D6.1
Report on existing data transfer procedures and gap analysis

Version 1.0 2019-06-25

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	D6.1: Report on existing data transfer procedures and gap analysis
Partner responsible	EU-OS
Author(s)	Ctibor Skuta, Milan Vorsilak, Sarka Simova, Katholiki Skopelitou

Introduction

Task 6.1 Implement IT solutions for efficient data transfer (M1-M24) (Lead: FVB-FMP, contributors: IMG, IME, MEDI, IMIM, CSC, IME) This task aims at improving data transfer procedures from EU-OS partner sites into the ECBD as well as into global data resources such as ChEMBL and PubChem. This includes the evaluation of existing procedures and the design and implementation of adapted data management tools and processes such as an improved deposition interface. Such an interface will allow submission of data and associated information in an interoperable format. Importantly, the data format applied for deposition should be flexible enough to accommodate extensions and customizations required to handle diverse types of biological information (e.g. numbers, structures, images). User support and help desk functions will be established, including online documentation illustrating user workflows and FAIR compliant data deposition guidelines for screening centres. The helpdesk will also have an important facilitating role in encouraging sites to share best practices and work towards common standards (D6.1). This task will be coordinated by the ERIC Central Office in collaboration with three EU-OS database-expert centres and three large representative screenings centres, as well consulting with the other screening centres of EU-OS. An ECBD curator will offer on-site visits to all screening centres to assist them in implementing the developed deposition interface and commonly agreed standards and procedures. In collaboration with the other biomedical ESFRIs, this WP will also contribute towards the aim to make high value primary data sets data publically available to facilitate machine learning and artificial intelligence paradigms. This effort will be coordinated with the planned European Open Science Cloud (EOSC) project EOSC-LIFE, where the partners IME, IMG, CSC and IMIM are partners in a thematically aligned demonstrator project. Task 6.1 meets objectives O6.1, O6.2, O6.3 and O6.4.

This report is focused on describing the existing procedures of data standardization and data management at different partner sites. The implemented survey followed by the gap analysis should recognize, characterize and distinguish these procedures. These tools will help to establish and share the best practices and work towards common standards.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

1.1 Survey

In order to identify and evaluate existing procedures at partner sites the survey “Data transfer procedures” was prepared and distributed to the DRIVE participants as a web-survey. The survey consisted of 13 questions (listed in 1.2.1) and it should map the contemporary procedures in data transfer.

1.1.1 Questions

*required fields

1. Please indicate your partner number and partner acronym in DRIVE.

Click or tap here to enter text.

2. Do you have a programmer/cheminformatician/bioinformatician in your lab*?

☐ yes

☐ no

Comments:

Click or tap here to enter text.

3. Do you have a person responsible for data management in your lab*?

☐ yes

☐ no

Comments:

Click or tap here to enter text.

4. What technology do you use for data storage*? (e.g., database, cloud, excel files, text files, etc.) – multichoice possible

☐ database

☐ cloud

☐ excel files

☐ text files

☐ Other:

Click or tap here to enter text.

5. Do you use any Laboratory Information Management System (LIMS)*?

☐ yes

☐ no

Comments:

Click or tap here to enter text.

6. What software do you use to process/analyze your data*? (LIMS - which?, Excel, Knime, Pipeline Pilot, etc.) – multichoice possible

☐ LIMS

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

If LIMS, which one?
Click or tap here to enter text.
☐ Excel
☐ Knime
☐ Pipeline Pilot
☐ Other:
Click or tap here to enter text.

7. How do you usually standardize assay data*? (e.g., Z-score, B-score, Efficacy, etc.)
Click or tap here to enter text.
Click or tap here to enter text.

8. Do you check for errors that can occur during the experiment*? (e.g., sample not transferred, contaminated pipette, etc.)
Click or tap here to enter text.
Click or tap here to enter text.

9. What quality standards for an assay development do you use*?
Click or tap here to enter text.
Click or tap here to enter text.

