

ONTOLOGIES FOR ONCOLOGICAL RADIOLOGY: CHALLENGES AND OPPORTUNITIES

MARIO DI LONARDO¹, CARLO SARTIANI²

¹Research Associate, University of Basilicata, Department of Engineering, Italy

² Associate Professor, University of Basilicata, Department of Engineering, Italy

E-mail: ¹mario.diloardo@studenti.unibas.it, ²carlo.sartiani@unibas.it

ABSTRACT

Medical imaging examinations, especially Magnetic Resonance Imaging (MRI), interpreted by radiologists in the form of narrative reports, are used to produce and confirm diagnoses in clinical practice at different levels. Being able to accurately and quickly identify the information scattered in the radiologists' narratives has the potential to reduce workloads, support clinicians in their decision processes, triage patients to get urgent care or identify and cluster patients for research purposes. This is especially critical within the context of the Tumor Boards, multidisciplinary groups made up of different specialists, who regularly meet to discuss oncological patients potentially needing pre/post-surgery treatments and to make diagnostic and therapeutic decisions for them.

Nowadays, it is still difficult to access and analyze radiology reports both effectively and efficiently at scale, due to their unstructured nature, the conciseness and often crypticity of the medical jargon used, and the background knowledge usually required for interpreting them correctly and making high-level correlations and conclusions. Privacy concerns introduce further difficulties and impose the adoption of local tools rather than cloud-based ones, hence preventing the use of popular LLMs.

Ontologies represent an important tool for easing the automatic process of medical records and radiology reports. Indeed, they can be used to add structure to otherwise unstructured texts; furthermore, they play a key role in data exchange and sharing, as they enable semantic interoperability.

In this review paper we will introduce the problems related to the use of ontologies in the medical domain, with particular emphasis on radiology reports, and discuss the prominent advantages that a systematic use of ontologies would generate.

Keywords: *Ontologies, RDF, Mappings, Annotations, ML*

1. INTRODUCTION

Medical imaging examinations play a pivotal role in modern clinical practice, with Magnetic Resonance Imaging (MRI) standing out as one of the most sophisticated and widely utilized diagnostic tools. Radiologists translate their findings into detailed narrative reports. These reports are essential as they guide the diagnostic process, aiding healthcare providers in confirming or establishing a diagnosis across various levels of medical care.

The ability to accurately and swiftly identify crucial information embedded within these radiologists' narrative reports can significantly impact multiple facets of healthcare delivery. One of the primary benefits is the potential reduction in workloads for both radiologists and clinicians. By efficiently extracting key data points, healthcare professionals can streamline their review processes,

thereby saving valuable time and resources. This efficiency is particularly beneficial in high-volume clinical environments, where timely diagnosis and treatment are critical.

In addition to easing workloads, precise identification of pertinent information supports clinicians in their decision-making processes. When critical diagnostic details are readily accessible, clinicians can make more informed choices regarding patient care. This capability enhances the overall quality of healthcare, as decisions are based on comprehensive and accurately interpreted data.

Moreover, the rapid and accurate extraction of information from narrative reports plays a vital role in patient triage. In emergency or urgent care settings, time is of the essence. Quickly identifying patients who require immediate attention ensures that they receive prompt and appropriate medical interventions. This not only improves patient

outcomes but also optimizes the use of healthcare resources.

Beyond immediate clinical applications, the ability to cluster and identify patients based on specific diagnostic criteria holds significant value for research purposes. Researchers can utilize this information to identify patient cohorts that share common characteristics, facilitating studies on disease patterns, treatment efficacy, and long-term outcomes. This data-driven approach accelerates medical research and contributes to the advancement of personalized medicine.

This capability is particularly critical within the framework of Tumor Boards. Tumor Boards are multidisciplinary groups composed of various specialists, including oncologists, surgeons, radiologists, pathologists, and other healthcare professionals. These groups convene regularly to discuss the cases of oncological patients, focusing on those who may require pre-surgical or post-surgical treatments. The collaborative nature of Tumor Boards ensures that diagnostic and therapeutic decisions are comprehensive and consider multiple expert perspectives.

In the context of Tumor Boards, having immediate access to accurately identified information from radiologists' narratives enhances the efficiency and effectiveness of case discussions. Specialists can quickly grasp the key diagnostic findings, enabling them to make well-informed decisions regarding patient management. This collaborative decision-making process is crucial for developing tailored treatment plans that address the unique needs of each cancer patient.

Accessing and analyzing radiology reports on a large scale remains a complex and challenging task. This difficulty primarily arises from several inherent characteristics of these reports. Firstly, the unstructured nature of radiology reports makes it hard to systematically process and interpret the data they contain. Unlike structured data, which is organized in predefined formats like tables or databases, unstructured data lacks a consistent format, rendering it less accessible for automated systems and more labor-intensive for human analysis.

Another significant factor contributing to this challenge is the concise and often cryptic medical jargon employed in these reports. Medical professionals frequently use specialized terminology and abbreviations that can be difficult for individuals outside the medical field to understand. Even for

those within the healthcare industry, deciphering the exact meaning can require considerable effort, as the language is designed to convey complex information succinctly, often omitting explanatory details that are assumed to be understood by the intended professional audience.

Furthermore, interpreting radiology reports accurately and drawing meaningful, high-level correlations and conclusions generally necessitates a substantial amount of background knowledge. This expertise is not just limited to understanding medical terminology, but also involves familiarity with various medical conditions, imaging techniques, and clinical contexts. Such background knowledge enables professionals to make connections between disparate pieces of information, derive insights, and support clinical decision-making.

