2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782443

Towards the development of a FAIR-compliant biomedical ontology for colorectal cancer

Vasileios Pezoulas
Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, and Biomedical Resarch Institute (BRI), FORTH-IMBB, GR 45110 Ioannina, Greece,
bpezoulas@gmail.com

Christos Androutsos
Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, and Biomedical Resarch Institute, FORTH-IMBB, GR 45110 Ioannina, Greece,
xristosandroutsos95@gmail.com

Dimitrios I. Fotiadis*
Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, and Biomedical Resarch Institute (BRI), FORTH-IMBB, GR 45110 Ioannina, Greece,
fotiadis@uoi.gr

***Abstract*— Despite the widespread development of ontologies in many domains of healthcare, the field of colorectal cancer (CRC) presents a notable gap considering the lack of standardized data models tailored to the CRC domain. To address this gap, we developed a structured, systematic approach to model CRC-related data. A reference data model (RDM) was first defined, including 250 terminologies related to sociodemographics, clinical biology, lifestyle, symptoms, among others. Each terminology was linked with globally recognized HL7 standards, including SNOMED-CT, LOINC, and ICD-10/11, among others, fulfilling the principles of findability, accessibility, interoperability, and reusability (FAIR) to facilitate its integration into global healthcare systems. The RDM was then transformed into an ontology using Protégé. The resulting ontology presents a unified framework for CRC data, enhancing data consistency across various healthcare platforms and research studies. The ontology's availability on GitHub invites global collaboration, ensuring its continual evolution and relevance in the rapidly advancing landscape of healthcare. The proposed standardized, accessible, and comprehensive data model has the potential to significantly impact CRC research and treatment, paving the way for more effective, personalized patient care and fostering a collaborative environment for ongoing advancements in the field.**

Keywords: colorectal cancer, FAIR principles, ontologies.

I. Introduction

Ontologies are hierarchical data models that enable the organization, classification, and systematic analysis of complex medical data [1-3]. These structured vocabularies and sets of concepts underpin the interoperable exchange of information across various healthcare systems, facilitating a shared understanding among diverse stakeholders, including clinicians, researchers, and policymakers. Ontologies support a multitude of functions, from advancing clinical decision support systems to enhancing the efficacy of biomedical research. However, the true potential of ontologies is fully realized only when they adhere to the FAIR (Findability, Accessibility, Interoperability, and Reusability) guiding principles [4-7]. The FAIRification of ontologies ensures that these rich data resources are discoverable by both humans and machines, accessible with clear usage licenses, compatible with other datasets, and equipped with comprehensive documentation for future applications [4-8]. In this context, FAIR ontologies are not just beneficial but essential for maximizing the utility of the data, promoting innovation, and accelerating the pace of research and discovery in healthcare.

Although ontologies have been widely used for data modeling across various areas of healthcare, the domain of colorectal cancer (CRC) stands out due to the significant absence of standardized data models specifically designed for CRC studies. CRC is the third most prevalent form of cancer and the fourth leading cause of cancer-induced mortality [9]. The likelihood of encountering CRC is estimated to be around 4%-5% [9], with risk factors tied to individual characteristics or lifestyle choices, including age, history of chronic illnesses, and lifestyle. While CRC is predominantly diagnosed in the 65-74 age group [10], there is an escalating trend of CRC in younger individuals which is attributed to factors like obesity, physical inactivity, poor dietary choices rich in fats and proteins, smoking, and other progressive lifestyle elements. By 2040, it is estimated that the incidence of CRC will rise to 3.2 million new cases annually, marking a 63% increase, and the death toll will reach 1.6 million per year, indicating a 73% rise compared to 2020 [11]. As a matter of fact, the development of a CRC-specific ontology would pave the way for more targeted and effective healthcare interventions to improve patient outcomes in this critical domain. Such an ontology would revolutionize the way that different types of data can be used for CRC prevention and treatment, thus offering opportunities for personalized medicine and epidemiological surveillance.

