Semantic Representation of Preclinical Data in Radiation Oncology

Olga Giraldo[1*], Abumansur Sabyrrakhim[1], Mareike Roscher[2], Rosemarie Euler-Lange[2], Michael Baumann[1,3,4], Ina Kurth[1,2,3,4] and Wahyu Wijaya Hadiwikarta[1,3]

[1] *Division of Radiooncology/Radiobiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany*

[2] *Service Unit for Radiopharmaceuticals and Preclinical Studies, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany*

[3] *German Cancer Consortium (DKTK), Core Center Heidelberg, 69120 Heidelberg, Germany*

[4] *Heidelberg Institute of Radiation Oncology (HIRO), 69120 Heidelberg, Germany*

Abstract

Background: In radiation oncology, the data generated from preclinical trials serve as initial validation for treatment effectiveness and optimizing clinical approaches by unraveling molecular mechanisms underlying different treatment responses. Therefore, it is important to standardize the practice in managing preclinical trial data to ensure consistency and reproducibility across studies, promoting collaboration, and facilitating regulatory review. The primary goal of this work is to standardize the representation of data collected from preclinical radiobiology and radiation oncology studies as a way to facilitate knowledge discovery. To achieve this goal, we combined ontology with semantic Web techniques to publish mapped data and easily query them using SPARQL Protocol and RDF Query Language (SPARQL). **Results**: We expanded the Radiation Oncology Ontology (ROO) to include terminology related to the exposure of animal models to treatment, animal model's demographic characteristics; as well as clinical information in live animals. The extended ROO contains 123 new entities (89 classes, 29 data properties and 5 object properties). We combined the extended ontology with Semantic Web technologies to demonstrate how to integrate and query data from different relational databases. **Discussion**: The use of ontologies and semantic web tools are a way to comply to the FAIR principles. FAIR preclinical data improve collaboration, transparency, and reproducibility in radiotherapy research.

1. Introduction

In radiation oncology, preclinical trials are conducted in animals prior to clinical trials to evaluate the safety and efficacy of radiation therapy effects, taking into account various aspects such as new radiation treatment techniques, radiation delivery methods, and novel therapeutic agents. The data generated from preclinical trials are very important, because they serve as initial validation for treatment effectiveness. Furthermore, considering the ethical and economical aspects of performing animal studies, preclinical data are highly valuable.

15th International Conference on Biological and Biomedical Ontology, July 17-19 2024, Enschede, The Netherlands

[*] Corresponding author.

olga.giraldo@dkfz-heidelberg.de (O. Giraldo); sabyrrakhim.amd@gmail.com (A. Sabyrrakhim); mareike.roscher@dkfz-heidelberg.de (M. Roscher); r.lange@dkfz-heidelberg.de (R. Euler-Lange); michael.baumann@dkfz-heidelberg.de (M. Baumann); ina.kurth@dkfz-heidelberg.de (I. Kurth); w.hadiwikarta@dkfz-hiedelberg.de (W. Hadiwikarta)

0000-0003-2978-8922 (O. Giraldo); 0000-0002-9340-974X (M. Baumann); 0000-0001-9261-5165 (I. Kurth); 0000-0002-5909-4107 (W. Hadiwikarta)

Therefore, it is important to standardize the practice in managing preclinical trial data to ensure consistency and reproducibility across studies, which is critical for advancing scientific knowledge, promoting collaboration, and facilitating regulatory review.

Some efforts focused on standardization of preclinical data exist in the field. An example is the Standard for Exchange of Nonclinical Data (SEND) from the Clinical Data Interchange Standards Consortium (CDISC), aimed at standardizing collected individual animal data in tabular data structures according to different nonclinical domains e.g., animal demographics, laboratory test results, treatment procedures, etc.[1]. Conversely, ontologies such as the Dependency Layered Ontology for Radiation Oncology (DLORO) [2], the Radiation Oncology Structures Ontology (ROS) [3], and the Radiation Oncology Ontology (ROO) [4], were developed for use cases in the clinical radiation oncology domain. Unfortunately, these aforementioned ontologies were designed to support human clinical trials and their relationships, therefore, unfortunately lack the required representation of preclinical data.

