Institute of Molecular Genetics
of the Czech Academy of Sciences
Vídeňská 1083
142 20 Prague 4, Czech Republic
www.img.cas.cz

Prague, 18th July 2023

A cloud-based chemical data repository for the "Proxidrugs" research project (1.9.2023 - 31.8.2024)

QUOTATION

ISSUED TO	ISSUED BY
Fraunhofer-Institute for Translational Medicine and Pharmacology ITMP Drug Discovery Research ScreeningPort Schnackenburgallee 114 22525 Hamburg Germany Phone: [redacted] Fax: [redacted] E-mail: [redacted]	**Institute of Molecular Genetics of the Czech Academy of Sciences** **CZ-OPENSCREEN** Vídeňská 1083 142 20 Prague 4 Czech Republic ID: 68378050 VAT ID: CZ68378050 Phone: [redacted] Fax: [redacted] E-mail: [redacted]

DESCRIPTION	PRICE (EUR)
The hosted solution of the chemical data repository for the Proxidrugs research project for one year (1.9.2023-31.8.2024) including: • daily backups • online support via service desk • system updates • software updates	20,000

Track record

We have strong expertise in cheminformatics/software development and a long (10 years+) track record in building up integrated solutions for the management of chemical and biological data, both in-house and for the wide scientific community.

Our projects in this area include:

1. **ScreenX** - Laboratory Information Management System (LIMS) and database for HTS and compound management. Used by the CZ-OPENSCREEN infrastructure and several other facilities in the Czech Republic and one abroad (Portugal).

2. **European Chemical Biology Database (ECBD)** - a central data repository for data generated within the EU-OPENSCREEN network, developed and managed at IMG since 2019. ECBD is developed in line with the FAIR principles ensuring Findability, Accessibility, Interoperability, and Reusability of the data.

   PUBLIC URL: *ecbd.eu*

3. **Probes & Drugs portal** (P&D) - a hub for the integration of high-quality bioactive compound sets enabling their analysis and comparison. Its main focus is on chemical probes and drugs but it also includes additional relevant sets from specialist databases/scientific publications and vendor sets. Upon these, established bioactive chemistry sources are utilized for compounds' biological annotation. P&D was released in 2017 and became one of the most comprehensive resources in the field of high-quality chemical tools.

   PUBLIC URL: *probes-drugs.org*

Technical implementation

Our proposed solution of the chemical data repository (CDR) for the PROXIDRUGS research project is a web application hosted at CZ-OPENSCREEN at the Institute of Molecular Genetics AS CR, v. v. i.

The server side of the CDR relies primarily on the Python programming language with Django web framework in combination with the PostgreSQL database. Both Python and Django are well-established programming tools used by the general public with a broad range of available scientific packages, including biology and chemistry. PostgreSQL is a high-performance relational database conforming to the ACID (Atomic, Consistent, Isolated, Durable) transaction properties. For the front-end, the VueJS JavaScript framework is utilized together with both svg- (vector) and canvas-based (raster) visualization libraries.

Regarding the security of the repository itself, the Django framework implements many security features that mitigate the threat of the most common malicious attacks, such as Cross-site scripting, Cross-site request forgery, SQL injection or Clickjacking. These security features are frequently updated together with the framework to prevent new potential threats.

Users are encouraged to use safe passwords and the login procedure is protected against automated login or account creation attempts using the Google CAPTCHA tool. All personal data are transferred in an encrypted form using HTTPS protocol.

The database is created and managed predominantly with the Django ORM; only more complex queries are performed with raw SQL. The database structure is designed using standard database normalization techniques (the third normal form, 3NF) with an emphasis on stability and performance.

IT infrastructure

Our CDR solution is operated with the utmost effort to ensure service availability. The system is running on a virtual server in an environment composed of two clusters of hardware servers. Individual hardware components of the environment are redundant and each cluster is housed in a different data center of the Institute of Molecular Genetics. The system is replicated from the primary to the secondary cluster every 30 minutes. In case the primary cluster becomes unavailable, the service can be restored in a short time on the secondary cluster by the system administrator. The long-term server availability (accessibility) is higher than 99%.

All communication of the users with the CDR system is secured via SSL. Database dumps are performed once a day and archived for 6 months, in case data restoration is necessary.

There is a dedicated time window used for regular maintenance of the system – Thursday, 8-10 PM local time in Prague (CET/CEST) - during which the system unavailability may occur. However, all scheduled maintenance services are announced in advance.

Compliance with the specifications

Our CDR solution is meeting a large majority of the requirements specified in **Section 5 (Requirements)** of the CDR specifications (A-7, A-16 and N-4 are currently not met, N-5 and N-6 are not applicable for our - hosted - solution). The following table contains the information about the compliance with the individual requirements:

A) Functionality Requirements

URS	Requirements	Priority	CDR
A-1	The repository enables storage of various data types chemical and bioactivity data such as bioactivity, and related chemical data	1	**YES**

	including structures		
A-2	The software database must be a secured repository within the EU region with the server hosted and maintained by the vendor	1	**YES**
A-3	It shall be possible to effectively implement FAIR data management principles	1	**YES**
A-4	It shall be possible to use ontologies for meta-data annotation	1	**YES**
A-5	The system shall provide templates and well-defined guidelines for submission of data/results	1	**YES**
A-6	The system shall allow submission of data via web or API	1	**YES**
A-7	It shall be possible to connect data upload/ download processes with KNIME Analytics Platform	2	**NO**
A-8	The software shall provide data provenance information, i.e. the information about data uploader and contact details	1	**YES**
A-9	It shall provide audit trail functionality i.e. creating, modifying and deleting entries must be traceable	1	**YES**
A-10	The software must provide a personalized login for the users	1	**YES**
A-11	It shall be possible to structure and group uploaded data/results, e.g. according to project partners	1	**YES**
A-12	It shall be possible to customize the user rights management	1	**YES**
A-13	The software shall warn if errors occur during submission of data and possible solutions shall be suggested by the system	1	**YES**
A-14	It shall provide the following file export formats: csv, xlsx, tab, sdf, mol	1	**YES**
A-15	The system shall be able to provide sharable links for specific datasets	1	**YES**
A-16	The software provides version-tracking of content and uploaded files	2	**NO**
A-17	The possibility shall be given to cross-reference with external database. For example, 1) it should be possible to be directed to the repository by clicking a URL in a graph database. 2) it should be possible to be directed to the graph database while being inside the repository	2	**YES**
A-18	It shall be possible to deactivate a user account if that user leaves the project, goes on long-term leave or moves to a department where use is not required, without loss of information within the system.	2	**YES**
A-19	Visualization of chemical structures shall be possible, e.g. drawing 2D/3D chemical structures from InChI keys or smiles provided	1	**YES**
A-20	Data visualizations such as scatter plots and bar plots shall be supported in the software	1	**YES**

B) Data requirements

URS	Requirements	Priority	CDR
B-1	Data must be assigned explicitly	1	**YES**
B-2	Configurable mandatory data entry fields for metadata must be possible	1	**YES**
B-3	Configurable optional data entry fields for metadata must be possible	1	**YES**
B-4	Rejected data must be excluded from reporting or marked as invalid	2	**YES**

C) Output Requirements

URS	Requirements	Priority	CDR

C-1	Exporting single datasets in the following file formats must be supported: csv, xslx or tab. In addition, chemical specific formats such as sdf or mol shall also be supported	1	**YES**
C-2	Export of the entire database must be possible in the following format:: sql dump or PostgreSQL	1	**YES**
C-3	Export functionality based on a search results shall be available ("filtered ex-port")	1	**YES**

D) System technology requirements

URS	Requirements	Priority	CDR
D-1	The repository server must run on common work group server hardware	2	**YES**

E) Authorization concept

URS	Requirements	Priority	CDR
E-1	Definition of administrators, superusers and users and a highly granular definition of access rights is required	1	**YES**
E-2	The software must allow to define roles and groups and assign users to groups and roles	1	**YES**
E-3	If a user leaves, the account should be de-activated without loss of information/data created by this user	1	**YES**
E-4	User access rights must be changeable during the run of the project	1	**YES**
E-5	Guest login with limited access for people outside the project should be supported	2	**YES**

