The radiation oncology ontology (ROO)

Citation for published version (APA):

Traverso, A., van Soest, J., Wee, L., & Dekker, A. (2018). The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. *Medical Physics*, 45(10), E854-E862. <https://doi.org/10.1002/mp.12879>

Document status and date:

Published: 01/10/2018

DOI:

[10.1002/mp.12879](https://doi.org/10.1002/mp.12879)

Document Version:

Publisher's PDF, also known as Version of record

Document license:

Taverne

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques

Alberto Traverso,^{a)} Johan van Soest,^{a)} Leonard Wee, and Andre Dekker

Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht 6062 NA, the Netherlands

(Received 26 October 2017; revised 26 January 2018; accepted for publication 17 February 2018; published 24 August 2018)

Purpose: Personalized medicine is expected to yield improved health outcomes. Data mining over massive volumes of patients' clinical data is an appealing, low-cost and noninvasive approach toward personalization. Machine learning algorithms could be trained over clinical "big data" to build prediction models for personalized therapy. To reach this goal, a scalable "big data" architecture for the medical domain becomes essential, based on data standardization to transform clinical data into FAIR (Findable, Accessible, Interoperable and Reusable) data. Using Ontologies and Semantic Web technologies, we attempt to reach mentioned goal.

Methods: We developed an ontology to be used in the field of radiation oncology to map clinical data from relational databases. We combined ontology with semantic Web techniques to publish mapped data and easily query them using SPARQL.

Results: The Radiation Oncology Ontology (ROO) contains 1,183 classes and 211 properties between classes to represent clinical data (and their relationships) in the radiation oncology domain following FAIR principles. We combined the ontology with Semantic Web technologies showing how to efficiently and easily integrate and query data from different (relational database) sources without *a priori* knowledge of their structures.

Discussion: When clinical FAIR data sources are combined (linked data) using mentioned technologies, new relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing. © 2018 American Association of Physicists in Medicine [https://doi.org/10.1002/mp.12879]

Key words: ontologies, radiation oncology, semantic web

1. INTRODUCTION

1.A. Motivation

Data-driven medicine has the potential to yield improved health outcomes¹ and is an integral component of value-based healthcare.² One of the biggest challenges for data driven medicine is to access and analyze clinical data with machine learning techniques to predict clinical outcomes combining all available information. Subsequently, these developed machine learning techniques can be used to build decision support systems for clinicians.

The current obstacle to be addressed is the availability of outcome information (e.g., tumor control and treatment-related toxicity) that must be provided to "train" the machine learning models. Many models have been built based on data from clinical trials. However, clinical trials recruit only a small part of the presenting cases, therefore questions about applicability to under-represented patient subgroups persist. In contrast, clinical data derived from routine care are known to have data quality issues (e.g., a high rate of missing values).

To overcome the potential sensitivity to missing values as well as to provide sufficient training samples for machine learning, a scalable "big data" architecture for the medical domain becomes essential. For such a scalable architecture, data standardization is imperative. In particular, clinical data should be transformed following FAIR (Findable, Accessible,

Interoperable, and Reusable) principles.³ To make healthcare data FAIR, Ontologies and Semantic Web technologies play a key role, and hence will address the issue of semantic interoperability. In⁴, the authors exploited the possibility to use ontological technologies to enable semantic interoperability with data coming from multicenter postgenomics clinical trials. In⁵, the authors focused on applying Semantic Web technologies to the medical imaging domain, developing an ontology for medical image annotations. In⁶, the authors investigated the possibility to use Semantic Web technology to store and represent metadata from DICOM image files. Both the studies showed the potential of ontologies technologies in allowing medical data interoperability.

However, the usage of ontologies and Semantic Web technologies applied to the field of radiation oncology are limited. In⁷, the authors converted clinical data of prostate cancer patients from a local database using a dedicated ontology, but they did not exploit the possibility to merge different datasets from different diseases combining different ontologies. In⁸, the authors stressed the concept of standardization of collected data (in rectal cancer) using ontological techniques to allow machine learning algorithms to build clinical prediction models.

In addition, they strongly suggested using Semantic Web technologies in order to allow data sharing while respecting the privacy protection of individual patients. Finally, ontologies and Semantic Web techniques represent the required

infrastructure for distributed learning.⁹ Compared to traditional centralized learning approach, in distributed learning clinical data do not leave the hospital, but after being transformed into FAIR, they are queried during the training of the model, while the model is “learned” from different centers.

