no standard way to report this information, preventing the integration and mining of these data for downstream knowledge extraction. In addition, usually other experimental details are presented in narrative form in articles, in line with publication policies of scientific journals, and typically not delivered in a uniform machine-readable form. While the “Methods” section of scientific publications is supposed to enable full understanding of the experimental procedures and support replication, similar experiments may be described in an inconsistent manner in different studies. The methods description may be partial and may leave room for multiple interpretations of the experimental details and procedures of data analysis.

With the ultimate aim of an open data ecosystem, the Cell Migration Standardisation Organisation (CMSO) was established in 2016 to define and implement standards for the cell migration community. The CMSO operates openly and transparently, is based on voluntary efforts from the community, and is open to anyone interested in contributing and/or providing feedback. The developed standards are designed and implemented with the aim of achieving participants’ consensus. The CMSO outputs can be found in GitHub [12], while general information and activities are available from the CMSO website [13].

The cell migration community standards are composed of 3 modules, corresponding to the CMSO working groups (WGs) (Fig. 1):

1. Reporting guidelines specifying the minimum information required when describing cell migration experiments and data (WG1);

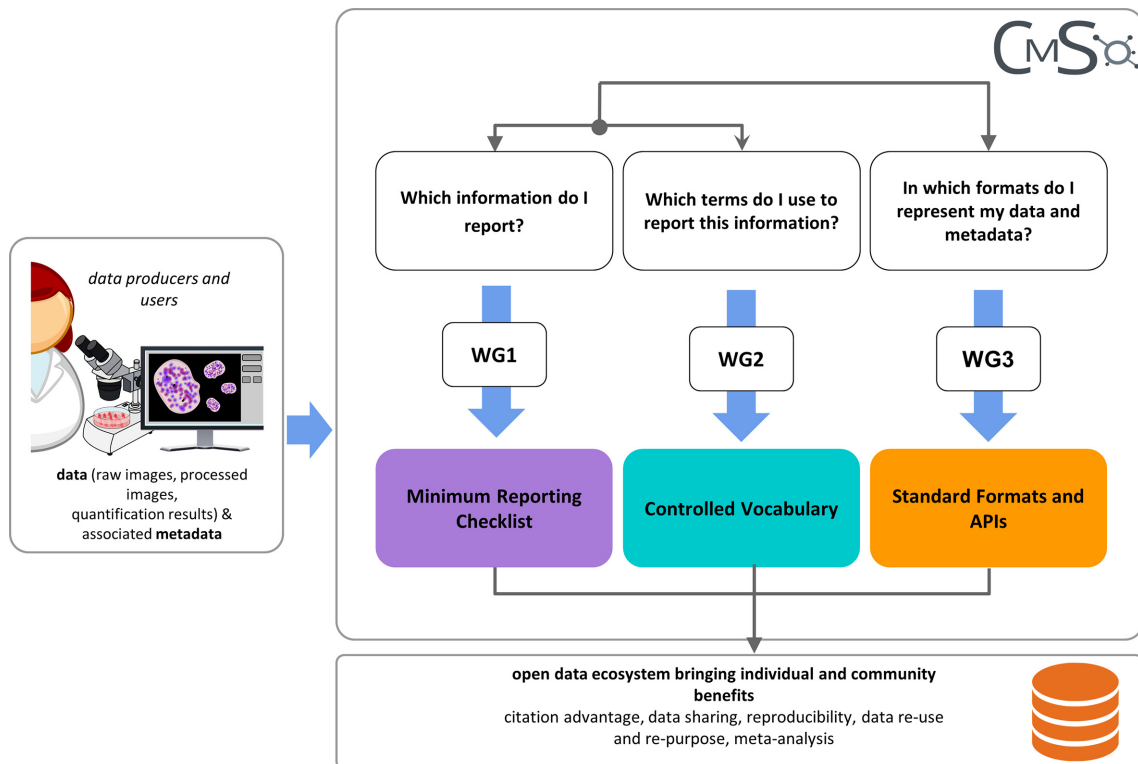


Figure 1: The Cell Migration Standardisation Organisation (CMSO). The 3 working groups (WGs) deliver specific standards in an interactive manner.

2. Controlled vocabularies (CVs) that unambiguously annotate these units of information (WG2);
3. Standard file formats for data and metadata, embodying the minimum reporting requirements and CV specifications, and APIs, which ensure that all data, results, and associated metadata can be read and interpreted by relevant software packages (WG3).

While CMSO consists of these 3 WGs (Fig. 1), interactions and synergies between them were essential to achieve the integrated model. For example, the information elements identified by the minimum reporting checklist (WG1) were annotated with the terms identified as CVs (in WG2), and both checklist and vocabularies were considered when developing formats and APIs (WG3).

In this article, we introduce the CMSO standards framework, providing an audit trail from the data to the machine-readable and interoperable metadata in a harmonized manner.

Results: CMSO Standards and Tools

Minimum reporting guidelines and controlled vocabularies

Minimum reporting guidelines, also termed “requirements” or “checklists,” aim to ensure that necessary and sufficient metadata are provided to enable the comprehension of an experiment, future data integration, data mining, and to ensure reproducibility. The CMSO has defined iterative versions of the Minimum Information About a Cell Migration Experiment (MIACME) guidelines, the latest version being MIACME 1.1 [14, 15], also registered [16] in the FAIRsharing portal [17]. Reporting guidelines, when enforced by journals, are an important factor to boost reproducibility according to 69% of researchers surveyed by *Nature*

[18] and have also been proven to improve the quality of experimental reporting [19].

MIACME consists of (i) generic information about an investigation, which can involve 1 or more studies, the associated publications, people, organizations, and grants; and (ii) specific information about the associated cell migration experiments. The cell migration-specific part of MIACME is partitioned into 3 conceptual domains (Fig. 2): (i) the experimental set-up: the assay, cell model, environmental conditions, and perturbations; (ii) the imaging condition: the microscopy settings; and (iii) the data: the raw images, summary information about the data (e.g., number of replicates, number of images), processed images, and the derived quantitative analysis outputs.

