OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Deliverable Report D3.6

Final data management, maintenance and sustainability plan

This project is funded by

the European Union

OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and in silico Analysis and Modelling in Risk Assessment

Project Number 731075

www.openrisknet.org

Project identification

Grant Agreement	731075
Project Name	OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and *in silico* Analysis and Modelling in Risk Assessment
Project Acronym	OpenRiskNet
Project Coordinator	Edelweiss Connect GmbH
Start date	1 December 2016
End date	30 November 2019
Duration	36 Months
Project Partners	P1 Edelweiss Connect GmbH Switzerland (DC) P2 Johannes Gutenberg-Universität Mainz, Germany (JGU) P3 Fundacio Centre De Regulacio Genomica, Spain (CRG) P4 Universiteit Maastricht, Netherlands (UM) P5 The University Of Birmingham, United Kingdom (UoB) P6 National Technical University Of Athens, Greece (NTUA) P7 Fraunhofer Gesellschaft Zur Foerderung Der Angewandten Forschung E.V., Germany (Fraunhofer) P8 Uppsala Universitet, Sweden (UU) P9 Medizinische Universität Innsbruck, Austria (MUI) P10 Informatics Matters Limited, United Kingdom (IM) P11 Institut National De L'environnement Et Des Risques, France (INERIS) P12 Vrije Universiteit Amsterdam, Netherlands (VU)

Deliverable Report identification

Document ID and title	Deliverable 3.6 Final data management, maintenance and sustainability plan
Deliverable Type	Report
Dissemination Level	Public (PU)
Work Package	WP3
Task(s)	Task 3.5 Dissemination, Exploitation, Data Management and Sustainability Plan
Deliverable lead partner	EwC
Author(s)	Thomas Exner, Daniel Bachler, Lucian Farcal, Oana Florean Tomaz Mohoric (EwC), Philip Doganis (NTUA), Egon Willighagen, Marvin Martens, Danyel Jennen (UM), Marc Jacobs (Fraunhofer) and Associated Partners
Status	Final
Version	V1.0
Document history	2019-11-11 Draft version 2019-11-29 Final version

Table of Contents

SUMMARY

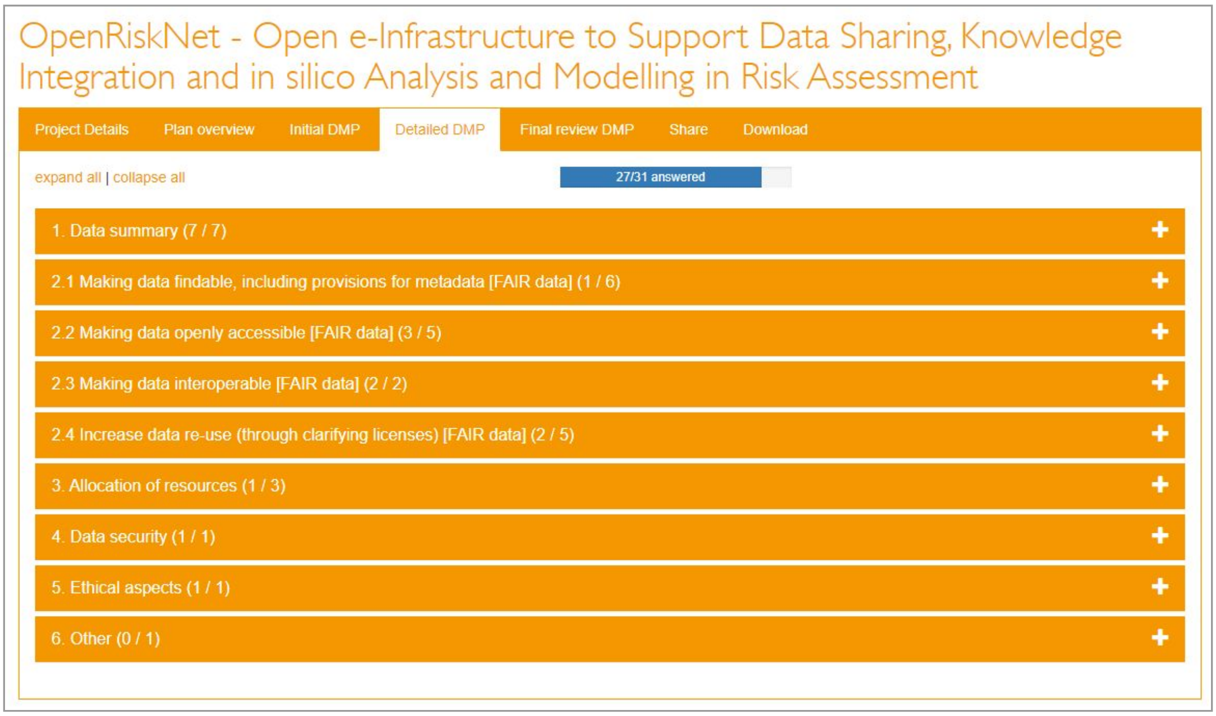
This report is based on the initial data management plan (DMP) (Deliverable 3.1) [1] and includes the final DMP for the OpenRiskNet e-infrastructure project. The current document covers the aspects of the OpenRiskNet data management based on the FAIR (findable, accessible, interoperable and reusable) guidelines, ethics considerations for re-sharing of public datasets, and details on the data sources made available and will be sustained after the project as OpenRiskNet data sources. These are alternative access to US EPA ToxCast/Tox21 data, intensities and fold changes obtained by processing the TG-GATEs and DrugMatrix transcriptomics data, BridgeDb, semantic annotated versions of WikiPathways, AOP-Wiki and AOP-DB in RDF format, ToxicoDB including the dataset for the TGX case study, the nano Daphnia dataset, ToxPlanet, SCAIview and as a specific feature of OpenRiskNet also the library of risk assessment workflows in form of Jupyter notebooks. Sustainability of OpenRiskNet developments and achievements other than data will be covered in the separate sustainability plan being part of the second Periodic Report.

INTRODUCTION

The European Commission is enforcing the openness of data produced in all projects under Horizon 2020 based on the Open Research Data Pilot (ORD Pilot). The ORD pilot aims to improve and maximise access to and re-use of research data and takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions [2]. Open data is data that is free to access, re-use, repurpose, and redistribute. The Open Research Data Pilot aims to make the research data generated by selected Horizon 2020 projects accessible with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access. Projects starting from January 2017 are by default part of the Open Data Pilot, including the Research infrastructures (including e-Infrastructures) are required to participate in the ORD Pilot. Since one of the main aims of OpenRiskNet is to allow for a simpler, more harmonised access to public, open data sources and workflows and enrich these with semantic annotation to improve their interoperability between each other and with predictive toxicology and risk assessment software, OpenRiskNet is fully supporting the ORD Pilot, is developing best-practice approach and trying to act as a role model for data management and sharing [2].

To help optimising the potential for future sharing and re-use of data, the OpenRiskNet Data Management Plan (DMP) helped the partners to consider any problems or challenges that may be encountered and helped them to identify ways to overcome these. Throughout the project, the DMP was handled as a “living” document and is available here in its final version as to November 2019. It outlines how the research data collected or generated, including redistribution of existing data sources as well as results from the *in silico* investigations performed as part of the case studies, were handled by the OpenRiskNet consortium and associated partners. It follows the Guidelines on FAIR Data Management in Horizon 2020 [2] and OpenAIRE guidelines [3], and is based around the resources available to the project partners in a realistic way taking the current knowledge into account. The activities of all OpenRiskNet partners to keep the DMP up to date followed an online, distributed approach as outlined in the above mentioned guidelines for creating an online DMP (see Figure 1).

In this report the general concepts for the description of datasets, as well as data sharing and archiving approaches adopted in the DMP are described first. This is then followed by the actual DMP in its final version as by the time of this writing. It covers the general aspects of the OpenRiskNet data management but also specific and clearly-defined measures for the data sources integrated by the project into the OpenRiskNet infrastructure specifically also including the library of toxicology and risk assessment workflows.



OpenRiskNet - Open e-Infrastructure to Support Data Sharing, Knowledge Integration and in silico Analysis and Modelling in Risk Assessment

27/31 answered

1. Data summary (7 / 7)

2.1 Making data findable, including provisions for metadata [FAIR data] (1 / 6)

2.2 Making data openly accessible [FAIR data] (3 / 5)

2.3 Making data interoperable [FAIR data] (2 / 2)

2.4 Increase data re-use (through clarifying licenses) [FAIR data] (2 / 5)

3. Allocation of resources (1 / 3)

4. Data security (1 / 1)

5. Ethical aspects (1 / 1)

6. Other (0 / 1)

Figure 1. DMPonline[1] tool used to guide the creation of the OpenRiskNet plan (https://dmponline.dcc.ac.uk/plans/16261)

DATA SET DESCRIPTION

This section give the general concepts of what is considered data in OpenRiskNet and a listing of what kind of data and information the project collected, made available for redistribution and sharing or generated as part of the case studies (*in silico* only), and to whom they might be useful later. More information and specific details are given in the DMP below.

Data refers to:

- Data generated in *in vivo*, *in vitro*, *in chemico* and *in silico* experiments, the first three coming from third-parties including other projects and associated partners, broadly related to toxicology and risk assessment in the form of raw, processed and summary data as well as metadata to describe the type of data and how it was produced (protocols and method descriptions for the experimental, processing and analysis procedures including the computational workflows);
- More specifically, data and metadata, which form part of an OpenRiskNet service or were needed to execute the case studies and validate results in scientific publications, and
- Other curated and/or raw data and metadata that may be required for validation purposes or with reuse value.

The metadata provided with the datasets allows to answer questions to enable data to be found and understood, according to the particular standards applied. Such questions include but are not limited to:

[1] https://dmponline.dcc.ac.uk/ DMPonline tool provided by the Digital Curation Centre (DCC) helps to create, review, and share data management plans that meet funder requirements

- What is the data about?
- Who created it and why?
- In what forms is it available?
- Which standards were applied.

Finally, the **metadata**, **documentation** and **standards** were selected and are now provided to make the data FAIR (Findable, Accessible, Interoperable and Re-usable) as well as to guide future project. They are not only satisfying the technical requirements like global, persistent identifier, clear access protocols (data application programming interfaces (APIs)) and licenses but are an attempt for harmonizing and improving the scientific interoperability of the data by semantic annotations and allowing combination and enrichment of datasets using linked-data approaches (Combined OpenAPI and JSON-LD description of data APIs).

Data integrated can be grouped into the following areas:

- Existing toxicology, chemical properties and bioassay databases for redistribution;
- Existing omics databases for redistribution;
- Existing knowledge bases for redistribution and information extracted by data mining;
- Document libraries;
- Computational workflows for redistribution and adaptation to specific scientific questions; and
- Intermediate or final results of *in silico* studies performed as part of the case studies as part of the Jupyter notebooks.

DATA SHARING

According to the ORD Pilot programme, the resulting data should be archived as Open Access by default as much as possible. Most data handled in OpenRiskNet is provided by international publicly funded projects, not-for-profit consortia or governmental and regulatory agencies. This data sources are already available under open-data licenses and will be redistributed by OpenRiskNet in a restructured and enriched form under the same license. Newly generated data are results from the improved processing, analysis and modelling workflows developed as examples in the case studies, including the workflows themselves, and OpenRiskNet makes these publicly available either as part of the publicly shared workflows or, if they have value outside the case studies, as a separate information and knowledge source. Working together with associated partners especially from the commercial sector (service providers and end users from SMEs and larger industry) has put some restrictions on the sharing of data generated in these collaborations. The legitimate reasons for not sharing resulting data is explained in the DMP in the one case they had to be applied (ToxPlanet chemical hazard and toxicology literature repository). OpenRiskNet was fully committed to protect personal data and IPR agreements and to responsible data sharing and has taken all steps reasonably necessary to ensure that data is treated securely and in accordance with the OpenRiskNet privacy policy (see section below on the Privacy Policy). No personal data was transferred to an organization or a country unless there were adequate controls in place including the security of data and personal information. Complementing these general data sharing policies, the DMP below describes any ethical or legal issue that has an impact on data sharing of specific data sources. Since in many cases, the data production was not under the control of the OpenRiskNet partners and the data was only redistributed by them, the obligation to guarantee that the data is generated from high quality, ethical research and can be shared under an open license is in the hand of the original data provider or the primary data distributor. This includes the obligation to operate in conformity with the requirements of their own institution, and fulfil all necessary national and international regulatory and ethical requirements. OpenRiskNet concortium members were working together and will continue to do so with the original data providers as well as ethical experts on producing workflows and a checklist (see attachments for a draft version) for the ethics evaluation and on documenting the measures adopted during the data generation process (ethical approval of the *in vivo* and *in vitro* experiments by the relevant authorities) as well as for protection personal data e.g. by anonymization of data before sharing. Whenever provided by the data provider and when technical feasible, this licensing, legal and ethics information was made available to the OpenRiskNet user as part of the data service description.

ARCHIVING AND PRESERVATION

To ensure that publicly funded research outputs can have a positive impact on future research, for policy development, and for societal change, it is also important to assure the availability of data for a long period beyond the lifetime of a project. This does not refer only to storage in a research data repository, but also to consider the usability of the data. One of the main goals of the infrastructure created by the OpenRiskNet project was to harmonise data, make it interoperable and sustainable and in some cases even enable data sharing or replace existing data sharing solutions. Therefore, the project had a special obligation for preserving data not only produced in the project but also from other projects redistributed by OpenRiskNet and software or any code produced to perform specific analyses or to render the data as well as being clear about any proprietary or open source tools that will be needed to validate and use the preserved data. OpenRiskNet was building on software engineering and infrastructure components developed, supported and adopted by a large community guaranteeing, on the one hand,stability and sustainability of the data sharing, accessing and processing solutions provided even in the relatively quickly changing field of microservice architectures and deployments. On the other hand, the containerization approach adopted by OpenRiskNet allows for the storage of the data and software in the version used during the execution of the analysing and modelling workflows allowing for complete and exact repeatability using the same code and improved reproducibility due to better documentation.