To bridge this gap, we developed a first, FAIR-compliant, CRC ontology which includes 250 terminologies related to sociodemographics, clinical biology, lifestyle, diagnosis, and symptoms, among others. Each terminology is interlinked with widely recognized HL7 standards like SNOMED-CT, LOINC, ICD-10/11, offering a standardized representation of CRC-related data. The proposed ontology adheres to the FAIR principles (under the aegis of FAIR CookBook and FAIR plus) and serves as a dynamic, easily accessible tool for global researchers and healthcare professionals. Through this way, the CRC ontology stands as a cornerstone in advancing personalized medicine in CRC treatment, allowing for

Funded by the European Union (DIOPTRA, 101096649). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI). Funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10056682].

interventions tailored to detailed, standardized patient profiles. Such a richly detailed ontological structure empowers researchers and healthcare providers to query complex datasets efficiently, draw meaningful insights, and contribute to advancing personalized patient care and evidence-based medicine in the domain of CRC. The CRC ontology is also accessible through GitHub to further enhance the scope and utility of the ontology for research purposes.

II. Materials And Methods

A. Definition of the CRC Reference Data Model (RDM)

A reference data model (RDM) was defined in cooperation with the clinical experts of the CRC domain. The RDM serves a structured tabular data model which has been designed to standardize the collection, organization, and interpretation of diverse data obtained from multiple clinical sites participating in CRC studies. The RDM offers a detailed understanding of the features, reflecting the CRC domain knowledge following a uniform approach to gathering, integrating, and analyzing the data. Figure 1 depicts all the elements that synthesize the data model for the data encoding process in each clinical site.

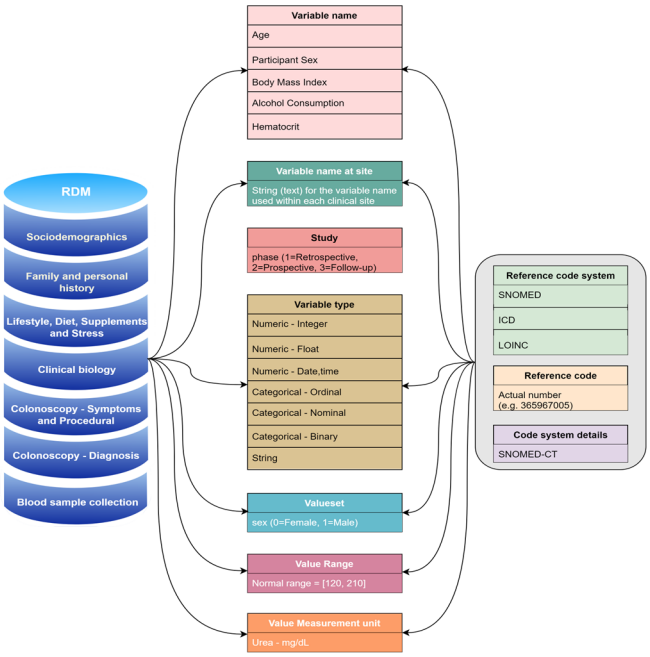

Fig. 1. The structure of the CRC RDM.

In the RDM, each column holds valuable information regarding the categorization and definition of variables to ensure uniformity and consistency across the data collection process. The variables are grouped into seven categories (or classes) related to sociodemographics, family and personal medical history, behavioral (e.g. lifestyle, diet, supplements), clinical biology, symptoms, diagnosis, blood sample collection. The RDM components are presented next, highlighting their significance in harmonizing the CRC data elements for efficient data preparation within the data lifecycle.

- **Variable name:** This refers to the standardized variable name within a CRC study. It ensures consistency and coherence across all clinical sites and helps to establish a common understanding of the variables.
- **Variable name at site:** The name used for the variable within each clinical site's data. Despite standardization, individual sites may refer to variables by different names, making this alignment crucial for data interlinking.
- **Study:** Denotes whether the variable is applicable to different phases of the study, Retrospective, Prospective, and potential use in Follow-up phases. Understanding variable relevance at each stage ensures proper utilization throughout the study.
- **Variable type:** This uniquely identifies each variable within the dataset, providing a reference for tracking and management purposes. It allows for efficient cross-referencing and identification of specific data elements.
- **Variable type:** Specifies the nature of the variable, whether it is numeric-integer, numeric-float, numeric-datetime, categorical-ordinal, categorical-nominal, categorical-binary, string.
- **Valueset and Value Range:** Includes information about the values or normal range that a variable can take, including the accepted format or structure. These components delineate the acceptable parameters and data structure for accurate representation and interpretation.
- **Value Measurement unit:** This is the unit of measurement associated with numerical or quantitative variables. Standardized units ensure consistency and accuracy in data interpretation.
- **Reference code system:** Specifies the coding system (e.g. SNOMED-CT [12], LOINC [13], ICD-10/11 [14]) or reference used to represent the variable, ensuring uniformity, and facilitating seamless data integration and understanding. Each code serves as a unique identifier, simplifying data cross-referencing, integration processes, and variable identification.