MIACME is presented as a specification accompanied by a spreadsheet that describes entities and properties, their expected values, cardinalities, and requirement levels. In addition, we also provide a machine-readable and actionable representation that can be validated in the form of JavaScript Object Notation (JSON) schemas [20] and configurations (or templates) for the ISA-Tabular (ISA-Tab) format, so that MIACME-compliant metadata can be created using the ISA framework [21] (see more details in the next section).

While the minimum information requirements determine the metadata elements to be reported, the community also needs to agree on the terms that will be used when describing cell migration experiments. A CV provides a standard terminology with unambiguous meaning for a particular domain with the goal of promoting consistent use of terms within a community [22]. These terms can then be included in an ontology that defines formal relationships between them. CVs and ontologies harmonize the data representation to perform queries across data repositories, enable data interoperability, and facilitate data integration, data mining, and knowledge discovery. A

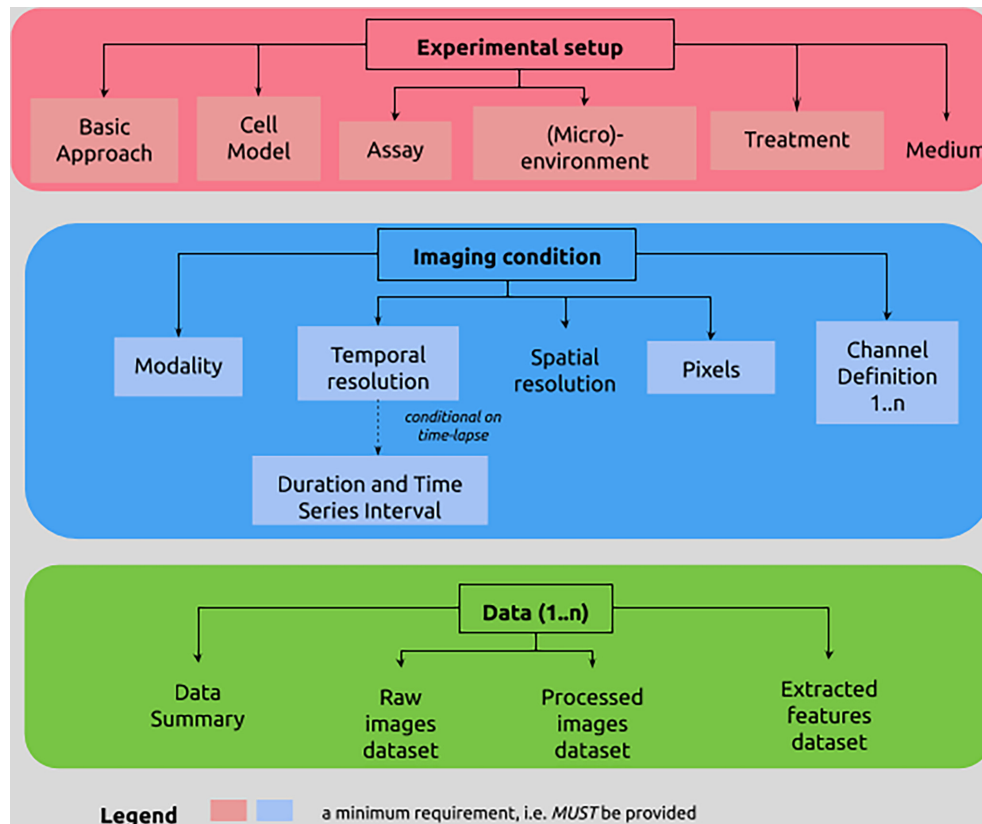


Figure 2: Overview of the cell migration-specific part of the MIACME specification (version 1.1 [16]). The figure presents an overview of the 3 main components of the cell migration experiment information: experimental set-up, imaging condition, and data. For more details about the MIACME guidelines, including the interrelationships among the 3 components, see the associated spreadsheet and schemas in the supporting data section, which specify the parameters for each conceptual area together with their requirement level, and illustrate them with examples.

typical data mining study could be based upon a set of competency questions, such as (i) find all *in vitro* cell migration experiments that make use of live-cell imaging, (ii) retrieve all experiments for which speed was recorded for cells migrating in a 3D collagen matrix, (iii) determine the migratory effect of knock-down of gene X in a breast cancer cell line, or (iv) find the dose response of compound Y on cell line Z in an invasion experiment.

The scope of a CV is therefore defined by a list of use cases as above. For cell migration, a CV requires, for example, terms for the cell line or type, gene names, and specific compounds used for molecular intervention as well as terms describing the type of cell migration assay or the manner in which the cells are presented at the start of the assay (currently termed “cellInput” in the specification). For particularly complex experiments, additional specific terms may be needed. For instance, for single-cell chemotaxis experiments, the CV needs to include terms for the directional chemoattractant application and the type of microscopy used.

The CMSO recommends the use of multiple ontologies for reporting cell migration experiments. The selection of relevant ontologies was based on an iterative strategy, as follows.

1. Determining the domain and scope of the terminologies, through a list of possible queries used for data mining in the domain, such as those mentioned above.
2. Reusing existing terminologies. Besides being more effective, reusing terminologies is also a best practice and a requirement to interoperate with other applications that

have already committed to particular CVs. The CMSO has identified existing controlled terminologies, recognized and maintained by the scientific community, that contain terms relevant for cell migration (see [23]).

3. Identifying missing terms (related to cell migration), specifying their definitions and relationships with existing terms. These terms are submitted to existing ontologies, when relevant, or will be created if necessary.

Standard formats, APIs, and tools

Once the content and terminology for reporting cell migration experiments have been defined, the community needs to reach consensus on the definition of a data exchange format to enable data sharing across researchers, institutes, software tools, and data repositories, as well as software libraries and APIs to interact with this format.

A single overarching file format may not suffice to capture the full complexity of the cell migration-associated data, and therefore CMSO opted for a collection of well-defined, open file formats. Each format is optimized toward different aspects of the experimental pipeline: (i) experimental metadata, (ii) imaging acquisition, and (iii) data analysis routines (Fig. 3). The existing APIs and formats from the Investigation/Study/Assay (ISA) [21, 24] and the Open Microscopy Environment (OME) [25] are used, respectively, for the experimental metadata and the image acquisition (left and middle boxes in Fig. 3).

