OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Deliverable Report D1.1

Report on requirements analysis and recommendations for WP2-4

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
Star 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 1.1 Report on requirements analysis and recommendations for WP2-4 including Data Management Plan
Deliverable Type	Report
Dissemination Level	Public
Work Package	WP1
Task(s)	Task 1.1
Deliverable lead partner	DC
Author(s)	Lucian Farcal (DC), Daniel Bachler (DC), Joh Dokler (DC), Thomas Exner (DC)
Status	Final
Version	V2.0
Document history	2017-02-14 First draft 2017-05-25 Consolidated draft 2017-06-01 Final version 1 2018-11-30 Version 2

History of revisions - Version 2.0

Description	Section
The summary was updated accordingly	Summary
The context, the aim and the approach taken on the requirement analysis were explained in detail and updated. Additional details on the stakeholders were added.	Introduction
Learning regarding the general requirements and also on specific ones extracted from the developers' or end users' answers, with reference to the complexity and diversity of requirements to be addressed	OpenRiskNet Survey
A new section on the ongoing requirements analysis was added, in order to document the current actions	Ongoing requirements analysis

Table of Contents

SUMMARY

This report describes the results obtained from the survey, interviews and interactions with associated partners and project stakeholders as part of the requirements analysis. The requirements analysis included surveys sent out to a large number of experts and designed to address issues relevant to:

- **End users** (e.g. members of academia, industry and regulatory agencies)
- **Developers** (tools developers, infrastructure provider and data managers)

Additionally, key persons were contacted for additional discussions to learn more about different infrastructure development or use in face-to-face and virtual interviews, at conferences like SOT, EuroTox and OpenTox as well as via webinars with question-and-answer sessions.

The aim of the requirements analysis was to challenge the proposed concepts and to identify gaps in the existing approaches and unmet needs in the different communities. These were then used for improving and correcting the design concepts of the general infrastructure, the container orchestration approach, the API definition and the semantic interoperability layer and were translated into case studies. Requirements analyses performed in other projects (e.g. eNanoMapper, ToxBank) and interactions with ongoing projects (EU-ToxRisk, NanoCommons, projects of the NanoSafety Cluster) have been used to identify eventual elements already covered by other initiatives, which can be adopted, adapted if necessary and integrated into OpenRiskNet.

Performance metrics:

- Feedback from all communities to requirements analysis survey
- Number of interviews ≥10 (by month 36)

To be noticed that this is a public document, therefore the details on the responders could not be disclosed and included in the report.

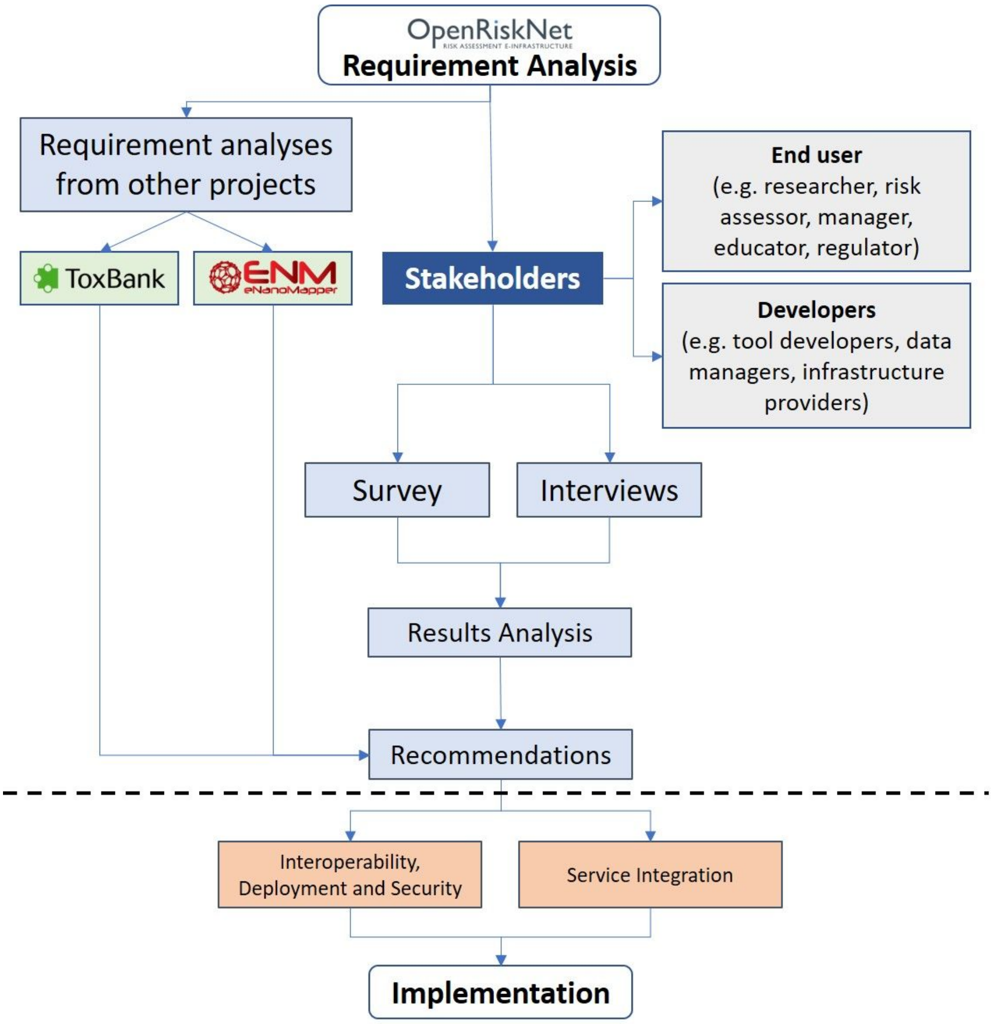

OpenRiskNet
RISK ASSESSMENT E-INFRASTRUCTURE
Requirement Analysis
Requirement analyses from other projects
ToxBank
ENM
eNanoMapper
Stakeholders
End user
(e.g. researcher, risk assessor, manager, educator, regulator)
Developers
(e.g. tool developers, data managers, infrastructure providers)
Survey
Interviews
Results Analysis
Recommendations
Interoperability, Deployment and Security
Service Integration
Implementation

Figure 1. Summary of OpenRiskNet requirements analysis approach

INTRODUCTION

OpenRiskNet aim is to provide open e-infrastructure resources and services to a variety of communities requiring chemical risk assessment, including chemicals, cosmetic ingredients, therapeutic agents, and nanomaterials. As proposed in the description of action, OpenRiskNet combines the achievements from earlier projects for generating modelling and validation workflows, knowledge integration and data management as well as including all ongoing projects and important stakeholders through an associated partner programme. The main components of the infrastructure is an interoperability layer added to every service to describe the functionality and guaranteeing technical and semantic interoperability, a discovery service, deployment options based on container technology, and packaging of the infrastructure into virtual instances. This is complemented by training and support on integration of specific services based on prototype implementation, usage of standard file formats for data sharing including the generation of templates for data and metadata, as well as the harmonised usage of ontologies. Case studies already demonstrate the applicability of the infrastructure in productive settings supporting research and innovation in safer product design and risk assessment. The requirements analysis described in this report was initiated to challenge these design concepts and to identify areas where they have to improved, optimised or corrected according to the stakeholders' feedback.

The detailed analysis of the user needs as well as of the perceived gaps for risk assessment and safe-by-design studies was carried out through a consulting progress using **surveys**, **face-to-face or online interviews**, **email conversations** as well as **presentations and webinars** on the OpenRiskNet concepts with dedicated **question-and-answer sessions**. The potential users were able to get more information on the developments but the consortium also asked the participants for their opinion on the different aspects of the infrastructure and to provide suggestions for additions and improvements. The target group did not include only individual researchers but also other key EU-funded consortia so as to ensure the alignment of OpenRiskNet with these efforts. This community-driven approach defines the communities to be included based on the expertise and research topic and will be continued throughout the project to be able to adapt to changes and new developments in the field.