10. Do you plan to upload supporting data (optional metadata)*?
☐ yes
☐ no
Comment:
Click or tap here to enter text.

11. How do you backup/archive your data*?
Click or tap here to enter text.
Click or tap here to enter text.

12. How many assays do you perform per year on average*?
Click or tap here to enter text.

13. Do you plan to provide some of your in-house compounds to the collection and how many*?
☐ yes
☐ no
If yes, how many?
Click or tap here to enter text.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

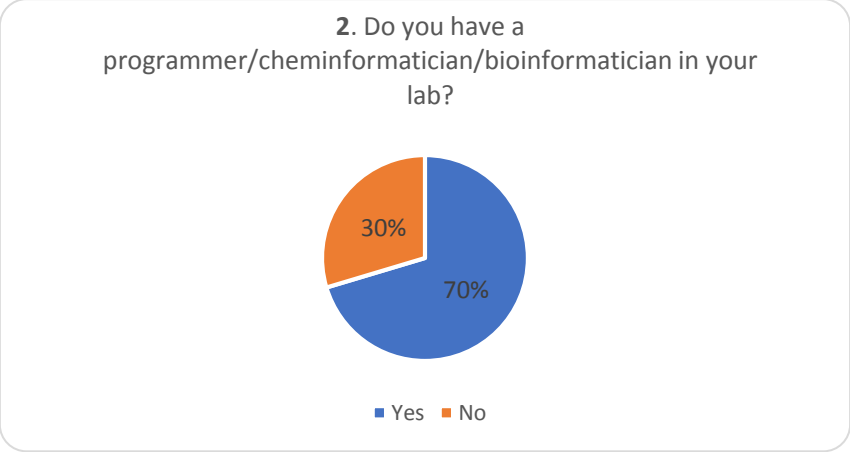
1.1.2 Survey results

1. Please indicate your partner site number and partner site acronym in DRIVE.

The total number of DRIVE participants participating (partner sites) in the survey was **27**. Of these **16** were EU-OPENSCREEN partner sites (EU-OS partner sites). Following graphs represent the summarized results.

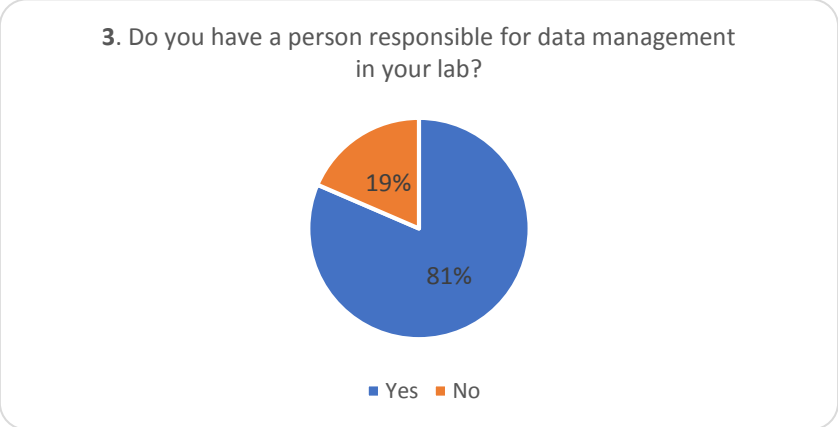
2. Do you have a programmer/cheminformatician/bioinformatician in your lab*?

Most of the partner sites (19/27) has a programmer/cheminformatician/bioinformatician in their lab.



3. Do you have a person responsible for data management in your lab*?

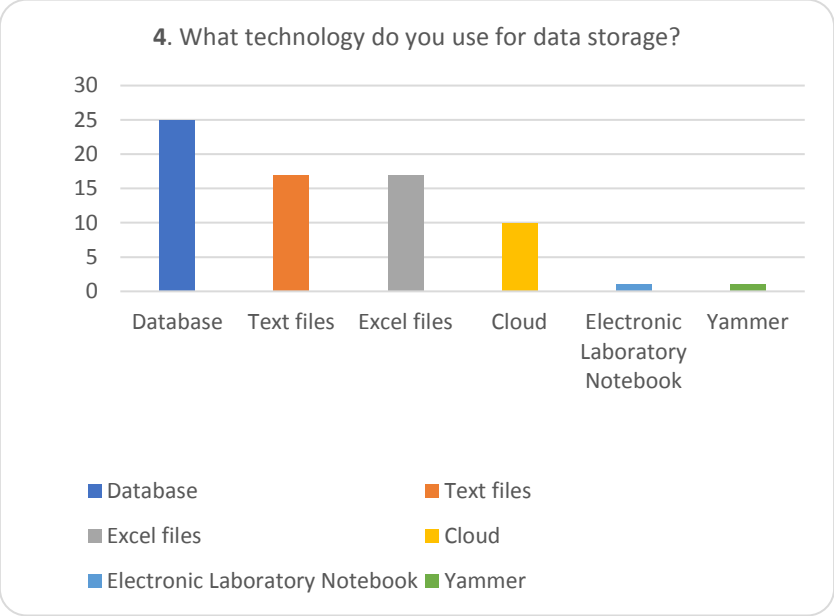
More than 80% (22/27) of the partner sites have a person responsible for data management in their lab.



