D6.5
Data Management Plan

for
H2020-INFRADEV-2018-1
(Development and long-term sustainability of new pan-European research infrastructures)

Research and Innovation Action (RIA)

Action Acronym:
EU-OPENSCREEN-DRIVE

Action Full Title:
“Ensuring long-term sustainability of excellence in chemical biology within Europe and beyond”

Grant Agreement No.: 823893

Dissemination level: confidential, only for members of the Consortium (including the Commission Services)

Document Properties:

Deliverable	D6.5 Data Management Plan
Partner responsible	EU-OS, IMG, IME, MEDI, IMIM, CSC, IME
Author(s)	Milan Vorsilak, Petr Bartunek, Sarka Simova, Katholiki Skopelitou

1. Data summary	
a)	What is the purpose of the data collection / generation and its relation to the objectives of the project?
	ECBD will store all datasets generated by EU-OPENSCREEN screening partner sites during the EU-OPESNCREEN DRIVE project for systematic investigation of chemical substances on biological systems. ECBD is operated by IMG, the database partner site of the EU-OPENSCREEN ERIC and EU-OPENSCREEN DRIVE project participant. EU-OPENSCREEN partner sites generate during the EU-OPENSCREEN DRIVE project substantial amount of data covering a variety of information on chemicals and their bioactivities, proteins, cellular pathways, assays, screens and chemical optimisation programs. Appropriate data management aligned with FAIR principles is an essential activity of the EU-OPENSCREEN DRIVE project. Data will be accessible to the scientific community upon the open access principles. Before the data go public, they will undergo necessary process of validation and inspection and only then they will be released to the scientific community.
b)	What types and formats of data will the project generate / collect?
	ECBD will be mainly web-based service, however, public part of the database will be accessible as csv file for each assay or as SQL dump for Postgres database. Assay data will be mostly link to ontologies and numerical data.
c)	Will you re-use any existing data and how?
	All data will be novel.
d)	What is the origin of the data?
	Data will be generated by EU-OPENSCREEN screening partner sites during the EU-OPESNCREEN DRIVE project.
e)	What is the expected size of the data?
	The size of one assay should not exceed 10 MB and during the project, 15 assays should be screened.
f)	To whom might the data be useful ('data utility')?
	The data will be available to the broad scientific community and it will be important mainly for researchers in the fields of chemical biology and drug discovery. The data will be available via ECBD portal at www.ecbd.eu.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement

2. FAIR data

2.1 Making data findable, including provisions for metadata

a) Are the data produced and / or used in the project discoverable and identifiable?

The data will be accessible through ECBD web portal as webpages, where user can browse measured assay data, i.e. activities for selected biological targets and substances.

The data search in ECBD will be implemented in a Google-like fashion, i.e., all data will be able to be searched from a simple persistent (i.e., always available) single field using an autocomplete function. During the text search, a list of prioritized terms, accompanied by their entity types (e.g., compound, target, assay, etc.), will be suggested to a user.

Compounds in screening set and screened assays will be identified by persistent and unique EOS number, which will be designed by ECBD staff. Compounds have reserved interval from EOS1 to EOS299999. Compounds will be linked by Unichem database especially to PubChem and ChEMBL.

Research data may be linked to the corresponding publications via DOI and in publications can be referenced via EOS.

b) What naming conventions do you follow?

Unique EOS number will be assigned to substances, assays and biological targets. EOS number will be persistent, the use of DOI is planned only for publications.

Naming convention will be derived from ontologies, i.e. for taxonomy and biological targets, and ontology terms.

c) Will search keywords be provided that optimize possibilities for re-use?

Ontology terms and their synonyms will serve as search keywords.

d) Do you provide clear version numbers?

Assay data will be uploaded at once with timestamp of creation and later with timestamp of validation. Revisions will be made as a new assay upload with new EOS number and during validation the old assay data will be marked as obsolete, however will stay accessible.

e) What metadata will be created?

Assay will be described with necessary metadata as creator, the dates of measurement, upload and validation; and description how and for which biological target was assay measured.

2.2 Making data openly accessible

a) Which data produced and / or used in the project will be made openly available as the default?

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement

	All validated assay data will be publicly available. If needed, a grace period of up to three years can be requested for particular data to allow researchers to protect their intellectual property rights. However, assay description and metadata will be accessible by default.
b)	How will the data be made accessible (e.g. by deposition in a repository)?
	The data will be accessible through ECBD web portal (www.ecbd.eu) and the upload of bioactive subset to ChEMBL database, which collects wide range of assay data from many sources, is also planned to provide higher re-usability.
c)	What methods or software tools are needed to access the data?
	Generally, data will be accessible via common (not older than 2 years) internet browser, however, it will be possible to download csv files, which will be human and computationally readable. More advanced way to obtain data will be through SQL dump, which will be available for Postgres database extended by RDKit with its Postgres cartridge (not for general use).
d)	Is documentation about the software needed to access the data included?
	Not for general methods.
e)	Is it possible to include the relevant software (e.g. in open source code)?
	Relevant software (Postgres and RDKit) is open source and is not necessary for data manipulation, thus links are provided.
f)	Where will the data and associated metadata, documentation and code be deposited?
	Data, metadata and documentation will be deposited at Cesnet infrastructure, where the portal will be served. Source code of the portal will be provided by IMG to ERIC, but is not planned for open access. All the data and associated metadata and documentation will be deposited in ECBD. During the project, bioactive data will be additionally uploaded to ChEMBL database. The transfer of fully curated public ECBD data to ChEMBL will follow data and format requirements defined in the ChEMBL Gateway deposition rules.
g)	Have you explored appropriate arrangements with the identified repository?
	ECBD is an in-house platform developed directly for the storage of data generated by chemical biology.
h)	If there are restrictions on use, how will access be provided?
	Restricted data such as not validated data or data within the grace period will be accessible only to authorized users, mainly to data uploaders. Metadata will be accessible without restrictions.
i)	Is there a need for a data access committee?
	All validated data will be publicly available after the grace period, therefore a data access committee is not necessary.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement

j)	Are there well described conditions for access (i.e. a machine-readable license)?
	Data will be accessible under Creative Commons Attribution-ShareAlike 4.0 International license (https://creativecommons.org/licenses/by-sa/4.0/).
k)	How will the identity of the person accessing the data be ascertained?
	The data will be accessible upon open access principles. The initial access to the data will be without any login or user account. ECBD will optionally provide login to registered users. Only privileged users (i.e. members of partner site and their contractors) can access data in the grace period.
2.3 Making data interoperable	
a)	Are the data produced in the project interoperable?
	During assay data upload will be annotated by data uploader with ontologies: • BioAssayOntology (http://bioassayontology.org/) • BRENDA (https://www.brenda-enzymes.org/) • Chembl Drug Target subset (http://geneontology.org/docs/go-subset-guide/) • Units of Measurement Ontology (https://github.com/HajoRijgersberg/OM) • Cellosaurus (https://web.expasy.org/cellosaurus/) • NIH Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) Relevant links will be provided to other information portals.
b)	What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
	The standard ontologies will describe assay data, units will be also specified by ontology. Chemical structures will be described by SMILES, Inchi and Mol files.
c)	Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
	As far as possible, standard vocabularies and machine-readable file formats (i.e. csv, SQL dump) are used when storing research data.
d)	In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
	Uploaders can use custom values and ECBD staff will try to link it to existing ontology, modify it internally or expand it externally with its authors.
2.4 Increase data re-use (through clarifying licences)	
a)	How will the data be licensed to permit the widest re-use possible?
	Data will be accessible under Creative Commons Attribution-ShareAlike 4.0 International license (https://creativecommons.org/licenses/by-sa/4.0/)
b)	When will the data be made available for re-use?

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement

	Assay data will be available as soon as the grace period expires at most 3 years after validation and data upload to ECBD, however uploader can reduce the grace period time. The grace period will be applied on data which needs or are requested to ensure the protection of intellectual property rights. Other data (organizational and chemical) will be without restrictions.
c)	Are the data produced and / or used in the project useable by third parties, in particular after the end of the project?
	Produced data will be useable by third parties. The data will be stored and usable independently of the project, until the ECBD is operated.
d)	How long is it intended that the data remains re-usable?
	ECBD server will be operated by IMG at least 2 years after the end of the project.
e)	Are data quality assurance processes described?
	The data will be checked automatically during the upload process and manually by ECBD staff and the uploader.

3. Allocation of resources	
a)	What are the costs for making data FAIR in your project?
	The implementation of FAIR principles is covered by the ECBD development contract, therefore minimal additional costs are anticipated for this project.
b)	How will these be covered?
	The costs will be covered by the EU-OPESNCREEN DRIVE project.
c)	Who will be responsible for data management in your project?
	ECBD is operated by IMG, the database partner site of the EU-OPENSCREEN ERIC. IMG is responsible for data management till April 2024. The period could be prolonged for an additional 5-year period or the responsibility can be transferred to another EU-OPENSCREEN partner site after that period. The correctness of uploaded data is responsibility of the uploader. To ensure further correctness, data are uploaded to intermediate store where are validated by uploader and ECBD staff.
d)	Are the resources for long term preservation discussed?
	During the project, bioactive data will be additionally uploaded to ChEMBL database therefore data should be preserved for long term with minimal costs.

4. Data security	
a)	Is the data safely stored in certified repositories for long term preservation and curation?

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement

	During the ECBD contract with IMG, data will be stored at CESNET servers with geographical backup. CESNET operates and develops the national e-infrastructure for science, research and education. CESNET is a contractual partner of IMG in ECBD operation. Sensitive data won't be publicly accessible and will be transferred only through private secured network. Long term preservation will be also ensured with upload to ChEMBL database and potentially to PubChem.
b)	What provisions are in place for data security?
	Geographical backup will serve to provide data security. Authentication, authorization, and access through HTTPS enable security of sensitive data.

5. Ethical aspects	
a)	Are there any ethical or legal issues that can have an impact on data sharing?
	Research data generated by the partner sites should not raise any ethical concerns. The user login information will not be publicly available.
b)	Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?
	Personal data are processed in accordance with the Regulation (EU) 2016/679 of the European Parliament and of the Council and Directive 95/46/EC.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement