

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter

FAIR Data Model for Chemical Substances: Development Challenges, Management Strategies, and Applications

Nina Jeliaskova, Nikolay Kochev and Gergana Tancheva

Abstract

Data models for representation of chemicals are at the core of cheminformatics processing workflows. The standard triple, (structure, properties, and descriptors), traditionally formalizes a molecule and has been the dominant paradigm for several decades. While this approach is useful and widely adopted from academia, the regulatory bodies and industry have complex use cases and impose the concept of chemical substances applied for multicomponent, advanced, and nanomaterials. Chemical substance data model is an extension of the molecule representation and takes into account the practical aspects of chemical data management, emerging research challenges and discussions within academia, industry, and regulators. The substance paradigm must handle a composition of multiple components. Mandatory metadata is packed together with the experimental and theoretical data. Data model elucidation poses challenges regarding metadata, ontology utilization, and adoption of FAIR principles. We illustrate the adoption of these good practices by means of the Ambit/eNanoMapper data model, which is applied for chemical substances originating from ECHA REACH dossiers and for largest nanosafety database in Europe. The Ambit/eNanoMapper model allows development of tools for data curation, FAIRification of large collections of nanosafety data, ontology annotation, data conversion to standards such as JSON, RDF, and HDF5, and emerging linear notations for chemical substances.

Keywords: FAIR, database, data model, chemical substance, nanomaterial, structure, molecular descriptors, linear notation, ontology

1. Introduction

Since the emergence of the term cheminformatics within the context of pharmaceutical industry activities around the end of the twentieth century, an adequate chemical structure representation has been essential for the efficient application of cheminformatics methodologies [1]. The chemical structure is at the core of various cheminformatics activities: molecular property prediction via Quantitative Structure-Property Relationships/Quantitative Structure-Activity

Relationships (QSPR/QSAR), searching new biologically active compounds, lead optimization, virtual screening, combinatorial chemistry, etc. The centrality of molecular structure gives the primary flavor that distinguishes these activities from the classical chemometrics approaches [2], focused on data mining of the analytical and experimental results in order to extract useful information for the chemical objects study (e.g. the popular structure elucidation task). The chemometrics techniques from the 70s were transferred, adapted, and further developed within the field of “mathematical chemistry” with a strong focus on graph theory applications for molecule structures representation in the 80s and in 90s, and together with the 3D structure information focus and movement toward big data, resulted in the birth of modern cheminformatics. The main motto “from data to knowledge” summarizes the data workflow from studying chemical objects toward gaining/formalizing chemical information and generation of chemical knowledge as models, classifiers, etc. An adequate representation of the structures is required for all stages of the data management workflow. The chemical object representation development is a dynamic process, which is strongly influenced by the practical needs of the industry and lately, regulatory bodies. The novel deep learning technologies are changing the ways the structure information is used (e.g. linear notations can be directly read by the artificial neural networks as well as vector representations of the structures generated). The chemical substance model is a logical extension of the traditional molecule representation and takes into account practical aspects of chemical data management and new emerging research challenges. Finally, the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [3] were widely popularized and strongly encouraged as a needed background for efficient ongoing interconnections and activities within the: academia, industry, and regulators. On the other hand, the substance paradigm is based on a more complex approach toward representation of the chemical objects and must handle multiple material compositions, enriched with mandatory metadata and corresponding ontology annotations in order to comply with FAIR principles.

In the following sections, the reader will be led into a journey, starting from the classical molecular data model, based on chemical structure, and going through complex representations of chemical substances, industry use cases, and nanomaterials (NMs). The logical evolution of the data model elucidation will be demonstrated within the context of various challenges. The importance of metadata will be discussed as well as the adoption of FAIR principles. The good practices will be exemplified by the Ambit/eNanoMapper data model and real chemical substances from ECHA REACH dossiers. This chapter also discusses the FAIRification of large collections of data and the importance of standard data formats and emerging linear notations for chemical substances.

2. Classical cheminformatics paradigm for molecular data: structure, properties, and descriptors

The cheminformatics is a vast interdisciplinary field with a large inheritance from the data mining, graph theory, and mathematical chemistry, enriched with modern methods for big data and artificial intelligence approaches. A common denominator of this methodological variety is the focus on the chemical structure. The centrality of chemical structure is also evident in other domains, strongly related to the cheminformatics, such as reaction informatics, bioinformatics (e.g. proteomics and metabolomics), toxicogenomics, etc. In QSPR/QSAR analysis, physicochemical

properties and biological activities are considered as functions of the molecular structure, i.e. $P = f(S)$ or $A = f(S)$. Also, equation reversal is observed for the chemometrics' structure elucidation task: $S = f^{-1}(P)$, e.g. structure is obtained out of the spectral data (spectrum is the property vector, P , in this case). The representation of the chemicals is the starting point for any of these activities. The molecular structure is the principle "model" that encompasses most important bits of the current chemical knowledge, used for further data processing and modeling.

The hierarchy of basic chemical objects' representations is shown in **Figure 1**. It starts from the smallest chemical objects, atoms, and bonds, which are the building blocks for the chemical structure. The connection table (CT) encodes the chemical graph and is the most widely used approach for structure information representation on a topological level. 2D coordinates and 3D coordinates together with the CT fully describe a chemical structure. Traditionally, the transition from structure to property is helped by an intermediate layer of descriptors, D , i.e. the first step is $D = f_1(S)$ and then $P = f_2(D)$.

The classical and widely adopted data model of a molecule representation is defined as a triple of the type (S-structure, D-descriptors, and P-properties), as illustrated in **Figure 2**. Different structure representations are systemized in several

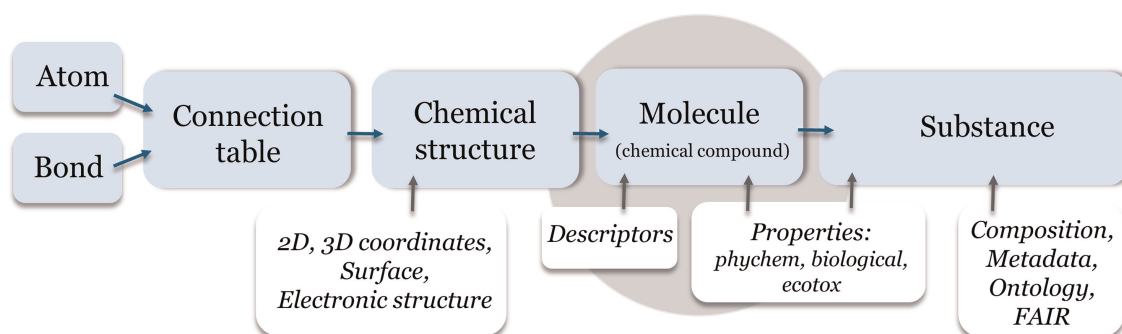


Figure 1. Hierarchy of chemical objects: From primitive/small objects (left) to larger and complex objects (right).

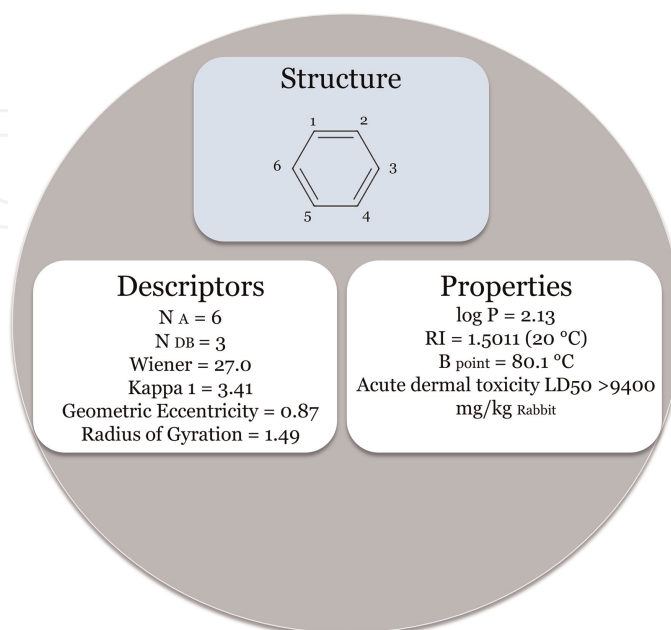


Figure 2. Classical triad model of a molecule: (structure, descriptors, and properties).

levels: 0D/1D – constitution, 2D – topology, 3D – geometry, 4D – conformation, and the QM (quantum mechanics) level with detailed electronic structure information. The intermediate layer of molecular descriptors is derived computationally or experimentally and represents useful information for the molecule. Structural descriptors are an important subset of descriptors, used as the principle interface between structure and properties. The structure is reduced to a simpler representation, namely a point in n-dimensional vector space. Variety of cheminformatics tasks, such as searching, classification, virtual screening, clustering, and measuring distance between the objects, can be performed in terms of points in the chemical space of so called “patterns.” Traditionally, the chemical patterns are considered more user-friendly to the classical machine learning methods than the original chemical objects.

Figure 3 shows various structure representations for the molecule of benzene: connection table, 2D and 3D coordinates (with corresponding graphical model), linear notations – SMILES, InChI and SLN, distance matrix as a topological descriptor and registry numbers CAS N, (EC) Number, and PubChem CID. Also, **Figure 2** exemplifies different descriptors: constitutional (N_A , N_{DB}), topological (Wiener index and kappa1 index), and geometrical (eccentricity and radius of gyrations) plus the third data layer with molecular properties: LogP, RI, BP, etc.

The majority of chemical database implementations are based on the classical structure paradigm – the molecule triad (S, D, P). This paradigm has been used for several decades, and even nowadays it is the predominant base layer for the public chemical databases. Naturally, the cheminformatics community and academic circles feel quite comfortable within the triad model. It has been like a “protecting bubble” and proved its usefulness as a “ground zero” for the chemical information workflow. However, staying in the limits of the classical (S, D, P) model may hinder or isolate the cheminformatics field evolution. The “conveniences” and simplicity of the (S, D, P) model may prevent the establishing of efficient interconnections between the

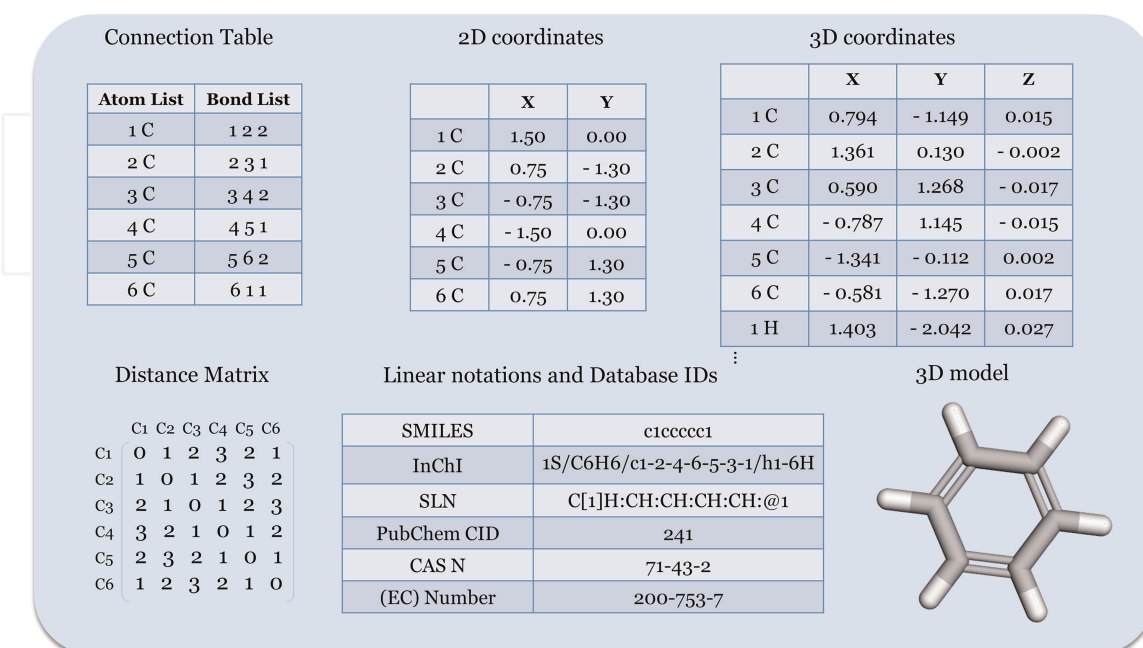


Figure 3. Various structure representations, descriptors, linear notations, and registry (database) indices for the molecule of benzene.

cheminformatics field and other scientific areas, especially in the context of industry and regulators. In the following sections, we describe the further development of the classical triad model into the paradigm of chemical substances (see the last element from the chemical objects chain in **Figure 1**).

3. Data models for chemical substances

The chemical structure describes a well-defined molecule. Unlike chemical structures, real chemical objects or industrially manufactured ones are not pure substances. Such substances are composed of several components; hence, they cannot be associated with a single unique structure. The regulatory authorities typically need information on chemicals as produced by industry. Another data gap emerges from the lack of tools to consider metadata about the performed experiments and measurements in cheminformatics use cases, e.g. QSAR model building, while such metadata is crucial for the toxicologists and regulators. The substance have to be represented as the entirety of the components with their roles and relations, include rich metadata to enable unambiguous description of experimental results from many biological assays, physicochemical characterizations, exposure, and environmental fate tracking. The challenges are increasing with representation of nanomaterials and advanced materials. Having a consensus on the chemical substance definition is a challenge also due to the discrepancies between the approaches of various regulatory institutions.

According to the International Union of Pure and Applied Chemistry (IUPAC) definition [4], a substance *“is matter of constant composition best characterized by the entities (molecules, formula units, atoms) it is composed of. Physical properties such as density, refractive index, electric conductivity, melting point etc. characterize the chemical substance.”* Under Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), the concept of substance is clearly described [5]: *“A substance is a chemical element and its compounds in the natural state or the result of a manufacturing process. In a manufacturing process, a chemical reaction is usually needed to form a substance.”* Under REACH, a chemical substance is composed of three types of components: constituents, impurities, and additives. Chemical substances can be mono-constituent (one main constituent is present to at least 80% (w/w)), multi-constituent (more than one main constituent is present in a concentration between 10% and 80% (w/w)), or UVCB (Substance of Unknown or Variable composition, Complex reaction products or Biological materials). The REACH definition of a substance encompasses all forms of substances and materials on the market, including nanomaterials.

The Government of Canada [6] defines a chemical substance as: *“Elements or compounds that are deliberately created, produced as a by-product of other processes or occurring naturally in the environment.”* The Canadian Environmental Protection Act (CEPA) also requires notification for new substances put in two lists: the Domestic Substances List (DSL) and the Non-DSL (NDSL). Toxic Substances Control Act (TSCA) [7] requires the United States Environmental Protection Agency (US EPA) to compile, keep current, and publish a list of each chemical substance that is manufactured or processed, including imports, in the United States for uses under TSCA. TSCA defines a “chemical substance” *as any organic or inorganic substance of a particular molecular identity, including any combination of these substances occurring in whole or in part as a result of a chemical reaction or occurring in nature, and any element or uncombined radical.* The Japanese Act on the Evaluation of Chemical Substances and Regulation of Their Manufacture is performed under Chemical Substance Control Law (CSCL) [8].

The underlying data model is of crucial importance for the efficiency of any cheminformatics, nanoinformatics, and bioinformatics workflow. Specifically, nanomaterial (NM) representations are the primary subject of the new and rapidly evolving field of nanoinformatics. According to ISO TS 80004-1:2015, definition of a nanomaterial is: “*a material with any external dimension in the nanoscale approximately 1 nm to 100 nm and/or having internal structure or surface structure in the nanoscale.*” [9]. The European Commission [10] definition of a nanomaterial is: “*A natural, incidental or manufactured material containing particles, in an unbound state or as an aggregate or as an agglomerate and where, for 50 % or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm-100 nm. In specific cases and where warranted by concerns for the environment, health, safety, or competitiveness, the number size distribution threshold of 50 % may be replaced by a threshold between 1 and 50 %.*” The substance definition in the European Union regulation REACH [5] and in the Classification, Labelling and Packaging (CLP) Regulation includes all forms of substances and materials on the market, including NMs, i.e. NM is treated as a particular case of a chemical substance.

There are several major data models highlighting the path for storing chemical substances in a database. IUCLID [11], the primary software for preparation and submitting REACH dossiers, stores and maintains data on the hazardous properties of chemical substances and mixtures, as well as their use and associated exposure levels. This is also the first system that fully implements the OECD harmonized templates (HT) [12] on the base of OECD guides of testing and agreed standards. The BioAssay Ontology (BAO) [13] provides a foundation for standardizing assay descriptions and endpoints with capabilities enabling the retrieval of data, relevant to a query. This is the first ontology to describe this domain, and certainly the first time that bioassay and HTS (high throughput screening) data have been represented using expressive description logic [14].

CODATA, the International Council for Science: Committee on Data for Science and Technology (www.codata.org), and VAMAS, an international pre-standardization organization, concerned with materials test methods (www.vamas.org), jointly foster the development of a uniform description system for NMs to address the diversity and complexity of nanomaterials. CODATA [15] encourages the interoperability and the usability of such data using a framework with four basic information categories General Identifiers, Characterization, Production and Specification and numerous subcategories and descriptors for detailed information. Most of the terms and concepts used in the descriptive system are easily understandable for people from different directions, as it is expected to be used by different groups of users for research reports, NM identification in regulations and standards, specifying NMs in commercial transactions, etc. [16].

ISA-TAB [17] defines three basic layers for sharing metadata, related to experiments: Investigation, Study and Assay, and the actual experimental data is stored on a separated forth layer and referenced by the ISA data [18]. Additional configuration settings and ontology annotations could be considered as additional layers to this complex multi-layered approach. The ISA model is non-standardized and user-defined and can include image files, spreadsheets, and protocol documents, forwarded to appropriate fields in the Study file table. The basic approach to present chemical compounds in ISA-TAB is an ontological record, which usually points to a single chemical structure. ISA model can be serialized via ISA-TAB [19] format as multiple spreadsheet files or ISA-JSON [20] – data is stored in more convenient fashion as JSON (JavaScript Object Notation [21]) files.

Although the technical approaches and the use case scenarios of the four data models differ, a unifying logic could be traced. The need of generally “larger”

chemical data object is not seen only in ECHA's REACH dossiers but also in all regulatory platforms (e.g. CEPA, TSCA, etc.) as well as such courses of action could be observed in public chemical databases evolution (e.g. PubChem has the notion of chemical substance). The foundation of a more sophisticated data model for substances is laid with three principal pieces of information: (1) identification, (2) material/substance description and composition, and (3) measurements records. This is practically illustrated in **Figure 4** for the substance of "benzene." The substance data model obviously includes a collection of standard triples:

$$\text{Structures} = \{(S_k, D_k, P_k) \mid k = 1, 2, \dots, m\},$$

However, a collection of new objects of the type "chemical substance" is needed to encompass the three principle levels. An identification layer may include identifiers and names. The challenge of unique substance identifiers and names is discussed in the last section of the book chapter.

A dynamic approach of material description is needed in at least two dimensions. Apart from multiple components, the industry and regulators, also, need to handle multiple compositions of the same substance, e.g. there may be different manufacturing processes for the same products. The latter is demonstrated with two different compositions of the "benzene" substance, as shown in **Figure 4**. The first one contains three components: benzene as main constituent, toluene as impurity, and some nonaromatic hydrocarbons. Toluene, which is an impurity in composition 1, is included in the second composition as well, but with a different role – it is a constituent of the "benzene" substance. The data model requires new data entities like:

$$\begin{aligned} \text{Substance} = \{ & \\ & \{\text{name}_1, \text{name}_2, \dots, \text{id}_1, \text{id}_2, \dots\}; \\ & \{\text{composition}_1, \text{composition}_2, \dots\}; \\ & \{\text{measurement}_1, \text{measurement}_2, \dots\}; \\ & \} \\ \text{and} \end{aligned}$$

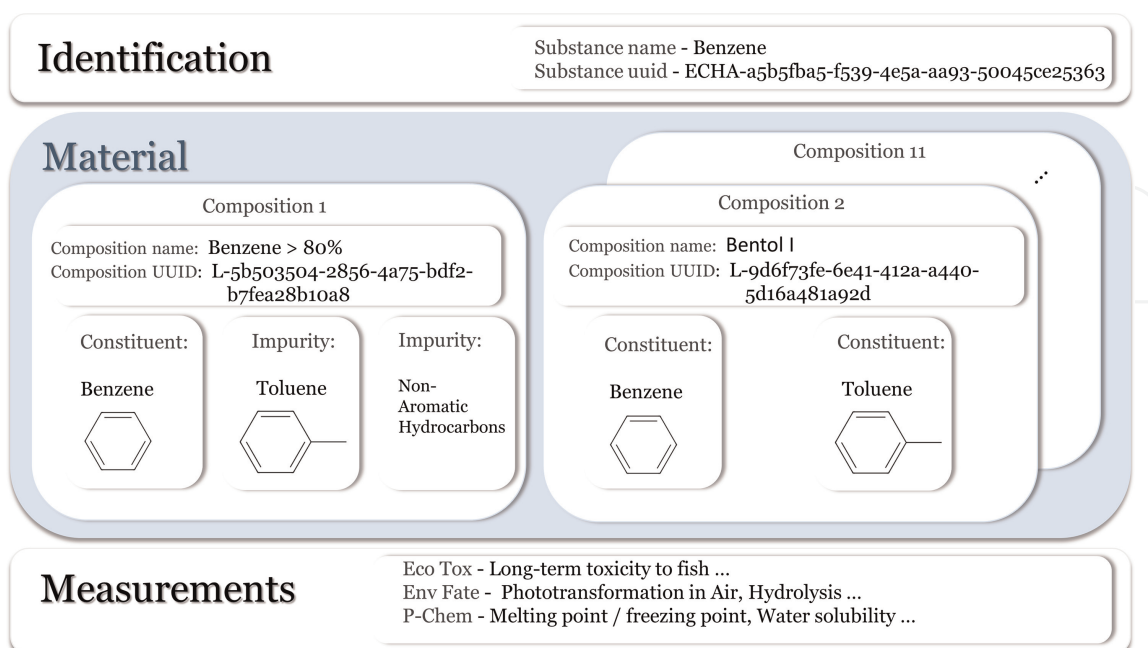


Figure 4. Substance of benzene with several different compositions and information grouped in three layers: Identification, material, and measurements (example is taken from the public records of the ECHA's dossiers and is also accessible via ambit-LRI database web interface).

Composition = {
 {(S₁, D₁, P₁), (S₂, D₂, P₂),...};
 {component relations};
 {component concentrations};
 }

In **Figure 4**, the term “benzene” is used for naming two different types of objects. There is a chemical structure of the benzene molecule which is the main constituent of the “benzene” substance. Hence, clear communication requires proper context in terms of data object types. On the other hand, the molecule of benzene could participate in other chemical substances with different roles. As it is illustrated in **Figure 5**, benzene molecule is an impurity component.

Also, the composition data entity should not be mistaken with the substance entity as well as the structure identifiers (e.g. benzene molecule CAS Number, 71–43–2 and InChI = 1S/C6H6/c1–2–4–6–5–3–1/h1–6H) should not be mistaken with the substance identifiers. The latter is a subtle error but is a common mismatch due to a long-term dominance of the structure-centered thinking. For example, in the nanoinformatics field, CAS number is wrongly associated with the whole nanomaterial instead of with a particular NM component. Also in **Figures 4** and **5**, identifiers of the substance compositions are shown as well. The complicated relationships between the three types of entities: structures, substances, and compositions require identifiers for all entity types. The shown examples utilize internal hash-based identifiers, uniquely

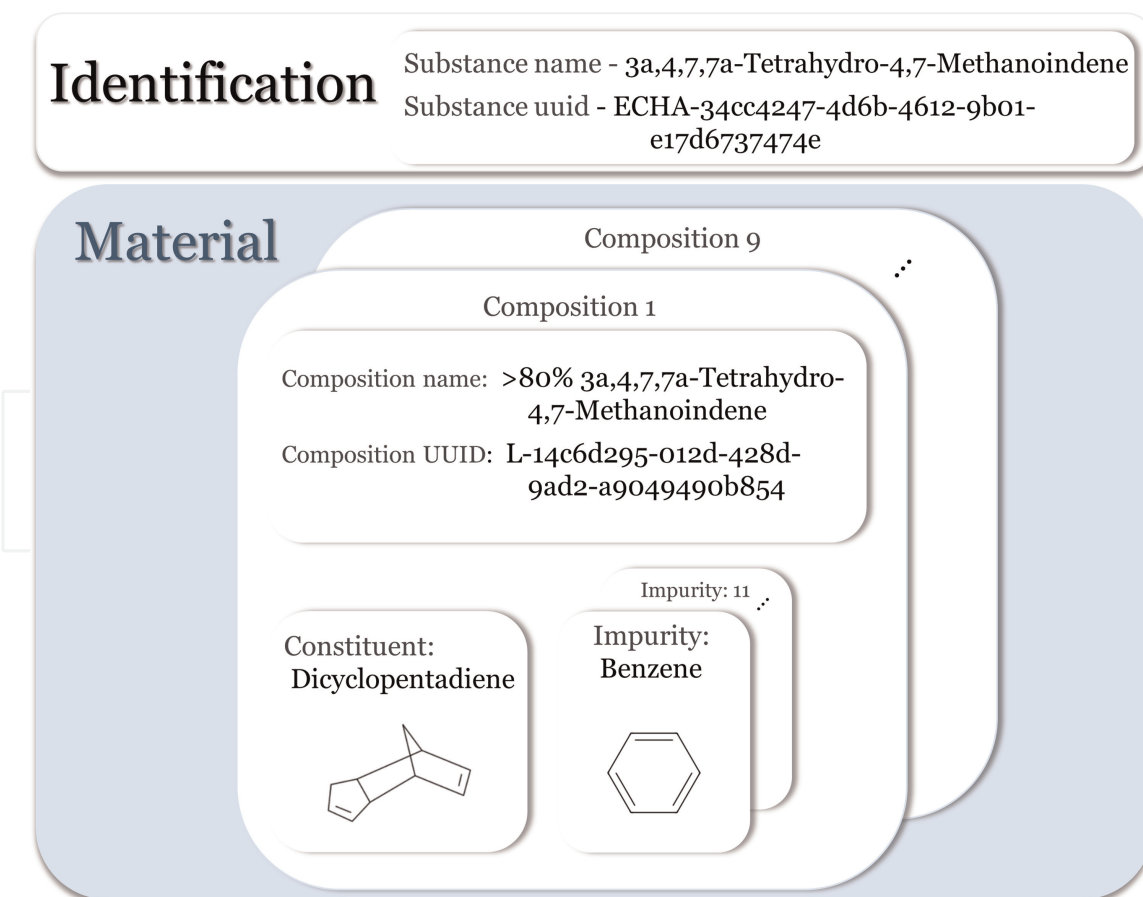


Figure 5. The substance of “3a,4,7,7a-tetrahydro-4,7-methanoindene” with a benzene molecule as an impurity component within the first composition (example is taken from the public records of the ECHA’s dossiers and is also accessible via ambit-LRI database web interface).

generated by the Ambit-LRI database system. The principle difference between structure and substance also dictates different approaches for database searching for these types of objects. Structure collections can be searched with the well-known cheminformatics methods (e.g. identity search, similarity, and substructure search), but the resulting hit list with structures should be logically related with the other types of data entities, i.e. substances and compositions. For instance, benzene structure is a component, playing different roles within about 200 different substances from the public ECHA's dossiers. Also, within the context of experimental data handling, there is a difference between the properties of the "entire" chemical substance, stored on the Measurements layer (see **Figure 4**) and the "nominal" properties of the component, as they are treated in the standard triad model.

4. FAIR principles

A chemical substance database is expected to facilitate analysis of chemical and physical properties, biological analyses, and human and environmental impacts, particularly in the context of safety and risk assessment. Integration of data from multiple sources (e.g. for the needs of read across) is only possible if original measurements are combined with rich metadata and obey a set of well-established good practices for data management. In 2016, Scientific data [3] published "FAIR Principles for Scientific Data Management." The authors provide guidance for improving the discoverability, accessibility, interoperability, and reuse of data popularized as FAIR (Findable, Accessible, Interoperable and Reusable). The principles (see **Figure 6**) emphasize machine capability (i.e. the ability of computing systems to find, access, interact with, and reuse data with no or minimal human intervention), since humans increasingly rely on computational support for data processing as a result of the increase of the volume, complexity, and speed of data creation.

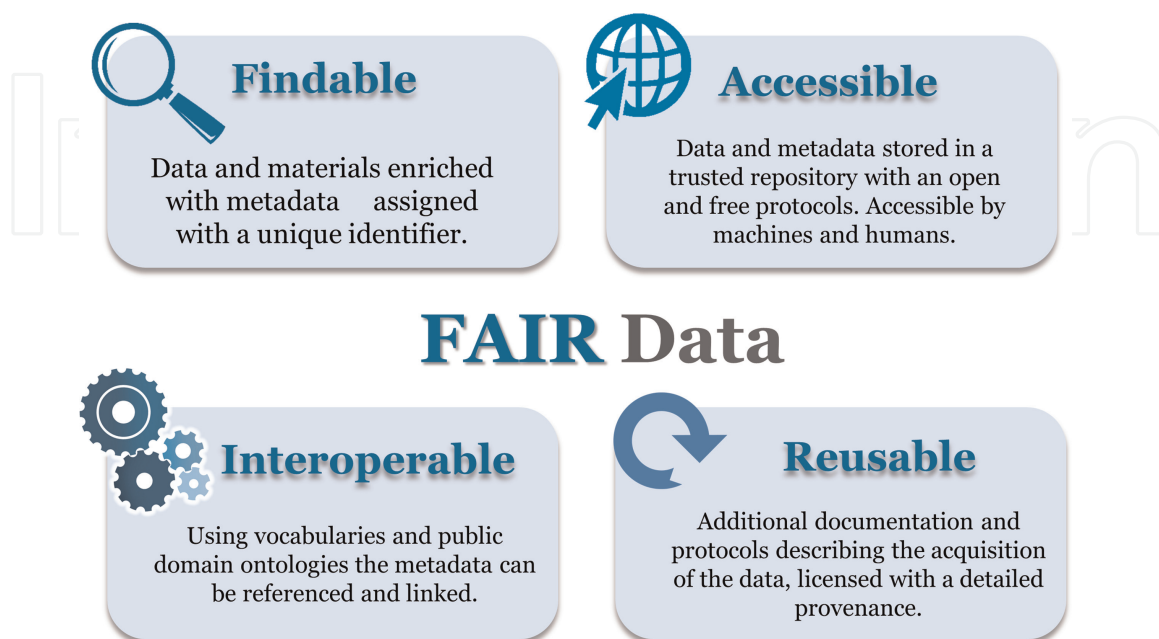


Figure 6. FAIR principles: findable, accessible, interoperable, and reusable.

The four foundational principles guide data producers and publishers in how to increase the value of modern scholarly digital publishing. The application of FAIR principles is, also, to algorithms, tools, and workflows used for data generation. GO-FAIR initiative (<https://www.go-fair.org/>) gained much popularity in the last few years and strongly endorses deployment of as much as possible FAIR data resources. On their dedicated site, GO-FAIR recommends a workflow of seven basic stages for transforming a non-FAIR data resource into a FAIR one (**Figure 7**). Rich and descriptive metadata, used by machines, is a key tool to evaluate and answer the questions being asked about the data. FAIR principles allow experimental data to be used beyond their origin to solve scientific problems, fill in missing data, reuse data in applications, do modeling, and provide tools for other scientific, industrial, and regulatory needs.

The step 3 of the FAIRification workflow is the most important one, namely definition of a semantic data model for chemical objects representation. In this sense, the efforts for substance data model elucidation are also efforts for FAIR data. The other pillars of a primary importance for the data FAIRness are inclusion of rich metadata (step 6), ontology annotations, and data linking with globally unique identifiers (step 4).

The FAIR principles are combined with so-called CARE (Collective benefit, Power to control, Responsibility, Ethics) principles [22]. CARE principles promote Indigenous Data Management to enhance machine functionality addressing concerns about rights and interests of the Indigenous people in their data throughout the data lifecycle (as a collective to have a say in how their data is actually used), trust, and accountability in the contexts of traditional knowledge and scientific data-oriented toward improving human well-being.

TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) is yet another set of principles [23] aligned with FAIR. To make data FAIR while preserving them over time requires trustworthy digital repositories (TDRs) with sustainable governance and organizational frameworks, reliable infrastructure, and comprehensive policies supporting community-agreed practices. TDRs may actively preserve data within dynamics of technology and stakeholder requirements.

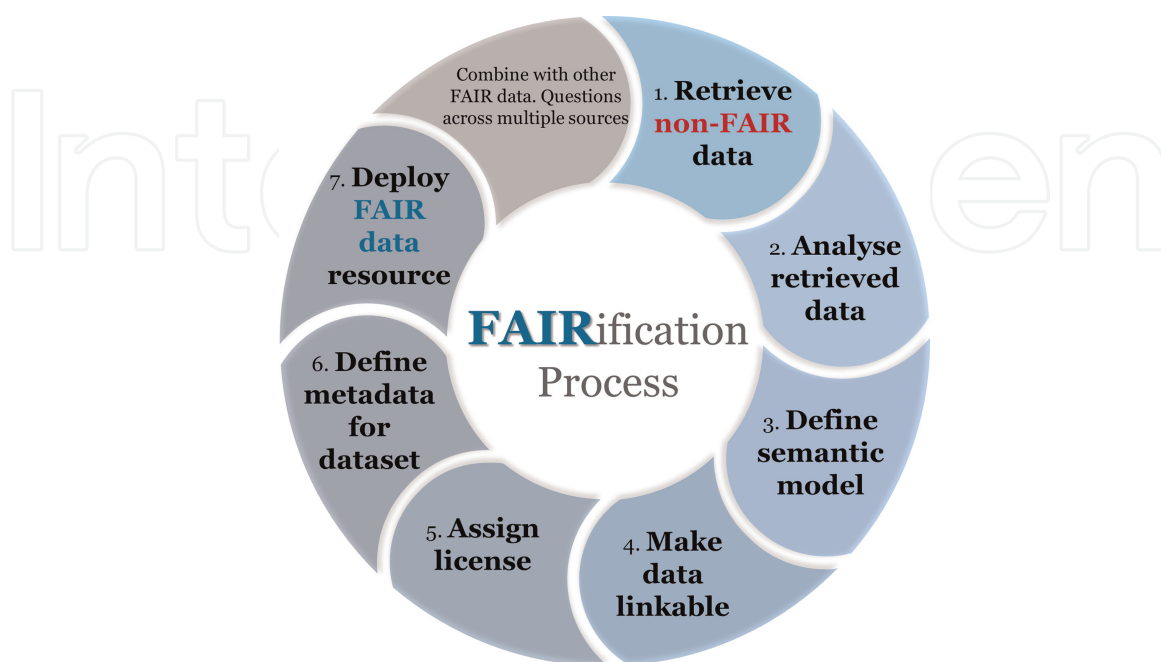


Figure 7. FAIRification process: a workflow of seven steps for transforming a non-FAIR data into a FAIR data resource.

The TRUST principles facilitate communication with all stakeholders, providing repositories and guidance to good practices.

5. Ambit/eNanoMapper data model

Safe by Design approaches are encouraged and promoted through regulatory initiatives and numerous scientific projects. Experimental FAIR data are the basis of risk assessment processing workflows. The Ambit/eNanoMapper [24, 25] database is an open-source chemical data management solution that currently holds the largest compilation of searchable nano-EHS (Environment, Health, and Safety) data in Europe from multiple completed and most of the ongoing H2020 Nano-EHS projects. Ambit is an open-source cheminformatics platform with over 30 modules implemented on the top of CDK [26, 27]. It is funded by CEFIC-LRI (<http://cefic-lri.org/>) for linking Ambit [28] system with the IUCLID substance database to support read across of substance data, category formation, REST APIs, web interface, substance and structure search facilities, toxicity prediction, and QSAR models. The eNanoMapper database is an extension of the Ambit cheminformatics platform.

The implementation of substance support in Ambit was inspired by the four data models discussed in previous section. The data model has been developed, tested, and improved for about 15 years, processing use cases and feedback from multiple users. The Ambit/eNanoMapper data schema is visualized in **Figure 8**. It contains a variety of data components (entities) serving different roles for the representation of items of information about substances and measurements. The data model entities may have different implementations at different stages of the data processing workflow:

- Serialization on input and output to the system (JSON, RDF or HDF5);
- Java classes in the server side of Ambit implementation;
- Relational database tables at the system back end;
- Python, R, Java, or JavaScript data structures within client libraries.

The data model is a conceptual representation of chemical substances and can be applied with different technologies, enabling interoperability and data linking, internally and externally via REST APIs.

The substances are characterized by their compositions and are identified by names and IDs. The model supports multiple compositions, with one or more components, each with a role assigned. Also, each component is treated via the standard triad approach. The results from physicochemical and biological measurements are treated as properties of the entire substance and are handled via the protocol applications. Efficient experimental protocol description is crucial for the correct communication of the scientific results and for creation of FAIR data resources. The latter is performed by means of a rich set of metadata parameters with a flexible logical organization (e.g. the full experimental data graph defined in ISA data model). The event of applying a test or experimental protocol to a chemical substance is described by a “protocol application” entity. Each protocol application consists of a set of “measurements” for a defined “endpoint” under given “conditions.” A measurement result can be a numeric, a string value, or a link to a raw data file (e.g. an IR spectrum,

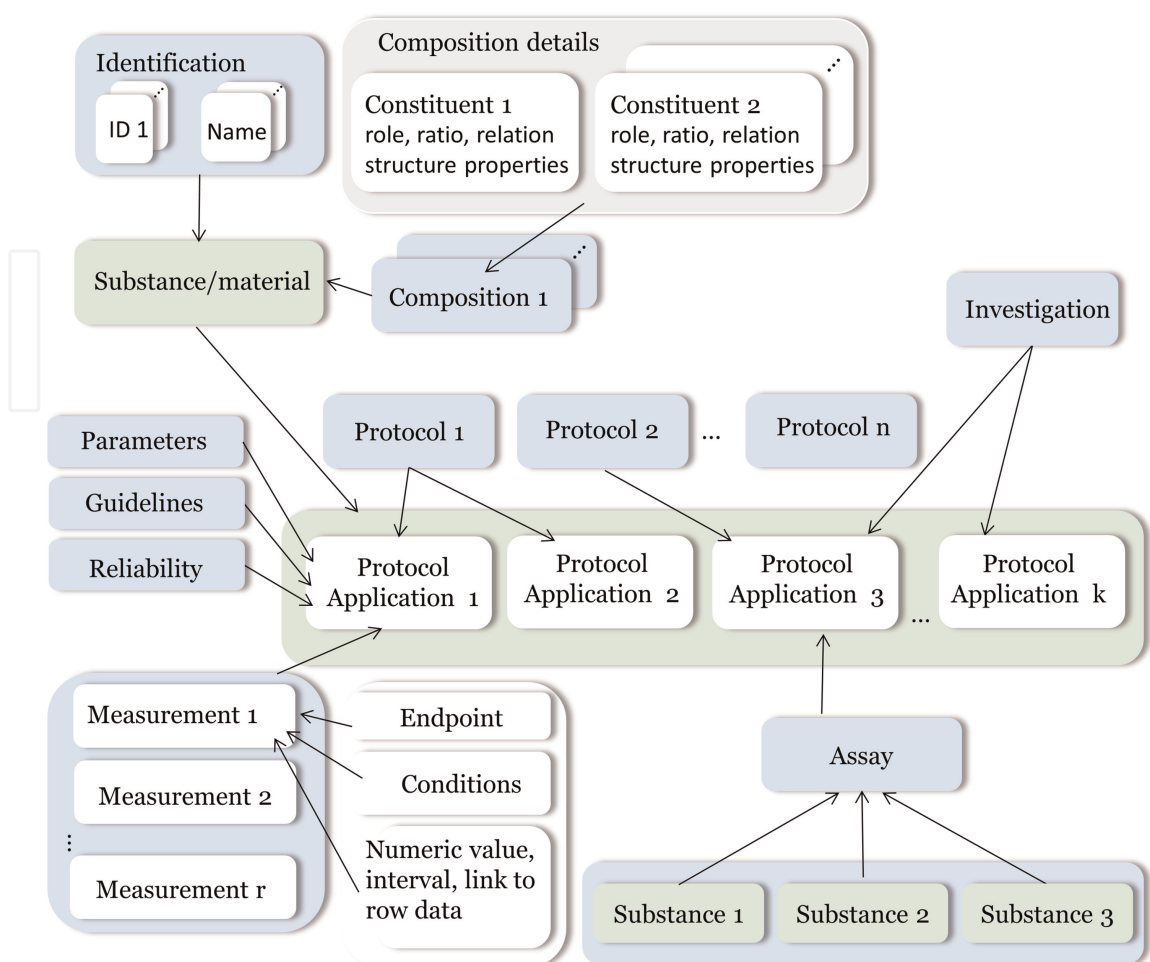


Figure 8.
 Schema of the ambit/eNanoMapper data model.

a microscopy image, HTS data, etc.). Measurement entity is also a dynamic data structure where a single number or an interval with lower and upper values together with specified qualifiers are supported. Ambit/eNanoMapper treats miscellaneous cases for a single datum storage, exemplified for the boiling point (BP) endpoint:

BP = 135°C, BP > 130°C, 120°C < BP ≤ 130°C, BP ~ 3°C, BP = 3 ± 0.5°C.

The measurement errors are represented via a separate qualifier, and different approaches for uncertainty are supported (e.g. SD – standard deviation, MAE – mean average error). The same flexibility is applied for storing metadata parameters. Each measurement is packed with a dynamic list of experimental conditions (or experiment factors such as concentration, time, etc.) which are considered as “lower” level metadata parameters. The “high” level metadata, namely the “protocol application”, is described by another dynamic list of parameters, links to Standard Operating Procedures (SOP), guidelines, publications, data quality, etc. The data for a particular substance may contain many “protocol applications.”

Figure 9 illustrates different levels of metadata: protocol parameters (Cell type = A549, Method = COMET, and Technical replicates = 3) and varied experimental conditions (Concentration and Exposure time). The same protocol “COMET” can be applied with different parameters (e.g. different cell line and replicates), and another protocol application will be obtained. The protocol applications that are related to one another are grouped to form an “Investigation” entity. Several different substances that have the same “protocol application” applied can be grouped via the

Substance	Protocol application data: Comet	Endpoint	Result	Concentration	Treatment	Exposure time
Substance name: BASO4 NM-220 Substance UUID: NRG2-2b94afbf-df44-3f76-ba5b-8973badd91b7 Public name: NM-220 Project: NanoReg2	Cell type: A549 Exposure time: 24 h, 3h Method: COMET SOP reference: link Input file: NR2_Scoring data_NILU_20210503.xlsx Number of technical replicates per conditions: 3	NET FPG SITES	3.65 % tail	0 ug/ml	sample	24 h
			12.35 % tail	0.16 ug/ml		
			11.59 % tail	0.48 ug/ml		
			12.30 % tail	1.6 ug/ml		
			13.21 % tail	4.8 ug/ml		
			11.99 % tail	16 ug/ml		
			13.75 % tail	48 ug/ml		
			6.94 % tail	120 ug/ml		3 h
			12.31 % tail	160 ug/ml		
			3.29 % tail	0 ug/ml		
			5.82 % tail	0.16 ug/ml		
			5.29 % tail	0.48 ug/ml		
			5.95 % tail	1.6 ug/ml		
			7.32 % tail	4.8 ug/ml		
			6.33 % tail	16 ug/ml		
			15.85 % tail	48 ug/ml		
			12.11 % tail	120 ug/ml		
13.48 % tail	160 ug/ml					

Figure 9. Protocol application data: COMET protocol with measurements of endpoint NET FPG SITES, applied for substance NM-220 from public database NanoReg2.

“Assay” entity. The higher level components of the model, such as Substance, Protocol Application, Investigation, and Assay, have automatically generated UUIDs which are used for linking and grouping the measurements.

A transition from the standard triples (S, D, P) to the extended substance data model is challenging for the experts from different domains due to various reasons. Typically the huge volume of metadata compared to the simple experimental data (the ratio of metadata volume to data volume reaches 10:1 or even higher) could be a stumbling block for experimentalists and cheminformatics experts, since a lot of effort is needed for the metadata generation and systematization. However, such a reluctance leads to non-FAIR data and poor findability, accessibility, interoperability, and reusability. Currently, the project funding institutions are challenged in the areas of project results sustainability and reusability as well as the issues of data curation of the results from past research projects and scientific publications.

6. Tools for data FAIRification

Huge volumes of already generated chemical substance data are non-FAIR. One of the predominant ways researchers store their scientific results is in the form of spreadsheets. We demonstrated that the FAIRification can be achieved through the multi-step FAIRification workflow (see **Figure 7**), using the semantic data model of Ambit/eNanoMapper. The analysis of data and metadata is an iterative process, requiring consultations with domain experts to explain the file content and layout, providing SOPs and correct ontology annotations. Generally, the original raw data needs to be converted to the substance data model. For this purpose, a dedicated software tool was developed, NMDataParser [29], to map the spreadsheets into the Ambit/eNanoMapper semantic model. The latter tool is essentially enabling the most important stage of the FAIRification process – mapping non-FAIR data (e.g. an Excel file) into an existing semantic model.

The NMDataParser is a configurable Excel file parser, developed in Java on top of the Ambit data model and with extensive use of the Apache POI library [30]. It was designed and developed as an open-source library to enable the import of substance data from the Excel spreadsheets with potentially unlimited layout permutations. Different row-based, column-based, or block-based spreadsheet data organizations

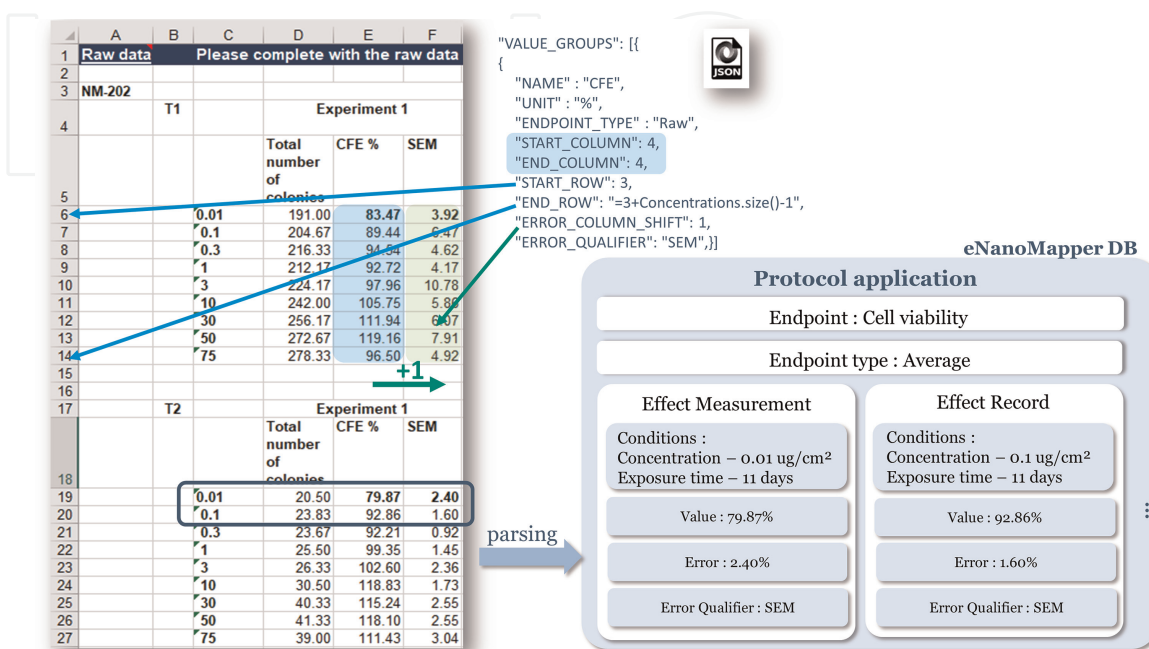


Figure 10. Parsing of an excel spreadsheet data for a CFE assay of measurements; part of the JSON configuration for relative addressing of the position of the error values is shown (top right); bottom right visualizes the data mapped within the ambit/eNanoMapper substance data model.

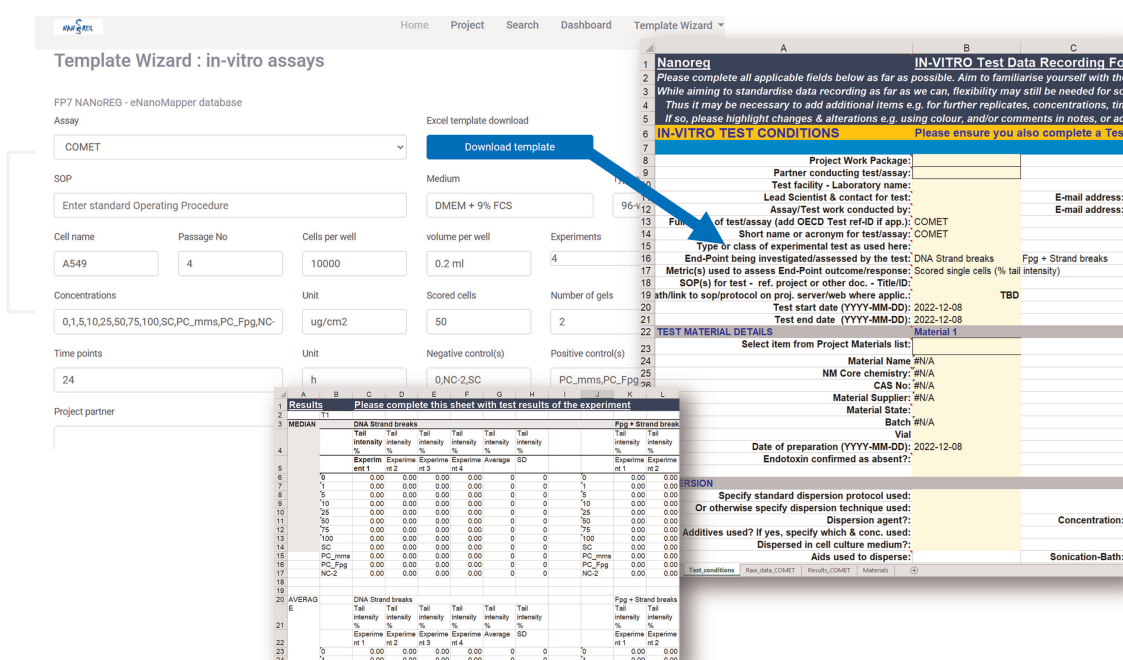


Figure 11. Web-based template wizard for automatic generation of standardized and harmonized templates with corresponding JSON configuration for NMDataParser.

are supported. The parser is configured via a separate JSON file with its own syntax for mapping the custom spreadsheet structure into the data model components (see **Figure 10**). The parser code, the JSON configuration syntax, documentation, and example files are available at <https://github.com/enanomapper/nmdataparser/>.

While one JSON configuration file can be applied to multiple Excel files with a similar layout, some complex spreadsheets (e.g. HTS) may require multiple JSON configurations for a single Excel file. The expertise, gained from many years of manual and exhausting configurations of excel file parsing, helped for developing a harmonized and continuously growing set of standard templates which are available via a web interface (see **Figure 11**) with an automatic template generation and corresponding JSON configuration attached.

7. Ambit/eNanoMapper applications, APIs, and services

Once the data is imported into an Ambit/eNanoMapper database instance, it is immediately available (publicly or with a restricted access) via the web user interface and machine readable via an API supporting multiple serialization formats. Non-public datasets are handled by an authentication and authorization system (API keys and OAuth2 plans for direct or delegated access grants are supported). Content from a variety of sources such as OECD HTs (IUCLID6 files or direct retrieval from IUCLID servers), custom spreadsheet templates, SQL dumps from other databases, and custom formats, provided by partners (e.g. the NanoWiki RDF dump [31]) is aggregated using the common semantic data model. A variety of options for export, data conversion, data retrieval, and data analysis are available (see **Figure 12**).

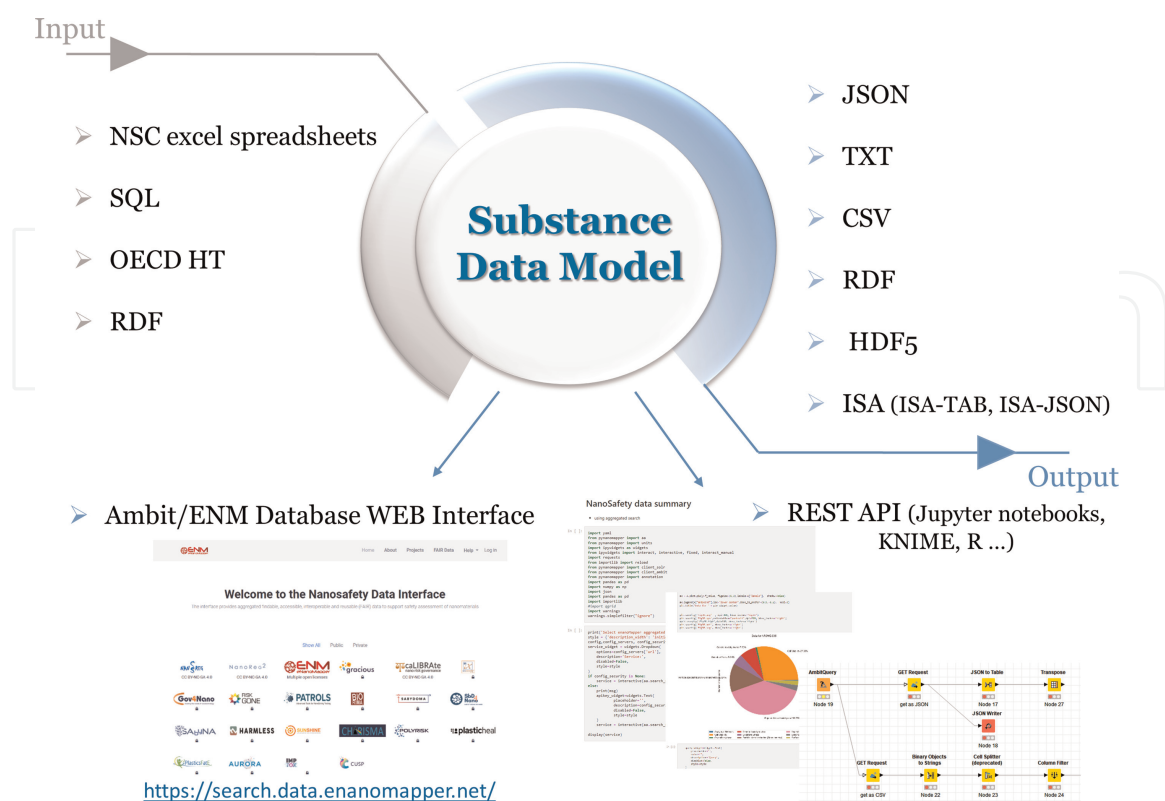


Figure 12.
Data input and output to ambit/eNanoMapper database.