In light of these challenges, ontologies emerge as a vital tool in simplifying the automatic processing of medical records and radiology reports. Ontologies, in the context of information science, refer to structured frameworks that organize information into categories and define the relationships between concepts. By applying ontologies, unstructured medical texts can be enriched with a level of structure that facilitates data organization, retrieval, and analysis.

The utility of ontologies extends beyond merely adding structure to unstructured data. They also play a crucial role in enhancing data exchange and sharing across different systems and institutions. This is achieved through semantic interoperability, which refers to the ability of systems to not only exchange data but also to interpret the meaning of that data consistently. Semantic interoperability ensures that the information remains meaningful and contextually accurate when shared between diverse healthcare systems, thereby supporting coordinated patient care, research, and public health initiatives.

In this survey, we will introduce the problems related to the use of ontologies in the medical domain, with particular emphasis on radiology reports, and discuss the prominent advantages that a systematic use of ontologies would generate.

2. ONTOLOGIES

2.1 The Crucial Role of Ontologies

In the last few years, significant advancements have been made in personalized medicine, genetic profiling, and machine learning. Consequently, the demand for high-quality big data has surged. Furthermore, studies of genetic and epigenetic

factors in cellular systems have revealed that the development of diseases encompasses billions of interacting processes.

To facilitate progress in disease treatment, it is imperative that knowledge acquired in one medical domain be correlated with knowledge gained in other domains. Researchers must engage in effective communication by comprehending areas beyond their expertise. These requirements underscore the necessity for organizing and standardizing knowledge in a manner that also facilitates its dissemination. In this context, ontologies emerge as a viable solution, particularly in fields where information is stored in electronic format.

2.2 Definition of Ontologies

The term “ontology” originates from Greek and translates to “study of being or what is.” While this definition has held true in the philosophical realm for millennia, it has evolved into a more contemporary understanding within the information sciences.

An ontology is essentially a “formal and explicit specification of a shared conceptualization.” This formal description adheres to defined rules, ensuring clarity and precision in defining concepts and relations. Furthermore, ontologies are “shared” because they are widely accepted and “computer-understandable” in the context of human knowledge related to a specific domain.

According to the realist school, an ontology must encompass a taxonomy of universals and defined classes, i.e., the entities, and the relationships between them. This taxonomy provides a tree or graph representation of the classification of entities. Universals are general qualities of entities that account for their similarities and represent natural qualities. They are instantiated by specific entities in reality, such as “The City of New York.” However, universals are often confused with defined classes. Defined classes are groups of entities that share a common selection criterion, like all smokers in New York City with lung disease, but they do not represent natural characteristics.

Relationships between entities are hierarchical, meaning child nodes inherit features from nodes above. These relationships can be of three types: between universals, between universals and particulars, and between particulars.

The community of scientific researchers in the oncology domain has already begun to use ontologies, standardize vocabulary, taxonomies, and interoperability modes to create knowledge sharing standards that converge into a single framework.

However, it was decided to represent knowledge in a specific domain with multiple ontologies.

In practice, it has been efficient to describe knowledge using *top-level* or *formal* ontologies, and *application* ontologies. The former establish the foundational entities and relationships that exist across all subdomains of a domain. Application ontologies, on the other hand, focus on a specific subset of the domain. In many cases, dividing the domain facilitates the implementation of the ontology. The formal ontology can serve as the basis for subdomains, and knowledge within these subdomains is limited to a manageable amount.

Ontologies are concerned with providing meaning to data and offering a representation of the knowledge within a specific domain. They achieve this by associating entities within the domain with definitions. These definitions can be either comprehensible only to humans or exclusively to machines: in the first case, these definitions are textual, while in the latter are based on logical axioms.

The preferred form for a textual definition is the “Aristotelian definition.” In this definition, entities are given the definition of their parent by adding the term “difference.” The term “difference” signifies the characteristics in which the child entities differ from the parent entities. Ontologies focus on standardizing terms and providing the semantic infrastructure necessary to ensure data interoperability. This requires standard ways of developing ontologies. In the 2000s, the Open Biomedical Ontologies (OBO) consortium proposed principles for the development of ontologies. These principles include: being open source; having unambiguous definitions and relationships; having a context; and using a namespace so that identifiers are kept unique.

These principles were selected, in part, to alleviate the confusion caused by overlapping ontologies and to facilitate the creation of a comprehensive schema that can accommodate various ontologies. Ontologies serve numerous purposes.

- They serve to share data and knowledge;
- They help in communication between people or between people and software;
- They are used in the realization of machine learning models;
- They are used for text extraction from clinical reports.

Multiple collaborative efforts have resulted in the development of tools and repositories dedicated to ontologies, providing invaluable resources to the biomedical community. The BioPortal, established

by the National Center for Biomedical Ontology, is an open repository of biomedical ontologies. These ontologies are developed in OWL, RDF [1] [2], and OBO formats and can be accessed through both web interfaces and web services [3]. Users can explore ontologies, add annotations, leave reviews, and view mappings between different ontologies through the web interface. Currently, the BioPortal hosts over 800 ontologies, including the Proteomics Standards Initiative, the OBO library, and the Semantic Type Ontology of the Unified Medical Language System (UMLS) [4].

UMLS is another source of biomedical vocabulary and ontologies, consisting of the Metathesaurus, Semantic Network, specialized lexicon, and lexical tools. The Metathesaurus includes various widely used terminologies, such as NCBI taxonomy, Medical Subject Headings (MeSH), and Digital Anatomist Symbolic Knowledge Base [5].

