EMBRIO INSTITUTE DATA MANAGEMENT & SHARING PLAN

1. Research Data Description

The data, samples, and materials expected to be produced consist of laboratory notebooks, image files
from gel imaging and microscopy experiments, and raw data files from protein and DNA concentration
experiments, colorimetric data from plate readers, and similar. Data also includes the following types.

Table 1: Data Types, Volume, Formats, and Standards

Type	Format	Average Size	No. of Files
Communications	Emails and text messages	0.1 MB	Thousands
Reports	*.docx, pdf, *.pptx	0.1 MB – 10MB	10 – 100
Experimental Data	ND2, CZI, LSM, IMS, JPEG, TIFF, OME-TIFF	1 GB – 100GB	Thousands
Modeling Data	*.m, *.dat, *.xyz, *.vtk, *.txt	10 MB – 1 GB	Thousands
Educational Materials	*.ipynb	10 MB – 1 GB	10 – 100
Figures	*.eps, *.tif	5 MB – 100 MB	Hundreds

Imaging data sets are uncompressed image files that span from single 2D images, 2D time series movies, 3D image volumes, 4D time series image volumes, and potentially 5D time series and wavelength series image volumes. Original raw data and modified/edited data sets will be stored. Formats include standard image formats (e.g. TIF), vendor-proprietary formats (e.g. Nikon ND2, Zeiss CZI and LSM, Imaris IMS), and Open Microscopy Environment Bio-Formats converted data (OME-TIFF)[1].

The data management scientist is responsible for curation of data to meet internal standards agreed upon by EMBRIO members. Individual labs (PIs and lab members) are responsible for formatting their data to meet the standards as well as publishing their data sets to the EMBRIO data repository.

2. Metadata Standards

Data will be analyzed with metadata extraction and annotation apps (e.g. MicroMeta App, OMERO MDEmic)[2] [3]. Hardware metadata provenance for microscopes will be defined, and microscope experimental conditions will be output in JSON format, then attached to the image data sets. Higher-level experimental information will be recorded into a formatted template spreadsheet that follows the REMBI standard[4,5]. These higher-level abstraction standards follow the ISA (Investigation, Study, Assay) architecture and are specifically designed for machine readability and reuse[6]. A future goal is to enhance the metadata with ontological mappings such that knowledge

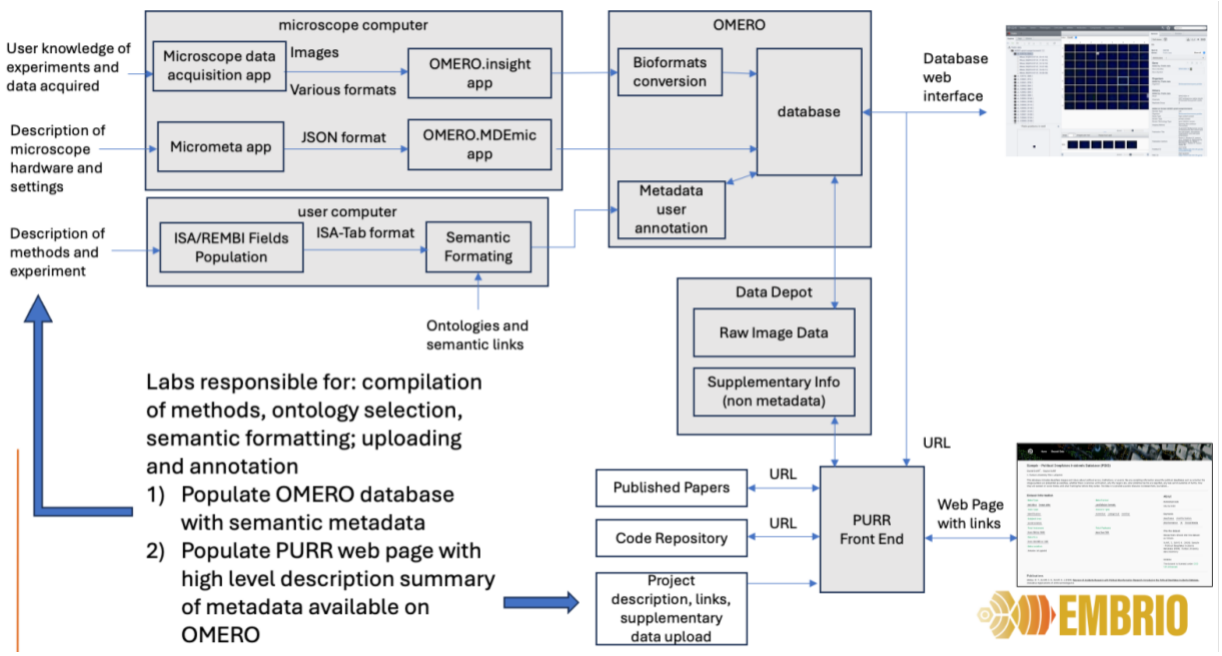
graphs may be built by machine learning data mining algorithms. The architecture for this work follows the data flow promoted by I3D:Bio[7]