Conversely, when looking at the radiation oncology domain, we could not find any study aiming at: (a) developing and validating a broader ontology to be used in the radiation oncology domain; (b) combining ontology and Semantic Web techniques to transform different clinical databases into FAIR and linked data.

The role of the ROO is to provide a detailed and broad coverage of main concepts used in the radiation oncology domain such as: classification of neoplasms, patients’ demographic characteristics; as well as clinical information like tumor’s classification or treatment. The ontology is strongly focused on re-using published ontologies and/or terminologies. The added value of the ROO is to re-used published ontologies/terminologies by defining new predicates, which establish relations between imported concepts. Combining different terminologies and expanding relationships between them is the path to guarantee the largest coverage.

The ROO allows transforming unstructured clinical data from following FAIR principles. In particular, data will become

- Findable (F): each data entity and their properties (F2), translated into universally concept via the ROO will have a globally unique identifier (F1) and will be indexed on the Web (F3). Metadata will include specification of the data identifier (F4)
- Accessible (A): data will be retrievable by means of RDF triples and queryable using a universal language (A1). A permanent de-centralized storage point will be permanently available (A2), even when the original database could not be anymore.
- Interoperable (I): data are represented by universally adopted RDF language (I1). Queries rely on concept from imported ontologies/vocabularies that follow FAIR principles (I2).
- Reusable (R): several attributes specific data properties and the relations between different concepts via ROO predicates (R1).

In this paper, we: (a) developed a broad ontology to cover the domain of radiation oncology; (b) combined ontology and semantic web techniques to transform clinical data from different disconnected databases into FAIR and linked data, allowing the discovery of new relationships.

1.B. Terminologies, vocabularies, and ontologies

Before going into the details of ontologies’ structure and properties, we provide the reader with some fundamentals regarding: terminologies, thesauri, vocabularies, and ontologies. Usually, a terminological system¹⁰ is an umbrella terms including the notions of: terminologies, thesauri,

vocabularies, and ontologies. Complexity increases from terminologies to ontologies:

- Terminology: a list of term referring to concept within a particular domain. For example, in the radiation oncology domain, concepts such as “patient” or “disease”. The terminology can be seen as a list of concepts, but without providing any definition or introducing any structures/relations between the terms.
- Thesauri: a thesauri is a terminology, where concepts are indexed according to a certain rule (usually alphabetically). Example of a thesauri is the International Classification of Diseases (ICD), which includes generic-related diagnostic terms (terminology), order alphabetically (thesauri).
- Vocabulary: in a vocabulary, indexed concepts are accompanied by a definition.

Conversely, an ontology is an explicit formal specification of the terms in the domain and relations among them¹¹ expressed in machine-readable language; therefore, they can be processed automatically. An ontology adds more complexity than a dictionary, since it explicitly defines the relationship, i.e., predicates, between unique entities. Classes (i.e., concepts), subclasses, and predicates between concepts represent an ontology. Inference rules (also called automated reasoning) in ontologies supply further knowledge, since (new) relationships between concepts, which can be discovered, since not formally defined *a priori*. An ontology is commonly used to model consensus in understanding a domain between different partners (e.g., different medical centers). Major advantages of ontologies are: (a) sharing common understanding; (b) re-using of domain knowledge, analyzing domain knowledge, and (c) inferring new knowledge starting from relationship between defined concepts. The standard for developing ontologies is the Web Ontology Language (OWL) as recommended by the W3C (World Wide Web Consortium) to represent ontologies.¹²

1.C. Semantic web technologies

Semantic Web is not a separate Web, but an extension of the current one, in which computers primarily interpret the data instead of humans. The current web provides rich-media content (e.g., written text, images, video’s,) which is not easy to interpret for computers. In the Semantic Web extension, the information is represented in well-defined structures and semantics in order to enable automated processing of the contents by computers.¹³ Hence, it can function as a computer representation of already available web content, next to the human-readable web content.

For the Semantic Web to function, computers must have access to structured collections of information. The basic building blocks are therefore provided by the Resource Description Framework (RDF) and the “SPARQL Protocol And RDF Query Language” (shorthand: SPARQL). Both RDF and SPARQL build on the existing web components of URIs and HTTP. URIs are the links to the actual resources, and can be represented as URLs (e.g., <http://mydomain.com/>)

rdf/patient/12345). These URIs are used to represent nodes (resources) and arcs (predicates) in the RDF graph. HTTP is used to publish RDF information on the web or to perform SPARQL queries on RDF stores. These RDF stores (also called SPARQL endpoints) are webpages which can be queried using the HTTP protocol. Most of these stores/endpoints also have human-readable web interfaces.

By using RDF as a universal graph data structure, the Semantic Web relies on ontologies to give domain-specific structure and interpretation to the represented data. In these ontologies, hierarchies of concepts can be defined, as well as relationships between certain concepts; all written in RDF. It is a common practice to add human-readable attributes to the URIs, as it enables the creation of human-readable views on an RDF endpoint. By creating instances of concepts defined in the ontology, users can create graphs of data for representing real-life concepts (e.g., “<http://mydomain.com/rdf/patient/12345> rdf:type <http://mydomain.com/ontology/patient>”) where the resource 12345 is an instance of the class patient).

In addition, ontologies can describe inferencing rules which are interpretable by inferencing-enabled RDF stores. In these stores, it is possible to query or show the inferred information, which is not hard-coded (or materialized) in the RDF store. Hence, updating inferencing rules in the ontology would enable users to query or show additional information without updating the RDF store itself.

This allows to uncover additional relationships in the actual data, and accommodates searches on different levels of data (e.g., patients are persons; therefore, searching for persons will include all patients in the database).

2. MATERIALS AND METHODS

2.A. Clinical database

We used a clinical database of oncological patients with a diagnosed rectal cancer from the THUNDER trial.¹⁴ The goal of the trial was to develop a prediction model of rectal tumor response after chemo-radiotherapy that might be helpful in individualizing treatment strategies, i.e., selecting patients who need less invasive surgery or another radiotherapy strategy instead of resection. The database includes 80 patients and contains a diverse range of information, combining demographic and clinical outcomes. Due to its heterogeneous nature, it represents a good validation for the ROO. The ROO was applied to convert each values in the database, mapping them through the concepts available in the ontology. Relations between individuals were mapped using a graph structure. The graph output was then transformed into RDF triples, published on a dedicated endpoint and queryable, in line with FAIR principles.

2.B. Radiation oncology ontology (ROO) development

We developed a radiation oncology ontology (ROO). The ontology was designed using the editor tool Protégé¹⁵ and publically published at the NCBO BioPortal (<https://>

bioportal.bioontology.org/ontologies/ROO). The ROO adheres to the Ontology Web Language (OWL) 2 Query Language (QL) profile (<http://www.w3.org/TR/owl2-profiles/>). The ontology provides basic concepts, relationships, and properties/attributes for radiation oncology.

The ontology was built following this procedure: (a) we identified variables of interest by collecting concepts and their definitions within the ontology using different datasets coming from several institutions belonging to the Euro CAT projects.¹⁶ Due to its multicenter nature, we could allow a broad coverage for different diseases with the aim of making the ontology as much detailed as possible; (b) we published and made publicly available on BioPortal several versions of the ontology during its development. This choice allowed users downloading, using and testing our ontology.

In addition, a dedicated section on GitHub permitted users highlighting inconsistencies and/or requiring enhancements. In this way, our ontology became a dynamic body of knowledge with the aim of guaranteeing the broadest possible coverage for the radiation oncology domain.

The high-level structure of the ROO is based on the Unified Medical Language System (UMLS) Semantic Network by the Semantic Types (classes) ontology (<http://bioportal.bioontology.org/ontologies/STY/?p=summary>) and the assertion of the Semantic Relations (properties) as specified by the UMLS (<https://uts.nlm.nih.gov/>).

The ROO re-uses as much as possible entities from other ontologies such as the National Cancer Institute (NCIT) Thesaurus or the International Classification of Disease (ICD) ontologies. The ROO makes only use of ontologies published at NCBO's BioPortal and provided without any restrictions. Common re-used ontologies were: NCIT (National Cancer Institute Thesaurus); Units of Measurement Ontology (UO); Foundational Model of Anatomy (FMA); Semantic Types Ontology (STY); Semantic DICOM Ontology (SEDI), and International Classification of Diseases, Version 10 (ICD10). The ROO uses the original Unique Resource Identifiers (URIs) for these imported entities: an example is the concept of lung cancer that is inherited using the concept code C34 from the ICD ontology.

2.C. Ontology validation

2.C.1. Mapping between database schemas and ontology

One of the most important test to validate the ontology is guaranteeing that every element in a clinical relational database and its properties can be fully mapped with respectively the concepts and predicates in the ontology. The basic idea of the mapping process is linking each component (row, columns, and values) of the database to its corresponding component (concept, property, relationship) of the ROO. The preliminary step is to identify a correspondence between the columns in the relational database and the ontology entities. A sketch representation of the mapping procedure is shown in Fig. 1.

