OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Deliverable Report D3.1

Initial version of data management 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	Douglas Connect GmbH
Start date	1 December 2016
End date	30 November 2019
Duration	36 Months
Project Partners	P1 Douglas 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 3.1 Initial version of data management plan
Deliverable Type	ORDP: Open Research Data Pilot
Dissemination Level	Public (PU)
Work Package	WP3
Task(s)	Task 3.5 Dissemination, Exploitation, Data Management and Sustainability Plan
Deliverable lead partner	DC
Author(s)	Lucian Farcal (DC), Oana Florean (DC), Philip Doganis (NTUA), Danyel Jennen (UM), Egon Willighagen (UM), Marvin Martens (UM), Daniel Bachler (DC), Noffisat Oki (DC), Thomas Exner (DC)
Status	Final
Version	V2.0
Document history	2017-03-17 First draft 2017-04-19 Consolidated draft 2017-05-31 Final version 1.0 2018-11-30 Version 2.0

History of revisions - Version 2.0

Description	Section
Major changes to the DMP were done at M24, considering the latest developments and tasks that involved data management, but also the feedback received on the first version. Detailed information and examples were added to the relevant chapters. DMP version 2 covers the general aspects of the OpenRiskNet data management based on the FAIR guidelines, ethics considerations for re-sharing of public datasets but also the first examples and specific information on individual shared data sources: diXa, BridgeDb, WikiPathways, AOP-Wiki and ToxCast. These include specific aspects depending on the data sources.	Summary, Introduction and the core part of the Data Management Plan

Four annexes were added, generated within the tasks and discussions related with ethics: 1) Ethics requirement for OpenRiskNet infrastructure data providers 2) Personal data protection and privacy policy 3) OpenRiskNet e-infrastructure terms of use 4) Ethical issues from the diXa project	Annexes

Table of Contents

SUMMARY

This report describes the first updated version of the data management plan (DMP) for the OpenRiskNet e-infrastructure projects. The current DMP covers the general 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 the first examples of shared data sources including diXa, BridgeDb, WikiPathways, AOP-Wiki and ToxCast/Tox21. More specific data source and clearly-defined measures will be added in parallel to their integration into the infrastructure, which will follow the time plan enforced by the case study requirements on data availability.

INTRODUCTION

The European Commission is running a flexible pilot under Horizon 2020 called the Open Research Data Pilot (ORD Pilot). The ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects 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 [1]. 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 [2]. 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,

To help optimising the potential for future sharing and re-use of data, the OpenRiskNet Data Management Plan (DMP) helps the partners to consider any problems or challenges that may be encountered and helps them to identify ways to overcome these. This DMP is a “living” document developed using the online tool described above and given as a snapshot of the current status (November 2018) in this document, 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, are handled during and after a research project. It follows the Guidelines on FAIR Data Management in Horizon 2020 [1] and is based around the resources available to the project partners in a realistic way taking the current knowledge into account. The ongoing activities to keep the DMP up to date follow an online, distributed approach as outlined in the Guidelines for creating an online DMP (see Figure 1) [3]. Here, we summarise the concepts for the description of data sets as well as data sharing and archiving approaches adopted in the DMP first followed by the DMP in relevant version of the time of this writing. Since the OpenRiskNet case studies are under ongoing development, which define the integrated data sources, the current DMP covers the general aspects of the OpenRiskNet data management but also specific and clearly-defined measures for the

first data sources integrated. Additional sources will be added in parallel to the data integration.

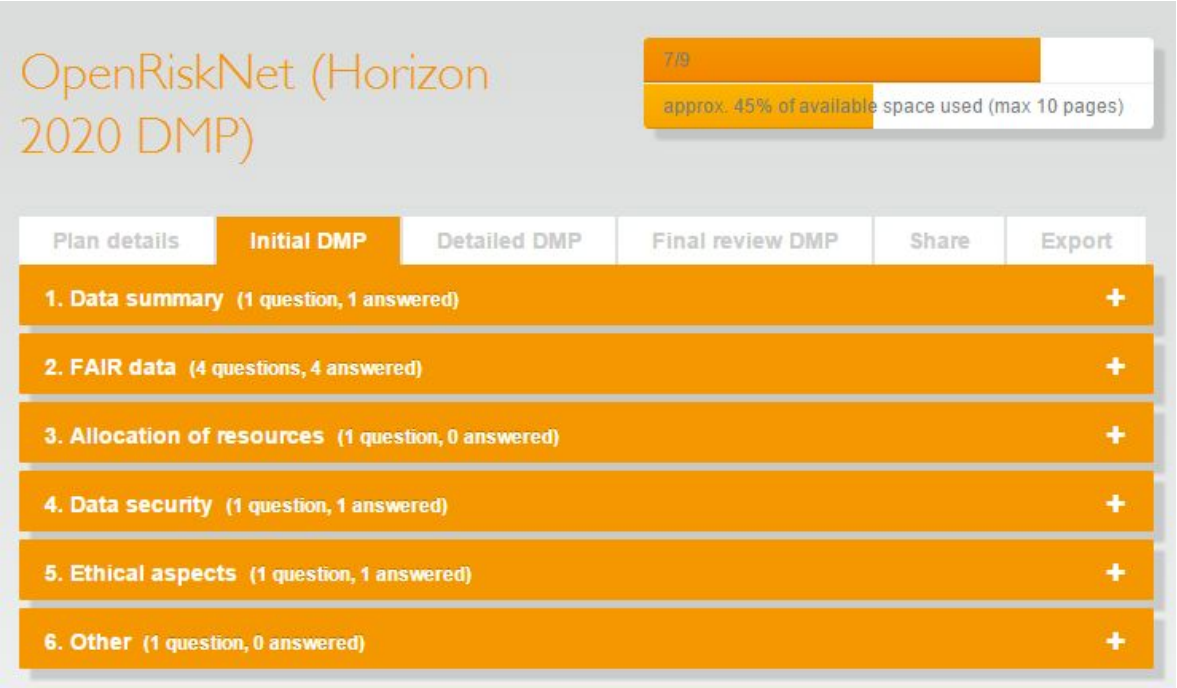


Figure 1. DMP tool interface used to create the OpenRiskNet plan [4]

