jove

gP2S, an Information Management System for CryoEM Experiments

Dorota Wypych[1], **Daniel Kierecki**[1], **Filip M. Golebiowski**[1], **Alexis Rohou**[2]

[1] Roche Polska Sp. z o.o. [2] Department of Structural Biology, Genentech

Corresponding Author

Alexis Rohou

rohou.alexis@gene.com

Citation

Wypych, D., Kierecki, D., Golebiowski, F.M., Rohou, A. gP2S, an Information Management System for CryoEM Experiments. *J. Vis. Exp.* (172), e62377, doi:10.3791/62377 (2021).

Date Published

June 10, 2021

DOI

10.3791/62377

URL

jove.com/video/62377

Abstract

Cryogenic electron microscopy (cryoEM) has become an integral part of many drug-discovery projects because crystallography of the protein target is not always achievable and cryoEM provides an alternative means to support structure-based ligand design. When dealing with a large number of distinct projects, and within each project a potentially large number of ligand-protein co-structures, accurate record keeping rapidly becomes challenging. Many experimental parameters are tuned for each target, including at the sample preparation, grid preparation, and microscopy stages. Therefore, accurate record keeping can be crucially important to enable long-term reproducibility, and to facilitate efficient teamwork, especially when steps of the cryoEM workflow are performed by different operators. To help deal with this challenge, we developed a web-based information management system for cryoEM, called gP2S.

The application keeps track of each experiment, from sample to final atomic model, in the context of projects, a list of which is maintained in the application, or externally in a separate system. User-defined controlled vocabularies of consumables, equipment, protocols and software help describe each step of the cryoEM workflow in a structured manner. gP2S is widely configurable and, depending on the team's needs, may exist as a standalone product or be a part of a broader ecosystem of scientific applications, integrating via REST APIs with project management tools, applications tracking the production of proteins or of small molecules ligands, or applications automating data collection and storage. Users can register details of each grid and microscopy session including key experimental metadata and parameter values, and the lineage of each experimental artifact (sample, grid, microscopy session, map, etc.) is recorded. gP2S serves as a cryoEM experimental workflow organizer that enables accurate record keeping for teams, and is available under an open-source license.

Introduction

Information management at cryoEM facilities

Starting in 2014 approximately, the number of cryogenic electron microscopy (cryoEM)[1] facilities has grown explosively, with at least 300 high-end systems installed around the world[2], including at a number at pharmaceutical companies, reflecting a growing role for cryoEM in drug discovery[3]. The missions of these facilities, and their requirements for data tracking and management differ[4]. Some, for example national cryoEM centers, are charged with receiving EM grids, collecting datasets, and returning data to the users for structure determination, perhaps after some automated image processing. In such facilities, tracking the provenance of the grid, its association with a user proposal or grant, and the lineage from grid to dataset is crucial, but other factors, such as the method of purification of the protein sample or the eventual structure determination process, are less, or not at all, relevant. In other facilities, such as local academic facilities, each end user is responsible for preparing their own samples and grids, conducting the microscopy, managing the raw data and its processing and publishing the results. There is no stringent need for metadata tracking on the part of such a facility because this role is fulfilled by the end user or their Principal Investigator.

In our cryoEM facility, the handling and optimization of samples, grids, data collection and processing protocols, and results (maps, models) is centralized across many projects onto a small group of practitioners. This presents challenges in experimental (meta)data management. The experimental lineage of structures, from atomic model all the way back to the exact identity of proteins and ligands, via grid preparation parameters and data collection protocols, must be accurately captured and preserved. These metadata must be made available to a number of human operators. For example, a person doing image processing may need to know which construct of a protein was used and what the imaging parameters were, even though they neither purified the protein nor collected the cryoEM data themselves; informatics systems such as automated data management daemons need to identify the project for which a microscope is currently collecting data in order to correctly and systematically assign directory names.

Several information management systems are available to support cryoEM facilities. Perhaps most complete among them is EMEN2[5], which combines features of an electronic lab notebook, an information management system, and some elements of a business process management tool. Used at many synchrotrons, ISPyB[6], originally built to support the x-ray beamlines for crystallography, now also supports cryoEM data collection. Scipion[7] is a rich and powerful wrapper around image processing packages, which allows users to record image processing workflows and share them, for example via the public repository EMPIAR[8,9], and is also integrated with ISPyB to enable on-the-fly cryoEM data processing.

Here we describe gP2S (for Genentech Protein to Structure), a modern and lightweight cryoEM information management system built to support the workflow from purified protein and small molecule ligand through to the final atomic model.

Overview of gP2S

gP2S is a user-friendly web-based cryoEM information management system that facilitates accurate record-keeping for cryoEM labs and multi-user, multi-project facilities. The following entities, their relationships and associated metadata

are tracked: projects, equipment, consumables, protocols, samples, grids, microscopy sessions, image processing sessions, maps, and atomic models. Users can also add free-text comments, optionally including file attachments, allowing for rich annotation of any entity registered in gP2S. The front-end has been designed to facilitate use with touchscreen devices and tested extensively on 12.9" iPad Pros, making it possible to use gP2S at the lab bench while preparing samples and grids (**Figure 1**), as well as at the computer when operating the microscope, processing images or depositing models. Each page in the front end aims to reduce manual data entry by pre-setting parameters to sensible default values when possible.

The backend of gP2S features a number of REST API (REpresentational State Transfer Application Programming Interface) endpoints, making it possible to integrate gP2S into existing workflows and scripts. The data model was designed to allow the accurate capture of negative stain and cryoEM workflows, including branching, for example with one sample used on several grids, data from several microscopy sessions being merged into a single data processing session, or one data processing session yielding several maps.

System architecture

gP2S is a classic three-tier application (**Figure 2**). In this modular architecture, the system is broken into three separate layers, each responsible for performing distinct duties, and each replaceable or modifiable independently of the others. (1) The presentation layer (or frontend) provides user access via web browser (extensively tested with Chrome and Safari), allows creating and modifying workflow elements (including data validation), and displays experimental data as individual entities, project-based lists and full workflow reports. (2) The service layer (or backend) serves as an intermediary layer between the user interface and the storage system - it holds core business logic, exposes the service API used by the frontend, integrates with data storage and LDAP (Lightweight Directory Access Protocol) system for user authentication, and provides a basis for additional integration with external systems. (3) The persistence layer (data access) is responsible for storage of experimental data, user comments, and file attachments.