Different views of substance data are implemented via a Web GUI based on the jToxKit [32] JavaScript library, as well as many customized methods for accessing the data through a REST API via external tools like Jupyter notebooks and the KNIME analytics platform.

Multiple data export formats are supported by the Ambit/eNanoMapper web interface and the API, including semantic formats (RDF and JSON-LD), Excel file formats [28], the native JSON serialization and HDF5 standard.

To facilitate data gap analysis and grouping, the aggregated search interface (see **Figure 13**) includes a number of options, allowing exporting all of the search hits into JSON or spreadsheet formats as well as flexible summary reports in Excel format.

All components of the Ambit/eNanoMapper data model (see **Figure 8**) are searchable. The model schema does not dictate a central entity unlike the standard triad, focused on the chemical structure. This way, more options for searching, storing, and viewing are available. **Figure 13** illustrates a faceted search. As it was pointed out, structure searching and substance searching are completely different features of the system. **Figure 14** shows an exact (identity) structure search for the benzene molecule and corresponding logical linking between the molecule of benzene and different chemical substances that contain it as a component.

A number of open-source libraries for accessing the eNanoMapper API are available: <https://github.com/enanomapper/rem> developed in R; <https://github.com/enanomapper/ambit.js> and <https://github.com/ideaconsult/jToxKit> developed in JavaScript; <https://github.com/ideaconsult/pynanomapper> developed in Python. Python library, in particular, is used for the set of open-source Jupyter notebooks that demonstrate the eNanoMapper API (<https://github.com/ideaconsult/notebooks-ambit/tree/master/enanomapper>).

The screenshot displays the H2020 NanoReg2 - eNanoMapper database search interface. The search criteria are 'A549' and 'Zeta potential'. The results list two entries:

- JRCNM02000a (NM-200 Synthetic Amorphous Silica PR-A-02) silicon dioxide nanoparticle**
 CONSTITUENT (1): SiO2, 7631-86-9, O=[Si]=O, VYPSYNLAIGMNEI-UHFFFAOYSA-N, InChI=1S/O2Si(Cl-3-2)
 Data completeness: 11% P-CHEM, 22% TOX, 43% ECOTOX
 P-CHEM: Particle size distribution (Granulometry) Zeta potential Bio-nano interaction
 TOX: Cell Viability Genetic toxicity in vitro Oxidative Stress
 ECOTOX: Short-term toxicity to aquatic invertebrates Short-term toxicity to fish Toxicity to aquatic algae and cyanobacteria
- JRCNM02001a (NM-201 Synthetic Amorphous Silica PR-B-01) silicon dioxide nanoparticle**
 CONSTITUENT (1): SiO2, 7631-86-9, O=[Si]=O, VYPSYNLAIGMNEI-UHFFFAOYSA-N, InChI=1S/O2Si(Cl-3-2)
 Data completeness: 11% P-CHEM, 22% TOX, 43% ECOTOX
 P-CHEM: Particle size distribution (Granulometry) Zeta potential Bio-nano interaction
 TOX: Cell Viability Genetic toxicity in vitro Oxidative Stress
 ECOTOX: Short-term toxicity to aquatic invertebrates Short-term toxicity to fish Toxicity to aquatic algae and cyanobacteria

Figure 13.

Faceted search for substances within NanoReg2 public database: Searching for NMs that have experiments with A549 cells and phys-chem characterization with zeta potential.

The screenshot shows the AMBIT web interface. At the top, there is a navigation menu with 'Search', 'Assessments', 'Import', 'Enhanced functions', 'Admin', and 'Help'. Below this is a search bar with the text 'Search structures and associated data'. The search criteria are set to 'Exact structure' and the search term is 'c1ccccc1'. The results are displayed in a table with columns: Diagram, CasRN, EC number, IUCLID 5, Names, Trade Name, IUPAC name, SMILES, Std. InChI key, and Std. InChI. The table shows four results, each with a 'Contained in as' column indicating the role of benzene in the substance.


Diagram	CasRN	EC number	IUCLID 5	Names	Trade Name	IUPAC name	SMILES	Std. InChI key	Std. InChI
									
- 1 -				Benzene					constituent
- 2 -				Distillates (petroleum), steam-cracked, C8-12 fraction	C9 resinfeed				constituent
- 3 -				acetone	Acetone				impurity
- 4 -				benzene	Benzene				constituent

Figure 14. Exact (identity) structure search in ambit/eNanoMapper database for the benzene structure; result structure is linked to a set of substances having the benzene mole as a component with different roles (rightmost column).

8. Linear notations and identifiers for chemical substances

Linear notations are representing chemical structure connectivity and other molecule features as a character string. Linear notations proved to be popular and efficient tools in the field of cheminformatics. The present-day mainstream notations, SMILES [33, 34] and InChI [35], are de facto standards and used in the majority of cheminformatics tools and structural databases. Naturally, linear notations played a significant role for establishing the classical triad model (S, D, P). Linear notation, InChI (International Chemical Identifier), as its name points it out, is originally designed to be a unique structure identifier. The methods for canonical atom numbering and canonical structural presentations are well known (e.g. canonical CTs and canonical SMILES) and together with hashing approaches (e.g. InChI-Key) are widely utilized for structure identification. Also, database and registry molecular numbers are another efficient means for molecule identification. The identification of the chemical substances is a huge challenge, especially in the field of nanoinformatics. Regulatory frameworks experience a lack of unique identifiers, since the traditional identifiers and the most popular linear notations are inadequate. One of the pillars for establishing the FAIR principles is utilization of globally unique and persistent identifiers (see points F1 and F3 from the FAIR principles [3]).

The chemical substance paradigm has been gradually adopted within the cheminformatics and nanoinformatics domains. The substances are serialized via data models with hierarchical organization (e.g. Ambit JSON or ISA model). With a proper canonicalization method, such data serialization (or parts of the data) could be hashed and used as a locally defined identifier, as it is the case with Ambit/eNanoMapper UUIDs (see **Figures 4** and **5**). The complexity of the substance data model justifies the

utilization of nonlinear techniques for serialization. However, lately, great effort has been put for developing a linear notation and universal identifiers for chemical substances and NMs.

The InChI Trust (<https://www.inchi-trust.org/>) works on developing and promoting the use of the IUPAC InChI [35] open-source chemical structure representation algorithm. InChI Trust projects cover versatile types of chemical objects and perform a pioneering work for developing lineation notations for mixtures (MInChI project), nanomaterials (NInChI – project), and Polymer InChI (PInChI – project) – to name a few of the most relevant projects to the chemical substances. Nano-InChI (NInChI) [36] project is a promising effort to integrate concepts of NMs intrinsic and extrinsic properties and to support a domain-specific language for nanoinformatics. NInChI is not intended to replace the chemical substance model but proposes to encode information (composition, size, shape, surface chemistry, etc.) required to unambiguously identify a specific NM as an extension of the IUPAC International Chemical Identifier, termed NInChI. NMs are particulates with specific relationships between the core and surface components that challenges traditional material naming and scientific data communication between researchers, modelers, industry, and regulators. Leveraging best practices with other InChI working groups, e.g. MInChI, Reaction InChI, and PInChI, is planned. NInChI development is a collaborative effort of domain experts from different fields. Currently, the NInChI is under active development, and there are some preliminary NInChI prototypes. For example: Fe₃O₄ core magnetic nanomaterial with diameter = 38 nm, coated with Glycine and shell thickness of 2 nm can be encoded as:

NInChI = 0.00.1A/C2H5NH2/C3-3-2(4)5/h1,3H2,(H,4,5)/msh/s2t-9!/3Fe.4O/msp/s38d-9/y2&1.

Another possible approach is utilization of SYBYL Line Notation (SLN) [37, 38]. SLN is unambiguous, nonunique linear notation developed by TRIPOS Inc. SLN supports syntax for specification of molecules, substructure queries, and reactions which cover the capabilities of SMILES [33], SMARTS [39], and SMIRKS [40] taken together. On top of the basic syntax, SLN includes other powerful features for the specification of user-defined attributes, macro and Markush [41] atoms for flexible definition of molecular fragments, search queries and structural libraries, as well as 2D and 3D coordinates. All that is accomplished through a unified syntax within a single notation. These features make SLN suitable for data storage and exchange. To our knowledge, SLN is the most comprehensive and rich linear notation for the representation of chemical objects of various kinds facilitating a wide range of cheminformatics algorithms. Though it is not the most popular linear notation nowadays, SLN has excellent capabilities for supporting the challenging tasks of present-day cheminformatics. SLN's rich syntax allows encoding of a comprehensive and versatile chemical information within the boundaries of a linear string representation otherwise manageable with complex data structures such as JSON [21] or XML [42] schemas.

Particularly, SLN is suitable for treating chemical objects with rich metadata (e.g. chemical substances). The SLN string defines one or more fully connected CTs plus a section with molecule attributes for each CT. One of the SLN advantages is the syntax extension, including comparison operations such as <, <=, >, and >=, while the SMILES/SMARTS standards support only attribute equality. The latter is in line with the substance model flexibility for storing experimental values. Within the existing notations from the past, SLN seems to have the most wide and flexible syntax features to support the chemical substance paradigm. A SLN example for the mentioned above Fe₃O₄ core magnetic NM, coated with Glycine:

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

9. Conclusions

The FAIR principles align with the global shift to open data by promoting governance criteria for increased data sharing. Cheminformatics, nanoinformatics, and bioinformatics methods are providing data-driven solutions in the field of chemical substance safety. The FAIR compliance calls for extension of the structure-centered data models to meet the challenges of chemical substance and materials data management. The substance must include not just a single structure, but a composition of many components with definite roles, corresponding interconnections, rich metadata, and ontology annotations. The variety of data sources, formats, and logical organizations challenges the aggregation of data from multiple projects into a common information system. Ambit/eNanoMapper data model has a well-defined semantics and full adoption of the FAIR principles in order to boost successful strategies for reusable and sustainable research results with efficient interconnections and collaboration between academia, industry, and regulators.

Acknowledgements

The work leading to this chapter has received funding from the European Union's Horizon 2020 Research and Innovation program, Grant Agreements no. 814426 NanoinformaTIX and LRI-EEM9.5 – IC AMBIT.

Author details


Nina Jeliaskova^{1*}, Nikolay Kochev^{1,2} and Gergana Tancheva²

1 Ideacconsult Ltd., Sofia, Bulgaria

2 Faculty of Chemistry, University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, Plovdiv, Bulgaria

*Address all correspondence to: jeliaskova.nina@gmail.com

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Gasteger J, Engel T, editors. Chemoinformatics Basic Concepts and Methods. Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA; 2018. p. 575
- [2] Massart D, Vandeginste BG, Kaufman L, Demin S, Michotte Y. Chemometrics: A Textbook. Elsevier Science (Verlag); 1988. p. 464. ISBN: 9780080868295
- [3] Wilkinson MD, Dumontier M, Ij A, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*. 2016;3:1-9. DOI: 10.1038/sdata.2016.18
- [4] McNaught AD, Blackwell AW. IUPAC. In: Compendium of Chemical Terminology Chemical Substance. 2014. 2nd ed. Available from: <https://goldbook.iupac.org/terms/view/C01039> . p. 2014. DOI: 10.1351/goldbook.C01039
- [5] ECHA (REACH). ECHA What is a substance? [Internet]. Available from: <https://echa.europa.eu/support/substance-identification/what-is-a-substance>. [Accessed: June 12, 2022]
- [6] Government of Canada, CEPA. Chemical Substances Glossary [Internet]. 1999. Available from: <https://www.canada.ca/en/health-canada/services/chemical-substances/chemical-substances-glossary.html>. [Accessed: June 12, 2022]
- [7] Epa A. TSCA Chemical Substance Inventory [Internet]. Available from: <https://www.epa.gov/tsca-inventory> [Accessed: June 12, 2022]
- [8] Japan CSCL. Japan CSCL – Chemical Substance Control Law [Internet]. Available from: <https://chemicalsubstancecontrol.jp/> [Accessed: June 12, 2022]
- [9] International Organization for Standardization. ISO/TS 80004-1:2015 - Nanotechnologies – Vocabulary – Part 1: Core-terms. ISO; 2015
- [10] The European Commission's Science and Knowledge Service [Internet]. Available from: https://joint-research-centre.ec.europa.eu/index_en [Accessed: June 12, 2022]
- [11] Chemicals European Agency in Association with the OECD. IUCLID 6 [Internet]. Available from: <https://iuclid6.echa.europa.eu/bg/project-iuclid-6>
- [12] OECD HT [Internet]. Available from: <https://www.oecd.org/ehs/templates/> [Accessed: June 12, 2022]
- [13] Abeyruwan S, Vempati UD, Küçük-McGinty H, Visser U, Koleti A, Mir A, et al. Evolving BioAssay ontology (BAO): Modularization, integration and applications. *Journal of Biomedical Semantics*. 2014; 5(Suppl. 1):1-22. DOI: 10.1186/2041-1480-5-S1-S5
- [14] Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schürer SC. BioAssay ontology (BAO): A semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*. 2011;12:257-273. DOI: 10.1186/1471-2105-12-257
- [15] Rumble J, Freiman S, Teague C. Towards a uniform description system for materials on the nanoscale. *Chemistry International* [Internet].

Available from: <https://www.degruyter.com/document/doi/10.1515/ci-2015-0402/html>. 2015;37(4):3-7.
DOI: 10.1515/ci-2015-0402

[16] Rumble J, Freiman S, Teague C. Uniform Description System for Materials on the Nanoscale Prepared by the CODATA-VAMAS Working Group On the Description of Nanomaterials. 2016. Available from: <https://zenodo.org/record/56720#.Y48ltMtBxD8>

[17] Assunta SS, Rocca-serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. National Public Grade. 2012;44(2): 121-126. DOI: 10.1038/ng.1054

[18] Robinson R, Cronin M, Richarz A, Rallo R. An ISA-TAB-Nano based data collection framework to support data-driven modelling of nanotoxicology. Beilstein Journal of Nanotechnology. 2015;6:1978-1999. DOI: 10.3762/bjnano.6.202

[19] Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-dantona E, et al. ISA-TAB-Nano: A specification for sharing nanomaterial research data in spreadsheet-based format. BMC Biotechnology. 2013;13:2-17. DOI: 10.1186/1472-6750-13-2

[20] ISA-JSON format [Internet]. Available from: <https://isa-tools.org/format/specification.html> [Accessed: June 12, 2022]

[21] ECMA. JSON (ECMA-404 The JSON Data Interchange Syntax). [Internet]. Geneva, Switzerland: ECMA International. Available from: <https://www.ecma-international.org/publications-and-standards/standards/ecma-404/2017> [Accessed: June 12, 2022]

[22] Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR principles for

indigenous data futures. Scientific Data [Internet]. 2021;8(1):8-13. DOI: 10.1038/s41597-021-00892-0

[23] Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST principles for digital repositories. Scientific Data. 2020;7(1):1-5. DOI: 10.1038/s41597-020-0486-7

[24] Jeliaskova N, Apostolova MD, Andreoli C, Barone F, Barrick A, Battistelli C, et al. Towards FAIR nanosafety data. Nature Nanotechnology. 2021;16(6):644-654. DOI: 10.1038/s41565-021-00911-6

[25] Jeliaskova N, Chomenidis C, Doganis P, Fadeel B, Grafström R, Hardy B, et al. The eNanoMapper database for nanomaterial safety information. Beilstein Journal of Nanotechnology. 2015;6:1609-1634. DOI: 10.3762/bjnano.6.165

[26] Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliaskova N, et al. The chemistry development kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics. 2017;9(1):1-19. DOI: 10.1186/s13321-017-0220-4

[27] Chemistry Development Kit [Internet]. Available from: <https://cdk.github.io/> [Accessed: June 12, 2022]

[28] Jeliaskova N, Koch V, Li Q, Jensch U, Reigl JS, Kreiling R, et al. Linking LRI AMBIT chemoinformatic system with the IUCLID substance database to support read-across of substance endpoint data and category formation. Toxicology Letters. 2016;258: S114-S115. DOI: 10.1016/j.toxlet.2016.06.1469

[29] Kochev N, Jeliaskova N, Paskaleva V, Tancheva G, Iliev L,

- Ritchie P, et al. Your spreadsheets can be fair: A tool and fairification workflow for the enanmapper database. *Nanomaterials*. 2020;**10**(10):1-23. DOI: 10.3390/nano10101908
- [30] Apache POI [Internet]. Available from: <https://poi.apache.org/> [Accessed: June 12, 2022]
- [31] NanoWiki RDF [Internet]. Available from: https://figshare.com/articles/NanoWiki_4/4141593_2016 [Accessed: June 12, 2022]
- [32] JToxKit [Internet]. Available from: <https://github.com/ideaconsult/jToxKit> [Accessed: June 12, 2022]
- [33] SMILES - A Simplified Chemical Language [internet]. Daylight Theory. Available from: <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html> [Accessed: June 12, 2022]
- [34] Weininger D, Weininger A, Weininger J. SMILES . 2 . Algorithm for generation of unique SMILES notation. *Chemical Information and Computer Science*. 1989;**29**(19):97-101. DOI: 10.1021/ci00062a008
- [35] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier [Internet]. *Journal of Cheminformatics*. 2015;**7**:1-34. DOI: 10.1186/s13321-015-0068-4
- [36] Lynch I, Afantitis A, Exner T, Himly M, Lobaskin V, Doganis P, et al. Can an inchi for nano address the need for a simplified representation of complex nanomaterials across experimental and nanoinformatics studies? *Nanomaterials*. 2020;**10**(12): 1-44. DOI: 10.3390/nano10122493
- [37] Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL line notation (SLN): A versatile language for chemical structure representation. *Journal of Chemical Information and Computer Sciences*. 1997;**37**(1):71-79. DOI: 10.1021/ci960109j
- [38] 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. *Journal of Chemical Information and Modeling*. 2008;**48**(12): 2294-2307. DOI: 10.1021/ci7004687
- [39] SMARTS - A Language for Describing Molecular Patterns [Internet]. Daylight Theory. Available from: <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> [Accessed: June 12, 2022]
- [40] SMIRKS - A Reaction Transform Language [Internet]. Daylight Theory. Available from: <https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html> [Accessed: June 12, 2022]
- [41] Barnard J, Wright PM. Towards in-house searching of Markush structures from patents. *World Patent Information*. 2009;**31**(2):97-103. DOI: 10.1016/j.wpi.2008.09.012
- [42] Extensible Markup Language (XML) 1.0 (Fifth Edition) [Internet]. 2008. Available from: <https://www.w3.org/TR/REC-xml/> [Accessed: June 12, 2022]