HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2022 April 14.

Published in final edited form as:

Nat Methods. 2022 March ; 19(3): 262–267. doi:10.1038/s41592-022-01415-4.

*These authors contributed equally

**Co-Corresponding Authors

The HTAN Consortium

Daniel L. Abravanel[28], Samuel Achilefu[18], Foluso O. Ademuyiwa[18], Andrew C. Adey[11], Rebecca Aft[18], Khung Jun Ahn[38], Fatemeh Alikarami[38], Shahar Alon[33], Orr Ashenberg[2], Ethan Baker[2], Gregory J. Baker[1], Shovik Bandyopadhyay[36], Peter Bayguinov[18], Jennifer Beane[23], Winston Becker[19], Kathrin Bernt[38], Courtney B. Betts[11], Julie Bletz[10], Tim Blosser[32], Adrienne Boire[9], Genevieve M. Boland[28], Edward S. Boyden[34], Elmar Bucher[11], Raphael Bueno[28], Qiuyin Cai[17], Francesco Cambuli[9], Joshua Campbell[23], Song Cao[18], Wagma Caravan[18], Ronan Chaligné[9], Joseph Chan[9], Sara Chasnoff[18], Deyali Chatterjee[18], Alyce A. Chen[1], Changya Chen[38], Chia-hui Chen[38], Bob Chen[17], Feng Chen[18], Siqi Chen[18], Milan G. Chheda[18], Koei Chin[11], Hyeyoung Cho[11], Jaeyoung Chun[9], Luis Cisneros[30], Robert J. Coffey[17], Ofir Cohen[2], Graham A. Colditz[18], Kristina A. Cole[38], Natalie Collins[12], Daniel Cotter[19], Lisa M. Coussens[11], Shannon Coy[24], Allison L. Creason[11], Yi Cui[34], Daniel Cui Zhou[18], Christina Curtis[19], Sherri R. Davies[18], Inode Bruijn[9], Toni M. Delorey[2], Emek Demir[11], David Denardo[18], Dinh Diep[40], Li Ding[18], John DiPersio[18], Steven M. Dubinett[29], Timothy J. Eberlein[18], James A. Eddy[10], Edward D. Esplin[19], Rachel E. Factor[30], Kayvon Fatahalian[19], Heidi S. Feiler[11], Jose Fernandez[19], Andrew Fields[11], Ryan C. Fields[18], James A.J. Fitzpatrick[18], James M. Ford[19], Jeff Franklin[17], Bob Fulton[18], Giorgio Gaglia[24], Luciano Galdieri[18], Karuna Ganesh[9], Jianjiong Gao[9], Benjamin L. Gaudio[1], Gad Getz[2], David L. Gibbs[7], William E. Gillanders[18], Jeremy Goecks[11], Daniel Goodwin[34], Joe W. Gray[11], William Greenleaf[19], Lars J. Grimm[30], Qiang Gu[11], Jennifer L. Guerriero[12], Tuhin Guha[19], Alexander R. Guimaraes[11], Belen Gutierrez[19], Nir Hacohen[33], Casey Ryan Hanson[19], Coleman R. Harris[17], William G. Hawkins[18], Cody N. Heiser[17], John Hoffer[1], Travis J. Hollmann[9], James J. Hsieh[18], Jeffrey Huang[38], Stephen P. Hunger[38], Eun-Sil Hwang[30], Christine Iacobuzio-Donahue[9], Michael D. Iglesia[18], Mohammad Islam[38], Benjamin Izar[12], Connor A. Jacobson[1], Samuel Janes[39], Reyka G. Jayasinghe[18], Tiarah Jeudi[12], Bruce E. Johnson[12], Brett E. Johnson[11], Tao Ju[18], Humam Kadara[35], Elias-Ramzey Karnoub[9], Alla Karpova[18], Aziz Khan[19], Warren Kibbe[30], Albert H. Kim[18] Lorraine M. King[30], Elyse Kozlowski[12], Praveen Krishnamoorthy[18], Robert Krueger[32], Anshul Kundaje[19], Uri Ladabaum[19], Rozelle Laquindanum[19], Clarisse Lau[7], Ken Siu Kwong Lau[17], Nicole R. LeBoeuf[28], Hayan Lee[19], Marc Lenburg[23], Ignaty Leshchiner[2], Rochelle Levy[12], Yize Li[18], Christine G. Lian[24], Wen-Wen Liang[18], Kian-Huat Lim[18], Yiyun Lin[35], David Liu[12], Qi Liu[17], Ruiyang Liu[18], Joseph Lo[30], Pierrette Lo[10], William J. Longabaugh[7], Teri Longacre[19], Katie Luckett[9], Cynthia Ma[18], Chris Maher[18], Allison Maier[31], Danika Makowski[38], Carlo Maley[30], Zoltan Maliga[1], Parvathy Manoj[9], John M. Maris[38], Nick Markham[17], Jeffrey R. Marks[30], Daniel Martinez[38], Jay Mashl[18], Ignas Masilionis[9], Joan Massague[9], Marciej A. Mazurowski[30], Eliot T. McKinley[17], Joshua McMichael[18], Matthew Meyerson[12], Gordon B. Mills[11], Zahi I. Mitri[11], Andrew Moorman[9], Jacqueline Mudd[18], George F. Murphy[24], Nataly Naser AL Deen[18], Nicholas E. Navin[35], Tal Nawy[9], Reid M. Ness[17], Stephanie Nevins[19], Ajit Johnson Nirmal[12], Edward Novikov[1], Stephen T. Oh[18], Derek A. Oldridge[36], Kouros Owzar[30], Shishir M. Pant[31], Wungki Park[9], Gary J. Patti[18], Kristina Paul[19], Roxanne Pelletier[1], Daniel Persson[11], Candi Petty[18], Hanspeter Pfister[32], Kornelia Polyak[12], Sidharth V. Puram[18], Qi Qiu[36], Álvaro Quintanal Villalonga[9], Marisol Adelina Ramirez[17], Rumana Rashid[24], Ashley N. Reeb[18], Mary E. Reid[37], Jan Remsik[9], Jessica L. Riesterer[11], Tyler Risom[19], Cecily Claire Ritch[24], Andrea Rolong[17], Charles M. Rudin[9], Marc D. Ryser[30], Kazuhito Sato[18], Cynthia L. Sears[17], Yevgeniy R. Semenov[33], Jeanne Shen[19], Kooresh I. Shoghi[18], Martha J. Shrubsole[17], Yu Shyr[17], Alexander B. Sibley[30], Alan J. Simmons[17], Anubhav Sinha[34], Shamilene Sivagnanam[11], Sheng-Kwei Song[18], Austin Southar-Smith[18], Avrum E. Spira[23], Jeremy St. Cyr[12], Stephanie Stefankiewicz[38], Erik P. Storrs[18], Elizabeth H. Stover[12], Siri H. Strand[19], Cody Straub[30], Cherease Street[18], Timothy Su[17], Lea F. Surrey[38], Christine Suver[10], Kai Tan[38], Nadezhda V. Terekhanova[18], Luke Ternes[11], Anusha Thadi[38], George Thomas[11], Rob Tibshirani[19], Shigeaki Umeda[9], Yasin Uzun[38], Sidharth V. Puram[18], Tuulia Vallius[1], Eliezer R. Van Allen[12], Simon Vandekar[17], Paige N. Vega[17], Deborah J. Veis[18], Sujay Vennam[19], Ana Verma[24], Sebastien Vigneau[12], Nikhil Wagle[12], Richard Wahl[18], Thomas Walle[9], Liang-Bo Wang[18], Simon Warchol[32], M. Kay Washington[17], Cameron Watson[11], Annika K. Weimer[19], Michael C. Wendl[18], Robert B. West[19], Shannon White[19], Annika L. Windon[17], Hao Wu[36], Chi-Yun Wu[36], Yige Wu[18], Matthew A. Wyczalkowski[18], Jason Xu[36], Lijun Yao[18], Wenbao Yu[38], Kun Zhang[40], Xiangzhu Zhu[17]