It has to be noted here that many of the data sources are only redistributed by OpenRiskNet. The primary data provider for e.g. ToxicoDB, ToxCast/Tox21 and TG-GATEs are international research groups or agencies. Raw data archiving and preservation have to be guaranteed by these institutions. However, OpenRiskNet and more specifically the OpenRiskNet partner responsible for the integration into the OpenRiskNet infrastructure are in charge of maintaining and updating the alternative method to access the data (OpenRiskNet-compliant data API), guaranteeing that the data available within OpenRiskNet is on the same technical and curation level and at the same version as in the primary source, and sustaining the solution beyond the OpenRiskNet project. The same is true for data sources, where OpenRiskNet also takes the responsibility of hosting the data and thus, becomes the primary data source. In the later case, archiving and preservation of the data source containers is of uttermost importance since otherwise there is the danger that the data is lost completely. Short-term sustainability (2 years) is secured by the reference infrastructure running at Johannes-Gutenberg Universität Mainz. Negotiations with the Birmingham Environment for Academic Research (BEAR) are underway to provide mid- to long-term archiving and preservation facilities for containerised data and software services for at least the next 5 years.

DATA MANAGEMENT PLAN (DMP)

This data management plan addresses all data-related measures adopted by the OpenRiskNet project including problems or challenges that were encountered by partners during the execution of the project. It consists of general guidelines and project-internal rules and regulations dealing with the type of data collected, data sharing following the FAIR principles, hard- and software resources as well as with data security, privacy and ethics. Additionally, more details for specific data sources on all these aspects are provided whenever necessary.

1. DATA SUMMARY

Summary of the data addressing the following issues:

- *State the purpose of the data collection/generation*
- *Explain the relation to the objectives of the project*
- *Specify the types and formats of data generated/collected*
- *Specify if existing data is being re-used (if any)*
- *Specify the origin of the data*
- *State the expected size of the data (if known)*
- *Outline the data utility: to whom will it be useful*

1.1 Purpose of the data collection

The main purpose in collecting and use of data and metadata in the OpenRiskNet project was to fulfill its main objectives in providing and improving solutions on data availability to the toxicology and risk assessment scientific community, data quality, interoperability, standardization and sustainability and overcome some of the data-related issues, e.g.:

- Fragmentation of data across different databases without common ways to query the data;
- Low quality, interpretability and reusability due to insufficient data curation;
- Poor explanation and insufficient details on experimental design and protocols applied;
- Data available in different formats and with different annotations.

Another goal was to generate guidelines and templates for data exchange, semantic annotations and harmonised use of ontologies as well as develop criteria and solutions for controlling the quality of a dataset or *in silico* tool for quantifying the uncertainty of predictive models and for improving the repeatability and reproducibility of processing, analysis and modelling workflows.

1.2 Relation to the objectives of the project

The OpenRiskNet project aimed to establish an e-infrastructure and services functions providing a centralised and standardised set of data and computing resources, accompanied by standardised operating procedures and guidance:

- Provision of quality sources of data to facilitate a more accurate evaluation of toxicity;
- Data infrastructure offering a centralised repository for data created during other research programs, including the import of relevant research *in vitro*, *in vivo* and human data from other sources.
- Well-designed data import facilities to support ongoing data collection according to quality guidance.
- Use and further development of data annotation and exchange standards for describing toxicity data based on application programming interfaces in order to reduce errors and enable data integration from different laboratories, including data sources outside the program
- Integrate regulatory reporting requirements with respect to metadata and documentation details and completeness.

The OpenRiskNet project aims also to develop and optimise computational models and automated and reliable analysis workflows in order to increase the mechanistic understanding of toxicity:

- Models permitting identification of mechanistic links between omics data at different levels of functional organisation;
- Models helping to advance the understanding of the relationship between toxicity, architecture, function and risk;
- Computational sensitivity analyses components aiding in identifying most sensitive parameters relevant to toxicity and guide further data acquisition and experiments towards increased chemical safety.

The data sources integrated during the project are highly relevant to the predictive toxicology and risk assessment community and therefore, are used to showcase and evaluate the concepts and solutions provided by OpenRiskNet and how these are addressing the aims just mentioned. Additionally, they were used in the case studies to provide the example workflows on how to apply and combine the different tools for effective problem solving for the different aspects of risk assessment.

1.3 Types and formats of data

OpenRiskNet was structured around the concept of semantic-annotated application programming interfaces, which can be used to search and access data from OpenRiskNet-compliant data sources. As serialised exchange format, JSON or the semantically annotated form JSON-LD is recommended and enforced whenever possible. These formats cover mainly the metadata associated to the data and in the case of small numbers of readouts (experimental toxicology endpoints) per sample also the data. Especially for omics data or imaging techniques, these files will be accompanied by data in standard file formats to keep the compatibility and interoperability with tools developed in these areas like gene- and pathway-enrichment approaches or image recognition software, respectively. Working together with other big projects (EU-ToxRisk and NanoCommons), the amount and content of metadata were defined, which has to be provided for each experimental assay or computational investigation, and data and protocol/test method description formats have been created in the cases standards didn’t exist so far providing the means for future additions and adaptations based on flexible data schema specifications to cover scientific advances. However, only the strict usage of ontologies in this data and metadata descriptions can guarantee that the information is easily understood by the user or automatically transferred between services.

1.4 Reuse of data

Since OpenRiskNet aimed at improving the reusability of data and software tools, it was the clear major goal of the project, to provide all input data independent of the original source as well as the results from the processing, analysis and modelling workflows under an open-data license and offer it in an easy way for reuse by others. On one hand, making sharing, accessing and reusing of data easier was achieved by the data solutions provided by OpenRiskNet and the integration of the reference data sources. On the other hand, results from the *in silico* investigations are considered as equally valuable for sharing and reuse especially with the goal to improve the evaluability, repeatability and reproducibility of these computational studies. Full documentation of the workflows including intermediate results and permanent storage of the final outcomes highly annotated by metadata describing the procedures was, therefore, organized and promoted for adoption by other projects using the capacities of workflow management tools like Jupyter storing not only the computational workflows but also the produced results.

1.5 Origin of the data

As mentioned before, the main data integrated, used and provided for easy reuse by OpenRiskNet are coming from other publicly funded research and infrastructure projects or institutions and are already in the public domain or will be made public available soon. However, users might also want to access commercial data services provided by associated partners or use their in-house data as part of the infrastructure and partly share them with a selected users under a specific license. These considerations lead to three different classes grouping the origins of the data:

- Data and models owned and provided by OpenRiskNet consortium members and associated Partners as part of the project work under an open-data license;
- Open Source data and models provided under the license mentioned by the owners;
- Data from third parties including associated partners and commercial services of OpenRiskNet partner, and not yet available in existing open databases provided under the conditions specified by the data owner and included in a formal agreement.

For all these data sources, the original license of data usage has to be considered and applied (in the original or more restricted form) to the version integrated in OpenRiskNet environments. To prevent unauthorised data access even in virtual environments shared by multiple users like the reference environment, an authentication and authorisation service was integrated in OpenRiskNet infrastructure, which also handles the license management. In the same way, commercial software or free software requiring a registration is handled. In cases, where even more protection was needed, the data services continued to be operated by the data provider and only restricted but harmonized access using the OpenRiskNet authentication and authorisation service was integrated into the infrastructure. In this way, the data provider keeps complete control over the data and can shield it against attacks to obtain unauthorised access, which would be easier possible if the containerised data is deployed into virtual environments on local machines.

1.6 Expected size of the data

The idea of the OpenRiskNet infrastructure was not to combine data from different sources into one data warehouse but to access the data from its original source and use the interoperability layer added to the data services to harmonise them. In this way, no additional capacity for storage of the original, mainly raw data was needed. However, two aims of the OpenRiskNet project led to additional requirements on data storage.

1) Some of the data sources considered for integration are not yet available in open-accessible databases, cannot be accessed via application programming interfaces from these original sources or don't comply with the FAIR principles.
2) Data sources were made available in a form suitable for in-house deployment. Even if the user system administrator setting up the in-house virtual environment (VE) is responsible for providing the required resources for such deployments, the data sources have to be containerised and provided to the users for download via the OpenRiskNet service catalogue.

Sizes for all data services are given in section 1.8 below together with other more specific details. This information can be used by OpenRiskNet users to assess the needed storage space for the containers and to give guidance on the needed computational resources for the VE.

1.7 Utility of data and models

OpenRiskNet solutions make data available to its main stakeholders (researchers, risk assessors and regulators) in an easy accessible, standardised and harmonised way in order to be able to base conclusions and recommendations about the safety of a chemical, drug, cosmetic ingredient and nanomaterial on all the available evidence. The same principles are applied also to data processing, analysis and modelling tools involved in risk assessment.

The access to the data infrastructure part of OpenRiskNet by academia, industry, risk assessors and regulators has the merit of providing a wide spectrum of data, with which users can perform parts of research and development activities and to lower the barriers to real innovation resulting in new products, processes and services. Close cooperation with the regulatory agencies is key to push the regulatory acceptance of the integrated tools and workflows.

Possible beneficiaries of the data, computational models and e-infrastructure:

- Software developers in academia and industry developing advanced risk assessment approaches based on the data and provide these to risk assessment experts;
- Industry represented by chemicals, pharma, food, cosmetics or other consumer products companies which are required to use all available information and to address the '3Rs' principles and report on alternative methods used (including *in silico*);
- Regulatory agencies (e.g. ECHA, EMA, EFSA);
- SMEs as they frequently do not have in-house tools and knowledge resources for the regulatory risk assessment requirements;
- R&D community as the translation of these methods to industrial and regulatory science will result in a deeper understanding of biological response to perturbations supporting e.g. better designed and safer drugs and clinical practice;

- Consumers as OpenRiskNet infrastructure support the integration and development of apps that can be used by consumers on their mobile phones supporting everyday activities, such as obtaining knowledge on ingredients in the products they are purchasing or using.

1.8 Specific information on individual shared data sources

In this section, specific information and requirements of individual data sources provided and sustained by OpenRiskNet with respect to their purpose, origin, relationship to the project, data type and format, size and potential users are summarized. Section 2.5 below is fulfilling the same purpose for issues on FAIR data sharing. These are meant as additions or clarifications to the general descriptions relevant only for the specific dataset/database. Points completely covered by the general remarks are not repeated here and thus some of the subsections above will not appear in the following descriptions.

OpenRiskNet sustained data sources are:

- Library of OpenRiskNet computational workflows
- BridgeDb
- WikiPathways
- AOP-Wiki
- AOP-DB
- ToxCast/Tox21
- TG-GATEs
- DrugMatrix
- Nano Daphnia dataset
- ToxicoDB
- ToxPlanet
- SCAIView annotated document corpora

1.8.1 Repository of OpenRiskNet computational workflows

OpenRiskNet has created a large number of computational workflows in relation to the work performed in the case studies and to demonstrate the functionality of single OpenRiskNet tools and how they can be combined to address complex risk assessment tasks profiting from the improved harmonization and interoperability. These build an important part of the achievements of OpenRiskNet and need to be sustained as an information and training resource.

1.8.1.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

As just described, the workflows build a central piece of the OpenRiskNet activities and achievements. They not only document the performed work but are also a way to offer the workflows composed of the computational procedure, the intermediate data and results for reuse by others. These users of the resource can rerun the workflow to show repeatability and more importantly modify the workflows to specifically address their risk assessment questions.

1.8.1.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The workflows are in the specific format used to exchange Jupyter, Squonk and nextflow workflows. Since they are highly related to the service development and integration, they are stored in a way in which they can be linked to other development resources and disseminated to computational toxicologies and tool developments. GitHub was created exactly for this purpose. An OpenRiskNet organisation has been set up in GitHub covering multiple repositories designated for the development, support and dissemination activities of the project. Besides source code control, issue tracking and developer documentation, one of the repositories is dedicated for sharing and disseminating workflows mainly in the form of Jupyter or Squonk notebooks and can be accessed at https://github.com/OpenRiskNet/notebooks. All OpenRiskNet workflows have been uploaded there and are publically available under open-source licences to allow reuse, modification and any publication of derived work. Sustainability of the content is guaranteed by the GitHub environment.

1.8.1.3 Additional details to 1.6 Expected size of the data

The overall size of the workflows is well below 1GB and is covered by the standard offering of GitHub.

1.8.1.4 Additional details to 1.7 Utility of data and models

Computational workflows offer the optimal way to present functionality of OpenRiskNet services and the e-infrastructure in total. They help users get easily started by allowing to rerun the predefined workflows (since all needed data and service access routes are defined) and adapt them to their questions at hand. Examples for specific risk assessment task are already available and the coverage of the full risk assessment framework will be continuously improved by additional workflows provided by the OpenRiskNet consortium members also after the end of the project and by opening up the workflows repository (read and write access) to all users to provide and share their workflows and even let others comment on and improve them.

1.8.2 BridgeDb

The BridgeDb project was set up to provide both identifier mapping data and a general framework that provides an API to access identifier mapping data [4]. BridgeDb is used in smaller and larger projects, the latter including WikiPathways, Cytoscape and Open PHACTS [5]. It is available in various forms, including an Open API web service, Java library, Docker image, and BioConductor package. The platform supports two kinds of identifiers. The first are simple identifier-data source combinations. The second is Internationalised Resource Identifiers, for use in semantic web technologies.

1.8.2.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

Data interoperability requires identifier mappings. The mapping data is collected by the BridgeDb project and reshared in OpenRiskNet (possible because of the open licenses). Availability of identifier mappings allows simplifications of workflows. Data is part of the BridgeDb Docker services, and either preloaded (as in the current OpenRiskNet services) or loaded when the service is fired up (this approach is currently not actively used in OpenRiskNet).

1.8.2.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

BridgeDb identifier mapping databases are commonly available in two formats: Derby data files and as link sets. Both formats have been developed for different use cases.

BridgeDb identifier mapping databases are available under open licenses or CC-Zero.

Identifier mapping is essential to data set interoperability. Existing identifier mappings databases suffice for the current needs, but mapping databases are expected to be needed for other entities, like nanomaterials and AOP entities (e.g. stressors, Key Events, outcomes).

1.8.2.3 Additional details to 1.6 Expected size of the data

Identifier mappings databases are released under the data management plan of the BridgeDb project. Data is shared in different ways depending on the type of entity. Metabolics identifier mappings databases are released on Figshare, and gene-variant databases are planned to be released on Figshare or Zenodo. The gene/protein and interaction mapping databases are currently still released using a custom approach, using a download server and not actively archived yet. The sizes of these databases vary, but typically are in the order of 500MB to 1GB in size. Exception are the gene-variant databases which are much larger. All sizes are still well within the scope of what archival websites allow.

1.8.2.4 Additional details to 1.7 Utility of data and models

Identifier mapping is essential to data set interoperability since there are multiple competing identifier systems available for labeling e.g. chemical compounds, genes and pathways. Existing identifier mappings databases suffice for the current needs, but mapping databases are expected to be needed for other entities, like nanomaterials and AOP entities (e.g. Key Events, outcomes). Additionally, access to these tools from other services for e.g. cross-database searches and data curation and enrichment will be facilitated by the OpenRiskNet integration.

1.8.3 WikiPathways

WikiPathways is a molecular pathway database, established by the WikiPathways team, a collaboration between the Department of Bioinformatics of Maastricht University and the Gladstone Institute, San Francisco. Its purpose is to facilitate the contribution and maintenance of pathway information by the biology community by utilizing the open, collaborative platform of WikiPathways.

1.8.3.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The contents of WikiPathways comprise of molecular pathways, consisting of nodes that are annotated for genes, proteins, and metabolites, which can be utilised for omics data analysis through pathway analysis in PathVisio. The WikiPathways database captures the biological knowledge in biological pathway diagrams, supported by scientific literature. Because molecular pathways can describe processes in any field of biology, it is relevant for toxicological risk assessment workflows. Pathways describe the connections between biological entities and show how a disturbance by a chemical or nanomaterial could cause downstream effects.

1.8.3.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The molecular pathways in WikiPathways are developed and curated by researchers, and are based on scientific literature. Pathways are available in multiple formats, including but not limited to the original Graphical Pathway Markup Language (GPML), Resource Description Framework (RDF), gene lists (GMT format), and nanopublications. The CC-Zero license puts no restrictions on reuse.

1.8.3.3 Additional details to 1.6 Expected size of the data

The complete collection of GPML files is less than 100MB.

1.8.3.4 Additional details to 1.7 Utility of data and models

Biological pathways are used for data analysis, biological interpretation of omics data, and data integration.

1.8.4 AOP-Wiki

The AOP-Wiki is the primary repository of qualitative, mechanistic Adverse Outcome Pathway (AOP) knowledge. It was developed by the Organisation for Economic and Co-operation and Development (OECD), representing a collaboration between the European Commission DG Joint Research Centre and US Environmental Protection Agency. The AOP-Wiki is part of the AOP-Knowledge Base, which was launched by the OECD to allow everyone to build AOPs.

1.8.4.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The AOP-Wiki data comprises of mechanistic toxicological knowledge relevant for risk assessment. While most of the knowledge is present as free-text, literature-supported descriptions, essential aspects, such as biological processes, objects, cell types, and stressor chemicals that cause a disturbance, among other things are annotated with ontologies and chemical identifiers. Therefore, the AOP-Wiki serves as a knowledge base for toxicological effects related to a variety of chemicals, which summarises relevant literature.

1.8.4.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

Knowledge in the AOP-Wiki data is stored partly as free text and partly as ontology annotations and chemical identifiers. The data originates from the AOP-Wiki database, and is supported by scientific literature that is gathered and written by researchers. The contents of the AOP-Wiki are reviewed by the OECD Extended Advisory Group on Molecular Screening and Toxicogenomics (EAGMST). Nightly exports of the AOP-Wiki contents are available, but only quarterly downloads are stored and maintained permanently on the Wiki which allows citation when the information is reused. For the OpenRiskNet service, the AOP-Wiki has been transformed into the Turtle-syntax, which described the data in RDF, providing semantic annotations of the database and its contents, and includes persistent identifiers to improve interoperability with external tools and resources.

1.8.4.3 Additional details to 1.6 Expected size of the data

While the contents of the AOP-Wiki are increasing rapidly on a daily basis, the latest permanent download of the data (October 2019) does not exceed 15MB. The RDF that is exposed in the Virtuoso SPARQL endpoint on the OpenRiskNet e-Infrastructure has a size of approximately 6MB.

1.8.4.4 Additional details to 1.7 Utility of data and models

In order to perform risk assessment, one has to gather all relevant knowledge about the mechanistic effects of a compound that requires assessment. The AOP-Wiki allows for reusing mechanistic knowledge of toxicological events upon disturbance by a stressor, often a chemical. As the AOPs are developed in a way that knowledge is separated in biological events (called Key Events) and are chemical-agnostic, their major purpose is the re-usability of toxicological knowledge. Therefore, the contents of the AOP-Wiki can be relevant for each risk assessment workflow, providing mechanistic information about biological processes and linking these together.

1.8.5 AOP-DB

The AOP-DB (The Adverse Outcome Pathway Database) serves to link molecular targets identified as molecular initiating events (MIEs) and key events (KEs) in the AOP-Wiki (https://aopwiki.org) to publically available data (e.g. gene-protein, pathway, species orthology, chemical, disease), in addition to ToxCast assay information.

1.8.5.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The AOP-DB service provides information related to AOPs that extend the existing AOP-Wiki service. Currently, the AOP-DB SPARQL endpoint contains a variety of resource types, all of which are linked to genes that are present in the AOP-Wiki. It has the links of those genes with diseases, ToxCast assays, and protein-protein interactions. By the integration of these different types of information from several databases allows for convenient extensions of knowledge captured in AOPs.

1.8.5.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The AOP-DB is a SQL database that integrates many types of data of several databases and repositories. The AOP-DB RDF that is developed in the OpenRiskNet project contains only a small section of the complete AOP-DB. The data that is captured in the AOP-DB RDF originates from a variety of public data sources, including NCBI genes, DisGeNet, Comparative Toxicogenomics Database (CTD), and EPA ToxCast Data available through the EPA Chemistry Dashboard (https://comptox.epa.gov/dashboard).

1.8.5.3 Additional details to 1.6 Expected size of the data

At the time of November 2019, the data that is loaded into the Virtuoso SPARQL endpoint on the OpenRiskNet e-Infrastructure has a size of approximately 320MB.

1.8.5.4 Additional details to 1.7 Utility of data and models

Similarly to the AOP-Wiki, the AOP-DB provides knowledge related to AOPs, which can be used to inform risk assessments. However, it can also generate hypotheses by creating links between chemical interactions and adverse effects by integrating the various resources that it captures. Therefore, the AOP-DB has multiple purposes, for several types of user communities.

1.8.6 ToxCast/Tox21

The United States Environmental Protection Agency (US EPA) Toxicity forecaster (ToxCast)[2] has generated toxicity screening data on thousands of chemicals in commerce and of interest to the agency and the general public. The project also uses computational approaches to prioritise and rank chemicals for risk assessments and regulatory decision making. Toxicology in the 21st Century (Tox21) was created to continue supporting regulations by developing better toxicity assessment methods and publicly sharing of the generated data and is a federal collaboration among US EPA, NIH, including the National Center for Advancing Translational Sciences and the National Toxicology Program at the National Institute of Environmental Health Sciences, and the Food and Drug Administration (FDA). The number of tested environmental chemicals was increased to 10,000 (called the Tox21 10K library) while the number of endpoints was reduced to 200 endpoints coming from about 70 quantitative high-throughput screening assays.

Both data sources are publicly available from the EPAas a MySQL database dump with additional supporting CSV files with information on the assays and chemicals used, are widely distributed and applied in risk assessment and can be annotated to fit into the OpenRiskNet data harmonisation and integration framework.

1.8.6.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The ToxCast research project data was generated on high-throughput in vitro toxicity screens for a variety of chemicals and biological targets. One of the goals of the project was to prioritise and evaluate the potential human health risk of chemicals in a cost efficient way. Processing and analysis workflows as well as computational and predictive models are also provided to estimate the toxicity potential of the chemicals in humans. The results of these analyses are being used actively to inform the context of decision making such as endocrine disruptor screening. To even expand these screening activities, the US EPA, National Toxicology Program (NTP) headquartered at the National Institute of Environmental Health Sciences, National Center for Advancing Translational Sciences (NCATS), and the Food and Drug Administration (FDA) formed the Tox21 Consortium. Tox21 is a US federal research collaboration focused on driving the evolution of Toxicology in the 21st Century by developing methods to rapidly and efficiently evaluate the safety of commercial chemicals, pesticides, food additives/ contaminants, and medical products.[3] To date, the Tox21 Consortium has been successful generating data on pharmaceuticals and thousands of data poor chemicals, developing a better understanding of the limits and applications of *in vitro* methods, and enabling the new data generated to be incorporated into regulatory decisions.

Tox21 data is publicly available through the National Library of Medicine's PubChem[4], the EPA's CompTox Dashboard[5] also providing the ToxCast data, and NTP's Chemical Effects in Biological Systems[6]. Even if the first can be accessed via APIs, PubChem doesn't offer the rich metadata and mechanistic annotation available from US EPA. The other two sources

[2] https://www.epa.gov/chemical-research/toxicity-forecasting
[3] https://tox21.gov/wp-content/uploads/2019/02/Tox21_FactSheet_Oct2018.pdf
[4] https://pubchem.ncbi.nlm.nih.gov/
[5] https://comptox.epa.gov/dashboard
[6] https://tripod.nih.gov/tox21

are currently only accessible via web frontends and no APIs are available at the moment even if the US EPA is planning to release an API to their CompTox database in the future. To allow for the data to be integrated with other data services of the OpenRiskNet infrastructure such as ontology mapping, pathway identification and mapping, and AOP development tools that already now, OpenRiskNet transferred the MySQL database provided by the US EPA as an alternative access point for computational toxicologists into a data management solution easy to access by less experienced users and automated workflows. Users of the OpenRiskNet service are able to take advantage of the information gaps filled through the integration of these datasets and models to develop predictive toxicology and risk assessment models e.g. read-across models. Examples of such uses are created in the case studies collecting evidence from all available data sources to create profiles of specific compounds, complementing omics data in bioinformatics workflows, data-driven developing and validating AOP, and model building based on chemical and biological data.

1.8.6.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The ToxCast and Tox21 data as provided by the US EPA includes over 9076 chemicals tested for 1473 toxicity endpoints (as at November, 2019) that map to hundreds of genes, biological pathways and cellular mechanisms in both humans and rats. The chemicals screened span various uses including industrial, individual, food additive and potentially safer alternatives to already existing older chemicals. The assays tested are usually of two types: i.) cell-based assays which measure changes in cellular response to the test substances; and ii) biochemical assays which measure the activity of a biological macromolecule. The cell typically used may be human or rat primary cells and cell lines. To inform chemical safety decisions, the computational toxicology research group at the US EPA makes both archived and current versions of the data available to the public through 1.) a database called invitroDB which is a MySQL download of all the data, 2.) summary data in flat-file format (e.g. comma-separated value files and tab separated files), 3.) concentration response plots in pdf format, and 4.) a CompTox dashboard (replacing the ToxCast dashboard) which serves as a portal for users to search and query the data. As the Tox21 project progresses and more chemicals are screened in more assays, the US EPA make periodic updates to the public release. At the time of this writing (November 2019), the invitroDB MySQL database is available in the version 3.2 (released 5 August 2019).

The data is released under the US government public domain license 1.0 and is not subject to domestic copyright protection under 17 U.S.C. § 105. It states that unless the work falls under an exception, anyone may, without restriction under U.S. copyright laws, reproduce the work in print or digital form, create derivative works, perform the work publicly, display the work and distribute copies or digitally transfer the work to the public by sale or other transfer of ownership, or by rental, lease, or lending. Therefore, the redistribution of the data by OpenRiskNet under the same conditions is completely covered by the license. With respect to ethics aspects, no personal data is collected during the data generation process as all data is from *in vitro* experiments and commercial cell lines are used for the testing.

OpenRiskNet has transferred the data into the OpenRiskNet-compliant EdelweissData management solution, with the advantage to provide the data via APIs in a semantically annotated form. For this, data was extracted from the MySQL dump of the invitroDB in version 3.1 and 3.2. Additional information on the assays and compounds were extracted from the provided CSV files and enriched by additional chemical identifiers taken from the PubChem service. It was then transferred into table format for the data and metadata in JSON format and the data schema was semantically annotated. These steps prepared the data for upload to the EdelweissData system fully integrated into the OpenRiskNet infrastructure, from which data and metadata can be accessed via the APIs in JSON format for re-use in OpenRiskNet services and workflows.

1.8.6.3 Additional details to 1.6 Expected size of the data

The current version of the invitroDB database available from EdelweissData has a total size of approximately 50 GB of data and metadata.

1.8.6.4 Additional details to 1.7 Utility of data and models

The high-throughput screening toxicity data and models available in ToxCast cover a wide chemical and biological space useful for risk assessment and as such of great value to the OpenRiskNet stakeholders. Integrating this dataset to the OpenRiskNet infrastructure will allow for easier access to the data by users who may not have background or expertise to setup and run the local databases and modelling pipelines. In addition being able to access this data from the OpenRiskNet service will also create greater utility for the data as it can be directly cross-referenced and used for modelling or analysis with other data in the service.

1.8.7 TG-GATEs

Open TG-GATEs [6] is a Japanese public toxicogenomic database resulting from two joint government-private sector projects [7] no date);[8], no date) organized by the National Institute of Biomedical Innovation, National Institute of Health Sciences and multiple pharmaceutical companies. The Japanese Toxicogenomics Project generated gene expression and toxicity data in rats and the primary cultured hepatocytes of rats and humans following exposure to 170 compounds (mainly pharmaceutical products). The follow-up Toxicogenomics Informatics Project discovered over 30 different safety biomarkers using the data and generated additional data for verifying the biomarkers and analyzing their mechanisms.

1.8.7.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

Open TG-GATEs is the public toxicogenomics database developed so that a wider community of researchers can utilize the fruits of TGP and TGP2 research. This database provides public access to data on 170 of the compounds catalogued in TG-GATEs. Data searching can be refined using either the name of a compound or the pathological findings by organ as the starting point. Gene expression data linked to phenotype data in pathology findings is available for download as a CEL file. Such data has to run through different processing steps before it can be used in risk assessment e.g. to run biological pathway enrichment analysis to identify areas of concern and get mechanistic insight. Even if these are standard procedures performed by bioinformaticians and are used to optimize the information available in the data, risk assessors first want to have a quick look at the data to understand the cellular mechanisms caused by a chemical. Therefore, OpenRiskNet provides processed data (intensities and fold changes) generated using a standardized approach, which can directly be used for gene, pathway and mechanism analysis and linked to AOPs using tools like AOP-Wiki and AOP-DB described above.

1.8.7.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

Open TG-GATEs datasets are based on the Affymetrix platforms Rat230-2 (for a rat) and HG-U133_Plus_2 (for a human) using *in vivo* and *in vitro* samples exposed to the chemicals in different concentrations and different timepoints. More information on the exposure scenario and the number of compounds are given in the table below.

Organ or cell type	Organism	Study type	Dose type	Dose level	No. of compounds tested
Liver	Rat	*In vivo*	Repeat dose	Control, low, middle and high	143
Liver	Rat	*In vivo*	Single dose	Control, low, middle and high	158
Kidney	Rat	*In vivo*	Repeat dose	Control, low, middle and high	41
Kidney	Rat	*In vivo*	Single dose	Control, low, middle and high	41
Liver	Rat	*In vitro*	Single dose	Control, low, middle and high	145
Liver	Human	*In vitro*	Single dose	Control, low, middle and high	158

For the generation of the processed data to be provided by OpenRiskNet, the raw data were downloaded from ftp.biosciencedbc.jp/archive/open-tggates/LATEST. Data was stored and processed locally using the R scripting language, which contains well maintained and documented libraries for gene expression analysis. Generally the workflow consisted of two independent parts - extraction of intensity readouts and calculation of fold changes.

The steps for intensity data file generations include (see also Figure 2 for a schematic representation):

1. For every CEL file, **the intensity readouts and probe IDs were extracted** and stored as a new dataset in the form of a CSV file. The CSV file contains only two columns - probe ID and intensity.
2. For every CEL file, the relevant **metadata associated with the particular assay was extracted** (such as organism, organ, study type, compound, dose level, dose type, route of exposure, duration, vehicle) and stored in the form of a JSON file. This is crucial information for easy and efficient searching/filtering in any subsequent data analysis.
3. Both databases provide only compound names as chemical identifiers. In order to

support easier searching/filtering through compounds, we **extracted additional chemical identifiers from the PubChem and CACTUS** (NIH/NCI) services, such as CAS number, SMILES string, InChI, InChI key, IUPAC name, PubChem ID. This information also became part of the metadata file.

4. For every CEL file, the **intensity data and metadata files were combined and uploaded to the EdelweissData** as a separate dataset. Along with that, a short description in a human-readable format was generated from the metadata and uploaded along with the dataset to EdelweissData

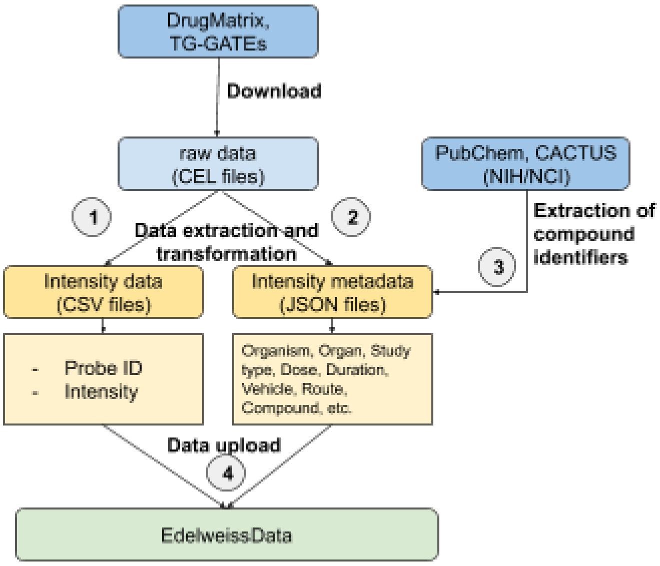

DrugMatrix, TG-GATEs
Download
raw data (CEL files)
PubChem, CACTUS (NIH/NCI)
1
2
Data extraction and transformation
Extraction of compound identifiers
3
Intensity data (CSV files)
Intensity metadata (JSON files)
- Probe ID
- Intensity
Organism, Organ, Study type, Dose, Duration, Vehicle, Route, Compound, etc.
Data upload
4
EdelweissData

Figure 2. Workflow to generate intensity readouts for DrugMatrix and Open TG-GATEs data and upload to EdelweissData. Critical steps in the workflow are numbered according to the description in the main text.

Fold changes were produced using the following workflow (see also Figure 3 for a schematic representation):

1. Firstly, **every CEL file was normalized** using the single-channel array normalization function of the SCAN.UPC library available through Bioconductor [9][10]. The latter has been shown to have the same or better performance as the other competing methods (such as RMA), while providing a crucial advantage due to a one-at-a-time normalization of CEL files (hence there is no need to reprocess all the affymetrix datasets, when an existing database is updated) [9]
2. For every unique set of conditions (compound, dose, organ or cell type, study type, vehicle, route, duration) the corresponding **treatment and control CEL files were identified**.
3. Normalized data of treatment and control CEL files were used as an input for the **differential expression analysis of microarray data** using the very well known limma library available through Bioconductor [11]. The empirical Bayes statistics for differential expression has been used to calculate the t- and p-values for every probe ID. Note that for the *in vivo* studies we used the usual t-test/ANOVA as it considers two independent groups of samples and fits the linear model to the expression data of each gene, while for the *in vitro* studies we used the paired t-test statistics, which consider dependent groups of samples.
4. Additionally, to aid the further analysis of fold changes, we have converted the probe ID column to the corresponding **gene identifiers** (gene symbol, Entrez ID,

Ensembl ID) using the AnnotationDbi library and rat2302.db or hgu133plus2.db array annotation data available through Bioconductor. Processed data has been stored in the form of CSV files with multiple columns (probe ID, gene symbol, Ensembl ID, Entrez ID, logarithm of fold change, average expression, t, p-value, adjusted p-value, B).

5. For every processed file (i.e. for every set of conditions) the **relevant metadata associated with the particular assay was extracted** (such as organism, organ, study type, compound, dose level, dosing type, route of exposure, vehicle, duration) and stored in the form of a JSON file. As for the intensities, this is crucial for easy and efficient searching/filtering in any subsequent data analysis.
6. The **additional chemical identifiers** generated in step 3 of the previous workflow were again used as part of the metadata file.
7. For every set of conditions the **processed data and metadata files were combined and uploaded to the EdelweissData** as a separate dataset.

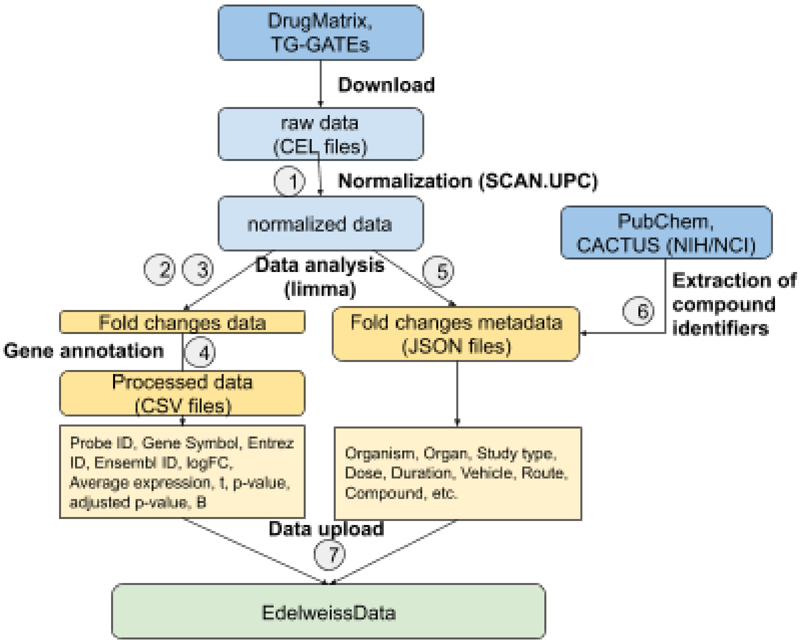

DrugMatrix, TG-GATEs
Download
raw data (CEL files)
1
Normalization (SCAN.UPC)
normalized data
PubChem, CACTUS (NIH/NCI)
2
3
Data analysis (limma)
5
Extraction of compound identifiers
6
Fold changes data
Fold changes metadata (JSON files)
Gene annotation
4
Processed data (CSV files)
Probe ID, Gene Symbol, Entrez ID, Ensembl ID, logFC, Average expression, t, p-value, adjusted p-value, B
Organism, Organ, Study type, Dose, Duration, Vehicle, Route, Compound, etc.
Data upload
7
EdelweissData

Figure 3. Workflow for the calculation of fold-changes of DrugMatrix and Open TG-GATEs datasets. Critical steps in the workflow are enumerated and harmonized with the description in the main text.

1.8.7.3 Additional details to 1.6 Expected size of the data

The overall size of the TG-GATEs processed data is approximately 160GB. The much larger raw data in CEL format is not managed by OpenRiskNet but can be accessed at the original source.

1.8.7.4 Additional details to 1.7 Utility of data and models

Processed data created using standardized normalization and processing procedures offer the advantage that they can be directly combined with similar data from other sources. At the moment, this can be done for TG-GATEs and DrugMatrix but will be extended to other sources in the future. The information can then be combined and re-used in OpenRiskNet services for gene and pathway enrichment analysis using tools like WikiPathways as well as linked to AOPs based on the knowledge covered e.g. in AOP-Wiki and AOP-DB.

1.8.8 DrugMatrix

DrugMatrix is one of the world's largest toxicogenomic reference resources provided by the National Toxicology Program of the US Department of Health and Human Services. It provides access to the toxicogenomic profiles of over 600 different compounds generated with Affymetrix and Codelink microarrays. While both types of microarray cover liver, kidney, thigh muscles, heart and cultured hepatocytes, the Codelink microarrays additionally cover bone marrow, spleen, intestine and brain.

1.8.8.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

As for TG-GATEs, OpenRiskNet provides processed data (intensities and fold changes) generated using standardized approach for normalization and processing starting from the DrugMatrix transcriptomics data, which can directly be used for gene, pathway and mechanism analysis and linked to AOPs using tools like AOP-Wiki and AOP-DB described above via simple to use APIs.

1.8.8.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

Only data from the Affymetrix microarrays is available as OpenRiskNet service up to now. All datasets are based on the Affymetrix microarray platform RG230-2.0 and male Sprague Dawley rats.

Organ or cell type	Study type	Dose type	Dose level	No. of compounds tested
Heart	*In vivo*	Repeat dose	Control, low and high	88
Kidney	*In vivo*	Repeat dose	Control, low and high	139
Liver	*In vivo*	Repeat dose	Control, low and high	200
Thigh Muscle	*In vivo*	Repeat dose	Control, low and high	21
Cultured Hepatocytes	*In vitro*	Single dose	Control and high	125

1.8.8.3 Additional details to 1.6 Expected size of the data

For the generation of the processed data to be provided by OpenRiskNet, the raw data were downloaded from https://ntp.niehs.nih.gov/results/drugmatrix/index.html, then processed using the workflows described for TG-GATEs above and uploaded to the EdelweissData system for easy access in OpenRiskNet workflows. The overall size of DrugMatrix datasets is 40 GB.

1.8.8.4 Additional details to 1.7 Utility of data and models

See section 1.8.7.4 for a general description of the utility of transcriptomics processed data.

1.8.9 KIT Daphnia data on nanoparticles

"KIT Daphnia data on nanoparticles" dataset is based on "Meta-analysis of Daphnia magna nanotoxicity experiments in accordance with test guidelines" study [12] and contains the raw “original_daphnia” data file and its eight derived processed files.

1.8.9.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The "original_daphnia" data is compiled from research articles in which nanotoxicity toward Daphnia Magna was measured according to the test guidelines from OECD and US EPA. Toxic response caused by nanomaterials have been assayed; however, the assay outcomes were varied between research articles. Therefore, this data set was compiled for meta-analysis to figure out what may be the cause of data heterogeneity between diverse assay outcomes.

This data contains physicochemical properties of nanomaterials, experimental conditions for assays and toxic response measurement. Most of physicochemical properties of nanomaterials were taken from the research articles without considering measurement details for them since measurement details were often absent in the articles. Therefore, further details could only be found in reference research article. Since nanomaterials were easily aggregated in media, dispersion methods were applied to make nanomaterials being nano-sized particle. Centrifuge, stir, sonication, and filter were four dispersion methods used among research articles.

The eight processed files are prepared by the authors and contain numeric values obtained from quantitative values of the raw "original_daphnia" data. They are “carbon_pec50”, containing the carbon-based nanomaterials with EC50 values, “coated_m_pec50” and “coated_m_class”, containing the coated metal nanoparticle data, “fullerene_class”, containing the fullerene nanomaterials data, “metal_pec50” and “metal_class”, containing the metal nanoparticle data, “meox_pec50” and “meox_class”, containing the metal oxide nanoparticle data. The eight processed files are used to build models, the ones with the label “pec50” are used for regression models and the ones with label “class” are used for classification models.

This nano datasets are very important for standardization and validation of test methods for regulatory usage and identifying reasons for discrepancies in results, even if test guidelines exist, and are a good example of a relatively small data source profiting highly from data management solutions integrated in the infrastructure and the harmonization and interoperability effort of OpenRiskNet. Integration into OpenRiskNet made it publicly available following the FAIR principles.

1.8.9.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The Daphnia data as used in the meta-analysis (Shin, Hyun Kil & Seo, Myungwon & Shin, Seong & Kim, Kwang-Yon & Park, June-Woo & No, Kyoung Tai. (2018)) was available only from the author composed out of the raw “original_daphnia” data file and its eight derived processed files, all being in tabular CSV (comma separated values) format. By providing it on the EdelweissData system, the data and associated metadata is now available through the APIs in JSOn format and can be reused in an interoperable way in combination with other data sources. Even if the data sources provided by the OpenRiskNet consortium cover nanomaterials only to a very small extent, the uptake of OpenRiskNet solutions e.g.

by the NanoCommons infrastructure will result in new interoperable resources in the near future. Models for nanomaterials using the data are already part of the infrastructure.

1.8.9.3 Additional details to 1.6 Expected size of the data

The total size of the data and metadata of this small data source are below 1 MB.

1.8.9.4 Additional details to 1.7 Utility of data and models

Even if this is a very small data source, it provides important information on the reproducibility of nanomaterial hazard data generated following a validated OECD guideline. It can now guide researchers with the task to perform equivalent experiments on new nanomaterials, and encourage them to provide the results in the same level of detail and preferable also to deposit the data on an OpenRisk data solution allowing combined analysis and monitoring progress with respect to reproducibility.