FIG. 1. Overview of the ROO structure and the relational database. The hierarchical structure of the ROO is presented in the rectangle a. Hierarchical Relationships (“is subclass of”) between classes, are expressed by dotted arrows. Mapping is performed to columns and values in a relational database (rectangle b).

At the top, the hierarchical structure of the ROO is presented in the rectangle A. Hierarchical Relationships (“is subclass of”) between classes, are expressed by dotted arrows. These relationships between more general classes (parents) and more specific classes (children) represent the ontology backbone since they allow properties inheritance. ROO concepts are expressed inside blue squares. Relationships between concepts (predicates) are expressed with arrows: they connect classes between each other. For example, patient and gender classes are connected by the property “has_gender”.

A sample table of one of the datasets is shown in the rectangle B. This table contains information about patient demographics (e.g., sex, age) as well as diagnosis (e.g., survival, tumor staging). The mappings are built between the table columns and the concepts in the ROO (shown as bold dotted double-headed arrows in the figure). For example, the column “Gen” is mapped to the concept gender (ncit:C17357) in the ROO. The link between a patient and the sex is made by the property “has_gender”.

Several languages and software tools are available to perform the mapping procedure from relational databases to RDF triples.¹⁷ We performed the mapping between the clinical data and the ontology using the D2RQ mapping language. D2RQ mapping language is a declarative language for mapping relational database schemas to RDF vocabularies and OWL ontologies. The language is read and interpreted by the D2RQ platform, which is written in Java and open-source available. We decided to use D2RQ because it represents one of the most common tools for database transformation from

relational database to network structures.¹⁸ In addition, the language is modular, easily allowing to link entities from the database to concepts and properties in ROO. The mapping defines a virtual RDF graph that contains instructions how to connect and map the information from the relational database. This is similar to the concept of views in SQL databases, except that the virtual data structure is an RDF graph instead of a virtual relational table. The mapping file, written in turtle (.ttl) syntax, contains the mapping between the database schema and the concepts defined in the ontology. The turtle syntax is the format for expressing data as RDF triples, then queryable using a dedicated language.

An example of the mapping file is shown in Fig. 2. The mappings between table columns and their corresponding concepts are created using the command `d2r:ClassMap`. The mapping between the table columns to their corresponding properties is performed by using the command `d2rq:PropertyBridge`. In addition, in the mapping script each entity is associated with a Unique Resource Identifier (URI) to facilitate publishing on the Semantic Web and data linking. In the example, the entity patient is mapped to the concept C16960 from the NCIT. The bridge between a patient and his/her gender is mapped through the predicate 100018 (“has_gender”) from the ROO.

2.C.2. Publishing and querying data on the semantic web

The mapped data, transformed into URIs, are then stored in a RDF store, which is web-enabled (HTTP) and can be

```

# PATIENT TABLE

map:patient a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:class nci:C16960; #patient
  d2rq:uriPattern
  "patient_@@derived_multidelineations.identifier@";

# PROPERTY BRIDGE FOR PATIENT

map:patient_label a d2rq:PropertyBridge;
  d2rq:BelongsToClassMap map:patient;
  d2rq:property rdfs:label;
  d2rq:column
  "derived_multidelineations.collection_identifier";
  d2rq:datatype xsd:String;

#CREATE NOW GENDER OBJECT

map:patient_gender_obj a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern
  "gender_@@derived_multidelineations.identifier@";
  d2rq:condition "derived.gender IS NOT NULL";

# Link to the gender object of the NCI thesaurus

map:patient_gender_uri_obj a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:patient;
  d2rq:property roo:100018; #has_gender
  d2rq:refersToClassMap map:patient_gender_obj;

```

FIG. 2. Example of the D2RQ mapping script. The first block (patient table) defines the mapping for each patient ID. The “ClassMap” property in a D2RQ script defines a mapping between a header in the relational database and the corresponding concept in the ontology. A “Property Bridge” is used in a D2RQ script to express relations between different concepts. In the example above, the “Property Bridge” has_gender is used to link the patient concept to his/her gender.

queried using SPARQL. Making these RDF stores web-enabled means that it is available internally or externally on a specific network, in the same way as webpages are. This does not per definition means that data are publically available, only that existing web techniques are used to represent semantically interoperable data. In our work, we used Blazegraph (www.blazegraph.com) as our RDF store (or SPARQL endpoint).

