OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Deliverable Report D4.1

Report of the Service Integration with OpenRiskNet (Initial Deployment)

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-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 4.1 Report of the Service Integration with OpenRiskNet (Initial Deployment)
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	UM
Author(s)	Danyel Jennen (UM), Tim Dudgeon (IM), Haralambos Sarimveis (NTUA), Philip Doganis (NTUA), Pantelis Karatzas (NTUA), Micha Rautenberg (JGU), Egon Willighagen (UM), Thomas Exner (DC)
Status	Final
Version	V1.0
Document history	2017-11-29 Draft version 2017-12-19 Final version

Table of Contents

SUMMARY

This report describes the status of selection of services of high priority for the OpenRiskNet infrastructure and their integration including active services provided by the consortium, associated partners and other third parties.

The work described in this report addresses the following tasks:

- Task 4.1 Toxicology, Chemical Properties and Bioassay Databases
- Task 4.2 Omics Databases
- Task 4.3 Knowledge Bases and Data Mining
- Task 4.4 Ontology Services
- Task 4.5 Processing and Analysis

In addition, work done in relation to Task 4.6 Predictive Toxicology and Task 4.7 Workflows, Visualisation and Reporting are also included as far as applicable. Due to their importance for service integration, we also reference to work performed in WP2 - Interoperability, Deployment and Security and WP3 - Training, Support, Dissemination.

INTRODUCTION

The OpenRiskNet functionality will be composed of and empowered by incorporating a variety of existing services (databases, knowledge bases, and preprocessing, analysis and modelling tools) that have either been developed by the consortium partners, prior to the start of the project or in other ongoing projects, or are publicly available. Work was started to harmonise these services and to add the semantic annotation for improved interoperability after various aspects such as the technical requirements of the programming interfaces, the deployment options, the security environment and the discovery services developed in WP2 have reached a state mature enough for the first version of service integration. Moreover, the definitions of standards and recommendations regarding file formats, ontology usage and technical and scientific descriptions proposed by WP3 support this integration and the generation of a powerful OpenRiskNet tool box. More and more services provided by the OpenRiskNet partners will now become part of the verified OpenRiskNet service directory and will be used as best-practice examples for guiding the developments by OpenRiskNet partners and third parties (associated members) needed to integrate additional services.

In the following, we will present the selection and prioritisation of the services, detail the first deployment of a service into the virtual research infrastructure and preliminary versions of semantic annotations of computational tools and data (*see Deliverable 3.2*) as well as highlight the relationship between the services and the case studies and use cases developed in WP1 for evaluating the usefulness of the OpenRiskNet e-infrastructure (*see Deliverable 1.3*).

SERVICE SELECTION

Work in WP4 officially started in month 7 of the project. However, the list of services available for implementation, first proposed in the description of action, was extended from the start of the project resulting in the updated list of **Table 1**. Not only adding additional services were selected but also information on available APIs, licences and integration status are now provided, will be made publicly available on the OpenRiskNet website shortly and updated there continuously.

Table 1. List of services ready for implementation provided by the partners and important external services

Service	Partner	Third Party	Service	Partner	Third Party
Toxicology, Chemical Properties and Bioassay Databases					
eChemPortal	UM	X	eNanoMapper	DC	
ChEMBL	UM	X	ToxCast	DC	X
ChEBI	UM	X	ToxRefDB	DC	X
ChemSpider	UM	X	FDA DILI	DC	X
PubChem	UM	X	ECHA DB	DC	X
DrugBank	UM	X	EPA CompTox Dashboard	UM	X
ToxNet	UM	X	Wikidata	UM	X
ToxBank	DC				
Omics Databases					
diXa	UM		ArrayExpress	NTUA	X
EGA	CRG		TG-GATEs	UM	X
Gene expression omnibus	NTUA	X			
Knowledge Bases and Data Mining					
AOP KB	UM/DC	X	KEGG	NTUA	X
CTD	UM/DC	X	Reactome	NTUA/UM	X
ConsensusPathDB (CPDB)	UM	X	JaqPot Quattro	NTUA	
Triple Annotator Tool	UM		SCAIView	Fraunhofer	
OmniPathDb	UM	X	UIMA	Fraunhofer	X
WikiPathways	UM		JProMiner	Fraunhofer	

Pathway Annotation Tool (PathVisio)	UM	X	BELIEF	Fraunhofer	
Identifier Mapping Tool (BridgeDb)	UM		Decision Tree Learners (J48, M5P)	JGU	partially
ID mappings from BridgeDb (Gene-Gene)	UM		Random Forests (RFs)	JGU	partially
ID mappings for BridgeDb (Gene-variant)	UM		Support Vector Machines (SVMs)	JGU	partially
ID mappings from BridgeDb (Compound-Cmpd)	UM		Partial Least Squares (PLS)	JGU	partially
ID mappings for BridgeDb (Reactions)	UM		Ridge Regression (RR)	JGU	partially
Chemistry Registration Service	UM		WEKA Machine Learning Algorithms	JGU	X
Gene Ontology	NTUA	X	Generic WEKA web service wrapper	JGU	X
Ontology Services					
Ontology Metadata Service	UoB		NCBITAXON		X
Ontology Reasoning	UoB		BioPortal		X
jQUDT	UM		Ontology Lookup Service		X
EDAM		X	AberOWL		X
CHEMINF	UM				
Processing and Analysis					
Toxygates	DC		Read-Across	NTUA	
Omics data analysis tools (TwinBoosting)	UoB		iClusterPlus	UM	X
Network modeling and analysis tools (DiME/CoCoMi)	UoB		ArrayAnalysis.org	UM	
Image Analysis	NTUA		ArrayQC	UM	
GO descriptor calculation	NTUA		MagiCMicroRna	UM	
RRegrs	NTUA/UM		RNA seq workflow	UM	
Optimal	NTUA		MeDIP-seq workflow	UM	

experimental design					
Dose-Response Modelling	NTUA		CDK	UM	
JaqPot Quattro	NTUA		RDKit	JGU	X
Java Modelling Algorithms	NTUA		OpenBabel	JGU	X
Python Modelling Algorithms	NTUA		gSpan'	JGU	
QPRF	NTUA		CPSign	UU	
Interlab Training	NTUA		GenePattern		X
Validation Services	NTUA		Babelomics		X
Predictive Toxicology					
Lazar	JGU	X	BBRC	JGU	
Nano-Lazar	JGU	X	LAST-PM	JGU	
PubChem Read Across	JGU	X	FCDE	JGU	
PBPK model	INERIS		PSCG	JGU	
Quantitative AOPs	INERIS		ELICIT QSAR	JGU	
PBPK Model	NTUA		Applicability Domain	JGU	
Workflow, Visualisation and Reporting					
Squonk	IM		CheS-Mapper	JGU	
Bioclipse	UU/UM		OpenTox Validation Services	JGU	
Risk21 App	DC		MDStudio	VU	

SERVICE INTEGRATION

Work has commenced on integrating partner applications into the OpenRiskNet infrastructure. This was initiated at a Workshop held at Uppsala on 25-26 September 2017 (coordinated by UU) where partners were provided with an overview of the OpenShift infrastructure that will be used to orchestrate the OpenRiskNet services and guidelines for what is needed to deploy applications to this infrastructure. More details of this can be found in the D3.2 report (First documentation of the core e-infrastructure).

Following this a development reference site has been deployed on the Swedish National resource SNIC Science Cloud (SSC) [1] and the first partner application, the Squonk Computational Notebook (IM) [2] has been successfully deployed to this site, and was demonstrated at the OpenRiskNet GA meeting in Basel on 20 November 2017.

Encouraged by this success, work is underway in all partner organisations towards integration, semantic annotation and deployment of additional applications, with the Lazar application (JGU) [3], JGU WEKA Rest service (JGU) [4], Jaqpot (NTUA) [5] and different data sources expected to be ready in the near future. Some highlights of the ongoing developments are:

1) The application services are implemented as close as it can be to the OpenAPI version 2 specifications and will incorporate version 3. The documentation of the API will use Swagger, which is basically a documentation tool for API's and provides many other tools for uses such as incorporating generating clients etc. A service discovery mechanism is under development for easy discovery and consumption of the services by end user applications including user interfaces such as Squonk or others like Jupyter notebook. This is achieved by extending the swagger tools with JSON-LD, in order to accommodate higher-level semantic annotation. JSON-LD is designed around the concept of a "context" to provide additional mappings from JSON to an RDF model. The context links object properties in a JSON document to concepts in an ontology. In order to map the JSON-LD syntax to RDF, JSON-LD allows values to be coerced to a specified type or to be tagged with a language. Following a predefined mapping, context can be embedded directly in a JSON-LD document or put into a separate file and referenced from different documents (from traditional JSON documents via an HTTP Link header). The OpenRiskNet approach is to create this context and pass it to the service discovery mechanism through the header of the http request that gets the swagger.json provided by the applications. Then a service discovery mechanism based upon SPARQL Protocol and RDF Query Language will be able to know the context of the services deployed on the OpenRiskNet infrastructure. The services will even be registered or discovered by this mechanism and the user can discover through SPARQL. Swagger will then let the user to digest a service provided in the application however he wishes. For usage of the services, a single-sign-on mechanism is provided and all the services that will be deployed will use this authentication mechanism. When a user is logged in through this mechanism, they will be able to use all the services provided by the OpenRiskNet.
2) Semantic annotation in the Jaqpot application in based on the adoption of the JSON-LD linked data format. Currently, JSON-LD is used for documenting and annotating models and algorithms. Specifically a JSON-LD serialisation is formatted when we send API calls to Jaqpot like getting an algorithm or a model schema with the HTTP header 'Accept: application/ld+json'. The API returns the

requested schema with the LD annotations and the context that JSON-LD dictates with all the semantic information included in that schema. Semantic annotation of Jaqpot applications and services will also use the JSON-LD format and is currently under development.

The examples below are working examples that highlight the differences between JSON and JSON-LD formats and the additional information that can be included in JSON-LD schema. Specifically, we present examples of an algorithm and a predictive model. The examples can be found in the swagger documentation of the API of the Jaqpot application [6].

JSON schema of an Algorithm produced by the Jaqpot API

```
{
  "meta": {
    "descriptions": [
      "An MLR algorithm by Weka"
    ],
    "titles": [
      "MLR - Weka Implementation"
    ],
    "subjects": [
      "mlr",
      "linear",
      "regression"
    ],
    "locked": false
  },
  "ontologicalClasses": [
    "ot:Algorithm",
    "ot:Regression",
    "ot:SupervisedLearning"
  ],
  "ranking": 2,
"trainingService":
"http://test.jaqpot.org:8090/algorithms
/mlr/training",
"predictionService":
"http://test.jaqpot.org:8090/algorithms
/mlr/prediction",
  "_id": "weka-mlr"
}
```

JSON-LD schema of an Algorithm produced by the Jaqpot API

```
{
"id":"http://test.jaqpot.org:8080/jaqpo
t/services/algorithm/weka-mlr",
  "type": "ot:algorithm",
  "title": [
    "MLR - Weka Implementation"
  ],
  "description": [
    "An MLR algorithm by Weka"
  ],
  "subject": [
    "mlr",
    "linear",
    "regression"
  ],
  "@context": {
    "date": "dc:date",
    "creator": "dc:creator",
    "audience": "dc:audience",
"enm":
"http://purl.enanomapper.org/onto/",
"owl":
"http://www.w3.org/2002/07/owl#",
"ot":
"http://www.opentox.org/api/1.1#",
    "subject": "dc:subject",
    "description": "dc:description",
"rdfs":
"http://www.w3.org/2000/01/rdf-schema#"
,
    "source": "dc:source",
    "type": "@type",
    "title": "dc:title",
    "seeAlso": "rdfs:seeAlso",
    "contributor": "dc:contributor",
    "rights": "dc:rights",
    "publisher": "dc:publisher",
"bibo":
"http://purl.org/ontology/bibo/doi",
    "comment": "rdfs:comment",
    "id": "@id",
    "parameters": "ot:parameters",
    "dc": "http://purl.org/dc/terms/",
    "doi": "bibo:doi",
    "sameAs": "owl:sameAs"
  }
}
```

Summary of JSON schema of a Model produced by the Jaqpot API

```
 {
  "meta": {
    "comments": [
      "Created by task lGQQt265w7be"
    ],
    "descriptions": [
      ""
    ],
    "titles": [
      ""
    ],
    "creators": [
      "guest"
    ], ……
"verbal.notes": [
"Ge  value  is:0.53.  Ge  for  optimal
design is 1.",
"Diagonality    value    is:    0.943.
Diagonality for minimal confounding is
1."
    ],
    "predictedFeatures": [
      "suggestedTrials"
    ]
  },
"transformationModels":[
"http://test.jaqpot.org:8080/jaqpot/ser
vices/model/yGEqlmZiZv7d46lbYcKM"
  ],
  "linkedModels": [],
  "_id": "6r7f1tPqIPaniPIqZwNX"
}
```

Summary of JSON-LD schema of a Model produced by the Jaqpot API

```
 {
"id":
"http://test.jaqpot.org:8080/jaqpot/ser
vices/model/6r7f1tPqIPaniPIqZwNX",
  "type": "enm:ENM_8000076",
  "title": [
    ""
  ],
  "description": [
    ""
  ],
  "creator": [
    "guest"
  ],
"source":[
"http://test.jaqpot.org:8080/jaqpot/ser
vices/dataset/corona-exp"
  ],
  "comment": [
    "Created by task lGQQt265w7be"
  ],
  ……
  },
  "@context": {
    "date": "dc:date",
    "transformationModels": {
      "@id": "enm:ENM_8000076",
      ...,
    "algorithm": {
      "@id": "ot:algorithm",
      "@type": "@id"
    },
    "creator": "dc:creator",
    "audience": "dc:audience",
features": {
      "@id": "enm:ENM_8000084",
      "@type": "@id"
    },
    "publisher": "dc:publisher",
    "linkedModels": {
      "@id": "enm:ENM_8000076",
      "@type": "@id"
    },
    "parameters": {
      "@id": "enm:ENM_8000088",
      "@type": "@id"
    },
    "dc": "http://purl.org/dc/terms/",
    "doi": "bibo:doi",
    "sameAs": "owl:sameAs"
  }
}
```

3) Existing data APIs for three well known toxicological data sets (ToxCast, ToxRefDB and Open TG-GATEs) were experimentally annotated with ontology terms on the OpenAPI layer. This proved to be a valuable stepping stone for semantic interoperability, but it was decided by the project partners that such an approach without a proper Linked Data support would be insufficient. Various standards were considered (e.g. HYDRA; see the WP 2 deliverables for more information) and the decision was made to annotate OpenAPI definitions as JSON-LD documents as described above. The work to upgrade the three data APIs to the new concept is currently ongoing. The DC Data Explorer, a tool that allows easy browsing and filtering of datasets exposed as APIs, is also being modified to utilise the richer semantic information provided by the updated data sources.
4) Work has been started to generate a general data model to semantically annotate datasets (a small part is shown in **Figure 1**). This will be used in the semantic interoperability layer to answer queries from other services on the available data provided by a data source, to select the requested data and to transfer it in the form agreed on by the provider and receiver service. Datasets from the EPA and the FDA are now manually annotated by OpenRiskNet partners to validate that this data model can cover these different data sources.

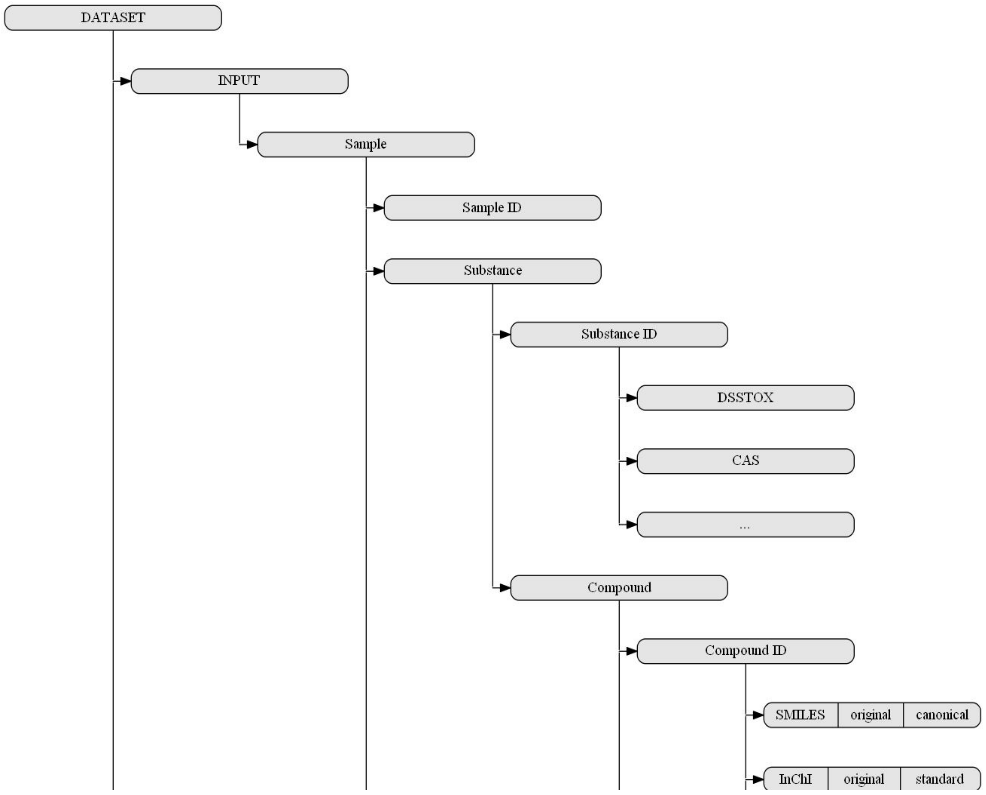


Figure 1. Schematic presentation of a part of the general data model for semantic dataset annotation

The status of the development reference site include a continuously updated list of available services can be seen here: https://home.dev.openrisknet.org/.

SERVICE PRIORITISATION

As becoming clear from **Table 1**, predictive toxicology and risk assessment require data and tools from a broad range of areas and many different services exist for each of these. Since these cannot be all integrated from the beginning, a rough time table has to be created to, on one hand, prioritise specific tools for early integration but, on the other hand, also to guarantee that all areas are covered since they could impose specific requirements on the e-infrastructure and the semantic interoperability layer. To identify areas in the risk assessment process, which would profit the most from OpenRiskNet solutions, to achieve full coverage of all these areas and to optimise the interconnection and interworking of the solutions, OpenRiskNet adopted the concept of case studies. Since these are described in *Deliverable D1.3*, we concentrate here on the links between the tasks of WP4 and the use cases (*see Deliverable D2.2*), probing small parts of the infrastructure, and case studies (*see Deliverable D1.3 and priority list therein*), demonstrating the features and functionality of the e-infrastructure (data, tools and services) using real-world application.

Use Cases

A detailed description of the use cases (UC) can be found on the OpenRiskNet project's website [7] and Deliverable report D2.2. Services provided are given per UC:

UC1 Merge existing data by a common structure identifier [8]

The purpose of the UC is to extract data from multiple sources and generate a merged and harmonised dataset that can be used for analysis or predictions. In doing so it is typically necessary to handle differences in data formats, terminology, units of measure and other factors as well as to identify and handle duplicate data points.

In this UC services from **Task 4.1** and **Task 4.2** are employed. Currently, information from ChEMBL can be retrieved which include IDs, structures and assay information. This UC can be extended with alternative assay providers such as PubChem. Furthermore, toxicity information can be added from services such as ToxNet. The output of this UC could be an input to UC2.

UC2 Building a (predictive) model [9]

In this UC existing data for a particular outcome (e.g. activity or toxicity of chemicals) might be used to generate a predictive model for that outcome. A variety of machine learning algorithms can be used (for instance, QSAR techniques [10]) to generate such a predictive model given suitable input data that covers the outcome being investigated. This input data could typically be generated by UC1.

The output would normally be a prediction of the outcome, and some assessment of the certainty of that prediction. For instance, given input data that defines hepatic toxicity of a range of chemical structures the output could be a predictive model that can be used to predict hepatotoxicity of new chemicals that are not part of the input data.

In this UC services from **Task 4.2, Task 4.3, Task 4.5** and **Task 4.6** are employed. A range of approaches (e.g. machine learning algorithms and their parameterisation) might typically be assessed to identify ones that perform well.

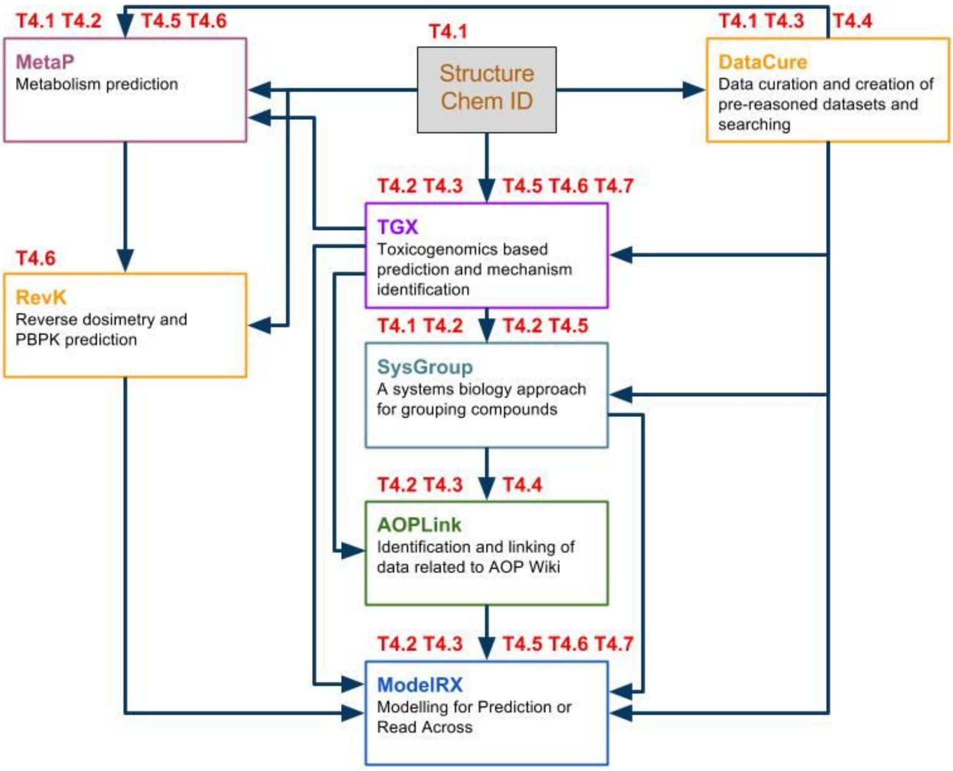
UC3 Search and Retrieve Assay Data based on Ontological Terms [11]

In this UC services from **Task 4.1, 4.2, 4.3,** and **4.4** are employed. It describes the process where a user wants to look up omics data for a particular bioassay. The user supplies a particular assay type, fetches related ontological terms from an ontology service, and retrieves all available assay data associated with assay types and terms. The UC is aimed at nanomaterials, but can be extended to other chemical substances, including compounds.

An initial implementation does not directly have to access omics databases (**Task 4.1**) but could take advantage of indexes, such as “bundle 8” in the eNanoMapper database [12]. The initial step is to find the appropriate ontology term for the biological assay, using BioPortal or OLS (see **Table 1**). Using this ontology identifier (IRI or OBO identifiers), it will query one or more databases, like eNanoMapper. For each query, it will return and summarise the lists of available datasets. Since returned data includes for which chemical compounds or substances (drug or nanomaterial) the measurement was performed, categorisation according to this information is possible.

Case Studies

The case studies (CS) are described in detail in the *Deliverable Report D1.3*. Since, work on these CS has just recently started, services have not yet been implemented. However, the services within the different WP4 tasks have already been appointed to the different CS as depicted in **Figure 2**. This overview will be updated continuously on the project’s website together with the description of individual CS including the status of all integrated services and results obtained by these.


T4.1 T4.2
T4.5 T4.6
T4.1
T4.1 T4.3
T4.4
MetaP
Metabolism prediction
Structure Chem ID
DataCure
Data curation and creation of pre-reasoned datasets and searching
T4.2 T4.3
T4.5 T4.6 T4.7
T4.6
TGX
Toxicogenomics based prediction and mechanism identification
RevK
Reverse dosimetry and PBPK prediction
T4.1 T4.2
T4.2 T4.5
SysGroup
A systems biology approach for grouping compounds
T4.2 T4.3
T4.4
AOPLink
Identification and linking of data related to AOP Wiki
T4.2 T4.3
T4.5 T4.6 T4.7
ModelRX
Modelling for Prediction or Read Across

T4.1 → Task 4.1 Toxicology, Chemical Properties and Bioassay Databases
T4.2 → Task 4.2 Omics Databases
T4.3 → Task 4.3 Knowledge Bases and Data Mining
T4.4 → Task 4.4 Ontology Services
T4.5 → Task 4.5 Processing and Analysis
T4.6 → Task 4.6 Predictive Toxicology
T4.7 → Task 4.7 Workflows, Visualisation and Reporting

Figure 2. Case studies and Services interconnection

CONCLUSION

The integration of services has started by the specification of the semantic annotation of data and modelling services pursuing a new concept based on the JSON-LD syntax and was tested on the Squonk and Jaqpot applications and different data sources. This work will now be continued by integrating services (data sources and software tools) relevant to the risk assessment process as listed in **Table 1**. Prioritization of specific tools will be done according to the requirements of the CS and UC defined in *Deliverable 1.3* as well as additional requirements coming from the associated partners.

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. SNIC Science Cloud. In: SNIC Science Cloud [Internet]. [cited 4 Dec 2017]. Available: https://cloud.snic.se/

2. Squonk Computational Notebook [Internet]. [cited 4 Dec 2017]. Available: https://portal-squonk.dev.openrisknet.org/

3. lazar Toxicity Predictions [Internet]. [cited 4 Dec 2017]. Available: https://lazar.in-silico.ch/predict

4. JGU WEKA REST services Swagger UI [Internet]. [cited 13 Dec 2017]. Available: https://cuttlefish.informatik.uni-mainz.de/

5. Jaqpot [Internet]. [cited 4 Dec 2017]. Available: http://jaqpot.org/

6. Swagger UI [Internet]. [cited 19 Dec 2017]. Available: http://test.jaqpot.org:8080/jaqpot/swagger/

7. Case Studies and Use Cases :: OpenRiskNet [Internet]. [cited 19 Dec 2017]. Available: https://openrisknet.org/development/case-studies/

8. Use Case 1 :: OpenRiskNet [Internet]. [cited 4 Dec 2017]. Available: https://openrisknet.org/development/case-studies/use-case-1/

9. Use Case 2 :: OpenRiskNet [Internet]. [cited 4 Dec 2017]. Available: https://openrisknet.org/development/case-studies/use-case-2/

10. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57: 4977–5010.

11. Use Case 3 :: OpenRiskNet [Internet]. [cited 4 Dec 2017]. Available: https://openrisknet.org/development/case-studies/use-case-3/

12. Data.eNanoMapper.net [Internet]. [cited 19 Dec 2017]. Available: https://data.enanomapper.net/ui/assessment?bundle_uri=https%3A%2F%2Fdata.enanomapper.net%2Fbundle%2F8