[27] Arizona State University, Tempe, AZ, USA

[28] Brigham and Women's Hospital, Boston, MA, USA

[29] David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

[30] Duke University Medical Center, Durham, NC, USA

[31] Harvard Medical School, Boston, MA, USA

[32] Harvard University, Cambridge, MA, USA

[33] Massachusetts General Hospital, Boston, MA, USA

[34] Massachusetts Institute of Technology, Cambridge, MA, USA

[35] MD Anderson Cancer Center, Houston, TX, USA

[36] Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[37] Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA

[38] The Children's Hospital of Philadelphia, Philadelphia, PA, USA

[39] Division of Medicine, University College London, London, UK

[40] University of California San Diego, San Diego, CA, USA

Author contributions

D.S., C.Y and A.S. initiated and implemented the MITI guidelines with extensive guidance from other authors and direct supervision by P.K.S. and S.S.. All authors contributed to and reviewed the final MITI guidelines. D.S., C.Y, A.S., P.K.S and S.S. wrote the manuscript with input from all authors.

MITI Minimum Information guidelines for highly multiplexed tissue images

Denis Schapiro[1,2,3,4,*], **Clarence Yapp**[1,5,*], **Artem Sokolov**[1,6,*], **Sheila M. Reynolds**[7], **Yu-An Chen**[1], **Damir Sudar**[8], **Yubin Xie**[9], **Jeremy Muhlich**[1], **Raquel Arias-Camison**[1], **Sarah Arena**[1], **Adam J. Taylor**[10], **Milen Nikolov**[10], **Madison Tyler**[1], **Jia-Ren Lin**[1], **Erik A. Burlingame**[11,26], **Human Tumor Atlas Network**, **Young H. Chang**[11], **Samouil L Farhi**[2], **Vésteinn Thorsson**[7], **Nithya Venkatamohan**[12], **Julia L. Drewes**[13], **Dana Pe'er**[9], **David A. Gutman**[14], **Markus D. Herrmann**[15], **Nils Gehlenborg**[6], **Peter Bankhead**[16], **Joseph T. Roland**[17], **John M. Herndon**[18], **Michael P. Snyder**[19], **Michael Angelo**[19], **Garry Nolan**[19], **Jason R. Swedlow**[20], **Nikolaus Schultz**[21], **Daniel T. Merrick**[22], **Sarah A. Mazzili**[23], **Ethan Cerami**[12], **Scott J. Rodig**[24], **Sandro Santagata**[1,24,**], **Peter K. Sorger**[1,25,**]

[1]Laboratory of Systems Pharmacology, Ludwig Center for Cancer Research at Harvard, Harvard Medical School, Boston, MA, USA

[2]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3]Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University Hospital and Heidelberg University, Heidelberg, Germany.

[4]Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany.

[5]Image and Data Analysis Core, Harvard Medical School, Boston, MA, USA

[6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[7]Institute for Systems Biology, Seattle, WA, USA

[8]Quantitative Imaging Systems LLC, Portland, OR, USA

[9]Program in Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[10]Sage Bionetworks, Seattle, WA, USA

[11]Oregon Health and Science University, Portland, OR, USA

[12]Dana-Farber Cancer Institute, Boston, MA, USA.

[13]Johns Hopkins University School of Medicine, Baltimore, MD, USA

[14]School of Medicine, Emory University, Atlanta, GA, USA

[15]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[16]Edinburgh Pathology, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK.

[17]Vanderbilt University School of Medicine, Nashville TN, USA

[18]Department of Surgery, Washington University School of Medicine, St. Louis, MO USA.

[19]School of Medicine, Stanford University, Stanford, CA, USA

[20]Division of Computational Biology and Centre for Gene Regulation and Expression, University of Dundee, Dundee, UK

[21]Department of Epidemiology & Biostatistics at Memorial Sloan Kettering Cancer Center, New York, NY, USA

[22]Pathology, University of Colorado, Aurora, CO, USA

[23]Boston University, Boston, MA, USA

[24]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA

[25]Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[26]Current Affiliation: Indica Labs, Albuquerque, NM, USA

Abstract

The imminent release of tissue atlases combining multi-channel microscopy with single cell sequencing and other omics data from normal and diseased specimens creates an urgent need for data and metadata standards that guide data deposition, curation and release. We describe a Minimum Information about highly multiplexed Tissue Imaging (MITI) standard that applies best practices developed for genomics and other microscopy data to highly multiplexed tissue images and traditional histology.

Highly multiplexed tissue imaging using any of a variety of optical and mass-spectrometry based methods (Supplemental Table 1) combines deep molecular insight into the biology of single cells with spatial information traditionally acquired using histological methods, such as hematoxylin and eosin (H&E) staining and immunohistochemistry (IHC)[1]. As currently practiced, multiplexed tissue imaging of proteins involves 20–60 channels of 2D data, with each channel corresponding to a different antibody or colorimetric stain (Figure 1). Multiple inter-institutional and international projects, such as the Human Tumor Atlas Network (HTAN)[2], the Human BioMolecular Atlas Program (HuBMAP)[3], and the LifeTime Initiative[4] aim to combine such highly multiplexed tissue images with single cell sequencing and other types of omics data to create publicly accessible "atlases" of normal and diseased tissues. Easy public access to primary and derived data is an explicit goal of these atlases and is expected to encompass native-resolution images, segmented single-cell data, anonymized clinical metadata and treatment history (for human specimens), genetic information (particularly for animal models), and specification of the protocols used to acquire and process the data. Given the imminent release of the first atlases, an urgent need exists for data and metadata standards consistent with emerging FAIR (Findable, Accessible, Interoperable, and Reusable) standards[5]. In this commentary, we establish the MITI (Minimum Information about highly multiplexed Tissue Imaging) standard and associated data level definitions; we also discuss the relationship of MITI to existing standards, practical implementations, and future developments.

Scope and target audiences

MITI covers biospecimen, reagent, data acquisition and data analysis metadata, as well as data levels for imaging with antibodies, aptamers, peptides, dyes and similar detection reagents (Supplemental Table 1). The standard is also compatible with images based on H&E staining, low-plex immunofluorescence (IF) and IHC. A working group is currently extending MITI to cover subcellular resolution imaging of nucleic acids using methods such as MERFISH[6]. While conceived with today's two-dimensional (2D) images in mind (these typically involve 5 – 10 μm thick sections of fixed or frozen specimens), MITI accommodates three-dimensional (3D) datasets acquired using confocal, deconvolution and light sheet microscopes[7]. MITI has been established as its own organization with its own GitHub repository, governing structure, and procedures for proposing and incorporating revisions. The definition of MITI is available in machine readable YAML format (https://github.com/miti-consortium/MITI) along with other relevant information. MITI has also been implemented in practice (https://github.com/ncihtan/data-models) and used to structure metadata available via the HTAN data portal (https://htan-portal-nextjs.vercel.app). However, MITI is independent of HTAN or any single research consortium.

Highly multiplexed imaging is derived from methods such as IHC and IF that are in widespread use in pre-clinical research using cultured cells and model organisms, and in clinical practice with human tissue specimens. Many standards and best practices have been established for these types of data (Supplemental Table 2), but high-plex imaging presents unique challenges: images are expensive to collect and can be very large (up to 1TB in size), specimens are often difficult to acquire and may have data use restrictions, and accurate clinical and genomic annotation is a necessity. Recent interest in highly multiplexed tissue imaging has been driven by applications in oncology, largely due to the importance of the tumor microenvironment in immuno-editing and responsiveness to immunotherapy, but the approach is broadly applicable to studying normal development, infectious disease, immunology and other topics. HuBMAP[3], for example, is using high-plex imaging to study a range of normal human tissues. MITI is also relevant to studies with model organisms and data tables have already been created to store data from genetically engineered mouse models (GEMMs) in a standardized manner.

Multiplexed imaging also promises to impact the pathological diagnosis of diseases, which is rapidly switching to digital approaches[8]. For over a century, histological analysis of anatomic specimens (from biopsies and surgical resection) has been the primary method of diagnosing diseases such as cancer[9], and this remains true today, despite the impact of gene sequencing. Multiplexed tissue imaging promises to augment conventional pathological diagnosis with the detailed molecular information needed to specify use of contemporary precision therapies. This is therefore an opportune time to seek alignment of research and diagnostic approaches by establishing public standards able to take full advantage of the detailed molecular information revealed by emerging imaging methods.

Existing standards and approaches

The Human Genome Project, the Cancer Genome Atlas (TCGA)[10] and similar large-scale genomic programs have developed several approaches to data management of immediate relevance to tissue atlases. The first is the concept of "minimum information" metadata, which has been employed in microarrays (the MIAME standard)[11], genome sequences (MIGS)[12], and biological investigation in general (MIBBI)[13]. The second is the idea of "data levels" (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/data-levels), which specify the extent of data processing (raw, normalized, aggregated or region of interest, corresponding to data levels 1–4) and access control. Access control is required because even anonymized DNA sequencing data pose a re-identification risk[14]. As a result, the database of Genotypes and Phenotypes (dbGaP), the NCI Genome Data Commons (GDC)[15], and the US Federal Register (79 FR 51345) control access to primary sequencing data (so-called level 1 & 2 sequencing data) based on policies set by a data access committee. Higher level genomic data, which are generally more consolidated, involve information aggregated form many patients, and pose little or no re-identification risk can be freely shared[16] (Figure 2). When datasets are combined, they acquire the most stringent restriction applied to any constituent element. While we are not aware of any policies addressing the anonymity of histological images, consultation with our Institutional Review Boards (IRBs, ethics committees) has led us to conclude that public release of tissue images does not constitute a risk to patient privacy. MITI data levels are nonetheless consistent with the existing GDC and dbGaP practice that data intended for unrestricted distribution are classified as level 3 and up. In the case of images adhering to the MITI standard, level 3 data have been subjected to quality control and some degree of human annotation, making them more useful in a shared environment than raw images. We anticipate that IRBs and government agencies will in the future provide further guidance on sharing of datasets that combine clinical history, sequence information, and tissue images; MITI will be adapted to accommodate such guidance.

The MITI standard also draws extensively on image formats developed for cultured cells and model organisms and on a wide variety of open-source software tools (Supplemental Table 3). Noteworthy among these are the Open Microscopy Environment (OME) TIFF standard[17] and the BioFormats[18] approach to standardization of microscopy data. MITI field definitions are harmonized with the QUality Assessment and REProducibility for Instruments and Images in Light Microscopy (QUAREP-LiMi)[19] effort, the Resource Identification Initiative[20], and antibody standardization efforts by the Human Protein Atlas[21] and are also compliant with the recently developed Recommended Metadata for Biological Images initiative[22]. Metadata on model organisms (particularly GEMMs - and patient derived xenografts - PDXs) are aligned with existing standards, many developed for genomic information (see Supplemental Table 2 for a full list of antecedent resources). Well-curated clinical information is essential for the interpretation of data from human specimens but standardizing such information has proven to be a major challenge in the past, for example in TCGA[23]. Thus, HTAN and other current NCI projects focused on human specimens are emphasizing standardization of clinical metadata, and the MITI standard is

designed to closely align with the Genomic Data Commons (GDC) Data Model[24] in this regard (Supplemental Tables 5–6).

All imaging methods generate data that comprises a sequence of intensity values on a raster; multi-spectral imaging simply adds new dimensions to the raster. The cameras that collect H&E and IHC images from bright-field microscopes or high-plex images from fluorescence microscopes generate a raster; ablation-based mass-spectrometry imaging (e.g. MIBI and IMC) is also raster based. As currently defined, MITI specifies that raster images should be stored in the OME-TIFF 6 standard, but OME formats are currently being migrated to a set of next generation file formats (collectively OME-NGFF)[25] to improve scalability and performance on the cloud. MITI will be updated to align with these new formats as they come into general use. Another area of translational and clinical research in which imaging is commonly encountered is radiology, which is almost entirely digital, and uses data interchange standards governed by DICOM (https://www.dicomstandard.org/). DICOM has recently been extended to accommodate both radiology data and OME-TIFF standards[26]. The NCI's ongoing program to create an Imaging Data Commons[27] is expected to be based on this dual standard, or on a successor using OME-NGFF. MITI is, or will be, compatible with these foundational data standards.

In highly multiplexed tissue imaging antibodies are either conjugated to fluorophores directly or via oligonucleotides, or are bound to secondary antibodies (Figure 1, Supplemental Table 4). Images are then acquired serially, one to six channels at a time, to assemble data from 20–60 antibodies. In ablation-based methods, antibodies are labelled with metals and vaporized with lasers or ion beams after which they are detected by atomic mass spectrometry (Supplementary Table 4). In all cases, the raw output of data acquisition instruments comprises Level 1 MITI data (Figure 2), analogous to the Level 1 FASTq files in genomics.

Whole slide imaging is required for clinical applications[28] and also necessary to ensure adequate power in pre-clinical studies[29]. However, resolution and field of view have a reciprocal relationship – both with respect to optical physics and the practical process of mapping image fields onto the fixed raster of a camera (or ablating beam). Whole slide images of histological specimens[8] must therefore be acquired by dividing a large specimen into contiguous tiles. This usually involves acquisition of ~100 to 1,000 tiles by moving the microscope stage in both X and Y, with each tile being a multi-dimensional, subcellular resolution, TIFF image. Tiles are combined at sub-pixel accuracy into a mosaic image in a process known as stitching. When high-plex images are assembled from multiple rounds of lower-plex imaging, it is also necessary to register channels to each other across imaging cycles and to correct for any unevenness in illumination (so-called flat fielding)[30]. Stitched and registered mosaics can be as large as 50,000 × 50,000 pixels × 100 channels and require ~500 GB of disk space. They correspond to Level 2 MITI data and represent full-resolution primary images that have undergone automated stitching, registration, illumination correction, background subtraction, intensity normalization and have been stored in a standardized OME format. The level of processing is analogous to BAM files, a common type of Level 2 data in genomics.

Level 3 data represent images that have been processed with some interpretive intent, which may include (i) full-resolution images following quality control or artifact removal, (ii) segmentation masks computed from such images, (iii) machine-generated spatial models, and (iv) images with human or machine-generated annotations. Level 3 MITI data is roughly analogous to Level 3 mRNA expression data in genomics. However, whereas many users of genomic data only require access to processed level 3 and 4 data, which are usually quite compact, quantitative analysis of tissue images adds a requirement for full-resolution primary images so that images and computed features can be examined in parallel[31]. Level 3 MITI data is intended to be the primary type of image data distributed by tissue atlases and similar projects.

Assembled level 3 images are typically segmented to identify single cells[31], which are quantified to produce a “spatial feature table” that describes marker intensities, cell coordinates and other single-cell features. The Level 4 data in spatial feature tables are a natural complement to count tables in single cell sequencing data (e.g. scRNA-seq, scATAC-seq, scDNA-seq) and can be analyzed using many of the same dimensionality reduction methods (e.g. PCA, t-SNE and U-MAP)[32] and on-line browsers such as cellxgene (Supplemental Table 3)[33]. These types of tabular data are all examples of “Feature Observation Matrixes” which are themselves being standardized across domains of biology to improve their utility and inter-compatibility. Level 5 MITI data comprise results computed from spatial feature tables or primary images. Because access to TB-size full-resolution image data is impractically burdensome when reading a manuscript or browsing a large dataset, a specialized type of Level 5 image data has been developed to enable panning and zooming across images using a standard web browser. In the case of Level 5 images viewed with MINERVA software, the aim is to exploit similar functionality and concepts as those in Google Maps or electronic museum guides[34]. The inclusion of digital docents with images makes it possible to combine pan and zoom with guided narratives that greatly facilitate comprehension of complex datasets and promote new hypothesis generation[35].

For any metadata standard to be used, a balance must be struck between ease of data entry, which minimizes non-compliance by data generators, and level of detail, which must be sufficient for data retrieval, analysis, and publication in a reproducible manner. Moreover, specifying a metadata standard is separate from the essential task of developing a practical and reliable means for capturing information needed to ensure adherence to the standard. Two approaches have proven most effective in addressing this requirement. One, exemplified by OMeta[36], involves a relational database and web interface that data generators use to input necessary information in a controlled manner. Another approach, exemplified by MAGE-TAB[37], involves a standardized format for collecting metadata via a series of structured documents, which are then used to populate web pages and databases[38]. As a practical test of MITI we have implemented the latter approach in a JSON schema (https://github.com/ncihtan/data-models) that also conforms to the design principles of SCHEMA.org. These principles focus on the creation, maintenance and promotion of schemas for structured data that is supported by major web search engines, thereby enhancing discoverability. In this TAB-like approach the MITI standard is exposed to data collectors as Google Sheets with dropdowns representing controlled vocabularies and highlighting required or optional elements; many fields are automatically validated upon

entry. These documents are ingested using SCHEMATIC (Schema Engine for Manifest Ingress and Curation; https://github.com/Sage-Bionetworks/schematic), automatically linked to primary imaging data, and stored as cloud assets. These implementations continue to evolve, and entirely different approaches are possible: nothing in a MITI-type standard constrains how data are collected.

Whereas many research agencies and countries have made a major investment in curating, storing, and distributing genomic data, fewer repositories exist for primary image data. The Image Data Resource[39] maintained by the European Bioinformatics Institute (EBI) is an exception, but as the volume of image data grows, other means of data distribution will almost certainly be required. In the U.S., in the absence of a major public investment in data storage, the development of "requester pays"[40] access to datasets is a promising development. The primary cost associated with creation and maintenance of a dataset on a commercial cloud service involves data download, not data ingress and storage. In a "requester pays" model, a user seeking access to a dataset pays the cost of data egress directly to the cloud provider making access both secure and anonymous (moreover, the cost of egress into another account on the same commercial cloud is low). Although the "requester pays" approach might appear to create an impediment to research, the actual cost of egress is quite low (currently, about hundred US dollars per TB) compared to any form of data acquisition and a key goal is to avoid a tragedy of the commons in which frequent, duplicate downloads overwhelm the system. A combination of a MITI implementation on a cloud service (as described above) with "requester pays" cloud access will also make it possible for individuals to distribute very large FAIR image datasets at relatively low cost. Such an approach does not obviate the need for public investments, such as those being made but EBI, but does represent a practical way forward to democratize release of standardized data – some of which can then be incorporated into publicly supported resources. Regardless, the MITI standard described here is available for immediate use, without being impacted by how access to the primary data is provisioned.

Public data and metadata standards have been essential for the success of genomics and other fields of biomedicine, but the creation of a new standard is no guarantee of successful adoption. An outpouring of effort 10–20 years ago led to the development of widely adopted and well maintained standards such as MIAME[11], MIGS[12] and MIBBI[13], and these have been consolidated and further documented by the Digital Curation Center (https://www.dcc.ac.uk/), FairSharing.org, and similar projects. However many other minimum information projects have been left unattended[41], and it remains unclear whether existing metadata adequately conform to user needs[42]. The development of MITI and of the initial HTAN implementation enjoys NCI support and is expected to become part of the NCI Cancer Research Data Commons[27], helping ensure its viability. However, individuals and organizations are invited to join in the further development of MITI and should make contact via the image.sc forum or submit pull requests (i.e. requests for inclusion in the MITI "code base" at https://github.com/miti-consortium/MITI). Because high high-plex tissue imaging is in its infancy and MITI has attracted the great majority of developers of existing high-plex tissue image acquisition methods, it represents a solid beginning for what will need to be an evolving standard. By having its own repository and governance structure,

independent of any particular research program or constituency, MITI also conforms with other requirements of successful open standards[43].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by the HTAN Consortium and the Cancer Systems Biology Consortium (CSBC). A list of all current Consortium members can be found at https://humantumoratlas.org/.

This work was supported by the following grants from the National Cancer Institute under the Human Tumor Atlas Network (HTAN) U2C CA233262 (Harvard Medical School), U2C CA233280 (OHSU), U2C CA233195 (Boston DFCI Broad), U2C CA233291 (Vanderbilt University Medical Center), U2C CA233311 (Stanford University), U2C CA233238 (Boston University Medical Campus), U2C CA233285 (Children's Hospital of Philadelphia), U2C CA233303 (Washington University St. Louis), U2C CA233280 (Oregon Health and Science University), U2C CA233284 (Memorial Sloan Kettering Cancer Center), U2C CA233254 (Duke University Medical Center) and by other public support including U54 CA225088 (SS, PKS) and U24 CA233243 (Dana-Farber Cancer Institute, Emory University, Institute for Systems Biology, Memorial Sloan Kettering Cancer Center, Sage Bionetworks). DS was funded by an Early Postdoc Mobility fellowship (no. P2ZHP3_181475) from the Swiss National Science Foundation and was a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRQ-03-20); DS is currently supported by the BMBF (01ZZ2004). NG was funded by the NIH Human BioMolecular Atlas Program (HuBMAP) OT2 OD026677 and MDH by NCI/NIH Task Order No. HHSN26110071 under Contract No. HHSN2612015000031.

Competing Interests Statement

PKS is a member of the SAB or BOD of Applied Biomath, RareCyte Inc., and Glencoe Software, which distributes a commercial version of the OMERO data management platform; PKS is also a member of the NanoString SAB and a consultant to Merck and Montai Health. In the last five years the Sorger lab has received research funding from Novartis and Merck. Sorger declares that none of these relationships have influenced the content of this manuscript. SS is a consultant for RareCyte Inc. NG is a co-founder and equity owner of Datavisyn. DS is a consultant for Roche Glycart AG. JRS is Founder and CEO of Glencoe Software, which distributes a commercial version of the OMERO data management platform. SR receives research funding from Bristol-Myers-Squibb, Merck, Affimed, and Kite/Gilead. SR is on the Scientific Advisory Board for Immunitas Therapeutics. DSu is employed by Quantitative Imaging Systems LLC. EAB is an employee of Indica Labs.

Data and Code Availability Statement

The detailed specification of the guidelines outlined in this manuscript are available at https://github.com/miti-consortium/MITI and https://www.miti-consortium.org/

References

1. Bodenmiller B Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. Cell Syst 2, 225–238 (2016). [PubMed: 27135535]
2. Rozenblatt-Rosen O et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell 181, 236–249 (2020). [PubMed: 32302568]
3. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature 574, 187–192 (2019). [PubMed: 31597973]
4. Rajewsky N et al. LifeTime and improving European healthcare through cell-based interceptive medicine. Nature 587, 377–386 (2020). [PubMed: 32894860]
5. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). [PubMed: 26978244]
6. Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090 (2015). [PubMed: 25858977]

7. Fischer RS, Wu Y, Kanchanawong P, Shroff H & Waterman CM Microscopy in 3D: a biologist's toolbox. Trends in Cell Biology 21, 682–691 (2011). [PubMed: 22047760]
8. Ghaznavi F, Evans A, Madabhushi A & Feldman M Digital imaging in pathology: whole-slide imaging and beyond. Annu Rev Pathol 8, 331–359 (2013). [PubMed: 23157334]
9. Amin MB et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging. CA Cancer J Clin 67, 93–99 (2017). [PubMed: 28094848]
10. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120 (2013). [PubMed: 24071849]
11. Brazma A et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet 29, 365–371 (2001). [PubMed: 11726920]
12. Field D et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26, 541–547 (2008). [PubMed: 18464787]
13. Taylor CF et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26, 889–896 (2008). [PubMed: 18688244]
14. Benitez K & Malin B Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 17, 169–177 (2010). [PubMed: 20190059]
15. Grossman RL et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med 375, 1109–1112 (2016). [PubMed: 27653561]
16. Byrd JB, Greene AC, Prasad DV, Jiang X & Greene CS Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet 21, 615–629 (2020). [PubMed: 32694666]
17. Swedlow JR, Goldberg I, Brauner E & Sorger PK Informatics and quantitative analysis in biological imaging. Science 300, 100–102 (2003). [PubMed: 12677061]
18. Li S et al. Metadata management for high content screening in OMERO. Methods 96, 27–32 (2016). [PubMed: 26476368]
19. Nelson G et al. QUAREP-LiMi: A community-driven initiative to establish guidelines for quality assessment and reproducibility for instruments and images in light microscopy. arXiv:2101.09153 [physics, q-bio] (2021).
20. Bandrowski A et al. The Resource Identification Initiative: A cultural shift in publishing. F1000Res 4, (2015).
21. Edfors F et al. Enhanced validation of antibodies for research applications. Nat Commun 9, 4130 (2018). [PubMed: 30297845]
22. Sarkans U et al. REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. Nat Methods 1–5 (2021) doi:10.1038/s41592-021-01166-8. [PubMed: 33408396]
23. Liu J et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 173, 400–416.e11 (2018). [PubMed: 29625055]
24. Zhang Z et al. Uniform genomic data analysis in the NCI Genomic Data Commons. Nat Commun 12, 1226 (2021). [PubMed: 33619257]
25. Moore J et al. OME-NGFF: scalable format strategies for interoperable bioimaging data. 2021.03.31.437929 https://www.biorxiv.org/content/10.1101/2021.03.31.437929v4 (2021) doi:10.1101/2021.03.31.437929.
26. Clunie DA Dual-Personality DICOM-TIFF for whole slide images: A migration technique for legacy software. Journal of Pathology Informatics 10, 12 (2019). [PubMed: 31057981]
27. Fedorov A et al. NCI Imaging Data Commons. Cancer Res canres.0950.2021 (2021) doi:10.1158/0008-5472.CAN-21-0950.
28. Health, C. for D. and R. Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices. U.S. Food and Drug Administration http://www.fda.gov/regulatory-information/search-fda-guidance-documents/technical-performance-assessment-digital-pathology-whole-slide-imaging-devices (2019).
29. Lin J-R et al. Multiplexed 3D atlas of state transitions and immune interactions in colorectal cancer. bioRxiv 2021.03.31.437984 (2021) doi:10.1101/2021.03.31.437984.