1.8.10 ToxicoDB

ToxicoDB is a web application to mine large and small scale toxicogenomics datasets. To better understand the molecular mechanisms underlying compound toxicity, great efforts have been made in screening of drugs/chemicals by various groups to generate datasets such as Open-TG GATEs and DrugMatrix.

1.8.10.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The inspiration for this application comes from the need for a common ground for analyzing toxicogenomics datasets with maximum overlap and consistency. The ToxicoDB will provide an intuitive interface for all users (including users that are not computational-savvy) to mine the complex toxicogenomic data. This is in line with OpenRiskNet’s objectives.

1.8.10.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

ToxicoBD will provide curated toxicogenomics datasets, which includes gene expression data and toxicological information on drugs and chemicals used to generate these gene expression data. These datasets are and will be obtained from publicly available resources, such as the diXa data warehouse, NCBI GEO, EBI’s ArrayExpress. All datasets are and will be preprocessed and checked for quality.

1.8.10.3 Additional details to 1.6 Expected size of the data

As a resource provided by an associated partner of OpenRiskNet, the data management and sustainability is in the hand of this institution. As work on ToxicoDB is still ongoing it is unclear what the size of the data will be. Currently, only data from TG-GATEs and DrugMatrix are available. However, by integrating the service into the OpenRiskNet infrastructure, OpenRiskNet guarantees that the metadata will be publicly available.

1.8.10.4 Additional details to 1.7 Utility of data and models

Users of the database can find drug and gene annotations, visualize gene expression data for datasets as well drugs of interest and will be able to download these data.

1.8.11 ToxPlanet

ToxPlanet aggregates and curates toxicology and chemical hazard information from over 500 sources. A web-based GUI as well as access via APIs to the commercial service are possible.

1.8.11.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

ToxPlanet gives access to regulatory reports and other structured and unstructured information sources. In this way, it is a comprehensive and well organized source for legacy data and documents the current state of regulation for a large number of compounds. This information and can be used in all application areas of OpenRiskNet partly automatically extracted by the text-mining workflows developed in OpenRiskNet.

1.8.11.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The data is mainly in the form of reports available in pdf format and originate from all major regulatory agencies and literature sources. Most of the documents are in the public domain while the search and browsing features of ToxPlanet are proprietary and their use needs a commercial license.

1.8.11.3 Additional details to 1.6 Expected size of the data

Due to the commerciality of the service, the data is managed by the provider. However, by integrating the service into the OpenRiskNet infrastructure, OpenRiskNet guarantees that the metadata will be publicly available.

1.8.11.4 Additional details to 1.7 Utility of data and models

OpenRiskNet provides text mining workflows to extract data from the ToxPlanet repository under the restriction that the user needs to acquire a license first.

1.8.12 SCAIView

The information retrieval system SCAIView allows for semantic searches in large text collections by combining free text searches with the ontological representations of entities derived by JProMiner. SCAIView gives answers to questions such as “Which genes / proteins are related to a certain disease, pathway or epigenetics?”. SCAIView´s key features are:

- A user-friendly search environment with a query builder supporting semantic queries with biomedical entities
- Fast and accurate search and retrievals, based on the newest technologies of

semantic search engines
- Visualization and ranking of the most relevant entities and documents
- Exportation of the search results in various file formats

Documents are retrieved by precisely formulated questions using ontological representations of biomedical entities. The entities are embedded in searchable hierarchies and span from genes, proteins, accompanied single-nucleotide polymorphisms to chemical compounds and medical terminologies. SCAIView supports the selection of the suitable entities by an autocompletion functionality and a knowledge base for each entity. This includes a description of the entity, structural information, pathways and links to relevant biomedical databases like EntrezGene, dbSNP, KEGG, GO and DrugBank.SCAIView represents the search results using a color-coded highlighting of the different entity-classes, statistical search results and various ranking functions. The selected biomedical entities are found by an approximate search algorithm implemented in the Fraunhofer-Gesellschaft information extraction tool JProMiner® which additionally disambiguates synonyms of entities to unique identifiers in public available entity databases. The SCAIView data collection comprises three different public document collections and an index of extracted biomedical entities from these documents.

1.8.12.1 Additional details to 1.1 Purpose of the data collection and 1.2 Relation to the objectives of the project

The main purpose of the data collection is to enable a semantic search functionality for researchers for public available full text collections. The documents are retrieved via free text queries in combination with semantic or ontological search of biomedical entities of interest. The biomedical entities are embedded in searchable hierarchies and span from genes, proteins, accompanied SNPs to chemical compounds and medical terminology. With Ontological Filtering, it is possible to restrict the result to a subset e.g. genes on a KEGG pathway or in a Cytoband region.

Advanced retrieval technology allows answering complex queries such as:
- Which genes/proteins are related to a certain context (e.g. disease/pathway/epigenetics)?
- Give me an overview of relevant biomedical concepts in my subcorpus
- Which drugs are relevant for this context?
- To which diseases is my gene associated?
- Which chromosomes show linkage to the disease?
- Which variations are mentioned in the context of the disease and could they be found in dbSNP?
- What other diseases are possibly co-occurring with my relevant disease?

The collection has been used in the Data Cure case study to find relevant information for chemical compounds and their cancer hazard to humans.

1.8.12.2 Additional details to 1.3 Types and formats of data, 1.4 Reuse of data and 1.5 Origin of the data

The data collection has been derived from publicly available xml collections from the U.S. National Library of Medicine (NLM: PubMed, PMC under Terms & Conditions of NLM) and the United States and Trademark Office (USPTO: PatFT). The annotated files are converted to JSON format and can be downloaded via the SCAIView API.

1.8.12.3 Additional details to 1.6 Expected size of the data

The size of the raw xml data is: PubMed 40G, PMC 61G, Patents 79G. The processed data comprises: PubMed 664G, PMC 431G, Patents 179G. The publication numbers increase constantly every year in the NLM and the USPTO. The processed data even grows at a larger rate since more terminologies and ontologies are indexed each year.

The processed data is loaded into a solr enterprise search engine and is hosted and regularly updated by Fraunhofer SCAI under the following Terms & Conditions.

2. FAIR DATA

2.1 Making data findable, including provisions for metadata

- *Outline the discoverability of data (metadata provision)*
- *Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?*
- *Outline naming conventions used*
- *Outline the approach towards search keyword*
- *Outline the approach for clear versioning*
- *Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how*

OpenRiskNet was integrating existing data sources and made them more easily findable, accessible and interoperable. This was based, on the one hand, on the metadata provided by the data sources and, on the other hand, on the interoperability layer and exploitation of semantic web standards (Resource Description Framework, RDF) , which harmonises these metadata leading to data service descriptions and data schemata, which can be queried through the OpenRiskNet discovery service.

The description of the capabilities of a database and the data schema allow for:

- Accessing specific search functionality, and
- Identify the data fields to be searched (e.g. where information on the biological assays are stored);
- Finding the best format for data exchange;
- Understanding all the data and tools, with transparent access to metadata describing the experimental setup or computational approaches.

In the case that the original data sources don't provide all the features required by the FAIR principles as e.g. unique persistent identifiers or clear access protocols, OpenRiskNet worked together with the data providers to either integrate this into the original service, transfer the data to more advanced data management solutions provided by OpenRiskNet or provide missing features as part of the interoperability layer added to the service in the context of OpenRiskNet all leading to improved quality of the data source.

2.2 Making data accessible

- *Specify which data will be made openly available? If some data is kept closed provide rationale for doing so*
- *Specify how the data will be made available*
- *Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?*
- *Specify where the data and associated metadata, documentation and code are deposited*
- *Specify how access will be provided in case there are any restrictions*

The OpenRiskNet approach enables the easy and transparent sharing and analysis of data between organisations involved in many sectors and programs. OpenRiskNet APIs and the used transfer formats were openly released immediately after their definition had reached a stable form at the end of the project. However, updates might be necessary to follow scientific and technical advances. In the prioritization of services to be integrated, open source tools were favoured for the use in the case studies and the reference workflows targeting specific question but commercial services are equally important to build in the specific requirements of restricted access rights and to help sustain the infrastructure in the long run. However, also these commercial services had to openly share their API definitions and data formats as well as provide features to make medatadata even of such restricted data findable to allow for integration and combination with other tools and comply with the FAIR principles.

Open Standards applied:

- Data and models are stored and served using well-developed and widely applied standards and technologies that promote data reuse and integration, such as JSON-LD, RDF and related semantic web technologies;
- OpenRiskNet resources are aligned with activities of toxicology communities like OpenTox, NanoCommons and EU-ToxRisk in developing open standards for predictive toxicology resources;
- Tools to access study data and metadata description in standards file formats already in use in a number of omics, toxicogenomic and nanosafety resources (e.g. ToxBank, diXa, eNanoMapper), further simplify the integration;
- Model descriptions are provided encoded guided by suitable open standards (e.g. QMRF, BEL, SBML) and annotated advancing appropriate minimal information standards (MIRIAM) for dissemination through appropriate repositories (e.g. BioModels) to cover the extended requirements of the semantic interoperability layer of OpenRiskNet.

OpenRiskNet did not create new file standards but rather employed existing approaches as to define a core set of information, on which the scientific community agrees that they are important to document, but which can also be modified and extended if necessary for a specific application. For defining this core set, regulatory files formats like **OECD harmonised templates (OECD HT)** [13] and **Standard for Exchange of Nonclinical Data (SEND)** [14] were included when collection the requirements for file transfer. Even if these file formats are too limited and do not have the flexibility to be used outside regulatory purposes and especially for early stage research and method development, the guidelines for data and metadata management proposed by OpenRiskNet and is continued to be developed in NanoCommons and other projects ensures that all relevant metadata and information needed for regulatory reporting is included in the data transfer templates.

2.3 Making data interoperable

- *Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability*
- *Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to*

more commonly used ontologies?

OpenRiskNet interoperability layer opens possibilities to provide data schemata, which describe the format of the data using a controlled vocabulary:

- Metadata standards and data documentation approaches consider the existing standards that can be consolidated and the equivalent data that can be retrieved independent of the file format;
- Developments towards the integration of ontologies under a single framework have been started and are ongoing together with partner projects mainly from the EU NanoSafety Cluster, which will contribute to the goal of automatic harmonization and annotation of datasets. The goal is not to develop new ontologies. Instead, already existing ontologies (e.g. OBI, ChEBI, PATO, UO, NCIT, EFO, OAE, eTOX, eNanoMapper, MPATH, etc.) are consolidated and integrated into applications ontologies for the toxicology community and specifically for the requirements of OpenRiskNet service annotation.
- Another requirements to establish the comprehensive use of ontologies and in this way foster the interoperability not only of the major data sources but also user-provided data are user-friendly capturing frameworks supporting the selection of ontology terms during data curation and an ontology mapping service resolving issues of using synonyms from different ontologies (e.g. CAS numbers can be annotated using the National Cancer Institute Thesaurus, the EDAM ontology or even the Chemical Information Ontology reused in the eNanoMapper ontology, where it is available under the term “CAS registry number”). OpenRiskNet was working with experts in the field to integrate such tools in the infrastructure and has partly integrated them into data management services provided by OpenRiskNet partners. However, since OpenRiskNet was not a major primary data provider, it is even more important that these tools are now made available and easily accessible to the community for integration in new and existing data sources of EU funded research projects.
- Allowing mapping between related items in different databases (e.g. different gene-identifiers, linking genes to proteins or RNA identifiers, or mapping between equivalent chemical structures in different databases. BridgeDb, which can perform such mappings and is part of the OpenRiskNet Services, is thus a core interoperability service.

Additionally, we provide guidelines and training on the usage of standard data transfer/sharing formats and ontologies in the context of OpenRiskNet. Best practice examples like ToxCast, AOP-Wiki and AOP-DB show how semantic annotation can be applied either directly by providing the data as RDF or via semantic annotated OpenAPI definitions or both and used to make the data and information understandable and, in this way, easier integrable and re-usable.

An important and heavily used part of improved data management throughout the OpenRiskNet project is searching and accessing data from different sources supported by the semantic annotation of the data sources based e.g on the Bioschemas and BioAssays ontology:

- The databases are accessible by the OpenRiskNet APIs (similar to the computational tools) including the interoperability layer;
- Searches throughout multiple databases are possible, removing the need to search

in everyone independently

- The interoperability layer can be used to inspect the data schema and find out if the needed information is available from the databank and if it can be provided in a form for further analysis.

All this work was based on and extended:

- OpenTox APIs, which were designed to cover the field of QSAR-based predictive toxicology with dataset generation, model building, prediction and validation;
- Open PHACTS APIs, which handle knowledge collection and sharing;
- Various other APIs for accessing databases like BioStudies, EGA, ToxBank, and PubChem;

which led to a fast uptake by the community due to existing familiarity with the underlying concepts.

2.4 Increase data re-use (through clarifying licenses)

- *Specify how the data will be licenced to permit the widest reuse possible*
- *Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed*
- *Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why*
- *Describe data quality assurance processes*
- *Specify the length of time for which the data will remain re-usable*

Most of the data sources are already available in the public domain. OpenRiskNet redistributes the data using the same license as the original data provider or, if this is demanded by the data provider, in a more restricted form. New data is made available through the OpenRiskNet access methods as soon as it is released by the original data provider, i.e. no additional embargo period is/was enforced by the OpenRiskNet project. Thus, users can access the same data with respect to the number of datasets and version of each dataset either from the original service provider, via e.g. a web interface specifically designed for the data warehouse, or through the OpenRiskNet mechanism, where the latter has the advantage of easy integration into workflows and interoperability with other data sources and software tools. Besides this simpler access, OpenRiskNet improved the quality of the data with the following measures:

Data quality assurance processes:

- Tools for performing automatic validation, analysis and (pre)processing are developed to find inconsistencies in the data and databases and, in this way, improve the quality of the source and are made available in OpenRiskNet, e.g. http://arrayanalysis.org/ and https://github.com/BiGCAT-UM. Additionally, efforts to establish a general, cross-database data curation framework, in which users can flag possible errors in the data and semantic annotation, is supported.
- Some partners (e.g. UM) developed their own pipelines for quality control and analysis of sequencing data (RNA-seq and MeDIP-seq).
- We also integrate tools for automatic or manual curation of datasets as well as deriving processed data. The modified dataset are stored (similar to the pre-reasoned datasets) in OpenRiskNet-compliant databases with a link to the

original source. Discussions were started, are underway and will continue with the original data providers to transfer the curated datasets back into the original database so that users preferring to use the data from the primary source are also profiting from the curation effort.

