UNIVERSITY OF PLOVDIV "PAISII HILENDARSKI"

FACULTY OF CHEMISTRY

DEPARTMENT OF ANALYTICAL CHEMISTRY AND COMPUTER CHEMISTRY

GERGANA ILIEVA TANCHEVA

APPLICATION OF CHEMOINFORMATICS METHODS TO MULTICOMPONENT SUBSTANCES AND NANOMATERIALS

ABSTRACT

OF A DISSERTATION

FOR THE AWARD OF EDUCATIONAL AND SCIENTIFIC DEGREE "PH.D."

Field of higher education: 4. Natural sciences, mathematics and informatics
Professional direction 4.2. Chemical Sciences
Doctoral program Theoretical Chemistry

Research supervisor: Assoc. Prof. Nikolay Kochev, Ph.D.

Plovdiv
2024

The dissertation was discussed and approved for defense at a meeting of the Departmental Council of the Department of Analytical Chemistry and Computational Chemistry, Faculty of Chemistry, at Paisii Hilendarski University of Plovdiv, held on October 14, 2024.

The dissertation consists of 196 pages and includes 65 figures, 6 tables, and 4 appendices, organized into 7 chapters and are cited 246 references.

The defense materials are available for review in the "Development of Academic Staff and Doctoral Studies" department at Paisii Hilendarski University of Plovdiv, the National Center for Information and Documentation at the Ministry of Education, Youth and Science, and the Central Library of Paisii Hilendarski University.

Scientific jury:

Prof. Ivanka Milosheva Tsakovska, Ph.D. - Bulgarian Academy of Sciences; Institute of Biophysics and Biomedical engineering, *Field of higher education: 4. Natural sciences, mathematics and informatics; Professional direction: 4.3 Biological sciences (Pharmacology)*

Prof. Irini Atanas Doichinova-Tsekova, D. Sci. - MU - Sofia, *Higher Education Department: 7. Health and Sports; Professional direction: 7.3. Pharmacy; Science major: Theoretical Chemistry*

Prof. Ivan Petkov Bangov, D. Sci. - retired, *Higher Education Department: 4. Natural Sciences, Mathematics and Informatics; Professional direction: 4.2 Chemical sciences (Theoretical Chemistry)*

Prof. Veselin Petrov Baev, Ph.D. - PU "Paisiy Hilendarski", *Higher Education Department: 4. Natural Sciences, Mathematics and Informatics; Professional direction: 4.3 Biological sciences (Molecular Biology)*

Prof. Vasil Borisov Delchev, D. Sci. - PU "Paisiy Hilendarski", *Higher Education Department: 4 Natural sciences, mathematics and informatics; Professional direction: 4.2 Chemical sciences (Theoretical Chemistry)*

I. Introduction

Chemoinformatics emerged as a response to the need to apply computer methods and information technologies in the overall cycle for processing the generated experimental or simulated chemical data and to combine it with other information resources for the purpose of subsequent transformation of the data into useful information for modeling and research activities, followed by the third stage - information formalization in the form of knowledge, with the ultimate goal of solving practical research problems and innovations. The classic three areas of application (problems) of chemoinformatics are well known: information and computer support in the discovery of chemical compounds with target properties, discovery of methods for their synthesis and structure elucidaton of unknown compounds[1]. Chemoinformatics has developed as an interdisciplinary science that covers design, creation, organization, management, search, analysis, dissemination, visualization and use of chemical information. The methods of chemoinformatics are related to presentation and storage of chemical objects in chemical databases, data processing, calculation of descriptors and modeling of physicochemical properties and biological activity. The basis of all described computer methods is the representation (formalization) of chemical compounds by means of a given data model.

The application of chemoinformatics methods to classical chemical entities – molecules (or chemical compounds) has been well researched for more than 40 years and has shown its usefulness over time. With regard to the exponentially developing production of multicomponent substances and nanomaterials, the direct application of classical chemoinformatics methods is questionable. This challenge motivates the purpose of the dissertation, namely: investigate the possibilities of applying chemoinformatics methods for processing and storing information about multicomponent substances, nanomaterials and advanced materials and discovering perspectives for effective information processing through semantic FAIR data model and its application in scientific experiments and modeling in nanomaterials and substances.

Conclusions from the literature review

In the last decade, production of nanomaterials and chemical substances has grown at an exponential rate. The concept Safe and Sustainable by Design (SSbD) is committed to the design of functional and safe chemical substances and nanomaterials at the early stages of their technological development and is promoted through regulatory initiatives and numerous scientific projects. The European Commission[2] encourages industry, academia and research centers to ensure that the methods, models and data, produced in line with the implementation of the European framework, comply with the guiding principles of Findability, Accessibility, Interoperability and Reusability (FAIR). In addition, the increase of high-quality FAIR data is also encouraged. The development of new risk assessment methods, models and tools is also encouraged.

It is clear from the literature review that the starting point of all chemoinformatics methods is the representation of the chemical compound in a machine-readable format. There is a well-defined model for representing molecules using three components: chemical structure, properties, and descriptors. This model has shown its usefulness in academia, but for the purposes of industry and regulatory agencies, it is not sufficient because, in reality, we do not deal with pure structures, but with multicomponent substances. Although the different regulatory agencies do not currently have a consensus on the definition of a chemical substance and on the definition of a nanomaterial, their proposals are united around a common understanding that a chemical substance is not made up of a single component. In conclusion, a new paradigm is needed for the presentation of chemical substances towards an adequate processing of data from nanomaterials, micro and nanoplastics and advanced materials.

The direct application of classical methods of chemoinformatics for chemical substances is a challenge due to the need of new approaches for adequate substance representation in accordance with the FAIR principles. It was these challenges and recommendations described by the European Commission that motivated the main objective of the current dissertation work.

II. Thesis main objective and work tasks

2.1 Main objective:

Researching the possibilities of applying chemoinformatics methods for processing and storing information about multicomponent substances, nanomaterials and advanced materials and discovering perspectives for effective information processing through a semantic FAIR data model and its application in scientific experiments and modeling of properties of nanomaterials and chemical substances.

2.2 Work tasks:

1. Study of existing software systems, published algorithms and technologies for chemical object representation and chemical information processing.
2. Exploring the FAIR principles and the possibilities of their application on data for multicomponent substances and nanomaterials.
3. Exploring existing semantic models for multicomponent substance representation and professional scientific data serialization formats.
4. Selection of a semantic model for presentation (formalization) of information (data and meta data) for chemical substances and nanomaterials.
5. Exploring existing ontologies for chemical substances and nanomaterials.
6. Adaptation and application of processing and modeling algorithms in chemoinformatics for chemical substances and nanomaterials.
7. Creation of algorithms for FAIRification of non-FAIR experimental data.
8. Application of FAIRification algorithms to nanomaterial and chemical substance data from previous and current European scientific projects.

9. Creation of algorithms and electronic notebooks for serialization of information from the underlying semantic data model.
10. Creation of algorithms and electronic notebooks for evaluating the completeness, quality and degree of FAIRification of the chemical objects represented by the selected semantic model.
11. Testing the capabilities of the selected semantic model for application on high-throughput screening (HTS) data from biological experiments.
12. Creation of improved and automated methodologies for processing HTS data and development of a methodology to integrate the FAIR principles for HTS data.
13. Prototyping of unique identifiers for multicomponent substances and nanomaterials.
14. Testing software analytics platforms for efficiency and speed for data preprocessing and modeling.
15. Testing the software modules, the created strategies and logic in terms of efficiency, speed and correctness of the generated chemical information and the created models.
16. Comparing our developed approaches and algorithms with other software systems.

III. Research

3. FAIR principles in data management.

Efficient management and aggregation of experimental data from different sources is one of the main goals of automated data processing. A basic requirement for achieving this goal is to combine the data from the original experiments tests of the chemical substances with rich metadata. One of the pillars of the scientific method is the possibility of independent confirmation and repetition of the obtained results, and this is where metadata is needed in order to describe all possible aspects of the experiments conducted.

In 2016, four basic principles of scientific data management, shown in Figure 1, were published, according to which data should be: Findable , Accessible, Interoperable and Reusable[3] (FAIR). These principles guide researchers and institutions that generate data, how to maximize the benefits of their data. Adherence to the FAIR principles is also important in the context of the applied algorithms, tools and workflows used to generate the data itself. The GO-FAIR initiative[4] and the European Commission[2] in implementation of the concept Safe and Sustainable by Design promote the application of FAIR principles.

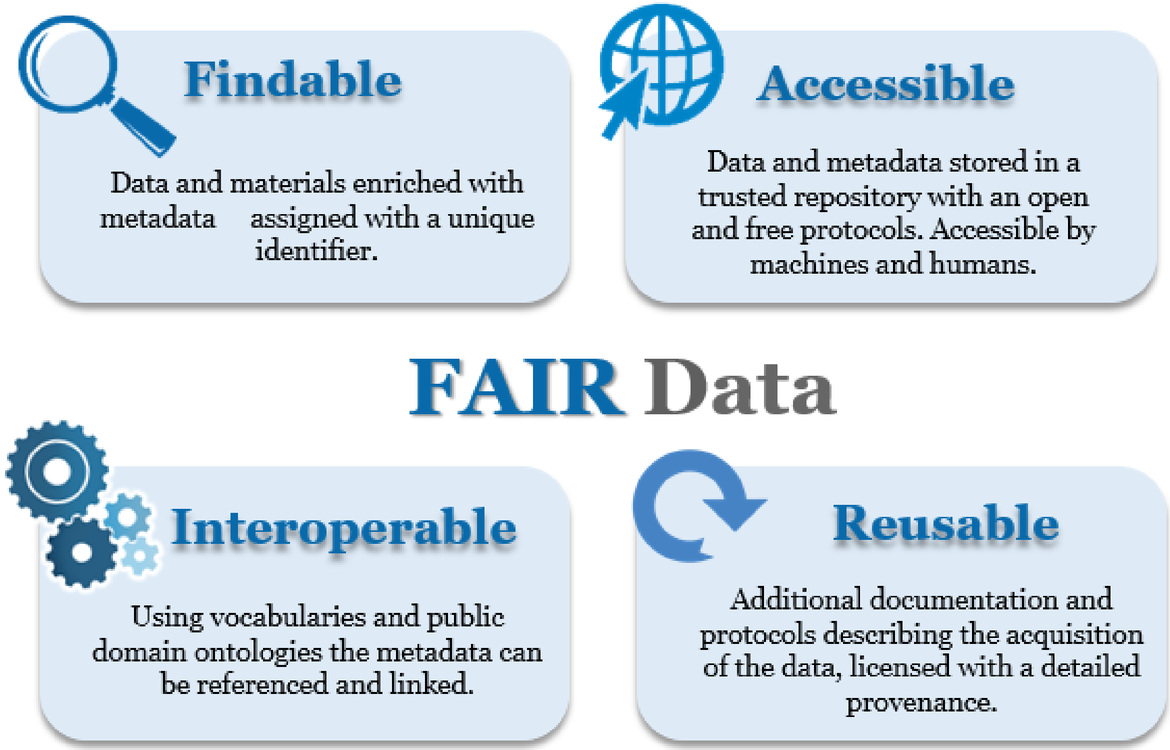

Findable
Data and materials enriched with metadata assigned with a unique identifier.
Accessible
Data and metadata stored in a trusted repository with an open and free protocols. Accessible by machines and humans.
FAIR Data
Interoperable
Using vocabularies and public domain ontologies the metadata can be referenced and linked.
Reusable
Additional documentation and protocols describing the acquisition of the data, licensed with a detailed provenance.

Figure 1: *FAIR principles: Findable, Accessible, Interoperable and Reusable*

The first step in reusing data is to make them findable. Machine-readable data and metadata are essential for the automatic discovery of datasets and online services. For each of the principles to be realized, certain data requirements must be met.

The FAIR principles allow experimental data to be used beyond their origin in order to solve scientific problems, fill missing data, reuse the data in applications, do modeling and provide tools for other needs of science, industry and regulators. The principles emphasize machinability (i.e., the ability of computing systems to find, access, interact, and reuse data with no or minimal human intervention) as humans increasingly rely on computational support for data processing as a result of the increase in volume, complexity and speed of data generation.

GO-FAIR recommends a seven-step workflow for transforming non-FAIR data into FAIR one shown in Figure 2.

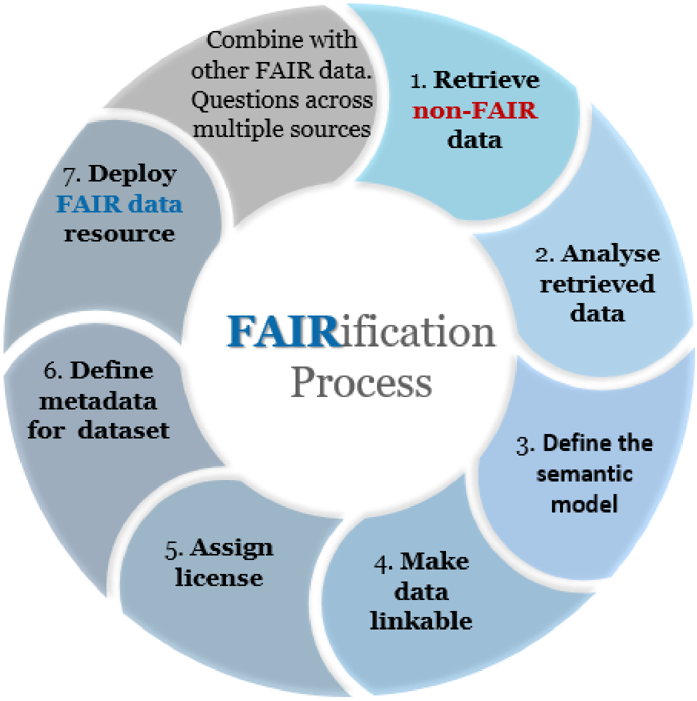

FAIRification Process
1. Retrieve non-FAIR data
2. Analyse retrieved data
3. Define the semantic model
4. Make data linkable
5. Assign license
6. Define metadata for dataset
7. Deploy FAIR data resource
Combine with other FAIR data. Questions across multiple sources

Figure 2*: FAIRification process: a workflow of seven steps for transforming a non-FAIR data into a FAIR data resource*

A semantic data model includes objects representing aspects of reality and their relationships. Researchers routinely describe experimental objects in the scientific literature as materials, methods and results. The computer representation of an experimental system (i.e. the materials and methods) requires the definition of data elements, their relationships, and the constraints between them. In computer science, this is known as **a data model** and serves as a blueprint for data storage, access and manipulation. A data model represents the principle (and abstract) logic behind the objects being described and is something different from the format used to store the data, because the same data model can be serialized in different formats.

FAIRification workflow efforts to clarify the data model for chemical substances are also efforts to implement the FAIR principles. Other steps of primary importance for data FAIRification are the inclusion of rich metadata (step 6), ontology annotations, and linking data with globally unique identifiers (step 4).

4. A data model for the representation and processing of chemical substances and nanomaterials

In the present dissertation, different approaches for describing information about multicomponent substances and nanomaterials were studied. As a result, three main metadata layers for FAIR description of multicomponent substances are recognized and are shown in Figure 3 as: (i) substance identification, (ii) description of composition and (iii) measurement records with rich metadata.

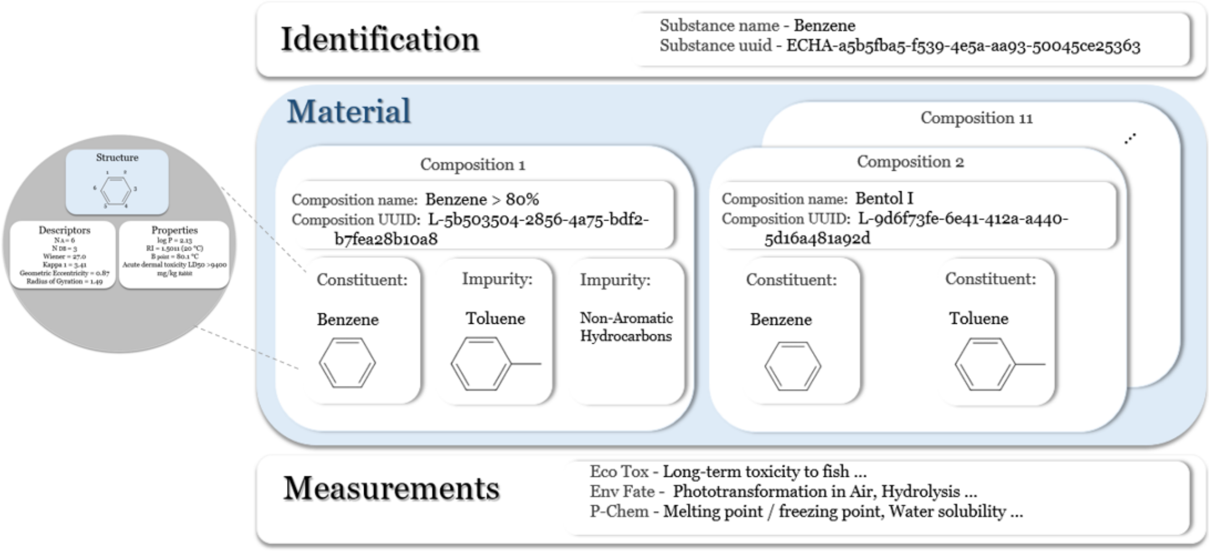

Identification
Substance name - Benzene
Substance uuid - ECHA-a5b5fba5-f539-4e5a-aa93-50045ce25363
Material
Composition 11
Composition 1
Composition name: Benzene > 80%
Composition UUID: L-5b503504-2856-4a75-bdf2-b7fea28b10a8
Constituent:
Benzene
Impurity:
Toluene
Impurity:
Non-Aromatic Hydrocarbons
Composition 2
Composition name: Bentol I
Composition UUID: L-9d6f73fe-6e41-412a-a440-5d16a481a92d
Constituent:
Benzene
Constituent:
Toluene
Structure
1
2
3
4
5
6
Descriptors
N A = 6
N DB = 3
Wiener = 27.0
Kappa 1 = 3.41
Geometric Eccentricity = 0.87
Radius of Gyration = 1.49
Properties
log P = 2.13
RI = 1.5011 (20 °C)
B point = 80.1 °C
Acute dermal toxicity LD50 >9400 mg/kg rabbit
Measurements
Eco Tox - Long-term toxicity to fish ...
Env Fate - Phototransformation in Air, Hydrolysis ...
P-Chem - Melting point / freezing point, Water solubility ...

Figure 3: *Substance of benzene with several different compositions and information grouped in three layers: identification, material, measurements (example is taken from the public records of the ECHA's dossiers and is also accessible via Ambit-LRI database web interface).*

The specific example in Figure 3 demonstrates that the chemical substance benzene has 11 registered compositions, two of which are: benzene with a purity >80% and bentol, where the chemical compound benzene is the main component of the substance. It must emphasized again that in the terminology of ECHA, as well as in the present dissertation, a significant difference is made between the concepts **of chemical compound** and **chemical substance.**

Ambit/eNanoMapper data model (Figure 4) is a conceptual representation of the chemical substances and can be implemented with different technologies, ensuring interoperability and data linkage. It contains various data components or objects performing specific roles to represent elements of information about chemical substances and measurements. Objects can have different implementations at different stages of the data processing workflow: JSON, RDF and HDF5 formats, Java and Python classes or SQL tables.

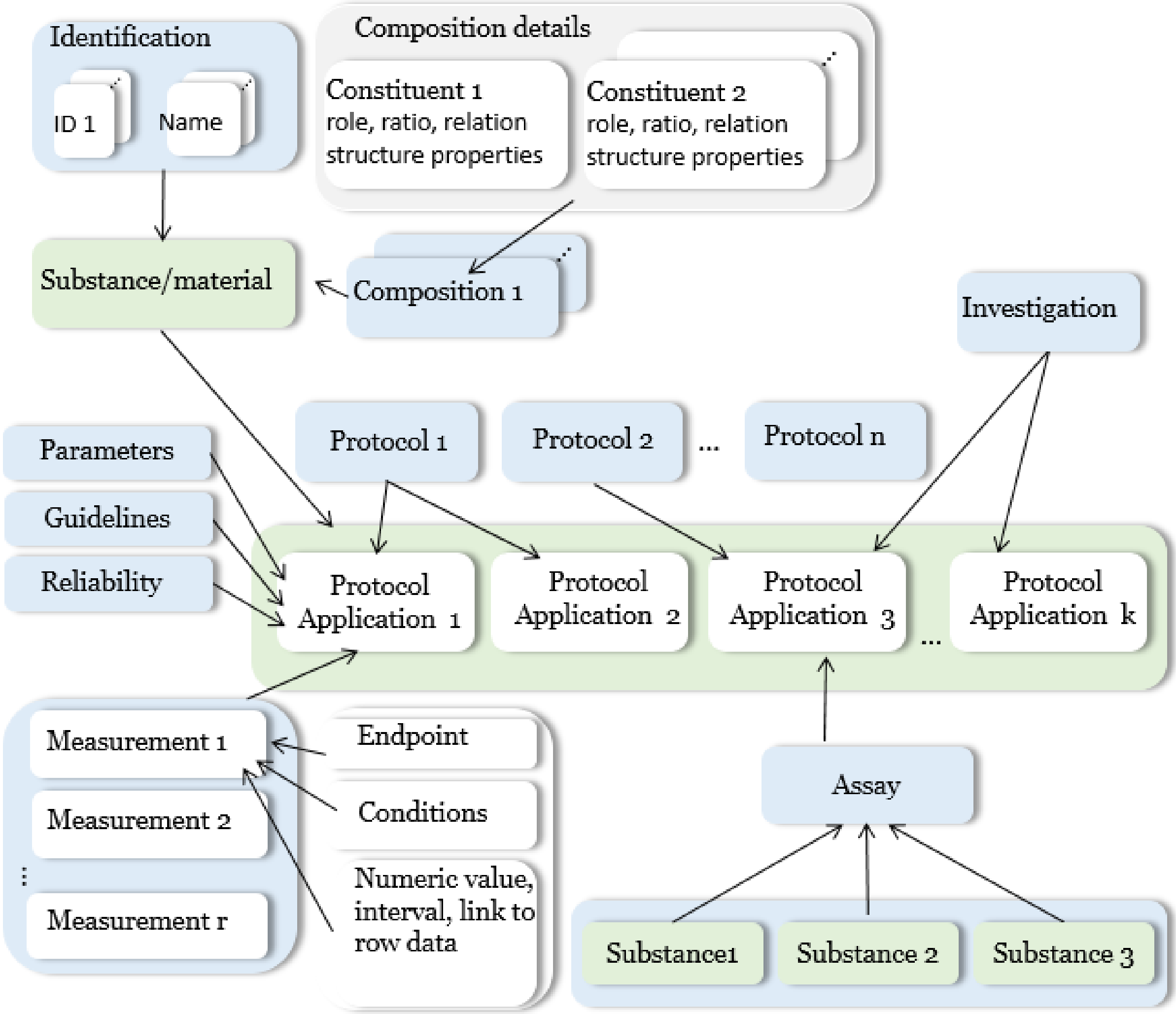


Figure 4: *Schema of the Ambit/eNanoMapper data model*

In the data model, substances are characterized by component concertation ratios in the composition and are identified by names and identifiers. The data model supports multiple compositions with one or more components, each one with a specific role such as main ingredient, additive, impurity, core, nanomaterial coating, etc. Each component is represented by the classical model for a chemical compound. The results from physicochemical and biological measurements are treated as properties of the entire chemical substance and are informationally managed through objects called "protocol applications". Effective description of the experiments conducted in a protocol is critical to the proper scientific results communication and to the creation of FAIR data resources. The latter is implemented using a rich set of parameters (metadata) with a flexible logical organization. Each application of a protocol consists of a set of measurements for endpoints under given experimental conditions. The measurement result can be represented as a numeric value, an interval, text or a link to a data file (e.g. infrared spectrum, microscopic image, HTS data, etc.). There is flexibility in storing parameters from the metadata. Each measurement is associated with a dynamic list of experimental factors (such as concentration, time, replicate number, etc.) considered as "lower" level parameters. "High" level metadata includes parameters, describing measurement equipment, sample preparation, guides, links to standard procedures, publications, and more.

5. Platform for FAIRification of chemical object data. JSON configurations, harmonized templates and ontologies

The semantic data model enables the integration of data from various sources, such as OECD harmonized templates, custom spreadsheet templates, SQL output, and more. Once the data has been imported into the eNanoMapper database, various options are available for data extraction, analysis and conversion to other formats (Figure 5). The data model is flexible and allows application of various customized methods for accessing the data through the REST API interface and the pynanomapper python library, using external tools, the machine learning and data analysis platforms, workflows in the Ploomber platform, etc.

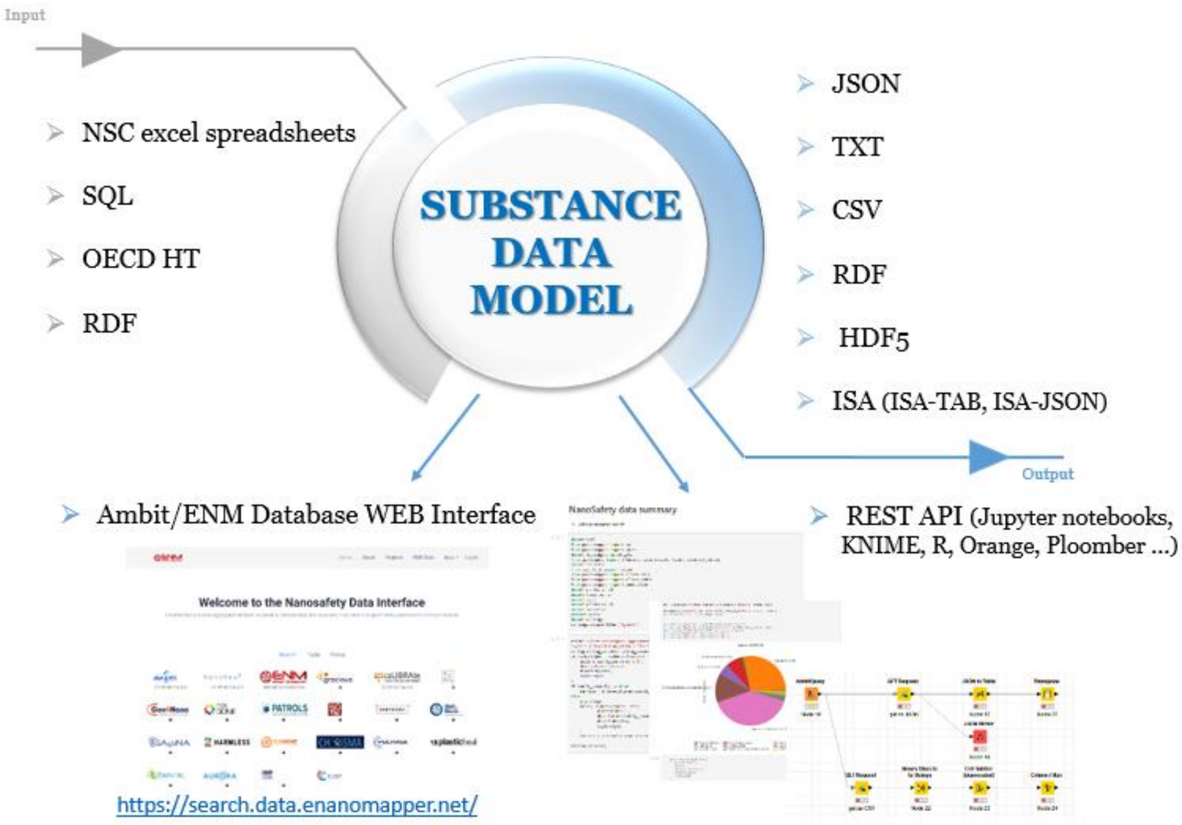

Input
NSC excel spreadsheets
SQL
OECD HT
RDF
SUBSTANCE DATA MODEL
JSON
TXT
CSV
RDF
HDF5
ISA (ISA-TAB, ISA-JSON)
Output
Ambit/ENM Database WEB Interface
Welcome to the Nanosafety Data Interface
https://search.data.enanomapper.net/
NanoSafety data summary
REST API (Jupyter notebooks, KNIME, R, Orange, Ploomber ...)

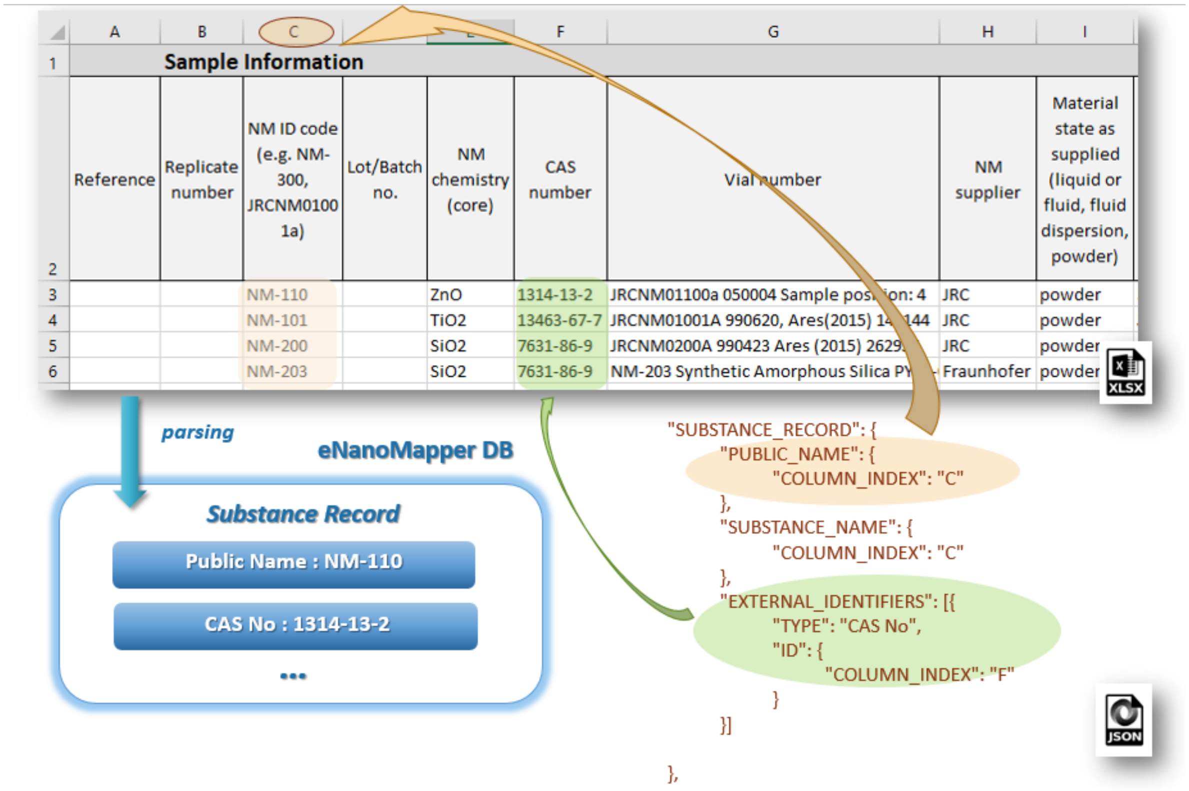
***Figure 5**: Data input and output to Ambit/eNanoMapper database*

The most preferred way by researchers to enter data is through spreadsheet templates, which are generally non-FAIR data resources and hence the need arises for an EXCEL files FAIRfication. For this purpose, a configurable tool NMDataParser[5] was developed, which facilitates data preparation and uploading to the eNanoMapper database i.e. data is converted to the Ambit/eNanoMapper data model . Different forms of EXCEL sheet data organization are supported – by rows, by columns or by blocks. Due to the wide variety of EXCEL spreadsheet files, it is necessary to configure the conversion via a separate JSON file. The NMDataParser tool works with two input files: a spreadsheet (*.xlsx file) and a JSON configuration file, and returns an iterator to a list of chemical substance records.

5.1 JSON configuration files for importing data

The JSON configuration file associates the individual elements of the spreadsheet with the components of the semantic data model and annotates the data with a domain-specific ontology. The JSON configuration file syntax for NMDataParser includes a set of keywords for reading data from spreadsheets and converting it to the Ambit/eNanoMapper data model. Keywords define different strategies for reading data, from one or multiple Excel sheets, and allow different combinations of data from different sheets, rows, columns, blocks of columns, and rows. A JSON configuration file consists of several main sections that represent first-level objects in the JSON schema.

Figure 6 illustrates configuring a JSON file to read from EXCEL the nanomaterial name and CAS number as an external identifier, and to convert the data into the Ambit/eNanoMapper model.



	A	B	C	E	F	G	H	I
1		Sample Information						
2	Reference	Replicate number	NM ID code (e.g. NM-300, JRCNM01001a)	Lot/Batch no.	NM chemistry (core)	CAS number	Vial number	NM supplier	Material state as supplied (liquid or fluid, fluid dispersion, powder)
3			NM-110		ZnO	1314-13-2	JRCNM01100a 050004 Sample pos ion: 4	JRC	powder
4			NM-101		TiO2	13463-67-7	JRCNM01001A 990620, Ares(2015) 14 144	JRC	powder
5			NM-200		SiO2	7631-86-9	JRCNM0200A 990423 Ares (2015) 2629	JRC	powder
6			NM-203		SiO2	7631-86-9	NM-203 Synthetic Amorphous Silica PY	Fraunhofer	powder

```
"SUBSTANCE_RECORD": {
     "PUBLIC_NAME": {
          "COLUMN_INDEX": "C"
     },
     "SUBSTANCE_NAME": {
          "COLUMN_INDEX": "C"
     },
     "EXTERNAL_IDENTIFIERS": [{
          "TYPE": "CAS No",
          "ID": {
               "COLUMN_INDEX": "F"
          }
     }]

},
```

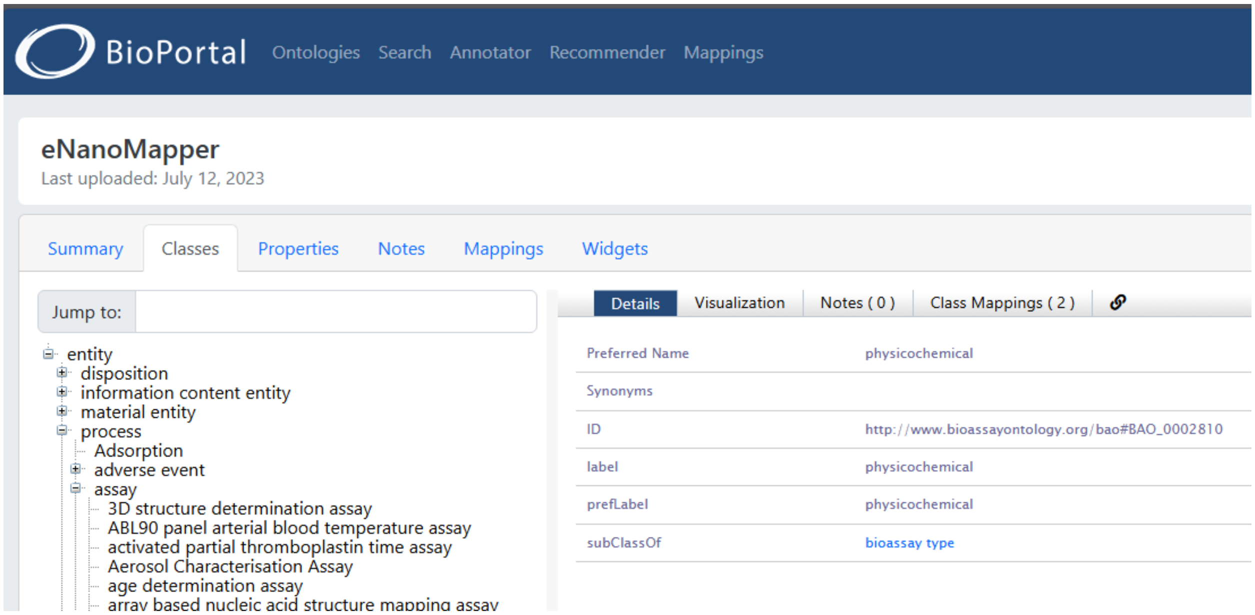
Figure 6: *Reference data from EXCEL in "ROW_SINGLE" iteration mode and SUBSTANCE_RECORD iteration.*

The described FAIRification process was applied to 1400 EXCEL files with experimental data. The converted data contained information from several European projects: NanoTest, NanoReg, NanoReg2, MARINA, ENPRA. The data imported in the eNanoMapper database includes a wide variety of physicochemical and biological assays such as: cell viability, oxidative stress, immunotoxicity , in-vivo/in-vitro toxicity, ecotoxicity, physicochemical characterization and others, for many nanomaterials such as: carbon nanotubes, silver nanoparticles, zinc and titanium dioxides, iron and cerium oxides, materials with different coatings and without coating.