Key technologies and frameworks

In order to facilitate development, building and maintenance of the gP2S application, several technologies and frameworks were used in the project. The most important ones are: Vue.js 2.4.2[10] for the frontend and SpringBoot 1.3[11] with embedded Tomcat 8 server for the backend. The application uses MySQL 5.7 and MongoDB 4.0.6 databases for storage and LDAP[12] for authentication. By default, all of these component parts are shipped and deployed as one application.

In total the application uses hundreds of different libraries either directly or indirectly. The most prominent ones are listed in **Table 1**.

Data model

Three types of entities can be distinguished in the gP2S data model (**Figure 3**): workflow entities related to data gathered during experiments (e.g., samples or microscopy sessions); equipment and protocol entities that describe data that are common across all projects (e.g., microscopes or vitrification protocols); other entities that play supportive or technical roles in the system (e.g., comments or default values).

The root of the workflow data tree is the Project entity. Every project consists of a number of Proteins and/or Ligands that are building blocks for creating Sample entities. Each Sample can be used to create multiple Grids which in turn are used

in Microscopy Sessions (one Grid per Microscopy Session). The latter are assigned to Processing Sessions that can yield one or more maps. The last entity in the tree is the atomic Model, created using one or many Maps. In consequence every workflow-related entity, from Protein to Model, is always bound to a particular Project via its ancestors. Such design creates data aggregates that are easy to process either by the frontend module or by external systems using the API.

In addition to workflow data there are entities that describe equipment used in experiments or protocols that were followed while preparing grids. Defining these entities is a prerequisite for creating experimental workflow entities such as Grids, Microscopy and Processing Sessions.

The last type of data entity, collectively named as "Other", is used for technical purposes (e.g., file attachments or default values). This category includes comment entities that can be linked to any workflow or equipment/protocol entities.

Software availability

The open-source version of gP2S is available under an Apache License Version 2.0[26], from https://github.com/arohou/gP2S. A Docker image to run gP2S is available from https://hub.docker.com/r/arohou/gp2s. A closed-source branch of gP2S is under continued development at Roche & Genentech.

Running the gP2S application

There are two ways to run gP2S: as a docker container or as a standalone Java application. The optimal choice will depend on the target deployment environment. For example, if the ability to customize or enhance the code to suit specific needs of the users is desired, the whole application must be re-built first. In this case, running gP2S as a standalone application might be recommended.

Docker container

The easiest way to start working with the gP2S application is to run it as a Docker service. For that purpose, a dedicated Docker image has been prepared and published in Docker Hub repository ("https://hub.docker.com/r/arohou/gp2s"). Running the gP2S image depends on access to MySQL and MongoDB databases, and to a LDAP server. For non-production environment, it is recommended to run all these dependencies as multi-container Docker applications along with the gP2S application. To make this seamless, a docker-compose file (https://github.com/arohou/gP2S/blob/master/docker-compose.yml) that includes all needed configurations of the final environment has been prepared and provided in the gP2S GitHub repository (https://github.com/arohou/gP2S). The following docker images are dependencies: mysql[27], mongodb[28], apacheds[29].

In the default configuration, all stored data, both entities and file attachments will be deleted upon removal of the docker containers. In order to keep the data, either docker volumes should be used, or the gP2S application should be connected to dedicated database instances (MySQL and MongoDB). The ApacheDS LDAP server container comes with a preconfigured admin user (password: secret). These credentials should be used to log in to the gP2S application when it is run as a Docker service. For production environments the same docker-compose file can be utilized to deploy gP2S (and other containers if needed) as services to a Docker Swarm container orchestration platform.

The full process of running gP2S as a Docker container, including all details regarding proper configuration is described in the gP2S GitHub repository and covers the following topics:

- Running the dockerized gP2S application with all dependencies.
- Accessing the gP2S application, database and LDAP.
- Updating gP2S service with a new version.
- Removing gP2S application.
- Configuring data persistence.
- Connecting the dockerized gP2S application to dedicated databases or a LDAP server.
- Configuration details

Standalone Java application

Another option to run the gP2S application is to build a self-contained Java package. This approach should be taken if running Docker containers is not possible. Building the gP2S application requires installing a Java Development Kit version 8 or above. The whole build process is managed by the Maven tool, which is provided in the codebase in GitHub repository. Build configuration is prepared to build the frontend part first, then copy it to backend sources, and then build it as a final application. This way, there is no need to install any other tools or libraries in order to prepare a fully functioning gP2S package. By default, the result of the build is a JAR package (stored locally) and Docker image (pushed to the repository configured in the Maven pom.xml file). It is important to remember that information required to connect to external systems (databases and LDAP server) needs to be provided in a proper configuration file before the package is built.

Once the gP2S JAR package has been created, it contains all dependencies and configuration information needed to run the application, including the Tomcat application server which hosts the system. If the package was built with multiple configuration files it can be run in different modes without rebuilding.

The gP2S GitHub repository includes a complete description of the process of building and running gP2S as a standalone application and covers the following topics:

- Building gP2S using the Maven tool
- Building and running with embedded databases
- Building and running with dependencies deployed as docker containers
- Building and running with dedicated databases
- Configuring authentication

Protocol

1. Setting up gP2S for work

1. Log on to gP2S. Upon successful login, the main screen is shown.

   NOTE: In the top right corner, the user name is shown - click on this to log out. The left-hand-side navigation bar consists of a project selector (top), a set of navigation items listing the experimental entity types that define the cryoEM workflow (Samples, Grids, Microscopy Sessions, Processing Sessions, Maps, and Models), and a link to the Settings section of the application.

2. Before any experiments can be logged, populate the Settings section with information about the Projects, Equipment, Consumables, Software and Protocols that are in use at the cryoEM facility. Settings can be updated at any time by adding new tools and projects and by editing the existing entries; however, just like all entities in gP2S, Settings entities cannot be deleted once they are created.

2. Configure at least one project

1. Navigate to **Settings > Projects**.

2. Click on **Create New Project**.
3. Type in a Project label.
4. Click **Save**.

3. Configure at least one Surface Treatment Machine.

NOTE: Surface treatment machines are used to modify the surface properties of EM grids - most commonly they are glow dischargers or plasma cleaners.

1. From the **Equipment** section, choose **Surface Treatment Machine**.
2. Click **Create New Machine**.
3. Enter a label, which will serve to identify the machine later on.
4. Provide its Manufacturer, Model and Location.
5. Click **Save**.

4. Register at least one Grid Type.

NOTE: Grid Types are meant to identify models of grids (e.g., "2-µm holey carbon film on 300-mesh copper grids"), not specific batches or lots of grids

1. From the **Consumables** section select **Grid Type**.
2. Click on **Create New Grid Type**.
3. Enter a Grid Type label, Manufacturer and Description.
4. Click **Save**.