This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

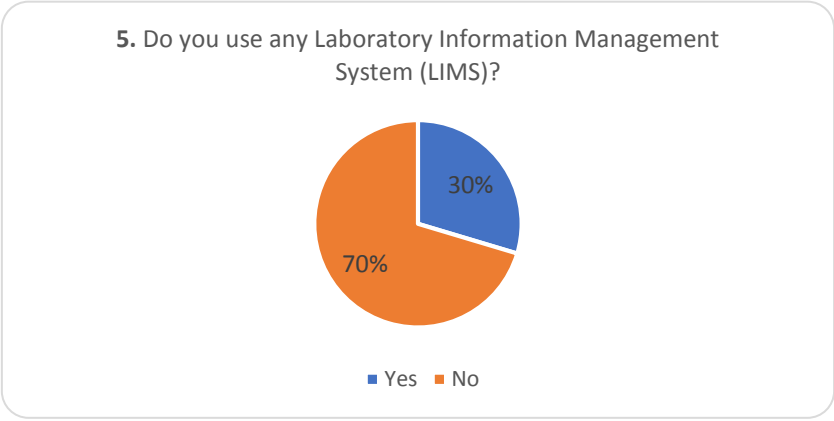
4. What technology do you use for data storage*? (e.g., database, cloud, excel files, text files, etc.)

Most of the partner sites (25/27) store their data in a database, 17/27 also uses Excel or text files, and 10/27 store their data in the cloud. One partner site uses Electronic Laboratory Notebook and one also Yammer. Most of the partner sites (22/27) uses more than one of these approaches to store their data.



5. Do you use any Laboratory Information Management System (LIMS)*?

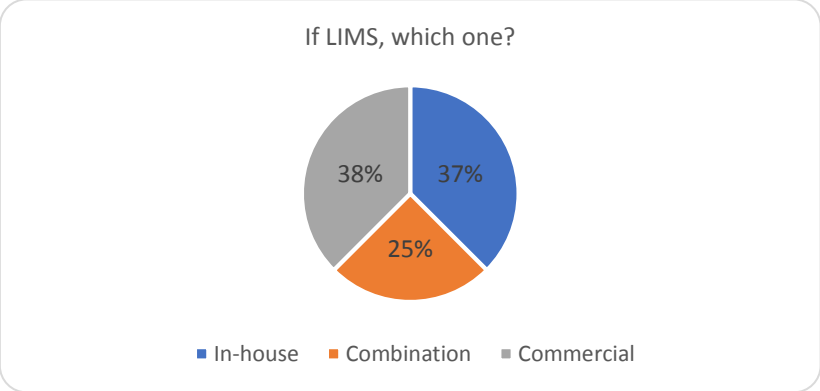
Only one third (9/27) of the partner sites use Laboratory information management system (LIMS) in their lab.
Three of the partner sites develop their own LIMS, three use a commercial one (Grit42, CRIMS, LABWARE) and two use a combination of commercial and in-house software (e.g., GeneData, CDD, etc.).