2.3 Data and knowledge sharing

There are several ways to enhance healthcare. Two particularly effective approaches leverage the data-driven approach: Learning Healthcare Systems and Distributed Learning.

A Learning Healthcare System involves sharing care and data on patient conditions to foster continuous improvement and knowledge acquisition. In contrast, a Distributed Learning System preserves patient privacy. It relies on infrastructure called *tracks*, where data holders like hospitals, referred to as *stations*, share information only with authorized applications called *trains*. These trains generate new knowledge by finding, accessing, and comprehending data from stations. Consequently, data must adhere to the FAIR principles: “Findable, Accessible, Interoperable, and Reusable.”

Ontologies can be very useful here as they provide an infrastructure for data and knowledge sharing, especially to minimize redundancy and discrepancies in the data itself. A number of ontologies have emerged in the field of radiology including the Dependency Layered Ontology for Radiation Oncology (DLORO), the Radiation Oncology Ontology (ROO), and the Radiation Oncology Structures Ontology (ROS).

Currently, efforts are being made to move toward a single radiology ontology updated and maintained by professional societies such as AAPM, ASTRO, and ESTRO, as well as the radiology community. Developing such an ontology, however, requires referring to ontologies adjacent to the radiology field

and relying on other very useful ontologies such as the FMA (Foundational Model of Anatomy), i.e. an ontology on the human anatomy; the NCI Thesaurus, which represents a vocabulary for clinical care, translational, and basic research; Common Terminology Criteria for Adverse Events (CTCAE), i.e., a standard, severity-scale classification of adverse events that have occurred in cancer therapy clinical trials and other oncology settings; the Radiomics Ontology, i.e., collection of radiomics features as well as other entities (e.g., software properties, filter properties, feature extraction parameters) involved in radiomics computation; and the Semantic DICOM Ontology, which is an ontology for DICOM.

Another interesting aspect of ontologies is that when linked to other ontologies, they are able to capture other knowledge. This is possible because the classes of some ontologies can be generalized or specialized by other ontologies. This characteristic can be exploited in some applications: for example, data providers can share a minimal dataset, which, enriched with ontologies, still allows the user to retrieve richer information.

Although an ontology is being developed to collect data on radiation oncology, a clear description of applications based on this knowledge is still lacking. For example, to use an application such as automatic AI contouring or a model to predict treatment complications, it is important to know who created it, what data it was trained and validated on, what patients it is suitable for, and what uncertainties it might have. In this sense, again, ontologies can lend a hand: having an ontology that systematically organizes this information would go a long way toward making models and software more secure and reliable.

2.4 Communication in radiotherapy

To facilitate effective communication among people, users, and software tools, as well as between different software tools, various standards have emerged that encapsulate radiotherapy terminology and concepts. One such standard is the DICOM-RT3 (Digital Imaging and Communications in Medicine) standard. As its name suggests, this standard is specifically designed for describing medical procedures. It divides each medical procedure into distinct parts and provides the syntax for describing each part.

Additionally, there are standards that regulate communication by defining the essential information to be exchanged. For instance, the HL7 standard, developed by the IHE-RO organization serves this

purpose. This standard defines the structure as well as the format of the messages to be exchanged. HL7 is primarily based on XML [6], [7], [8], [9], but there also exist extensions that exploit JSON and JSON Schema [10] [11] [12] [13] [14]. These extensions allow for a safer handling of messages, by exploiting the inherent advantages of JSON over XML [15] [16] [17] [18] [19]. However, these standards often have differing perspectives on the concepts they encompass and may present entirely different descriptive structures for them. This discrepancy can lead to confusion. In such cases, a community-accepted ontology, which provides a taxonomy, can help clarify concepts and formalize vocabulary. Furthermore, relationships between concepts within an ontology can provide additional detail, enhancing communication.

Ontologies also find practical applications in retrieving data from various databases created for diverse purposes and developed by different vendors. These databases include oncology information systems, electronic medical records (EMRs) used at the hospital or system level, treatment planning systems, and image archiving and communication systems (PACS). Typically, these databases are relational and possess distinct schemas. Ontologies, encompassing classes and properties, can be instrumental in mapping these databases, enabling data extraction from any database. Subsequently, the extracted data needs to be transformed into HL7 or DICOM messages, based on XML or JSON [20] [21] [22] [23] [24], or more contemporary formats that inherently support ontologies. An example of such standards is FHIR (*Fast Healthcare Interoperability Resources*), which is similar to DICOM in that it encapsulates a conceptual model of healthcare delivery and outlines the types of processes and procedures that should be described. As for HL7 v.3 and DICOM, FHIR is based on XML as well as on JSON [25] [26] [27] [28].

2.5 Implementation of machine learning models

Ontologies can significantly aid in the development of machine learning (ML) models, particularly probabilistic models. In recent years, machine learning has gained substantial interest in radiation oncology, particularly as a potential tool for automated care [29]. However, the development of ML-based applications faces challenges due to the lack of standardized data in terms of content, format, structure, and nomenclature. This problem can be addressed by utilizing an 'ontology that maps ML

models to their respective clinical data, creating a reproducible and transparent link.'

DLORO is an example of such an ontology that proves useful for this purpose. It is designed to automate the construction of Bayesian networks (BNs) [30] [31] [32]. As an ontology, DLORO provides the semantic properties necessary to map ontological concepts and relational schemas to databases. It serves as a standardized layer that connects data from various information systems (OISs) to the variables of the designed Bayesian network. This automation facilitates the extraction of data from OISs, thereby streamlining the training and updating of the network itself.