5. Register at least one Vitrification Machine

1. From the **Equipment** section, select **Vitrification Machine**.
2. Click on **Create New Machine**.
3. Provide its Manufacturer, Model and Location.
4. Click **Save**.

6. Register at least one Blotting Paper

1. From the **Consumables** section select **Blotting Paper**.
2. Click on **Create New Blotting Paper**.
3. Type in a Blotting Paper label, Manufacturer and Model.
4. Click **Save**.

7. Register at least one Cryo Storage Device

1. From the **Equipment** section, select **Cryo Storage Device**.
2. Click on **Create New Storage Device**.
3. Enter the device's Manufacturer, Model and Location.
4. Set the toggle switches to specify whether the added storage device features cylinders, tubes and/or boxes.
   NOTE: If it does, gP2S will let users specify relevant cylinder, tube and/or box identifiers later on when users log the storage locations for individual grids. With the above pieces of Equipment and Consumables set up, it is possible to create three types of Protocols - Surface Treatment, Negative Stain and Vitrification.

8. Register at least one Surface Treatment Protocol

1. From the **Protocols** section, select **Surface Treatment**.
2. Click on **Create New Protocol**.
3. Enter a label to identify the Protocol.
4. Select one of the Surface Treatment Machines.

5. Specify settings used during this protocol: duration, current and polarity of the discharge, and pressure as well as any additives in the atmosphere.
6. Click **Save**.

9. Create at least one negative stain protocol

1. From the **Protocols** section, select **Negative Stain**.
2. Click on **Create New Protocol**.
3. Enter a protocol label.
4. Describe the stain by giving values for its Name, the pH, and concentration of heavy metal salt.
5. Specify the incubation time of stain before blotting.
6. Enter free-text description of the protocol.
7. Click **Save**.

10. Register at least one grid-freezing protocol

1. From the **Protocols** section, select **Vitrification**.
2. Click on **Create New Protocol**.
3. Enter a protocol label.
4. Choose the relevant Vitrification Machine from the drop-down list.
5. Choose the Blotting Paper used in this protocol.
6. Then, provide the remaining experimental information: relative humidity, temperature, blot force, number of blots, blot time, wait time, drain time, number of sample applications.
7. Enter a free-text description.
8. Click **Save**.
   NOTE: After configuring the Protocols, it is possible to create both cryo and negative-stain grids. To use gP2S to record the next steps in the workflow, starting from Microscopy sessions, it is necessary to configure a Microscope, an Electron Detector and a Sample Holder.

11. Register at least one microscope

1. From the **Equipment** section, select **Microscope**.
2. Click **Create New Microscope.**
3. Type in a Microscope label.
4. Provide its Manufacturer, Model and Location.
5. Select which acceleration voltages are configured and usable on this microscope, out of the preset list of 80, 120, 200 and 300 kV.
6. Specify the list of condenser ("C2") and objective apertures installed.NOTE: for each type, up to 4 aperture slots can be configured, one of which is designated as the default aperture for this microscope. In the case of the objective apertures, indicate that one or more of the slots are taken up by a phase plate, in which case the diameter parameter is disabled.
7. Indicate whether this microscope is equipped with an autoloader or requires a side-entry holder.
8. Indicate whether the microscope is fitted with an energy filter.
9. Provide default values for extraction voltage, gun lens setting, spot size, and energy filter slit width (if relevant). The provided values will be used when users create Microscopy Sessions.

12. Register at least one electron detector

1. From the **Equipment** section, select **Electron Detector**.
2. Click on **Create New Electron Detector**.
3. Enter a label, manufacturer and model.

4. Select from a drop-down list the Microscope onto which this detector is mounted.
5. Add at least one magnification calibrated for this microscope-detector combination:
    1. Under magnifications, select **Add New**.
    2. Provide both nominal and calibrated magnification values.
    3. Repeat these steps for all magnification settings expected. These magnification settings will later be available in a drop-down selector for users logging Microscopy Sessions.
6. Use checkboxes to specify whether the detector is capable of electron counting, dose fractionation, and super resolution.
7. Finally, provide additional specifications of the detector: counts-per-electrons factor (the average number of counts registered by incident electron), the linear dimension of each pixel (in µm), and the numbers of rows and columns of pixels.
8. Click **Save**

13. If there are one or more microscopes that require side-entry sample holders, register available sample holders in gP2S.

1. From the **Equipment** section, select **Sample Holder**.
2. Click on **Create New Holder**.
3. Enter a label, manufacturer, model and location.
4. Specify the maximum tilt (in degrees) for the sample holder.
5. Use the checkboxes to specify whether it is capable of holding cryogenic EM grids, and whether it is capable of dual-axis tilting.
6. From a drop-down list, select all microscopes with which this holder can be used.

    NOTE: this will ensure that only relevant holders are listed when users register Microscopy Sessions using side-entry microscopes.
7. Click **Save**.

14. Specify the pattern that gP2S will follow in setting the directory name associated with each Microcopy Session.

NOTE: It can be very useful to have gP2S automatically generate a directory name for the storage of image data recorded during a Microscopy Session. This ensures systematic, information-rich naming of storage directories. Specify the pattern that gP2S will follow in setting the directory name associated with each Microscopy Session.

1. From the Admin section, select **Settings**.
2. Edit the directory name pattern string.

    NOTE: this string may contain the following variables: project label, Grid ID, Grid label, Microscopy Session label, Microscopy Session ID, Microscopy Session start date, Microscopy Session start time, and Microscope label, delimited by ${}. Other than these variables, directory name patterns may contain most characters. The default directory name pattern, for example, is ${GridLabel}_${MicroscopyStartDate}_ ${ProjectLabel}_${MicroscopeLabel}_grid_ ${GridID}_session_${MicroscopySessionID}. Now, sufficient Settings are configured to enable the registration of experimental entities up to and including Microscopy Sessions.

15. Register image processing software available to the users.

NOTE: This will enable the registration of Processing Sessions and later entity types (Maps and Models).

1. Select **Image Processing**.
2. Click on **Create New Image Processing Software**.
3. Type in the name of the software
4. List all versions available to users:
    1. Under software version(s), select **Add New**.
    2. Enter the software version.
       NOTE: This will enable users to specify exactly which version of the software they used to reach their results when registering Image Processing Sessions. This completes the necessary configuration of gP2S. Users should now be able to accurately capture key metadata describing their electron microscopy experiments, as described in the following section.

