Imaging and curation of human TB lung tissue with linked data

Gordon A. Wells[1,*], Kapongo Lumamba[1], Kievershen Nargan[1], Delon Naicker[1], Ashendree Govender[1], Rajhmun Madansein[2], Kameel Maharaj[2], Mpumelelo Msimang[3], Paul V Benson[4], Threnesan Naidoo[1,5], Zoi Katsirea[6], Tannia Gracia[6], Fani Memi[6], Josh Moore[7], Andra Waagmeester[8] and Adrie J. Steyn[1,9,10]

[1]*Steyn Lab, Africa Health Research Institute, Durban, South Africa*

[2]*Inkosi Albert Luthuli Central Hospital and University of KwaZulu-Natal, Durban, South Africa*

[3]*Department of Anatomical Pathology, National Health Laboratory Service, Inkosi Albert Luthuli Central Hospital, Durban, South Africa, South Africa*

[4]*Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA*

[5]*Department of Lab Medicine and Pathology, Walter Sisulu University, Eastern Cape, South Africa*

[6]*Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK*

[7]*German BioImaging – Gesellschaft für Mikroskopie und Bildanalyse e.V. (GerBI-GMB), Constance, Germany*

[8]*Micelio, 2180 Ekeren, Belgium*

[9]*Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL, USA*

[10]*Centers for AIDS Research and Free Radical Biology, University of Alabama at Birmingham, Birmingham, AL, USA*

Abstract

We describe a linked data model for a tissue repository comprising samples from more than 900 patients that is being used to characterise the complex manifestations of tuberculosis. This repository comprises tissue samples, formalin-fixed paraffin-embedded blocks, and microscope slides, each with associated metadata. These samples, blocks, and slides are subjected to various imaging modalities to characterise TB at the cellular and molecular level. These include immunohistochemistry, traditional histology, RNAScope, spatial transcriptomics, metallomics, spatial proteomics and micro-computed tomography. All of these samples and imaging outputs have associated data and metadata that need to be linked for maximum scientific value. We are developing a linked data model that combines metadata capture in REDCap with imaging data stored in OMERO and related platforms.

Keywords

Tuberculosis, Imaging, Linked Data, OMERO, REDCap, Histopathology,

1. Introduction

Tuberculosis (TB) is caused by infection with *Mycobacterium tuberculosis* and remains the deadliest infectious disease in the world, resulting in approximately 1.3 million deaths annually[1]. In the early 20th century, the disease was studied by examination of autopsy tissue. This abated as successful antibiotics were discovered and animal models replaced autopsy examination. Nonetheless, with the

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025

*Corresponding author.

gordon.wells@ahri.org (G. A. Wells); kapongo.lumamba@ahri.org (K. Lumamba); kievershen.Nargan@ahri.org (K. Nargan); delon.Naicker@ahri.org (D. Naicker); ashendree.govender@ahri.org (A. Govender); rajhmadansein@gmail.com (R. Madansein); drkameelmaharaj@gmail.com (K. Maharaj); Mpumelelo.Msimang@nhls.ac.za (M. Msimang); pvbenson@uabmc.edu (P. V. Benson); threnesan.naidoo@ahri.org (T. Naidoo); zk3@sanger.ac.uk (Z. Katsirea); tannia.gracia@sanger.ac.uk (T. Gracia); fm13@sanger.ac.uk (F. Memi); josh@openmicroscopy.org (J. Moore); andra@micelio.be (A. Waagmeester); adrie.Steyn@ahri.org (A. J. Steyn)

https://anatpath.ukzn.ac.za/staff/academic-leader/ (M. Msimang); https://scholars.uab.edu/6975-paul-benson (P. V. Benson); https://www.sanger.ac.uk/person/katsirea-zoi/ (Z. Katsirea); https://uk.linkedin.com/in/tannia-gracia-92904368 (T. Gracia); https://www.sanger.ac.uk/person/memi-fani/ (F. Memi); https://joshmoore.github.io/ (J. Moore); https://www.ahri.org/scientist/adrie-steyn/ (A. J. Steyn)

0000-0003-2328-5208 (G. A. Wells); 0000-0002-2649-0795 (K. Lumamba); 0000-0002-4247-981X (K. Nargan); 0000-0001-8778-5725 (D. Naicker); 0000-0002-3819-6884 (P. V. Benson); 0000-0002-1864-4301 (T. Naidoo); 0009-0009-3815-4922 (Z. Katsirea); 0000-0003-4028-811X (J. Moore); 0000-0001-9773-4008 (A. Waagmeester); 0000-0001-9177-8827 (A. J. Steyn)

advent of drug resistance, TB remains difficult to treat and diagnose. TB pathology is complex and heterogenous, and carriers can remain infectious without showing symptoms (latent and sub-clinical TB). This requires a better understanding of the tissue states at a cellular and molecular level.