This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

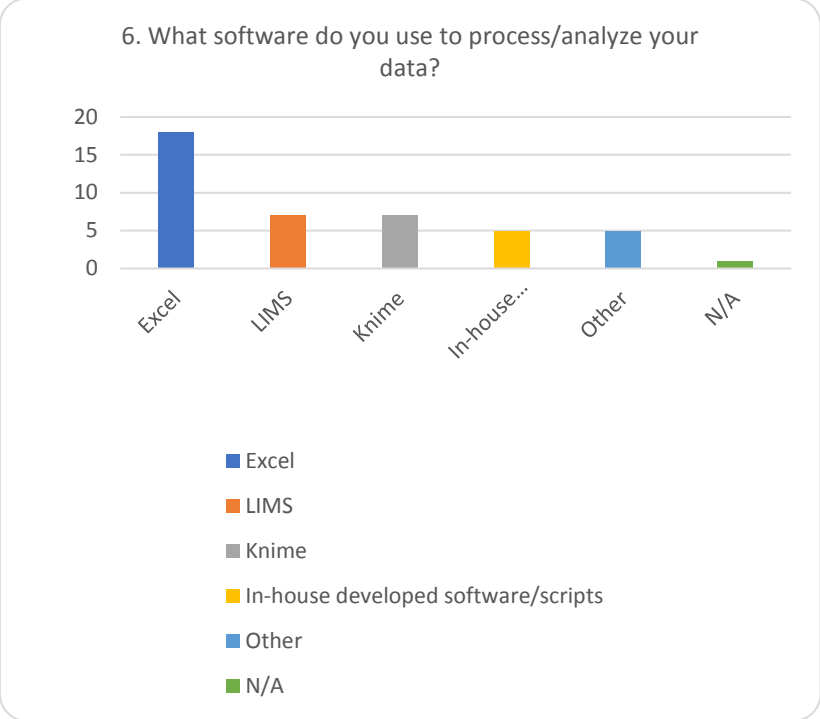
If you use LIMS, please indicate which one.

Three of the partner sites develop their own LIMS, three use a commercial one (Grit42, CRIMS, LABWARE) and two use a combination of commercial and in-house software (e.g., GeneData, CDD, etc.).



6. What software do you use to process/analyze your data*? (if LIMS – which one?, Excel, Knime, Pipeline Pilot, etc.)

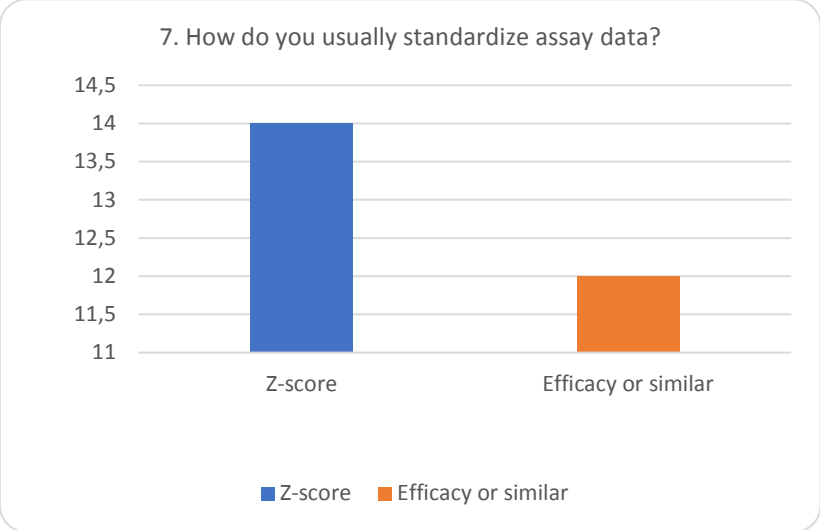
Almost 70% (18/27) of the partner sites use Excel as one of the options to analyze their data, 7/27 perform the analysis in their LIMS or in the KNIME Analytics platform, 5 use their in-house developed software/scripts. Other software, such as Pipeline Pilot, Graphpad Prism, Breeze, TDB or Scilslab is also employed by at least one partner site.



This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

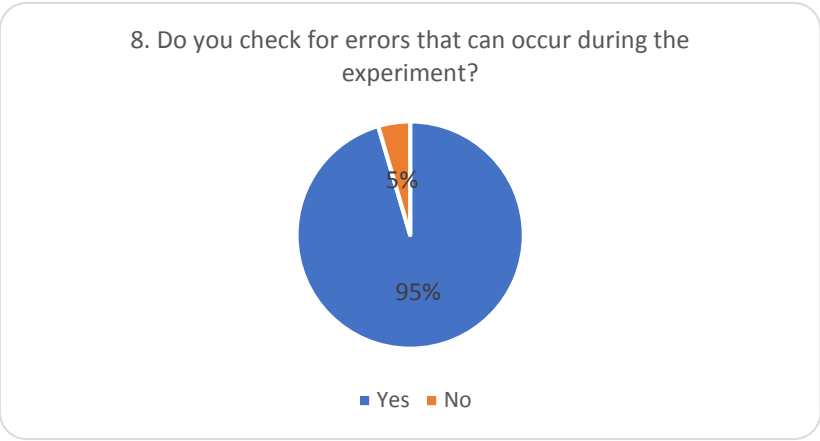
7. How do you usually standardize assay data*? (e.g., Z-score, B-score, Efficacy, etc.)