F) User interface requirements

URS	Requirements	Priority	CDR
F-1	An intuitive web client must be available	1	**YES**
F-2	During the data upload process intermediate versions shall be saved to continue the upload procedure at a later point of time	2	**YES**
F-3	The UI should be accessible via standard web browsers such as Microsoft Edge, Firefox and Chrome. Any performance issues with standard browsers should be informed.	1	**YES**
F-4	It must be possible to prevent the loading of encrypted files or password-protected files to a write-up.	2	**YES**

G) Error Processing Requirements

URS	Requirement	Priority	CDR
G-1	Uploading data should not influence the accessibility of the system by other us-ers	1	**YES**

G-2	Errors must be collected in a log file and human-readable	2	**YES**

H) Data migration requirements

URS	Requirements	Priority	CDR
H-1	The provider ensures that the repository/database is exportable and accessible by standard database applications	1	**YES**

I) Data protection requirements

URS	Requirements	Priority	CDR
I-1	Data backup and recovery must be possible for both hosted and in-house installations of the system	1	**YES**
I-2	The system must be able to detect and report attempts at unauthorized access and have safeguards built in to prevent such attempts	1	**YES**

J) System operation requirements

URS	Requirements	Priority	CDR
J-1	The system allows access by multiple users and from multiple sites in parallel	1	**YES**
J-2	The system shall be available on-line for 95% of the core hours. A lower level of availability will be acceptable outside these hours	1	**YES**
J-3	Full operation of the system must be recovered after a major failure within three working days	2	**YES**
J-4	In the event of system recovery, no more than 1 day of data should be lost	1	**YES**
J-5	The back-end system should support a replicated/redundant configuration on two different sites or operating in a master-slave fashion with automatic failover	1	**YES**
J-6	Any planned backups or maintenance should be informed at least a week in advance	2	**YES**
J-7	Any planned backups or maintenance should ideally take place outside office hours	2	**YES**

K) Operating system requirements

URS	Requirements	Priority	CDR
K-1	The application must be able to run on a standard server OS	1	**YES**
K-2	Repository must be able to be accessed under Windows, Linux and Mac OS	1	**YES**
K-3	The application must run on desktop PCs and Laptops equipped with standard computer hardware	1	**YES**

L) Server requirements

URS	Requirements	Priority	CDR
L-1	The repository must use a standard database (ORACLE, MySQL, PostgreSQL) for data storage	2	**YES**
L-2	The server of the database must be located in the European area (EU).	**1**	**YES**

M) Manufacturer requirements

URS	Requirements	Priority	CDR
M-1	The provider must provide support to users and administrators during normal working hours in Europe	2	**YES**
M-2	The system must be operational within 3 months after contract signature	1	**YES**
M-3	On-site support must be available within defined reaction times	2	**YES**
M-4	System updates must be supported by the provider	1	**YES**
M-5	It must be possible to deploy the repository in a phased approach, deploying first to an initial set of users and then bringing additional groups of users online at a later time. In other words, there must not be any constraints related to the design and construction of the system that prevents this being possible	1	**YES**

N) Documentation Requirements

URS	Requirements	Priority	CDR
N-1	The provider must deliver a technical specification of the system	2	**YES**
N-2	The software must have an online help capability	1	**YES**
N-3	The database structure with all relations must be documented and accessible for administrators	2	**YES**
N-4	The provider must deliver an English operating manual for administration	2	**NO**
N-5	If the provider will not host the resource, the provider shall provide a list of the common maintenance tasks required to keep the system running optimally, along with the regularity with which such tasks should be performed. Such routine maintenance should be possible without the need to go back to the provider for consultancy.	1	**-**
N-6	The provider will also provide information on additional system maintenance routines that need to be performed periodically, based on configuration changes, feature requests, and fixes. If some changes are needed for a specific defect, provider representatives are expected to be available and onsite (if required) to resolve.	1	**-**

Price

The price for our CDR solution is **20,000 EUR per year** for service and support.