In our lab, we have accumulated a repository of human tissue infected with TB in collaboration with hospitals and pathologists in South Africa and the USA. The TB burden in South Africa is particularly high and exacerbated by co-infection with HIV. This has allowed us to study the disease, as was done in the early 20th century. But in traditional histology, we and our collaborators also employ modern imaging and spatial omics modalities. These include RNAscope, spatial transcriptomics, spatial proteomics, immuno-histochemistry, metallomics and micro-computed tomography.

This results in the generation of vast amounts of heterogeneous data that need to be linked to be optimally exploited to better understand TB. This includes patient meta-data, tissue sampling, formalin fixed paraffin embedded (FFPE) blocks and glass slides, imaging outputs, and regions of interest (ROIs). Specifically, ROIs expertly curated by experienced pathologists can be used to filter physiologically relevant regions for further analysis. These can also be used to train machine learning algorithms for large-scale analysis.

Here we describe our ongoing project to link patient and sample meta-data from our tissue repository captured in REDCap (Research Electronic Data Capture)[2] with derived imaging outputs from various modalities served in platforms such as OMERO[3].

2. Rationale

In the Steyn lab we have accumulated a substantial repository of human tissue from lung resections and autopsies in order to better understand the pathology of TB in the human host using modern experiments. Processing these samples results in the generation of thousands of blocks and glass slides. Most of these slides are histochemically stained, and a subset are used for other modalities of -omics imaging.

The metadata for samples, blocks and slides is being stored in REDCAP, and the imaging outputs are being indexed and presented in OMERO, with similar platforms being considered for future implementation.

To fully unlock the scientific value of these repositories, it is essential to enable seamless cross-querying between REDCap, OMERO, and potentially other databases. Historically, these types of analyses were either impossible or impractical for research teams due to the significant time and complexity involved. By addressing these limitations, we aim to make previously unattainable insights accessible and actionable. Additionally, querying capabilities could be extended to include resources in the linked-data cloud or other data repositories that follow the FAIR data principles, enabling a broader and more integrative approach to TB research.

Some illustrative examples of queries that could be performed include:

- What is the size distribution of necrotic TB lesions in HIV positive male patients aged between 20-45?
- What is the correlation between necrotic and adjacent fibrotic surface area by size distribution in necrotic lesions from HIV negative male patients aged between 45-60?
- Which transporter proteins are up or downregulated by a certain factor in T-cells comparing between necrotic and non-necrotic granulomas?
- Which metals and metal-binding proteins correlate above a certain threshold on the surface of non-necrotic granulomas in patients not symptomatic for TB?
- Select matched/nearly adjacent slides from symptomatic and asymptomatic TB patients with necrotic granulomas of a size range x-y for ROI paired spatial transcriptomics, proteomics and fluorescent IHC in order to train machine learning models to distinguish between symptomatic and asymptomatic TB from untrained samples.