BNs consist of a graph [33] [34] [45], where nodes represent variables and arcs represent dependencies between them. Additionally, a probability table, known as a *conditional probability table* (CTP), expresses the strength of these dependency relationships. In the field of radiation oncology, BNs are employed for various applications, including *diagnostic reasoning*, biomedical data meta-analysis, modeling, and clinical decision support systems.

The DLORO ontology organizes radiation oncology knowledge into classes and subclasses, establishing dependencies between concepts. Additionally, it uses the *dependsOn* relationship to establish dependencies between concepts outside the same hierarchy. By selecting nodes of interest in the Bayesian Network (BN), such as variables, the class-subclass and *dependsOn* relationships can define the arcs of the BN, enabling its automatic realization. To complete the model, a probability table must be added, which requires training on data. Typically, data is manually extracted using SQL queries on Open Information Systems (OIS) databases. However, at this stage, DLORO can still be used to generate SQL queries. However, mapping between DLORO and each database to be used is necessary. These mappings should be added to DLORO as annotations for the mapped classes.

In cases where ontology classes do not match one-to-one with relational database columns, annotations that bind those classes to database variables must be manually added. Once again, queries can be automatically created based on these annotations.

In summary, then, by combining the SQL query generator and the BN generator, based on DLORO it is possible to implement Bayesian networks

automatically, based on a choice of variables desired by DLORO.

Table 1: Example of mapping between DLORO and the Aria (Arian Medical System, Palo Alto, USA) and Mosaik (Elekta AB, Stockholm, Sweden) databases [36].

Ontology	Mosaik Variable	Mosaik Table	Aria Variable	Aria Table
Anatomic_tumor_loc	Category	Medical	Site	Prescription
Dose_Per_Fraction	Dose_Tx	Site	PrescribedDose	RTPlan
Beam_Energy	Energy	TxFieldPoint	Energy	EnergyMode
Table_Angle	Couch_Ang	TxFieldPoint	PatientSupportAngle	ExternalFieldCommon

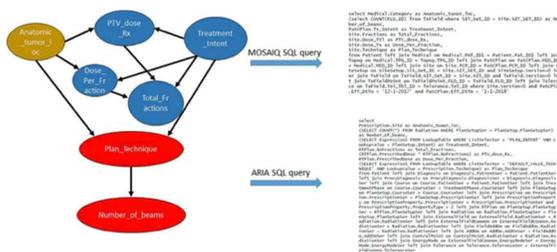


Figure 1: Example of Automatic Query Generation from the DLORO and the ARIA and Mosaik Databases, Based on the Chosen Variables.

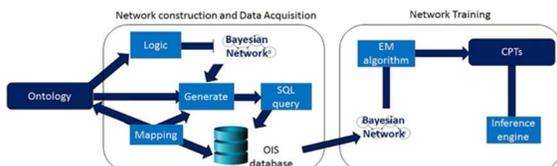


Figure 2: Flow Chart of BN Network Realization from an Ontology.

2.6 Text extraction from clinical notes

Information extraction (IE) is a branch of natural language processing (NLP) that focuses on automatically extracting specific, structured information from unstructured or semi-structured texts, such as documents, articles, emails, or web pages. In radiology, IE from unstructured clinical notes remains an unresolved challenge, although it can enhance the evaluation and improvement of radiotherapy treatments and support clinical decision-making or research [37].

The primary sources of unstructured clinical notes are *Electronic Healthcare Records* (EHRs), which are commonly used to collect, store, and display patient information. Currently, machine learning or rule-based methods are employed for IE from clinical notes. Rule-based approaches involve a domain expert identifying the necessary knowledge and rules for extracting information from specific types of notes. These approaches require less data and can be more easily adapted to new

domains. In contrast, machine learning-based approaches necessitate training data from which statistics or rules are automatically derived for use in information extraction. Alternatively, rules for IE could also be extracted through a domain-specific ontology, where rules are based on ontological concepts, which often perform better than rules based on textual elements. These types of rules are particularly useful in *Named Entity Recognition* (NER), whose objective is to identify specific words or phrases and categorize them.

There are several tools employed for information extraction in the clinical setting. Among the most widely used are cTAKES, MetaMap, and MedLEE. cTAKES is an open-source NLP pipeline that employs a hybrid approach combining rule-based and machine learning techniques. In contrast, MetaMap and MedLEE are purely rule-based tools. cTAKES offers a comprehensive platform for performing various information extraction tasks within the clinical domain, including syntactic, lexical, and semantic parsing. It utilizes clinical terminologies and ontologies, such as the SNOMED CT ontology graph [38] [39] [40] and RadLex, to extract concepts from medical texts. Additionally, the US ONC SHARPN project [41] [42] further enhanced clinical text mining by defining semantic standards that have been integrated into cTAKES. YTEX, a widely used extension of cTAKES, provides a general framework for mapping clinical sentences with any ontology.

MedLEE, a tool primarily used for developing vocabularies, was initially developed for processing torso radiology reports. However, its applications have expanded to other areas. Finally, MetaMap was introduced to map academic biomedical text with UMLS, employing symbolic techniques of natural language processing and computational linguistics. As mentioned, information extraction rules can be manually generated, derived from knowledge bases, or implemented in hybrid modes. Knowledge bases typically contain domain-specific ontologies. An example of a domain-specific ontology relevant to radiation oncology is PCD [43]. The concepts in PCD were developed by integrating the knowledge of clinicians and experts, analyzing the Electronic Health Record (EHR) database through formal and informal analysis of clinical terms, and incorporating medical ontologies from the UMLS. Consequently, PCD serves as a comprehensive domain-specific ontology for various information extraction tasks in radiation oncology. It not only represents concepts with related Semantic Groups but also provides a framework for patient clinical data contained in EHR databases. The PCD ontology represents concepts as

classes and *individuals*, where each class has a relationship to other classes and adheres to the HL7 standard for properties. Efforts to further integrate this ontology are ongoing, including integrating a more comprehensive set of EHR records and leveraging worldwide crowdsourcing activities.

Despite progress over the past decade, there are still many unresolved challenges in information extraction from clinical notes. The lack of well-formatted data to create annotated corpora hinders the effective training of pipeline-NLPs. Additionally, institutional barriers to data sharing further complicate the use of machine-learning approaches, particularly in radiation oncology. Identifying critical reports, disambiguation between body position and anatomical structure, proper mapping in DICOM files, and extracting specific abstract concepts are still open research problems. Metadata of radiology reports are often ignored or used implicitly due to the lack of methodologies to model them, which could aid in pipeline-NLP development. The field of word sense disambiguation research is not yet mature, especially in radiation oncology, where a standardized ontology is lacking. Ambiguity at the word or concept level significantly hinders the rapid annotation of text blocks using pipeline-NLPs. Currently, systems performing word meaning disambiguation rely on manually created rule sets and hand-labeled example sets. However, some automated methods have recently been proposed, including unsupervised learning, learning from UMLS [44], or using ontology-based semantic similarity measures. Further research is needed to reduce ambiguity at the word/concept level.

Still, open challenges encompass understanding abbreviations, linguistic errors, pronoun usage, and chronological comparisons with previously observed conditions. These challenges extend beyond traditional NLP tasks and are crucial for accurate information extraction.

2.7 The Standard Ontology

From our observations, developing a unified ontology shared by all stakeholders in healthcare could significantly enhance *computer-driven discovery and reasoning*. In 2017, during the Big Data Workshop, key data essential for computer-driven reasoning and a minimum set of relationships among these data were defined. These relationships and data are crucial for developing a taxonomy to organize them. The objective is to create an ontology containing concepts in a standardized format, developed in accordance with the principles outlined

by the OBO Foundry, and aligned with the emerging HL7-FHIR standards for data transmission. This approach ensures scalability in mapping clinical data to the ontology.

Furthermore, enabling the use of the standard ontology necessitates the adoption of recognized and approved nomenclatures by professional societies, such as those developed by the AAPM Working Group-263. Without these nomenclatures, the data necessary for the ontology to fully realize its potential would remain inaccessible. The AAPM is actively collaborating with ASTRO, ESTRO, COMP, and industry leaders to develop the standard ontology. Additionally, the AAPM is working with ASTRO, ESTRO, COMP, and CPQR to define taxonomies and nomenclatures required to enrich the standard ontology with data, such as those pertaining to disease status.

3. ONTOLOGIES AND SEMANTIC INTEROPERABILITY

3.1 Semantic interoperability

Semantic interoperability is achieved when computer systems communicating with each other exchange data with a specific syntax and their semantics. This is accomplished by enriching the data with metadata, which associates each data element with a link to a shared, controlled vocabulary. The meaning of the data is conveyed along with the data itself, creating a “self-descriptive package of information.” Ontologies are particularly useful in achieving semantic interoperability because they provide terms and relationships, and they can facilitate semantic interoperability through various mechanisms.

The creation of basic semantic representations. Ontologies enable the creation of basic semantic representations by defining classes and categories that organize concepts within a domain. At the class level, ontologies define general categories, while knowledge bases (at the instance level) represent concrete or specific examples of those categories. This allows systems to have a shared, structured understanding of information.

Definition of semantic mappings and transformations. Ontologies facilitate the creation of semantic mappings, that is, links between equivalent or related concepts found in different ontologies. These mappings enable the transformation and integration of data from different systems, ensuring that the meaning of information remains consistent

even when crossing boundaries between different data models.

Use of algorithmic methods for semantic similarity. Algorithms can be used to analyze ontologies and identify similarities between concepts. These methods help facilitate mapping between different ontologies, enabling easier integration and information exchange between systems using different terminologies or data structures.

The 2009 Ontology Summit [45] proposed a framework for utilizing ontologies, encompassing various aspects related to semantic interoperability. One of these areas is *information integration*, where ontologies facilitate the combination of multiple information resources. This enables the matching of concepts with similar meanings, thereby enhancing information aggregation and data fusion. Another area of application is software interoperability, which enables different systems to comprehend each other by exchanging messages using their own or shared ontologies.

3.2 Semantic Heterogeneity

Semantic interoperability is closely related to *semantic heterogeneity*. Semantic heterogeneity occurs when, across digital ecosystems, the same information is described with different schemas, vocabulary with varying and locally developed semantics, markup languages, and models based on different conceptualizations. This makes it difficult for information to be shared among software, for data from multiple sources to be combined, and for end users to interact effectively with the submitter. In fields like medicine, where collaboration between specialists and different departments is crucial, semantic heterogeneity poses a significant challenge to interoperability. Efficient use of clinical information could reduce costs and save thousands of lives annually. Additionally, a better experience for physicians and patients is an intangible benefit of using highly interoperable systems.

There are three types of interoperability: syntactic interoperability, schematic interoperability, and semantic interoperability. Syntactic and schematic interoperability can be achieved through the use of standard data sharing protocols and formats. Achieving semantic interoperability, however, is more complex. Simply sharing a controlled vocabulary is not sufficient; one must rely on a common interpretation of the exchanged messages and data, ensuring that their meaning remains unchanged during the exchange between digital domains and systems. Semantic

heterogeneity can be resolved through the use of ontologies.