Representative Results

Overall design and navigation pattern

The gP2S application is project oriented, such that an entity can only be created in the context of a project. The relevant project is first selected from the dropdown located near the top left corner of the application. For convenience, the list of projects is filterable and it is sorted with the recently used projects shown at the top. When selecting a project, the number of entities of each type which are associated with this project is displayed in the workflow section of the left-hand-side navigation bar. The user can then click on any of the workflow entity types (e.g., Microscopy Sessions) to display a list of those entities within the selected project (**Figure 4**). This list consists, for each entity, of a label, date and time of creation, the name of the user who created it, an indication of whether any comments have been made about this entity, and up to six key metadata fields (for example, for each Microscopy Session: Grid, number of images, starting and finishing times, and what Microscope and Detector were used). Selecting one of the listed entities opens a details page listing all the information available for this item, including a summary list of all ancestor entities (for example, for a Microscopy Session, its parent Grid and Sample are listed). This allows for very quick navigation through the "lineage" of an entity, for example enabling single-click navigation from an atomic Model to the details of the Sample (**Figure 5**). In addition, any entity in gP2S can be commented on, by selecting "Comments" in the upper right part of its details page, entering a free-text comment, and optionally attaching one or more files.

Sample preparation

In the first step of the workflow describe the Sample. To do so, first define at least one component: Protein or Ligand.

Adding a new Protein requires only a protein label, but to help in better describing the protein add a PUR ID (for purification identifier). This field accepts any text and can for example contain a lot/batch number or serve as a place for a barcode label. If gP2S has been customized to integrate with a protein registration system (see Discussion), the PUR ID can be validated automatically and used to retrieve and display detailed information about this lot of protein. For Ligands, a label and stock concentration are mandatory information. All other fields are optional, and include: concept (barcode, common name or other ligand identifier) and batch/lot identifier. Again, if gP2S has been configured to integrate with a ligand registration system, the concept and lot

identifiers can be used to fetch and display externally-stored data describing the ligand (e.g. its chemical structure, assay results).

A Sample is defined by any combination of Proteins and Ligands and their final concentrations. Optionally, specify other experimental details of the sample such as incubation time and temperature, buffer and a free-text protocol description.

Grid preparation

When the Sample is ready, navigate to Grids. In the list, under each Grid's label find one or two colored tags that indicate the grid type (cryo or stain) and whether that grid is available for use. To create a new Grid, select **Create New Grid**. Type in a label, select the Grid Type and the Surface Treatment Protocol (e.g., glow discharge) used. Then, indicate whether preparing a cryo or negative stain grid, and select one of the pre-configured preparation protocols from the dropdown list, which is populated with Negative Stain Protocols or Vitrification Protocols, depending on the grid preparation type selected earlier. Next, select the appropriate Sample from the dropdown list and use a toggle switch to indicate if the sample remains available (described in more detail below). If choosing to dilute or concentrate the selected sample, indicate this using the "diluted/concentrated?" toggle and specify the relevant dilution or concentration factor. Specify the volume applied onto the grid (in µL) and optionally can also record an incubation time. Finally, define the Grid's storage location. For negative stain grids, record the storage box label/number and the Grid's position within the box. For cryo grids, first select a storage device from the list and then provide information for the available and appropriate fields (cylinder, tube, and/or box, depending on the Cryo Storage Device properties previously defined in the Settings).

The parts of the workflow that were described above, Samples and Grids, are part of an inventory management system. This feature keeps track of whether the components are still available for use.

1. A Protein or Ligand can be made unavailable from the Sample level. When creating a Sample, selecting "last drop" for any of that Sample's components marks those components as unavailable for future use: they will no longer be available in the drop down when creating Sample, and they will not be marked by the "Available" tag in the list view.
2. A selected Sample can be marked as unavailable by using one of the two toggle switches - "Available for grid-making?" (under Samples) or "Sample is available for further use?" (under Grids).
3. To manage the grid's availability, use the "Grid returned to storage?" toggle (under Microscopy Sessions). By default, this value is set to "Yes" for all negative stain grids and to "No" for cryoEM grids.

Data collection

Once the grids are registered, register data collection experiments by creating Microscopy Sessions in gP2S. Microscopy Session is the most complex experimental entity tracked by the application and it is organized into four sections: basic information, microscope settings, exposure settings and microscope control.

The first section contains basic information: a Microscopy Session label, its start and finish dates and times, what Grid was imaged, which Microscope, Detector and Sample Holder (if applicable) were used, and how many images were collected. When creating a new Microscopy Session, the system automatically fills in the starting date and time. Finishing date and time are optional. This is because a

Session may be registered in the system while the experiment is still ongoing and therefore its ending time would not be precisely known. If the finish date and time are not known, type it in manually or use the "now" button to enter the current date and time. Another way is to take advantage of the fact that gP2S does not allow more than one unfinished Microscopy Sessions on any given Microscope. Starting a new Microscopy Session on the same Microscope automatically marks any previously-started Session as finished.

In the next step, choose the Grid. The dropdown list will have all available Grids in the current project. After choosing a Grid, some of its basic information will be seen: who created it and when, and what Sample was applied to it. Depending on what type of grid is selected, the Microscopy Session will be marked as "stain" or "cryo" on the list view.

By default, the Microscope most recently used in the current project is pre-selected. If a particular Microscope has a sample insertion mechanism defined as an autoloader, this is the information displayed as the Sample Holder. However, if the selected Microscope requires the use of side entry holders, select the holder used from the list of Sample Holders configured to work with this microscope (if the selected grid is a cryo grid, only cryo-capable holders are listed).

The second section of a Microscopy Session form contains information about Microscope settings such as extraction and acceleration voltages, gun lens, diameter of C2 aperture, objective aperture and energy filter slit width. During routine usage, these settings are rarely changed because users commonly do not have to deviate from default values.

The third section of the Microscopy Session contains information about exposure settings. In this section the following metadata are recorded: magnification (pixel size), spot size, diameter of illuminated area, exposure duration, and whether nanoprobe, counting mode, dose fractionation and super resolution were used (counting mode, dose fractionation and super resolution settings are only enabled if the selected Detector has these features). If dose fractionation was used, the number of frames and exposure rate are also recorded.

For convenience, a number of experimentally important parameters are calculated on the fly and displayed within the form: the final image pixel size (Å), exposure rate (electrons/Å^2/s), total exposure (electrons/Å^2), frame duration (s) and exposure per frame (electron/Å^2).

The fourth and final section of the Microscopy Session can be used to record the minimum and maximum target underfocus, and the number of exposures per hole.