The most commonly employed standardization procedures are Z-Score (14/27) and Efficacy (12/27) or generally other methods calculating value relative to top/bottom control values (reference compounds).



8. Do you check for errors that can occur during the experiment*? (e.g., sample not transferred, contaminated pipette, etc.)

More than 80% (21/27) of the partner sites check for errors that can occur during an experiment with only 1/27 partner site indicating it doesn't perform any check for errors. For the rest (4/27), this question is irrelevant. The most common procedures are 1) the analysis of logs, 2) detection of row/column and edge effects, and 3) visual inspection of results.



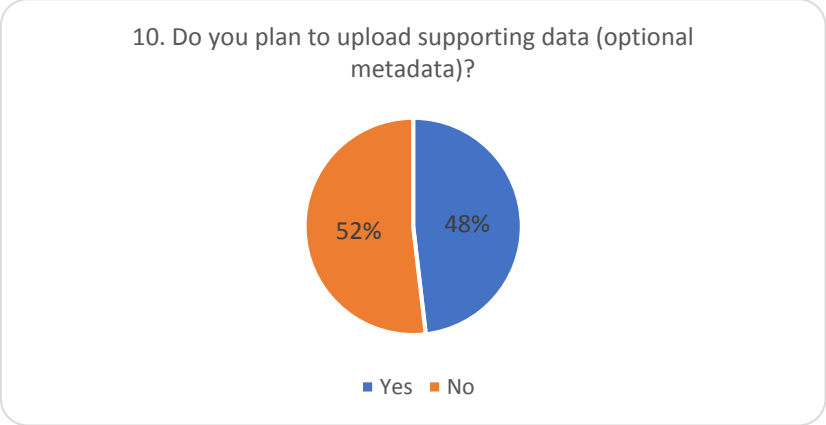
This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

9. What quality standards for an assay development do you use*?

All of the screening sites (19/27) employ some kind of quality standards during the assay development. The most common standards are Z`factor and reproducibility in time (e.g., day-to-day reproducibility)

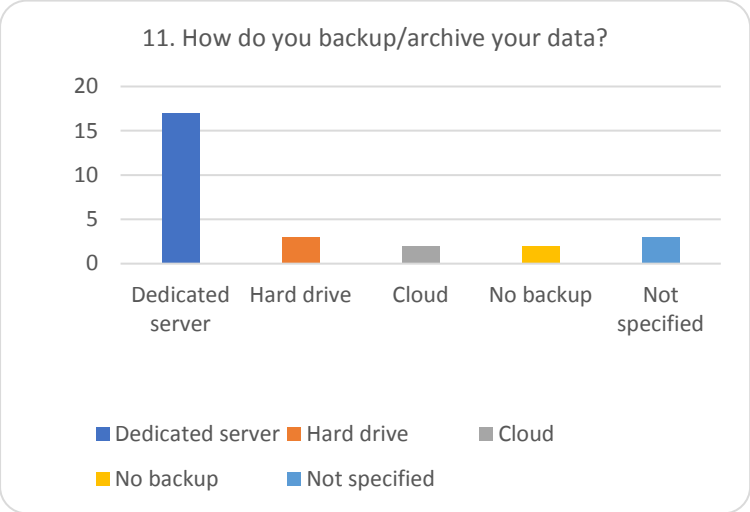
10. Do you plan to upload supporting data (optional metadata)*?

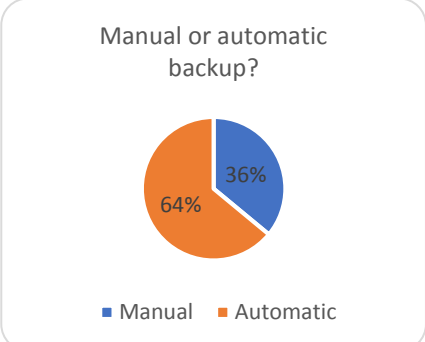
Almost 50% (13/27) of the partner sites is planning to upload supporting metadata along with the experimental data.



11. How do you backup/archive your data*?

Most of the partner sites (19/27) backup their data on a dedicated server or in a cloud. Only 3/27 backup their data on hard-drives and 2/27 use no backup at all (3/27 did not specify their backup procedure). In 16/27 cases the backup is performed automatically and in 9/27 manually/not specified.