Quality assurance in the processing, analysis and modelling tools:

- Protocolling of the performed calculations increasing the repeatability and reproducibility of the studies, is supported by the automatic logging and auditing functionalities of modern microservices frameworks as well as the integrated workflow management systems.
- Validation of the services were enforced by the consortium and appropriate measures of uncertainty were requested for all models.

2.5 Specific information on individual shared data sources

2.5.1 Workflows

All OpenRiskNet workflows created as part of the case studies or demonstrating the functionalities of individual services or combinations of services are available under open licenses from the Github repository. These can be downloaded, re-executed to demonstrate repeatability and modified to answer specific scientific questions of the user.

2.5.2 BridgeDb

The BridgeDb software is available under the OSI-approved Apache License 2.0. Identifier mappings files are available under open licenses too, following the open licenses of the upstream resources (Ensembl, Rhea) or CCZero in case of the metabolite mapping database. The BridgeDb web service and data for identifier mappings is made available on the OpenRiskNet cloud using an OpenAPI specification wrapped around a REST services.

2.5.3 WikiPathways

All contents of WikiPathways are licenced with the Creative Commons CC0 waiver, which states that all contents of the database are free to share and adapt. WikiPathways adopts a customised quality assurance protocol to curate the database, which is done on a weekly basis.

2.5.4 AOP-Wiki

The AOP-Wiki provides quarterly downloads for the complete database, which are permanently maintained by the OECD. The AOP-Wiki does not provide licence information, but states that the data can be reused. All AOPs undergo review by EAGMST to ensure the quality of the contents of the AOP-Wiki.

FAIR Principles	WikiPathways	AOPWiki
F1. (Meta)data are assigned and globally unique and persistent identifiers	2	1
F2. Data are described with rich metadata	1	1
F3. Metadata clearly and explicitly include the identifier of the data they describe	2	2
F4. (Meta)data are registered or indexed in a searchable resource	2	2
A1.1. The protocol is open, free and universally implementable	2	2
A1.2. The protocol allows for an authentication and authorization where necessary	2	2
A2. Metadata are accessible, even when the data are no longer available	2	2
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	2	2
I2. (Meta)data use vocabularies that follow FAIR principles	1	1
I3. (Meta)data include qualified references to other (meta)data	2	1
R1.1 (Meta)data are released with a clear and accessible data usage license	2	1
R1.2. (Meta)data are associated with detailed provenance	1	1
R1.3. (Meta)data meet domain-relevant community standards	1	2

Table 1. Compliance to FAIR principles [15] by AOP-Wiki and WikiPathways. Score meanings: 1 = partial compliance, 2 = compliance

2.5.5 AOP-DB

The AOP-DB provides data that originates from a variety of data sources. The OpenRiskNet services that exposes the AOP-DB RDF contains data from NCBI, ToxCast, CTD and DisGeNET. NCBI and ToxCast are produced by the U.S. Government and the information is by default in the public domain. The DisGeNET database is made available under the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. CTD is subject to the terms of reuse by the MDI Biological Laboratory and NC State University (http://ctdbase.org/about/legal.jsp).

2.5.6 ToxCast

All data produced by the U.S EPA including ToxCast/Tox21 is by default in the public domain (U.S. Public Domain license) and is not subject to domestic copyright protection under 17 U.S.C. § 105. This allows to reproduce the work in print or digital form, create derivative works, perform the work publicly, display the work and distribute copies or digitally transfer the work to the public by sale or other transfer of ownership, or by rental, lease, or lending. The release currently provided as part of OpenRiskNet is using the EdelweissData system including a REST API developed by Edelweiss Connect and is based on the MySQL database dump and additional downloadable CSV files. The data was extracted and restructured to fit the OpenRiskNet concept and semantic annotated.

2.5.7 TG-GATEs

Open TG-GATEs is provided as a public source by the Japanease National Institutes of Biomedical Innovation, Health and Nutrition. OpenRiskNet used standard procedures to normalize and processed and now provides the data under the Attribution CC BY 4.0 license based on the EdelweissData platform, which is designed to comply to the FAIR principles.

2.5.8 DrugMatrix

DrugMatrix is provided as a public source by the National Toxicology Program of the US Department of Health and Human Services. As for TG-GATEs, OpenRiskNet used standard procedures to normalize and processed and now provides the data under the Attribution CC BY 4.0 license based on the EdelweissData platform, which is designed to comply to the FAIR principles.

2.5.9 KIT Daphnia data on nano particles

KIT Daphnia data on nano particles dataset is provided by the Department of predictive toxicology, Korea Institute of Toxicology. OpenRiskNet provides the data under the Attribution CC BY 4.0 license based on the EdelweissData platform, which is designed to comply to the FAIR principles.

2.5.10 ToxicoDB

ToxicoDB is provided by the University Health Network, Toronto, Canada and is available under GNU Lesser General Public License 3 (LGPLv3.0). The database provides toxicological data and toxicogenomics datasets (incl. TG-GATEs and DrugMatrix) from publicly available sources and thereby complying to the FAIR principles.

2.5.11 ToxPlanet

ToxPlanet is available under a commercial license. However, metadata is publicly available to comply with the FAIR principles.

2.5.12 SCAIview

The main resources on which SCAIView builds (namely PubMed and US Patents) are very metadata rich and SCAIView makes this metadata searchable and accessible via API. SCAIView harmonizes the metadata of the different sources into a common JSON schema and therefore makes the data more interoperable and reusable. SCAIView adds additional metadata from text mining by adding semantic annotations. These annotations are derived from publically available ontologies (eg. OBO foundry). Each annotation contains a referable identifier (URI), a source, a preferred label and workflow provenance on how this annotation has been produced. I.e. the generated data is by definition FAIR.

3. ALLOCATION OF RESOURCES

Explain the allocation of resources, addressing the following issues:

- *Estimate the costs for making your data FAIR. Describe how you intend to cover these costs*
- *Clearly identify responsibilities for data management in your project*
- *Describe costs and potential value of long term preservation*

Making data FAIR was a central task of the integration of data source in the OpenRiskNet infrastructure. Many of the sources integrated already follow the FAIR principles at least to some extent and limited additional budget was needed other than for the effort needed to make the sources OpenRiskNet-compliant. For data sources not at a sufficient high level, the OpenRiskNet partners owning the data or responsible for its integrating covered the costs of the integration from their allocated budget. In the case of third-party data sources, the integration was performed in collaboration with the associated partners partly financially supported through the Implementation Challenge with an OpenRiskNet partner designated as main contact point of the associated partner.

4. DATA SECURITY

- *Address data recovery as well as secure storage and transfer of sensitive data*

The OpenRiskNet approach on data recovery, secure storage and transfer of sensitive data included:

- Responsible and secure management processes for personal data including anonymisation, encryption, logging of data usage as well as data deletion after usage are implemented;
- To ensure that all ethical guidelines are followed by all OpenRiskNet Partners and Associated Partners and implemented in every step of the infrastructure, a *privacy by design* approach was followed in the project, documented in the OpenRiskNet privacy policy (see below) and controlled by an independent Data Protection Officer;
- The most sensible way to protect sensitive data offered by the OpenRiskNet infrastructure was to bring the virtual environment and all data sources behind a company's firewall by in-house deployment;
- All these data protection measured were documented and distributed as part of the terms of usage to guide existing and future data providers and research projects in need for secure and ethical data management.

5. ETHICAL ASPECTS

- *To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former*

The ethical aspects are covered by the Protection Of Personal Data (POPD) requirements and deliverables:

- D6.1 - NEC - Requirement No. 3: Statements regarding the full compliance of third country participants to the H2020 rules
- D6.2 - POPD - Requirement No. 4: Copies of the previous ethical approvals of data to be collected and used in the project, approvals which must also allow the possible secondary use of data
- D6.3 - POPD - Requirement No. 5: Consent forms and information sheets before interviews and surveys and any other data and personal data collection activity in the project
- D6.4 - POPD - Requirement No. 6: A statement by third party testers that they will comply with the applicable EU law and H2020 rules
- D6.5 - POPD - Requirement No. 7: Data Protection Officer Report 1
- D6.6 - POPD - Requirement No. 8: Data Protection Officer Report 2
- D6.7 - POPD - Requirement No. 9: Data Protection Officer Report 3

Based on the ethics report received as part of the evaluation of the proposal and the three reports of the **Data Protection Officer (DPO)** (D6.5, D6.6 and D6.7) summarising the relevant regulations and legislation for the different data types and test models, OpenRiskNet has worked also together with an external expert to develop elements on an ethics review framework and a document in the form of a “Proposal for ethics requirement for data providers to European e-infrastructures like OpenRiskNet” was generated (version 3 in **Annex 1**). This document was further reviewed and improved based on the recommendation received from the DPO.

The OpenRiskNet platform is based on a federated data management system that makes existing data sources provided by other European projects in this field or from international consortia mainly from *in vitro* human and animal and *in vivo* animal experiments available to all stakeholders in a harmonised way. Such a federated systems imposes highest demands on data management and sharing with respect to data security, data integrity and ethics becoming even more relevant with the new EU General Data Protection Regulation (GDPR) in effect since May 2018. Thus, not only OpenRiskNet but the complete scientific community is in demand of support and recommendations to achieve the highest impact without jeopardising the ethics integrity of the e-infrastructure.

As a starting point and to fulfil the requirements from the ethics review, a step-by-step decision process was first designed addressing how important legacy data sources need to be handled by the project. On top of the workflows provided in the reviews of the Data Protection Officer, a hierarchical data source analysis and evaluation of the ethical implications for OpenRiskNet was performed. Different categories of data sources have been analysed, including references to the legislation in place and the conditions for primary, secondary and tertiary data collection and use. Also special measures that need to be considered for some specific cases are included.

Data type and additional aspects considered:

- General requirements
- Additional Requirements for Human Data
    - Basic requirements for ALL types of human data
    - Further requirements according to type of biomaterial being provided
        - Commercial Cell Lines
        - Research Using, Producing or Collecting Human Cells and Tissues (EXCLUDING Embryonic Stem Cells (hESC) and Human-Induced Pluripotent Stem Cell (hIPSC)
        - Research using Human Embryonic Stem Cells hESCs and Human-induced Pluripotent Stem Cell (hiPSC) Lines
        - Research using Clinical Trial Data
- Additional Requirements for Animal Data

Reasoning
The European Commission is enforcing strong data management guidelines following the FAIR principles (Findable, Accessible, Interoperable and Re-usable) on all funded projects and the open, public sharing of all data generated in these projects. However, this sharing of research data needs to be following all ethics requirements and highest standards of privacy protection. To support the definition of such standards, OpenRiskNet has developed this checklist to help data providers understand the requirements for privacy-protecting data sharing and to guide them in fulfilling the existing guidelines and the ethics requirements for the specific data supplied. The checklist has general requirements that apply to all data types, and additional requirements specifically for data used in the toxicology and risk assessment fields. The final goal is to allow for an evaluation following the relevant criteria and to generate a statement confirming that the applicable requirements are fulfilled before a dataset can be used.

Since data management and sharing is becoming a more centralised, European or even global effort spanning different infrastructures and disciplines, recommendations and guidelines adopted by OpenRiskNet cannot be developed autonomously and have to be aligned and harmonised with the ongoing discussions on the EU level and with changing regulations. Therefore, this document needs to be considered only as an initial attempt to cover issues in the specific scientific area and it will be provided to working groups established within the governance of the European Open Science cloud (EOSC) on FAIR data requirements. Until additional feedback is collected and other players are consulted (i.e. EOSC) the information included in Annex 1 will not be implemented yet as an online tool linked to the OpenRiskNet infrastructure, but kept as an input document to the e-infrastructure community.

The current draft document of the checklist proposes specific ethics evaluation steps for each data source to be included and/or used in the OpenRiskNet platform based on the assessment to one of the data categories listed above (animal *in vitro* and *in vivo*, human *in vivo*,...) aligned with specific regulatory and ethical requirements. Even if the obligation to fulfil all necessary national and international regulatory and ethical requirements including obtaining legal and ethics clearance of all experiments and to operate in conformity with the institutional regulations is ultimately in the hands of the original data producer, OpenRiskNet is committed to continue providing the framework for the ethics evaluation described above to all providers of data services (OpenRiskNet internal, associated partners and other third-parties) and supported the execution and documentation of the data source evaluation. However, as already mentioned, all scientific disciplines are now facing the same challenge how to document ethical and privacy-protecting generation, management and sharing of research data within the open research data and open science goals of the European Union. Therefore, we did not enforce the adoption of the checklist as a requirement for data services on OpenRiskNet since this would be very specific to this e-infrastructure while the services are also of interest in a more general setting like e.g. the European Open Science Cloud. This makes it clear that discussions on ethics standard have to pushed to a higher level and harmonised across all relevant European infrastructures or even better globally to avoid that data providers have to deal with multiple parly contradicting, incomplete or outdated requirements, guidelines and checklists depending on the setting their services are used in.

To funnel our experience, know-how and results of the ethics reviews into these discussions, we opened the checklist now available in a draft version for comments, raised the importance of clear licenses, privacy-protection and ethics guidelines at different project but even more important ELIXIR and EOSC meetings (including the Building EOSC through the Horizon 2020 projects current status and future directions" workshop, 9 – 10 Sep 2019 in Brussels), and engaged with interested parties and the EOSC-Secretariat to improve and extend the OpenRiskNet guidelines and checklist to make them fit for multiple disciplines and infrastructures and to foster the uptake of by the scientific community.

5.1 Privacy Policy

The privacy policy[7] is implemented (version 3 from 20 November 2019) that discloses the ways OpenRiskNet website manages the content, the personal data or analytics on website usage. Specifically, the disclaimer implemented refers to the following aspects related to the Personal Data Protection and Privacy Policy:

- Website content disclaimer
- External links disclaimer
- Copyright and acknowledgement of sources
- Data Protection and Privacy Policy
  - Responsibility for the processing of Personal Data
  - Categories of Personal Data processed by OpenRiskNet
  - Storage of Personal Data
  - Sharing of Personal Data
  - Data Protection Rights under the General Data Protection Regulation (GDPR)
    - Service Providers
    - Analytics
  - Security and Integrity
  - Cookies
  - SSL or TLS encryption
  - Email communication
  - Changes of the Data Protection and Privacy Policy
- Legal effect of disclaimer
- Contact details

The latest version of the Privacy Policy is included in **Annex 2**.