While Microscopy Sessions in gP2S can be used to register any type of microscopy work, be it for screening or data collection purposes, we have found that it is sufficient and more efficient to ask users to focus on registering data collection sessions, and that screening sessions, wherein a grid is only briefly inspected for quality control need not necessarily be registered as Microscopy Sessions.

Image processing

Image processing work is recorded in gP2S as Processing Session entities. Each Processing Session is related to one or more Microscopy Session, which must be selected from a dropdown list. Indicate which Software packages (programs and versions) were used, the number of micrographs and number of particles picked. Optionally, record the name of the directory of the processing.

Map deposition

Once one or more three-dimensional reconstructions have been obtained, the Maps can be deposited into gP2S. Each Map is associated with a Processing Session, and consists of the actual map file (typically an MRC-formatted file, but gP2S allows for any file type) and key metadata: size of the pixel (Å), recommended isocontour level for surface rendering, what symmetry is applied, number of images used to create the map, and the estimated resolution: in its best and worst parts as well as the average global resolution. Maps may be associated with each other using the following types of relationships: filtered, masked, resampled, or refined versions. When registering such an association, select the type of relationship (e.g., "is filtered version of " or "has filtered version").

Model deposition

Once an atomic model has been obtained, it can be deposited into gP2S's Model section for the relevant project. The Model feature in the first release of gP2S is barebones: other than the actual model file (typically a PDB or mmCIF file), only the resolution (in Å) and the Map (or list of Maps) from which the model was derived, are required. Additionally, it is possible to indicate that a Model is a refined version of a previously-deposited Model. Additional features, including model validation, are under development and may be added to the open-source version of gP2S in future.

Reports

It may be necessary to generate summary documents to be distributed to collaborators, who may not have access to gP2S, or to be archived on a filesystem. gP2S provides a report functionality for this purpose, available via a printer icon at the top right of each entity details view page. This generates a printable PDF file that includes all metadata describing the entity and each of its ancestor entities, including all comments. This feature is particularly valuable following Model deposition, since all data and metadata tracing the lineage of the final atomic model all the way back to specific protein and small molecule ligand lots via Microscopy Session(s) and Grid(s) will be available in a single document.

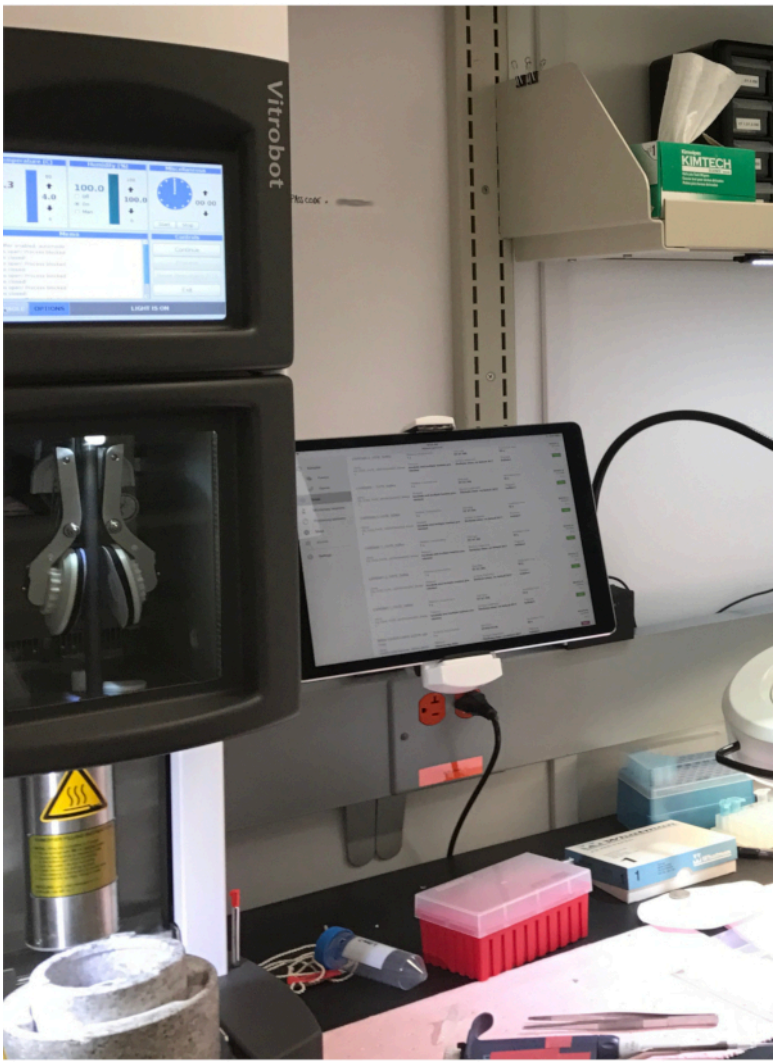

Figure 1. gP2S running on an iPad at a vitrification lab bench. The user interface has been designed for operation using touch screens, which facilitates in-lab use and accurate metadata entry. Please click here to view a larger version of this figure.