3. RESULTS

3.A. Radiation oncology ontology (ROO)

The ROO contains 1,183 classes, with an average number of four children per class; two classes have more than 25 children. The classes cover the most common concepts in radiation oncology, including cancer diseases, cancer-staging systems, and oncology treatments.

Besides the classes, 211 predicates are introduced to express relationships between different classes. We divided the properties into five big categories: (a) conceptually_related_to; (b) functionally_related_to; (c) physically_related_to; (d) spatially_related; and (e) temporally_related_to. Examples of mentioned categories are respectively: (a) diagnosed_by; (b) has_result; (c) connected_to; (d) has_location; (e) follows.

All entities and predicates in the ROO have a URI, which can be resolvable as a link since hosted on www.cancerdata.org. A web RDF viewer allows the users inspecting a concept by typing on an internet browser the address [www.cancerdata.org/roo/\[URI\]](http://www.cancerdata.org/roo/[URI]), where URI is the URI of the ontology entity. For example, the user will type www.cancerdata.org/roo/100287 for the predicate “has_pathological_stage”. In addition, the users are able to transverse the full tree of the ontology through the Web RDF viewer. The latest version of the ontology has been published on BioPortal, totally Open Source and available for the user to download. The ROO is available in the most common format, including OWL, which can be opened by the users using the software Protégé’.

3.B. Ontology validation

3.B.1. Mapping between database schemas and ontologies

A wiki page on how to perform the mapping between relational database schemas and the ontology is publically available on the GitHub (<https://github.com/jvsoest/Data-Integration-Tutorial/wiki/conversionClinicalData>). The users can follow the guide to convert part of the Thunder dataset into RDF triples with the ROO using the example scripts provided.

3.B.2. Query formulation

After having mapped the data, it is possible to query them using SPARQL language. Users could query the data without having any prior knowledge of the relational database, since SPARQL queries are based on universal concepts defined by the ontology.

Following the example in the Wiki (<https://github.com/jvsoest/Data-Integration-Tutorial/wiki/queryClinicalData>), let us suppose we want to search all the patients with rectal cancer and retrieve following information: age at diagnosis, ECOG (Eastern Cooperative Oncology Group) performance status score, clinical TNM stage, pathological TNM status, and prescribed dose in Gray.

The example query is available at <https://github.com/jvsoest/Data-Integration-Tutorial/blob/master/queries/queryClinicalData.sparql> and it is shown in Fig. 3(a). The system returns all the patients and displays the results in the SPARQL result window on the web browser. Each object shown is associated with an URI, universally and unambiguously defining it when published on the Web. Furthermore, all triple patterns to find a certain variable are grouped in curly brackets. This creates the opportunity to make some variables optional or to specify some filters. For example, we could have asked for patients with an age at diagnosis below a certain value, by modifying the original query with a filter [highlighted in blue in Fig. 3(b)].

```

1 prefix roo: <http://www.cancerdata.org/roo/>
2 prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
3 prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4 prefix icd: <http://purl.bioontology.org/ontology/ICD10/>
5 prefix uo: <http://purl.obolibrary.org/obo/uo/>
6
7 SELECT ?patient ?gender ?ageDiagnosis ?clnT ?clnN ?ecogStatus ?presDoseGy
8 WHERE {
9   ?patient rdf:type ncit:C16968.
10  ?patient roo:100008 ?disease.
11  ?disease rdf:type icd:C20.
12
13  ?patient roo:100301 ?rtRes.
14  ?rtRes rdf:type ncit:C15313.
15  ?rtRes roo:100402 ?disease.
16
17  ?patient roo:100010 ?genderRes.
18  ?genderRes rdf:type ?gender.
19
20  # Get age at diagnosis
21  {
22    ?patient roo:100016 ?ageResDiagnosis.
23    ?ageResDiagnosis rdf:type roo:100002.
24    ?ageResDiagnosis roo:100027 ?ageDiagnosisUnitRes.
25    ?ageDiagnosisUnitRes rdf:type uo:0000036.
26    ?ageResDiagnosis roo:100042 ?ageDiagnosis.
27  }
28
29
30  # Get ECOG performance status
31  {
32    ?patient roo:100218 ?ecogRes.
33    ?ecogRes rdf:type ?ecogStatus.
34  }
35
36  # Get clinical TNM values
37  {
38    ?disease roo:100243 ?clnTnnRes.
39    ?clnTnnRes rdf:type ncit:C40001.
40    ?clnTnnRes roo:100244 ?clnTRes.
41    ?clnTRes rdf:type ?clnT.