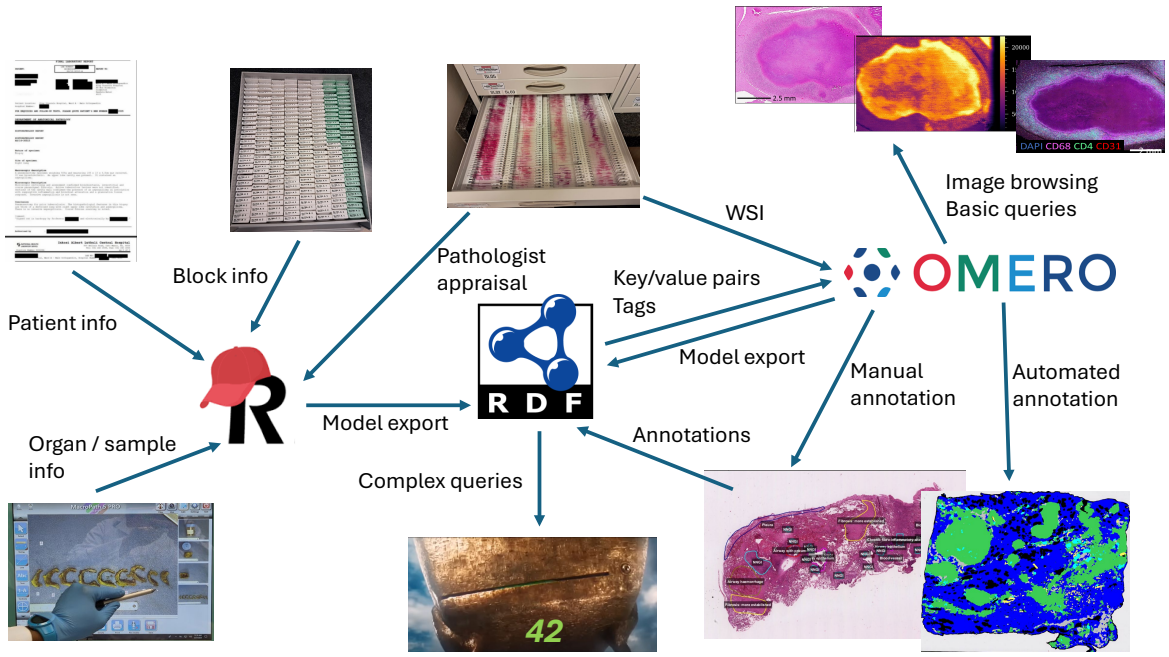

Figure 1: TB tissue data curation and integration at the AHRI Steyn Lab

Extending this querying capability to FAIR-aligned and linked-data resources could provide even greater opportunities for integrating knowledge. For example, by connecting to databases in the linked-data cloud, researchers could investigate correlations with environmental, socio-demographic, or genetic datasets, enriching TB research with new dimensions of contextual information. Similarly, FAIR-compliant resources in genomics, transcriptomics, or epidemiological studies could offer critical insights, such as linking observed phenotypes in TB lesions to global molecular pathways or epidemiological trends.

3. Methodology

3.1. General overview

Formalin fixed tissue samples from lung resection are first photographed and processed for blocking. From the paraffin embedded blocks various microscope slides are generated for histology, immuno-histochemistry and various -omics experiments. This generates a large amount of data that needs to be integrated, annotated and curated. This involves close collaboration between microscopists, histologists, semantic experts, image analysts etc. Ultimately the aim is to link and integrate this data across disparate data capture platforms (RedCap, Omero, etc) to facilitate complex queries and enable sophisticated machine learning of TB tissue states (Fig. 1).

3.2. Data Sharing using Marfa

All data sharing in this study was made possible through Wellcome Leap Delta Tissues' Master Research Funding Agreement (MARFA)[1], which creates a framework covering intellectual property (IP) rights, confidentiality, and publication processes for all consortium partners. Projects can then be initiated with minimal administrative overhead. With MARFA in place, data as well as tissue samples could quickly and efficiently be shared internationally while protecting IP. Samples from AHRI were sent to

[1]https://wellcomeleap.org/wellcome-leap-hbnet/

Sanger for the analysis with additional modalities. These data products were then similarly linked to the original sources via the mechanisms described above.

3.3. Sharing of tissue samples

Tissue samples in the form of blocks and slides were shipped from AHRI in Durban (SA) to Sanger in Cambridge (UK). The blocks and slides were prepared for shipment after initial analysis and bioimaging with hematoxylin and eosin.

Blocks are shipped with the expectation that after the analysis and enough tissue sampling was done, the remaining blocks are returned to sender. Slides are intented to remain at the receiver. The data was uploaded to Globus in accordance with the Data Sharing agreement alligned with Marfa.

3.4. Semantic bootstrapping

Semantic bootstrapping in our project begins with a careful selection of ontologies, ensuring alignment with existing frameworks and community standards such as NCIT and other OBO Foundry ontologies. A comprehensive review of the available ontological landscape guides this selection, enabling the integration of compatible models while identifying gaps that may require custom solutions. Initially, the Delta Tissue repository operates as a closed ecosystem, allowing us to establish a controlled environment for testing and refining our semantic models. This approach ensures that foundational elements are robust before expanding interoperability with external systems and broader data-sharing networks.

To achieve this, we leverage the principles of the Semantic Web, which provides a robust framework for defining entities and their relationships, enabling a consistent and machine-readable network of linked data. Central to this approach is the Resource Description Framework (RDF), a standardized method for representing knowledge as triples: each triple consists of a subject, a predicate, and an object. For example, in the context of tuberculosis (TB) research, a triple might describe how a necrotic TB lesion contains a specific immune cell type, such as macrophages:

$$\underbrace{\text{Necrotic TB Lesion}}_{\text{subject}} \quad \underbrace{\text{Has Part}}_{\text{predicate}} \quad \underbrace{\text{Macrophage}}_{\text{object}}$$

Each component of the triple is assigned a Uniform Resource Identifier (URI), a globally unique and resolvable identifier that ensures consistency and interoperability. For instance, the triple above can be represented in RDF as:

$$\underbrace{\texttt{<http://example.org/tb/lesion/12345>}}_{\text{subject}} \quad \underbrace{\texttt{<http://purl.obolibrary.org/obo/BFO\_0000051>}}_{\text{predicate}} \quad \underbrace{\texttt{<http://purl.obolibrary.org/obo/CL\_0000235>}}_{\text{object}}$$

Here, the URI for the “necrotic TB lesion” is a local identifier specific to the repository, the predicate “has part” comes from the OBO Relations Ontology (RO), and the object “macrophage” is identified through the Cell Ontology (CL). This structure not only ensures that data can be linked across datasets but also makes each component uniquely and universally resolvable, enhancing interoperability and facilitating automated reasoning.

The use of RDF allows us to represent complex relationships between biological entities, such as the connection between clinical metadata and molecular profiles. In the Delta Tissue repository, RDF triples capture information about tissue samples, associated clinical data, and molecular annotations, creating a network of semantically linked data. For example, triples can describe how a specific tissue sample from a TB patient is linked to metadata about patient demographics and clinical symptoms, associated with imaging data showing lesion structure, and connected to molecular measurements such as cytokine levels or gene expression profiles.

By leveraging the Semantic Web, we can also extend the utility of our repository to integrate external FAIR-aligned resources, such as those in the linked-data cloud. For instance, ontologies and identifiers from resources like UniProt, ChEBI, or the Gene Ontology (GO) can be seamlessly integrated into the

RDF graph, enabling queries that connect tissue-level observations to broader biological knowledge. For example, a query might retrieve all necrotic TB lesions linked to macrophage activity and correlate those findings with cytokine pathways from external ontologies.

Additionally, RDF provides flexibility for human-readable descriptions through annotations, such as labels, definitions, and metadata, which can be encoded as RDF triples. These annotations enhance the usability of the data for both humans and machines, ensuring that the information is both accessible and interpretable.

By starting within a controlled ecosystem and gradually extending to external systems that follow Semantic Web standards, we can create a scalable and interoperable framework for data sharing and integration. This allows researchers to address complex questions that span diverse datasets and modalities, paving the way for deeper insights and more impactful discoveries in tuberculosis research and beyond.

3.4.1. REDCap

REDCap is a secure, web-based application designed for the collection, management, and analysis of research data. Originally developed at Vanderbilt University, REDCap has become a widely adopted platform in clinical and biomedical research environments. It supports diverse use cases, from small single-site projects to large, multicenter clinical studies, offering customizable data forms, automated workflows, and integration options with other research tools.

REDCap supports a wide variety of export formats, catering to structured data formats commonly used in research. These include CSV, Excel, and specialized formats for statistical analysis software such as R, SPSS, and SAS. However, its export capabilities are limited when it comes to semantic layers of data, as the current focus is primarily on structured formats rather than ontological or semantic representations. While REDCap provides some level of ontological support during data entry, such as using controlled vocabularies and terminologies, the export of these semantic enrichments remains limited in scope, restricting the integration of REDCap data with broader semantic frameworks.

To address this gap, we developed a Python script leveraging the REDCap API to extract structured data and store the results in Resource Description Framework (RDF) format. RDF is a flexible, graph-based data model designed for representing and linking data in a machine-readable way. It encodes data as triples in the form of subject-predicate-object relationships, enabling semantic interoperability across datasets. This RDF export currently mirrors the structure of the data forms in REDCap, capturing the design and content of the forms but without mapping the data to external ontologies or adding higher-order semantic context.