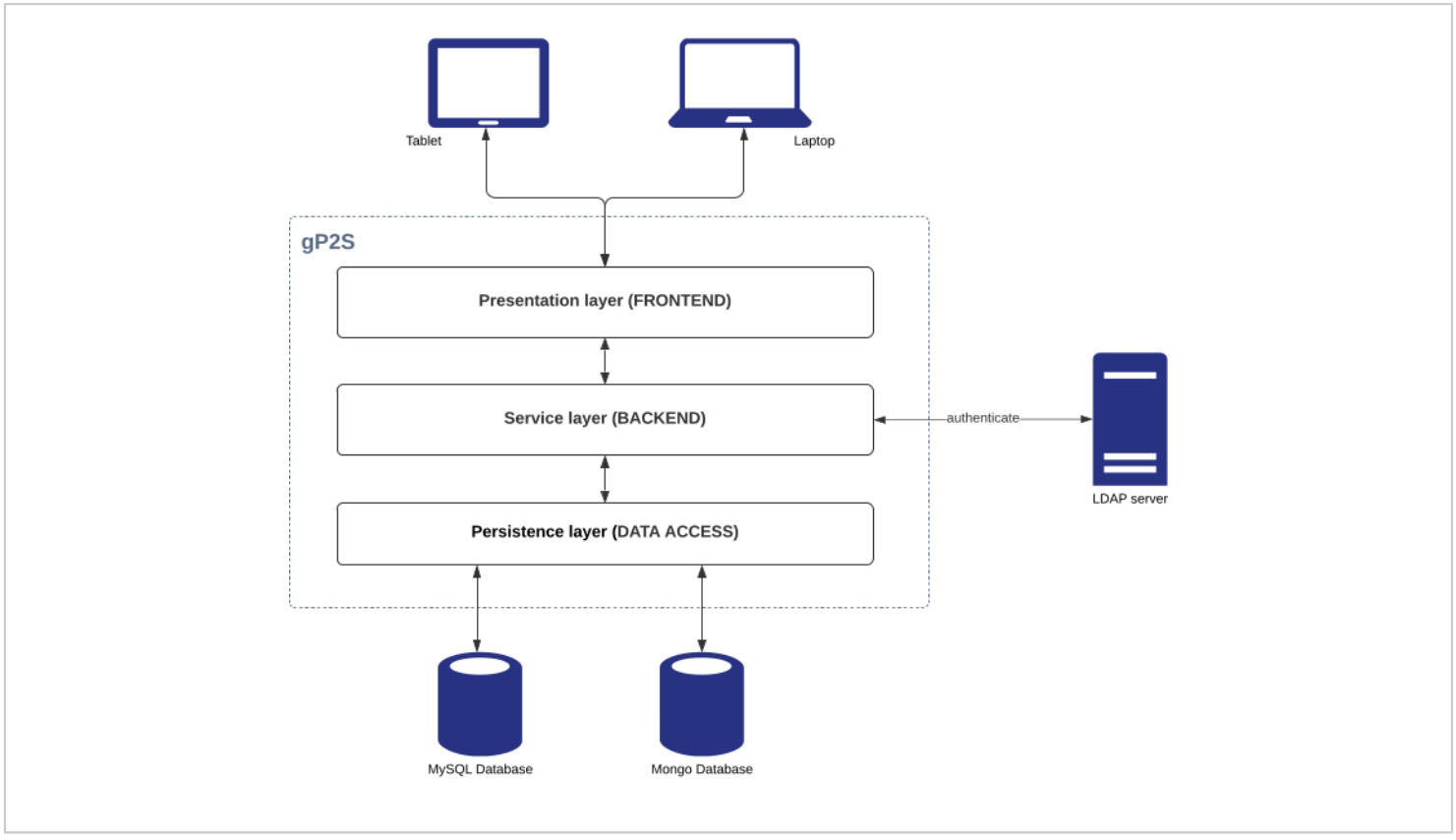

Figure 2: gP2S system architecture. gP2S follows a classic three-tier organization and relies on two database servers for data storage and an LDAP server for user authentication. Please click here to view a larger version of this figure.

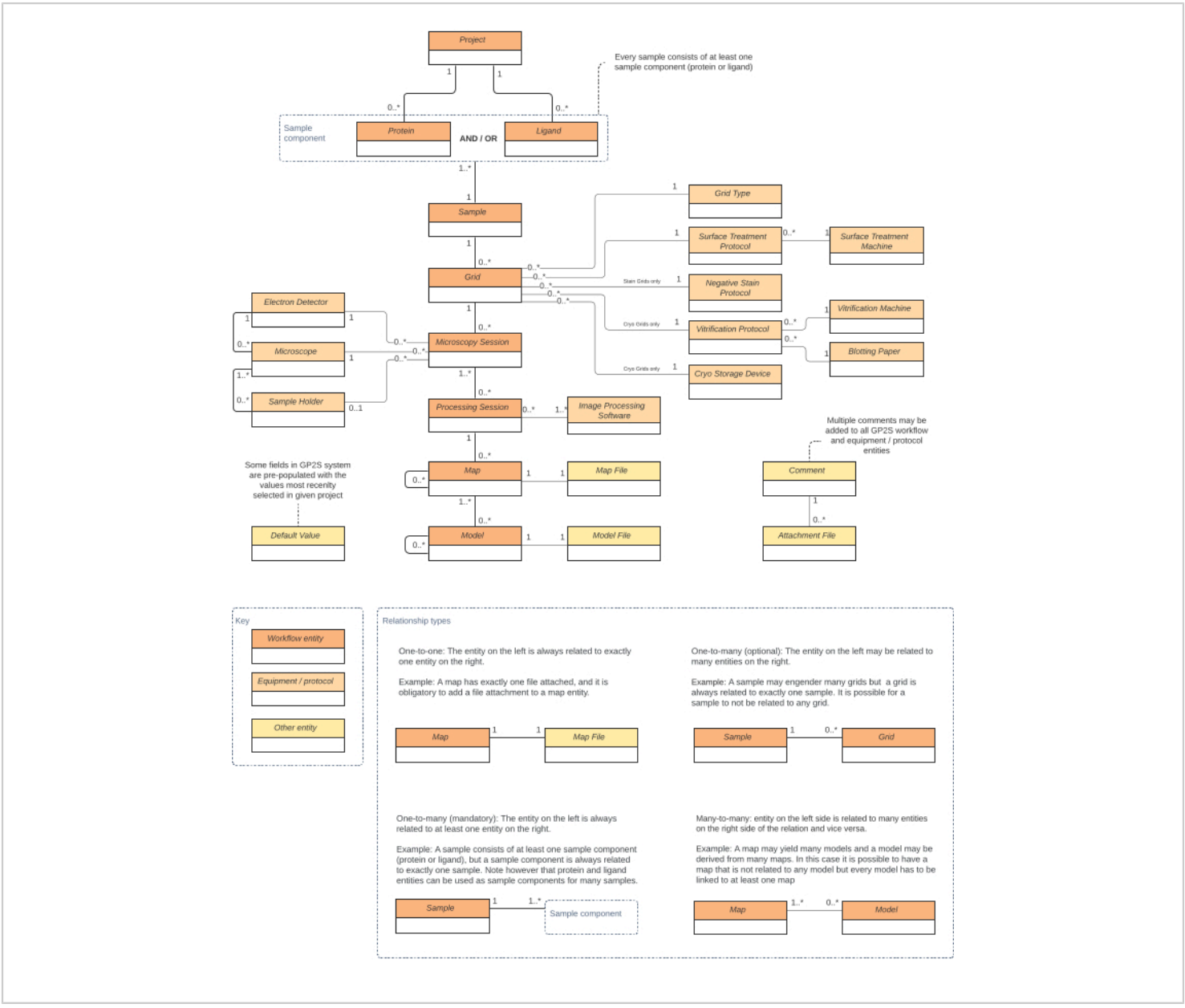

Figure 3: The gP2S data model. Entities are depicted as rectangles (dark orange for workflow entities, orange for equipment and protocols, yellow for other entity types), and their relationships are (one-to-one, one-to-many, many-to-many) denoted by continuous lines. Please click here to view a larger version of this figure.