3.3 Types of ontologies

Although ontologies are a helpful tool in resolving semantic heterogeneity and achieving interoperability, their diverse content, formal language, and detail can still lead to semantic chaos, also known as “semantic mess.” Therefore, it is crucial to use ontologies effectively. Ontologies can be categorized into three classes based on their intended usage: (i) *top-level* ontologies; (ii) *reference* ontologies; (iii) *application* ontologies; and (iv) *bridge* ontologies.

3.3.1 Top-level ontologies

Top-level ontologies are tools that facilitate the resolution of semantic heterogeneity by explicitly defining concepts shared by a diverse range of more specialized ontologies. By adopting a top-level ontology and its modeling practices, more specific ontologies can reduce the introduction of semantic inconsistencies. However, there are some challenges associated with the use of top-level ontologies.

- There are many top-level ontologies, and choosing which one to use can be difficult.
- Many top-level ontologies are complex, with complicated axioms and abstractions too far removed from real data.
- Top-level ontologies must be harmonized with domain, reference, and application ontologies.
- A top-level ontology developed in top-down mode could impose ontological assumptions that might have specific local vocabularies and meanings.
- Top-level ontologies can be too rigid, such that it is difficult to incorporate small changes without compromising their semantic coherence.

3.3.2 Reference Ontologies

Reference ontologies represent the knowledge base related to a specific domain and are designed for reuse. Unlike ontologies created for specific use cases, they are not rigidly bound to the requirements of a particular application. Their primary function is to facilitate integration between different systems, repositories, and data sources. In contrast to top-level ontologies that mediate between different ontologies, reference ontologies map terminology

from various information systems and data sources to a set of shared concepts. An example of a reference ontology is the Foundational Model of Anatomy (FMA), which represents the structure of the human body and consists of approximately 75,000 classes, 120,000 terms, and 168 types of relationships. Well-designed reference ontologies can be considered orthogonal (non-overlapping) and interoperable resources. The OBO Foundry proposes an ontology model specifically designed for reference ontologies.

3.3.3 Application Ontologies

Application ontologies are “ontologies specifically designed for a particular use case or application domain.” They are created to meet the requirements of a project and can be applied to either a local domain or multiple domains. However, when these ontologies are created without reference ontologies, it becomes challenging to link them to other ontologies. This limitation can hinder their usefulness in integration scenarios. To address this issue, mappings are employed, connecting the concepts of the ontologies to be integrated. Once these mappings are generated, they can also be utilized to create reference ontologies.

3.3.4 Bridge Ontologies

Bridge ontologies, similarly to what happens for reference ontologies, facilitate the mediation between terminologies in different systems. While mapping can resolve mediation between ontologies, it may require the addition of useful concepts to link related concepts present in the source ontologies. These concepts can be retrieved from bridge ontologies.

Although top-level ontologies serve as the primary link, bridge ontologies can better capture the similarities between different application and local ontologies within the same domain. In such cases, bridge ontologies act as a cohesive entity, integrating and harmonizing source ontologies. Unlike reference ontologies, bridge ontologies may not be as comprehensive and can sometimes be more concise, with fewer axioms. The use of a bridge ontology underscores the need for a reference or application ontology with broader requirements.

3.4 Methods to ease semantical interoperability

There are several useful methods to facilitate semantic interoperability. The main ones are: (i) ontology reuse and modular design; (ii)

harmonization of vocabularies; (iii) mapping between ontologies; and (iv) the use of design patterns for ontologies.

3.4.1 Reuse of ontologies and modular design

According to the Ontology Summit 2014, reuse involves incorporating content from one source into another or drawing inspiration from a source's content. While reusing concepts and semantics of an ontology can enhance semantic interoperability, it is not preferred over creating new ontologies for several reasons. First of all, existing ontologies are too large and complex, not sufficiently granular and documented. Furthermore, it is difficult to determine the useful aspects of an ontology. Finally, some semantic languages such as OWL, which are used to develop ontologies, do not support partial import of ontologies, which often forces the inclusion of many more concepts than are considered appropriate or correct.

Creating smaller, cohesive, and self-contained ontologies would significantly enhance interoperability because they can serve as the foundation for developing more intricate ontologies. However, implementing this approach today involves the challenging task of decomposing existing ontologies into more manageable building blocks, which is a complex process.

3.4.2 Vocabulary harmonization

When different organizations work on separate projects, they often develop vocabularies (sets of terms and definitions) that describe similar concepts using different terms and formats. This leads to confusion and hinders data integration and usage. Vocabulary harmonization is the process of standardizing these terms and definitions so that everyone can use the same language or, at least, compatible vocabularies. For instance, in water quality, vocabularies may combine multiple concepts (such as the substance measured, the environment, the method of measurement, and the units of measurement) into a single term. This makes data comparison across different systems challenging because it's unclear which part of the term refers to what.

To address this issue, some standardized vocabularies and their definitions have been made available online through Linked Data, using unique identifiers (URIs). This makes terms easier to access and understand because they are uniformly defined. Tools are available to collect these vocabularies,

facilitating easier access and management of definitions. However, despite these improvements, creating automatic links between different vocabularies remains challenging due to the lack of clearly defined relationships between concepts. In essence, despite progress, harmonization of vocabularies persists as a problem primarily because there are insufficient tools (standard schemas illustrating relationships between concepts in different vocabularies) and common ontologies that can serve as bridges between different vocabularies (bridge ontologies, reference ontologies, or top-level ontologies).