The data management scientist will curate the library of metadata terms and depth to meet the requirements agreed upon by the EMBRIO group. Individual labs (PIs and lab members) are responsible for annotating their data in the EMBRIO data repository.

3. Data Sharing Plan

Data management and cross-institutional data sharing throughout the project's lifecycle will leverage two resources from Purdue IT: the Purdue University Research Repository (PURR), and the OMERO rich imaging database server deployed by the Rosen Center for Advanced Computing (RCAC).

Figure 1: Bioimaging Data Capture and Metadata Annotation Schematic



Raw imaging data storage and metadata annotation is managed by a software database product, OMERO, that is maintained by the Open Microscopy Environment (OME), and is hosted at Purdue with persistent public access that is managed by EMBRIO[8].

The OMERO server is deployed at Purdue on a composable platform with extensible storage provided by Data Depot (https://www.rcac.purdue.edu/storage/depot). Funding for this implementation has been supported by NSF Award 2232872, "CC* Data Storage: Software Defined Storage for Composable and HPC Workflows", where the EMBRIO deployment serves as a user case under this award.

Although the OMERO server provides managed access to data to support FAIR principles, having the ability to provide web search capability for this data by higher level project context will assist in making the data findable. To meet this need, a set of EMBRIO web front end pages have been built

that displays the context for each project, including project summary, authors, publication DOI, OMERO raw data DOI, keywords, code repository DOI, and persistent storage for supplementary information that falls outside OMERO. This front end web resource has been deployed using the project management resources of the Purdue University Research Repository (PURR), Purdue's institutional data repository[9]. Built on the HUBzero® platform for scientific collaboration, PURR will support EMBRIO institution faculty, staff, and students throughout the research life-cycle with tools, information and support that meet the requirements of many funding agencies such as the National Science Foundation (NSF) and the National Institutes of Health (NIH). All published datasets are preserved and made accessible for long-term access in a secure, trustworthy digital repository and maintained as university library collections. After publication, PURR provides regular reports on views, downloads, and citations to measure dataset impact.

Each Research and Educational Thrust PI will be responsible for the management and retention of research data. Their role will be to ensure that all members of the project keep reliable records that will be
immediately accessible to both them and the NSF.

4. Schedule for Data Archiving and Sharing

Preliminary data and internal documents will be stored privately across PURR and OMERO. This space will be accessible to all research group members working on this project. However, these preliminary and internal resources will not be available to the general public upon being deposited. Only
once the results are published in the peer-reviewed scientific literature will we consider these public
products (suitable for dissemination beyond our research group) and make the data and requisite documents public digital assets on PURR. All computational biology models and curriculum will be accessible through nanoBIO HUB, which is one of three "nodes" under the NSF's Network for Computational Nanotechnology nanoHUB project.

5. Data Storage and Preservation of Access

Storage will utilize PURR and Data Depot, Purdue's institutional data repositories. Datasets will be deposited into OMERO and will be referenced by PURR and retained therein, as digital assets, for a standard 10-year period. Software under version control will be retained indefinitely.

6. Data Sharing Policy

6.1. Default Data Sharing Status:

By default, a work(data) is considered "all rights reserved," which means others are not permitted to use the work(data) without the owner's or owners' permission.

6.2. Data Use for Publication:

When utilizing another researcher's data for publication purposes, it is imperative to establish communication with the data generator prior to any data analysis is performed. This ensures awareness and allows for inclusion in discussions related to the data's use.

6.3. Principles for Shared Project Data in Publications

We uphold the following principles for the use of shared project data in publications:\

6.3.1 Priority Rights for Data Generators:

- Data generators retain priority rights to determine how the data they produced will initially appear in timely publications, such as in a publication led by the data generator.

- In the spirit of collaboration within EMBRIO, we anticipate and actively strive for publications that incorporate multiple datasets contributed by various generators. Such collaborations will involve negotiations for a mutually agreed-upon co-authorship arrangement.

- Our overarching goal is to strike a balance that provides opportunities for all teams to lead collaborative papers, particularly facilitating leadership roles for students and postdocs in fulfilling publication credit requirements for their theses or early career projects. Collaborative papers may take the form of co-first and/or co-last authored by groups bringing different key datasets or through an agreed-upon sequence of collaborative papers led by various generators in the collaboration.

6.3.2 Transparent Planning for Collaborative Papers:

- To enhance the production of collaborative papers under conditions of free data availability, we commit to regular and transparent discussions to revisit plans for project publications.

- The aim is to clarify the "right-of-way" for authors or groups to proceed with specific analyses for publication, particularly relevant for students and postdocs designing papers for dissertations and postdoc projects. Recognizing the dynamic nature of data analysis and scientific ideas, we anticipate that publication plans may evolve over time. Regular group discussions and a curated list of proposed papers are tools employed to achieve transparent evolution with the active participation of all affected parties.

6.3.3 Inclusion of Data Generators for Already-Published Data:

- Generators of already-published data are integral to discussions about developing papers using such data.

- Once data is published, the data generator is not necessarily an author on subsequent manuscripts. Further authorship requires additional contributions.

- Data generators should be given the opportunity to participate as co-authors on papers if they express a desire to contribute, such as further interpretation of data or text editing, qualifying them for co-authorship. This policy fosters community building within our

collaboration and captures the unique insights that data generators often possess about their datasets.

-However, data generated within the EMBRIO collaboration is considered subject to the guidelines outlined in the White House Memorandum to Federal Agencies Ensuring Free, Immediate, and Equitable Access to Federally Funded Research dated 08/25/2022. According to this directive, scientific data underlying peer-reviewed scholarly publications resulting from federally funded research should be freely available and publicly accessible at the time of publication, unless subject to limitations as described in Section 3(c)(i) of the memo. The National Science Foundation Grant General Conditions include the following from Article 54. Sharing of Findings, Data, and Other Research Products

a. NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages the grantee to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.

b. Adjustments and, where essential, exceptions may be allowed to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate legitimate interests of investigators.

6.4. Enforcement of the Policy

If Institute Leadership determines that a clear violation of this policy has occurred, it may result in loss of EMBRIO funds held by the research

References

(1) Linkert, M.; Rueden, C. T.; Allan, C.; Burel, J.-M.; Moore, W.; Patterson, A.; Loranger, B.; Moore, J.; Neves, C.; MacDonald, D.; Tarkowska, A.; Sticco, C.; Hill, E.; Rossner, M.; Eliceiri, K. W.; Swedlow, J. R. Metadata Matters: Access to Image Data in the Real World. *J. Cell Biol.* **2010**, *189* (5), 777–782. https://doi.org/10.1083/jcb.201004104.

(2) Rigano, A.; Ehmsen, S.; Öztürk, S. U.; Ryan, J.; Balashov, A.; Hammer, M.; Kirli, K.; Boehm, U.; Brown, C. M.; Bellve, K.; Chambers, J. J.; Cosolo, A.; Coleman, R. A.; Faklaris, O.; Fogarty, K. E.; Guilbert, T.; Hamacher, A. B.; Itano, M. S.; Keeley, D. P.; Kunis, S.; Lacoste, J.; Laude, A.; Ma, W. Y.; Marcello, M.; Montero-Llopis, P.; Nelson, G.; Nitschke, R.; Pimentel, J. A.; Weidtkamp-Peters, S.; Park, P. J.; Alver, B. H.; Grunwald, D.; Strambio-De-Castillia, C. Micro-Meta App: An Interactive Tool for Collecting Microscopy Metadata Based on Community Specifications. *Nat. Methods* **2021**, *18* (12), 1489–1495. https://doi.org/10.1038/s41592-021-01315-z.

