Deliverable 4.1

Requirements analysis for the cell migration repository

Grant Agreement number:	634107
Project acronym:	MULTIMOT
Project title:	Capture, dissemination and analysis of multiscale cell migration data for biological and clinical applications
Deliverable type and status:	Report - Public
Workpackage:	4
Deliverable responsible:	Benny Geiger
Organisation:	Weizmann
Email:	Benny.geiger@weizmann.ac.il
Due Date:	31st of July 2016
Delivery Date:	
Project coordinator name, title:	Prof. Dr. Lennart Martens
Organization:	VIB, Computational Omics and Systems Biology Group - Department of Biochemistry, Ghent University (UGent) Department of Medical Protein Research, VIB Albert Baertsoenkaai 3 B-9000 Gent Belgium
Tel:	+32 9 264 93 58
E-mail:	Lennart.martens@vib-ugent.be
Project website:	www.multimot.org

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 634107

1. Table of Contents

2. Summary of Deliverable

2.1 Background

The creation of the Cell Migration Repository is part of MULTIMOT WP4 (Development of a community accessible cell migration data repository and knowledge base).

2.2 Goal

The goal is to create a modern, standards-compatible and community accessible repository for cell migration research.

The repository should fully implement the standards developed in WP2, provide support to the experimental data generated in WP6, and the corresponding data analysis tasks and approaches developed in WP5.

Furthermore, the repository should implement high-level federation with the Image Data Repository of the Open Microscopy Environment (IDR: https://idr-demo.openmicroscopy.org/about/), it should as well implement a bidirectional metadata query application programming interface (API), and moreover provide extensive links to existing, relevant external (bioinformatics) databases.

On top of this Data Repository, there should be a Knowledge Base layer containing metadata. These metadata will allow for searching and linking to data located either in the local repository or on remote servers like the IDR. This Knowledge Base layer will inherit from the existing Cell Migration Gateway (https://www.cellmigration.org).

3. Description of work

The Cell Migration Repository is being designed as a central space to support and encourage cell migration data storage, analysis and sharing. Constructed on top of a distributed multi-layer architecture, the repository will manage access to data under a role-based model, complying with requisites of patient privacy for sensitive data (*e.g.* subset of data generated in WP5).

During the design phase of the Cell Migration Repository, WIS contacted experienced developers and analyzed well-established related projects, including OMERO (https://www.openmicroscopy.org/site/products/omero), Zegami (http://zegami.com) and XNAT (https://www.xnat.org).

Moreover, WIS can also build on extensive in-house experience with the development and maintenance of Bioimaging (http://miw.weizmann.ac.il/srv/bioimg), a campus-wide repository running already for five years, developed at the Bioinformatics Unit of the Weizmann Institute of Science.

The following sections of this document report on the requirements for the repository, and furthermore introduces some technical choices that were made for their implementation.

3.1 Repository capacity

The Data Repository should be able to store and manage cell migration data produced in the MULTIMOT consortium, and in the scientific community at large. For the first stage of development, during the pilot phase of the Data Repository, we are targeting to serve the equivalent output of ten research groups, with an estimated overall data size of about 20 TB. Following the experience gained with the modular design of Bioimaging (http://miw.weizmann.ac.il/srv/bioimg/), we plan to gradually increase hardware processing capacity and data storage to support additional research groups and larger data volume.

3.2 Data formats

The repository should accept, store and manage raw image data, processed image data and analysis results, allowing downstream data analysis to take place at multiple levels. Data should be stored as is (*i.e.* in their native formats), because these files/formats, as well as the file systems in which these are structured, are required for some of the existing software tools. The repository metadata should represent the information contained in the minimum reporting requirements (MIACME), and this information should be compliant with the controlled vocabularies (CVs).

Both the MIACME and the CVs are currently under development by the Cell Migration Standardization Organization (MULTIMOT WP2, Deliverables D2.2 and D2.3; and CMSO, https://cmso.science). Moreover, the goal is to automate MIACME compliance validation for data sets stored in the repository (see also Deliverable D4.4).

It should be noted here that it is expected that the standards will continue to evolve during and after the project running time, and we therefore expect to keep the repository up-to-date with these changes through a close connection to the CMSO and its output.

3.3 Data loader

The varying data sizes of cell migration experiments demand special attention to the mechanism by which the data will be uploaded to the Data Repository. We would like to implement a resumable data loader with a friendly graphical user interface (GUI) and programmable API. If possible, we would like for the data loader to be a server side mechanism, avoiding the need to install dedicated clients on the scientist’s computer.

It should also be noted that CellMissy will be adapted in such a way that it can provide its data in standardized formats (see Deliverable D3.2), which will then be ready for submission to the repository through the data loader we will develop for the repository, and/or through the data loader API. In the latter case, data submission could ideally even be integrated directly in the CellMissy software.

3.4 Data access

The Data Repository should have the option to define both public and private level for data access. The uploaded data belong to the depositor, who has the right to declare all or part of the data as being of public access, and to grant read access to users that request so. We need to implement a solid mechanism to clearly identify users at the time of uploading and accessing data. To ensure that the stored data can be used as Data Reference, the Data Repository should be a WORM (write once read many) storage.