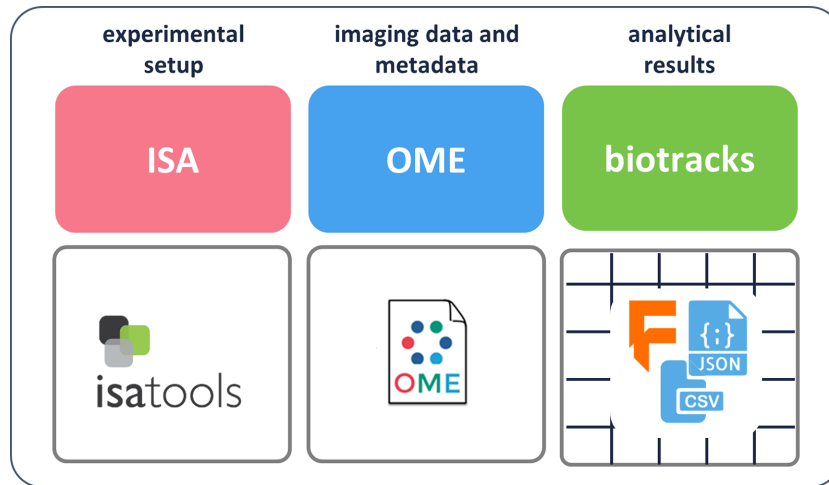


Figure 3: The first standardization products assembled and developed by CMSO WG3. Experimental set-up: Investigation Study Assay (ISA); image data and metadata: Open Microscopy Environment (OME); analytical results: biotracks.

The ISA model [26, 27] provides a rich description of the experimental metadata (e.g., sample characteristics, technology and measurement types, sample-to-data relationships). This ISA feature served as basis for the conceptual “experimental set-up” section in the MIACME guidelines (Fig. 2, left box).

The OME Data Model [28] is a specification for the exchange of image data. It represents images as 5D entities: the 2D plane (x, y), the focal position (z), the spectral channel, and the time. The OME format also includes metadata such as details of the acquisition system and experimental parameters related to acquisition. These metadata are related to the “imaging condition” conceptual area within MIACME (Fig. 2, middle box).

With the ISA and OME models established and publicly available, the remaining challenges are around standardized reporting of routine analyses, such as cell tracking or quantification of cell shape. Here, we report the specification and implementation of a new open tracking data format named “biotracks” [29] (Figs 3 and 4). The biotracks format was designed to accommodate the time-resolved tracking information of various objects observed in cell migration experiments. These tracked objects can either be cells, specific organelles, or cellular structures (e.g., leading cell edges, nuclei, microtubule organizing centers, filaments, single molecules, and signals that report signaling dynamics). Generally, an object could be any region of interest (ROI), including an arbitrary mask or even a single point in space. As such, the specification includes 3 levels of information: (i) objects identified during cell segmentation or detection tasks, (ii) links that linearly connect objects across frames of the acquired time sequence, and (iii) tracks that connect links across events such as splitting or merging (Fig. 4A). This abstraction enables the standardized description of a wide variety of biological tracking data.

Whereas **biotracks** follows the general strategy of any tracking software, its specification focuses on enabling data interoperability in a simple way by specializing the Tabular Data Package [30] container format. In the data package, the data (objects, links, and optionally tracks, Fig. 4B) are stored in tabular form as comma-separated values (CSV) files, while metadata and schema information are stored as a JSON file (Fig. 4C). The development of the **biotracks** format is complementary to the OMEGA system for particle tracking data, which has particular

emphasis on results from viral and vesicular trafficking experiments [31].

Standardization is also required in other parts of the experimental process, such as standards to report how the analytical results were obtained, methods for segmentation and other image-processing tasks, descriptions of post-image data exclusion and curation, and descriptions of statistical analyses [32]. These aspects are left for future work by the CMSO community.

Each of the aforementioned formats (ISA, OME, and biotracks) have associated software to manipulate them, which we introduce below, together with other APIs and tools that facilitate in building machine-actionable and FAIR cell migration data.

The ISA model has an associated set of open source tools [26]. In particular, the ISAcceptor desktop-based tool allows for the creation, parsing, and validation of experiments described with the ISA model. A version of ISAcceptor is made available including MIACME configurations or templates [33]. On the other hand, the ISA-API Python-based software [34] supports the programmatic creation and manipulation of experimental metadata.

The OME model for imaging metadata is supported by several software packages, most notably the Java Bio-Formats library [35], which can read and write OME’s own OME-TIFF standard as well as convert a wide variety of proprietary file formats into OME-TIFF [36] and, more recently, OME-Files [37], which serves as a reference implementation of OME-TIFF in C++ and Python.

For the analytical routines downstream, we have developed a library for cell tracking data: the **biotracks** API [38, 39]. As shown in Fig. 5, the library takes as input a cell tracking file from a tracking software (such as TrackMate [40], CellProfiler [41], Icy [42], and MosaicSuite [43]) and produces a data package where objects and links are stored in the standardized format depicted in Fig. 4.

The CellMissy software package [44, 45, 46], a cross-platform data management and analysis system for cell migration/invasion data, was extended to import and export datasets whose experimental metadata are available in MIACME-compliant ISA-Tab format and whose cell tracking data are represented with biotracks.