(3) Kunis, S.; Hänsch, S.; Schmidt, C.; Wong, F.; Strambio-De-Castillia, C.; Weidtkamp-Peters, S. MDEmic: A Metadata Annotation Tool to Facilitate Management of FAIR Image Data in the Bioimaging Community. *Nat. Methods* **2021**, *18* (12), 1416–1417. https://doi.org/10.1038/s41592-021-01288-z.

(4) Miller, C.; Bialy, N.; Strambio De Castillia, C. [ASCB CellBio 2024] What Is Image Metadata Anyway? Hands On Workshop for Cell Biology Image Data Life Cycle, 2024. https://doi.org/10.5281/zenodo.14518379.

(5) Sarkans, U.; Chiu, W.; Collinson, L.; Darrow, M. C.; Ellenberg, J.; Grunwald, D.; Hériché, J.-K.; Iudin, A.; Martins, G. G.; Meehan, T.; Narayan, K.; Patwardhan, A.; Russell, M. R. G.; Saibil, H. R.; Strambio-De-Castillia, C.; Swedlow, J. R.; Tischer, C.; Uhlmann, V.; Verkade, P.; Barlow, M.; Bayraktar, O.; Birney, E.; Catavitello, C.; Cawthorne, C.; Wagner-Conrad, S.; Duke, E.; Paul-Gilloteaux, P.; Gustin, E.; Harkiolaki, M.; Kankaanpää, P.; Lemberger, T.; McEntyre, J.; Moore, J.; Nicholls, A. W.; Onami, S.; Parkinson, H.; Parsons, M.; Romanchikova, M.; Sofroniew, N.; Swoger, J.; Utz, N.; Voortman, L. M.; Wong, F.; Zhang, P.; Kleywegt, G. J.; Brazma, A. REMBI: Recommended Metadata for Biological Images—Enabling Reuse of Microscopy Data in Biology. *Nat. Methods* **2021**, *18* (12), 1418–1422. https://doi.org/10.1038/s41592-021-01166-8.

(6) Sansone, S.-A.; Rocca-Serra, P.; Field, D.; Maguire, E.; Taylor, C.; Hofmann, O.; Fang, H.; Neumann, S.; Tong, W.; Amaral-Zettler, L.; Begley, K.; Booth, T.; Bougueleret, L.; Burns, G.; Chapman, B.; Clark, T.; Coleman, L.-A.; Copeland, J.; Das, S.; de Daruvar, A.; de Matos, P.; Dix, I.; Edmunds, S.; Evelo, C. T.; Forster, M. J.; Gaudet, P.; Gilbert, J.; Goble, C.; Griffin, J. L.; Jacob, D.; Kleinjans, J.; Harland, L.; Haug, K.; Hermjakob, H.; Sui, S. J. H.; Laederach, A.; Liang, S.; Marshall, S.; McGrath, A.; Merrill, E.; Reilly, D.; Roux, M.; Shamu, C. E.; Shang, C. A.; Steinbeck, C.; Trefethen, A.; Williams-Jones, B.; Wolstencroft, K.; Xenarios, I.; Hide, W. Toward Interoperable Bioscience Data. *Nat. Genet.* **2012**, *44* (2), 121–126. https://doi.org/10.1038/ng.1054.

(7) Kunis, S.; Dohle, J. Structuring of Data and Metadata in Bioimaging: Concepts and Technical Solutions in the Context of Linked Data, 2022. https://doi.org/10.5281/zenodo.7018750.

(8) Allan, C.; Burel, J.-M.; Moore, J.; Blackburn, C.; Linkert, M.; Loynton, S.; MacDonald, D.; Moore, W. J.; Neves, C.; Patterson, A.; Porter, M.; Tarkowska, A.; Loranger, B.; Avondo, J.; Lagerstedt, I.; Lianas, L.; Leo, S.; Hands, K.; Hay, R. T.; Patwardhan, A.; Best, C.; Kleywegt, G. J.; Zanetti, G.; Swedlow, J. R. OMERO: Flexible, Model-Driven Data Management for Experimental Biology. *Nat. Methods* **2012**, *9* (3), 245–253. https://doi.org/10.1038/nmeth.1896.

(9) Purdue University Office of Research; Boehm, R. Purdue University Research Repository (PURR). *Univ. Res. Core Facil. Boil. Descr.* **2023**.