Because CellMissy will ultimately be able to import data from the repository (see Deliverable D3.5), we will ideally also be able to provide an option in CellMissy to access private data sets if the correct user credentials are provided. In this way, apart from the implicit access to public data that will be provided by CellMissy, we will also be able to provide explicit access to private data. Such a feature will be primarily useful for journal editors and peer reviewers, who will then be able to inspect and interact with the data related to a submitted manuscript. For this to function, we will however, need to be able to provide automatically generated 'reviewer' accounts that provide access to individual data sets, and that can be safely shared with journal editors, and peer reviewers.

3.5 Data finding

Data and metadata in the Data Repository should be findable through both GUI and API (notably a REST service API, see also Deliverable D4.3). Both interfaces should allow users to select the information of interest by terms from the minimum reporting requirements and controlled vocabularies, experimental conditions, data analysis techniques and algorithms, related publications, data depositor, gene and protein names, amongst others.

It should be noted that our initial plans to also implement a BioMART (see also Deliverable D4.3) are currently being reconsidered, in light of the gradual decline in BioMART installations, which is especially noticeable at the European Bioinformatics Institute. It is clear that REST services have become much more popular and more easily implemented by third parties. Moreover, the main benefit from BioMART is its ability to interface with other BioMARTs, a feature that is rapidly losing its appeal due to the shuttering of many of the most relevant BioMARTs. We will therefore likely shift the development effort from BioMART to improved and more performant REST services in Deliverable D4.3.

3.6 Data retrieval

Unless usage experience demonstrates the need for additional methods, the initial data retrieval would be implemented over HTML5. If a need for large-scale data download is identified, alternative approaches such as FTP or Aspera will be looked into.

Note that data download will also be possible through CellMissy (see Deliverable D3.5 and section 3.9 'CellMissy integration' below).

3.7 Data identification

Each uploaded dataset should receive a unique and permanent identifier from the Data Repository. New versions of the same dataset, if any, will receive distinct unique permanent identifiers. These

Cell Migration Data Repository unique identifiers will allow for cross-linking from other databases and for referencing in scientific publications. We will also work towards integration of the repository's identifier namespace and identifiers in the MIRIAM registry (http://www.ebi.ac.uk/miriam/main) so that these can be resolved easily.

3.8 User Authentication and Authorization

As the Identity Provider for User Authentication and Authorization to the Data Repository, we are considering incorporating the existing eduGAIN system for authentication (http://services.geant.net/edugain/About_eduGAIN/Pages/Home.aspx), complemented with a local Identity Management System for users not supported by eduGAIN.

3.9 CellMissy integration

There should be a close integration between CellMissy and the Data Repository. For the researcher this means to manage and work with their data in the familiar environment of CellMissy, and to upload these data for external sharing and permanent storage to the Data Repository when desired (see Deliverable D3.2, and section 3.3 'Data loader' above). This communication should be bidirectional, allowing CellMissy to search for and retrieve data and metadata from the Data Repository (see Deliverable D3.5, and sections 3.5 'data finding' and 3.6 'data retrieval' above). This bidirectional communication will allow users to send their data to the repository easily, but also to obtain an overview of other projects in the repository that may be related to their work. If any relevant projects are indeed found in the repository, the user will then be able to download these data for detailed local processing and interpretation within CellMissy, and will even be able to integrate these analyses with those of the local data for comparison or added even added power.

4. Future plans

These requirements provide the basis for our development work and software architecture (including relevant third party services or tools). And over the coming months, we will implement a pilot Data Repository and Database Knowledge layer (to be reported in D4.2, at M24) based on this overall framework.

As mentioned above, we will first target storage facilities to the equivalent output of ten research groups, but we will increase processing capacity and data storage of the repository to scale with the needs of the MULTIMOT consortium, and the cell migration research community at large. Based on data growth in existing repositories and databases in other fields (most notably the omics fields) it is expected that these requirements will at first rise exponentially, but will then continue to grow in a more linear fashion over time.

We also anticipate a close connection with the work of the CMSO throughout the project running time and beyond, as the standards developed within the CMSO will likely continue to evolve over time. Especially the MICAME reporting requirements and the controlled vocabulary will likely see continued development over time. Moreover, we will participate actively in the CMSO to ensure that the experiences from running the repository are taken into account by the standards community. Relevant examples of the benefits of such feedback from the repository to the standards community include missing CV terms that will need to be curated and added to the CV at regular intervals (see also Deliverable D2.2), the ability of the community to adhere to MIACME requirements, and the possible expansion of MIACME requirements should a dearth of adequate metadata comprise efficient reuse of data sets held in the repository.

The integration of CellMissy with the repository will be important to provide a reference implementation of an end-user oriented software tool that is capable of bidirectional interaction with the repository. As such, it will also ensure that end users have at least one freely available, open source software tool available for these tasks. It is however, also the intention to stimulate as much as possible the uptake of similar types of repository integration in both commercial and freely available tools in the domain of cell migration research. This integration can be direct, or indirect. In the latter case, standards compliance of third-party tools (see also Deliverable D3.4 for the specific case of Idea Biomedical software) will allow data to de sent to the repository with relative ease, and data downloaded from the repository in standard formats to be loaded into the software.

Finally, interaction with the community (notably: researchers, journals, funders, and instrument vendors) will be a key outreach activity, and a relevant position paper will be written on this specific topic when the system is sufficiently mature (see Deliverable D7.4).

It should be noted that future work on the repository itself will be documented in stages, in Deliverables D4.2-D4.6.