In this report, we first define the stakeholder groups, which are relevant as potential users of the offerings of the OpenRiskNet e-infrastructure. Then we present the results of the first version of the Requirements Analysis survey, which are based on the answers we got until M6 of the project. This includes the learning the OpenRiskNet consortium took from the survey results and how this influenced and changed the design and implementation of the infrastructure. This is followed by a short descriptions of the individual interviews. The main outcomes of these interviews and the feedback from the presentations to larger groups have been integrated in the description of the results and the learning from the survey. Finally, we describe the ongoing activities to continuously integrate stakeholder requirements and feedback into the project and the infrastructure and how these are determining the development and the course of the case studies.

OpenRiskNet stakeholders

Users and other stakeholders targeted by the OpenRiskNet infrastructure and services include, on one hand, end-users representing researchers performing studies used in early stage product development up to regulatory registration (safe-by-design), risk assessors and regulators and, on the other hand, database managers, software and tools developers from relevant areas, as well as workflow integrators (both developers of tools for workflow management and researchers implementing workflows e.g. in industry settings):

- End users
  - Members of academia
  - Industry
  - Consultants
  - Risk assessors
  - Regulators
- Tools developers
- Data managers
- Infrastructure providers

The services provided by OpenRiskNet aim at relevant scientific communities (pharmaceutical, chemical, cosmetics and nanomaterial sectors) and should be optimised to cover current approaches and address specific needs with regards to standard risk assessment protocols based on *in vitro* and *in silico* methods. Specific requirements also need to be addressed to define access guidelines and user experience design concepts for the infrastructure by the scientific community, industry and regulatory bodies. Therefore, a close collaboration between these categories of knowledge providers/integrators and knowledge users is a key to push acceptance especially in regulatory settings of the integrated tools and workflows.

The relevant communities for OpenRiskNet infrastructure developments include but are not limited to:

- Relevant European research communities and institutions (e.g. NanoSafety Cluster, OpenTox, EC-Joint Research Centre)
- Relevant manufacturing industries through their organizations as well as individual members:
  - Pharmaceutical industry (European Federation of Pharmaceutical Industries and Associations - EFPIA)
  - Chemical industry (European Chemical Industry Council - CEFIC)
  - Cosmetic industry (Cosmetics Europe)
  - Nanomaterial industry (Nanomaterial Industry Association - NIA)
- Relevant regulatory agencies:
  - European Medicines Agency (EMA)
  - European Chemicals Agency (ECHA)
  - Scientific Committee on Consumer Safety (SCCS)
  - European Food Safety Authority (EFSA)
  - Organization for Economic Co-operation and Development (OECD)
  - International equivalent agencies (e.g. US EPA)

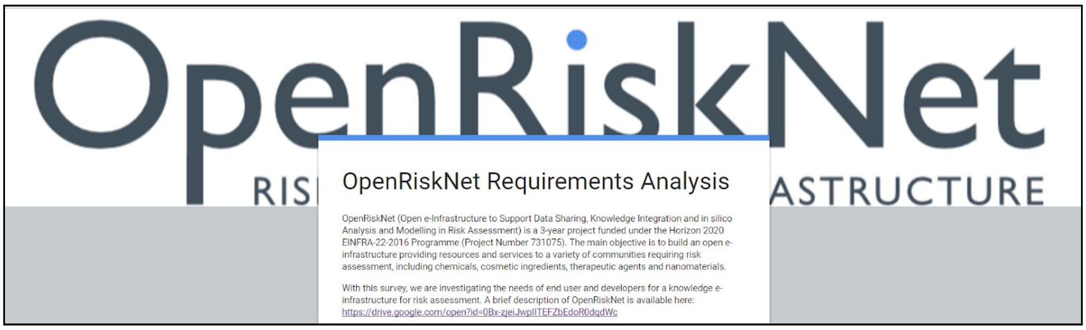
Due to the involvement of various OpenRiskNet partners in other EU-funded projects, we aim also to interact, exchange information and get input from ongoing or past projects which generated different information systems, e.g.:

- EU-ToxRisk
- PhenoMeNal
- HeCaTos
- diXa
- eNanoMapper
- ToxBank
- eTox
- Open PHACTS
- EU NanoSafety Cluster projects including ACEnano and NanoCommons

All these interactions and feedback received from potential users will be helpful to organise the e-infrastructure framework, to raise awareness of its development, to simulate early external testing of the functionality and usability, and to lead to acceptance of the OpenRiskNet concepts on openness with respect to data, interfaces and source code as well as technical and semantic interoperability.

OPENRISKNET SURVEY

The requirements analysis was mainly based on a survey and interviews performed among representative potential users of the OpenRiskNet e-infrastructure in the area of safety assessment and predictive toxicology as well as presentations and discussions with larger groups at conferences, workshops and meetings of the above mentioned group. With the survey, we aimed to investigate the needs of end users and developers for a knowledge e-infrastructure for risk assessment. An important goal was also to identify individuals, institutions and initiatives interested in the OpenRiskNet developments and to initiate a more detailed discussion on specific requirements (see the interview section below) and finally to win them as associated partners. Based on experience from previous projects including much larger ones like EU-ToxRisk and their problems to receive large numbers of responses, we designed the survey in a way that conclusions can still be drawn from identifying and collecting representative answers from different communities and to be able to follow-up on specific needs and identified gaps.



OpenRiskNet Requirements Analysis

OpenRiskNet (Open e-Infrastructure to Support Data Sharing, Knowledge Integration and in silico Analysis and Modelling in Risk Assessment) is a 3-year project funded under the Horizon 2020 EINFRA-22-2016 Programme (Project Number 731075). The main objective is to build an open e-infrastructure providing resources and services to a variety of communities requiring risk assessment, including chemicals, cosmetic ingredients, therapeutic agents and nanomaterials.

With this survey, we are investigating the needs of end user and developers for a knowledge e-infrastructure for risk assessment. A brief description of OpenRiskNet is available here: https://drive.google.com/open?id=0Bx-zjeiJwpllTEFZbEdoR0dqdWc

Figure 2. OpenRiskNet Requirements Analysis survey (version 1). This version was closed at M18 of the project and replaced by a new survey with additional questions more specific to the OpenRiskNet infrastructure (see below)

In the following, we will present the results from the responses obtained in the first six months of the project. Whenever relevant, we include also feedback and learning from the interviews, presentations, email communication and other discussions during this time. From 28 participants in the survey, 61% identified themselves as end users and 39% as developers. As became obvious from answers to questions below, the participants are in many cases the academic (professors or independent PIs) and industry (laboratory or unit) group leaders. Thus, they should be considered as representative for their groups or even for complete projects, due to their strong involvement in national, European and international projects and consortia, and not necessarily as individuals responders. The answers are clearly guided by the experience from these projects including the expertise and requirements of the consortium in total.

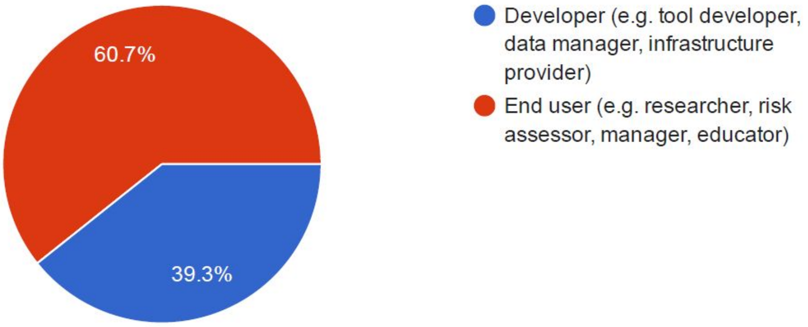


Figure 3. Distribution of responders between the two pre-defined categories (developers and end users)

The first part of the survey consists of general questions on identification of responder, affiliation, country and email address for further communication and follow-up discussions as well as on the work framework and context. Additionally, it includes a statement on how the information provided will be used within the project and that personal data will not be shared outside the consortium. This statement had to be acknowledged by the participants to fulfil the consent and information requirements from the ethics evaluation of the project.