DATA SET DESCRIPTION

This section give the general concepts of what is considered data in OpenRiskNet and a listing of what kind of datal the project collects, makes available for redistribution and sharing oor generates as part of the case studies (*in silico* only), and to whom might they 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 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);
- More specifically, data and metadata 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 allow to answer questions to enable data to be found and understood, ideally according to the particular standards applied. Such questions include but are not limited to:

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

Finally the **metadata**, **documentation** and **standards** will help in making the data FAIR (Findable, Accessible, Interoperable and Re-usable) by not only provide the technical requirements like global, persistent identifier and clear access protocols (data application programming interfaces (APIs)) and licenses but by harmonizing and improving the scientific interoperability of the data by semantic annotations and allowing combination and enrichment of data sets using linked-data approaches (Combined OpenAPI and JSON-LD description of data APIs).
Data integrated, for which the integration is in progress or produced 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
- Intermediate or final results of *in silico* studies performed as part of the case studies

DATA SHARING

According to the ORD Pilot programme, by default as much of the resulting data as possible should be archived as Open Access. Most data handled in OpenRiskNet is 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 and OpenRiskNet is fully committed to make 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) might put some restriction on the sharing of data generated in these collaborations. Such legitimate reasons for not sharing resulting data will be explained in the DMP in the rare cases they have to be applied. Additionally, OpenRiskNet is committed to protect personal data and IPR agreements and to responsible data sharing and is taking 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 will be transferred to an organization or a country unless there are adequate controls in place including the security of data and personal information. Complementing these general data sharing policies, the DMP describes any ethical or legal issues that can have an impact on data sharing. Since in many cases, the data production is not under control of the OpenRiskNet partners and is 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 is working together with the original data providers as well as ethical experts on producing workflows and checklist (see attachments) 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 agreed on by the data provider and technical feasible, this data is 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 is 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 has 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 is build upon software engineering and infrastructure components developed, supported and adopted by a large community guaranteeing, on one hand, some 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. diXa and ToxCast are big European infrastructures or US agencies, ELIXIR and US EPA, respectively. For these, 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 is 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. Negotiations with the Birmingham Environment for Academic Research (BEAR) are underway to provide archiving and preservation facilities for containerised data and software services for at least the next 5 years, for which this cannot be guaranteed by the service provider.

DATA MANAGEMENT PLAN (DMP)

This data management plan will address all data-related problems or challenges that may be 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 will be 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 is to fulfill its main objectives in providing and improving solutions on data availability to the 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;
- Low quality 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 is to generate guidelines and templates for data exchange, harmonise the 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 aims to establish the 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 more accurate evaluation of toxicity;

- Data infrastructure will offer 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 as well as to export options into file formats like ISA, ToxML and OECD harmonised templates.

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:

- The models will permit identification of mechanistic links between omics data at different levels of functional organisation;
- The models will help to advance the understanding of the relationship between toxicity, architecture, function and risk;
- Computational sensitivity analyses components will aid 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 are 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 is structured around the concept of semantic-annotated application programming interfaces, which will also 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 will 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. Additionally, to be able to integrate legacy data and provide the final results in the format required for e.g. regulatory reporting, OpenRiskNet is supporting import and export of standard file formats like ISA (-tab or -json), ToxML and OECD harmonised templates. However, such technical file format conversion solutions require a scientific harmonization of the metadata completeness and data description levels, which are better defined for summary data used for regulatory purpose than for raw data, where the reporting style is very dependent on the individual data provider and application area. Therefore, we are working together with other big projects (EU-ToxRisk

and NanoCommons) to define the amount and content of metadata, which has to be provided for each experimental assay or computational investigation, and data formats if no standards exist so far as well as providing the means for future additions and adaptations based on flexible data schema specifications. 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

For data and computational models, we use, as much as possible, existing data, software tools, open and readily available to all partners. We will aim at re-usable and extensible tools. OpenRiskNet is not producing any new experimental data but is reusing data from publicly available data sources or if absolutely needed for interaction with an associated partner or a case study, data from other projects not yet under an open license. It is the clear 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 provide it in a easy way for reuse by others. On one hand, making sharing, accessing and reusing of data easier is the main goal of 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 is, therefore, another central goal of the OpenRiskNet infrastructure.

1.5 Origin of the data

As mentioned before at many places, 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 and 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 Partners 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 used 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 as to be considered and also 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 is integrated in OpenRiskNet infrastructure, which also handles the license management. In the same way, commercial software or free software requiring a registration is handled.

1.6 Expected size of the data

As described above the idea of the OpenRiskNet infrastructure is 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 data is needed. However, two aims of the OpenRiskNet project might lead to additional requirements on data storage.

1) Some of the data sources considered for integration are not yet available in open-accessible databases, these cannot be accessed via application programming interfaces or don't comply with the FAIR principles. In such cases, OpenRiskNet will negotiate with the data owners if the data should be either transferred to standard data repositories or if the existing solution should be improved within the framework of the associated partner programme.
2) Most if not all of the data sources should also be provided in a form suitable for in-house deployment. Even if the user system administrator setting up the in-house virtual research environment (VRE) 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. To assess the needed storage space for the containers and to give the users guidance on the needed computational resources for the VRE, expected sizes for all data services are given in section 1.8 below.

1.7 Utility of data and models

OpenRiskNet solutions will work towards making 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 safeness of a chemical, drug, cosmetic ingredient and nanomaterial on all 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 industry has the merit of providing a wide spectrum of data, with which industry could 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 also key to push the regulatory acceptance of the integrated tools and workflows.

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

- Industry represented by chemicals, pharma, food, cosmetics or other consumer products companies which are required 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: 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: OpenRiskNet aims to support integration of apps that can be used by

consumers on their mobile phones supporting everyday activities, such as obtaining knowledge on ingredients in products they are purchasing or using.