5.3 eNanoMapper Ontology

The eNanoMapper[6] data management is based on semantic web standards and ontologies. The eNanoMapper project started the development of a comprehensive ontology[7] in order to annotate nanosafety database , which is an important step to address the challenge of unified annotation of nanomaterials and their relevant biological properties, experimental model systems, conditions, protocols and environmental impact data[8]. The description of information concerning the safety of nanomaterials includes diverse subfields such as biological analyses, determination of physicochemical and ecological characteristics, and in this regard, developing an ontology "from scratch" would be quite laborious and time-consuming[9].

Initially, we reviewed the terms from the MESOCOSM ontology and compared them with existing ontologies, with the main goal to select those with the closest definition to the target domain of the MESOCOSM ontology. Through the BioPortal website we searched the largest repository of biomedical and natural mathematical ontologies. The second step towards the integration of the MESOCOSM ontology was to add new terms to the eNanoMapper ontology. Adding terms to the eNanoMapper ontology was done using "slimmer" library and specialized configuration files (.props and .iris files). Figure 7 demonstrates the addion of the term "physicochemical" to the eNanoMapper ontology, accessible through the BioPortal platform at the link: https://bioportal.bioontology.org/ontologies/ENM?p=classes&conceptid=http%3A%2F%2Fwww.bioassayontology.org% 2Fbao%23BAO_0002810



eNanoMapper
Last uploaded: July 12, 2023

- entity
  - disposition
  - information content entity
  - material entity
  - process
    - Adsorption
    - adverse event
    - assay
      - 3D structure determination assay
      - ABL90 panel arterial blood temperature assay
      - activated partial thromboplastin time assay
      - Aerosol Characterisation Assay
      - age determination assay
      - array based nucleic acid structure mapping assay

Preferred Name	physicochemical
Synonyms	
ID	http://www.bioassayontology.org/bao#BAO_0002810
label	physicochemical
prefLabel	physicochemical
subClassOf	bioassay type

Figure 7: *Screen from eNanoMapper ontology from BioPortal with the newly added term "physicochemical".*

Uploading configuration files to the ontology Git repository does not mean that the term is automatically added to the eNanoMapper ontology. Proposed terms are subject to review, discussion and consensus by a team of experts working to improve the ontology, and for each new term this may take a certain amount of time, even considerable time, if there is no consensus on the definition of the term.

6 Representation (serialization) for multicomponent substances using linear notations

The development of identifiers for multicomponent substances is a major challenge, as the approaches used for traditional identifiers of chemical compounds, as well as the most popular linear notations SMILES and InChI, are not directly applicable to the description of multicomponent substances. One of the key FAIR principles is the use of globally unique and persistent identifiers.

SYBYL Line Notation[10, 11] (SLN) is unambiguous non-unique linear notation and supports syntax for specification on molecules, requests for substructure search and reactions that cover the possibilities on the standard notations SMILES, SMARTS and SMIRKS. In addition, the SLN syntax includes others effective means for specification user defined attributes of atoms, bonds, structures and reactions, as well as macro and Markush atoms for flexible definition on molecular fragments, representation of structural libraries and description of 2D and 3D atomic coordinates .

An open source software library, Ambit - SLN, has been developed for processing chemical substance information using the SLN linear notation . Ambit-SLN is a software module, part of the AMBIT chemoinformaics platform. The Ambit-SLN library includes several core functionalities such as an internal representation of SLN information, a parser for the full SLN substructure search query syntax with support for macro and Markush atoms, global and local dictionaries, and user-defined properties. The latter could be used to represent the principle elements of the Ambit/eNanoMapper data model. Figure 8 illustrates the basic workflow of the Ambit-SLN library and the two main use cases of SLN notation application.

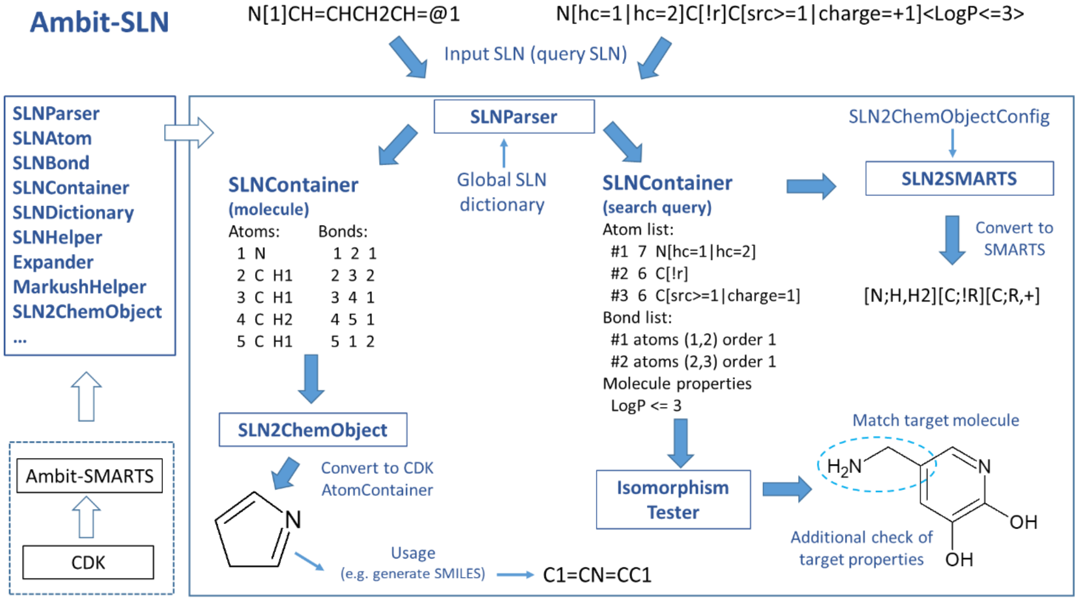


Figure 8*Ambit-SLN workflow applied to a standard connectivity table and substructure query.*

Ambit-SLN covers most of the SLN syntax. We tested the capabilities of the Ambit-SLN library for describing substances and nanomaterials. The prototype SLN identifier for Fe_3O_4 material with d=38nm and a glycine coating of 2nm is shown below and demonstrates the concept of Nano-SLN:

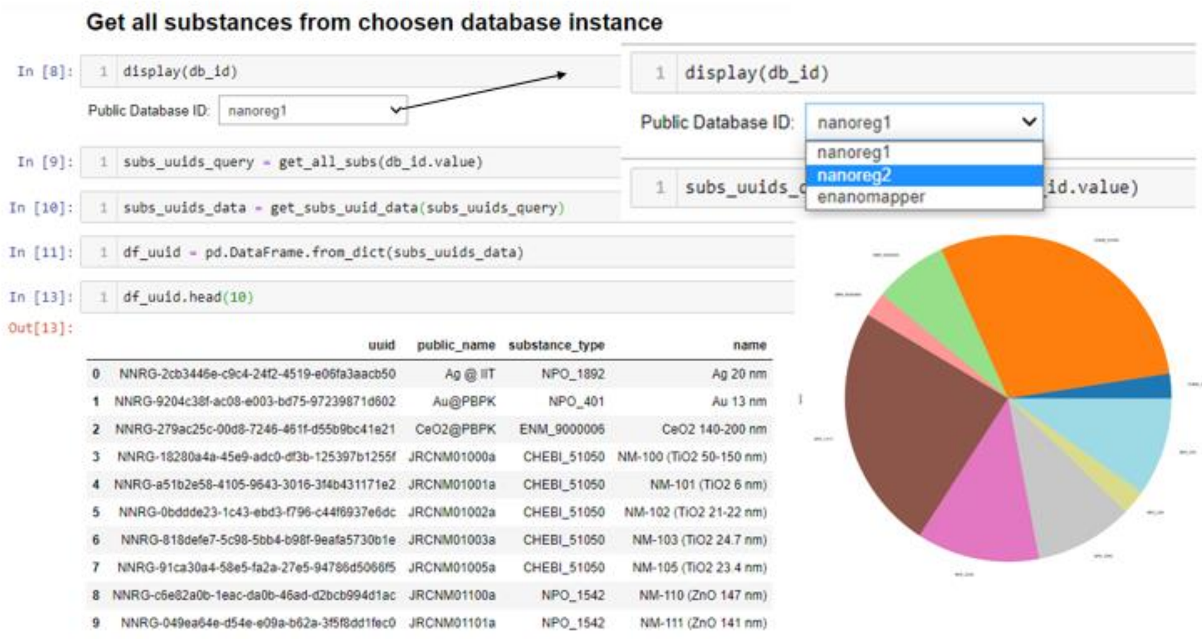
O[1]Fe[2]OFeOFe@1O@2 <role= core ;size =38nm>
CH2(C(=O)OH)NH2 <role= coating; size =2nm>

7 Data Processing

Experimental data is the basis of information processing workflows in the activities for risk assessment of used nanomaterials. A big part of the time and effort in developing models for physicochemical properties and biological activity is spent analyzing, collecting, filtering, and preparing an appropriate samples of data. Effective data mining allows scientists to formulate correct hypotheses and create better models.

7.1 Electronic notebook (Jupyter notebook) for retrieving, processing and visualizing nanosafety data from the eNanoMapper database via REST API requests.

The electronic notebooks enable dynamic selection of instances from the eNanoMapper database (instance is a distinct subset of chemical substances generated within a given scientific project) and tabular visualization of metadata for all chemical substances with their unique identifiers (uuids), ontological (substance type) and public names. Figure 9 shows a selection of a particular instance, Nanoreg2, from the database, along with a table of all available substances and a pie chart with the data availability aggregated by substance type. Substances are filtered according to "substance type" CHEBI 51050 [12] – titanium dioxide. Figure 10 shows subsequent filtration for a selected measurement value "cell viability" with available experimental data for 6 nanoforms of the filtered substance - titanium dioxide.



Get all substances from choosen database instance

```
In [8]:  display(db_id)
```

Public Database ID: nanoreg1

```
In [9]:  subs_uuids_query = get_all_subs(db_id.value)

In [10]: subs_uuids_data = get_subs_uuid_data(subs_uuids_query)

In [11]: df_uuid = pd.DataFrame.from_dict(subs_uuids_data)

In [13]: df_uuid.head(10)
Out[13]:
```

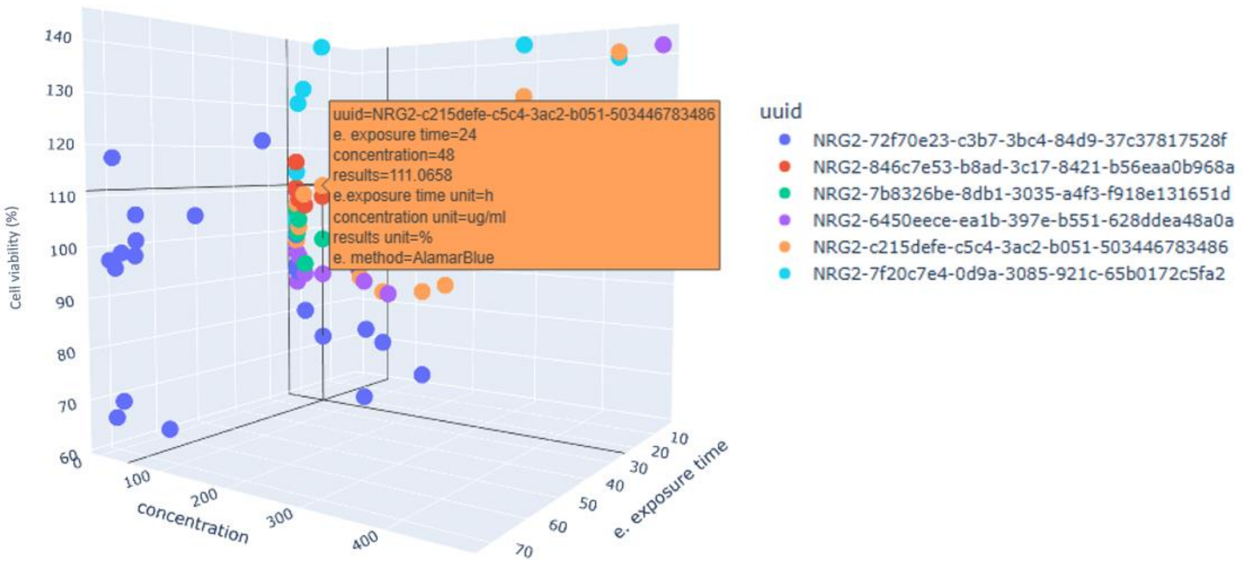
	uuid	public_name	substance_type	name
0	NNRG-2cb3446e-c9c4-24f2-4519-e06fa3aacb50	Ag @ IIT	NPO_1892	Ag 20 nm
1	NNRG-9204c38f-ac08-e003-bd75-97239871d602	Au@PBPK	NPO_401	Au 13 nm
2	NNRG-279ac25c-00d8-7246-461f-d55b9bc41e21	CeO2@PBPK	ENM_9000006	CeO2 140-200 nm
3	NNRG-18280a4a-45e9-adc0-df3b-125397b1255f	JRCNM01000a	CHEBI_51050	NM-100 (TiO2 50-150 nm)
4	NNRG-a51b2e58-4105-9643-3016-3f4b431171e2	JRCNM01001a	CHEBI_51050	NM-101 (TiO2 6 nm)
5	NNRG-0bddde23-1c43-ebd3-f796-c44f6937e6dc	JRCNM01002a	CHEBI_51050	NM-102 (TiO2 21-22 nm)
6	NNRG-818defe7-5c98-5bb4-b98f-9eafa5730b1e	JRCNM01003a	CHEBI_51050	NM-103 (TiO2 24.7 nm)
7	NNRG-91ca30a4-58e5-fa2a-27e5-94786d5066f5	JRCNM01005a	CHEBI_51050	NM-105 (TiO2 23.4 nm)
8	NNRG-c6e82a0b-1eac-da0b-46ad-d2bcb994d1ac	JRCNM01100a	NPO_1542	NM-110 (ZnO 147 nm)
9	NNRG-049ea64e-d54e-e09a-b62a-3f5f8dd1fec0	JRCNM01101a	NPO_1542	NM-111 (ZnO 141 nm)

```
display(db_id)
```

Public Database ID: nanoreg1

```
subs_uuids_q                    id.value)
```

Figure 9: *Customized data filtration according to the available database instances and summary visualization of the substances available in the selected instance.*



uuid=NRG2-c215defe-c5c4-3ac2-b051-503446783486
e. exposure time=24
concentration=48
results=111.0658
e.exposure time unit=h
concentration unit=ug/ml
results unit=%
e. method=AlamarBlue

Figure 10: *3D interactive graphic showing the distribution of substances in space (dose, result, treatment time) with color designation of the nanoforms and additional information for each substance.*

7.2 Electronic notebook for checking data completeness

We developed an electronic notebook to assess data completeness within the terminology used in the harmonized templates and the information available in the eNanoMapper database. The terms used to describe the data from a selected instance of the database are compared to those from the harmonized templates, assessing their similarity by means of the Levenstein distance[13], which is the minimum number of character edits (insertions, deletions or substitutions), needed to change one word to another. Table 1 demonstrates a concrete example of the work of the notebook.

Table 1: *Comparison for a set of terms used to describe data for the endpoint „specific surface area by BET".*

Terms in database	Terms in a template	Terms comparison
MATERIAL _ STATE,	MATERIAL _ TATE ,	Wrong term in template
INSTRUMENT _ TYPE, OUTGASSING _ TIME, OUTGASSING _ TEMP,	INSTRUMENT_TYPE, OUTGASSING _ TIME, OUTGASSING _ TEMP,	Completely matching terms
END_RELATIVE_PRESSURE START_RELATIVE_PRESSURE	-	Missing terms in template
-	SAMPLE _ DENSITY	Missing terms in the database

8 Data processing from high-throughput screening (HTS) of biological experiments and subsequent modeling

Applying the FAIR principles in the context of HTS data is challenging because of the need to automatically link a large volume of experimental data with descriptive metadata, harmonize the terminology used, and convert it into a machine-readable format that will allow the data to be easily findable, accessible and reusable. Traditional approaches to storing and documenting HTS results, such as the use of spreadsheets, are time-consuming and prone to technical errors. The integration of external software in the workflow, such as Toxicological Prioritization Index (ToxPi)[14,15], introduces additional complexity, especially due to the need to transfer significant data sets from one software to another. Although, ToxPi is a valuable software tool for material grouping and visualization, its capabilities are limited by the lack of data preprocessing functions and limited options for presenting and storing the obtained results.

8.1 Cell-based high-throughput screening and Tox5-Score methodology for clustering in-vitro toxicity data of chemical substances and nanomaterials.

Screening methodology, developed in the department of Toxicology at Misvik Biology, allows rapid toxicity assessment of multiple materials using five well-established toxicity assays:

- CellTiter-Glo ® for determining cell viability by luminescence measurement on a plate reader;
- Dapi for determining the number of living cells by fluorescence microscopy;
- gammaH2AX for detection of DNA damage by fluorescence microscopy;
- 8OHG for sensing nucleic acid oxidative stress by fluorescence microscopy;
- Caspase - Glo 3 to determine apoptosis by fluorescence microscopy and specific Caspase - Glo assay ® 3/7 but by luminescence measurement on a plate reader.

The analyzes are carried out by varying several time points, different concentration levels and different cell lines, adapted from the described methodologies [16, 17]. Traditional toxicity testing is based on determining the concentration of a given substance at which cell growth is inhibited by 50% (GI_{50}),. Combining toxicity results from multiple time points and multiple assays can provide more sensitive and specific toxicity estimates. The GI_{50} could not be calculated or is not optimal for some of the selected assays used in the Misvik methodology. For this reason, the GI_{50} – independent Tox5-score scoring system was developed and used (Figure 11). The Tox5-score approach[18], is based on calculated parameters from the dose-response curve, integrating these indicators into a single overall toxicity score, expressed as a toxicological prioritization index that preserves transparency to the contribution of each specific analysis.

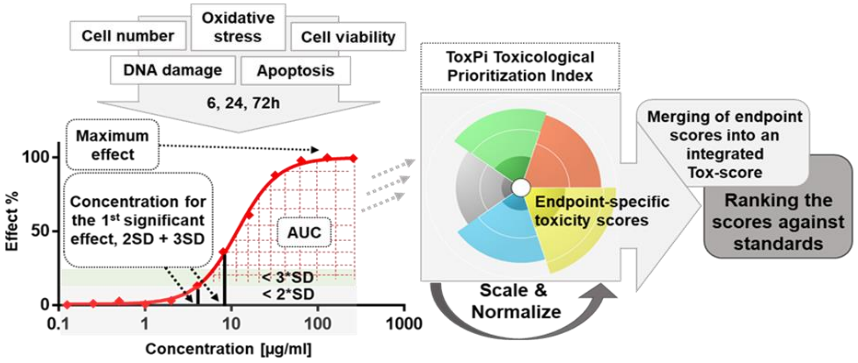

Cell number
Oxidative stress
Cell viability
DNA damage
Apoptosis
6, 24, 72h
Maximum effect
100
50
0
Effect %
Concentration for the 1st significant effect, 2SD + 3SD
AUC
< 3*SD
< 2*SD
0.1
1
10
100
1000
Concentration [µg/ml]
ToxPi Toxicological Prioritization Index
Endpoint-specific toxicity scores
Scale & Normalize
Merging of endpoint scores into an integrated Tox-score
Ranking the scores against standards

Figure 11: *Tox5-score approach for scoring in-vitro toxicity data.*

The toxicity profile represented by a pie chart that describes the similarity in toxicity responses. Complexity and visualization varies depending on the amount of information included. The approach allows a clear visualization of the overall assessment and the materials could be ranked and compared to control substances of known toxicity. Using Tox5-score, data from other experiments with similar parameters can be combined, which also expands the method applicability.

Based on the developed methodology for high-throughput screening and the approach for calculating the toxicological prioritization index Tox5-score, we introduced a general concept for pretreatment and grouping of materials. On top of the general concept, we developed a ToxFAIRy software library and an Orange3-ToxFAIRy user interface.

8.2 A general concept for high-throughput screening data preprocessing and Tox5-Score clustering

Screening data preprocessing and toxicological index calculation steps of Tox5-Score are shown in Figure 12. The latter can be implemented with various computational techniques, such as MS Excel or other software.

The first step is related to collecting and annotating experimental data with metadata, as well as recalculating the applied doses according to the specific surface area of nanomaterials and relative to the cell growth area. The second step involves normalization and data preprocessing, with the aim of minimizing the impact of systematic errors caused by human, biological or technical factors during the experimental process. This reduces unwanted variation and noise level in the data. The third step involves calculating parameters from the dose-response curves: maximum effect, area under the curve, and first significant dose. The final step is the calculation of a toxicological prioritization index and grouping of materials.

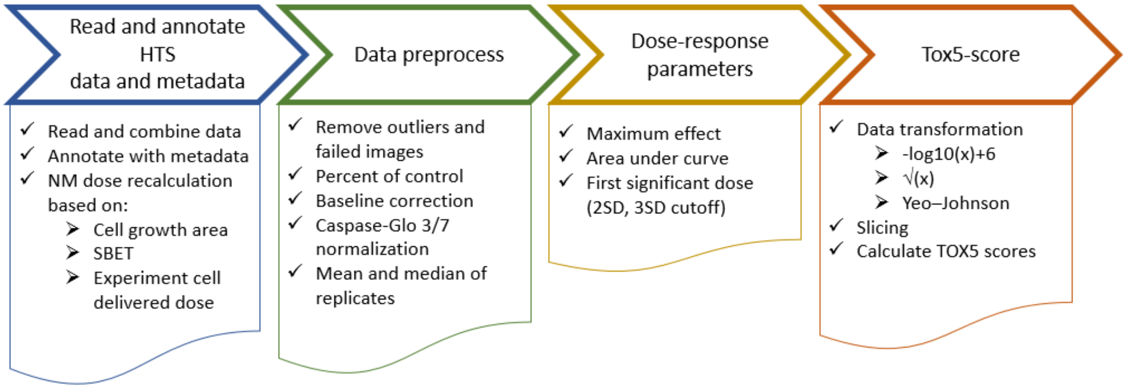


Figure 12: *General steps in the workflow for HTS data preprocessing and toxicity scoring.*

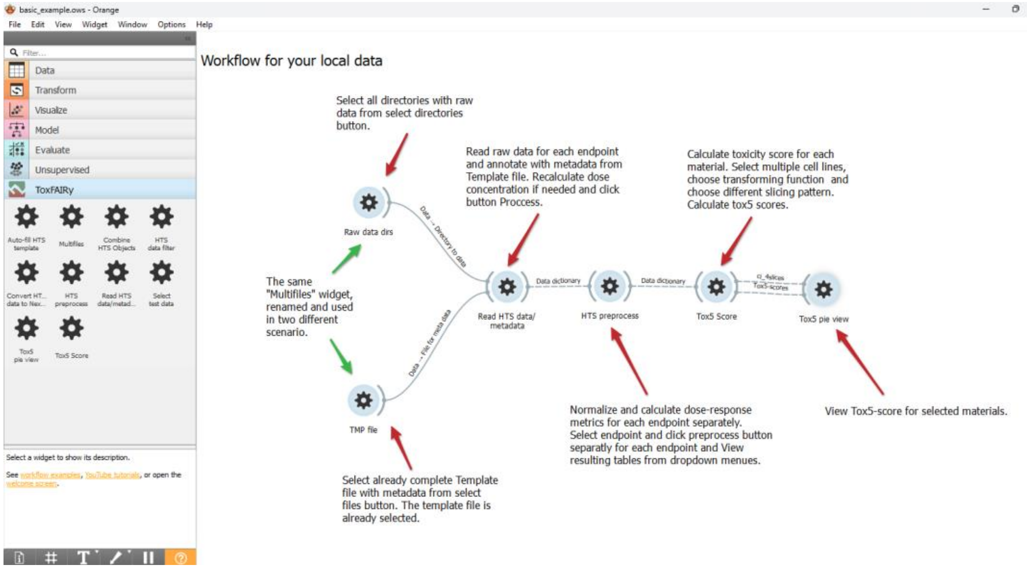
8. 3 ToxFAIRy software library

We developed a python library ToxFAIRy for: (i) reading and annotation of HTS data, (ii) pre-processing, (iii) data FAIRification and (iv) toxicity assessment. The ToxFAIRy library is available in the platform's orange3-toxfairy repository GitHub via the link:

https://github.com/ideaconsult/orange3-toxfairy.

8.4 Orange3-ToxFAIRy Add-on

We have developed an Orange3-ToxFAIRy add-on to the Orange machine learning and data analysis software. Orange3-ToxFAIRy can be used as a user interface to the ToxFAIRy library. The individual software components allow building of a complete automated workflow (Figure 13). The main advantage is that the user can create complex workflows in visual programming manner and be saved as OWS files and uploaded for reuse and/or shared with other users. We have also developed a visual guide for the user interface, available from the link: https://doi.org/10.5281/zenodo.13685297.



Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

Select all directories with raw data from select directories button.

Read raw data for each endpoint and annotate with metadata from Template file. Recalculate dose concentration if needed and click button Proccess.

Calculate toxicity score for each material. Select multiple cell lines, choose transforming function and choose different slicing pattern. Calculate tox5 scores.

The same "Multifiles" widget, renamed and used in two different scenario.

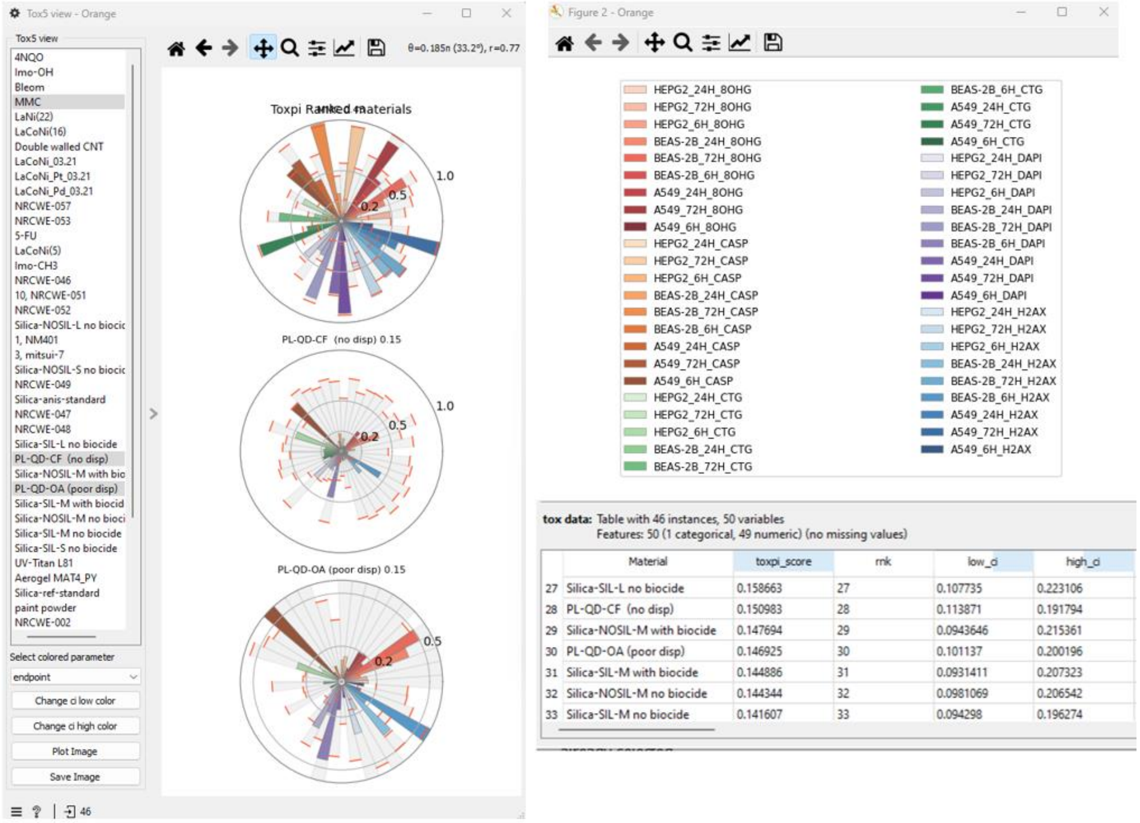
Normalize and calculate dose-response metrics for each endpoint separately. Select endpoint and click preprocess button separatly for each endpoint and View resulting tables from dropdown menues.

View Tox5-score for selected materials.

Select already complete Template file with metadata from select files button. The template file is already selected.

Figure 13: *Orange Tox5-score workflow. Basic example of workflow for automatic HTS data processing and Tox5-scoring.*

Toxicity profiles are visualized for each material separately using the Tox5 view software component (Figure 14). Confidence intervals were calculated and are shown in gray for each specific slice. For user convenience and more efficient work, we have also developed an automated approach to color sections according to assays, cell lines or treatment time. In the example shown, the sections are colored based on the different analyses. In addition to the graphical visualization of the material toxicity results, the information is also available in tabular form.



4NQO
Imo-OH
Bleom
MMC
LaNi(22)
LaCoNi(16)
Double walled CNT
LaCoNi_03.21
LaCoNi_Pt_03.21
LaCoNi_Pd_03.21
NRCWE-057
NRCWE-053
5-FU
LaCoNi(5)
Imo-CH3
NRCWE-046
10, NRCWE-051
NRCWE-052
Silica-NOSIL-L no biocic
1, NM401
3, mitsui-7
Silica-NOSIL-S no biocic
NRCWE-049
Silica-anis-standard
NRCWE-047
NRCWE-048
Silica-SIL-L no biocide
PL-QD-CF (no disp)
Silica-NOSIL-M with bio
PL-QD-OA (poor disp)
Silica-SIL-M with biocid
Silica-NOSIL-M no bioci
Silica-SIL-M no biocide
Silica-SIL-S no biocide
UV-Titan L81
Aerogel MAT4_PY
Silica-ref-standard
paint powder
NRCWE-002

tox data: Table with 46 instances, 50 variables
Features: 50 (1 categorical, 49 numeric) (no missing values)

	Material	toxpi_score	rnk	low_ci	high_ci
27	Silica-SIL-L no biocide	0.158663	27	0.107735	0.223106
28	PL-QD-CF (no disp)	0.150983	28	0.113871	0.191794
29	Silica-NOSIL-M with biocide	0.147694	29	0.0943646	0.215361
30	PL-QD-OA (poor disp)	0.146925	30	0.101137	0.200196
31	Silica-SIL-M with biocide	0.144886	31	0.0931411	0.207323
32	Silica-NOSIL-M no biocide	0.144344	32	0.0981069	0.206542
33	Silica-SIL-M no biocide	0.141607	33	0.094298	0.196274

Figure 14: *Visualization of the Tox5-scores as a pie chart.*

8.5 Application of the automated workflow to HTS data from caLIBRAte and HARMLESS EU projects.

All data used for development and testing are publicly available from the Zenodo link: https://doi.org/10.5281/zenodo.13683162

ToxFAIRy library have been used to develop a scripted automated workflow with the ploomber library. The workflow is built from specific tasks (https://github.com/ideaconsult/orange3-toxfairy/tree/main/toxfairy_workflow/tasks) with well-established relationships between them, presented in Figure 15. The overall configuration of the workflow is set via a JSON file with specific instructions.

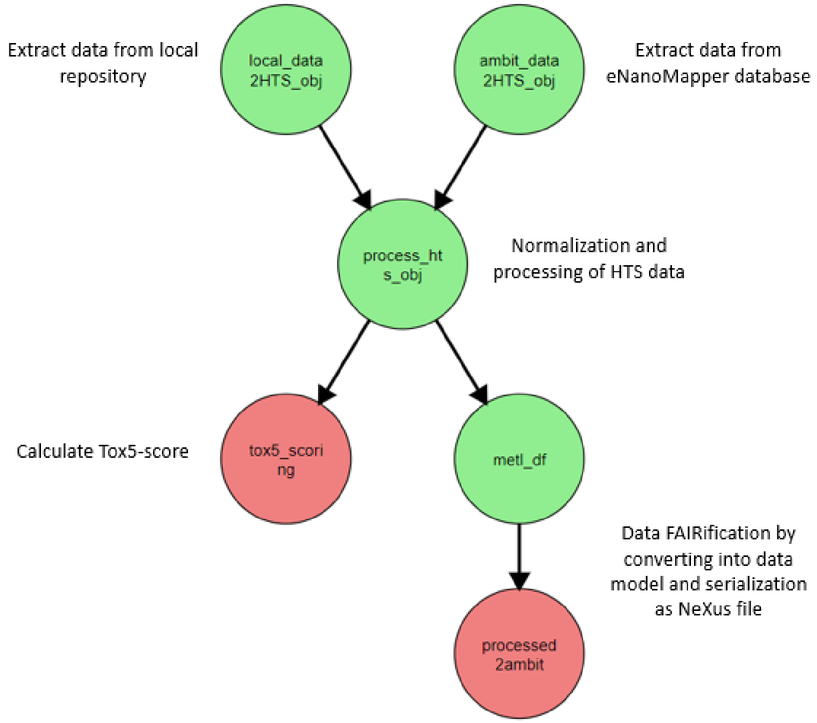

Extract data from local repository
local_data 2HTS_obj
ambit_data 2HTS_obj
Extract data from eNanoMapper database
process_hts_obj
Normalization and processing of HTS data
Calculate Tox5-score
tox5_scoring
metl_df
Data FAIRification by converting into data model and serialization as NeXus file
processed 2ambit

***Figure 15**: Automated workflow for processing, FAIRification and grouping of HTS data.*

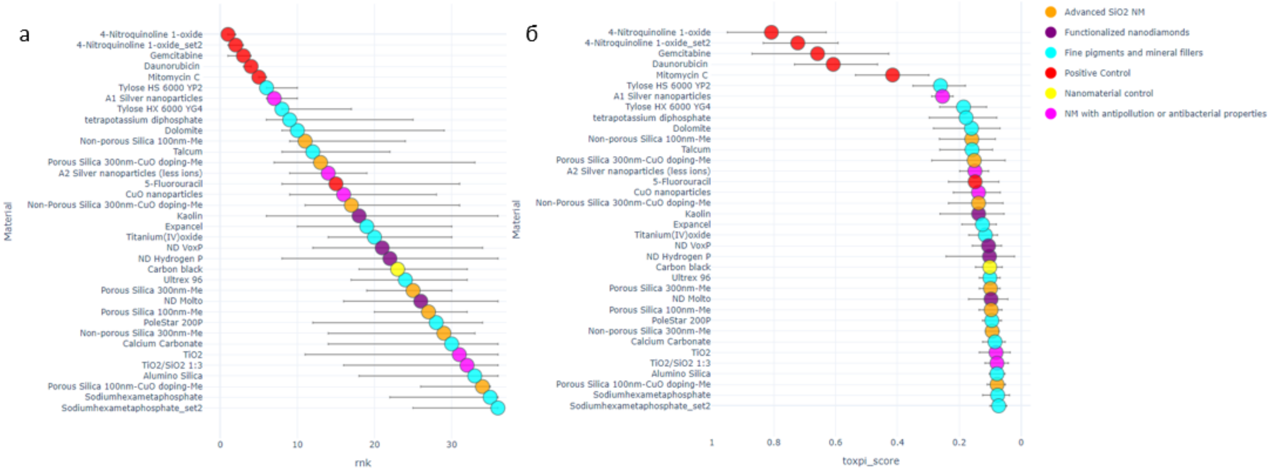


***Figure 16:** Ranked caLIBRAte materials and controls with bootstrap confidence interval a) by ranks and b) by Tox5-score.*

Figure 16 presents the results of the automatic processing and ranking of a set of materials according to data from the caLIBRAte project. Confidence intervals were calculated for the ranks and for the Tox5-score indices and are presented as error bars. The material groups are color-coded. The included positive controls, marked in black, and the nanomaterials control, marked in yellow, create a reference field for the relative toxic rankings of the materials and allow comparison with new data sets with similar controls. Tylose HX 6000 YG4, chemically modified hydroxyethyl cellulose, from the group of pigments and mineral filters, marked in blue, and TiO2/SiO$_2$, from the group of nanomaterials with anti-pollution and antibacterial properties, marked in light purple, are respectively ranked the most the toxic and least toxic nanomaterial tested.

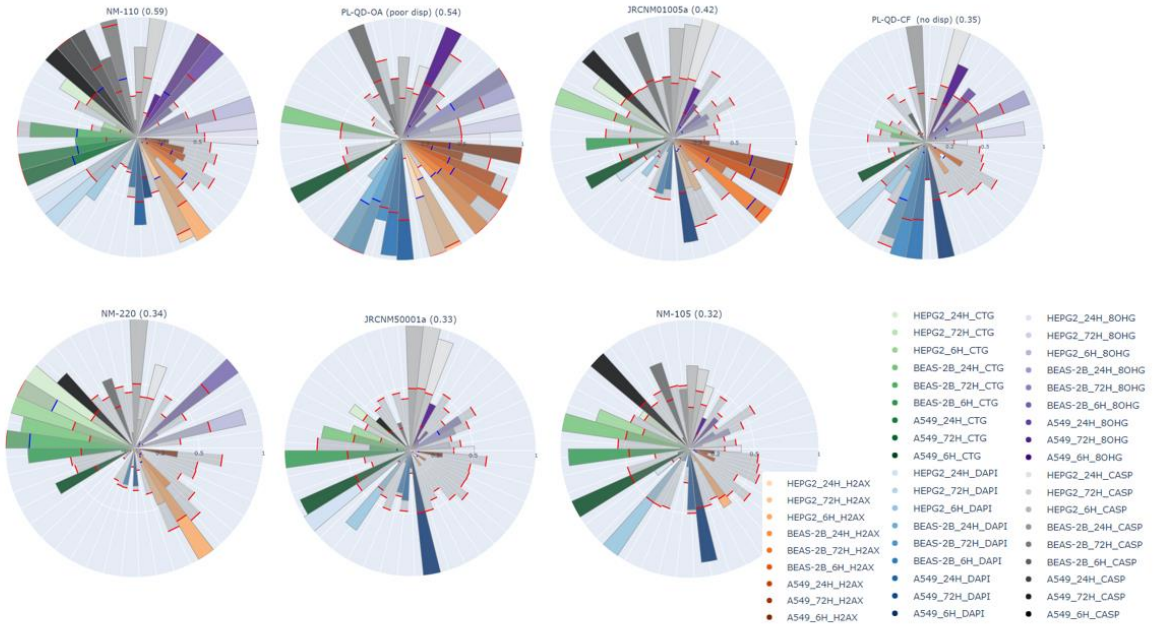


Figure 17: *Graphical representation of toxicity profiles for quantum dots and controls.*

Figure 17 demonstrates clustered quantum dots (ZnCuInS core/ZnS shell (PL-QD-CF) and ZnCdSeS no shell (PL-QD-OA)) and nanomaterials as positive and negative controls via a pie chart, where each pie represents the total Tox5-score for each material. Each slice corresponds to a specific result for a group of parameters, and is colored according to the analysis. For example, CTG analysis groups are colored green, with a gradient indicating the specific cell lines and time points. Also, the confidence interval for each slice is plotted with the upper bound shown in red and the lower bound in blue. Quantum dots reduce the number of viable cells (blue sectors) and cause oxidative stress to nucleic acids (purple sectors), while zinc oxide (NM-110), ranked as the most toxic, is more likely to induce apoptosis (gray sectors) and loss of cell viability (green sectors).

Figure 18 illustrates hierarchical clustering, for the calibrate project data, using Euclidean distance as metric and clustering method - Ward[19]. The optimal number of clusters was determined using the Elbow method, which demonstrated better statistical performance in terms of cluster significance indicators (Silhouette[20] , Davies-Bouldin[21] and Calinski-Harabasz[22] scores), compared to the Silhouette method. These clustering metrics are automatically calculated and visualized on the dendrogram.

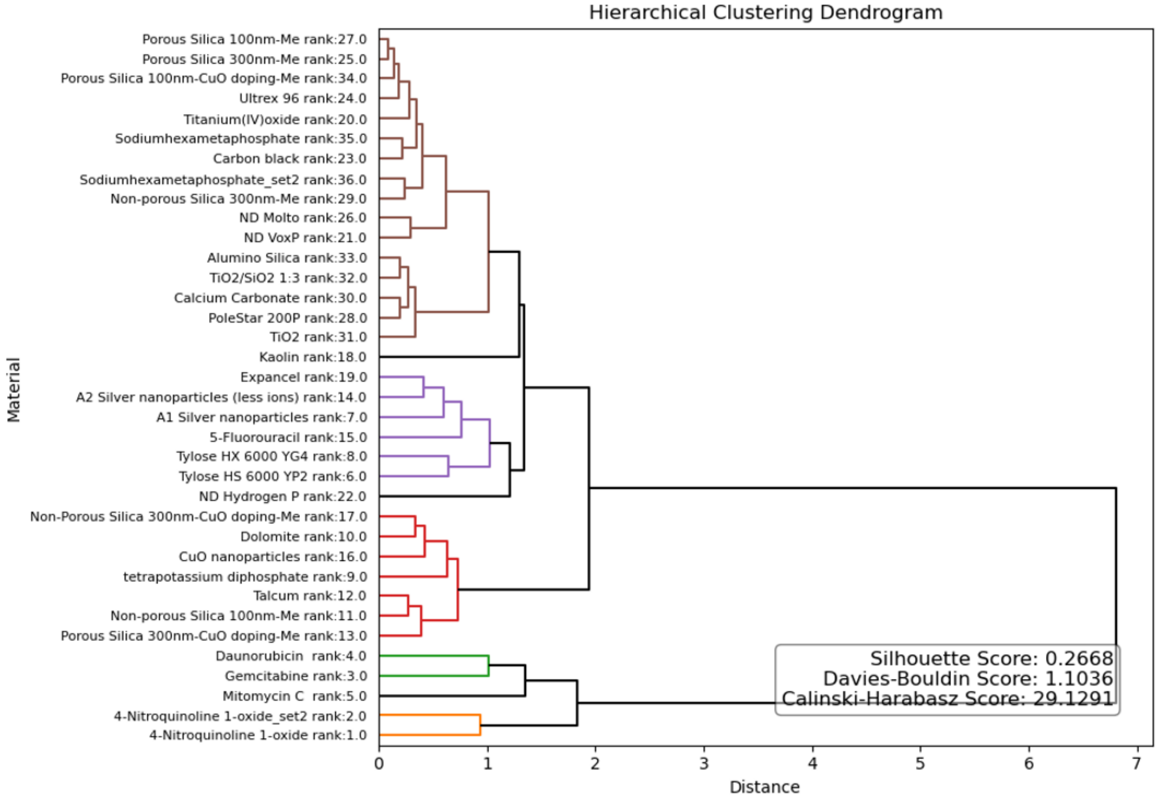


Figure 18: *Hierarchical clustering of caLIBRAte nanomaterials.*

Despite the small and heterogeneous data set, statistically significant clusters are formed, for example between nanodiamonds VoxP and Molto and porous silica particles.

The TOPSIS method was applied to the quantum dots data sample and compared to the results from the Tox5-score approach (see Table 2). A high value for TOPSIS preferences indicates how close the given alternative is to a less toxic material, while a high Tox5–-score represents potentially highly toxic material. The two methods show quite similar behavior in the clustering of the materials except for the quantum dot ZnCuInS core/ZnS shell and JRCNM01005a, which have swapped positions. We chose the TOPSIS method as an alternative and verification of the Tox5-score approach, also TOPSIS is noted as one of the five most popular multicriteria decision-making methods, described in the latest SSbD manuals of JRC[23].

Table 2*: Comparison of TOPSIS and Tox5-score prioritizations methods applied to quantum dots.*

Material	TOPSIS estimation of preferred effect	Tox5-score	TOPSIS rank	Tox5-score rank
NM-110	0.336	0.591	1	1
ZnCdSeS no shell	0.441	0.535	2	2
ZnCuInS core /	0.555	0.351	3	4

ZnS shell				
JRCNM01005a	0.631	0.425	4	3
NM220	0.635	0.343	5	5
JRCNM50001a	0.702	0.330	6	6
NM-105	0.807	0.320	7	7

8.6 FAIRification of highthroughput screening (HTS) data

After preprocessing and calculation of dose-effect parameters, the data structures are converted to the Ambit/eNanoMapper data model using the pynanomapper library and serialized as a NeXus file. This enables database import and access to high-throughput screening (HTS) data as a quality FAIR resource. The resulting NeXus file contains all materials, analyses, including raw and processed data. Data is stored as a hierarchical tree structure of multidimensional arrays. Figure 19 presents a Nexus file for Material: Porous Silica 300nm-CuO doping-Me, assay: 8OHG with cell line: BEAS-2B. Figure 19 is a heatmap plot of normalized data with a selected subset of data from replicates D2.

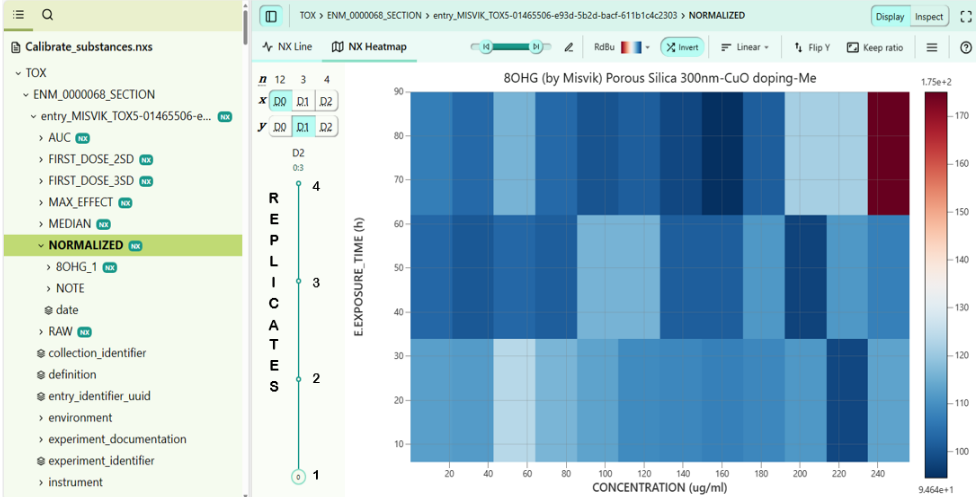


Figure 19*FAIR HTS data serialized as NeXus format*

IV Conclusions

9.1 Scientific contributions

1. FAIRification methodoly for experimental data for nanomaterials was developed, based on the representation of information on multicomponent substances through the Ambit/eNanoMapper semantic data model and the configurable software tool NMDataParser.
2. Nanomaterials prototype identifier that demonstrates the capabilities of SLN linear notation for representing chemoinformatics and nanoinformatics objects with the potential to

generate a globally unique substance identifier. The created prototype allows rich metadata to be encoded together with the structural information, which is important in the context of chemical information management for academic, regulatory and industrial needs.