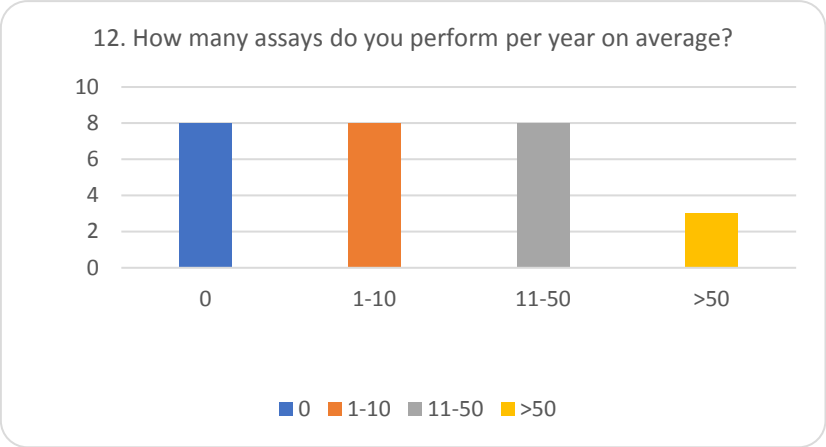




This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

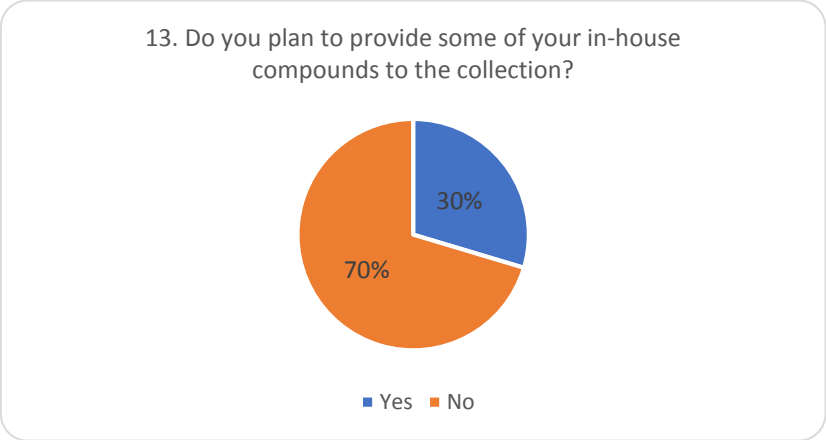
12. How many assays do you perform per year on average*?

The screening sites (19/27) perform on average 73 assays/experiments per year with a median equal to 14 assays (8/27 between 1 and 10, 8/27 between 11 and 50 and 3/27 more than 50 assays per year).



13. Do you plan to provide some of your in-house compounds to the collection and how many*?

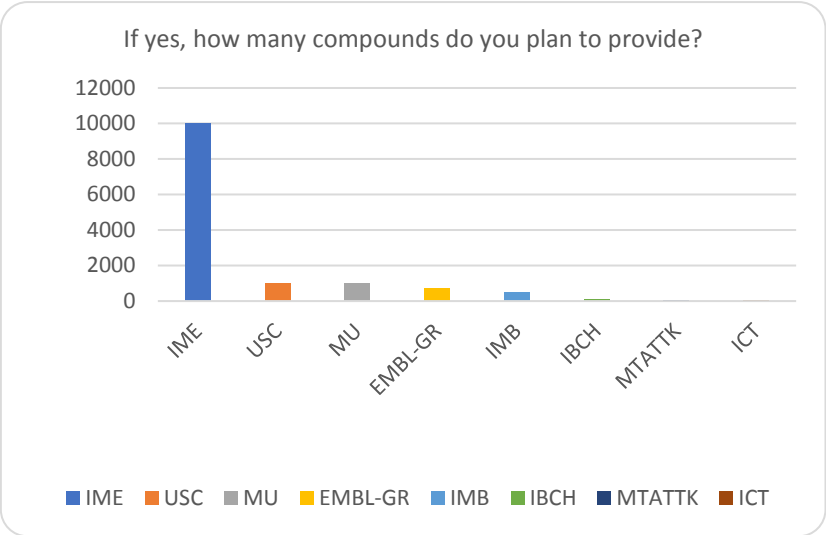
Almost one third (8/27) of the partner sites is planning to donate some of their in-house compounds to the EU-OPENSCREEN compound collection.