Requirements analysis results: general questions

Specific questions and answers are:

Question: Type of Organisation: Industry, SME, Regulatory Agency, Governmental Institution, International Organisation, Academia, Non Governmental Organisation, Other

Answers:

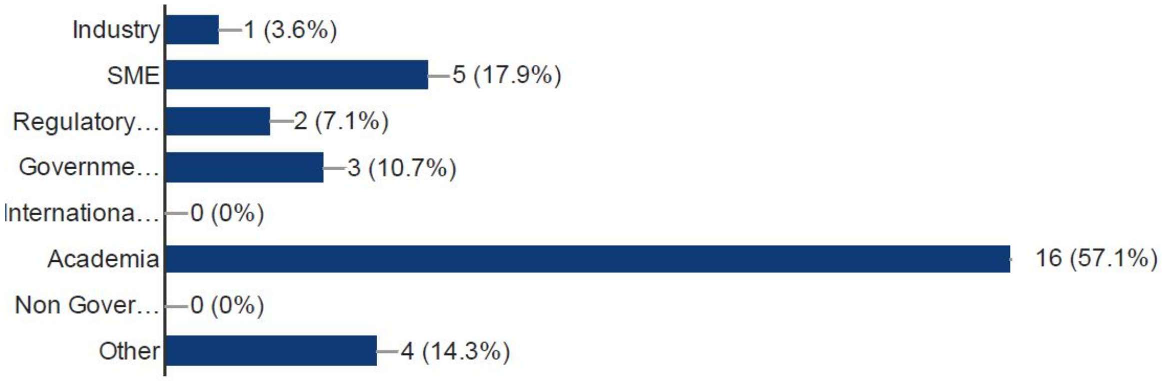


Figure 3. Type of Organisations represented

Learning: Academia clearly dominate the group of participants. This was also true for the participants listening to the presentations of the OpenRiskNet infrastructure or answered email correspondence. This is not unexpected since this group builds also the majority in other projects, is often involved in developing of new concepts, approaches and early stage tools and is more willing to accept sub-optimal, pre-mature solutions. In contrast, industry and especially regulatory agencies want to use high-standard, fit-for-purpose and preferably validated solutions and we, therefore, expect that these groups will become interested when the infrastructure is running stably and more services are available. This highlights the need for focused dissemination first to industry and finally to regulators throughout the project and a good exploitation concepts, since sustainability of the infrastructure will depend extremely on getting acceptance by these two groups. Having the Diamond Light Source as an official associated partner and first user to deploy the infrastructure in-house is a good demonstration of successful service provision to an commercial organisation and can be used as role model during dissemination to other industries. Additionally, it is encouraging to see that SMEs developing services are the second largest group in the survey, showing large interest in all meetings and have been strong in applying to the implementation challenge. This shows that they see the benefit of the OpenRiskNet approach to promote their commercial offerings and to attract new customers.

Question: Is the activity developing within the framework of an existing EU or international projects or initiatives?

Answers: Various projects were mentioned in different areas (safety assessment, biology and biotechnology, infrastructure, etc.). This list of projects provides information on initiatives where OpenRiskNet could focus on and/or develop partnerships.

Learning: Due to the strong involvement of the participants in national, European and international projects and consortia, the answers to the questions are not only representing personal opinions but are also guided by the experience from these projects and, in this way, represent a larger group of the toxicology and risk assessment community. It also shows that in many cases, the academic (professors or independent PIs) and industry (laboratory or unit) group leaders participated in the survey and the answers represent the expertise and requirements of the complete group.

Question: What type of methods and data do you regularly use? *In silico, In vitro, In vivo*

Answers:

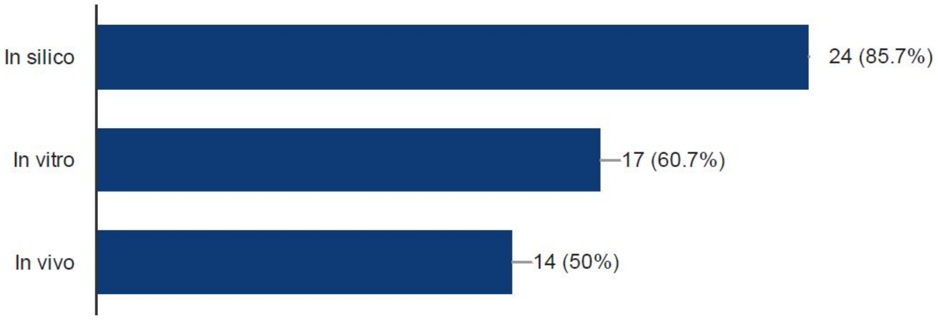


Figure 4. Type of methods and data regularly used by responders

Learning: The results are probably biased by the large amount of developers, who are more likely to responds to a survey coming from an infrastructure project in its early stage. However they still show that all different data types and tools are needed for the daily work in predictive toxicology and risk assessment by the majority of participants. Combined with the more specific questions below, it becomes clear that OpenRiskNet cannot concentrate on specific areas but needs to be able to support the integration of services from all areas with their different requirements on data and computational approaches. This can only be guaranteed with a very flexible design of the API definition and the semantic annotation concept.

After these general question, the participant was forwarded to the specific questions addressed to the two categories of stakeholders, developer and end user, covered in the next sections.

Requirement analysis results: developers' requirements

The aim was to learn on current practices in different organisations, experience of developers with different development systems and finally to identify requirements from the perspective of a tools developer, infrastructure provider or data manager.

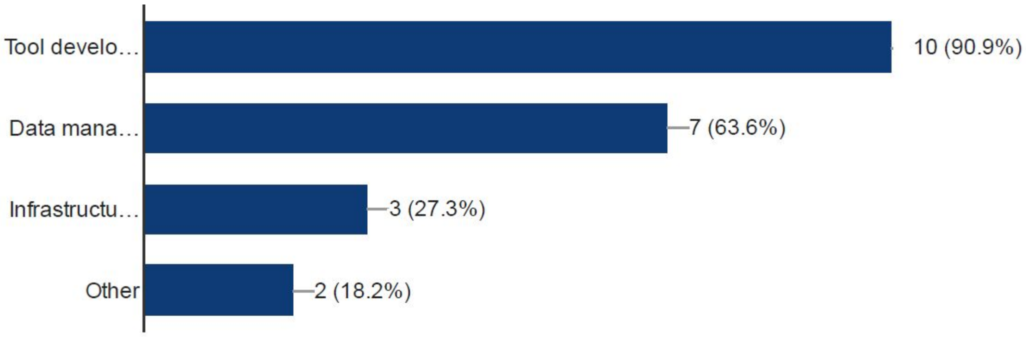


Figure 6. Developers roles within their organisations

Additionally, we tried to identify actual tools, data management and programming systems used, as well as to get recommendations for additional applications, which would be suitable for integration in the OpenRiskNet e-infrastructure.

Question: What is your experience working with HTTP based APIs (REST or similar)?

Answers:

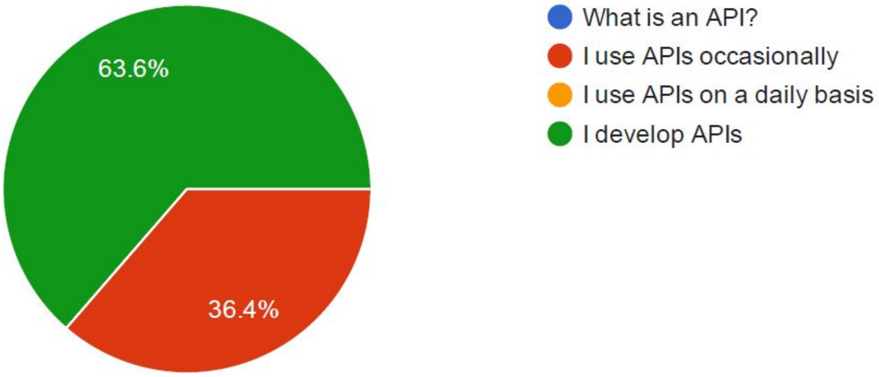


Figure 7. Experience with HTTP based APIs (REST or similar)

Learning: All participants identified themselves as developers have at least occasionally used application programming interfaces (APIs) and almost 2/3 are actively developing them for their services. This clearly shows that APIs are now well established and an infrastructure based on harmonization and annotation of APIs and providing means to combine the services in a user-friendly way can be successful in reducing the fragmentation of available data and software. This will more and more remove the need for manual manipulations when preparing the data and going from one service to another and will make the services available to a much larger group of stakeholders independent of their expertise in programming/scripting.

Question: What is your experience working with service containerization?

Results:

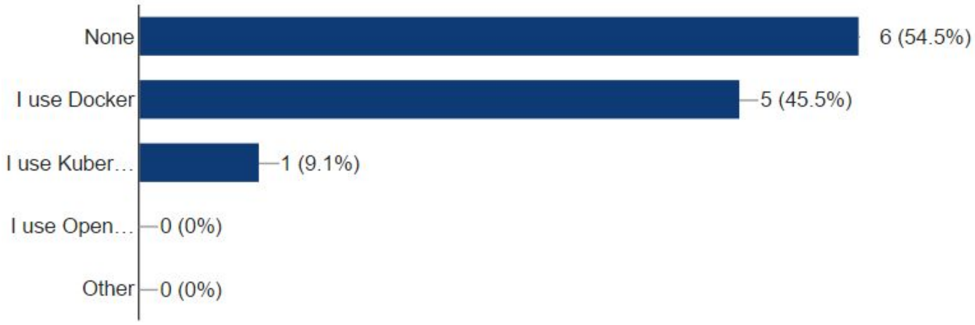


Figure 8. Experience with service containerization (Docker, Kubernetes, OpenShift)

Learning: Containerization and container orchestration environments (see also next questions) are slowly recognised as a useful way to allow deployment of software to very different hardware setting and to provide (commercial) services to clients with sensitive data, which cannot be transmitted to external services. Nevertheless, more than 50% of the developers are still not providing their software in this way. This shows that there is still high demand on extensive training options to equip more developers in academia and software companies with the know-how to create containers and complex deployment scripts as well as on good implementation examples to be provided by the OpenRiskNet consortium.

Question: In your organization, do you use service containerization?

Results:

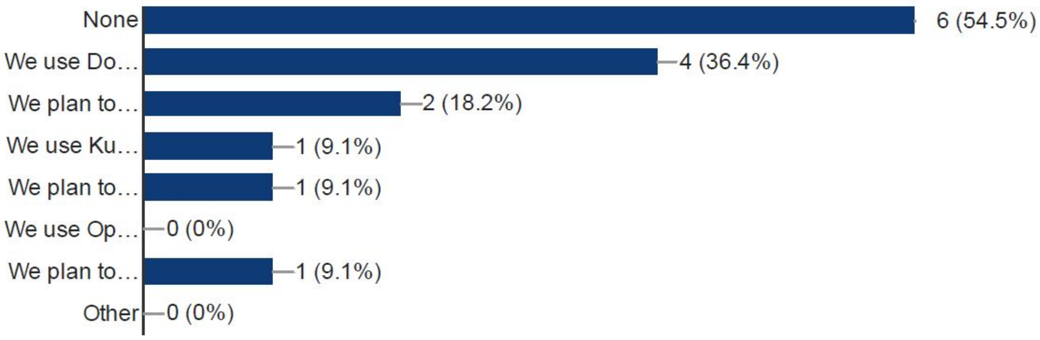


Figure 9. Use of service containerization (Docker, Kubernetes, OpenShift) within the organisations

Question: What programming language(s) are you using in your development?

Answers:

- Java
- Python
- JavaScript
- R
- Ruby
- C
- SQL
- GO
- PHP
- Scala

Learning: Many different programming languages are used in the different groups. This will not change in the future since the groups have collected large expertise with the chosen language, considerable amount of code exists and reused for new projects, and different languages have features making them especially useful for a specific application. This means, however, that a harmonisation of services to make them more interoperable cannot be performed on a code bases and trying to force new services into one of the existing modelling platforms will not be successful. Seeing this point clearly highlighted in the answers to this question, strengthens our confidence that the combination of microservices with well defined APIs, containerization and container orchestration is the more sensible way and building the necessary infrastructure including easy deployable VRE will be most beneficial to the developers and their task to integrate their services.

Question: What version control system are you using?

Answers:

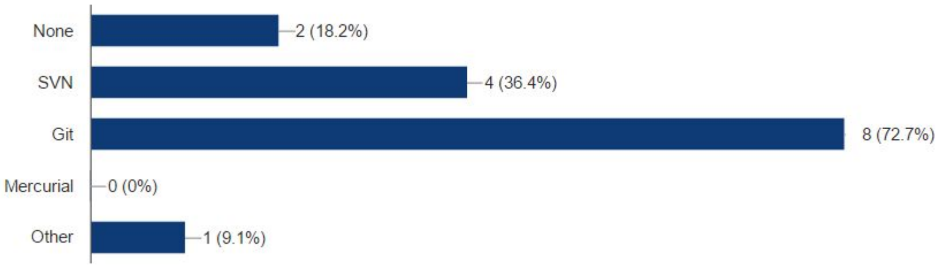


Figure 10. Version control system used in development

Learning: Git and SVN are the two most used version control systems and and it was decided to also use the first in OpenRiskNet to allow for collaborative development and sharing of code, scripts, documentation and workflows.

Question: Are you using a continuous integration and continuous deployment system and if so which?

Answers:

- Jenkins
- Manually solutions

Learning: This is clearly a subject where improvements can be made among many participants. We see from previous questions that several providers work with containerization, but that deployment strategies are commonly manual. We note that moving towards a CICD system will undoubtedly improve the stability and maintenanancibility for many tool providers and developers. To facilitate the adoption, the CI/CD features available in a VRE (provided by OpenShift) are comprehensive and include the use of Jenkins.

Question: What kind of *in silico* approach and data management tools do you develop?

Answers:

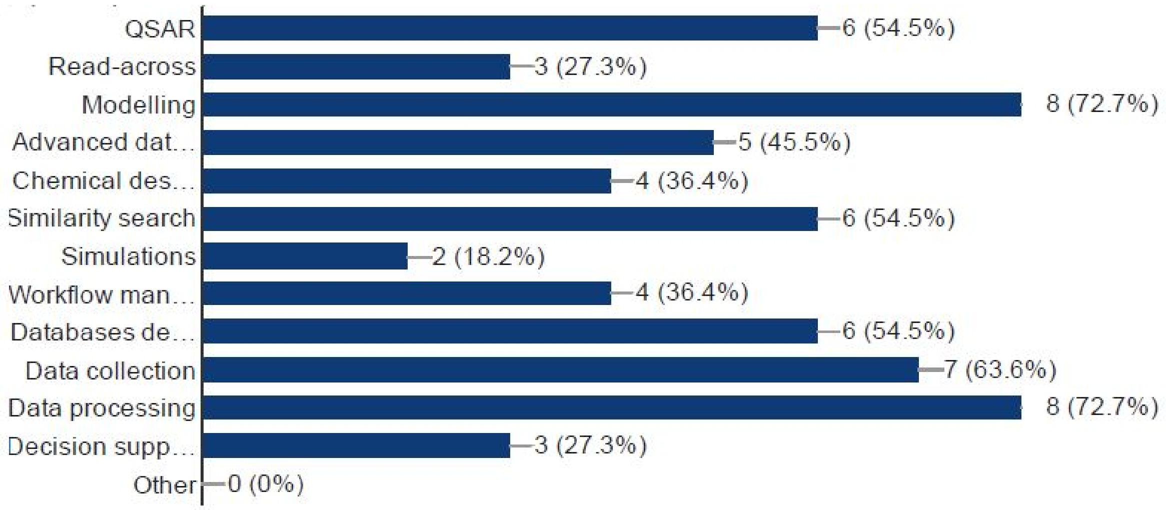


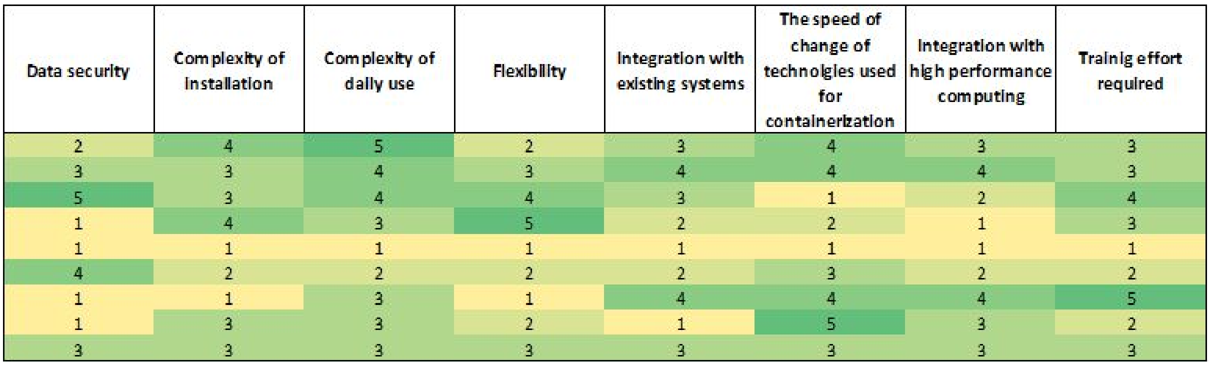
Figure 11. Type of *in silico* approach and data management tools developed

More specifically, these modelling tools are used to develop web interfaces, build data warehouse and models, as well as to answer research questions by integrating computational toxicology for prediction of chemical or nanomaterial safety.