A cell migration data repository [47] was created, which accepts submissions of experimental metadata in ISA-Tab for-

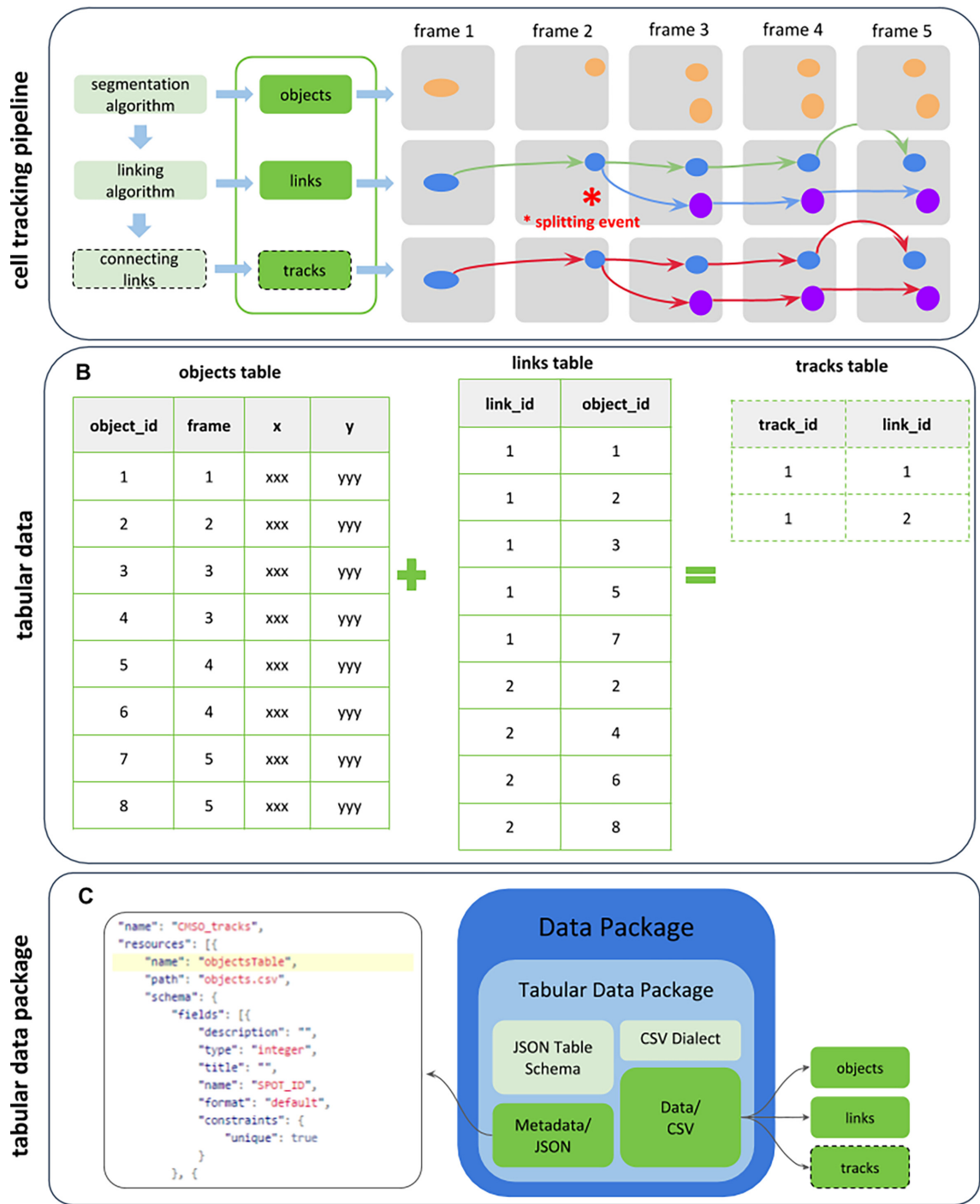


Figure 4: Schematic view of the biotracks format developed by the CMSO. In **A**, a segmentation algorithm identifies objects in the raw images, annotating them with the frame information, coordinates, and any other features that the algorithm extracts. These are described in the objects table in **B**. A linking algorithm then connects the objects across frames in a parent-child relationship. Among the possible events, the linking algorithm can then identify a split, where a parent has >1 child. This information is reported in the links table in **B**. The tracks table in **B** can finally be inferred from the objects and links tables. The tabular data package format is represented in **C**. Here, objects, links, and tracks data tables are saved as comma-separated values (CSV) files. The accompanying file in the JSON format contains both the general metadata of the data package and the metadata of the CSV files.

mat compliant with the MIACME guidelines together with raw data submission, and supports searching across the deposited data.

The CMSO standards were incorporated into the WiSoft platform, a commercial software package developed and distributed by IDEA Bio-Medical Ltd, which consists of 2 software tools—

Athena and Minerva—designed to support the addition of experimental parameters, imaging properties, and analysis modules. This extensibility feature facilitated the adoption of the MIACME elements and allows easy adaptation of algorithms contributed by researchers as part of the ongoing effort of analyzing dynamic biological data.

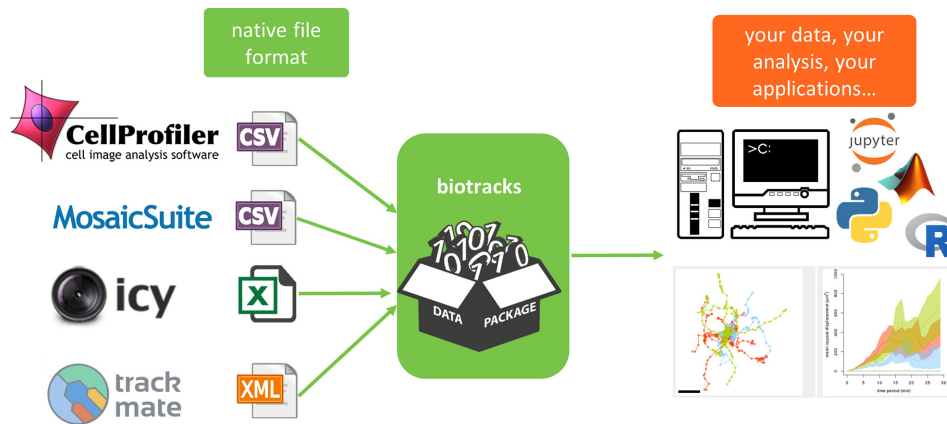


Figure 5: The biotricks library. The library receives cell tracking data as input from multiple tracking software and converts them to the biotricks format (see Fig. 4), which can be further visualized and analysed with downstream applications within this framework.