If yes, how many compounds do you plan to provide?

The partner sites will provide from 50 to 10 000 compounds.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893



1.2 Gap analysis

In general a gap analysis is used to identify the discrepancies between the current state and the desired state of particular existing procedures in the organization. We applied the gap analysis on the results of the survey collecting relevant data of data transfer procedures (data standardization and data management). Thanks to the involvement of the majority of the DRIVE project participants (27 out of 33) in the conducted survey, the collected data represents the valuable view of the existing procedures in the participating organizations. The subsequent gap analysis will be beneficial to improve the current state of the analyzed procedures and achieve the desired state in the near future.

The following table "**Gap analysis**" lists for each of the survey question:

- "current state" column summarizing the current processes
- "desired state" column explaining the reasons for the improvement and advantages of it
- "actions required" column suggesting the steps and solutions that can be implemented to achieve the desired states of the procedures
- "priority" column prioritizing the steps based on their importance

IMG as a WP6 leader and the ECBD host will support the other DRIVE participants to implement the suggested action to improve the current state.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

GAP analysis

QUESTION	QUESTION	CURRENT STATE	DESIRED STATE	ACTION/S REQUIRED	PRIORITY
1	**Please indicate your partner number and partner acronym in DRIVE.**	The total number of DRIVE participants participating (partner sites) in the survey was 27. Of these 16 were EU-OPENSCREEN partner sites (EU-OS partner sites).			
2	**Do you have a programmer/cheminformatician/bioinformatician in your lab*?**	Most of the partner sites (19/27) has a programmer/cheminformatician/bioinformatician in their lab.	From a general point of view, it is advantageous to have at least one programmer in the lab with an intermediate or higher skill to write software/scripts in one of the common programming languages (e.g., Python, Java, R, C, etc.). Such a person is less dependent on third-party software and is able to write custom functions to process, analyze or format data. Only data in a pre-defined custom format will be accepted for upload to ECBD.	The ECBD site will prepare a data template in a tabular form that can be used to upload data to ECBD. Along with the template, there will be a detailed description of the format and an example data. In a case that any of the partner sites is unable to prepare the data in the pre-defined format, the ECBD site will provide a support and/or tools that will help them to process and prepare the data.	*High*
3	**Do you have a person responsible for data management in your lab*?**	More than 80% (22/27) of the partner sites have a person responsible for data management in their lab.	From a general point of view, it is advantageous to have a person/s responsible for data management in a lab. It lowers the possibility of a data-loss, enhances consistent data formatting and simplifies data-related procedures not only inside the lab but also between the lab and its partners/customers.	The ECBD site will ask each partner site to assign a person dedicated to data management in their lab. This person will be responsible for the upload of the data to ECBD as well as their preparation. He/She will be also the contact person for data-related questions from the ECBD team and ECBD users.	*High*
4	**What technology do you use for data storage*? (e.g., database, cloud, excel files, text files, etc.) – multichoice possible**	Most of the partner sites (25/27) store their data in a database, 17/27 also uses Excel or text files, and 10/27 store their data in the cloud. One partner site uses Electronic Laboratory Notebook and one also Yammer. Most of the partner sites (22/27) uses more than one of these approaches to store their data.	Store data in a database with a backup on a separate machine (server, disk, cloud). The database storage ensures that the data are stored in the same format and with the same internal IDs which enhances the possibility to analyze/compare data from different experiments/projects. The backup on a separate machine lowers the risk of a data loss due to a hardware failure or human mistake.	All data uploaded to ECBD will be stored in the database using underlying ontologies for their description (e.g., BioAssay Ontology, Gene Ontology, BRENDA tissue ontology, NCBI taxonomy, etc.). This approach will simplify comparison and analysis of data from different partner sites, and also their linking to external data sources.	*High*
5	**Do you use any Laboratory Information Management System (LIMS)*?**	Only one third (9/27) of the partner sites use Laboratory information management system (LIMS) in their lab.	-	-	-
	If LIMS, which one?	Three of the partner sites develop their own LIMS, three use a commercial one (Grit42, CRIMS, LABWARE) and two use a combination of commercial and in-house software (e.g., GeneData, CDD, etc.).	-	-	-