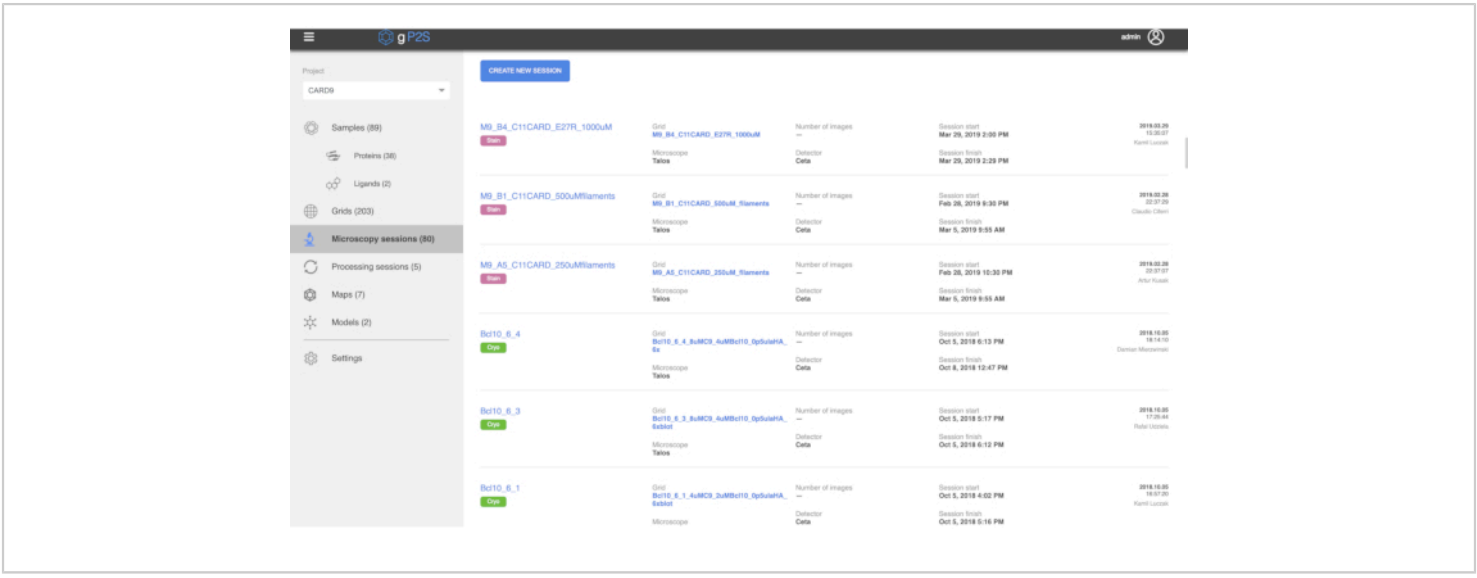

Figure 4. Microscopy Session list view. In this view, all Microscopy Sessions registered under the selected project ("CARD9" in this screenshot) are listed. A green or purple tag differentiates between room-temperature (negative stain) and cryogenic Microscopy Sessions, and a few key metadata describing each session is listed (e.g. the user who registered it, at the far right). Clicking on the name of a Microscopy Session opens a detailed view of that Session (a detailed view of a Model is shown in **Figure 5**). Please click here to view a larger version of this figure.

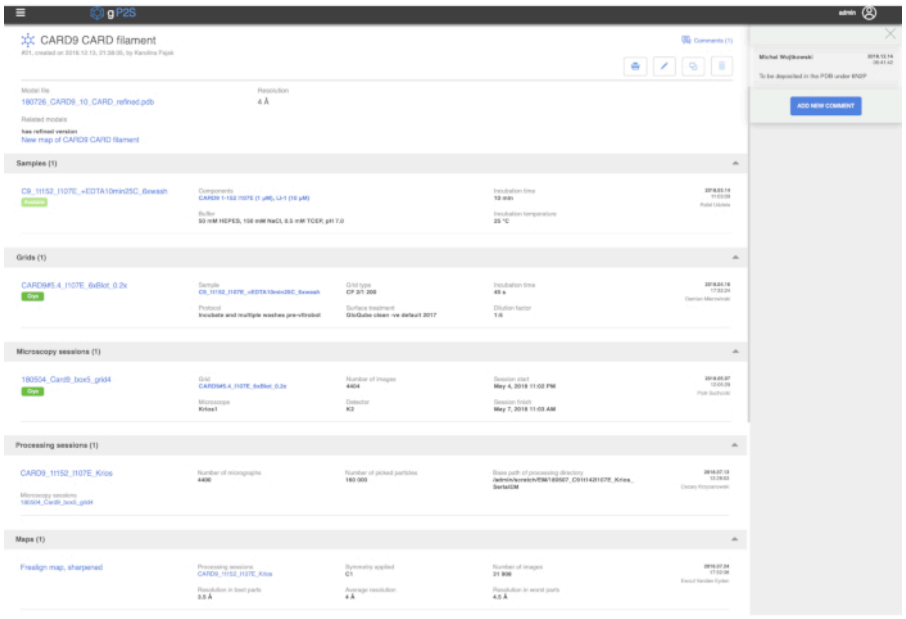

Figure 5. Model detail view. The top part of the page shows available metadata for the selected model. The comment pane on the right can be hidden by clicking on the cross (top right) or the "Comments (1)" to its left. Below, a set of icons enables the generation of a PDF report (printer icon, see main text), editing of the entry (pencil icon), or duplicating it (double rectangles icon). The lower part of the page contains a structure list of all of the entities from which this Model is descended, from Samples to Maps. Please click here to view a larger version of this figure.

Name of the library or framework	Type	Version
ApacheDS	LDAP server	0.7.0
Docker	development tool	n/a
Element	library	1.4.10
Hibernate	library	5.0.12
Java	programming language	1.8+
JavaScript	programming language	EcmaScript 2017
JUnit	library	4.12
Karma	library	1.4.1
Maven	development tool	3+
MongoDB	DB server	4.0.6
MySQL Database	DB server	5.7
Node.js	framework	6.9.1
SASS (node-sass)	library	4.5.3
SpringBoot	framework	1.3
Swagger UI	library	2.6.1
Tomcat	application server	8.5.15
Vue.js	framework	2.4.2
vue-cli	development tool	2.6.12

Table 1. Libraries and frameworks used by gP2S

Discussion

When used properly and consistently, gP2S helps achieve proper record keeping of high-quality metadata by enforcing the recording of critical experimental metadata using structured data models and defined vocabularies, but the added value of this is only fully realized when a high level of compliance is achieved in the lab. The above protocol does not cover how to achieve this. We found that an effective enforcement technique was to have microscope operators refuse to collect data on grids not registered in gP2S. This drove compliance up very quickly and laid the ground for the emergence, over the following months, of a large body of detailed and accurate experimental details and corporate memory. After a few months of usage, the value of the corpus of metadata stored in gP2S became so obvious to most users that compliance remained high without explicit intervention.