1.8 Specific information on individual shared data sources

In this section, we will address specific information and requirements of individual data sources provided by OpenRiskNet with respect to their purpose, origin, relationship to the project, data type and format, size and potential users. Section 2.5 and 5.3 below are fulfilling the same purpose for issues on FAIR data sharing and ethics. These are meant as additions or clarifications to the general descriptions relevant only for the specific data set / database. Points completely covered by the general remarks will not be repeated here again and thus some of the subsections above will not appear in the database descriptions

1.8.1 diXa

Like OpenRiskNet, diXa was an e-infrastructure project for collecting data from different research projects and make them publicly available via a common interface. Thus, its goals and objectives fit those of the OpenRiskNet project. As part of the integration of diXa Data Warehouse into the OpenRiskNet infrastructure, the data access will be semantically annotated and further harmonised to other OpenRiskNet services.

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

The Data Infrastructure for Chemical Safety (diXa) project (http://www.dixa-fp7.eu/) was funded by EU FP7 to provide a single resource for the capture of toxicogenomics data produced by past, present and future EU research projects, and to ensure sustainability of such a resource for use by the wider research community. Therefore, the diXa Data Warehouse was established (http://wwwdev.ebi.ac.uk/fg/dixa/index.html).

Data from the diXa Data Warehouse as well as other sources (i.e. NCBI GEO and EBI ArrayExpress) is currently being used in a meta-analysis for genotoxicity prediction using data from multiple *in vitro* cell models as part of the TGX case study. The results using only the human data have been presented as poster at EUROTOX 2018 (https://doi.org/10.1016/j.toxlet.2018.06.608).

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 diXa Data Warehouse comprises of 95 studies, 29609 samples and 469 compounds (including solvents). *In vitro* human and rodent data and *in vivo* rodent data were collected from the EU FP6 projects carcinoGENOMICS, PredTox, NewGeneris, Predictomics, the EU FP7 projects ESNATS, Predict-iv, the Dutch project of the Netherlands Toxicogenomics Centre, the Japanese project Open TG-GATEs and the US project DrugMatrix. Most studies consist of transcriptomics data, whereas a few also contain metabolomics and/or proteomics data. It should be noted that the 2 *in vivo* human studies of the EU FP7 Envirogenomarkers project do not contain data as these have been retracted based on objections from the Swedish biobank with regard to personal data protection.

In addition, 188 human disease transcriptomics data sets have been added to the data warehouse.

Metadata for all studies and disease data sets are captured in the ISA-tab format.

In addition to this 'omics data collection links to other globally-available chemical/toxicological databases were provided.

The diXa Data Warehouse has been further used in the EU FP7 project HeCaToS coordinated by UM. In this project the data warehouse has become part of EBI's BioStudies (https://www.ebi.ac.uk/biostudies/) and the data generated in HeCaToS were directly uploaded to BioStudies. Upon public release of these data they can also be used within OpenRiskNet.

1.8.1.3 Additional details to 1.6 Expected size of the data

The raw data of the ~30,000 samples are at least 400 GB in size.

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 [5]. BridgeDb is used in smaller and larger projects, the latter including WikiPathways, Cytoscape and Open PHACTS [6]. 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.

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 2018) does not exceed 12Mb.

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 ToxCast

The United States Environmental Protection Agency Toxicity forecaster (ToxCast)[1] has generated toxicity screening data on thousands of chemicals in commerce and of interest to the agency and general public. The project also uses computational approaches to prioritise and rank chemicals for risk assessments and regulatory decision making. The data is publicly available, widely distributed and can be annotated to fit into the OpenRiskNet data harmonisation and integration framework.

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 ToxCast research project data is generated on high-throughput in vitro toxicity screens for a variety of chemicals and biological targets. One of the goals of the project is to prioritise and evaluate the potential human health risk of chemicals in a cost efficient way. The data generated also includes computational and predictive models to predict toxicity potential of the chemicals in humans. The results of these analysis are being used actively to inform the context of decision making such as endocrine disruptor screening. The data can be integrated with other data services of the OpenRiskNet infrastructure such as ontology mapping, pathway identification and mapping, and AOP development tools that. Users of the OpenRiskNet service will be 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. The US EPA is also making the data publicly available and accessible through various means. As the project progresses, and more chemicals are screened, the agency makes periodic updates to the public release as more data is generated.

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 data currently available on the ToxCast dashboard includes over 9076 chemicals tested in 1192 assays (as at November 26, 2018) that map to hundreds of genes 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 the 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 ToxCast dashboard web application which serves as a portal for users to search and query the data. No personal data is collected during the data generation process as all data in this set is in vitro and commercial cell lines are used for

[1] https://www.epa.gov/chemical-research/toxicity-forecasting

the testing.

1.8.5.3 Additional details to 1.6 Expected size of the data

The current version of the invitroDB database download is less than 8GB. The accompanying summary files and analysis pipeline and concentration response plots have a cumulative size of less than 30GB.

1.8.5.4 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.

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 is integrating existing data sources and make them easier findable, accessible and interoperable. This is based, on one hand, on the metadata provided by the data sources and, on the other hand, on the interoperability layer, which harmonises these metadata into 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 allows 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 features required by the FAIR principles as e.g. unique identifiers, the interoperability layer added to the service in the context of OpenRiskNet can add these features and in this way improve the 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 will enable the early 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 first stable form and updates will be made available throughout the project. In the prioritization of services to be integrated, open source tools are favoured for the use in the case studies and the reference workflows targeting specific question but commercial services will be equally important to sustain the infrastructure in the long run. However, also these commercial services are required to openly share their API definitions and data formats to allow for integration and combination with other tools.

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 in developing open standards for predictive toxicology resources;
- Tools to access study data and metadata description in standards file formats such as ISA, 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 does not propose to create new file standards rather to employ the 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)** [7] and **Standard for Exchange of Nonclinical Data (SEND)** [8] are considered. 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 OpenRiskNet partners developing the guidelines are including as much information needed for these reports as possible 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 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 is working with experts in the field to integrate such tools in the infrastructure.
- Allowing mapping between related items in different database (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 already 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 diXa and ToxBank are used to create templates for data storage and sharing;
- Data schemata for different endpoints as already available in file formats like ISA and ToxML will be transformed into more flexible data transfer approach able to accompany modifications needed due to changed and enhanced experimental and computational protocols.

An important part of these procedure 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 will be accessible by the OpenRiskNet APIs (similar like the computational tools) including the interoperability layer;
- Searches throughout multiple databases will be 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.

OpenRiskNet provides to its potential data managers or developers complete control over the provided data and associated functionalities but requires from them to describe the interfaces and transfer formats in a generally understandable OpenRiskNet-compliant way through the interoperability layer. This work is based on and extends:
- 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.

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 re-use. 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 useable 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 will redistribute 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 will be made available through the OpenRiskNet access methods as soon as it is released by the original data provider, i.e. no additional embargo period will be 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 design 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 aims to improve 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 manual curation of datasets. The modified dataset are stored (similar to the pre-reasoned datasets) in the original databases as a new version of the dataset or in other OpenRiskNet-compliant databases with a link to the original source.

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 are enforced by the consortium and appropriate measures of uncertainty are requested for all models.

2.5 Specific information on individual shared data sources

2.5.1 diXa

The diXa Data Warehouse is open access. Data retrieval as well as upload takes place via the diXa Data Warehouse webpage http://wwwdev.ebi.ac.uk/fg/dixa/index.html using the ISA-tab format for the metadata. There is no API available for diXa.

Since the diXa Data Warehouse has become part of BioStudies diXa's data can also be accessed here. Data uploaded via BioStudies using the PageTab format (Page layout Tabulation format) will be part of BioStudies and therefore, will not be visible via the diXa Data Warehouse. The same is true the other way around.

BioStudies also provides other formats, such as JSON and XML. Furthermore, BioStudies is part of the ELIXIR infrastructure and an Rest API[2] is available.

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	AOP-Wiki
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

[2] https://github.com/EBIBioStudies/ribs/tree/master/src/main/java/uk/ac/ebi/biostudies/api

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 [9] by AOP-Wiki and WikiPathways. Score meanings: 1 = partial compliance, 2 = compliance

2.5.5 ToxCast

All data produced by the U.S EPA including ToxCast 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 used in OpenRiskNet is a REST API developed by Douglas Connect of the official summary data release made available as downloadable CSV files. These were restructured to fit the OpenRiskNet concept and to make them fit for semantic annotation.

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 is a central task of the integration of data source in the OpenRiskNet infrastructure. Many of the sources considered for integration already follow the FAIR principles and no additional cost are foreseen other than 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 covers the costs of the integration from their allocated budget. In the case of third-party data sources, the integration will be performed in collaboration of the associated partners partly financially supported through the Implementation Challenge and 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 includes:

- 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 is 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 is to bring the virtual environment and all data sources behind a company's firewall by in-house deployment.

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 submitted 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

And the two future deliverables:

- 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 first report of the Data Protection Officer summarizing the relevant regulations and legislation for the different data types and test models, OpenRiskNet has worked together with an external expert to develop elements on an ethics review framework and ethics requirements checklist, which have to be considered in the project and the infrastructure implementation before integrating a data source.

One important component for working on the project case studies and the future usage of the e-infrastructure is access to existing data provided by other European projects in this field or from international consortia. The OpenRiskNet platform is based on a data management system that will make existing data sources mainly from *in vitro* human and animal and *in vivo* animal experiments available to all stakeholders in a harmonised way. In this context, one important additional task of the DMP, becoming even more relevant with the new EU General Data Protection Regulation (GDPR) in effect since May 2018, is to support and give recommendations in achieving the highest impact without jeopardising the ethics integrity of the OpenRiskNet infrastructure.

To fulfil the requirements from the ethics review, a step-by-step decision process was developed addressing how important legacy data sources need to be handled by the project. On top of the workflows provided in the first review 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 sources considered:

I. Human Biomaterial Use, Collection or Storage (Donors)
II. Primary Results Data processing (Clinics)

III. a) Compound Storage/Processing (Commercial cell lines data providers), and b) Secondary Results Processing (Experimentalists)
IV. Tertiary Use - Storage/Provider - No Processing (Database)

Based on the assessment of each of these categories, a checklist is proposed for each data source to be included and/or used in the OpenRiskNet platform (**Annex 1**), and 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 will provide the framework for the ethics evaluation described above to all providers of data services (OpenRiskNet internal, associated partners and other third-parties) and will support the execution and documentation of the data source evaluation and will publish the results on the OpenRiskNet service catalogue together with the service descriptions as a certificate for best practice in including ethical aspects in data management.

5.1 Privacy Policy

A privacy policy[3] was implemented 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
    - Types of Data Collected
    - SSL or TSL encryption
    - E-mail Communication
    - Service Providers
    - Analytics
    - Verifying, modifying or deleting information
- Legal effect of disclaimer
- Changes To This Privacy Policy

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

5.2 Terms of use

The terms of use[4] regulate the use of the OpenRiskNet infrastructure. Most relevant for the report presented here are:

- Obligations of data providers and data users to operate in conformity with the requirements of their own institution, fulfil all necessary national and international regulatory and ethics regulations and to carry out high quality, ethical research.
- Obligation of users of confidentiality and conformity to data protection principles to ensure that data is processed in compliance with the legal and ethical requirements.
- Obligations of providers of *in vivo* data to operate in conformity of with relevant guidelines for animal welfare and care.

[3] https://openrisknet.org/privacy-policy/
[4] https://openrisknet.org/terms-of-use/

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

5.3 Specific information on individual shared data sources

5.3.1 diXa

The ethics workflows and checklists were created based on the data for different models (*in vivo* animal models, *in vitro* human and animal models) which are the same types of data included in the diXa Data Warehouse. These data also have different requirements and legislations relevant for the ethics review. A complete evaluation according to the checklist is ongoing.

