OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Deliverable Report D4.4

Report on Re-Identification Risks and Private by Design Risk Management

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-Universitat 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 Universitat 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 4.4 - Report on Re-Identification Risks and Private by Design Risk Management
Deliverable Type	Report
Dissemination Level	Public (PU)
Work Package	WP4
Task(s)	Tasks 4.1, 4.2, 4.3, 4.4 and 4.5
Deliverable lead partner	DC
Author(s)	Lucian Farcal (DC), Thomas Exner (DC)
Status	Final
Version	V1.1
Document history	2017-10-06 Draft version 2017-12-13 Final version 1.0 2018-02-14 Revised version 1.1

Table of Contents

SUMMARY

This report describes the possible risks with respect to security, privacy and re-identification of personal data of relevance to the OpenRiskNet infrastructure and community, as well as presenting our private by design risk management concept supporting our data solution development.

INTRODUCTION

The OpenRiskNet Consortium develops the OpenRiskNet e-infrastructure for the harmonisation and improved interoperability of data and software tools in the area of predictive toxicology and risk assessment. The approach combines interoperable web services providing data or analysis, processing and modelling tools communicating over well-defined and harmonized application programming interfaces (APIs) supplemented by a semantic interoperability layer added to every service to describe the functionality whilst guaranteeing the technical and semantic interoperability. OpenRiskNet is following a multi-domain community-driven approach by analysing the unmet needs and requirements of all relevant communities either defined based on expertise (data manager, tools developer, workflow integrator and end user community, where the latter is composed out of researchers, risk assessors, regulators and the general public) or based on research topic (chemistry, agrochemistry, pharma, cosmetic ingredients or nanomaterials).

The e-infrastructure developed in the project requires access to existing publicly-available datasets of *in vitro* and *in vivo* experimental studies based on animals or cell models (in electronic form) from other projects, third parties and stakeholders. Even if the generation, use and storage of personal data in relation to the biological data is not foreseen within the case studies of the project, i.e. no research involving human participants (patients or healthy volunteers) for medical studies or studies based on human cells or tissues will be conducted, the design of the infrastructure will provide all necessary components to enable responsible and secure data management processes to make it fit for possible future applications in the field of personalized toxicology.

Details on the protection of personal data measures already included in the confidential **D6.2 Ethics deliverable report,** will be reproduced here and partly updated.

SECURITY, PRIVACY AND PREVENTION OF RE-IDENTIFICATION OF PERSONAL DATA

The OpenRiskNet consortium is integrating and will continue to integrate measures into the e-infrastructure to guarantee security, privacy and prevent re-identification of individuals based on the provided data. These activities are targeting two specific but still interconnected areas:

1. Public and commercial data sources: We are working closely together with the Data Protection Officer (DPO) assigned to the OpenRiskNet project to develop criteria and workflows to evaluate the ethics standards of datasets grouped into the four categories of 1) human personal data, 2) data from primary and stem cell lines, 3) data from standard human cell line models and 4) animal data including data from *in vivo* and *in vitro* models. These workflows will be first reported in the confidential ethics deliverable report D6.5 prepared by the DPO but will be made publicly available shortly after the M18 review meeting, which will also include an ethics review.

   Applying these workflows on examples of databases prioritised for integration into OpenRiskNet (diXa data warehouse and Tox21 from the U.S. Environmental Protection Agency), it became obvious that while security and data integrity have to be considered during the interface development, privacy and prevention of re-identification are not relevant for the public and commercial data sources planned to be integrated during the course of the project, since these data sources don't include personal data (category 1) and only commercial human cell lines which have full ethical clearance were used (category 2 and 3).

   For each individual data source, considered for integration, a dedicated OpenRiskNet partner will evaluate if it complies to high standards regarding security and privacy. Additionally, OpenRiskNet will guarantee during the interface development that these standards cannot be bypassed using the new access routes (data application programming interfaces) or by manipulation of the user management in local deployments.

2. User-provided data: Users might want to analyse data, which includes confidential and private human data even coming from individuals. OpenRiskNet offers through the deployment options and the virtual environments ways to do these analyses in private, based on local instances of the OpenRiskNet e-infrastructure. In this way, no data has to leave the premises of the user so that it cannot be compromised during data transfer. If a specific service needs to access public resources, which cannot be deployed to the local instance, the user will be notified by the service that the provided data will be transferred to an external site and the user can choose an alternative service to avoid this.

More details on how this will be achieved, are given in the next section.

PRIVATE BY DESIGN RISK MANAGEMENT APPROACH

Public and commercial data sources

OpenRiskNet partners integrating a data source are responsible for obtaining the clear consent and permission of the data owner for the anticipated usage and transferring access restrictions from the original data source to the corresponding OpenRiskNet service. During this process, both parties evaluate together what specific requirements have to be fulfilled to guarantee security and privacy and how these factors influence the integration of the data source. Additionally, clear regulations regarding the access (license) and the re-usability are formulated.