B. Expressing the RDM into an ontology

The RDM was transformed into an ontology following a structured process with respect to the FAIR CookBook [15] and FAIRplus principles [5] through Protégé [16]. Initially, the entities, relationships, and properties of the RDM were identified. This information was then translated into a hierarchical data model by creating entities that correspond to the categories of the RDM. Relationships and entity attributes were represented as object and data properties, respectively, with careful consideration of domains and ranges. Constraints from the RDM were expressed through OWL axioms, reflecting cardinalities and other restrictions. To ensure broader interoperability, the ontology was aligned with existing HL7 standards, such as, SNOMED-CT [12], LOINC [13], and ICD-10/11 [14]. Annotations were added for clarity, serving as documentation for each element within the ontology. The ontology was then validated for logical consistency using Protégé's reasoning tools, with adjustments made as necessary. Once finalized, the ontology was exported in a standard format

such as RDF/XML or OWL, ready for sharing and integration into data management systems.

C. Achieving FAIR compliance

The ontology has been designed to be FAIR following the FAIR CookBook and FAIRplus principles. These principles are essential for ensuring that the data and metadata are well-managed and can be easily shared and used by others. The terminology interlinking with the SNOMED-CT [12], LOINC [13] and ICD-10/11 [14] code systems indicates that the CRC ontology uses standardized clinical terminologies, which helps with interoperability. Accessibility is ensured by providing open access to the ontology in standard formats [17] such as RDF, OWL, or JSON, along with maintaining a stable retrieval URL and clear access terms. Lastly, making the ontology available on GitHub suggests a commitment to openness and collaboration, allowing other researchers to access, use, and possibly contribute to the ontology.

III. Results

A. The CRC ontology

The different components of the CRC ontology are depicted in Fig. 2, including the hierarchical structure of the entities (classes) and their relationships (the lines indicate subclass (is-a) relationships) and the data properties using Protégé OntoGraf [18] (Fig. 2(B)) and WebVOWL [19] (Fig. 2(C)). The layout is a directed acyclic graph where edges point from subclasses to classes, reflecting the ontology's taxonomy. According to Fig. 2 (A), (B), the classes of the CRC ontology are presented next:

- **Individual**: The root class of the CRC ontology.
- **Study_information**: Includes data properties specific to individual patients or participants in the study.
- **Behavioral**: Includes data properties related to alcohol consumption, physical activity, and smoking.
- **Blood_samples_collection_metadata**: Includes metadata from the collection of blood samples during examination.
- **Clinical_biology**: Includes biological data relevant to the clinical aspects of CRC.
- **Diagnosis**: Includes information related to CRC diagnosis.
- **Family_and_personal_medical_history**: This class includes medical data of the individual and their family.
- **Sociodemographics**: Includes demographic data of the individuals participating in the study.
- **Symptoms_and_colonoscopy_procedural_measurements**: Includes data about symptoms/medical conditions of the individuals and findings from colonoscopy procedures.