CMSO standards in action

To demonstrate the application of the CMSO standards we applied them to the study by Masuzzo et al. [44], which proposed an end-to-end software solution for the visualization and analysis of high-throughput single-cell migration experiments. The authors used 2 datasets to demonstrate their software. Here we reuse their Ba/F3 cells experiment to demonstrate CMSO standards in action as follows. As a first step, we annotated the data using the MIACME guidelines (v1.1) (see example in Supplementary Table 1 [48]). The 2 columns of this table represent the 2 layers of information of Fig. 2: specific entities (column 1) are annotated using (controlled) terms (column 2). As shown through this example, the MIACME schema converts a large imaging-based study into an easy-to-interpret structured description. The resulting metadata are available in GitHub [49].

As shown above, when reporting an experiment, researchers will need to complete the information indicated in MIACME, and for those fields that require a CV or ontology term, they will need to select the most appropriate terms, considering the suggested ontologies for each field (e.g., for organisms, MIACME recommends the NCBI Taxonomy). The criterion for selecting a term is to consider the most specific description available in the ontology. If no term exists in current ontologies (designated with asterisk in Supplementary Table 1), the CMSO community has been submitting it to a relevant ontology. If there is no relevant ontology, the CMSO community has been gathering the terms to create a specific ontology in the future if necessary.

This MIACME information is included in an ISA-Tab representation of the dataset, which also expands it with more information about the processes performed in the experiment and their intermediate inputs and outputs.

Second, to demonstrate the use of the data formats and APIs, we have prepared an interactive Jupyter notebook [50]. The notebook uses the set of 3 software libraries discussed above: ISA-tools to manipulate the experimental set-up information, OME-files for the imaging data and metadata, and biotricks for the cell tracking data. The pipeline presented in the notebook bundles the 3 libraries together, showing the interaction of the CMSO standards in a complete experimental and analytical workflow.

Discussion: The Overall Vision

Genomics, proteomics, and structural biology have greatly benefited from well-developed data standards [51] that contributed to rapid progress in these fields. However, in the field of cell migration, the lack of unifying standards and repositories has limited the opportunities to make similar progress. Thus, expensive and difficult-to-generate imaging data are stored at local laboratories with no further access for the community and with no standardized descriptions of the experiments that generated them. CMSO has taken initiatives to increase accessibility and reproducibility of cell migration data across models, by developing an open access reporting structure that aims to accommodate diverse types and complexities of cell migration data. The use of standardized and unambiguous terminology and structured metadata in reporting experiments will allow other researchers to more easily reproduce these experiments. We consider this effort as a first step towards standardization of cell migration data, which will facilitate integration, validation, and meta-analysis of cell migration data across models, and foster progress across study and model comparison, enabling the validation of new discoveries.

In this work, we presented a framework around community-driven standards and tools, developed by CMSO through an open process, for managing cell migration data along its data life cycle. The CMSO framework relies on established standards for experimental metadata (ISA) and imaging data (OME), both complemented with models and tools developed by CMSO. We introduced reporting guidelines that identify what elements should be reported for cell migration experiments (MIACME), and a format for cell tracking data. We also provide APIs and software tools supporting the description and publication of cell migration experiments, their workflows, and results.

Open cell migration data following data standards and the tools to manipulate them will enhance the performance and relevance of the field and deepen insight into this complex biological progress beyond the impact of primary research. This necessitates future actions and tools moving forward: the community must improve the user-friendliness of the routine processes of data curation, deposition, and exchange. It also must be ensured that the data standards continuously evolve to meet community needs. Eventually, the community must strive toward making maximal use of the data: contributing public data; developing

new data-driven computational tools; mining for patterns that can drive new biological hypotheses; testing new hypotheses in experimental, computational, and clinical models; and unlocking new knowledge that drives scientific progress and yields new therapies and strategies to improve health.

Outlook: CMSO sustainability

Proper standards and broad community support are crucial for the establishment of a long-term open data sharing ecosystem for cell migration research. Therefore, the output of CMSO is a crucial cornerstone for the implementation of such an ecosystem. The European Union Horizon 2020 MULTIMOT project [52] has been the initial driving force for the development of CMSO, from establishing the organization to arranging the first meetings, enabling the involvement of the broader cell migration community. From the start CMSO was planned as an independent entity, including people beyond MULTIMOT, with its own governance structure [53] and a community-driven decision mechanism. MULTIMOT members are required to implement the CMSO standards, and therefore they constitute the first users and quality assurance testers. CMSO members are responsible for the dissemination of the materials produced and for the sustainability of the organization in the future through funding.

One of the prime goals of CMSO is to raise community awareness on best practices for data stewardship, to promote and disseminate the use of community standards. CMSO is managed and run by volunteers from the community and is open to participation from anyone interested. Current CMSO participants include cell biologists, immunologists, cancer researchers, medical professionals from laboratory medicine, microscopists, computational biologists, and data scientists [54].

The CMSO also welcomes cell migration data contributions from the scientific community and provides guidelines for creating MIACME-compliant descriptions of experiments and using CVs to annotate them.

Methods

Compilation of use cases

To provide incentives for cell migration researchers to invest the time and effort required to structure their data and make it FAIR, CMSO identified a series of use cases where applying the CMSO framework would enable data integration and data reuse and would drive further scientific discoveries.

Combining harmonized data (e.g., from the same cell culture model retrieved from different studies [55–58]) will facilitate data analysis and mining across imaging acquisition techniques, set-ups, and cell lines, with applications such as (i) comparison of results from 2D and 3D culture environments that make use of the same cell model, (ii) validation of *in vitro* results against datasets from *in vivo* experiments in model organisms (e.g., zebrafish embryos, mouse models of disease) to ascertain *in vivo* relevance, or (iii) systematic comparison of the relative dose effects of growth factors or chemical compounds on migration behaviors across cell models.

As an important further opportunity, the reuse of existing primary image data with new analyses can reveal previously unexplored patterns contained in such complex data [55–58]. Typical examples for secondary reuse [7] of multiparametric imaging datasets result from applying novel computational algorithms to derive kinetic shape features (e.g., leading edge oscillations or membrane curvature changes) or functional components such

as switch behaviours and stochasticity in cell population behaviour [9, 59]. Furthermore, public datasets can serve as benchmarks for comparing the performance of computational and analytical methods. When they include proper metadata annotation, such datasets are invaluable for developing new methods [60–62] and training machine learning algorithms in a variety of analytical tasks [63]. As resources for the development and comparison of methods, the value of large amounts of publicly available image data cannot be overstated [64–66].