30. Peng T et al. A BaSiC tool for background and shading correction of optical microscopy images. Nat Commun 8, (2017).
31. MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging | bioRxiv. https://www.biorxiv.org/content/10.1101/2021.03.15.435473v1.
32. Heiser CN & Lau KS A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. Cell Rep 31, 107576 (2020). [PubMed: 32375029]
33. Megill C et al. cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv 2021.04.05.438318 (2021) doi:10.1101/2021.04.05.438318.
34. Gutman DA et al. The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research. Cancer Res. 77, e75–e78 (2017). [PubMed: 29092945]
35. Rashid R et al. Interpretative guides for interacting with tissue atlas and digital pathology data using the Minerva browser. Nat Biomed Eng. 2020.03.27.001834 (2020) doi:10.1101/2020.03.27.001834.
36. Singh I et al. OMeta: an ontology-based, data-driven metadata tracking system. BMC Bioinformatics 20, 8 (2019). [PubMed: 30612540]
37. Rayner TF et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7, 489 (2006). [PubMed: 17087822]
38. Martínez-Romero M et al. Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. Database 2019, (2019).
39. Williams E et al. The Image Data Resource: A Bioimage Data Integration and Publication Platform. Nat. Methods 14, 775–781 (2017). [PubMed: 28775673]
40. Using Requester Pays buckets for storage transfers and usage - Amazon Simple Storage Service. https://docs.aws.amazon.com/AmazonS3/latest/userguide/RequesterPaysBuckets.html.
41. Kacprzak E et al. Characterising dataset search—An analysis of search logs and data requests. Journal of Web Semantics 55, 37–55 (2019).
42. Löffler F, Wesp V, König-Ries B & Klan F Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? PLOS ONE 16, e0246099 (2021). [PubMed: 33760822]
43. Swedlow JR et al. A global view of standards for open image data formats and repositories. Nat Methods 1–7 (2021) doi:10.1038/s41592-021-01113-7. [PubMed: 33408396]