3. A general concept for annotating HTS data with metadata, preprocessing and calculation of a toxicological prioritization index using the Tox5-Score approach. The overall concept enables the development of software tools for automated data processing from HTS, which is consistent with regulatory recommendations and widely expressed industry needs.

9.2 Applied science contributions

1. FAIRification of 1400 EXCEL files, enriching the eNanoMapper database with high-quality FAIR data, with information on the safety of nanomaterials from several major European projects. The data includes a wide variety of physicochemical and biological analyzes for multiple nanomaterials.
2. eNanoMapper ontology is enriched in the area of ecotoxicity and environmental protection by adding 15 new terms.
3. Based on the developed general concept for annotation and processing of HTS data, we developed an open source software library ToxFAIRy, for annotation and data processing, calculation of the prioritization index and FAIRification of HTS data.
4. Implementation of Orange 3-ToxFAIRy module to the Orange analytical platform, which can be used as a user interface to the ToxFAIRy software library.
5. We have developed an automated workflow on the Ploomber platform that fully implemented the functionalities of the ToxFAIRy library and enables fast simultaneous processing of a large volume of data. The automated workflow has been applied to HTS data for nanomaterials and advanced materials generated by the European projects caLIBRAte and HARMLESS.

9.3 Directions for future development

The results described in this thesis outline some main directions for future development:

1. Development of software tools for the calculation of descriptors for nanomaterials, advanced materials and multicomponent substances based on the representation of objects through the FAIR semantic data model.
2. Application of new modern algorithms and vector representations (AI embedding), based on artificial intelligence, for modeling, QSPR/QSAR analysis, read across methodologies and linking the Ambit/eNanoMapper data model with semantic knowledge-based systems.
3. Refinement and development of the proposed SLN prototype identifier for multicomponent substances and nanomaterials. Development of algorithms to make the SLN identifier unique. This would facilitate not only the scientific community, but also regulatory agencies and industry in managing data and storing information about nanomaterials.

4. Expanding the functionality of the ToxFAIRy software library in several directions:

- development of new toxicological indices including physicochemical indicators and results of in-vivo analyzes integrated from the eNanoMapper database .
- development of approaches for integrating information from graph databases, based on AOPs (Adverse Outcome Pathways)[24], to serve for validation of the obtained results of substances and nanomaterials prioritization.

V Scientific announcements on the dissertation

10.1 Articles

A1 Kochev, N.; Jeliazkova, N.; Paskaleva, V.; Tancheva, G.; Iliev, L.; Ritchie, P.; Jeliazkov, V. "Your Spreadsheets Can Be FAIR: A Tool and FAIRification Workflow for the eNanoMapper Database". Nanomaterials 2020, 10, 1908. https://doi.org/10.3390/nano10101908
with 16 citations

A2 Kochev N, Jeliazkova N, Tancheva G. "Ambit-SLN: an Open Source Software Library for Processing of Chemical Objects via SLN Linear Notation". Mol Inform. 2021 Nov;40(11):e2100027. doi: 10.1002/minf.202100027. Epub 2021 Aug 3. PMID: 34342942.
with 1 citation

A3 Jeliazkova N, Kochev N, Tancheva G. "FAIR Data Model for Chemical Substances: Development Challenges, Management Strategies, and Applications". Data Integrity and Data Governance. IntechOpen; 2023. Available from : http://dx.doi.org/10.5772/intechopen.110248

1 0.2 Posters and Reports

P1 G. Tancheva, N. Kochev, N. Jeliazkova, V. Paskaleva, *DATA PROCESSING FOR CHEMICAL SUBSTANCES AND NANOMATERIALS*, scientific conference Current regulatory requirements for chemical analysis and modern instrumental technologies for their coverage, 06.2019, Plovdiv

P2 G. Tancheva, N. Kochev, V. Paskaleva, *QSAR modeling of melting points of organic compounds. Methods comparison.* Fifth scientific conference for students, doctoral students and young scientists "Challenges in Chemistry", October 2019, Plovdiv.

P3 G. Tancheva, N. Kochev, N. Jeliazkova, V. Paskaleva, L. Iliev, P. Ritchie, V. Jeliazkov, *FAIRification Workflow for Handling Nano Safety Excel Spreadsheet templates in eNanoMapper database* . International FAIR Convergence Symposium, 12.2020

P4 G. Tancheva, N. Kochev, V. Jeliazkov, L. Iliev, N. Jeliazkova, *Fairification workflow for integrating nanosafety data : Enanomapper database*, ACM2 seminar 06.2022, Plovdiv

P5 G. Tancheva, N. Kochev, N. Jeliazkova *Electronic notebook for nanosafety data interactive preprocessing analysis and visualisation*. 6 Scientific conference for students, doctoral students and young scientists "Challenges in Chemistry", 10. 2022, Plovdiv

P6 G. Tancheva, P. Nymark, V. Hongisto, N. Kochev, N. Jeliazkova, *Automatic workflow for HTS data FAIRification, preprocessing and Tox5 in-vitro toxicity scoring.* QSAR - 2023, 06. 2023, Copenhagen

P7 G. Tancheva, P. Nymark, V. Hongisto, N. Kochev, N. Jeliazkova , *Automatic workflow for in vitro high-throughput screening data FAIRification, preprocessing and scoring* , 12 Chemical Conference, 11. 2023, Plovdiv

P8 G. Tancheva, V. Hongisto, K. Patyra, L. Iliev, N. Kochev, P. Nymark, R. Grafström, N. Jeliazkova*, Automatic workflow for in vitro high-throughput screening data FAIRification, preprocessing and scoring*, General Assembly HARMLESS, 10-11. 01.2024, Ludwigshafen, https://www.harmless-project.eu/harmless-general-assembly-2024/

P9 G. Tancheva, V. Hongisto, K. Patyra , L. Iliev, N. Kochev , P. Nymark, R. Grafström , N. Jeliazkova , *Automatic workflow for in in vitro high-throughput screening data FAIRification, preprocessing and scoring* , NANOTOX 2024, 09., Venice.

R1 *Fairification workflow for integrating nanosafety data: Enanomapper database.* SciDataCon - IDW Seoul 06. 2022. Oral report online form.

R2 *Automatic workflow for HTS data FAIRification, preprocessing and Tox5 in-vitro toxicity scoring,* M30 General Assembly, HARMLESS project, Turku (Finland) 13 – 14. 06. 2023. Oral report online form, https://www.harmless-project.eu/general-assembly-m30-in-turku/

R3 *Automatic workflow for HTS data FAIRification, preprocessing and toxicity scoring. Case study: Quantum dots,* Summary meeting of collaborative projects under the HORIZON 2020 program (DIAGONAL, HARMLESS, SUNSHINE) NMBP-16 ambassadors, 18. 06. 2024, Oral report online form.

10.3 Projects

Project 1 **Transnational Access (TA) into NanoCommons** project in collaboration with Maastricht University (HORIZON 2020 EU, agreement #731032) , https://www.nanocommons.eu/e-infrastructure/awarded-ta-projects/

Project 2 **HARMLESS** - Advanced High Aspect Ratio and Multicomponent materials: towards comprehensive intelligence Testing and Safe by design Strategies, (HORIZON 2020 EU agreement ID: 953183) https://www.harmless-project.eu/project-summary/

Project 3 **NanoReg2**-Development & Implementation of Grouping & Safe-by-Design Approaches within Regulatory Frameworks, (Grant agreement ID: 646221)

Project 4 **NanoinformaTIX**-Development and Implementation of a Sustainable Modelling Platform for NanoInformatics (Grant agreement No 814426) https://www.nanoinformatix.eu/

Project 5 **POLYRISK**- Understanding human exposure and health hazard of micro- and nanoplastic contaminants in our environment (Grant agreement No 964766) https://polyrisk.science/

VI References

1. Engel, T; Gadteiger J. *Chemoinformatics: A Textbook*.; 2003. doi:10.1002/3527601643
2. EU Commission. COMMISSION RECOMMENDATION (EU) 2022/2510.; 2022. doi:10.1890/0012-9623(2004)85[163:po]2.0.co;2
3. Mark D. Wilkinson; et all; Comment: The FAIR Guiding Principles for scientific data management and stewardship. 2016:1-9.
4. MESOCOSM database. https://aliayadi.github.io/MESOCOSM-database/.
5. Kochev N, Jeliazkova N, Paskaleva V, et al. Your spreadsheets can be fair: A tool and fairification workflow for the enanomapper database. *Nanomaterials*. 2020;10(10):1-23. doi:10.3390/nano10101908
6. eNanoMapper. https://www.enanomapper.net/.
7. Онтология eNanoMapper. https://github.com/enanomapper/ontologies.
8. Janna Hastings (EMBL-EBI) EW (UM). Deliverable Report D2.1 Framework and Infrastructure for ontology development, versioning and dissemination. 2014;(February 2014):1-30.
9. Hastings J, Jeliazkova N, Owen G, et al. eNanoMapper : harnessing ontologies to enable data integration for nanomaterial risk assessment. 2015:1-15. doi:10.1186/s13326-015-0005-5
10. Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *J Chem Inf Comput Sci*. 1997;37(1):71-79. doi:10.1021/ci960109j
11. Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD. SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. *J Chem Inf Model*. 2008;48(12):2294-2307. doi:10.1021/ci7004687
12. CHEBI 51050. https://www.ebi.ac.uk/chebi/searchId.do?chebiId=51050.
13. Levenshtein distance. https://en.wikipedia.org/wiki/Levenshtein_distance#cite_note-1.
14. To KT, Fry RC, Reif DM. Characterizing the effects of missing data and evaluating imputation methods for chemical prioritization applications using ToxPi. *BioData Min*. 2018;11(1):1-12. doi:10.1186/s13040-018-0169-5
15. Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. ToxPi Graphical User Interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. *BMC Bioinformatics*. 2018;19(1):1-7. doi:10.1186/s12859-018-2089-2
16. Kohonen P, Ceder R, Smit I, et al. Cancer biology, toxicology and alternative methods development go hand-in-hand. *Basic Clin Pharmacol Toxicol*. 2014;115(1):50-58. doi:10.1111/bcpt.12257
17. Grafström RC, Nymark P, Hongisto V, et al. Toward the replacement of animal experiments through the bioinformatics-driven analysis of “omics” data from human cell cultures. *ATLA Altern to Lab Anim*. 2015;43(5):325-332. doi:10.1177/026119291504300506
18. Hongisto V, Nymark P. Systems toxicology to support development of adverse outcome pathways, Abstracts of the 55th Congress of the European Societies of Toxicology (EUROTOX 2019) TOXICOLOGY SCIENCE PROVIDING SOLUTIONS. *Toxicol Lett*. 2019;314(October):S25. doi:10.1016/j.toxlet.2019.09.002
19. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc*.

1963;58(301):236-244. doi:10.1080/01621459.1963.10500845

20. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20(C):53-65. doi:10.1016/0377-0427(87)90125-7
21. Davies DL, Bouldin DW. A Cluster Separation Measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227. doi:10.1109/TPAMI.1979.4766909
22. Calinski T, Harabasz J. Communications in Statistics A dendrite method for cluster analysis. *Commun Stat*. 1974;3(1):1-27.
23. Abbate E, Garmendia Aguirre I, Bracalente G, et al. *Safe and Sustainable by Design Chemicals and Materials - Methodological Guidance*.; 2024. doi:10.2760/28450
24. Adverse Outcome Pathway. https://aopwiki.org/.