Fully leveraging this collective memory requires that the metadata stored in gP2S be accessible to external systems and easily associated with the experimental data (micrographs) and results (maps and models). The above protocol does not describe how to integrate gP2S with other informatics and data processing systems. Most straightforward are potential integrations via gP2S's backend REST API, which do not require any modification of gP2S. For example, each computer controlling our data collection detectors runs a script which periodically queries gP2S's endpoint "getItemByMicroscope" under the microscopy session management REST controller, to check whether a Microscopy Session is ongoing on its microscope. If so, the script retrieves from gP2S the appropriate data storage directory name (as configured in the Settings page, see above), and creates a directory on the local data storage device using this name. This ensures systematic naming of data storage directories and reduces the risk of error due to typos.

Although they have been commented out in the source of the public version of gP2S, further integrations involving gP2S consuming external systems' data are also possible. In our lab, our deployment of gP2S integrates with (i) a project management system, so that each project configured in gP2S can be linked to a company-wide portfolio project, and metadata from the portfolio can be displayed within gP2S; (ii) a protein registration system, so that each protein added to gP2S is linked, via an identifier stored locally, to a complete set of records detailing the provenance of the protein, include details of the relevant molecular biology, expression system and purification; (iii) a small molecule compound management system, allowing gP2S to display key information about each ligand, such as its chemical structure. The code modifications necessary to enable these integrations are described in the "Integration" section of the README-BUILD.md document available from the gP2S repository (https://github.com/arohou/gP2S).

The current version of gP2S has limitations, first among which is the overly simplistic data model and frontend for structure (Model) deposition. This was intentionally left in a "barebones" state in the released version of gP2S because a fully-fledged structure deposition and validation feature is currently under development together with support for X-ray crystallography. Another design decision was to not implement any privilege or permission system: all users in gP2S have equal access to its features and data. This may make it a poor choice for facilities who serve user groups with competing interests and confidentiality requirements, but was not a concern for our facility.

Development of our in-house version of gP2S is ongoing and it is our hope that the open-source version described here will be useful to other cryoEM groups, and that some may contribute suggestions, or code improvements in future. Future high-value developments could for example focus on integrations with lab equipment (vitrification robots, electron microscopes), software (e.g. to harvest image processing metadata) and external public repositories (e.g. to facilitate structure depositions).

The systematic collection of high-quality metadata enabled by routine use of gP2S in the lab and cryoEM facility can have a significant, positive impact on the ability to prosecute multiple projects in parallel over a period of years. As more and more shared and centralized cryoEM groups and facilities are established, we anticipate the need for information management systems such as gP2S will continue to grow.

Disclosures

All authors are contractors with or employees of Roche or of its subsidiary Genentech.

Acknowledgments

The authors thank all the other members of the gP2S development team who have worked on the project since its inception: Rafał Udziela, Cezary Krzyżanowski, Przemysław Stankowski, Jacek Ziemski, Piotr Suchcicki, Karolina Pająk, Ewout Vanden Eyden, Damian Mierzwiński, Michał Wojtkowski, Piotr Pikusa, Anna Surdacka, Kamil Łuczak, and Artur Kusak. We also thank Raymond Ha and Claudio Ciferri for helping assemble the team and shape the project.

References

1. Cheng, Y., Grigorieff, N., Penczek, P.A., Walz, T. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell.* **161** (3), 438-449 (2015).
2. *High-End Cryo-EMs Worldwide.* at https://www.google.com/maps/d/u/0/viewer?mid=1eQ1r8BiDYfaK7D1S9EeFJEgkLggMyoaT (2021).
3. Renaud, J.-P. *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature Reviews Drug Discovery.* **17** (7), 471-492 (2018).
4. Alewijnse, B. *et al.* Best practices for managing large CryoEM facilities. *Journal of Structural Biology.* **199** (3), 225-236 (2017).
5. Rees, I., Langley, E., Chiu, W., Ludtke, S.J. EMEN2: An Object Oriented Database and Electronic Lab Notebook. *Microscopy and Microanalysis.* **19** (1), 1-10 (2013).
6. Delagenière, S. *et al.* ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics.* **27** (22), 3186-3192 (2011).
7. Rosa-Trevín, J.M. de la *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *Journal of Structural Biology.* **195** (1), 93-99 (2016).
8. *EMPIAR deposition manual.* at <https://www.ebi.ac.uk/pdbe/emdb/empiar/deposition/manual/#manScipion>. (2021).
9. Iudin, A., Korir, P.K., Salavert-Torres, J., Kleywegt, G.J., Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods.* **13** (5), 387-388 (2016).
10. *Vue.js.* at https://vuejs.org/ (2021).
11. *Spring Boot.* at https://spring.io/projects/spring-boot (2021).
12. *Lightweight Directory Access Protocol.* at https://ldap.com/ (2021).
13. *Vue CLI.* at https://cli.vuejs.org/ (2021).
14. *Element, A Desktop UI Library.* at https://element.eleme.io/ (2021).
15. *Sass.* at https://sass-lang.com/ (2021).
16. *Karma.* at http://karma-runner.github.io/ (2021).
17. *Node.js.* at https://nodejs.org/ (2021).
18. *Java.* at https://www.java.com/ (2021).
19. *Apache Tomcat.* at http://tomcat.apache.org/ (2021).
20. *Hibernate.* at https://hibernate.org/ (2021).
21. *Swagger UI.* at https://swagger.io/tools/swagger-ui/ (2021).

22. *JUnit.* at <https://junit.org/junit4/> (2020).

23. *Apache Maven Project.* at <https://maven.apache.org/> (2020).

24. *MySQL.* at <https://www.mysql.com/>. (2021).

25. *mongoDB.* at <https://www.mongodb.com/>. (2021).

26. *Apache license, version 2.0.* at <https://www.apache.org/licenses/LICENSE-2.0> (2004).

27. *mysql Docker Official Image.* at https://hub.docker.com/_/mysql (2021).

28. *mongo Docker Official Image.* at https://hub.docker.com/_/mongo (2021).

29. *openmicroscopy apacheds.* at https://hub.docker.com/r/openmicroscopy/apacheds (2021).