Learning: The answers to this question show a good balance with tools from all areas. This gives us confidence that the conclusions drawn from the survey are not biased to a specific type of application and that the proposed solutions are of equal interest. It also shows that the participants build a good reference group to further discuss requirements to make the infrastructure relevant for all areas of the complete risk assessment framework.

Question: Do you worry about any of the following for using containerised solutions for your day to day work?

Answers:



Data security	Complexity of installation	Complexity of daily use	Flexibility	Integration with existing systems	The speed of change of technolgies used for containerization	Integration with high performance computing	Trainig effort required
2	4	5	2	3	4	3	3
3	3	4	3	4	4	4	3
5	3	4	4	3	1	2	4
1	4	3	5	2	2	1	3
1	1	1	1	1	1	1	1
4	2	2	2	2	3	2	2
1	1	3	1	4	4	4	5
1	3	3	2	1	5	3	2
3	3	3	3	3	3	3	3

Figure 12. Concerns for different situations or areas by using containerised solutions (1=very worried; 5=less worried)

Learning: The answers to this question draws a very mixed picture. Probably the answers of one individual person show more their general opinion with respect to containerization. For example, one person is very worried about all aspects, which shows a general skepticism. Data security seems to be the top issue in most replies, which is addressed by OpenRiskNet with the option to deploy the infrastructure with all requested services inside the premises of the user avoiding any transfer of data outside the intranet.This was a concept followed by the PhenomeNal project, which was dealing with personal data with its very strict data protection requirements.

Question: Can you name any tools you developed which would be suitable for integration in the OpenRiskNet e-infrastructure?

Answers:

Table 1. Tools identified by responders as suitable for integration in the OpenRiskNet e-infrastructure

Tool	Description	Link
AP-Portal	Action Potential prediction online part of which is a decoupled web service WSDL to ApPredict invocation	https://bitbucket.org/gef_work/ap_predict_online/overview
ChemDIS	Chemical-disease inference system based on chemical-protein interactions	http://cwtung.kmu.edu.tw/chemdis
SkinSensDB	Curated database for skin sensitization assays	http://cwtung.kmu.edu.tw/skinsensdb
MetFrag	*In silico* fragmentation for computer assisted identification of metabolite mass spectra	https://msbi.ipb-halle.de/MetFragBeta/
Online chemical database	Online chemical modelling environment (OCHEM): web platform for data storage, model development and publishing of chemical information	http://ochem.eu
EPA’s Chemistry Dashboard	The Chemistry Dashboard is a part of a suite of databases and web	https://comptox.epa.gov/dashboard

	applications developed by the US Environmental Protection Agency's Chemical Safety for Sustainability Research Program.	
N/A	A nonlinear least squares and entropy-based distance metrics built on top of the CERN-Colt matrix algorithms for high speed matrix operations in Java. These are linked to JFreeChart and the Apache poi Excel interface as UI's.	
SCAIView	A user-friendly search environment with a query builder supporting semantic queries with biomedical entities	http://www.scaiview.com
JProMiner	ProMiner is a tool for specific terminology recognition and addresses several fundamental issues in named entity recognition in the field of life sciences	https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/products/prominer.html
Lazar and Nano-lazar		
QSAR-report gem library		
CPSign		
Bioclipse		
MetaPrint2D		

Learning: A number of very interesting tools were proposed, with some of them covering areas not targeted by the services of the OpenRiskNet consortium. The developer of all these tools were contacted and they were included in the OpenRiskNet implementation roadmap. Planning of the needed steps and the implementations has also already started in some cases. More details on the status of implementation and the release of OpenRiskNet-complaint version are available in the deliverables of WP4. Making this high-demand services available was sometime delayed by important developments needed in the original services. In such cases, the OpenRiskNet consortium was creating was to at least provide some features of the services. One example is the chemistry dashboard developed at the US EPA, for which the release of APIs was postponed and is now planned for Spring 2019. Even if the full functionality of this service can only be integrated after this release, access to the ToxCast and Tox21 data via APIs was already provided by OpenRiskNet.

Requirement analysis results: end users' requirements

In the case of possible e-infrastructure end users (e.g. researcher, risk assessor, manager, educator) the aim was to learn on the data type, modelling, analysis and data management tools (including eventual experience in programming) and computational systems used for daily work. Further, we aimed to identify the type of data generated by users and the approach for data storage or sharing, including the security requirements. Further, we investigated in detail the requirements with regards to ontologies, chemical descriptors and data annotations.

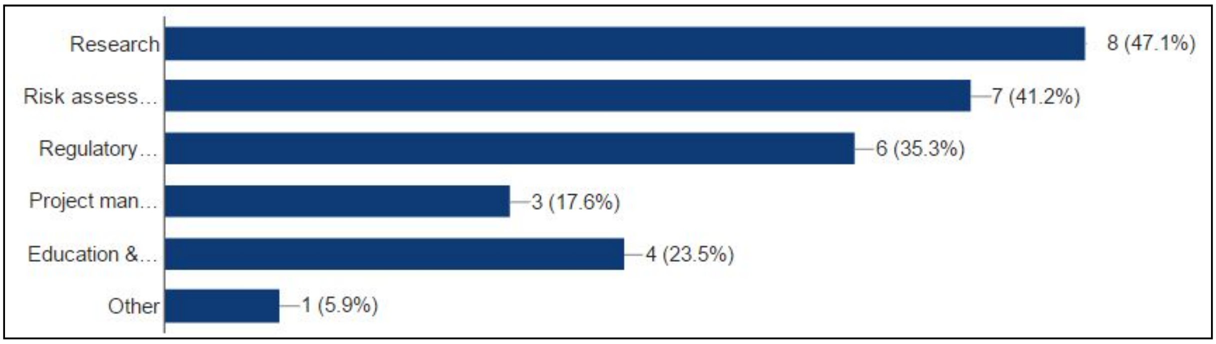


Figure 13. Distribution of roles of responders

Question: What kind of data are you using (e.g. physicochemical properties, exposure, hazard or other biological data) and from which sources?

Answers:

Table 2. Type of data in use and their sources

Data type	Data sources
A mix of data (physicochemical properties, exposure, hazard or other biological data) is used by majority of end-users. More specific type of data: ● Transcriptomics data ● Phys-chem properties and toxicological information on nanomedicines ● Ion channel screening data	● EPA ● IARC ● WERCS ● REACH ● NANoREG ● Scientific publications ● Own experiments ● Directly from company records ● Other online data repositories not specified

Learning: As already seen from the answers from the service providers, a large number of different data sources is of relevance for the predictive toxicology and risk assessment area. However, besides the public sources, features to bring data in different format from own experiments, company records or extracted from literature into the infrastructure have to be provided. To allow for the combination of the internal and external data sources, data curation and semantic annotation has to be supported.

Question: What kind of analysis and modelling tools are you using? Give few examples.

Answers:

Table 3. Type of analysis and modelling tools used

Analyses type or purpose	Tools
● Toxicological analysis, read-across, model prediction analysis, chemical structures, NOAEL/BMD and uncertainty factors for hazard assessment	● OECD QSAR Toolbox ● 3D-QSAR ● YASARA molecular modelling suite. ● ToxTree ● VEGA ● TEST ● Derek ● Leadscope ● Multicase ● EPA’s EPI ● ChemAxon Instant JChem
● Risk assessment	● ECETOC TRA ● Stoffenmanager nano ● CB ISO12901-2 ● EUSES ● IUCLID6
● Mathematical models of electrophysiology	● ApPredict software
● PBPK and Reaction Kinetics models	● SBML
● Multicell modelling	● COmpucell3D
● Statistical modelling	● JMP (SAS)
● Bayesian network inference ● ODE modelling ● Data integration and Visualisation ● Statistical analysis ● Kinetic models for exposure assessment	● Not specified

Learning: The answers show again the fragmentation of the predictive toxicology and *in silico* risk assessment field. Even if there is a very small number of software used by a larger subset of the users, many tools are only mentioned once or twice. OpenRiskNet is providing the technical solutions to bring all these into one common platform/infrastructure but cannot integrate all these tools during the run time of the project even with the help of the associated partner and the implementation challenge. Therefore, the developments will focus on a specific set of services based on the requirements of the case studies, which will then function as examples guiding the implementation of additional services as part of the sustainability efforts outlined in the dissemination and exploitation plan.