The ethical issues statements from the diXa project and an example of the ethics check of one of the studies in the diXa Data Warehouse are provided in **Annex 4.**

5.3.2 BridgeDb

BridgeDb only stored different identifiers and their mappings, which are not related to any kind of personal or experimental data. Therefore, no ethics evaluation is necessary.

5.3.3 WikiPathways

The pathway information stored in WikiPathway is fully based on knowledge extracted from the scientific literature without any links to the experimental data, which was used to arrive at the conclusions. Therefore, no ethics evaluation is necessary

5.3.4 AOP-Wiki

The knowledge that is present in AOP-Wiki is completely based on scientific literature and is therefore not directly related to any experimental data. Therefore, no ethics evaluation is necessary.

5.3.5 ToxCast

Ethics evaluation of the *in vitro* data available from ToxCast following the checklist has started.

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

2. OpenAIRE - OpenAIRE [Internet]. Available: https://www.openaire.eu/

3. How to create a DMP Plan | Open Research Data Pilot [Internet]. Available: https://www.openaire.eu/opendatapilot-dmp

4. DMPonline [Internet]. Available: https://dmponline.dcc.ac.uk/

5. 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.

6. 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.

7. OECD Harmonised Templates - OECD [Internet]. Available: https://www.oecd.org/ehs/templates/

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

9. 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. Ethics requirement for OpenRiskNet infrastructure data providers

Annex 2. Personal Data Protection and Privacy Policy

Annex 3. OpenRiskNet e-infrastructure terms of use

Annex 4. Ethical issues from the diXa project

Ethics requirement for OpenRiskNet infrastructure data providers

Version 1 from **29 November 2018**

OpenRiskNet has developed a **checklist** to make sure that data providers meet required ethics guidelines for the specific data to be supplied. The checklist has general requirements that apply to all data types, and additional requirements for specific data types. Before a dataset can be used in this project, an evaluation following the relevant criteria listed in this document must be done and a statement confirming that all outlined requirements applicable to the data provider and for the specific datasets must be submitted along with supporting documentation (e.g. copies of ethical approvals, response sheets, etc.).

General requirements for COMPANIES, LABORATORIES, and INSTITUTIONS providing data:

- ❑ Confirm the responsibility in the collection, use and processing of biomaterial and data being provided.
- ❑ 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 EC, 2018, 10:37-38)[1].
- ❑ Are liable for any subsequent problem arising from the material that has not been disclosed explicitly previously.
- ❑ Have guaranteed no harmful material was collected or processed without the full informed consent of the donor and the partner handing the material.
- ❑ 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 25.8.2018. These differences are stated in the requirements for the individual data types.

Basic requirements for ALL types of human data:

- ❑ The contractor/data provider must be in compliance with EU Council Directive (EC) 2004/23 of 31 March 2014[2] and EU Council Directive (EC) 2006/86 24 October 2006 implementing Directive 2004/23/EC[3].
- ❑ Free and fully informed consent of the donors for primary and secondary use of material or additional material must be obtained.
- ❑ No personal data on donor of biomaterial is allowed.

Further requirements according to type of biomaterial being provided

1. **Commercial Cell Lines:**

[1] European Commission (February 2018). Horizon 2020 Programme Guidance on How to complete your ethics self-assessment (version 5.3.). EC DG Research and Innovation. Brussels. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf [retrieved 7.7.2018]

[2] 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.

[3] 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.

- ❑ No personal data
- ❑ Fully anonymized data
- ❑ Non re-identifiable

2. **Research Using, Producing or Collecting Human Cells and Tissues (EXCLUDING Embryonic Stem Cells (hESC) and Human-Induced Pluripotent Stem Cell (hIPSC):**
   - ❑ Statement confirming that relevant ethics approvals were obtained for accreditation, designation, authorization 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
     - procurement, processing, labelling, packaging, distribution, traceability and imports and exports of cells and tissues from and to third countries
   - ❑ Compliance with EU and national legislation regarding protection of data, privacy policies.
   - ❑ Free and fully informed consent of the donors for secondary use of material or from derived clinical practices (e.g. waste material from surgery or operations) or additional material with option to erase further use at any time (this option is for data generated after 25.8.2018).
   - ❑ Material is fully anonymised.
   - ❑ If data provider is a COMPANY/LABORATORY/INSTITUTION/BIOBANK, the following must also be provided
     - Details of provider (including name and country).
     - Origin of the 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.
     - In addition to obtaining informed consent, clear exclusion/inclusion criteria (capacity to discriminate, children, emergency, etc) must be stated.
     - Confirmation of general principles of ethics at EU or national level are respected.

3. **Research using Embryonic Stem Cells hESCs and Human-induced Pluripotent Stem Cell (hiPSC) Lines:**
   - ❑ For new data associated with hESC or hiPSC (samples established after 25.5.2018) and data derived from them can be pseudonymized but must not be fully anonymized and must be traceable to sample donor.
   - ❑ New or existing samples after 25.5.2018, where the sample donor is not fully anonymized therefore relate to identifiable legal/natural persons and so fall within the scope of the GDPR and as such consent on new purpose/usage must be sought from donor(s).
   - ❑ Compliance with EU and national legislation regarding protection of data, privacy policies.
   - ❑ Relevant ethics approvals were obtained for accreditation / designation / authorisation / licensing for using, processing or collecting the human cells and tissues (if relevant).
   - ❑ For samples established before 25.5.2018, the following information is also required:
     - Origin and line of cells.
     - Details of licensing and control measures by DPO.
     - hESCs/hiPSC are registered in EU registry (www.hescreg.eu).

- Statement confirming that the 6 specific conditions (EC, 2018, 1: 4-5))[1] below are met and that ethics approval with informed consent and information sheets is obtained.
    I. cells were NOT derived from embryos specially created for research or by somatic cell nuclear transfer.
    II. The project uses existing cultured cell lines only.
    III. Cell lines were derived from supernumerary non-implanted embryos resulting from in vitro fertilisation.
    IV. informed consent has been obtained for using donated embryos for the derivation of the cell lines.
    V. Personal data and privacy of donors of embryos for the derivation of the cells are protected.
    VI. NO financial inducements were provided for the donation of embryos used for derivation of the cell lines.