6	**What software do you use to process/analyze your data*? (LIMS - which?, Excel, Knime, Pipeline Pilot, etc.) – multichoice possible**	Almost 70% (18/27) of the partner sites use Excel as one of the options to analyze their data, 7/27 perform the analysis in their LIMS or in the KNIME Analytics platform, 5 use their in-house developed software/scripts. Other software, such as Pipeline Pilot, Graphpad Prism, Breeze, TDB or Scilslab is also employed by at least one partner site.	-	-	-
7	**How do you usually standardize assay data*? (e.g., Z-score, B-score, Efficacy, etc.)**	The most commonly employed standardization procedures are Z-Score (14/27) and Efficacy (12/27) or generally other methods calculating value relative to top/bottom control values (reference compounds).	Using the same standardization methods enhances the possibilities to compare data from different assays/experiments and ensures the consistency of used methods.	All data uploaded to ECBD will be processed using the same standardization procedures and software tools in order to ensure consistency and to enable analyzing data coming from different partner sites.	*High*
8	**Do you check for errors that can occur during the experiment*? (e.g., sample not transferred, contaminated pipette, etc.)**	More than 80% (21/27) of the partner sites check for errors that can occur during an experiment with only 1/27 partner site indicating it doesn't perform any check for errors. For the rest (4/27), this question is irrelevant. The most common procedures are 1) the analysis of logs, 2) detection of row/column and edge effects, and 3) visual inspection of results.	Checking for errors is a necessary step in order to detect and prevent hardware malfunction and bottlenecks in the instrumentation. As a consequence, this procedure can detect and prevent the propagation of wrong data into both in-house as well as public data repositories, and also prevent the propagation of false-positive hits from primary assays to later stages of hit-to-lead campaigns.	To track errors in measurement occurring during the experiment, ECBD will enable to tag errors in the experimental data during their upload. Based on these data (if available), ECBD will track the most common errors and will try to make the partner sites/community aware of them.	*Medium*
9	**What quality standards for an assay development do you use*?**	All of the screening sites (19/27) employ some kind of quality standards during the assay development. The most common standards are Z`factor and reproducibility in time (e.g., day-to-day reproducibility).	Checking for quality standards is a necessary step in order to evaluate the ability of an assay to distinguish between active and inactive compounds and also to ensure the reproducibility of the data in the future.	All of the screening sites employ some kind of quality standards during the assay development. No action required.	*High*
10	**Do you plan to upload supporting data (optional metadata)*?**	Almost 50% (13/27) of the partner sites is planning to upload supporting metadata along with the experimental data.	Relevant supporting metadata can play a key role in the data analysis of a single assay as well as finding patterns in large data repositories. In many cases, metadata are as important as the data themselves.	ECBD will enable its users to simply upload any kind of additional metadata accompanying experimental data. ECBD users will be also able to use additional attributes to describe their data (assays) either through underlying ontologies or custom attibutes the user deems to be important.	*Medium*
11	**How do you backup/archive your data*?**	Most of the partner sites (19/27) backup their data on a dedicated server or in a cloud. Only 3/27 backup their data on hard-drives and 2/27 use no backup at all (3/27 did not specify their backup procedure). In 16/27 cases the backup is performed automatically and in 9/27 manually/not specified.	Automatic and periodic (e.g., daily) backup on a separate machine (ideally on a different electric circuit) or cloud is the best option how to prevent data loss due to a hardware failure or a human mistake.	All data uploaded to ECBD will be protected and archived through daily and weekly scheduled backups. There will also be a copy of each backup stored on a second independent site.	*High*
12	**How many assays do you perform per year on average*?**	The screening sites (19/27) perform on average 73 assays/experiments per year with a median equal to 14 assays (8/27 between 1 and 10, 8/27 between 11 and 50 and 3/27 more than 50 assays per year).	-	-	-
13	**Do you plan to provide some of your in-house compounds to the collection?**	Almost one third (8/27) of the partner sites is planning to donate some of their in-house compounds to the EU-OPENSCREEN compound collection.	To create a large academic compound collection as possible with the focus on diversity and novelty of donated compounds.	To argument the advantages of donating compounds in the academic collection (e.g., qaulity control, bioprofiling, etc.).	*Low*
	If yes, how many compounds do you plan to provide?	The partner sites will provide from 50 to 10 000 compounds.	-	-	-