3.4.3 Mapping between ontologies

In the absence of a universal reference or top-level ontology, mappings between local ontologies used by various information systems can be established to achieve semantic interoperability. There is extensive and growing research on ontology mapping, particularly for OWL ontologies. However, there has been limited research on mapping ontologies written in more expressive languages like Common Logic. New languages such as *Distributed Ontology and Modelling and Specification Language* (DOL) are emerging and are expected to provide bases for mapping between ontologies written in the same language, as well as between ontologies written in different languages (e.g., between OWL ontologies and ontologies in Common Logic). Despite the existence of metalanguages and techniques for expressing ontology mappings, their automatic generation remains a significant challenge, especially for more complex languages.

3.4.4 Using design patterns for ontologies

In the 1970s, Christopher Alexander introduced the concept of *design patterns*, which are general solutions and shared guidelines for addressing recurring design problems in specific contexts. In 2009, Gangemi and Presutti [46] introduced *Ontology Design Patterns* (ODPs), which are modular, self-contained building blocks that represent reusable solutions to recurring problems in ontology modeling and development. ODPs help organize, name, and understand parts of knowledge that are closely related to each other. Essentially, an ODP is a predefined solution that can be applied to similar situations in different contexts, facilitating the creation and management of ontologies. When creating ODPs, it is useful to follow these guidelines.

First of all, a well-designed ODP is based on specific definitions and data, but is abstract enough to be applied generally. A pattern formed in this way can be used as the basis for building a more complex ontology. Over time, the community can enrich and specialize the basic pattern with additional axiomatic commitments. Sometimes, it is necessary to create subtypes of general patterns to respond to particular use cases and local complexities.

Second, it is important to carefully consider the domains and relationships within an ODP to avoid overly limiting the formal semantics of the pattern. A slightly axiomatized pattern can be linked to a complex ODP, which can be used as a framework for creating larger ontologies that address broader problems.

Finally, modular patterns are flexible and extensible, allowing local data and domains to be integrated effectively. In response to specific needs, local applications can extend the pattern by adding or modifying axioms and aligning them with local data and needs. In this process, the development of local ontologies becomes the link between a data source and patterns through specific and explicit mapping. To handle different and/or more refined interpretations, an appropriate mapping/alignment can be created between a local vocabulary and the pattern. This approach separates conceptualization from nomenclature issues, as local vocabulary terms may be specific to a data repository and need not be the same as those used to label pattern concepts.

A well-designed set of ODPs provides a useful tool for developing easily integrated ontologies.

4. CONCLUSIONS

In this paper, we presented and discussed the most important challenges related to the use of ontologies in the biomedical field, with particular emphasis on the oncological radiology domain. We also illustrated the opportunities and the benefits that a systematic use of ontologies would bring.

Our future work will root on this survey to improve the treatment of medical data, in particular narrative medical reports, in the context of the Italian National Health System.

REFERENCES:

- [1] Mohammad Amin Farvardin, Dario Colazzo, Khalid Belhajjame, Carlo Sartiani: Scalable Saturation of Streaming RDF Triples. Trans.

- Large Scale Data Knowl. Centered Syst. 44: 1-40 (2020)
- [2] Mohammad Amin Farvardin, Dario Colazzo, Khalid Belhajjame, Carlo Sartiani: Streaming saturation for large RDF graphs with dynamic schema information. DBPL 2019: 42-52
- [3] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Biportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37(suppl_2):W170–W173, 2009.
- [4] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. Biportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research, 39(suppl_2):W541–W545, 2011.
- [5] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270, 2004.
- [6] Nicole Bidoit, Dario Colazzo, Noor Malla, Carlo Sartiani: Evaluating Queries and Updates on Big XML Documents. Inf. Syst. Frontiers 20(1): 63-90 (2018)
- [7] Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Linear Time Membership in a Class of Regular Expressions with Counting, Interleaving, and Unordered Concatenation. ACM Trans. Database Syst. 42(4): 24:1-24:44 (2017)
- [8] Nicole Bidoit, Dario Colazzo, Carlo Sartiani, Alessandro Solimando, Federico Ulliana: Andromeda: A System for Processing Queries and Updates on Big XML Documents. ADBIS (Short Papers and Workshops) 2015: 218-228
- [9] Dario Colazzo, Giorgio Ghelli, Luca Pardini, Carlo Sartiani: Almost-linear inclusion for XML regular expression types. ACM Trans. Database Syst. 38(3): 15 (2013)
- [10] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, Stefanie Scherzinger: Validation of Modern JSON Schema: Formalization and Complexity. Proc. ACM Program. Lang. 8(POPL): 1451-1481 (2024)
- [11] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, Stefanie Scherzinger: Negation-closure for JSON Schema. Theor. Comput. Sci. 955: 113823 (2023)
- [12] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, Stefanie Scherzinger: Witness Generation for JSON Schema. Proc. VLDB Endow. 15(13): 4002-4014 (2022)
- [13] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, Stefanie Scherzinger: An Empirical Study on the "Usage of Not" in Real-World JSON Schema Documents. ER 2021: 102-112
- [14] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Yunchen Ding, Michael Fruth, Giorgio Ghelli, Carlo Sartiani, Stefanie Scherzinger: A Test Suite for JSON Schema Containment. ER Demos/Posters 2021: 19-24
- [15] Dario Colazzo, Giorgio Ghelli, Luca Pardini, Carlo Sartiani: Efficient asymmetric inclusion of regular expressions with interleaving and counting for XML type-checking. Theor. Comput. Sci. 492: 88-116 (2013)
- [16] Nicole Bidoit, Dario Colazzo, Noor Malla, Federico Ulliana, Maurizio Nolé, Carlo Sartiani: Processing XML queries and updates on map/reduce clusters. EDBT 2013: 745-748
- [17] Nicole Bidoit, Dario Colazzo, Noor Malla, Carlo Sartiani: Partitioning XML documents for iterative queries. IDEAS 2012: 51-60
- [18] Dario Colazzo, Carlo Sartiani: Precision and complexity of XQuery type inference. PPDP 2011: 89-100
- [19] Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Schemas for safe and efficient XML processing. ICDE 2011: 1378-1379
- [20] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Francesco Falleni, Giorgio Ghelli, Cristiano Landi, Carlo Sartiani, Stefanie Scherzinger: A Tool for JSON Schema Witness Generation. EDBT 2021: 694-697
- [21] Michael Fruth, Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, Stefanie Scherzinger: Challenges in Checking JSON Schema Containment over Evolving Real-World Schemas. ER (Workshops) 2020: 220-230
- [22] Mohamed-Amine Baazizi, Clément Berti, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Human-in-the-Loop Schema Inference for Massive JSON Datasets. EDBT 2020: 635-638
- [23] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: A Type System for Interactive JSON Schema Inference (Extended Abstract). ICALP 2019: 101:1-101:13