To enhance the semantic utility of the RDF export, we rely on SPARQL CONSTRUCT queries for data transformation. SPARQL, the query language for RDF, supports flexible and expressive queries to retrieve and manipulate graph data. The CONSTRUCT clause in SPARQL enables the creation of new RDF graphs by transforming existing RDF triples into a desired structure. This makes SPARQL CONSTRUCT a powerful tool for aligning data with ontologies, adding mappings, and creating interoperable datasets without altering the original data source.

Currently, our efforts focus on exploring other RDF transformation methods to further streamline and automate the process of enriching REDCap data with semantic annotations. These transformations aim to make the data more interoperable and reusable in broader research contexts, bridging the gap between structured data exports and fully semantic representations.

3.4.2. OMERO

OMERO is an open-source software platform designed to manage and analyse complex biological data, including images, matrices, and tables. By providing a unified interface through a server-based application, it supports diverse research areas such as light microscopy, high-content screening, and electron microscopy. OMERO's flexible, model-driven architecture facilitates efficient data handling

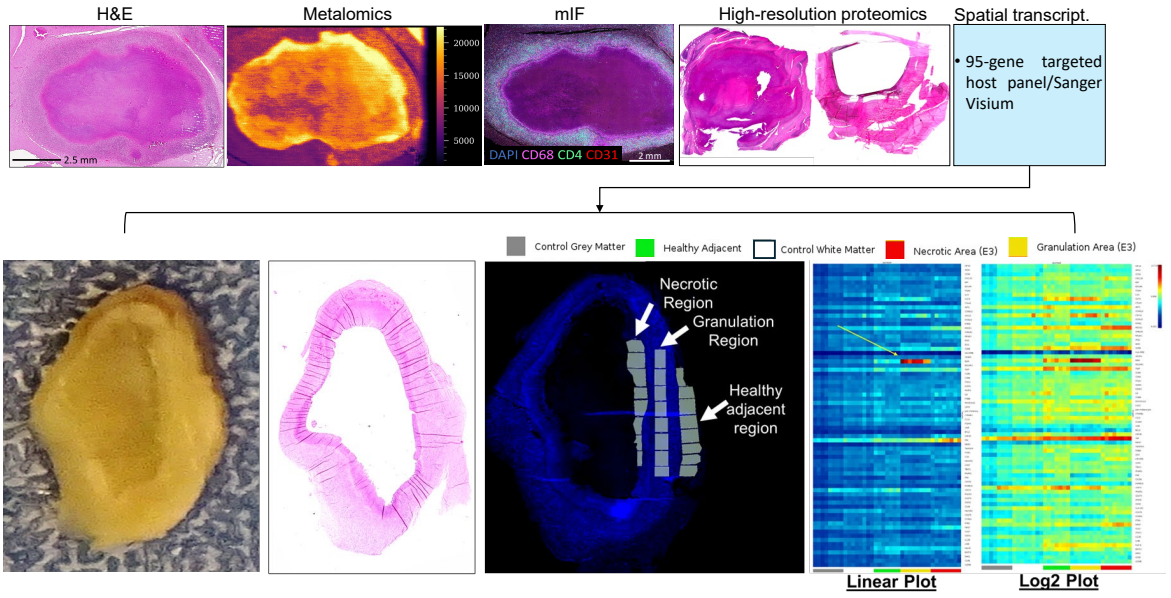

Figure 2: Multi-modality imaging of TB tissue states

and integration across various experimental contexts, enhancing the accessibility and reproducibility of biological research.

The OME Model, defined using XML Schema, encompasses over 160 classes that constitute a standardized vocabulary for describing multi-dimensional imaging experiments. In the OMERO platform, these data are stored within PostgreSQL tables structured with extensive foreign key constraints and are accessible through a custom-built API. However, extracting comprehensive information about an imaging experiment can be both labor-intensive and error-prone. The integration of the omero-rdf library[4] addresses this challenge by providing a semantic mapping, facilitating interoperability with various external platforms.

Specifically for this study, OMERO provides a platform to store and annotate whole slide images if TB infected tissue. Annotations and regions of interest (ROIs) created by pathologists can then be shared and queried.

By providing links in OMERO to REDCap information and links from REDCap to OMERO, the basis for a linked data infrastructure loosely coupled via patient identifiers was made possible.

4. Results

The Steyn lab at AHRI has accumulated a large repository of tissue samples from about 900 patients to study active TB and TB in patients that died from other causes. From this tissue thousands of FFPE blocks and slides and their accompanying whole slide images (11 000) have been generated. This tissue is being used to study TB tissue states using multiple modalities (Fig. 2). These imaging outputs are catalogued in OMERO and in turn linked with patient metadata stored in RedCap (Fig. 3).