4. Research using Clinical Trial Data

- ❏ No personal data
- ❏ Fully anonymized and Non re-identifiable data
- ❏ Data provider complies with European Commission Ethics requirement (EC, 2018, 1: 6-19))[1] in line with national law and ethics committees. To show this the following must be provided:
    - A statement confirming that an ethical review committee approved the research conditions and that ethics approval with informed consent and information sheets were provided
    - Confirmation of general principles of ethics for the EC[4]
    - Privacy requirements
    - Data protection statements (for data generated before 25.5.18. classic ethical statement applies, if data is generated after 25.5.18 then GDPR should be applied)
    - All applicable health and safety requirements were/are met.
    - Data is collected using electronic encoding tools.

- ❏ Compliance with the Declaration of Helsinki[5] and the Oviedo Bioethics Convention[6]
- ❏ Respect of EU Regulation No 536/2014 on clinical trials on medicinal products for human use[7]
- ❏ 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)[8]

---

[4] Horizon 2020 Ethics Procedure http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm

[5] 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

[6] 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

[7] EU Regulation No 536/2014 of the European Parliament and of the Council on clinical trials on medicinal products for human use, repealing Directive 2001/20/EC (OJ L 158, 27.5.2014) https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX:32014R0536 / https://ec.europa.eu/health/human-use/clinical-trials/regulation_en

[8] 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

Additional Requirements for Animal Data

Additional requirements for animal data are listed below.

Basic requirements for ALL types of animal data:

Your research must comply with:

- ❑ Respect of EU Directive 2010/63/EU[9] and other applicable national and international law guiding use of animals.
- ❑ The relevant/necessary national authorisations for the supply of animals and the animal experiments (and other specific authorisations, if applicable).

Definition of terms[10]

- **Anonymisation** = data are considered to be anonymised where they are fully (unlinked) anonymised or linked (coded, pseudo-) anonymised where the linkage code (cipher) is not held by, or accessible to, the researchers/research establishment. 'Anonymised' data do not contain any identifiable information such as, for example, name, address, phone number, full date of birth, national health or social insurance numbers, full postcode, etc., and it is not reasonably possible for the researcher to identify the individual to whom the data relates.
  Linked anonymised (or pseudo-anonymised or coded) data are fully anonymous to the researchers who receive or use them, but contain information or codes that would allow others (e.g., the clinical team who collected them or an independent body entrusted with the safekeeping of the code) to link them back to identifiable individuals.
  Unlinked anonymised data contain no information that could reasonably be used by anyone to identify the individuals who donated them or to whom they relate.
- **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** = data which may be used to identify a research participant. (Note: although in some EU jurisdictions personal data may also be used to describe human biosamples, in the context of this document, it relates to identifiable data only).
- **Data Owner** = the 'data owner' is the individual researcher or investigator or body of researchers or investigators that produced the original 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.

[9] 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).

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

Personal Data Protection and Privacy Policy

Effective date: 21 November 2018

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 assumes no responsibility for errors or omissions in the contents on the Service.

This information is

- of a general nature only and is not intended to address the specific circumstances of any particular individual or entity;
- not necessarily comprehensive, complete, accurate or up to date;
- sometimes linked to external sites over which the 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, whether in an action of contract, negligence or other tort, arising out of or in connection with the use of the Service or the contents of the Service. OpenRiskNet reserves the right to make additions, deletions, or modification to the contents on the Service at any time without prior notice.

OpenRiskNet does not warrant that the Service is free of viruses or other harmful components.

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.

Copyright and acknowledgement of sources

The author aims to observe the copyright of any graphic, audio document, 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. We collect several different types of information for various purposes to provide and improve our Service to you.

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 details as well 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.

Types of Data Collected

Personal Data

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:

- Email address
- First name and last name
- Affiliation
- Cookies and Usage Data

Usage Data

We may also collect information 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.

Tracking & Cookies Data

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.

Use of Data

OpenRisKNet uses the collected data 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

Transfer Of Data

Your information, including 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.

Disclosure Of Data

Legal Requirements

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.

Security Of Data

The security of your data is important to us, but remember that no method of transmission over the Internet, or method of electronic storage is 100% secure. While we strive to use commercially acceptable means to protect your Personal Data, we cannot guarantee its absolute security.

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.

E-mail Communication

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

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 analyzing 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 the Google Privacy & Terms web page: https://policies.google.com/privacy?hl=en

Verifying, modifying or deleting information

If you want to verify, modify or delete your personal data stored by the responsible controllers for the OpenRiskNet website and its sub-domains, please notify the Coordinator Office by email. In your email, clearly state your request and include the URL of the website/webpages your request refers to.

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.

Changes To This 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.

Contact Us

If you have any questions about this Privacy Policy, please contact us by email: at openrisknet@douglasconnect.com.

OpenRiskNet e-infrastructure terms of use

Effective date: 27 November 2018

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

Anonymisation = data are considered to be anonymised where they are fully (unlinked) anonymised or linked (coded, pseudo-) anonymised where the linkage code (cipher) is not held by, or accessible to, the researchers/research establishment. 'Anonymised' data do not contain any identifiable information such as, for example, name, address, phone number, full date of birth, national health or social insurance numbers, full postcode, etc., and it is not reasonably possible for the researcher to identify the individual to whom the data relates.

Linked anonymised (or pseudo-anonymised or coded) data are fully anonymous to the researchers who receive or use them, but contain information or codes that would allow others (e.g., the clinical team who collected them or an independent body entrusted with the safekeeping of the code) to link them back to identifiable individuals.

Unlinked anonymised data contain no information that could reasonably be used by anyone to identify the individuals who donated them or to whom they relate.