```

```

1 prefix roo: <http://www.cancerdata.org/roo/>
2 prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
3 prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4 prefix icd: <http://purl.bioontology.org/ontology/ICD10/>
5 prefix uo: <http://purl.obolibrary.org/obo/uo/>
6
7 SELECT ?patient ?gender ?ageDiagnosis ?clnT ?clnN ?ecogStatus ?presDoseGy
8 WHERE {
9   ?patient rdf:type ncit:C16968.
10  ?patient roo:100008 ?disease.
11  ?disease rdf:type icd:C20.
12
13  ?patient roo:100301 ?rtRes.
14  ?rtRes rdf:type ncit:C15313.
15  ?rtRes roo:100402 ?disease.
16
17  ?patient roo:100010 ?genderRes.
18  ?genderRes rdf:type ?gender.
19
20  # Get age at diagnosis
21  {
22    ?patient roo:100016 ?ageResDiagnosis.
23    ?ageResDiagnosis rdf:type roo:100002.
24    ?ageResDiagnosis roo:100027 ?ageDiagnosisUnitRes.
25    ?ageDiagnosisUnitRes rdf:type uo:0000036.
26    ?ageResDiagnosis roo:100042 ?ageDiagnosis.
27    FILTER (?ageDiagnosis <= '75'^^xsd:integer).
28  }
29
30  # Get ECOG performance status
31  {
32    ?patient roo:100218 ?ecogRes.
33    ?ecogRes rdf:type ?ecogStatus.
34  }
35
36  # Get clinical TNM values
37  {
38    ?disease roo:100243 ?clnTnnRes.
39    ?clnTnnRes rdf:type ncit:C40001.
40    ?clnTnnRes roo:100244 ?clnTRes.
41    ?clnTRes rdf:type ?clnT.

```

FIG. 3. (a) On the left, example query without any filter; (b) on the right, example query introducing a filter on the diagnosis age. The query is written using the SPARQL language. Lines 1 to 5 are used to import the required ontologies. The query starts from line 7 (select query) asking for following information: patients gender, age at diagnosis, tumor T and N stages, overall health status, and prescribed dose. Then, each variable is queried in different blocks by making use of ROO concepts and predicates.

Finally, this example query can be used directly in programming languages/statistical languages to request valid data matrices. For example, in R using the SPARQL package, or in any other language using a representational state transfer (REST) interfacing package. Results from query shown in Fig. 3, where compared with data available in the original database to verify the correctness of the mapping. Data comparison and visualization were performed and no differences were found when comparing the information available in the database with respect to the one available as SPARQL queries. The advantage with respect to a standard excel file, is that RDF data could be queried without any knowledge of the original data structures, by mean of SPARQL queries based on universal concepts defined by the ROO.

3.C. Combining different databases: linked data

One of the biggest benefits of Semantic Web and ontology technologies is the possibility to query different databases and make connections within them. For example, in radiation oncology it can be interesting for clinicians to investigate some properties (e.g., survival) of patients: (a) with a certain disease AND (b) treated according to a predetermined protocol AND (c) associating the publications of the clinical trial related to the protocol.

Performing such a query using traditional relational databases is a real issue, since it not only requires combining different databases, but also a prior knowledge of their schemas. We solved the problem using ontologies and Semantic Web technologies.

3.C.1. Query formulation

In particular, we made use of Bio2RDF: an open-source project that uses Semantic Web technologies to build and provide the largest network of linked data for the life sciences.¹⁹ It contains among others the RDF versions of ClinicalTrials.gov and PubMed.

In our query, the first part is equal to the query provided in Section 3.B.2. This query retrieves the patients available in the RDF store, and characteristics of these patients (e.g., age, gender, ECOG performance status), their disease (e.g., tumor classification), and the prescribed treatment. Based on this information, we linked the patients to matching treatment protocols, as we defined the protocols and linked them to the correct ClinicalTrials.gov entry in Bio2RDF.

Afterwards, the query contains a section to query the ClinicalTrials.gov linked data representation from Bio2RDF, and a URL generation for a PubMed query. To link the clinical information to public ClinicalTrials.gov (CTgov) information,

we used the prescribed treatment variable (containing a unique URL) which was available on both internal (clinical) data, and the Bio2RDF CTgov linked data. From the CTgov linked data, we queried in which trials the same treatment protocol URLs were used. From this relation, we could retrieve information regarding the specific clinical trials, such as the CTgov identifier, the time period when the trial was conducted, which institutes were involved, and trial contact persons. Based on the CTgov identifier, we generated a link to the related manuscripts which have been indexed in PubMed.