Question: Do you have experience in programming and/or workflow management tools?
Answers: Half of responders have no programming experience, while the other half has experience in using some tools (e.g. C++, Matlab)

Learning: The answers to this and the following two questions show the diversity of the end users. The can be placed on a continuous scale going from users, who use their programming and scripting skills to develop analysis and modelling approaches optimised for their current needs using combinations of existing software with self-made core (mainly academic and industry researchers), to users, who use graphical user interfaces of specific software (often risk assessors and regulators). To have most impact both in research and regulatory setting, OpenRiskNet has to serve all of these different expectations (flexibility vs. ease to use) and backgrounds.

Most users are familiar with single tools, most of them stand-alone software and databases or individual web services. Combination of tools into workflows other than hand-made scripts are only rarely used direct calls to APIs are only done by very experienced programmers. Additionally, moving to other tools is seen as complicated and often avoided, even if this new tool would provide better functionality and accuracy.

To flatten the learning curve when moving between tools and to promote combinations of tools to optimise the results (e.g. consensus modelling), the OpenRiskNet consortium came to the decision to strongly concentrate, beside the integration of databases and modelling tools, on the development and integration of workflow tools. Workflows presenting the capabilities of tools, their optimal combination and the benefits with respect to advanced and customised functionality when accessed via APIs can be shared between the consortium and the users as well as between users. Additionally, these workflows are excellent tools for training courses and webinars since they can be easily reproduced by the participants. Users more experienced with programming and scripting should be equipped by the consortium with workflows e.g. for retrieval of data or generating a predictive model in the form of Jupyter notebooks. These are easy to adopt to the specific needs and, in this way, open up the full potential of the integrated services. Users with less programming experience will still be able to run the standard workflows to perform tasks often needed in predictive toxicology and risk assessment. However, for such users, workflow management tools with more advanced user interfaces like Squonk will be a better starting point to access the OpenRiskNet services and training material requiring that the same functionality as the Jupyter notebook (to the extend this is possible) can be provided. A third alternative somewhat in between Jupyter and Squonk are KNIME workflows, since KNIME has a high popularity in the community especially also with industry stakeholders and offers additional methods, which can be combined with the OpenRiskNet services.

Question: What is your experience in working with the command line?
Answers:

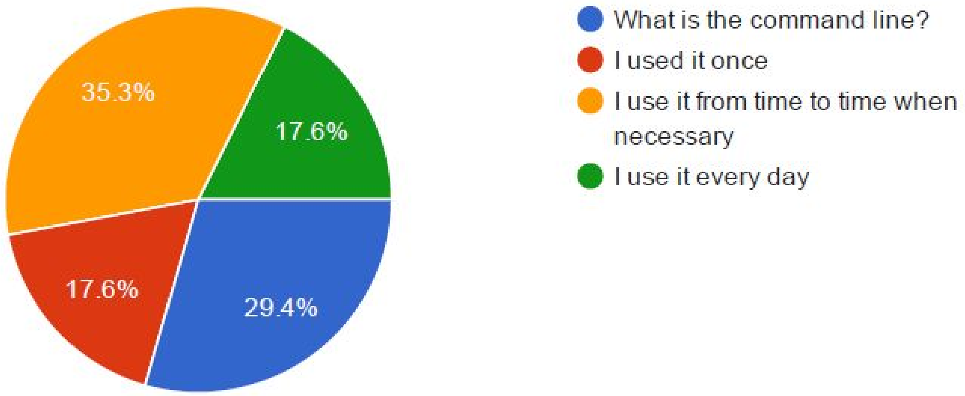


Figure 14. Experience in working with the command line

Question: What kind of computational infrastructure do you use (hardware / software)?

Answers: This question resulted in a diverse set of answers with references to different operating systems and software tools.

- C++ software on linux servers
- Toolbox on Windows 10
- Microsoft Office
- PC, Mac & Linux.
- SBML compliant tools
- Ontology tools such as Protege.
- Macintosh, PC, Microsoft Excel, Matlab, fortran, C++
- Linux and Windows workstations
- Cosmetopeia, Office, ECHA/REACH database
- MacOS, Windows

Even if workstations and servers were sometimes mentioned, this question has to be replace by a more specific one on the experience with HPC and cloud systems.

Question: What kind of data do you generate and to which database do you submit it?

Answers:

Table 4. Type of data generated by end-users and the database used for submission

Data type	Database
● Concentration-effect curves ● Models ● Docking and molecular dynamics data ● Data analysis ● OMICS data	● eNanoMapper ● NANoREG ● CEINT NIKS database ● BioModels ● ECHA database ● AP-portal database ● Internal databases

Frequently, the answer was that the data generated is not submitted to any database, and is stored internally.

Learnings: The answers show the large number of different data sources and methods used in the risk assessment field. The OpenRiskNet consortium already anticipated this during proposal writing due to their expertise in many previous and ongoing project. A number of the data sources correspond to the services listed in the proposal to be integrated by the consortium partners and their integration has already been started. This includes also options to bring internal data into the system. Others will be covered by associated partners integrating third-party tools partly supported by the implementation challenge.

Question: Do you see a benefit in accessing all data processing steps from raw data to final results? Do you see requirements differences for different type of experiments?

Answers:

- Yes - majority
- Important for understanding the gap between measurement limitations and actual levels that result in physiologic response
- Data should be accessible for traceability
- Access to raw and intermediate data may be necessary on case-by-case basis to improve confidence in risk assessment outcome
- For regulators all data processing steps in generating data for chemical risk assessment should be transparent, however, not every regulator needs to access each and every step her-/himself.

Learning: Improving repeatability and reproducibility of scientific results is of foremost importance and a struggle in all scientific areas even before the postulation of the reproducibility crisis in 2017[1]. Risk assessment performed in regulatory setting has to be fully traceable and the regulatory agencies demand that all information for computational analysis, QSAR and read-across applications are provided to allow to repeat (getting exactly the same result running the same code) and reproduce (ending up with same scientific conclusions form different approaches based on the same data) the results. OpenRiskNet will support these efforts to improve the quality of in silico method documentation. Following the workflow concept described above with methods to safe and share processing procedures including all information on individual steps and including the input and output as well as intermediate results will guarantee full traceability and improve repeatability and reproducibility.

Question: What level of data security would you require?

Answers:

[1] https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

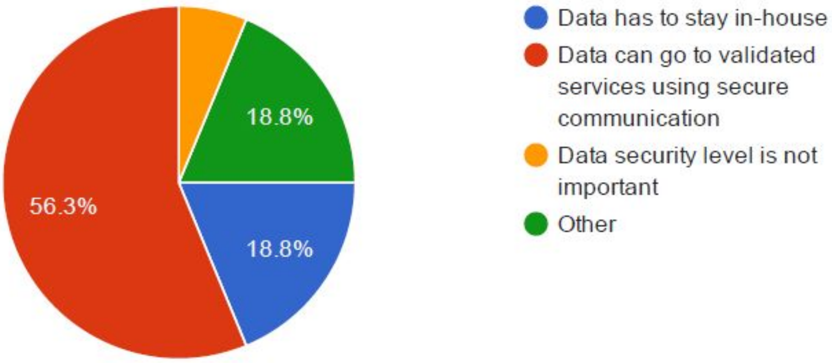


Figure 15. Level of data security required

Learning: For all stakeholders, data security if important and especially for industry participants clear requirements to keep the data local were stated, not to jeopardise intellectual properties in form of confidential chemical structures or specific formulations. OpenRiskNet has from the beginning specifically integrated such requirements in its infrastructure concept. The use of virtual research environments, which can be deployed in-house providing the basic infrastructure with all requested services, guarantees data security since all services will run locally and there is no need to transfer data outside the premises of the user.