[7] https://openrisknet.org/privacy-policy/

5.2 Terms of use

The terms of use[8] (version 3 from 20 November 2019) regulate the use of the OpenRiskNet infrastructure and is structured as follows:

- Definition of terms used
- About the OpenRiskNet e-infrastructure
- Data providers and data users of OpenRiskNet e-infrastructure
- Use of OpenRiskNet e-infrastructure
- Confirmation of Acceptance of the terms of use

The latest version of the OpenRiskNet e-infrastructure terms of use is included in **Annex 3**.

[8] https://openrisknet.org/terms-of-use/

GLOSSARY

The list of terms or abbreviations with the definitions, used in the context of OpenRiskNet project and the e-infrastructure development is available at:

https://github.com/OpenRiskNet/home/wiki/Glossary

REFERENCES

1. Farcal L, Florean O, Doganis P, Jennen D, Willighagen E, Martens M, et al. Initial version of data management plan (Deliverable 3.1). 2019 [cited 19 Sep 2019]. doi:10.5281/zenodo.2558117

2. Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, 26 July 2016. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

3. How to create a Data Management Plan. In: OpenAIRE [Internet]. Available: https://www.openaire.eu/how-to-create-a-data-management-plan

4. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, et al. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics. 2010;11: 5.

5. Batchelor C, Brenninkmeijer CYA, Chichester C, Davies M, Digles D, Dunlop I, et al. Scientific Lenses to Support Multiple Views over Linked Chemistry Data. In: Mika P, Tudorache T, Bernstein A, Welty C, Knoblock C, Vrandečić D, et al., editors. The Semantic Web – ISWC 2014. Cham: Springer International Publishing; 2014. pp. 98–113.

6. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. 2015;43: D921–7.

7. Construction of Drug Safety Prediction System by Toxicogenomics Technology and Related Basic Research. Report No.: H14-Toxico-001.

8. Drug Safety Assessment Based on Toxicity Mechanisms using Toxicogenomics Database. Report No.: H19-Toxico-001.

9. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 2012;100: 337–344.

10. Piccolo SR, Withers MR, Francis OE, Bild AH, Johnson WE. Multiplatform single-sample estimates of transcriptional activation. Proceedings of the National Academy of Sciences. 2013. pp. 17778–17783. doi:10.1073/pnas.1305823110

11. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.

12. Shin HK, Seo M, Shin SE, Kim K-Y, Park J-W, No KT. Meta-analysis of Daphnia magna nanotoxicity experiments in accordance with test guidelines. Environ Sci: Nano. 2018;5: 765–775.

13. OECD Harmonised Templates - OECD. Available: https://www.oecd.org/ehs/templates/

14. Standard for Exchange of Nonclinical Data (SEND). In: CDISC [Internet]. Available: https://www.cdisc.org/standards/foundational/send

15. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.

ANNEXES

Annex 1. Proposal for ethics requirement for data providers to European e-infrastructures like OpenRiskNet

Annex 2. Personal Data Protection and Privacy Policy

Annex 3. OpenRiskNet e-infrastructure terms of use

Proposal for ethics requirement for data providers to European e-infrastructures like OpenRiskNet

Version 3** from **19 November 2019

The European Commission is enforcing strong data management guidelines following the FAIR principles (Findable, Accessible, Interoperable and Re-usable)[1] on all funded projects and the open, public sharing of all data generated in these projects. However, this sharing of research data needs to be following all ethics requirements and highest standards of privacy protection. To support the definition of such standards, OpenRiskNet[2] has developed a checklist to help data providers understand the requirements for privacy-protecting data sharing and to guide them in fulfilling the existing guidelines and the ethics requirements for the specific data supplied. The checklist has general requirements that apply to all data types, and additional requirements specifically for data used in the toxicology and risk assessment fields. The final goal is to allow for an evaluation following the relevant criteria and to generate a statement confirming that the applicable requirements are fulfilled before a dataset can be used.

Since data management and sharing is becoming a more centralised, European or even global effort spanning different infrastructures and disciplines, recommendations and guidelines adopted by OpenRiskNet cannot be developed autonomously and have to be aligned and harmonised with the ongoing discussions on the EU level and with changing regulations. Therefore, this document needs to be considered only as an initial attempt to cover issues in the specific scientific area and it will be provided to working groups established within the governance of the European Open Science cloud (EOSC) on FAIR data requirements. Until additional feedback is collected and other players are consulted (i.e. EOSC) the information included below will not be implemented yet as an online tool linked to the OpenRiskNet infrastructure, but kept as an input document to the e-infrastructure community. Also, this document does not aim to duplicate the existing **Horizon 2020 Ethics checklist**[3], but create a more specific implementation of it updated whenever the underlying legislation changes.

Data submitted to OpenRiskNet however, is still subject to acceptance of the **Privacy policy**[4] and the **Terms of use.**[5]

[1] H2020 Programme - Guidelines on FAIR Data Management in Horizon 2020 https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] https://openrisknet.org/

[3] Horizon 2020 - Ethics Appraisal Procedure and Guidance ‘How to complete your ethics self-assessment’ (Version 6.1, from 4 February 2019) https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm

[4] OpenRiskNet Privacy policy https://openrisknet.org/privacy-policy/ (version from 7 May 2019)

[5] OpenRiskNet Terms of use https://openrisknet.org/terms-of-use/ (version from 7 May 2019)

General requirements

- ❏ Confirm the responsibility in the collection, use and processing of biomaterial and data being provided;
- ❏ Confirm that general principles of ethics at EU or national level are respected;
- ❏ Risk assessment of materials/methods/technologies used has been conducted in compliance with international, EU and national laws addressing concerns relating to potential harm or misuse of materials, technologies and information;
    - *Note: the data provider should specify which regulations were enforced during data generation; this could then be presented to the data user, who would decide if this fulfils their standards for reusing of data.*
- ❏ Are liable regarding quality of information, as well as material, data, products, processes used, non-harmful handling of biomaterial;
- ❏ Have guaranteed no harmful material was collected or processed without the full informed consent of the donor and the partner handling the material;
- ❏ Confirm if data is provided by international publicly funded projects, not-for-profit consortia or governmental and regulatory agencies;
- ❏ Confirm if data is already available under open-data licenses.
- ❏ Transfer of data was done according to international legislation and authorisations.

Additional Requirements for Human Data

Additional requirements for human data are listed below. Please note that for certain types of data/biomaterial, consent and privacy rules differ for data collected/generated after 25 May 2018, as the GDPR[6] adoption. These differences are stated in the requirements for the individual data types.

Basic requirements for ALL types of human data

- ❏ Compliance with EU Council Regulations and Directives as well as other legislation applicable to the Research data provided (see below the respective section);
- ❏ Free and fully informed consent not applicable for commercial cell lines;
- ❏ Relevant ethics approvals were obtained for accreditation / designation / authorisation / licensing for using, processing or collecting the human cells and tissues (if relevant);
- ❏ No personal data[7] on data subject and/or donor of biomaterial is allowed.

[6] General Data Protection Regulation https://gdpr.eu/

[7] What is considered personal data under the EU GDPR? https://gdpr.eu/eu-gdpr-personal-data/ (retrieved 19 Sept 2019)

Further requirements according to type of biomaterial being provided

1. Commercial Cell Lines

- The contractor/data provider must be in compliance with EU Council Regulations and Directives (EC) 2004/23 of 31 March 2014[8], EU Council Directive (EC) 2006/86 24 October 2006 implementing Directive 2004/23/EC[9].
- No additional requirements on top of basic mentioned above.

2. Research Using, Producing or Collecting Human Cells and Tissues (EXCLUDING Embryonic Stem Cells (hESC) and Human-Induced Pluripotent Stem Cell (hIPSC)

❏ The contractor/data provider must be in compliance with EU Council Regulations and Directives (EC) 2004/23 of 31 March 2014[9], EU Council Directive (EC) 2006/86 24 October 2006 implementing Directive 2004/23/EC[10].

❏ Statement confirming that relevant ethics approvals were obtained for accreditation, designation, authorisation and licensing for using, processing or collecting the human cells and tissues (if relevant), in particular:
  - origin of the cells and tissues;
  - tissue establishments and tissue/cell preparation processes;
  - quality management of cells and tissues;
  - detail of legislation under which the material will be stored;
  - duration of use, storage, transfer and purpose of use of cell line;
  - procurement, processing, labelling, packaging, distribution, traceability and imports and exports of cells and tissues from and to third countries.

3. Research using Human Embryonic Stem Cells hESCs and Human-induced Pluripotent Stem Cell (hiPSC) Lines

❏ The contractor/data provider must be in compliance with EU Council Regulations and Directives (EC) 2004/23 of 31 March 2014[9], EU Council Directive (EC) 2006/86 24 October 2006 implementing Directive 2004/23/EC[10], and national law (in particular, the Statement of the Commission related to research activities involving human embryonic stem cells[10]);.

❏ Statement confirming that relevant ethics approvals were obtained for accreditation, designation, authorisation and licensing for using, processing or collecting the human cells and tissues (if relevant).

❏ For research activities involving hESC means that:

  ❏ Cells were NOT derived from embryos specially created for research or by somatic cell nuclear transfer;

  ❏ The project uses existing cultured cell lines only;

[8] Council Directive (EC) 2004/23 of 31 March 2014 on setting standards of quality and safety for the donation, procurement, testing, processing, preservation, storage and distribution of human tissues and cells, OJ L102/48

[9] Council Directive (EC) 2006/86 24 October 2006 implementing Directive 2004/23/EC of the European Parliament and of the Council as regards traceability requirements, notification of serious adverse reactions and events and certain technical requirements for the coding, processing, preservation, storage and distribution of human tissues and cells, OJ L294/32.

[10] Article 19(1) of the Horizon 2020 Framework Programme Regulation (EU) No 1291/2013 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013R1291

    - ❏ The cell lines were derived from supernumerary non-implanted embryos resulting from *in vitro* fertilisation;
    - ❏ Informed consent has been obtained for using donated embryos for the derivation of the cell lines
    - ❏ Personal data and privacy of donors of embryos for the derivation of the cells are protected;
    - ❏ NO financial inducements were provided for the donation of embryos used for derivation of the cell lines.

4. Research using Clinical Trial Data

- ❏ Existence of a clinical trial registry number in one public platform of clinical trials (e.g. ClinicalTrials.gov, EudraCT);
- ❏ Data provider complies with European Commission Ethics requirements[11] in line with national law and ethics committees regarding clinical trials;
- ❏ Compliance with the Declaration of Helsinki[12] and the Oviedo Bioethics Convention[13] ;
- ❏ Respect of EU Regulation No 536/2014 on clinical trials on medicinal products for human use[14];
- ❏ Respect of EU Directive 2005/28/EC of 8 April 2005 on principles and detailed guidelines for good clinical practice investigating medicinal products for human use and manufacturing or importation of such products (OJ L 91, 9.4.2005, p. 13)[15].

Additional Requirements for Animal Data

Compliance with:

- ❏ EU Directive 2010/63/EU[16] and other applicable national and international law guiding use of animals;
    - ❏ The relevant/necessary national authorisations for the supply of animals and animal experiments (and other specific authorisations, if applicable).

[11] Horizon 2020 - Ethics Appraisal Procedure and Guidance 'How to complete your ethics self-assessment' (Version 6.1, from 4 February 2019) https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm

[12] World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013 Nov 27; 310(20): 2191–2194. doi: 10.1001/jama.2013.281053

[13] The Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine (ETS No 164), 4 April 1997 in Oviedo (Spain) https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/164 / https://www.coe.int/en/web/bioethics/oviedo-convention

[14] Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX:32014R0536 / https://ec.europa.eu/health/human-use/clinical-trials/regulation_en

[15] EU Directive 2005/28/EC of 8 April 2005 laying down principles and detailed guidelines for good clinical practice as regards investigational medicinal products for human use as well as the requirements for authorization of the manufacturing or importation of such products (OJ L 91, 9.4.2005, p. 13) https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/dir_2005_28/dir_2005_28_en.pdf

[16] EU Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes (OJ L 276, 20.10.2010, p. 33).

OpenRiskNet Privacy Policy

Effective date: 20 November 2019

Website content disclaimer

The information contained on https://openrisknet.org/ website (the "Service") is for general information purposes only.

OpenRiskNet consortium (“OpenRiskNet”, the “author”, "us", "we", or "our") maintains this website to enhance public access to information about the project and its outcomes. Our goal is to keep this information timely and accurate. If errors are brought to our attention, we will correct them as soon as possible. However, we assume no responsibility for errors or omissions in the content on the Service.

This information about the project and its outcomes is

- of a general nature only and not intended to address the specific circumstances of any particular individual or entity;
- not necessarily comprehensive, complete, accurate or up to date;
- sometimes linking to external sites over which we have no control and for which we assume no responsibility.

In no event shall OpenRiskNet be liable for any special, direct, indirect, consequential, or incidental damages or any damages whatsoever. OpenRiskNet reserves the right to make additions, deletions, or modifications to the content on the Service at any time without prior notice.

OpenRiskNet has appropriate technical and organisational measures to ensure a level of security appropriate to the risk.

External links disclaimer

The Service may contain links to external websites that are not provided or maintained by or in any way affiliated with OpenRiskNet.

Please note that OpenRiskNet does not guarantee the accuracy, relevance, timeliness, or completeness of any information on these external websites. OpenRiskNet does not warrant that these external websites are free of viruses or other harmful components.

Copyright and acknowledgement of sources

The author aims to observe the copyright of any graphics, audio documents, video sequence or text in all publications, to use his/her own graphics, audio documents, video sequences or texts or to make use of license free graphics, audio documents, video sequences or texts.

All trademarks and brands mentioned on the website, including those protected by third parties, are without limitation subject to the provisions under the respective labelling law and the rights of the copyright holder. The sole mentioning of a trade mark on this website should not lead to the assumption that it is not protected by the rights of a third party.

The author of the website has the exclusive copyright to all published objects created by him-/herself. The reproduction or use of any such graphics, audio documents, video sequences or texts in other electronic or printed publications is allowed under the Creative Commons License Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

The licenses of each application, tool, dataset or dissemination material (e.g. publication) part of the OpenRiskNet e-infrastructure are mentioned elsewhere and included within their own description. These individual licences needs to be considered when the applications, tools, datasets or dissemination materials are used or shared.