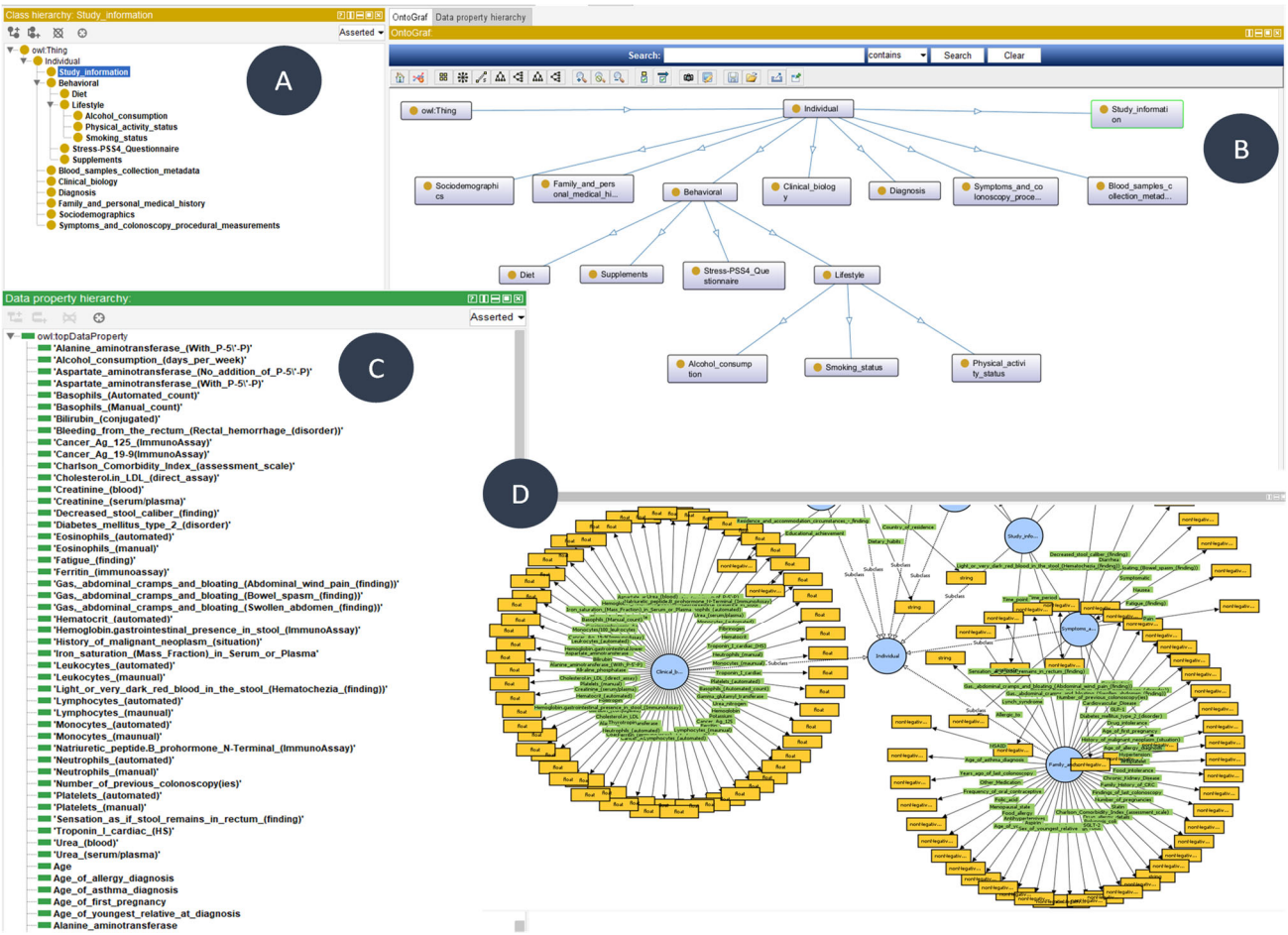

Fig. 2. Different components of the proposed CRC ontology. (A) The class hierarchy of the individual. (B) An illustration of the class hierarchy using OntoGraf [18]. (C) The data property hierarchy. (D) Instances of different classes using WebVOWL [19].

- **Diet**: Includes dietary information (from questionnaires).
- **Supplements**: Includes various dietary supplements.
- **Stress-PSS4_Questionnaire**: Includes questions under the Perceived Stress Scale (PSS) questionnaire.
- **Lifestyle**: Includes information about the lifestyle of the participant which intersects with behavioral data.

B. Data properties

The CRC ontology currently includes 250 terminologies which are related to the classes presented in Fig. 2. Each terminology has been linked with internationally recognized standards such as SNOMED-CT, LOINC, and ICD-10/11 to ensure interoperability and to facilitate a seamless integration into global healthcare systems. Emphasis has been given to the definition of data types for each property, with distinctions made between strings, integers, and other relevant data types.

C. Compliance with the FAIR plus guidelines

The proposed ontology's compliance with the FAIR CookBook criteria [15] is comprehensive. In modeling the domain, it utilizes metadata extraction, relational mapping, and data modeling stages, employing identifiers from widely recognized HL7 data models such as SNOMED-CT [12], LOINC [13], and ICD-10/11 [14] for constructing reference ontologies. The ontology addresses identifier mapping to make the data models interoperable. It applies data standards through reusing, developing, applying, and validating HL7 standards. For the selection of data vocabularies, the solution emphasizes on the selection, annotation, and management based on the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [20]. For data interoperability, it focuses on identifier mapping, vocabulary alignment, and data model mapping based on HL7 FHIR-based data models. The CRC ontology is publicly available for collaboration on our GitHub repository (https://github.com/vpz4/CRC_Ontology). It has been released in OWL/XLS formats to cover different use cases and preferences for interacting with CRC-related data.