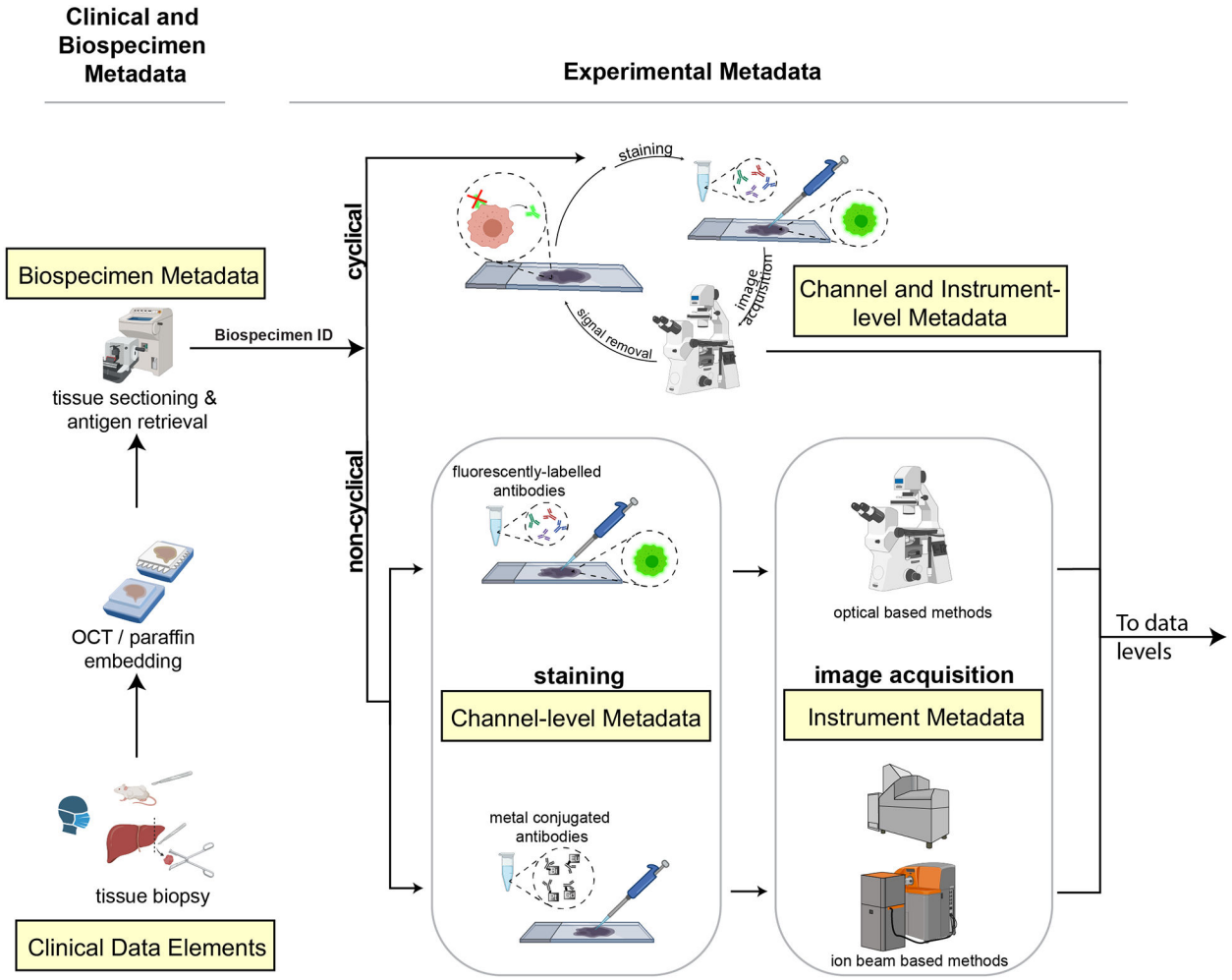

FIGURE 1: Schematic diagram of the steps in a canonical multiplexed tissue imaging experiment and the associated metadata

In a typical workflow, samples collected from patient biopsies and resections or from animal models are formaldehyde fixed and paraffin embedded (FFPE) or frozen and then sectioned and mounted onto either a standard glass microscope slide (for CyCIF, mIHC, IMC, MELC or mxIF), fluidic chamber (for CODEX) or specialized carriers (for MIBI). Clinical and biospecimen metadata (extracted from clinical records, for example) is linked to all other levels of metadata via a unique ID (Biospecimen ID). Data is acquired using cyclical or non-cyclical staining and imaging methods and both reagent and experimental metadata collected (consisting of antibody, reagent and instrument metadata). In both cyclic and non-cyclic methods, sections undergo pre-processing, antigen retrieval, and antibody incubation and images are acquired. In cyclical imaging methods, fluorophores or chromogens are inactivated or removed and additional antibodies and/or visualization reagents are applied and data acquisition repeated. Channel and instrument metadata capture these essential details. Created with BioRender.com.

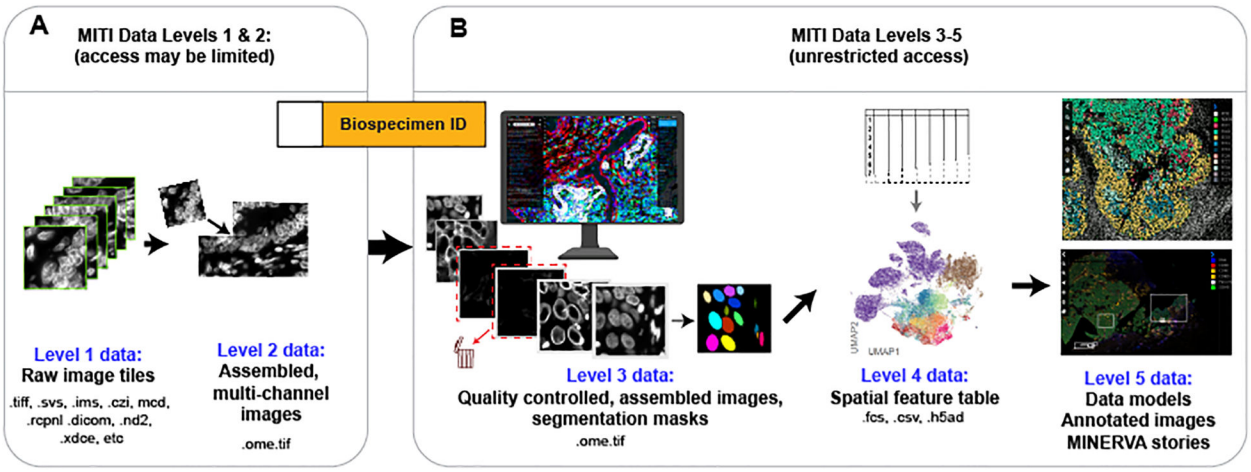

FIGURE 2: MITI data levels and formats

Data levels specify the extent of data processing and, in the case of sequencing data, whether access requires the approval of a data access committee. In common practice, data at levels 3 and up are freely shared. Primary data arising from microscopes and data acquisition instruments corresponds to level 1 data. Because the raw image data acquired from one slide usually consists of separate image fields, possibly from proprietary formats, they are processed to correct for uneven illumination and other instrumentation artifacts and assembled into a single multi-channel image in the OME-TIFF format (level 2 data). OME-TIFF image mosaics undergo quality control (including artefact removal, channel rejection, evaluation of staining quality) to generate full-resolution, assembled and curated level 3 image data; segmentation algorithms generate one or more label masks that also comprise level 3 data. The great majority of users will want to access these level 3 images. Each label mask (e.g., nuclei, cytoplasmic-regions, whole cells, organelles, etc.) is used to compute quantitative features, such as the mean signal intensity, spatial coordinates of individual cells and morphological features, which are stored as level 4 spatial feature tables (where rows represent single cells and columns the extracted cellular features); these data are suitable for analysis using the dimensionality reduction and visualization tools used for other types of single-cell data (e.g. UMAP plots). Spatial models computed from images and spatial feature tables, or by direct application of machine learning to images, as well as images annotated by humans, comprise level 5 data.