FAIR standards for cell migration

Since the turn of the century, there have been standardization efforts in different domains: genomics [67, 68], proteomics [69–71], and metabolomics [72–74]. Traditionally, these efforts considered each of the aspects (content, terminology, formats) independently (see the FAIRsharing repository of standards [51]). Different communities developed minimum information guidelines in narrative form (e.g., MIAME for microarray [75], MIAPE for proteomics [69]), and while they encouraged the use of ontologies [76] and formats [70] for annotation and representation, respectively, they developed them separately and emphasized that the guidelines were implementation independent [69]. Whereas independent reporting guidelines imply that they can be implemented in different formats with different semantic models, the importance of producing data that are truly FAIR in cell migration and the breadth of assays and technologies that are used in the field mandate that a reporting guideline in narrative form is no longer sufficient. The need for FAIR data necessitates a reference implementation, i.e., a standard software tool(s) that reads and writes in a standardized format that meets the specification, along with clear, usable examples that show how to implement and use the format. Finally, it is essential to provide validation tool(s) so that scientists and technology developers attempting to adopt the standard can be assured that their work is compliant with the reporting guidelines. A comprehensive approach that encompasses the content, terminology, and format and thus considers the full machine-actionable model for data description is needed. This is the strategy chosen by CMSO. In this way, the minimal reporting checklist is a first step towards the identification of the metadata elements required for a full data description model, especially in view of the FAIR data principles [8] and FAIR data models [77, 78]. We chose a variety of formats to represent the checklist, including a machine-actionable and FAIR representation based on JSON-schemas for JSON-LD data.

To produce these models and tools following a community-driven approach, during face-to-face workshops as well as via online tools, we run the following activities:

Identified metadata descriptors for cell migration experiments based on the model used by the CellMissy software tool. These metadata descriptors were divided into 3 categories: experimental set-up, imaging condition, and cell migration data. Then, a group of cell migration researchers ranked the descriptors according to 3 values representing categories of important, somewhat important, and not so important. The analysis of the ranking provided the first guidance to build the initial MIACME guidelines.

After choosing a representative article on cell migration [79], a survey was prepared and distributed to researchers asking them to complete the values of the identified metadata descriptors considering the experimental description given in the article. By asking them to complete the values, we could check whether it was easy to identify the relevant element in the arti-

cle, and how clear the explanation about the descriptor was. The descriptors were split in multiple sections: general experiment overview and description, cell system description, cell culture conditions, assay description, vessel, plate and environment information, perturbation and intervention, imaging, image analysis information, licensing and terms of use, and request for feedback. Again, we requested researchers to rate each element on a 1–5 scale, from essential to useless.

The above steps and discussions with researchers allowed us to refine the metadata descriptors to be included in MIACME, and in this way we developed several versions of the checklist.

During this iterative process, we also identified the values for the different elements for a variety of experiments, some published and some unpublished. The feedback and discussions among researchers about the importance of each descriptor were crucial to keep refining the checklist and settle on a set of descriptors deemed minimal but also sufficient to enable the comprehension and replicability of the experiment.

In terms of semantic annotations, we found terms in existing ontologies when available and otherwise requested the addition of terms in relevant ontologies (e.g., [80])

The development of a common standard format to represent cell tracking data associated with cell migration experiments aimed to produce a simple and extensible format, reuse existing standard formats where possible, and support both human- and machine-readable metadata. The adopted solution relies on the Tabular Data Package, which supports the association of data with a JSON file to specify metadata and schema information, and resulted in a specification [29] and a Python-based software tool (biotracks) to manage the new format.

We also compiled experimental datasets to demonstrate the application of the different standards and how they integrate in the CMSO framework.

Availability of Supporting Data

Snapshots of our code and other supporting data are available in the GigaScience repository, GigaDB [81].

Additional Files

Supplementary Table 1: A MIACME-annotated cell migration study.

Abbreviations

API: Application Programming Interface; CHEBI: Chemical Entities of Biological Interest; CLO: Cell Line Ontology; CMSO: Cell Migration Standardisation Organisation; CSV: comma-separated values; CV: controlled vocabulary; ERO: eagle-i Research Resource Ontology; FAIR: Findable, Accessible, Interoperable, and Reusable; GDPR: General Data Protection Regulation; GO: Gene Ontology; ISA: Investigation, Study, Assay; JSON: JavaScript Object Notation; JSON-LD: JavaScript Object Notation for Linked Data; JSON-schema: JavaScript Object Notation Schema; MIACME: Minimum Information for Cell Migration Experiments; MIAME: Minimum Information About a Microarray Experiment; MIAPE: Minimum Information About a Proteomics Experiment; NCBI: National Center for Biotechnology Information; OBI: Ontology for Biomedical Investigation; OME: Open Microscopy Environment; OME-TIFF: Open Microscopy Environment Tagged Image File Format; WG: Working Group.

Authors' Contributions

A.G.B. and P.M. contributed equally to writing the manuscript and addressed all authors' comments and contributions. A.G.B. led the development of the MIACME reporting guidelines, coordinating the community feedback and updates, with contributions from P.R.S., P.M., M.V.T., A.Z., R.H.E., and L.M. and feedback from cell migration researchers involved in CMSO via face-to-face meetings and online communications. A.G.B. wrote the MIACME specification; P.R.S. and A.G.B. produced MIACME-compliant ISA-Tab configurations. A.G.B. produced MIACME JSON-schemas and JSON-LD context files. P.M. and S.B. led the development of the biotracks specification with contributions from S.L., G.S., J.M., and the CMSO community. P.M. led the development of the biotracks package with contributions from S.L. and G.S. G.S. led the extension of CellMissy to support the CMSO standards. J.P. developed the cell migration repository. Y.P. extended the IDEA Bio-Medical Ltd. software to support the CMSO standards. P.R.S. produced the ISA-Tab representation of cell migration studies published by P.M. P.R.S. provided the list of recommended ontologies. P.M., A.G.B., and S.L. produced the examples of CMSO datasets with their MIACME, ISA-Tab, OME, and biotracks representations and code to validate them. S.B., A.G.B., S.L., P.M., and M.V.T. produced the training material. A.G.B. created the FAIRsharing collection for CMSO and included its widget in the CMSO website. A.G.B., S.B., L.M., and P.M. produced the CMSO website. All authors contributed to, read, and approved the final manuscript.