The full query to run this linked data example is available at <https://gist.github.com/jvsoest/eb015abfb0efd5c669fd36915ce2487d>. For example purposes, this query can be executed at <http://sparql.cancerdata.org/>.

4. DISCUSSION

4.A. Rationale for the ROO

Patients' demographics and clinical information are important for radiation oncology prediction/modeling studies. In particular, it is of interest of the radiation oncology community to explore the maximum amount of available clinical data to improve semantic interoperability during patient referral, and for models aiming at predicting outcomes such as overall survival or toxicities after a treatment.

To reach this goal, data integration from different sources (internal/external relational databases) becomes a key factor, since most of the data are usually located in different relational databases.

Since relational databases can present different structures, querying them to access information without having a prior knowledge of the structures becomes a real issue. To tackle this issue, there is the need to transform clinical data following FAIR principles.³ Ontologies and Semantic Web technologies could represent the right choice to achieve this goal.

We developed the Radiation Oncology Ontology (ROO) with the aim to provide an ontology of use within the radiation oncology field to be used to transform clinical data following FAIR principles.

4.B. Advantages of ontologies and semantic web data integration compared to relational databases

As presented in previous sections, the ROO has been used to transform clinical traditional database schemas into graph databases relying on ontologies.

There are some differences between graph and database schemas. First, ontologies represent a domain on knowledge. Conversely, database schemas are conceived for (and linked to) particular applications, making their structures very diversified and difficult to be made interoperable. In fact, only users knowing the schemas structure (usually the owner of the data) can easily access them.

On the contrary, ontologies transform data into universal concepts that can be queried by the users using SPARQL, without knowing the structures of the data themselves. In fact, data are transformed on universal concepts defined by the ontology itself, and available using URIs (and URLs).

The usage of ontologies adds to transforming data from database schemas into FAIR data. An ontology, combined with Semantic Web technologies, is a stable conceptual interface on top of the relational database system. In fact, it can be scaled for data integration among multiple domains.

Individual database schemas are mapped to the concepts of the ontology and it is relatively easy to integrate new database systems (when mapped/converted into Semantic Web data). The only modification required would be to update the mapping file.

Overall, ontologies increase the semantic interoperability of already available data sources. This outcome has a direct impact on several clinical applications. In particular, it represents the underlying infrastructure for developing multicenter prediction models for clinical outcomes in radiation oncology. In fact, if every medical center transformed their data into FAIR through the ontology, data analytics can be performed on a broader dataset reducing possibilities of overfitting.

Ontologies and Semantic Web technologies will provide the infrastructure to query in an easy way the data needed by the model. Data will not need to leave the hospital, since being now FAIR, will be queried using SPARQL during the model training/validation. This application, known as distributed learning has been recently presented in literature as a promising application in radiation oncology^{16,20,21}.

In addition, Semantic Web and ontologies allow connecting different databases. In fact, data are transformed into universal concepts connected between each other: linked data. New relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing.

In the examples we showed in the result section, we successfully integrated data coming from different sources: clinical databases, clinical trials bank, and scientific literature databases to answer questions of clinical interest.

Finally, semantic databases (e.g., a collection of RDF records) have all the advantages from relational databases, but could provide the possibility of artificial intelligence to query and analyze the data, since these have been transformed into machine-readable records. Recently, we faced a transition from relational databases to semantic databases. The reason is that, semantic databases utilize an expanding semantic model that readily incorporates new varieties of data sources and more easily adjusts to changed requirements as they arise. Subsequently, linking disparate datasets is far easier in a semantic graph setting. In addition, semantic graphs allow to discover hidden relationships between underlying data. In fact, the granular nature of semantics allows to determine relationships between different elements.

4.C. Dynamic body of knowledge

We decided to put the ontology publically available on BioPortal, so that users could test and validate it with the aim of (a) developing a dynamic and growing body of knowledge; (b) guaranteeing the broadest coverage for the radiation oncology data domain.

In addition, the latest version of the ROO is published on the GitHub: users are able to insert enhancements and open issues, making the ontology development a collaborative process.

4.D. Limitations

In this work, we explored the ontology-based data integration with data from rectal cancer databases. We were able to map all the entities present in the databases with concept and properties from the ROO ontology. However, the ROO should be tested also on larger databases, other diseases and routine clinical data to check if all the main information could be covered.

In addition, this work lacks of the system evaluation. Further investigations on evaluating the system performance need to be considered such as comparing the query time between SPARQL and traditional databases.

4.E. Future developments

The first future development is to extend the ROO to guarantee a particular a broader coverage for an extensive use in the radiation oncology field. In particular, we would like to expand our ontology with:

- Detailed concepts for mapping radiation oncology annotations including organ at risks, nodals.
- Detailed concepts for mapping treatment-related concepts and properties such as Dose Volume histograms (DVH).

The second future development wants to expand the number of users. In this sense, we will continue proposing the ontology as underlying architecture for advance modeling applications such as distributed learning. In addition, we will try to use the ROO combined with other ontologies under development to combine and link: DICOM information, clinical data and quantitative features computed on patients' images and variables.

5. CONCLUSION

We successfully demonstrated that is possible to convert clinical data following FAIR principles using the combination of ontologies and Semantic Web technologies.

We developed a broad Radiation Oncology Ontology that can be used in the domain of radiation oncology for data integration.

In addition, we showed how Semantic Web technologies based on developed ontologies allows to efficiently and easily

query data from different (relational database) sources without knowing *a priori* their structures.

This outcome opens the possibility to use ontologies and Semantic Web technologies to further produce and analyze linked data in radiation oncology.

CONFLICT OF INTEREST

The authors declare not to have any conflicts of interests.

Alberto Traverso and Johan van soest equally contributed to the manuscript.

^{a)}Authors to whom correspondence should be addressed: Electronic mails: johan.vansoest@maastro.nl; alberto.traverso@maastro.nl.

REFERENCES

1. Aspinall MG, Hamermesh RG. Realizing the promise of personalized medicine. *Harv Bus Rev*. 2007;85:108–117, 165.
2. Duffy MJ, Crown J. A personalized approach to cancer treatment: how biomarkers can help. *Clin Chem*. 2008;54:1770–1779.
3. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
4. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claerhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. *Comput Methods Progr Biomed*. 2015;118:322–329.
5. Rubin DL, Rodriguez C, Shah P, Beaulieu C. iPad: Semantic annotation and markup of radiological images. *AMIA Annu. Symp. Proc. AMIA Symp*. 626–630; 2008.
6. Van Soest J, Lustberg T, Grittner D, et al. Towards a semantic PACS: using semantic web technology to represent imaging data. *Stud Health Technol Inform*. 2014;205:166–170.
7. Min H, Manion FJ, Goralczyk E, Wong Y-N, Ross E, Beck JR. Integration of prostate cancer clinical data using an ontology. *J Biomed Inform*. 2009;42:1035–1045.
8. Meldolesi E, van Soest J, Alitto AR, et al. VATE: validation of high technology based on large database analysis by learning machine. *Color Cancer*. 2014;3:435–450.
9. Jochems A, Deist TM, van Soest J, et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. *Radiother Oncol*. 2016;121:459–467.
10. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: terminology and typology, methods. *Inf Med*. 2000;39:16–21.
11. Gruber TR. A translation approach to portable ontology specifications. *Knowl Acquis*. 1993;5:199–220.
12. Bechhofer S. OWL: web ontology language. In: Liu L, Özsu MT, eds. *Encyclopedia of Database Systems*. Boston, MA: Springer US; 2009:2008–2009.
13. Berners-Lee T, Hendler J. Publishing on the semantic web. *Nature*. 2001;410:1023–1024.
14. van Stiphout RGPM, Valentini V, Buijsen J, et al. Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: a multicentric prospective study with external validation. *Radiother Oncol*. 2014;113:215–222.
15. Noy NF, Crubezy M, Fergerson RW, et al. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu. Symp. Proc. AMIA Symp*. 953; 2003.
16. Deist TM, Jochems A, van Soest J, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol*. 2017;4:24–31.
17. Hert M, Reif G, Gall HC. A comparison of RDB-to-RDF mapping languages, in (ACM Press; 2011), pp. 25–32.

18. Yunianta A, Barukab OM, Yusof N, Dengen N, Haviluddin H, Othman MS. Semantic data mapping technology to solve semantic data problem on heterogeneity aspect. *Int J Adv Intell Inform.* 2017;3:161.
19. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform.* 2008;41:706–716.
20. Lambin P, Roelofs E, Reymen B, et al. Rapid learning health care in oncology' – an approach towards decision support systems enabling customised radiotherapy. *Radiother Oncol.* 2013;109:159–164.
21. El Naqa I, Li R, Murphy MJ, eds. *Machine Learning in Radiation Oncology.* Cham: Springer International Publishing; 2015.