However, the integration of validated public services into workflow to obtain the best results e.g. for the case studies seems to be also an option for more than half of the participants, as long as the data transfer uses secure communication. This could be a temporary solution until the service is ready to be deployed in-house or a permanent solution for services too big for efficient deployment like e.g. the ToxCast/Tox21 datasets or omics data.

Question: What kind of terminologies / ontologies do you use for annotating and retrieval of data?

Answers:

- Multiple bio-ontologies available, for example BioPortal.
- eNanoMapper ontology (for nanoparticles ontology)
- OECD, IUCLID endpoint terminology, WHO IPCS Risk Assessment Terminology
- Ontologies and platforms developed by end-users, OWL compliant, own keyword systems

Learning: The answers to this and the next two question show that there are still many confusions about identifiers and particular chemical (substance and compound) and gene IDs as well as the benefits of semantic annotation of data and tools. Almost everyone agrees that semantic annotation is a very important topic. However there are still many open questions and the benefits have still to be shown more widely before being able to overcome the reluctance to do additional work. As a consequence, OpenRiskNet has initiated the ontology task force providing the concepts how this annotation can be done technically and partly automated and it is now reaching out to other projects to align with other ongoing ontology efforts and organizing common training offers. Service and data annotation are performed now on the OpenRiskNet services in a stepwise manner. In this way, we will be able to show working solutions with some real-world examples as soon as possible.

Questions: What kind of identifiers / descriptors would you prefer for retrieval of data? e.g. SMILES, CAS-RN, HGNC, ENTREZ GENE, etc. Do you use automatic tools to convert / harmonise these identifiers?

Answers:

- CAS-RN and SMILES (majority)
- SDF
- EC numbers
- Since descriptors are domain specific (eg. SMILES makes no sense for a gene sequence) it is not possible to have a "preferred" identifier/descriptor. As long as the identifier is uniquely linked to a particular URI/URN then it is acceptable to me.

Question: Would you like to add meta-data / annotations to data points, data sets, tools or workflows (this could be comments, keywords, descriptions, references/links to other objects, discussions, etc.)?

Answers:

- Yes (majority)
- Some of this would be useful - raw dataset->postprocessing to get simulation inputs->simulation inputs->mathematical model->simulation outputs.
- Meta-data / annotation is currently the weakest aspect of data sharing.

INDIVIDUAL INTERVIEWS

The last question of the survey was if the participant would agree to a follow-up in form of an interview. Most of the participants agreed and the interview process has been started and will be continued throughout the project. Minutes of each interview are available to all partners. Here, we highlight the outcome of one interview, which was used to define specific risk assessment workflow requirements.

Risk assessment workflow requirements

With this requirements analysis we aimed to identify some specific needs of end users working in risk assessments, in terms of workflows, data and software used. The risk assessment process usually involves more expert groups, performing different steps, e.g.:

- Group 1 setup the risk assessment workflow, establish the search terms and keywords, identify and list databases containing that information and write instructions on how to curate the information);
- Group 2 follow and apply these instruction to collect relevant publications, curate, extract and organise the relevant information;
- Group 3 performs the risk assessment based on the information provided in the previous steps, asks for clarifications and concludes on the assessment (expert judgment step).

The whole process and the experts involved in this workflow require automatic and interoperable tools which can facilitate their daily work, and supports the assessment of information, drawing the conclusions and taking the decision.

Primary sources of information

As the survey results shown, the primary sources used for retrieving information useful in the risk assessment process are represented by the **publicly available databases provided by the European and national (regulatory) agencies or other recognised international institutions, like OECD and WHO**. These sources represent a rich source of information, in a form of case study reports. These detailed reports are also using data and toxicological information from other sources, most commonly curated from the available scientific publications. Thus, for a risk assessor, one primary source of information is represented also by these scientific publications (reviews and original papers), where details on the studies can be identified. **PubMed**[2] represents an important and usually a first entry point for searching such publications.

When the search starts from a compound or a chemical identification is needed, the information is curated from databases like **ChemIDPlus**[3], **OECD QSAR toolbox**[4], **IUCLID**[5], but always the approach depends on the case studied.

One important requirement is related to the **repeating of searches, which would be valuable for retrieving updates**, but is not commonly done because of missing functionality.

[2] https://www.ncbi.nlm.nih.gov/pubmed/
[3] https://chem.nlm.nih.gov/chemidplus/
[4] http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm
[5] https://iuclid6.echa.europa.eu/

Data sets

The risk assessor has also to work with specific values and parameters, but they need also to look at how the data was produced and get access to the original data sets. Such data sources are curated manually at the moment (if at all) and no workflows to automate this are available. Some dashboards are available and could be used to some extent (e.g. **iCSS ToxCast Dashboard**[6]).

Systematic review

The workflow includes accessing the original source and extracting the information and data needed:

- The initial steps include a preliminary data filtration, source filtering and search by specific term and keywords;
- The starting points are most often the scientific reviews from which the primary literature is extracted;
- The process depends on how much relevant literature there is on that specific subject;
- Filtering options and semantic searches for having more condensed information is an important and useful feature;
- Also, standardised forms for performing the systematic review can be used (e.g. **ToxRTool - Toxicological data Reliability Assessment Tool**[7])

Integrated testing strategies (ITS) in regulatory decision

To accept results from *in vitro* studies for regulatory decisions, a few requirements have to be fulfilled in terms of their reliability. Thus, for considering such studies, an existing consensus from international bodies is required (validated tests by international institutions like EC-JRC ECVAM and ICCVAM, opinions or recommendations of a recognised scientific committee, or already included in an OECD test guideline. However, for regulatory purposes in some areas the *in vitro* data is not yet the mainstream of the risk assessment process. The ITS should support such integration but is still at an early stage of development, at least when a regulatory decision is concerned. For *in silico* data, some parameters are generally well accepted e.g. logP, protein binding, and using tools like the **OECD QSAR toolbox** is very useful for generating such parameters. However, in both cases (*in vitro* and *in silico*) a combination of methods is usually needed in order to support the expert's judgement and to be used in a regulatory decision process. Also, the Adverse Outcome Pathway (AOP)-based risk assessment and the development of the AOP approach (which provides information on the adverse outcome of regulatory concern), facilitates such data integration and supports the use of ITS for regulatory decisions. However, the confidence in the output of a method is very important in drawing a risk assessment conclusion and taking a regulatory decision.

Thus, OpenRiskNet can help by providing **multiple tools for consensus building** and **reporting confidence levels**.

[6] https://actor.epa.gov/dashboard/

[7] https://eurl-ecvam.jrc.ec.europa.eu/about-ecvam/archive-publications/toxrtool

ONGOING REQUIREMENTS ANALYSIS AND ITS INFLUENCE ON THE IMPLEMENTATION AND WORK ON THE INFRASTRUCTURE AND CASE STUDIES

As described above, the first requirements analysis performed until M6 of the project confirmed that OpenRiskNet is targeting the highly relevant problem of the fragmentation of data and software tools in the area of predictive toxicology and risk assessment. This leads to large amounts of manual work to bring together the needed data in a form acceptable for a specific tool, steep learning curves to be able to use a new tool, replace an old one or get multiple tools working together in a consensus approach, and makes it almost impossible to get the optimum out of one tool or the combination of multiple tools if this is not implemented in specialised workflows most of the time only accessible through a tool-specific graphical user interface. The general concepts described in the proposal in which the infrastructure is based on virtual research environments, to which containerised microservices can be deployed, which communicated via semantic annotated APIs, provides all necessary features to overcome barriers towards more harmonised, interoperable and in this way, easier to use analysis, modelling, prediction and assessment services. However, the requirements analysis also showed that some of the proposed approaches have to be refined and optimised. For example, defining a standardised and harmonised API for all services is not possible since, on one hand, this would put a large burden on the service developers, who would have to adapt their software to this API definition, and, on the other hand, the large amount of different areas, from which the tools are coming, would make the API definition very complex to be able to fulfill all the different requirements and endpoints. Therefore, we adapted the API concept to the bottom-up approach described in report D2.2 Initial API version provided to providers of services, which provides more flexibility on the API endpoints as well as input and output options but puts more pressure on the semantic interoperability layer, which has to provide all information on how to link services via the semantic annotation. A second area, in which the requirement analysis together with the experience from other projects and a literature review directly influenced the direction of the project in a large extent was the selection of the first set of 7 case studies. These are trying to group the data sources and software tools mentioned in the answers to the survey into application domains and relate these to recently developed risk assessment frameworks and approaches followed in the risk assessment community formalised e.g. in the read-across case studies of the EU-ToxRisk project. A publication summarizing the results from the survey and interviews, the learnings extracted from these and the resulting concept changes and case study design is in preparation.