Acknowledgements

The Cell Migration Standardisation Organisation acknowledges funding from the European Union's Horizon 2020 Programme under the MultiMot project, Grant Agreement 634107 (PHC32–2014).

References

1. Meijering E, Carpenter AE, Peng H, et al. Imagining the future of bioimage analysis. *Nat Biotechnol* 2016;**34**:1250–5.
2. Peng H, Zhou J, Zhou Z, et al. Bioimage informatics for big data. *Adv Anat Embryol Cell Biol* 2016;**219**:263–72.
3. Macklin P. Key challenges facing data-driven multicellular systems biology. *Gigascience* 2019;**8**(10), doi:10.1093/gigascience/giz127.
4. Chervitz SA, Deutsch EW, Field D, et al. Data standards for Omics data: the basis of data sharing and reuse. *Methods Mol Biol* 2011;**719**:31–69.
5. Gomez-Cabrero D, Abugessaisa I, Maier D, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol* 2014;**8** (Suppl 2):11.
6. Masuzzo P, Martens L2014 Cell Migration Workshop Participants, et al., 2014 Cell Migration Workshop Participants An open data ecosystem for cell migration research. *Trends Cell Biol* 2015;**25**:55–8.
7. Zaritsky A. Sharing and reusing cell image data. *Mol Biol Cell* 2018;**29**:1274–80.
8. Wilkinson MD, Dumontier M, Jan Aalbersberg JJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
9. Te Boekhorst V, Preziosi L, Friedl P. Plasticity of cell migration in vivo and in silico. *Annu Rev Cell Dev Biol* 2016;**32**:491–526.
10. Zaritsky A. Cell biologists should specialize, not hybridize. *Nature* 2016;**535**:325.

11. Shafqat-Abbasi H, Kowalewski JM, Kiss A, et al. An analysis toolbox to explore mesenchymal migration heterogeneity reveals adaptive switching between distinct modes. *eLife* 2016;5:e11384.
12. CMSO Github page. <https://github.com/CellMigStandOrg>. Accessed 4 May 2020.
13. Main website for the Cell Migration Standardisation Organisation. <http://cmso.science>. Accessed 4 May 2020.
14. Gonzalez-Beltran A, Rocca-Serra P. CellMigStandOrg/MIACME: MIACME v1.1. Zenodo 2019, doi:10.5281/zenodo.3457561.
15. Minimum Information About Cell Migration Experiments (MIACME). <http://cmso.science/MIACME/>. Accessed 4 May 2020.
16. Cell Migration Standardisation Organisation. MIACME; Minimum Information About Cell Migration Experiment. 2018, doi.org/10.25504/FAIRsharing.vh2ye1.
17. Sansone S-A, McQuilton P, Rocca-Serra P, et al. FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol* 2019;37:358–67.
18. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533:452–4.
19. Turner L, Shamseer L, Altman DG, et al. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev* 2012;1:60.
20. JSON Schema homepage. <https://json-schema.org/>. Accessed 4 May 2020.
21. Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. *Nat Genet* 2012;44:121–6.
22. Orchard S, Montecchi-Palazzi L, Hermjakob H, et al. The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. *Pac Symp Biocomput* 2005:186–196.
23. FAIRsharing collection of standards and databases relevant for the CMSO. <https://fairsharing.org/collection/CellMigrationStandardisationOrganisation>. Accessed 4 May 2020.
24. Rocca-Serra P, Brandizi M, Maguire E, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 2010;26:2354–6.
25. Goldberg IG, 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 Biol* 2005;6:R47.
26. ISA tools. Standardizing metadata for scientific experiments. <http://isa-tools.org>. Accessed 4 May 2020.
27. ISA Model and Serialization Specifications. <http://isa-specs.readthedocs.org>. Accessed 4 May 2020.
28. OME. <http://www.openmicroscopy.org/>. Accessed 4 May 2020.
29. CMSO Tracks. <http://cmso.science/Tracks/>. Accessed 4 May 2020.
30. Tabular Data Package. <http://specs.frictionlessdata.io/tabular-data-package>. Accessed 4 May 2020.
31. Rigano A, Galli V, Clark JM, et al. OMEGA: a software tool for the management, analysis, and dissemination of intracellular trafficking data that incorporates motion type classification and quality control. *BioRxiv* 2018, doi:10.1101/251850.
32. Strömblad S, Lock JG. Using Systems Microscopy To Understand The Emergence Of Cell Migration From Cell Organization. *Methods Mol Biol* 2018;1749:119–34.
33. Templates to build MIACME-compliant metadata in ISAcreeator. <https://github.com/CellMigStandOrg/ISAcreeator-MIACME>. Accessed 4 May 2020.
34. ISA tools API. <https://github.com/isa-tools/isa-api>. Accessed 4 May 2020.
35. Linkert M, Rueden CT, Allan C, et al. Metadata matters: access to image data in the real world. *J Cell Biol* 2010;189:777–82.
36. Bio-Formats supported formats. <https://docs.openmicroscopy.org/bio-formats/5.9.2/supported-formats.html>. Accessed 4 May 2020.
37. Leigh R, Gault D, Linkert M, et al. OME Files - An open source reference library for the OME-XML metadata model and the OME-TIFF file format. *BioRxiv* 2016, doi:10.1101/088740.
38. Leo S, Besson S, Sergeant G, et al. CellMigStandOrg/biotracks: Release of biotricks 0.5.0. Zenodo 2019, doi:10.5281/zenodo.3355530.
39. Biotricks specification. <https://github.com/CellMigStandOrg/biotricks>. Accessed 4 May 2020.
40. Tinevez J-Y, Perry N, Schindelin J, et al. TrackMate: An open and extensible platform for single-particle tracking. *Methods* 2017;115:80–90.
41. Bray M-A, Carpenter AE. CellProfiler Tracer: exploring and validating high-throughput, time-lapse microscopy image data. *BMC Bioinformatics* 2015;16:368.
42. Chenouard N, Bloch I, Olivo-Marin J-C. Multiple hypothesis tracking for cluttered biological image sequences. *IEEE Trans Pattern Anal Mach Intell* 2013, doi:377077B0-C943-4376-B928-2872987BC8F1.
43. Szbalzarini IF, Koumoutsakos P. Feature point tracking and trajectory analysis for video imaging in cell biology. *J Struct Biol* 2005;151:182–95.
44. Masuzzo P, Huyck L, Simiczyjew A, et al. An end-to-end software solution for the analysis of high-throughput single-cell migration data. *Sci Rep* 2017;7:42383.
45. Sergeant G, Hulstaert N, Masuzzo P, et al. compomics/cellmissy: Cell Migration Invasion Storage System. Latest version. Zenodo. 2019, doi:10.5281/zenodo.1493927
46. Cellmissy: Cell Migration Invasion Storage System. <https://github.com/compomics/cellmissy>. Accessed 4 May 2020.
47. Cell Migration WIS Repository. <https://repo.cellmigration.org/>. Accessed 4 May 2020.
48. Rochelle T, Daubon T, Van Troys M, et al. p210bcr-abl induces amoeboid motility by recruiting ADF/destrin through RhoA/ROCK1. *FASEB J* 2013;27:123–34.
49. Example of dataset metadata described with CMSO standards. <https://github.com/CellMigStandOrg/CMSO-datasets/tree/master/cmsodataset0001-masuzzo>. Accessed 4 May 2020.
50. Notebook demonstrating the interaction of CMSO standards. https://github.com/CellMigStandOrg/CMSO-training/blob/master/notebooks/CMSO_PM.ipynb. Accessed 4 May 2020.
51. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database (Oxford)* 2016, doi:10.1093/database/baw075.
52. MULTIMOT. <https://multimot.org/>. Accessed 4 May 2020.
53. Roles and responsibilities of the CMSO. <https://cmso.science/roles-and-responsibilities/>. Accessed 4 May 2020.
54. How to get involved in the CMSO. <https://cmso.science/how-to-get-involved/>. Accessed 4 May 2020.
55. Pasquetto IV, Randles BM, Borgman CL. On the reuse of scientific data. *Data Sci J* 2017;16, doi:10.5334/dsj-2017-008.
56. Zaritsky A, Welf ES, Tseng Y-Yu, et al. Seeds of locally aligned