The primary goal of this work is to standardize the representation of data collected from preclinical radiobiology and radiation oncology studies as a way to facilitate knowledge discovery. Formalized preclinical data will serve as a critical basis for the conduct and interpretation of clinical trial data stored in the database system RadPlanBio [5,7]. The data are stored according to the CDISC SEND standard. To achieve this goal, we propose populating the ROO with terminologies related to the exposure of animal models to treatment, animal model's demographic characteristics; as well as clinical information in live animals. We decided to reuse and extend the ROO because this ontology contains classes that cover the most common concepts in radiation oncology, including oncological diseases, cancer staging systems, and oncological treatments. To reuse and extend an existing ontology is in principle aligned to the open world assumption (OWA).

In this article we: i) present the materials and methods used to populate the ROO with preclinical terminologies; ii) describe the validation and evaluation process of the ontology, iii) show the current state of the extended ontology, and iv) conclude with the discussion and outlook for future work.

2. Materials and Methods

2.1. Preclinical database

As a use case, we analyze information from databases collected in a study focused on investigating the effect of nimorazole combination treatments on hypoxic tumor areas in mice [6]. The databases are available in the German Cancer Consortium (DKTK) RadPlanBio platform, a web-based platform which supports the collection and the exchange of radiotherapy research data in clinical and preclinical studies [7].

The databases analyzed in this work include >2000 mice and contain information on: i) demographic characteristics of each individual animal; ii) details of an animal's exposure to treatment; iii) body weights of animals during the study and at the end of the study; iv) diagnosis of the cause of death of animals, and; v) laboratory test data per animal. The information in these databases is organized and structured according to the format suggested by the SEND standard for tabulation of nonclinical datasets.

Due to the heterogeneous nature of the data, they provide a good validation for the extended ROO. The extended ROO was applied to represent each value in the database and to map them through the concepts available in the ontology.

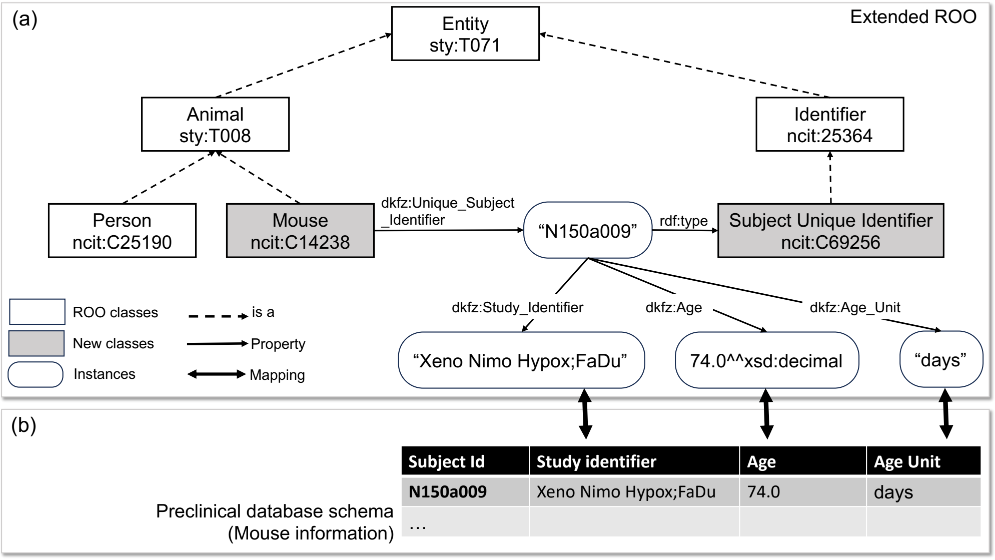
2.2. The ROO extension process

The process of enriching and extending the Radiation Oncology Ontology (ROO) with preclinical concepts consists of three steps: i) collection of preclinical concepts; ii) semantic analysis of existing vocabularies, and; iii) ontology extension.

In the first step, terms were collected from the preclinical databases mentioned in section 2.1. The next step was to identify reusable terminologies from other ontologies. BioPortal [8] and the UMLS Metathesaurus Browser [9] were used throughout this stage to find references and definitions for each terminology. In the last step, the ROO was extended with terminologies that come primarily from the National Cancer Institute Thesaurus (NCIt) [10]. New terms that are not coming from existing ontologies use the prefix 'roo' and a local ID that starts with the letters DKFZ followed by 6-digit numbers; as an example, the identifier for the class 'Animal Identifier' is roo:DKFZ000006. Protégé v. 5.6.1 [11] was used to create new concepts and manage the ontology.

2.3. Ontology validation and evaluation

The ontology validation procedure ensures that the ontology can effectively represent and capture the knowledge and data from the preclinical relational databases. This validation process involves mapping the elements (rows, columns, and values) of the database to the concepts and properties (predicates) in the ontology. Figure 1 shows a correspondence between the columns in the relational database and the ontology entities. At the top (rectangle a) the hierarchical structure of the extended ROO is illustrated. White boxes represent existing concepts from original ROO. Grey boxes represent new concepts proposed in this work.



Subject Id	Study identifier	Age	Age Unit
N150a009	Xeno Nimo Hypox;FaDu	74.0	days
…			

Figure 1. Overview of the extended ROO structure and the relational database. The hierarchical structure of extended ROO is presented in (a). The mapping performed to columns and values in a database is presented in (b).

The added concept “Mouse (ncit:C14238)” is a subclass of “Animal (sty:T008)” in description logic syntax, it can be expressed as $\mathtt{Mouse} \sqsubseteq \mathtt{Animal}$. In addition, mouse or person are animals ($\mathtt{Mouse} \sqcup \mathtt{Person} \sqsubseteq \mathtt{Animal}$) and mouse is not a person ($\text{Mouse} \sqsubseteq \neg\, \text{Person}$).

Boxes with rounded corners represent instances or individuals. Hierarchical relationships (“is subclass of”) between classes, are represented by dotted arrows. Properties are represented with arrows; they connect classes or instances between each other.

At the bottom (rectangle b) demographic information about the mouse (e.g., age, age unit) and the study identifier of which the animal was enrolled are presented as examples. In the extended ROO, the column “Subject Id” is mapped to the concept “Subject Unique Identifier (ncit:C69256)”. The link between a mouse and the subject identifier is the property “Unique Subject Identifier (DKFZ000009)”. In description logic syntax, any mouse that has a unique subject identifier can be expressed as $\mathtt{Mouse} \sqcap \exists\, \mathtt{Unique\_Subject\_Identifier}.\top$.

Several languages and software tools are available to perform the mapping procedure from relational databases to RDF triples [12]. We use RDF Mapping Language (RML), an extension of R2RML to map columns and rows of preclinical databases and our ontology. R2RML is a W3C standard for mapping relational databases to RDF. RML follows exactly the same syntax as R2RML; therefore, RML mappings are themselves RDF graphs [13]. The stages we implemented to generate linked data between the extended ROO and our preclinical relational databases are illustrated in Figure 2 and explained below.

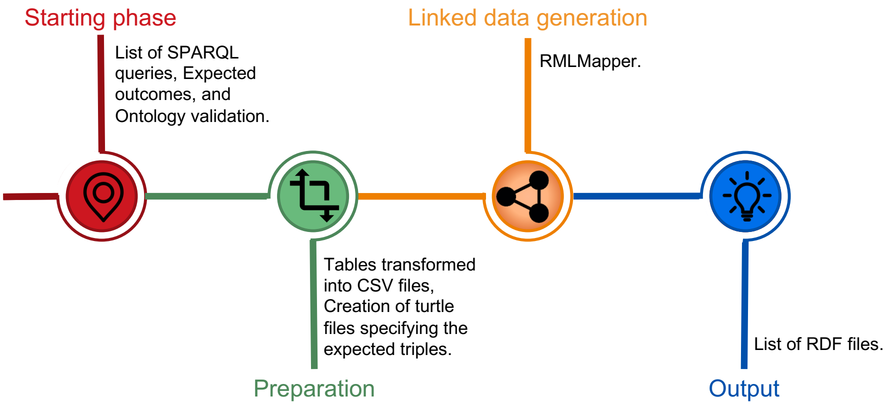

Starting phase
List of SPARQL queries, Expected outcomes, and Ontology validation.
Linked data generation
RMLMapper.
Tables transformed into CSV files, Creation of turtle files specifying the expected triples.
Preparation
List of RDF files.
Output

Figure 2. Linked data generation process.

2.3.1. Starting phase

In the first stage, we gathered a set of SPARQL queries and the corresponding expected outcomes (triples and query result). We focused on the functional aspects that we wanted the ontology to represent. The queries we gathered include, “*retrieve the Subject Unique Identifier of the animals tested*”.

2.3.2. Preparation

As preparation, we exported the analyzed preclinical databases to CSV formats. Then, we created turtle files specifying expected triples. Some of the expected triples we specified include, <N150a009> **rdf:type** <Subject Unique Identifier (ncit:C69256)>; <N150a009> **study identifier (roo:DKFZ000008)** <Xeno Nimo Hypox;FaDu>.

2.3.3. Linked data generation

To generate linked data, we use the RMLMapper [14] which executes RML rules to achieve its task. We used Docker [15] to run RMLMapper and storing data.

2.3.4. Output

An ontology validation process was considered done with valid result if the generated outcomes are not different from the expected outcomes. Therefore, in this stage the ontology evaluation is done and we compare the generated triples with the expected triples specified in the preparation step.

3. Results

3.1. Extended Radiation Oncology Ontology (ROO)

The extended ROO contains 123 new entities (89 classes, 29 data properties and 5 object properties). The new terminologies represent: i) attributes that are common across the used databases e.g., "subject unique identifier (ncit:C69256)", study identifier (roo:DKFZ000008); ii) demographic characteristics e.g., "strain (roo:DKFZ000041)"; iii) findings or information collected during a study e.g., "body weight (roo:DKFZ000017)", "cause of death (roo:DKFZ000021)" and "clinical observation (roo:DKFZ000036)"; iv) exposure information e.g., "treatment name (roo:DKFZ000013)", "route of administration (roo:DKFZ000011)" and "treatment vehicle (roo:DKFZ00012)". We followed the design principles from ROO. The extended ROO is saved as OWL and available on GitHub [16].

3.2. Ontology validation and evaluation

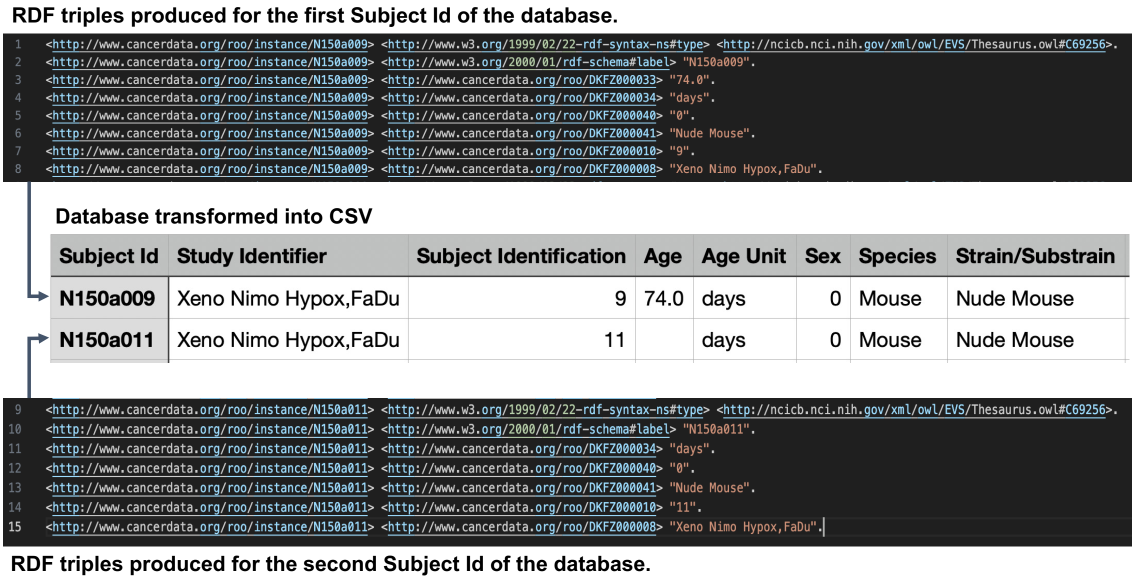
3.2.1. SPARQL queries

The ontology represents and captures the knowledge and data from the preclinical relational databases. The expected SPARQL queries were executed by using a Protégé desktop plug-in that provides support for writing and executing SPARQL queries. All the queries returned the expected results. The complete list of the queries is available on GitHub [17].

3.2.2. Linked data

Based on Linked Data principles, an ontology enables semantic interoperability across preclinical data available in relational databases. Our ontology facilitates data sharing and transparent access to data. Figure 3 shows a database transformed into CSV format and the RDF triples produced for the first and second subject Ids of the transformed database. The database presents demographic information of two mice. The first mouse has the unique subject identifier "N150a009". The second mouse has the unique subject identifier "N150a011". Both were registered in the same study "Xeno Nimo Hypox.FaDu". Each mouse was given an identifier used within the study; "9" is the study identifier for the first mouse and "11" for the second mouse. The age is available for the first mouse "74.0"; the age unit is "days". Both mice are male (represented as "0") and belong to the strain/substrain "Nude Mouse". As seen in Figure 3, the RDF triples obtained after running RMLMapper capture the data described above.

CSV tables reflecting the content of the databases, the Turtle files specifying expected triples, and the RDF files generated by running RMLMapper are available on GitHub [18].



RDF triples produced for the first Subject Id of the database.

```
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C69256>.
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.w3.org/2000/01/rdf-schema#label> "N150a009".
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.cancerdata.org/roo/DKFZ000033> "74.0".
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.cancerdata.org/roo/DKFZ000034> "days".
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.cancerdata.org/roo/DKFZ000040> "0".
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.cancerdata.org/roo/DKFZ000041> "Nude Mouse".
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.cancerdata.org/roo/DKFZ000010> "9".
<http://www.cancerdata.org/roo/instance/N150a009> <http://www.cancerdata.org/roo/DKFZ000008> "Xeno Nimo Hypox,FaDu".
```

Database transformed into CSV

Subject Id	Study Identifier	Subject Identification	Age	Age Unit	Sex	Species	Strain/Substrain
N150a009	Xeno Nimo Hypox,FaDu	9	74.0	days	0	Mouse	Nude Mouse
N150a011	Xeno Nimo Hypox,FaDu	11		days	0	Mouse	Nude Mouse

```
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C69256>.
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.w3.org/2000/01/rdf-schema#label> "N150a011".
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.cancerdata.org/roo/DKFZ000034> "days".
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.cancerdata.org/roo/DKFZ000040> "0".
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.cancerdata.org/roo/DKFZ000041> "Nude Mouse".
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.cancerdata.org/roo/DKFZ000010> "11".
<http://www.cancerdata.org/roo/instance/N150a011> <http://www.cancerdata.org/roo/DKFZ000008> "Xeno Nimo Hypox,FaDu".
```

RDF triples produced for the second Subject Id of the database.

Figure 3. RDF triples capturing demographic characteristics from mice.

4. Discussion and future work

Semantic representation of preclinical data in radiobiology and radiation oncology involves structuring and encoding information about demographic characteristics of animals, findings and treatments in a machine-readable format that facilitates data integration, analysis, and interpretation of outcomes, such as, overall survival or toxicities after treatment.

To achieve this goal, we have expanded the ROO to describe preclinical data [16]. This ensures semantic interoperability and enabling integration with other datasets and knowledge resources. Our extended ontology supports publishing preclinical data as linked data using RDF to enable integration and interoperability with other datasets. The use of ontologies and semantic web tools are a way of adhering to the FAIR principles [19]. FAIR preclinical data enhances collaborations, transparency, and reproducibility in preclinical research.

In this work, we were able to map all the entities present in the analysed databases with concepts and properties from the extended ROO. Nevertheless, it is not without limitations. The extended ontology should be validated against other preclinical data to ensure robustness. Additionally, improving the ontology extension strategy is crucial, e.g., by utilizing owl:imports. Currently, the extension was performed manually, while preserving the existing ROO entities to maintain the organizational structure of preclinical terminology derived from our relational databases. Improving the extension strategy will address issues such as the lack of unique URIs for the preclinical entities from the analyzed RDBs, and enabling to index the ontology on a repository such as BioPortal. Further step includes testing our ontology against competency questions that retrieve information from two or more databases and establish interconnections. For example, "*survival of mice when are exposed to a particular treatment such as cisplatin*". Then will be to integrate the ontology to the semantic layer of the RadPlanBio platform, through a knowledge graph to allow semantic querying, reasoning, and inference. The final step will be to develop a plan to maintain the ontology over time. This plan will involves addressing issues such as ontology evolution, version control, and alignment with evolving domain knowledge.

Acknowledgements

The first author of this paper has received funding from the European Federation for Cancer Images (EUCAIM), Project ID: 101100633. The authors acknowledge the DKTK funding for the operation of RadPlanBio platform in DKFZ. We also acknowledge Dr. Freddy Priyatna for his support in the creation of mapping files and in the transformation/generation of triples. The authors also wish to thank to Thomas Früchtel, Myta Pristanty and Betül Çakir, for the digitization and standardization of the preclinical data, and the reviewers of this article for their valuable comments.

References

[1] CDISC SEND Standard. URL: https://www.cdisc.org/standards/foundational/send

[2] Kalet AM, Doctor JN, Gennari JH, *et al.* Developing Bayesian networks from a dependency-layered ontology: A proof-of-concept in radiation oncology. Med Phys. 2017. doi: 10.1002/mp.12340.

[3] Bibault JE, Zapletal E, Rance B, *et al.* Labeling for Big Data in radiation oncology: The Radiation Oncology Structures ontology. PLoS One. 2018. doi: 10.1371/journal.pone.0191263.

[4] Traverso A, van Soest J, Wee L, *et al.* The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018. doi: 10.1002/mp.12879.

[5] RadPlanBio. URL: https://helmholtz.software/software/radplanbio

[6] Koi, L., Bitto, V., Weise, C. *et al.* Prognostic biomarkers for the response to the radiosensitizer nimorazole combined with RCTx: a pre-clinical trial in HNSCC xenografts. *J Transl Med* 21, 576 (2023). https://doi.org/10.1186/s12967-023-04439-2

[7] T. Skripcak *et al.*, "Toward Distributed Conduction of Large-Scale Studies in Radiation Therapy and Oncology: Open-Source System Integration Approach," in *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1397-1403, Sept. 2016, doi: 10.1109/JBHI.2015.2450833.

[8] Whetzel PL, Noy NF, Shah NH, *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011.

[9] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004. doi: 10.1093/nar/gkh061.

[10] NCI Thesaurus (NCIt). URL: https://ncithesaurus.nci.nih.gov/ncitbrowser/

[11] Protégé. URL: https://protege.stanford.edu/

[12] Hert M. *et al.* A comparison of RDB-to-RDF mapping languages, in (ACM Press; 2011).

[13] RDF Mapping Language (RML). URL: https://rml.io/specs/rml/

[14] RMLMapper. URL: https://github.com/RMLio/rmlmapper-java

[15] Docker. URL: https://www.docker.com

[16] ROOext.owl. URL: https://github.com/DKFZ-E220/ROOx/blob/main/ROOext_v0.1.owl

[17] SPARQL queries from ROOx. URL: https://github.com/DKFZ-E220/ROOx/blob/main/SPARQL%20queries%20examples%20from%20ROOx.txt

[18] Linked data. URL: https://github.com/DKFZ-E220/ROOx/tree/main/mapping_ROOx_v0.1

[19] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18