IV. Conclusions

The development of a first, FAIR-compliant, ontology for the CRC domain represents a significant innovation with the potential to profoundly impact healthcare. Its novelty lies in providing a structured model for categorizing and sharing CRC-related data. In addition, the CRC ontology promotes interdisciplinary collaboration by offering a unified, HL7 driven set of terminologies for clinicians, researchers, and data scientists. From a clinical point of view, the proposed CRC ontology can enhance decision support systems and facilitate more accurate diagnoses and personalized treatment plans. Considering that the CRC ontology adheres to FAIR principles also ensures that data are not only easily accessible but also available for reuse in ongoing and future studies.

Moreover, by aligning with recognized standards like SNOMED-CT, it promotes interoperability with other health systems and research, enabling a seamless exchange of information. Beyond its immediate utility in research and clinical practice, the proposed ontology can serve as a vital educational resource, underpinning training, and development in oncology. The fact that the ontology has been made available on GitHub enables the global community to contribute to its refinement, ensuring that the components of the ontology are aligned with the latest scientific discoveries. We also plan to upload the proposed CRC ontology to https://fairsharing.org/ to further enhance its visibility and adoption.

References

[1] Tudorache, T., "Ontology engineering: Current state, challenges, and future directions," Semantic Web, vol. 11, no. 1, pp. 125-138, 2020.

[2] de Mello, B. H., Rigo, S. J., da Costa, C. A., da Rosa Righi, R., Donida, B., Bez, M. R., and Schunke, L. C, "Semantic interoperability in health records standards: a systematic literature review," Health and technology, vol. 12, no. 2, pp. 255-272, 2022.

[3] Pezoulas, V., Exarchos, T., and Fotiadis, D. I., "Medical data sharing, harmonization and analytics," Academic Press, Elsevier, 2020.

[4] Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., et al., "FAIR principles: interpretations and implementation considerations," Data intelligence, vol. 2, no. 1-2, pp. 10-29, 2020.

[5] Harrow, I., Balakrishnan, R., McGinty, H. K., Plasterer, T., & Romacker, M., "Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q," Drug Discovery Today, vol. 27, no. 5, pp. 1441-1447, 2022.

[6] Inau, E. T., Sack, J., Waltemath, D., and Zeleke, A. A., "Initiatives, concepts, and implementation practices of FAIR (findable, accessible, interoperable, and reusable) data principles in health data stewardship practice: protocol for a scoping review," JMIR research protocols, vol. 10, no. 2, pp. e22505, 2021.

[7] Kim, J. W., et al. "Scalable Infrastructure Supporting Reproducible Nationwide Healthcare Data Analysis toward FAIR Stewardship," Scientific Data, vol. 10, no. 1, pp. 674, 2023.

[8] Touré, V., et al., "FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network," Scientific Data, vol. 10, no. 1, pp. 127, 2023.

[9] Mármol, I., et al., "Colorectal carcinoma: a general overview and future perspectives in colorectal cancer," International journal of molecular sciences, vol. 18, no. 1, pp. 197, 2017.

[10] Granados-Romero, et al., "Colorectal cancer: a review," Int J Res Med Sci, vol. 5, no. 11, pp. 4667, 2017.

[11] https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer.

[12] https://www.snomed.org/.

[13] https://loinc.org/.

[14] https://icd.who.int/en.

[15] Rocca-Serra, et al., "The FAIR Cookbook-the essential resource for and by FAIR doers," Scientific data, vol. 10, no. 1, pp. 292, 2023.

[16] https://protege.stanford.edu/.

[17] Hendler, J., Gandon, F., and Allemang, D., "Semantic web for the working ontologist: Effective modeling for linked data, RDFS, and OWL," Morgan & Claypool, 2020.

[18] https://protegewiki.stanford.edu/wiki/OntoGraf.

[19] https://service.tib.eu/webvowl/.

[20] Reich, C., et al., "OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization," JAMIA, pp. ocad247, 2024.

[21] Duda, S. N., Kennedy, N., Conway, D., Cheng, A. C., Nguyen, V., Zayas-Cabán, T., and Harris, P. A., "HL7 FHIR-based tools and initiatives to support clinical research: a scoping review," JAMIA, vol. 29, no. 9, pp. 1642-1653, 2022.