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 = data which may be used to identify a research participant. (Note: although in some EU jurisdictions personal data may also be used to describe human biosamples, in the context of this document, it relates to identifiable data only).

Data Owner = the 'data owner' is the individual researcher or investigator or body of researchers or investigators that produced the original 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 finally 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 provided by the data owner.

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 a 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 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 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. This does not limit the use of the OpenRiskNet infrastructure installed locally with suitable security measures prohibiting unauthorised access.

- ❏ 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 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@douglasconnect.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.

Ethical issues from diXa project

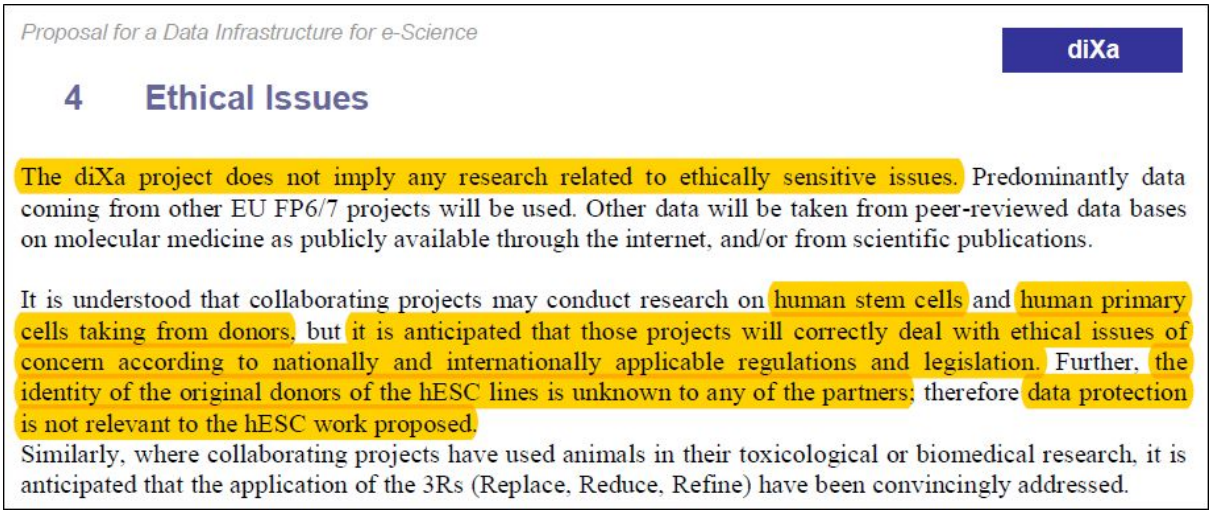
The Data Infrastructure for Chemical Safety (diXa) project (http://www.dixa-fp7.eu/) was a e-infrastructure project funded by EU FP7 to provide a single resource for the capture of toxicogenomics data produced by past, present and future EU research projects, and to ensure sustainability of such a resource for use by the wider research community.

The captured data was placed in the diXa Data Warehouse, which served as a "Pilot Database" for ethical review by the OpenRiskNet project. In D6.5 - "POPD - Requirement No. 7" the final outcome of this evaluation is reported, thereby summarizing the ethical requirements needed and providing minimal requirements.

Ethical issues statements from diXa project

The ethical issues statements for gathering animal and human data (including data from human stem cells and human primary cells) were obtained from the diXa project proposal.

Part of the ethical issues from the diXa project proposal



Proposal for a Data Infrastructure for e-Science

diXa

4 Ethical Issues

The diXa project does not imply any research related to ethically sensitive issues. Predominantly data coming from other EU FP6/7 projects will be used. Other data will be taken from peer-reviewed data bases on molecular medicine as publicly available through the internet, and/or from scientific publications.

It is understood that collaborating projects may conduct research on human stem cells and human primary cells taking from donors, but it is anticipated that those projects will correctly deal with ethical issues of concern according to nationally and internationally applicable regulations and legislation. Further, the identity of the original donors of the hESC lines is unknown to any of the partners; therefore data protection is not relevant to the hESC work proposed.
Similarly, where collaborating projects have used animals in their toxicological or biomedical research, it is anticipated that the application of the 3Rs (Replace, Reduce, Refine) have been convincingly addressed.

The ethical governance under the diXa project will see to the following:

a. The research by the collaborating projects will be undertaken within the domestic requirements for ethics committee approval. This is the case for both the data from human participants, and the data from animals. All the decisions of the various regional ethics committees will be provided to the project's Scientific Office in the European Commission.

b. DiXa will only gather animal data from other research projects. Before such data are included in the DiXa work, DiXa will require that the original project where the data has been gathered evidences the ethical approvals that were gained for that original work, and an undertaking that the work was carried out within the requirements of the relevant domestic law and ethics.

c. diXa gathers human data in two ways: first, and predominantly, diXa will receive anonymous data (or anonymised data, i.e. information from which the identifiers have been completely removed and without any access to any remaining key) from already running projects; and second, and if necessary, diXa will itself gather data from patients. Both these situations raise important ethical and legal considerations. diXa will conduct its research within the requirements of the European and local laws and ethics for the countries where the work will be undertaken and from where information is originally gathered. At the European level, care will be taken to ensure that the project conforms to the requirements of, in particular, the Data Protection Directive (95/46/EC) and general requirements and expectations for medical research. Concerning the ethical management and use of the human data, this means that:

Example: the carcinoGENOMICS project

The carcinoGENOMICS project serves as an example for providing the necessary ethical information for the OpenRiskNet project. In particular the ethical issues concerning the human embryonic stem cell models will be addressed.

Part of the ethical issues from the carcinogenomics project proposal for the human embryonic stem cell lines

- **Use of human embryonic stem cell lines**

Regulations

The research on hES cells is regulated differently in different countries. Most countries, such as UK and Sweden, consider the whole procedure of hES cell derivation and research as ethically acceptable, while other countries have more stringent regulations. The CARCINOGENOMICS project is carried out in accordance to the Swedish national legislation described below, because the hESC research activities to be performed during the 1st period will only be done at Cellartis AB in Sweden.