None of the existing data sources provided until now are accompanied with complete documentation on the ethics status for the complete database or individual datasets. Together with the data protection officer, OpenRiskNet developed a catalog of required information, which includes documentation of the ethical approval of the study, in which the dataset was produced, statement on security and privacy measures as well as on approval of secondary usage, and the license (open or commercial) under which the data can be accessed and reused. OpenRiskNet is currently preparing best-practice guidelines on how ethics requirements should be integrated into the design of databases and how to ensure privacy and security of e-infrastructures. We are working together with data providers from the field of predictive toxicology and risk assessment as well as other related projects like NanoCommons to include feedback from all stakeholders. The resulting recommendations will concentrate first on *in vitro* and *in vivo* data from human cell cultures and animals and are meant to become part of data management best practices e.g. to be put down and agreed on in data management plans. Simultaneously, we are on the way to establish interactions with neighboring disciplines like drug design and personal medicine facing the similar ethics requirements with the goal to extend the guidelines towards preventing re-identification of personal data, which is a prerequisite to be able to open up the e-infrastructure also to these disciplines.

In addition to the task performed in cooperation with the data providers, sensitive data at rest will be encrypted and remotely accessed services will use secure protocols like SSH and HTTPS for data transfer in the OpenRiskNet infrastructure. Depending on the results of the Implementation Challenge [1], data sources with personal data from third parties might be selected because of their relevance for the OpenRiskNet project towards the impact goal outlined in section 2.2.5. Since these will be existing resources, secure data management processes regarding anonymisation eliminating the risk of re-identification, encryption and logging as well as preventing malevolent / criminal / terrorist abuse will by default already be in place. The partner from the OpenRiskNet consortium providing the technical and implementation support for the integration of the source will take the responsibility that these processes are also followed and implemented into the OpenRiskNet-compliant service. This task has to be done in intense cooperation with the original data provider since they have to provide the technical documentation of the measures implemented in the data warehouse including the user interface. If such processes are not established in the original service, the OpenRiskNet partners and the third party will work together to implement these or, if this is not possible, stop the integration until an appropriate data management is in place to keep the high ethical standards for all OpenRiskNet services. All data sources anticipated so far to be integrated by the OpenRiskNet consortium do not include any human personal data, were generated by major consortia or institutions with high ethics standards and are already publicly accessible.

Having security and privacy measures implemented in the original service is only the first step. The

OpenRiskNet infrastructure will open up additional access routes and make the resource available in local environments. Therefore, the implementation has to take care that the measures cannot be bypassed in these new settings. The authorisation and authentication services, which are a central part of OpenRiskNet, will guarantee adoption of the required security protocols in five main directions:

- Confidentiality: ensuring that information is not accessed by unauthorised persons.
- Authentication: ensuring that the users of critical functions are the persons they claim to be.
- Integrity: ensuring that external occurrences cannot alter the information in such a way to be not detected by authorised persons.
- Prohibiting changing of access rights by local users with superuser privileges.
- Limiting deployment to datasets accessible to the user to prohibit data security breaches possible by technical insecurities of the local (or cloud) system.

Data provided by the user

One reason to follow the concepts of virtual environments and easy deployments of these in OpenRiskNet is to improve the security of data provided by users. When running the risk assessment workflows on a local system or a cloud solution behind a company/institution firewall, sensitive data does not have to leave the institution generating the data and being responsible for keeping them secure. Public available data and services to perform analyses and modelling tasks on combined private and public data will also be accessible in a containerized form and integrated in a virtual environment on the private network of the institution, which is, in this way, protected against unauthorised access from external parties including the providers of the services. Such an approach is already successfully implemented in the PhenoMeNal project [2]. If the user decides to use public cloud solutions, e.g. the OpenRiskNet reference instance, they will be warned that the security and privacy of user-supplied data cannot be guaranteed by the OpenRiskNet infrastructure. However, we were able to win Red Hat, Inc. as an associated technology partner of OpenRiskNet, which can provide secure cloud solutions to the user if more computing power is needed than available at the institution. Additionally, if a service, e.g. because it is in a preliminary, development status, cannot be completely deployed to a local system and needs to access some resources over the internet, the user will be notified about this fact, and will be informed what kind of data will be exchanged and if there is a similar service deployable in-house.

CONCLUSION

Measures to guarantee security and privacy for public and commercial data sources and additionally to prevent re-identification of individuals based on the provided data for user-provided data were established and are currently implemented into the e-infrastructure. To assure an independent assessment of these measures, an external Data Protection Officer (DPO) was assigned to the OpenRiskNet project with the role to develop criteria and workflows to evaluate the ethics standards of datasets. Details on the protection of personal data measures were included also in the Deliverable Report D6.2 Protection of Personal Data.

GLOSSARY

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

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

REFERENCES

1. Associated Partner Programme :: OpenRiskNet [Internet]. [cited 4 Dec 2017]. Available: https://openrisknet.org/associated-partner-programme/

2. PhenoMeNal – Large-scale Computing for Medical Metabolomics [Internet]. [cited 4 Dec 2017]. Available: http://phenomenal-h2020.eu/home/

3. EUR-Lex - 31995L0046 - EN. OPOCE; Available: http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML