Paper DS01

FAIRification of preclinical data in radiation oncology through data standardization and formalized representation

Olga Giraldo, German Cancer Research Center (DKFZ), Heidelberg, Germany
Wahyu Wijaya Hadiwikarta, German Cancer Research Center (DKFZ), Heidelberg, Germany

ABSTRACT

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 and formalize data collected from preclinical radiation oncology studies to facilitate knowledge discovery. To achieve this goal, data is stored in relational databases and organized according to the CDISC SEND standard, and formalized by utilizing ontologies and semantic web technologies. FAIRdifying preclinical data could enhance personalized medicine initiatives and drive innovation in therapeutic development.

1. INTRODUCTION

In radiation oncology, preclinical trials generate vital data that serves as initial validation for treatment effectiveness by unraveling molecular mechanisms underlying different treatment responses and further allowing optimizing of clinical methods. Standardizing the practice in managing preclinical trial data ensures uniformity and reproducibility across multiple studies. This fosters collaborations, streamlines regulatory review, and enhances the reliability of findings.

Some efforts are focused in the standardization of preclinical data, such as the Standard for Exchange of Nonclinical Data (SEND) from the Clinical Data Interchange Standards Consortium (CDISC). SEND provides a standardized framework for structuring and presenting data from preclinical studies that involve animal models [1].

In contrast, ontologies like the Dependency Layered Ontology for Radiation Oncology (DLORO) [2], Radiation Oncology Structures Ontology (ROS) [3], and Radiation Oncology Ontology (ROO) [4] focus on clinical radiation oncology, supporting human trials. However, these ontologies lack the necessary representation of preclinical data, limiting their applicability in preclinical contexts.

Conversely, when looking at the radiation oncology domain, we could not find any study aiming at: (a) developing and validating a ontology to be used to represent preclinical data in the radiation oncology domain; (b) use of a Knowledge Graph (KG) based in a ontology to integrate different preclinical databases into a FAIR environment (Findable, Available, Interoperable, Reusable) [5] . Building on this foundation, we developed the Ontology for Preclinical Trials in Radiation Oncology (PTRO) [6] to represent terminology concerning animal model exposure to treatments, demographic characteristics of these models, body weights, causes of death, and laboratory test results. To achieve interoperability with other biomedical and healthcare ontologies, the PTRO ontology leverages the Semantic Types Ontology (STY) [7] as a framework for high-level categorization of concepts. This categorization is based on the Unified Medical Language System (UMLS), a comprehensive resource from the U.S. National Library of Medicine (NLM) [8], facilitating the alignment and integration of PTRO with established biomedical terminologies and enhancing its utility across diverse datasets.

This work focuses on the standardization and formalization of data collected from preclinical studies in radiation oncology to support advanced knowledge discovery. To achieve this, the data are stored in relational databases, structured according to the CDISC SEND standard, and formalized using the PTRO ontology alongside semantic web technologies. By combining SEND's data standardization with PTRO's ontology formalization, the FAIR principles are effectively applied to preclinical data in radiobiology and radiation oncology. This structured and interconnected framework aims to enhance knowledge discovery, maximize data reusability, stimulate innovation in therapeutic development, and advance personalized medicine efforts.

In this article we: i) outline the process of standardizing and formalization of preclinical data in radiobiology and radiation oncology to ensure alignment with FAIR principles, ii) present the current progress of this process; and iii) end with a conclusion and prospects for future advancements.

2. DATA STANDARDIZATION AND FORMALIZATION PROCESS

The stages of the process are outlined below and depicted in Figure 1.

2.1. DATA CLEANING

This step started with the digitalization of preclinical data from paper forms about:

- Project overview of the tumour model (treatment arms)
- List of the mouse numbers
- Data of Measurements (tumor size, weight etc.)

- Treatment schemes (irradiation and chemo therapy details)

Inconsistencies, or incomplete entries were rectified, or removed.

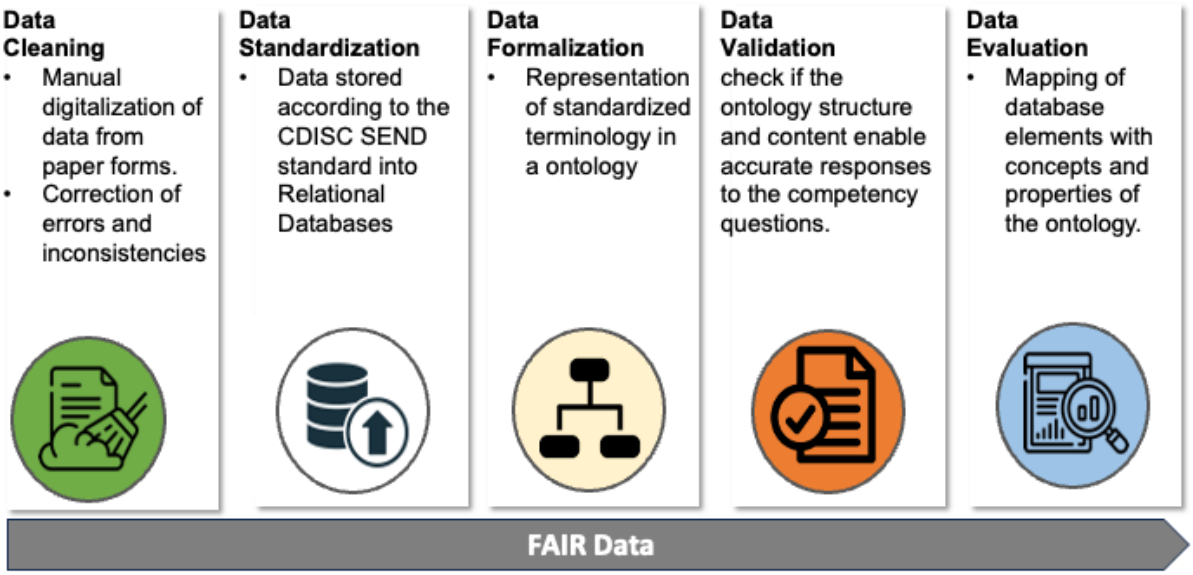

Figure 1. Data Standardization and formalization process

2.2. DATA STANDARDIZATION

The digitized data was organized and structured according the SEND standard for the tabulation of nonclinical datasets. Key advantages of the SEND standard include: i) its tailored approach for managing nonclinical study data, making it particularly effective for preclinical applications, and (ii) its role in regulatory compliance, especially with FDA requirements, as SEND adherence is mandatory for nonclinical submissions. Compliance with SEND not only enhances data consistency but also enables regulatory reviewers to more effectively navigate and assess datasets, ultimately streamlining the review process.

The standardized data include critical data from multiple studies, which are part of the German Cancer Consortium's (DKTK) RadPlanBio platform [9,10,11] operated under the German Cancer Research Center (DKFZ), in Heidelberg.

2.3. DATA FORMALIZATION

The standardized terminology according the SEND standard is being represented in an ontology called the Ontology for Preclinical Trials in Radiation Oncology (PTRO). This ontology represents an advancement of a prior model [12], incorporating improvements to achieve greater accuracy and interoperability. The development process consists of three steps: i) collection of preclinical concepts available in RadPlanBio; ii) semantic analysis of existing vocabularies, and; iii) ontology extension. The PTRO represent terminology concerning animal model exposure to treatments, demographic characteristics of these models, body weights, causes of death, and laboratory test results.

2.4. DATA VALIDATION

The validation process ensures that the ontology can effectively represent and capture the knowledge and data from the preclinical relational databases. This validation process involves formulating SPARQL queries based on each competency question and executing them on the ontology. This allows the validator to check if the ontology structure and content enable accurate responses to the questions.

2.5. DATA EVALUATION

Once the ontology can address all competency questions, next step is to proceed with the evaluation phase. This process entails mapping the database elements (rows, columns, and values) to the concepts and properties (predicates) in the ontology. Figure 2 illustrate the alignment between relational database columns and ontology entities. To achieve this mapping, we utilize the RDF Mapping Language (RML), an extension of R2RML to link columns and rows of preclinical databases and our ontology. Details on the evaluation process is available at [12].

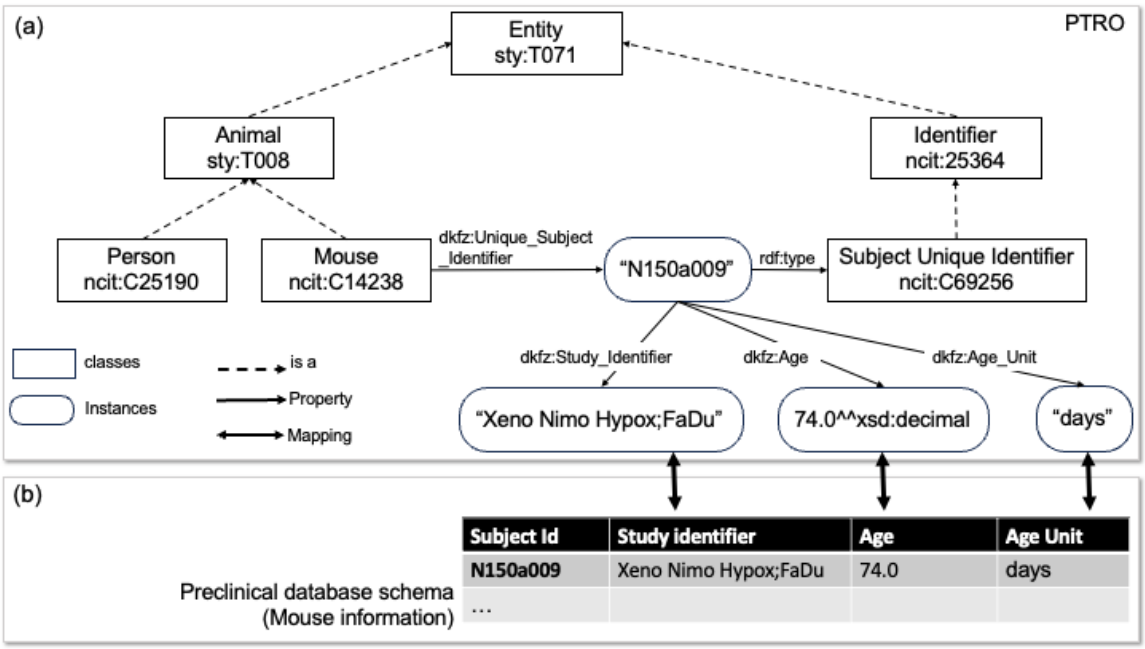

Figure 2. Overview of the PTRO structure and the relational database. The hierarchical structure of PTRO is presented in (a). The mapping performed to columns and values in a database is presented in (b).

3. RESULTS

3.1. DATA STANDARDIZATION

The RadPlanBio platform hosts a comprehensive collection of databases from various studies, such as tumour volume measurements, biomarker expression results, and treatment protocols from over 5000 mice. These databases are structured according to the SEND standard and include the following datasets:

- **Demographic Characteristics (DM)**: Detailed animal demographics.
- **Comments (CO)**: Observational notes and relevant remarks.
- **Exposure Details (EX)**: Information on the animals' treatment exposures.
- **Body Weights (BW)**: Records of animal body weights both during and at the study's conclusion.
- **Cause of Death (DD)**: Diagnoses regarding causes of mortality.
- **Laboratory Test Data (LB)**: Laboratory results for individual animals.
- **Clinical Observations (CL)**: Clinical observations recorded throughout the study.

This structured approach in RadPlanBio ensure consistency and facilitate integration and analysis across studies.

3.2. DATA FORMALIZATION

3.2.1. The ontology for preclinical trials in radiation oncology (PTRO)

The PTRO ontology include 958 classes, 8 object properties and 55 data properties. To ensure interoperability with other biomedical and healthcare ontologies, the PTRO ontology incorporates the Semantic Types Ontology (STY) as a framework for broad concept categorization. Leveraging STY enables PTRO to align seamlessly with established biomedical terminologies, enhancing its integrative capabilities and utility across various datasets.

PTRO reuse terminology that come primarily from the National Cancer Institute Thesaurus (NCIt) [13] as the CDISC Controlled Terminology [14] is part of the NCIt. New terms not derived from existing ontologies use the prefix ‘ptro’ and a local ID that starts with the letters DKFZ followed by 6-digit numbers. New defined terms represent: i) attributes that are common across the used databases e.g., “subject unique identifier (ncit:C69256)”, study identifier (ptro:DKFZ000008); ii) demographic characteristics e.g., “strain (ptro:DKFZ000041)”; iii) findings or information collected during a study e.g., “body weight (ptro:DKFZ000017)”, “cause of death (ptro:DKFZ000021)” and “clinical observation ptro:DKFZ000036)”; iv) exposure information e.g., “treatment name (ptro:DKFZ000013)”, “route of administration (ptro:DKFZ000011)” and “treatment vehicle (ptro:DKFZ00012)”.

PTRO is saved as OWL, Protégé v. 5.6.1 [15] was used to create new concepts and manage the ontology. The ontology has been publicly released at github [16].

3.3. DATA VALIDATION

3.3.1. SPARQL queries

The ontology represents and captures the knowledge and data from the preclinical databases. SPARQL queries were tested using a Protégé desktop plug-in that facilitates both the composition and execution of SPARQL queries. All queries returned the expected results, verifying the ontology's utility in accessing precise data points. For instance, to locate "all male mice of a specific age (74 days)," a query retrieves all unique identifiers for male mice aged 74 days. The full set of queries used to validate this ontology is accessible on GitHub [17].

```
PREFIX ptro: <http://www.radplanbio.org/ptro/>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX uo: <http://purl.obolibrary.org/obo/UO>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?Subject_Unique_Identifier

WHERE
{
  ?Subject_Unique_Identifier ptro:DKFZ000040 ?Sex .
  FILTER (?Sex = 0).
  ?Subject_Unique_Identifier ptro:DKFZ000033 ?Age .
  FILTER (?Age = 74.0)
  }
```

3.4. DATA EVALUATION

3.4.1. Linked data

Following Linked Data principles, our ontology supports semantic interoperability for preclinical data stored in databases, enhancing data sharing and transparent access within the RadPlanBio framework. Figure 3 illustrates RDF triples derived from a demographic database containing information on 20 mice, with two examples shown here for brevity. Initially The database was transformed into CSV format; then, RDF triples were generated for each subject ID. For instance, the first mouse, "N150a009," and the second, "N150a011," both participated in study “Xeno Nimo Hypox.FaDu,” with identifiers "9" and "11," respectively. 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. CONCLUSION AND FUTURE WORK

The standardization of preclinical data with the SEND standard and formalizing it using the PTRO ontology creates a structured, interoperable, and semantically rich framework that enhances data findability, accessibility, interoperability, and reusability (FAIR) in radiation oncology. Specifically, by formalizing the SEND-standardized data with the PTRO ontology, researchers achieve greater semantic interoperability, making the data machine-readable and allowing for complex queries. PTRO enriches the dataset with domain-specific terms and relationships that are critical for radiation oncology, such as those involving animal demographics, findings, and treatments exposures.

Combining SEND standardization with PTRO ontology formalization enables comprehensive data integration across diverse studies available on RadPlanBio. This facilitates more robust knowledge discovery, allowing for more effective translation of preclinical findings to clinical applications, and supports personalized treatment approaches by providing clear links between preclinical data and clinical outcomes. Future work will focus on integrating the ontology into the semantic layer of the RadPlanBio platform through a knowledge graph, to enable semantic queries, reasoning and inferences.

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. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data 3(160018) (Mar 2016). https://doi.org/https://doi.org/10.1038/sdata.2016.18
6. Ontology for Preclinical Trials in Radiation Oncology (PTRO). URL: https://github.com/DKFZ-E220/PTRO
7. The Semantic Types Ontology (STY). URL: https://bioportal.bioontology.org/ontologies/STY
8. The Unified Medical Language System (UMLS). URL: https://www.nlm.nih.gov/research/umls/index.html
9. 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
10. RadPlanBio. URL: https://helmholtz.software/software/radplanbio
11. Hadiwikarta W., Ebert N., Roscher M., Kurth I., Baumann M., Knowledge Management System in Preclinical Radiooncology / Radiobiology Research, CEUR-WS.org/Vol-2721 - ISWC 2020 Posters, Demos, and Industry Tracks, 237-242, http://ceur-ws.org/Vol-2721/paper560.pdf. 2020.
12. Olga Giraldo, Abumansur Sabyrrakhim, Mareike Roscher et al. Semantic Representation of Preclinical Data in Radiation Oncology., 03 November 2024, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-5373454/v1]
13. NCI Thesaurus (NCIt). URL: https://ncithesaurus.nci.nih.gov/ncitbrowser/
14. CDISC Controlled Terminology. URL: https://www.cdisc.org/standards/terminology/controlled-terminology
15. Protégé. URL: https://protege.stanford.edu/
16. Ontology for Preclinical Trials in Radiation Oncology (PTRO). URL: https://github.com/DKFZ-E220/PTRO
17. SPARQL queries – PTRO. URL: https://github.com/DKFZ-E220/PTRO/blob/main/SPARQL%20queries%20examples%20from%20PTRO.txt
18. Mappings – PTRO. URL: https://github.com/DKFZ-E220/PTRO/tree/main/mapping_PTRO_v0.1

ACKNOWLEDGMENTS

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, as well as to Abumansur Sabyrrakhim, Mareike Roscher, Rosemarie Euler-Lange, Dr. Michael Baumann, and Dr. Ina Kurth for their valuable insights and feedback.

CONTACT INFORMATION

Feedback and inquiries are welcome. Below is the contact information for the authors of this article:

Author Name: Olga Giraldo
Company: German Cancer Research Center (DKFZ)
Address: Neuenheimer Feld 280, 69120 Heidelberg, Germany
Email: olga.giraldo@dkfz-heidelberg.de

Author Name: Wahyu Wijaya Hadiwikarta
Company: German Cancer Research Center (DKFZ)
Address: Neuenheimer Feld 280, 69120 Heidelberg, Germany
Email: w.hadiwikarta@dkfz-heidelberg.de