Swedish legislations specifically concerned are: *Swedish Government Bill on Stem Cell Research (Prop. 2003/04:148, 1 April, 2005)* and *Swedish Government Bill on Genetic Integrity (Prop. 2005/06:64, 1 July, 2006).* For full details on the Swedish legislation see: http://www.riksdagen.se/webbnav/index.aspx?nid=3110&titel=&rm=2003%2F04&bet=148&doktyp=&org=&s=S%C3%B6k and http://www.riksdagen.se/webbnav/index.aspx?nid=3110&titel=&rm=2005%2F06&bet=64&doktyp=&org=&s=S%C3%B6k

In conclusion, Swedish law permits hES cell research and commercialisation. A summary of this information can be found at: http://www.regeringen.se/sb/d/183/a/26332. All hES cell lines at Cellartis have been derived in line with this document, and in accordance to the above described laws.

Cellartis warrants that:

- All research and handling of hES cells as well as adult stem cells is based on careful ethical considerations;
- The company is guided by and follows all existing laws and regulations in the major markets (including Europe, USA and Japan);
- The company does not see any need for somatic cell nuclear transfer (therapeutic cloning) in the foreseeable future;
- The company finds cloning of human beings (reproductive cloning) unethical and supports initiatives aimed at a global ban.

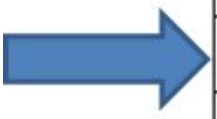
Cellartis has identified the following principal requirements regarding human embryonic stem cell research and the procurement of embryonic stem cells from supernumerary embryos:

- Free and informed consent from the donating couple or woman;
- Approval of the research by an ethics committee;
- No financial gain for the donors or the health care system;
- Anonymity of the donors and protection of the confidentiality of personal information of the donors;
- Transparency regarding donation and all general handling of the donated material.

hES cell lines used in CARCINOGENOMICS

Cellartis is the only partner that will use hES cells. The proposed work does thus NOT include the establishment of new human ES cell lines. Instead, only hES cells and derivates from cell lines previously established over the years 2001-2004, will be used. Moreover, other participants of CARCINOGENOMICS who will only use materials (supernatants, cell extracts, etc...) from studies on hES cells at Cellartis (partners 1, 2, 3, 5, and 8), will do this in accordance with the various national regulations that are relevant in their home country (e.g. the Netherlands, Spain, Belgium, and the UK).

In particular, the following hES cell lines will be used:



Cell line nr	Ethics approvals	Establishment date	Tracable?
SA001	Ö 507-00, Ö026-03, 067-04	20-3-2001	Yes
SA002	Ö 507-00 , Ö026-03, 067-04	21-5-2001	Yes
SA046	Ö 507-00, Ö026-03, 067-04	3-12-2001	Yes
SA094	Ö026-03, 067-04	25-2-2002	Yes
SA111	Ö026-03, 067-04	15-4-2002	Yes
SA121	Ö026-03 , 067-04	29-4-2002	Yes
SA167	Ö026-03, 067-04	16-9-2002	Yes
SA181	Ö026-03, 067-04	30-9-2002	Yes
SA191	Ö026-03, 067-04	20-10-2002	Yes
SA196	Ö026-03, 067-04	29-10-2002	Yes
SA202	Ö026-03, 067-04	7-11-2002	Yes
SA218	Ö026-03, 067-04	26-11-2002	Yes
SA240	Ö026-03, 067-04	19-12-2002	Yes
SA348	Ö025-03,	26-8-2003	No
SA352	Ö025-03	28-8-2003	No
SA399	Ö025-03	20-10-2003	No
SA461	Ö025-03	27-1-2004	No
SA502	Ö025-03	23-3-2004	No
SA506	Ö025-03	30-3-2004	No
SA521	Ö025-03	19-4-2004	No
SA540	Ö025-03	17-5-2004	No

**the arrow indicates the finally used hES cell line in the carcinoGENOMICS project*

The finally used hES cell line in the carcinoGenomics project, SA002, has been described in the European Stem Cell Registry. Furthermore, this cell line has been made fully anonymous by destroying any information connecting the cell line to the original donors (see below Englund *et al.* 2010).

Fragments from paper showing registration and description of the hES cell line from Cellartis.

In Vitro Cell.Dev.Biol.—Animal (2010) 46:217–230
DOI 10.1007/s11626-010-9289-z

REPORT

The establishment of 20 different human embryonic stem cell lines and subclones; a report on derivation, culture, characterisation and banking

Mikael C. O. Englund · Gunilla Caisander · Karin Noaksson · Katarina Emanuelsson · Kersti Lundin · Christina Bergh · Charles Hansson · Henrik Semb · Raimund Strehl · Johan Hyllner

To date, we have established 31 hES cell lines and several subclones of these hES cell lines, whereof several are presented at different hES cell registries worldwide and also made available via different hES cell banks. For more information regarding this, please visit the European Stem Cell Registry (http://www.hescreg.eu/), the International Stem Cell Forum (http://www.stemcellforum.org/), the US National Stem Cell Bank (http://www.nationalstemcellbank.org/) or contact Cellartis AB directly (http://www.cellartis.com/)

Regulatory regime. All hES cell lines derived by Cellartis originate from surplus human embryos from clinical *in vitro* fertilisation (IVF) treatment. The blastocysts were donated after informed consent and approval of the local ethics committees at Gothenburg University and Uppsala University. Cell lines derived from blastocysts sourced from Gothenburg University has the prefix SA* (Sahlgrenska Academy) and cell lines derived from blastocysts sourced from Uppsala University has the prefix AS* (Akademiska Sjukhuset, i.e., Uppsala University Hospital).

The two hES cell lines, SA001 and SA002, and thus also the subclone, SA002.5, have been made anonymous, i.e., all records connecting the cell lines to the original blastocyst donating couple have been irreversibly destroyed.

Conclusion

The required ethics information of the hES cell model from the carcinoGENOMICS project could be traced back, including some additional information on the used model from literature.