- motion and stress coordinate a collective cell migration. *Biophys J* 2015;**109**:2492–500.
57. Maiuri P, Rupprecht JF, Wieser S, et al. Actin flows mediate a universal coupling between cell speed and cell persistence. *Cell* 2015;**161**:374–86.
 58. Lavi I, Piel M, Lennon-Duménil A-M, et al. Deterministic patterns in cell motility. *Nat Phys* 2016;**12**:1146–52.
 59. Haeger A, Krause M, Wolf K, et al. Collective invasion of mesenchymal tumor cells imposed by tissue confinement. *Biochim Biophys Acta* 2014;**1840**:2386–95.
 60. Zaritsky A, Obolski U, Gan Z, et al. Decoupling global biases and local interactions between cell biological variables. *eLife* 2017;**6**, doi:10.7554/eLife.22323.
 61. Eibl RH, Benoit M. Molecular resolution of cell adhesion forces. *IEEE Proc Nanobiotechnol* 2004;**151**:128–32.
 62. Eibl RH, Moy VT. Atomic force microscopy measurements of protein-ligand interactions on living cells. *Methods Mol Biol* 2005;**305**:439–50.
 63. Sullivan DP, Lundberg E. Seeing more: A future of augmented microscopy. *Cell* 2018;**173**:546–8.
 64. Williams E, Moore J, Li SW, et al. The image data resource: A bioimage data integration and publication platform. *Nat Methods* 2017;**14**:775–81.
 65. Sharing images. *Nat Methods* 2017;**14**:753.
 66. Ellenberg J, Swedlow JR, Barlow M, et al. A call for public archives for biological image data. *Nat Methods* 2018;**15**:849–54.
 67. Field D, Sterk P, Kottmann R, et al. Genomic standards consortium projects. *Stand Genomic Sci* 2014;**9**: 599–601.
 68. Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008;**26**:541–7.
 69. Taylor CF, Paton NW, Lilley KS, et al. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* 2007;**25**:887–93.
 70. Martens L, Chambers M, Sturm M, et al. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011;**10**, doi:10.1074/mcp.R110.000133.
 71. Deutsch EW, Albar JP, Binz PA, et al. Development of data representation standards by the Human Proteome Organization Proteomics Standards Initiative. *J Am Med Inform Assoc* 2015;**22**:495–506.
 72. MSI Board Members et al. The metabolomics standards initiative. *Nat Biotechnol* 2007;**25**:846–8.
 73. Salek RM, Steinbeck C, Viant MR, et al. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience* 2013;**2**, doi:10.1186/2047-217X-2-13.
 74. Sumner LW, Amberg A, Barrett D, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 2007;**3**:211–21.
 75. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;**29**:365–71.
 76. Mayer G, Montecchi-Palazzi L, Ovelheiro D, et al. The HUPO Proteomics Standards Initiative- mass spectrometry controlled vocabulary. *Database (Oxford)* 2013;**2013**:bat009.
 77. Data models to GO-FAIR. *Nat Genet* 2017;**49**:971.
 78. González-Beltrán A, Maguire E, Sansone S-A, et al. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics* 2014;**15** (Suppl 14):S4.
 79. Lock JG, Mamaghani MJ, Shafqat-Abbasi H, et al. Plasticity in the macromolecular-scale causal networks of cell migration. *PLoS One* 2014;**9**:e90593.
 80. Open issue on the IAO Github on missing CV terms. <https://github.com/information-artifact-ontology/IAO/issues/212>. Accessed 4 May 2020.
 81. Gonzalez-Beltran AN, Masuzzo P, Ampe C, et al. Supporting data for “Community Standards for Open Cell Migration Data”. *GigaScience Database* 2020.<http://dx.doi.org/10.5524/100738>.