Data Protection and Privacy Policy

OpenRiskNet is committed to user privacy and complies with its obligations under the GDPR. The goal of this privacy policy is to help you understand how OpenRiskNet, as data controller, deals with any personal data you provide to us. In the context of the relationship with the Data Subject, OpenRiskNet ensures that Personal Data will be processed in a lawfully, fairly and transparent manner and that only process Personal Data that are adequate, relevant and limited to what is strictly necessary for the purposes for which they are processed.

Processing your personal data under your consent is necessary for our legitimate interest of allowing the day-to-day management, operation and functioning. On some pages (e.g. catalogue of services and service description, dissemination and training materials, etc.) you have the possibility to enter personal or business data. The disclosure of this data is voluntary. If technically feasible and where reasonable, all services offered can be used without disclosing personal information or by use of anonymised data or aliases. The information provided in forms, surveys or questionnaires including the personal data will only be made available to the full partners of the OpenRiskNet consortium and will only be used to define the requirements or select services for the OpenRiskNet e-infrastructure.

Responsibility for the processing of Personal Data

We collect several different types of information (see also Categories of Personal Data processed by OpenRiskNet) for various purposes:

- To provide and maintain the Service
- To notify you about changes to our Service
- To allow you to participate in interactive features of our Service when you choose to do so
- To provide customer support
- To provide analysis or valuable information so that we can improve the Service
- To monitor the usage of the Service
- To detect, prevent and address technical issues

Categories of Personal Data processed by OpenRiskNet

While using our Service, we may ask you to provide us with certain personally identifiable information that can be used to contact or identify you ("Personal Data"). Personally identifiable information may include, but is not limited to:

- Identification data
- Personal contact details
- Access to the website data
- Data on preferences
- Data on the use of information technology

We may also collect information about how the Service is accessed and used ("Usage Data"). This Usage Data may include information such as your computer's Internet Protocol address (e.g. IP address), browser type, browser version, the pages of our Service that you visit, the time and date of your visit, the time spent on those pages, unique device identifiers and other diagnostic data.

Special categories of personal data will be uploaded anonymised by data provider.

Storage of Personal Data

The criteria used to determine the period of storage of Personal Data is the respective statutory retention period. After expiration of that period, the corresponding data is routinely deleted, as long as it is no longer necessary for the fulfillment of the contract or the initiation of a contract.

However, OpenRiskNet may be obliged to storage some personal Data for a longer period, taking into account factors such as:

- Legal obligations, under current laws, to keep personal data for a certain period;
- Limitation periods, under the laws in force;
- Judicial and administrative Proceedings and Procedures;
- Guidelines issued by the data protection supervisory authorities;

During the processing period, OpenRiskNet guarantees that Personal Data is processed in accordance with this Data Protection and Privacy Policy. Once the personal data is no longer necessary, OpenRiskNet will proceed to its erasure in a safe way.

Sharing of Personal Data

Personal Data may be transferred to - and maintained on - computers located outside of your state, province, country or other governmental jurisdiction where the data protection laws may differ than those from your jurisdiction.

Your consent to this Privacy Policy followed by your submission of such information represents your agreement to that transfer.

OpenRiskNet will take all steps reasonably necessary to ensure that your data is treated securely and in accordance with this Privacy Policy and no transfer of your Personal Data will take place to an organization or a country unless there are adequate controls in place including the security of your data and other personal information.

OpenRiskNet may disclose your Personal Data in the good faith belief that such action is necessary

- to comply with a legal obligation;
- to protect and defend the rights or property of OpenRiskNet;

- to prevent or investigate possible wrongdoing in connection with the Service;
- to protect the personal safety of users of the Service or the public; and
- to protect against legal liability.

Data Protection Rights under the General Data Protection Regulation (GDPR)

If you are a resident of the European Economic Area (EEA), you have certain data protection rights. If you want to exercise your data subject rights please contact us by email. In your email, clearly state your request and include the URL of the website/webpages your request refers to. Please note that we may ask you to verify your identity before responding to such requests.

You have the following data protection rights:

- The right to access, update or delete the information we have on you: whenever made possible, you can access, update or request deletion of your Personal Data directly within your account settings section. If you are unable to perform these actions yourself, please contact us by email to assist you;
- The right of rectification: You have the right to have your information rectified if that information is inaccurate or incomplete;
- The right to object: You have the right to object to our processing of your Personal Data;
- The right of restriction: You have the right to request that we restrict the processing of your personal information;
- The right to data portability: You have the right to be provided with a copy of the information we have on you in a structured, machine-readable and commonly used format;
- The right to withdraw consent: You also have the right to withdraw your consent at any time where OpenRiskNet relied on your consent to process your personal information.

Please note that the above rights, particularly the deletion, are only available whenever the processing of your personal data is not necessary to:

- Comply with a legal obligation;
- Perform a task carried out in the public interest;
- Exercise authority as a data controller;
- Archive for purposes in the public interest, or for historical research purposes, or for statistical purposes;
- Establish, exercise or defend legal claims.

You have the right to complain to a Data Protection Authority about our collection and use of your Personal Data. For more information, please contact your local data protection authority in the European Economic Area (EEA).

Service Providers

We may employ third party companies and individuals to facilitate our Service ("Service Providers"), to provide the Service on our behalf, to perform Service-related services or to assist us in analysing how our Service is used.

These third parties have access to your Personal Data only to perform these tasks on our behalf and are obligated not to disclose or use it for any other purpose.

The login to OpenRiskNet reference site is using social authentication Service Providers for managing logins. Currently the following providers are supported:

- LinkedIn
- GitHub

Using a social authentication provider means that we never see your password. You authenticate with the social provider and if successful they forward you back to the OpenRiskNet website. We only store minimal information about you: name and email.

Analytics

We may use third-party Service Providers to monitor and analyze the use of our Service:

- Google Analytics

Google Analytics is a web analytics service offered by Google that tracks and reports website traffic. Google uses the data collected to track and monitor the use of our Service. This data is shared with other Google services. Google may use the collected data to contextualize and personalize the ads of its own advertising network.

You can opt-out of having made your activity on the Service available to Google Analytics by installing the Google Analytics opt-out browser add-on. The add-on prevents the Google Analytics JavaScript (ga.js, analytics.js, and dc.js) from sharing information with Google Analytics about visits activity.

For more information on the privacy practices of Google, please visit Google Privacy & Terms web page.

Security and Integrity

Personal data will be treated by OpenRiskNet only in the context of the purposes identified in this Policy, in accordance with the internal policies of OpenRiskNet and using technical and organizational measures designed according to the risks associated with the specific treatment of Personal Data. The technical and organizational measures designed to ensure, to the maximum extent possible, the security and integrity of Personal Data, in particular in relation to unauthorized or unlawful treatment and its accidental loss, destruction or damage.

Cookies

We use cookies and similar tracking technologies to track the activity on our Service and hold certain information.
Cookies are files with small amount of data which may include an anonymous unique identifier. Cookies are sent to your browser from a website and stored on your device. Tracking technologies also used are beacons, tags, and scripts to collect and track information and to improve and analyze our Service.
You can instruct your browser to refuse all cookies or to indicate when a cookie is being sent. However, if you do not accept cookies, you may not be able to use some portions of our Service.

Examples of Cookies we use:

- Session Cookies. We use Session Cookies to operate our Service.

- Preference Cookies. We use Preference Cookies to remember your preferences and various settings.
- Security Cookies. We use Security Cookies for security purposes.

SSL or TSL encryption

This site uses SSL or TLS encryption for security reasons and for the protection of the transmission of confidential content, such as the inquiries you send to us as the site operator. You can recognize an encrypted connection in your browser's address line when it changes from "http://" to "https://" and the lock icon is displayed in your browser's address bar.

If SSL or TLS encryption is activated, the data you transfer to us cannot be read by third parties.

Email communication

Security gaps can occur in email communication, if the connection is not encrypted. An email sent to a recipient can be intercepted and read by experienced Internet users. Emails are received by the Coordinator Office at Edelweiss Connect which processes the messages on behalf of OpenRiskNet. If you send an email to the Coordinator Office, we assume that the staff is authorised to reply by email. If you do not wish to receive an email, we kindly ask you to consider alternative ways of communication.

Changes of this Data Protection and Privacy Policy

We may update our Privacy Policy from time to time. We will notify you of any changes by posting the new Privacy Policy on this page.

We will let you know via a prominent notice on our Service, prior to the change becoming effective and update the "effective date" at the top of this Privacy Policy.

You are advised to review this Privacy Policy periodically for any changes. Changes to this Privacy Policy are effective when they are posted on this page.

Legal effect of disclaimer

This disclaimer is part of the website linked to this page. If parts of this text or certain wordings are not, no longer or not completely in line with current legislation, it will not prejudice the rest of the document in terms of content or validity.

Contact Us

If you have any questions about this Privacy Policy or any other issue related to personal data protection, please contact by email the Coordinator Office and Data Protection Officer of OpenRiskNet.

OpenRiskNet Terms of use

Effective date: 20 November 2019

OpenRiskNet is a project funded by the European Commission within Horizon 2020 EINFRA-22-2016 Programme (Project number 731075) aiming to develop an open e-infrastructure providing resources and services to a variety of scientific communities requiring risk assessment.

Definition of terms used in this document

Processing = means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

Controller = means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.

Anonymisation = the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject. Note: data uploaded in OpenRiskNet are fully anonymous to the researchers who receive or use them and also to data provider, who doesn't have the link back to identify individuals.

Pseudoanonymisation = the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. Note: data uploaded in OpenRiskNet are fully anonymous to the researchers who receive or use them, but contain information or codes that would allow others (e.g. data provider) to link them back to identifiable individuals.

Data = the term 'data' in this document may refer to biological data, toxicity data, genomic data, anonymised images, metadata, etc. It does not refer to data that contains identifiable information such as name, phone number, or date of birth.

Personal Data = any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Data Owner = the ‘data owner’ is the individual researcher or investigator or body of researchers or investigators that produced the original data (‘data controller’). The ‘data owner’ has the responsibility of authorising further use of the data.

Data Provider = the ‘data provider’ is the individual researcher or investigator or body of researchers or investigators that makes data available for access and use within the OpenRiskNet e-infrastructure.

Data User = the ‘data user’ is the individual researcher or investigator or body of researchers that processes data through the OpenRiskNet e-infrastructure.

About the OpenRiskNet e-infrastructure

OpenRiskNet is an e-infrastructure for the harmonisation and improved interoperability of data and software tools in predictive toxicology and risk assessment. It aims at supporting safe-by-design product development and risk assessment of drugs, chemicals, cosmetic products and nano materials by integrating existing toxicology databases and in silico tools and combine them to workflows for predicting hazard, exposure and risk.

It combines:

- Web services providing data or analysis, processing and modelling tools communicating over well-defined and harmonized application programming interfaces (APIs);
- An interoperability concept and framework for general and specific services integration by consortium members and associated partners;
- A featured platform for predictive toxicology and risk assessment for end users (e.g. toxicologists, risk assessors and regulators using case studies).

OpenRiskNet e-infrastructure includes data management systems that make available existing and open data sources mainly from *in vitro* human and animal and *in vivo* animal experiments to all stakeholders in a harmonised way.

OpenRiskNet e-infrastructure commitment and privacy policy strive for making computational tools and data as accessible as possible to the scientific community, while protecting the interests of participants from whom the data originate with regard to Ethical, Legal and Social Implications (ELSI) and within the scope of their consent. These Terms of Use reflects OpenRiskNet commitment to provide this service and impose no additional constraints on the use and transfer of the contributed data than those authorised by the data owner and provided by the data provider.

Data providers and data users of OpenRiskNet e-infrastructure

Data providers and users of OpenRiskNet e-infrastructure have a number of responsibilities and obligations, such as the obligation to respect participant confidentiality. Researchers and data managers accessing the data and even more providing data as an OpenRiskNet service have a custodian role, to ensure the careful and responsible management of the information. They have an obligation to operate in conformity with the requirements of their own institution, and fulfil all necessary national and international regulatory and ethical requirements during data generation (*in vivo*, *in vitro* and *in silico*), preparation for sharing and ongoing management of the resources. They also have obligations to the OpenRiskNet e-infrastructure, the integrity of their own research, as well as the funders and the wider research community, to carry out high quality, ethical research.

Use of OpenRiskNet e-infrastructure

❏ All users have an obligation of confidentiality and must conform to the data protection principles to ensure that data is processed in compliance with the legal and ethical requirements.

❏ The data owners must ensure that they have sought and obtained, where necessary, all appropriate approvals, ethical and legal, for the data collected. OpenRiskNet will provide information on the approval procedure and status, whenever this is provided by the data owner or data provider and collection is technical feasible. Listing of the information does not imply that OpenRiskNet guarantees the accuracy of any provided information.

❏ For *in vivo* animal data, the data owner must ensure that national guidelines for their welfare and care during the collection of data have been followed.

❏ OpenRiskNet does not guarantee the accuracy of any provided data.

❏ OpenRiskNet has implemented appropriate technical and organisational measures to ensure a level of security which we deem appropriate, taking into account the sensitivity of data we handle. However, the data provider holds sole responsibility for the usage and distribution of data.

❏ OpenRiskNet requires all data provided or used in the OpenRiskNet infrastructure being anonymised before submission.

❏ Computing of personal and sensitive data on OpenRiskNet e-infrastructure should be run internally by the users on their secure cloud infrastructures under appropriate firewalls. OpenRiskNet will not hold any liability for any loss or damage to data.

❏ While we will retain our commitment to the privacy of sensitive data, we reserve the right to update these Terms of Use at any time. When alterations are inevitable, we will let you know by placing a notice on our website and update the "effective date" at the top of this Terms of Use, but you may wish to check each time you use the website. The date of the most recent revision will appear on the ‘OpenRiskNet e-infrastructure terms of use’ page. If you do not agree to these changes, please do not continue to use our services. We will also make available an archived copy of the previous Terms of Use for comparison.

❏ Any questions or comments concerning these Terms of Use can be addressed to: openrisknet@edelweissconnect.com.

Confirmation of Acceptance of the terms of use

By ticking the box, data providers and users certify that they will abide by this Ethical Governance Framework, Terms of Use and its stipulations, and that appropriate ethical approval and/or consent are in place prior to use of the data within the project. The acceptance of these conditions along with other registration data will be collected by the project coordinator and stored centrally.