A challenge that remains is overcoming the ”inertia of familiarity”, that is convincing specialists in other domains (pathologists, prosectors, PIs, histologists etc) to contribute to platforms they find unfamiliar. This includes using systems such as RedCap and OMERO.

5. Outlook

Among the long-term goals are:

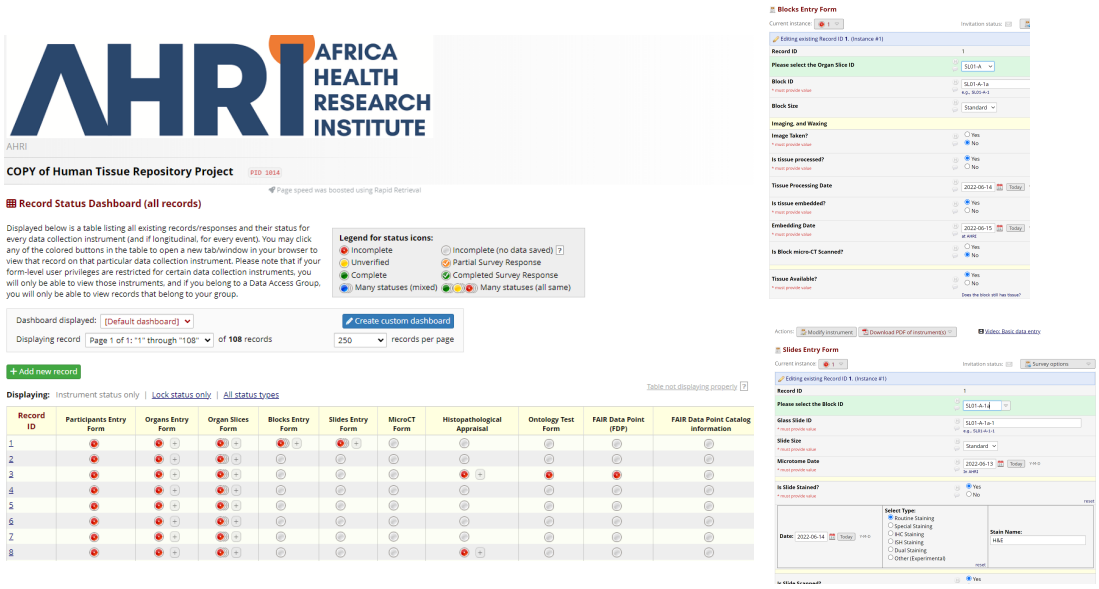

Figure 3: Metadata cature in RedCap

- Enable complex queries across multiple imaging modalities that can also serve as an input to machine learning methods.
- Create a tissue data repository that is internationally available to TB researchers.
- Contribute to a standardised vocabulary for TB researchers.
- Create a semantically capable and reusable TM image RDMS.

Acknowledgments

This work was supported by the Delta Tissue program of Wellcome Leap[2], a global ARPA for health. To all tissue donors and their family members for their humble gesture so that this research could be performed.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Global tuberculosis report 2024, World Health Organization, Geneva, 2024. URL: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024.

[2] P. A. Harris, R. Taylor, B. L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby, S. N. Duda, REDCap Consortium, The REDCap consortium: Building an international community of software platform partners, J. Biomed. Inform. 95 (2019) 103208.

[3] C. Allan, J.-M. Burel, J. Moore, C. Blackburn, M. Linkert, S. Loynton, D. MacDonald, W. J. Moore, C. Neves, A. Patterson, M. Porter, A. Tarkowska, B. Loranger, J. Avondo, I. Lagerstedt, L. Lianas, S. Leo, K. Hands, R. T. Hay, A. Patwardhan, C. Best, G. J. Kleywegt, G. Zanetti, J. R. Swedlow,

[2] https://wellcomeleap.org/

OMERO: flexible, model-driven data management for experimental biology, Nature Methods 9 (2012) 245–253. URL: https://doi.org/10.1038/nmeth.1896. doi:10.1038/nmeth.1896.

[4] J. Moore, et al., German-bioimaging/omero-rdf: v0.4.1, 2024. URL: https://doi.org/10.5281/zenodo.13380095. doi:10.5281/zenodo.13380095.