- [24] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Schemas and Types for JSON Data: From Theory to Practice. SIGMOD Conference 2019: 2060-2063
- [25] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Schemas And Types For JSON Data. EDBT 2019: 437-439
- [26] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Parametric schema inference for massive JSON datasets. VLDB J. 28(4): 497-521 (2019)
- [27] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Counting types for massive JSON datasets. DBPL 2017: 9:1-9:12
- [28] Mohamed-Amine Baazizi, Housseem Ben Lahmar, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani: Schema Inference for Massive JSON Datasets. EDBT 2017: 222-233
- [29] Alan M Kalet, Samuel MH Luk, and Mark H Phillips. Radiation therapy quality assurance tasks and tools: the many roles of machine learning. *Medical physics*, 47(5):e168–e177, 2020.
- [30] Alan M Kalet, John H Gennari, Eric C Ford, and Mark H Phillips. Bayesian network models for error detection in radiotherapy plans. *Physics in Medicine & Biology*, 60(7):2735, 2015.
- [31] Alan M Kalet, Jason N Doctor, John H Gennari, and Mark H Phillips. Developing bayesian networks from a dependency-layered ontology: a proof-of-concept in radiation oncology. *Medical Physics*, 44(8):4350–4359, 2017.
- [32] Samuel MH Luk, Juergen Meyer, Lori A Young, Ning Cao, Eric C Ford, Mark H Phillips, and Alan M Kalet. Characterization of a bayesian network-based radiotherapy plan verification model. *Medical physics*, 46(5):2006–2014, 2019.
- [33] Dario Colazzo, Vincenzo Mecca, Maurizio Nol , Carlo Sartiani: PathGraph: querying and exploring big data graphs. SSDBM 2018: 29:1-29:4
- [34] Maurizio Nol , Carlo Sartiani: Regular Path Queries on Massive Graphs. SSDBM 2016: 13:1-13:12
- [35] Maurizio Nol , Carlo Sartiani: A Distributed Implementation of GXPath. EDBT/ICDT Workshops 2016
- [36] Mark H Phillips, Lucas M Serra, Andre Dekker, Preetam Ghosh, Samuel MH Luk, Alan Kalet, and Charles Mayo. Ontologies in radiation oncology. *Physica Medica*, 72:103–113, 2020.
- [37] Sima Ajami and Tayyebe Bagheri-Tadi. Barriers for adopting electronic health records (ehrs) by physicians. *Acta Informatica Medica*, 21(2):129, 2013.
- [38] Dario Colazzo, Carlo Sartiani: Typing regular path query languages for data graphs. DBPL 2015: 69-78
- [39] Maurizio Nol , Carlo Sartiani: Processing Regular Path Queries on Giraph. EDBT/ICDT Workshops 2014: 37-40
- [40] Dario Colazzo, Carlo Sartiani: Typing query languages for data graphs. ICDE Workshops 2014: 28-31
- [41] Jyotishman Pathak, Kent R Bailey, Calvin E Beebe, Steven Bethard, David S Carrell, Pei J Chen, Dmitriy Dligach, Cory M Endle, Lacey A Hart, Peter J Haug, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the sharpn consortium. *Journal of the American Medical Informatics Association*, 20(e2):e341–e348, 2013.
- [42] Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A Oniki, Les Westberg, Calvin E Beebe, Cui Tao, Craig G Parker, Peter J Haug, Stanley M Huff, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of ehr data: the sharpn project. *Journal of biomedical informatics*, 45(4):763–771, 2012.
- [43] Hussein Boshnaka, Sayed Abdel-Gaber, and Engy Yehiad Amany-Abdoc. Ontology-based knowledge modeling for clinical data representation in electronic health records. *Int J Comp Sci Info Sec (IJCSIS)*, 16(10), 2018.
- [44] Adam Wright, Elizabeth S Chen, and Francine L Maloney. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics*, 43(6):891–901, 2010.
- [45] Donna Fritzsche, Michael Gr ninger, Ken Baclawski, Mike Bennett, Gary Berg-Cross, Todd Schneider, Ram Sriram, Mark Underwood, and Andrea Westerinen. Ontology summit 2016 communiqu : Ontologies within semantic interoperability ecosystems. *Applied Ontology*, 12(2):91–111, 2017.
- [46] Aldo Gangemi, Valentina Presutti, *Ontology Design Patterns. Handbook on Ontologies 2009*: 221-243