After integrating all the feedback and learnings into the design, it is now important to have constant feedback on if the provided solutions are really addressing the requirements and are fit for purpose. Additionally, new requirements will probably surface once the infrastructure is used for real-world applications. Therefore, we have designed a system outlining measures to foster stakeholder involvement to cover additional requirements and evaluate the infrastructure regarding its fitness for purpose.

1) We have updated the requirements survey to include more specific questions on solutions provided by OpenRiskNet. This **second version**[8] will be kept online during the whole duration of the project and further extended with additional specific questions if needed. Added questions include e.g. "OpenRiskNet is prioritising and testing its approaches based on case studies related to specific areas of risk assessment. Which case studies are relevant to your area of research? And would you be willing to support their development and execution by providing expertise on how the requirements and challenges could be met?", "OpenRiskNet offers multiple support options. Which one is most relevant to you? What else would you like to see?" and "OpenRiskNet offers two main ways of accessing the services (home.prod.openrisknet.org): visual workflow managers like squonk and easy ways to develop and share scripts (python/R) executing workflows in Jupyter notebooks. Which one would you use considering that examples will be provided by OpenRiskNet?“. Users accessing the reference infrastructure for the first time will be asked to fill in this survey or at least a shorter, anonymous version.
2) OpenRiskNet has organised a set of introduction and demonstration webinars, which attracted around 70 participants in total. These meetings included questions-and-answers sessions, which we used to probe for additional requirements and opinions on the chosen solutions. Additional potential users profited from the recordings of the webinars to get more information on the project triggering mail exchange with the consortium expressing specific requirements and possible interactions. The webinar series will be continued with more specific topics on the setup of the virtual environment and the usage of data, tools and workflows.
3) One question asked in the extended survey and in the introduction webinars is if there is the need to start additional case studies. The presentation given to the US National Cancer Institute Nanotechnology Working Group and follow-up discussions with the ACEnano and NanoCommons project resulted in the request for a physicochemical characterization case study as part of nanomaterial risk assessment, which will be performed in collaboration of OpenRiskNet with the two nano projects. Similar initiatives are also possible with other EU or international projects.
4) OpenRiskNet was presented at major conferences like SOT and EuroTox, with posters specifically about the infrastructure layout and as part of the Douglas Connect booth in the exhibition area, as well as within specific sessions at OpenTox Euro 2017 and 2018. These opportunities were used to have short interviews with people interested in the project and its solution. This will continued with the next major event at SOT 2019 in Baltimore.
5) Organizing additional longer interviews with specific stakeholders will also be continued. This will especially include scientific advisory board and participants of the implementation challenge. The applications to the first round of the challenge showed that especially SMEs providing services are highly interested in integrating their tools in the infrastructure with the additional requirement to handle commercial licenses.
6) Finally, other EU projects like Eu-ToxRisk, NanoCommons and ACEnano start to integrate OpenRiskNet concepts and tools in their knowledge infrastructure. Additionally, the first in-house virtual research infrastructure was deployed at the Diamond Light source. These real-world applications will show in the near future if

[8] https://goo.gl/forms/4O2DuF8Fy4suf7gq2

the OpenRiskNet solutions are fit for purpose and/or where the concepts have to be adapted, optimised or even redesigned to cover new, unexpected requirements.

CONCLUSION

The information on the status quo with respect to the used software tools (development tools as well as predictive toxicology and risk assessment software) and data sources, technical knowledge of the users, safety concerns, data annotation and protocoling requirements obtained from the survey and the interviews can now be used to optimise specific parts of the infrastructure concept improving the usability and the user-friendliness of the solution. All participants in the requirements analysis agreed that more harmonization and improved interoperability are of major importance for lowering the level of expertise to run predictions, open these approaches to a broader group of users, collecting more experience on the quality of the approaches, compare different approaches and, in this way, provide the information for validation of new integrated testing methods for regulatory usage. For example, the access to multiple tools providing predictions for one endpoint open up the possibility to compare the outcome and consensus building was highlighted in an interview as a way to improve the confidence level in these approaches. Additionally, also the need for better data annotation with ontologies and the benefit in accessing all data processing steps from raw data to final results were acknowledged by the majority of participants. One participant answered that traceable and reproducible data treatment is “important for understanding the gap between measurement limitations and actual levels that result in physiologic response” and another that “access to raw and intermediate data may be necessary on case-by-case basis to improve confidence in risk assessment outcome”. Data quality and reproducibility were in the main focus when designing the infrastructure in the proposal writing process, a decision, which is now confirmed by the requirements analysis.

One very positive outcome from the analysis is that the technical expertise of the developers is high. The majority develop APIs with their software and many have experience with advanced deployment options. Docker is more or less the only tool used for containerization, which conforms the decision to concentrate on this option in OpenRiskNet. Container orchestration (Kubernetes, OpenShift) and continuous integration and continuous deployment system are less well known and the communities could profit from training units provided by OpenRiskNet.

The main outcome from the end-user survey is that a large variety of tools for many different applications is used. Even if it is clear that not all these tools can be integrated by the OpenRiskNet consortium, at least one representative tool for each area is available from the partners or was proposed by the developers answering the survey. These will be used to demonstrate and document the steps needed for integration and hopefully inspire additional developers to provide their tools as similar services. The answers also show that, in contrast to the developers, the computer expertise of end users is frequently limited to the use of stand-alone tools on the Windows or MacOS platform and to the usage of web interfaces for database access. It is very important to consider this point in the user experience design of the infrastructure. The goal has to be that at the final stage of the infrastructure development combining tools into workflows is possible by just a few clicks and deployment is as simple as possible even if this task could be delegated to the system administrator of an organisation. An additional issue becoming evident in the survey is that many researchers do not submit their data to public databases. Even if this is not in the focus of OpenRiskNet, it clearly shows that there is a need to make this upload process easier and provide tools for preparing the data in a form ready for submission including annotation with ontologies. Challenges to sustain infrastructures arising from workflows that were too complicated were experienced in the OpenTox and ToxBank projects. The OpenTox approach to create validated QSAR models was only adopted by few developers outside the consortium. Similarly, data upload to the ToxBank data warehouse was slow and had to be supported by the ToxBank team even if the chosen approach complied to the highest quality standards.

Even if the answers provided valuable feedback and the conclusions were integrated in the design and implementation of the infrastructure including the API concept, the semantic interoperability layer but also, perhaps even more important, the case studies, it has now to be constantly monitored if the provided solutions are fit for purpose and are really able to fulfill all these requirements. To achieve this, multiple ways to cover stakeholder feedback including an updated survey, additional interviews (virtual and face-to-face at different conferences, exhibitions, workshops and project meeting) will be followed throughout the project with a shifting focus on service providers from academia and SMEs in the first finished phase of the project to end users from industry, risk assessment consultancies and regulatory agencies in the remaining time. One specific planned event is a shared booth in the exhibition area of SOT 2019 in Baltimore probably in combination with a sponsored session for industry stakeholders.

This ongoing stakeholder involvement will help to collect additional requirements and especially also first user experiences now that the reference environment is operational. We have and will continue to include more specific questions in the updated survey and the interviews to ask the users explicitly about the usability of the chosen concepts and approaches. Requirements analysis performed in other projects (e.g. eNanoMapper, ToxBank) are also continuously integrated more closely for identifying specific requirements related to the areas for which they were prepared (e.g. nanosafety, alternative methods to animal testing). This will help to more convincingly show the benefits of the new infrastructure and the multitude of possibilities opened up by the integrated services and, in this way, to be able to attract more participants and potential users from larger companies and from the risk assessment and regulatory sectors, which will then perform a deeper usability analysis and validation required for acceptance of the provided solutions in these sectors and the sustainability of the OpenRiskNet infrastructure.

GLOSSARY

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

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