Computers in Biology and Medicine 43 (2013) 822–832

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

An ontology of cancer therapies supporting interoperability and data consistency in EPRs

Claudio Eccher [a,*], Alessandro Scipioni [b], Alexis A. Miller [c], Antonella Ferro [d], Domenico M. Pisanelli [e]

[a] Fondazione Bruno Kessler-Center for Information Technology, via Sommarive 18, 38050 Povo, Trento, Italy
[b] Department of Information Engineering and Computer Science, University of Trento, Italy
[c] Radiation Oncology, Illawarra Cancer Care Centre, Wollongong NSW, Australia
[d] Medical Oncology, S.Chiara Hospital, Trento, Italy
[e] CNR Institute of Cognitive Science and Technologies, Rome, Italy.

ARTICLE INFO

Article history:
Received 27 November 2012
Accepted 16 April 2013

Keywords:
Ontology
Semantic web
OWL and SWRL
Cancer therapy
EPR
DSS
Medical errors

ABSTRACT

Ontologies can formally describe the semantics of the medical domain in an unambiguous and machine processable form, acting as a conceptual interface between different applications that must interoperate.

In this paper we present an ontology of cancer therapies originally developed to bridge the gap between an oncologic Electronic Patient Record (EPR) and a guideline-based decision support system. We show an application of the ontology complemented by rules to classify therapies recorded in the EPR. The results show how such an ontology can be used also to discover possible problems of data consistency in the EPR.

1. Introduction

Medicine is a complex domain from the point of view of modeling and representing intended meaning. There are different activity domains (e.g. clinical vs. administrative knowledge), different scientific granularities (e.g. molecular vs. organic detail), different user requirements for the same service (e.g. physician-oriented vs. patient-oriented views), and ambiguous terminology (polysemy).

Many people today acknowledge that the development and use of biomedical ontologies represent a key advance in building better Information and Communication Technology (ICT) tools [1–5]. On the other hand, many outside the academic world are skeptical about the real impact that ontologies may have on the design and maintenance of working information systems.

Nevertheless, biomedical ontologies have become the essential element in a variety of semantic based applications: natural language processing [6], decision support systems [7,8], and data integration and system interoperability [7,9–13]. As a matter of fact, the inter-connection of data storage systems and decision support systems is a long-standing challenge for medical informatics research [14–17].

Among medical sub-domains, cancer is a complex group of diseases that affects a significant portion of the population worldwide. The provision of cancer care requires the coordinated action of healthcare professionals of different disciplines over an extended period. In this context, an important role can be played by ICT tools in favoring and supporting evidence-based, coordinated care provision through shared electronic folders for patient history management, guideline-based decision support systems, collaborative work and workflow management support tools.

The seamless integration of these systems in clinical practice; however, can be restrained by semantic conflicts, i.e. conflicts caused either by using different terms in heterogeneous systems in order to express the same entity: a drug trade name (e.g., *Oncovin*) and generic name (vincristine); or by denoting different real entities with the same term in the medical language: e.g., “therapy” means, according to the context, either an abstract plan or the enacted care; or, finally, by using ambiguous or ill-defined clinical terms and concepts: e.g., treating metastatic disease with *palliative* intent using *palliative* chemotherapy.

These problems of semantic interoperability can be overcome by using shared vocabularies formalized in ontologies, which describe the intended meaning of the domain terms in an

* Corresponding author. Tel.: +39 04 6131 4161; fax: +39 04 6130 2040.
E-mail addresses: cleccher@fbk.eu, claudio.eccher@gmail.com (C. Eccher).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.04.012

unambiguous and machine processable form. The advantages of specifying the conceptualization of the medical domain in ontologies are well known: well-constructed ontologies codify knowledge in a way independent from any particular application, and provide a rich, predefined vocabulary that can serve as a conceptual interface to different systems.

In this paper, we present the implementation and use of an axiomatic ontology of therapies, specifically focused on cancer. The ontology was initially developed to facilitate the interoperability between a guideline-based decision support system for breast cancer and an oncological Electronic Patient Record (EPR). This paper shows an application of the ontology in a real clinical setting to classify therapies data in the EPR. The ontology is validated by comparing our results and the classification of therapies made by oncologists at the time of therapeutic decisions. This allowed us to identify compilation errors and clinical misclassifications in the EPR.

2. The medical problem

2.1. The context

Cancer is a complex group of pathologies whose appropriate treatment requires the collaboration of different specialists working together in a coordinated fashion for the lifetime of the patient. Treatment of cancer is actually a sequence of therapies decisions and execution, outcome evaluation, with clinical assessments including laboratory tests, which are cyclically repeated as necessary. The process of cancer care is usually initiated by a family physician outside the oncology ward where the diagnosis of cancer becoming apparent after a set of clinical visits and tests for the investigate symptoms or signs suffered by the patient or from the result of a screening campaign.

The patient is deferred to an oncologist who verifies that a clear diagnosis has been reached, and that the stage of the disease is adequately described. The oncologist assesses the ability of the patient to endure treatment, and may decide on the first set of therapies if appropriate, or undertake wider discussion in a multidisciplinary meeting or 'tumor board'.

The therapies available include surgery, pharmaceuticals (chemotherapy, small molecules, monoclonal antibodies, hormone therapy) hyperthermia and radiotherapy (external beam, brachytherapy). Publications of therapy outcome inform the current best evidence paradigm for each tumor type, with different therapies being applied singly or combined in a variety of sequences. Different specialists (radiation, medical and surgical oncologists) supervise the different therapies applied in the care process, which highlights the need for an integrated approach to treatment by a coordinated multidisciplinary team.

In their everyday work, oncologists refer to treatments employing a professional "jargon", using several categorization levels for therapies. The first immediately identifiable level refers to the kind of therapy (surgery, pharmaceuticals or radiotherapy), with subcategorizations: e.g. surgery on the breast may involve removal of part (lumpectomy, breast conserving surgery) or the entire breast (mastectomy). Pharmaceuticals may include chemotherapy (cytotoxic chemicals), biological therapy with monoclonal antibodies (against tumor markers such as HER2 in breast cancer), receptor modifiers (e.g., small molecules targeting epidermal growth factor receptor), hormone therapy, or anti-angiogenesis agents.

Further levels of categorization, which we could call meta-levels, exist: by intent (cure, palliation), by importance (definitive, primary, radical), by execution order (neo-adjuvant, concurrent, adjuvant), by mechanism (eradicate local disease, achieve regional control), and by the extent of action (local versus systemic).

2.2. The problematics of medical terminology

Ambiguities and confusion arise when one analyses the actual usage of terms in the everyday activities. For example, while the intent of a treatment may be palliation, it is frequently described as a "palliative intent". However, the therapy chosen to achieve this intent might be described as "palliative hormone therapy". The same descriptor is used for two semantic entities. Even the notion of "treatment" and "therapy" is blurred, as patients will be told that they will be "treated with radiotherapy". Oncologists vary in their preferred terms. The most important treatment for a cancer may be called the "primary" or the "definitive therapy". Some believe that only a single therapy can be described as definitive, i.e. "definitive surgery" means that no other therapy will occur before or after the operation. While terms similar to "primary" and "adjuvant" are widely understood, there is no agreement that their use is even necessary. The design and development of effective interoperable ICT applications require unambiguous and precise definition of medical concepts, in order they can be correctly managed and used by the applications.

Consider, for example, the following two fragments of the NCNN guideline [18] for the treatment of invasive non-metastatic and metastatic breast cancer:

> *Breast irradiation may be omitted in those 70 y of age or older with estrogen-receptor positive, clinically node negative, T1 tumors who receive adjuvant endocrine therapy (category 1).*

> *This recommendation also applies to the relatively new class of patients who are diagnosed with HER2-positive metastatic disease following prior exposure to Trastuzumab in the adjuvant setting.*

As a matter of fact, the term 'adjuvant' is used hundreds of times in the entire guideline, but it is never formally defined because its meaning is considered obvious in the oncologists' community. Indeed, in the first guideline fragment, both the breast irradiation and the endocrine therapy are "adjuvant", as there is the presumption of prior 'primary' surgery, but the term is not specifically applied to the irradiation. In respect to the second guideline, oncologists frequently refer to the drug Trastuzumab using its commercial name Herceptin (see e.g. [19]).

When formalizing the guidelines for use by a decision support system, however, the terms must be precisely defined to allow the system to correctly identify the adjuvant therapies and the drugs used in the patient EPR. If the term "adjuvant" applies to the breast irradiation, then it should be explicitly inserted to differentiate it from "primary" breast irradiation, which can also occur in breast cancer. Furthermore the surgery implied in the first fragment must also be made explicit with defined terms. A more accurate guideline would be:

> *Following* primary surgery *which reveals estrogen-receptor positive, clinically node negative, T1 tumor*, adjuvant *Breast irradiation may be omitted in those 70 y of age or older with estrogen-receptor positive, clinically node negative, who receive adjuvant endocrine therapy (category 1).*

These difficulties are not diminished by the analysis of knowledge sources publicly available. In the NCI Thesaurus [20], the most comprehensive reference terminology for cancer, for example, *Adjuvant Therapy*, *Chemotherapy*, and *Systemic Therapy* are sibling classes, children of *Therapeutic or Preventive Procedure*. Chemotherapy, however, is a pharmacological therapy that uses cytotoxic chemicals, is a systemic therapy, and can be adjuvant, if administered after primary therapy to kill cancer cells spread throughout the organism, or neoadjuvant if administered to shrink the tumor.

2.3. The need for a therapies ontology

The beginning of this work dates back to 2008 with the OncoCure project [21], which was aimed to design and develop a prescriptive guideline-based Decision Support System (DSS) for giving active support to medical therapy decisions, through the execution of the Asbru-encoded protocols of pharmacological therapies for breast cancer. The guideline model was based on the breast cancer treatment protocols in use in the Medical Oncology Unit of the regional hospital of Trento (Northern Italy), which was involved in the project. The treatment protocols are prepared by an oncology domain expert distilling the knowledge in national and international guidelines and consensus meetings' recommendations, in order to create an agile instrument that can be easily used during the everyday care delivery activities. Although more structured than the textual guidelines, these cancer protocols still contain a lot of implicit knowledge regarding cancer therapies, since the terms used imply different treatment intents (cure, palliation), temporal relations and cancer phases (e.g., neoadjuvant, adjuvant, etc.), different procedures (medical therapy, radiation therapy), and different classes of drugs (hormone therapy, chemotherapy).

We decided to develop a therapies ontology with two aims: (i) disambiguate the meaning of terms for facilitating the discussion between the oncologists involved in the OncoCure project, and the computer scientists who model the guideline knowledge base, and (ii) enable the interoperability between the EPR in use in the oncological ward and the DSS, which needed to operate unambiguous concepts of higher-level with respect to raw clinical data stored in the EPR database.

3. The therapies ontology

3.1. Knowledge capture and modeling

Many languages can be used to represent ontologies, from highly informal to rigorously formal. The Web Ontology Language (OWL) is part of the growing stack of W3C recommendations related to the Semantic Web. To build our ontology we used the OWL DL sublanguage, which allows the maximum expressiveness while retaining computational completeness. Moreover, we used the Semantic Web Rule Language (SWRL), which extends the set of OWL axioms, in order to include in our ontology Horn-like rules expressed in terms of OWL concepts to reason about OWL individuals. SWRL is based on OWL DL and provides more expressive power than OWL DL alone, at the expense of decidability. Finally, we wrote the ontology in the ontology editor Protégé (v. 3.4.4).

The precise definition of cancer concepts is necessary to understand the protocol recommendations at each step of the care process, including the instantaneous position of the patient in the process (which describes the next treatment also), preceding history of treatments, and patient and tumor eligibility conditions. Hence, the acquisition of therapy-related knowledge occurred within the context of the oncological workflow, that is, a “care process” knowledge, which has a physical site (in our case the oncology ward for treating breast cancer patients). To this end, regular meetings between computer scientists (CE) and a breast cancer specialist (AF) were held. In these meetings, semi-structured interviews were conducted to clarify the content of the protocols and their application in the care process. The discussion was transcribed and later analyzed by the computer scientists. The analysis of the knowledge collected in each meeting was the basis for the questions posed at the next one.

Contextually to the care process definition, we created an initial set of therapy concepts and their relations with the explanation of each term. The meanings of terms were checked against the definitions given in different cancer knowledge sources such as UMLS terminologies, National Cancer Institute Thesaurus (NCIT) and web sites dedicated to breast cancer patients. The list of concepts and their combinations that define oncological therapies (e.g., curative adjuvant hormone therapy) were further discussed and validated with the oncologist and related to the characteristics of the diseases they are administered for (e.g. non-metastatic breast cancer). From the list of concepts defined at this step we used Protégé to develop the first version of the OWL ontology focused on breast cancer [22]. Subsequently, the content of the ontology was discussed with a second oncologist (AM) with training in medical informatics, to extend the domain coverage from breast cancer therapies to cancer therapies in general. The discussions produced a concept map in which categories and relations were revised and new concepts and links were added.

From the conceptual map we produced the final version of the therapies ontology.

3.2. Classes and relations

The ontology developed is based upon the DOLCE top-level ontology [23] and so it inherits the basic distinction between *endurants* and *perdurants*. Classically, endurants (also called *continuants*) are characterized as entities that are ‘in time’, they are ‘wholly’ present (all their proper parts are present) at any time of their existence [24]. On the other hand, perdurants (also called *occurrents*) are entities that ‘happen in time’; they extend in time by accumulating different ‘temporal parts’, so that, at any time t at which they exist, only their temporal parts at t are present. For example, in the context of this ontology, the description of a therapy is an endurant and its enactment a perdurant. The most significant perdurants are processes and here we find the different kind of actions (not to be mistaken for their descriptions): clinical, decisional, laboratory-exam and surgical.

It is natural for a realist-based ontology of particulars to take into account physical and tangible objects, but other immaterial and non-tangible entities must be considered too. Therefore, DOLCE distinguishes between *physical* and *non-physical* endurants. The former class is basically devoted to represent resources, which may be human, permanent (e.g. surgical theatre, linear accelerator, chemotherapy chair) or disposable (scalpel, cannula). Non-physical entities relevant for this ontology are information entities, i.e. descriptions of intents, treatments, therapies, actions and pathologies that can be found in a medical record. Fig. 1 shows the main endurants defined in the ontology of therapies.

The core information classes in our ontology are *Treatment_description* and *Therapy_description* (in the following, ontology entities will be written in italics). A treatment is a high level action plan that combines several therapies in a particular sequence to treat the cancer of a particular stage. The treatment has an intent such as ‘cure’, ‘palliation’, or ‘surveillance’. In our ontology the registration of treatments was modelled through the *Treatment_description* class, whose instances represent the descriptions of the plans to treat the disease. A restriction on *Treatment_description* that acts along the property *has_action_plan* with a filler of *Therapy_description* specifies a treatment is composed by at least one therapy. The information entity, *Therapy_description*, has instances that represent the registration of therapies administered to patients. In fact, the object property *describes* links individuals of *Therapy_description* to individuals of *Therapy_enacted*, which is a perdurant representing the actual process of administering the therapy. A restriction on *Therapy description* that acts along the property *has_temporal_relation* with a filler of *Therapy_description*

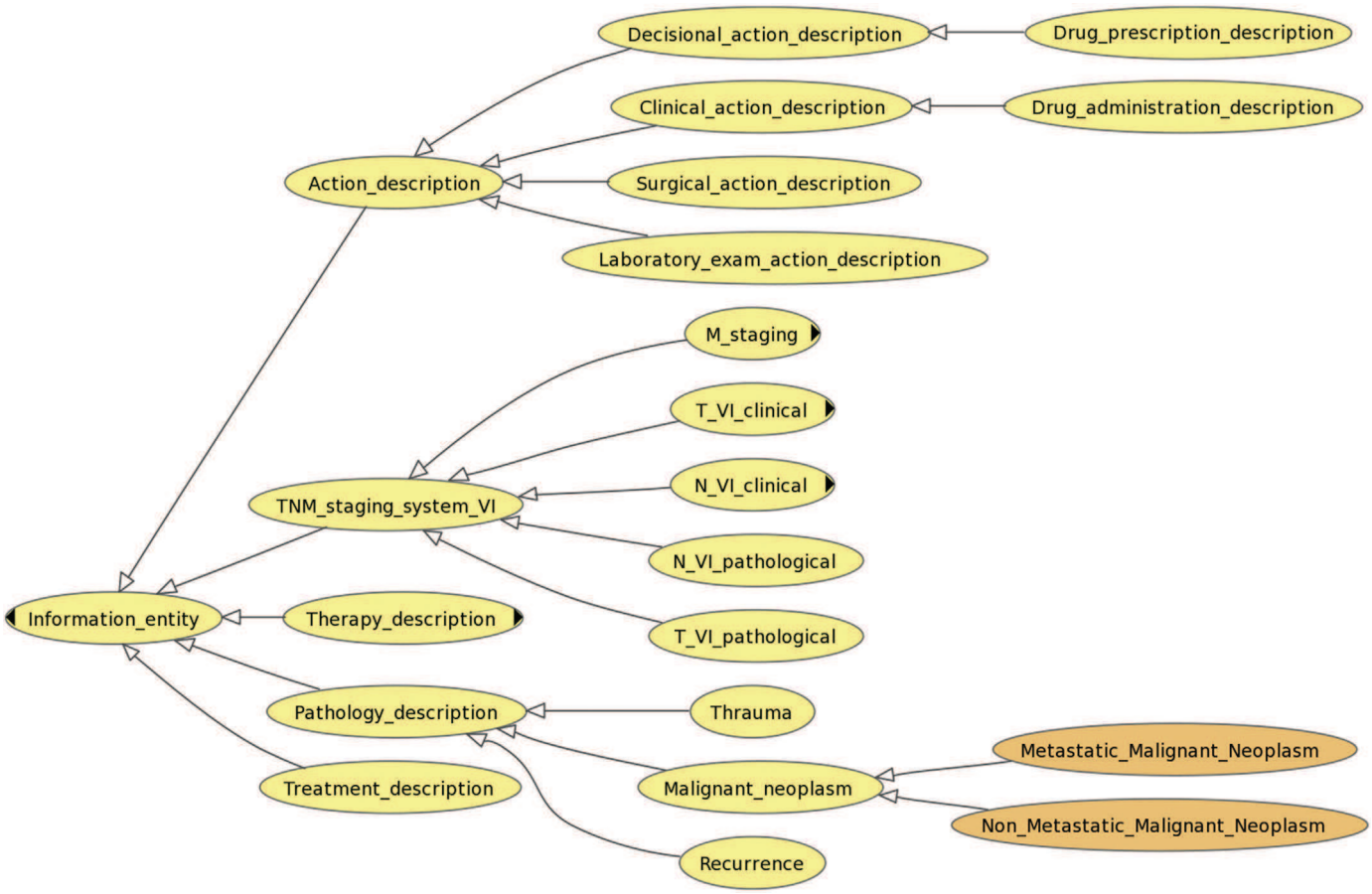

Fig. 1. The main endurants in the therapies ontology are information entities representing the description of clinical events that are registered in patient folders, either paper-based or electronic-based. For example, an individual of *Therapy_description* represents the registration of an enacted therapy.

specifies that a therapy description is temporally related to other therapy descriptions. In fact, the sub-properties of *has_temporal_relation* represent the Allen temporal operators between the time intervals in which the therapy described by an individual of *Therapy_description* was performed. For example an individual of *Therapy_description* describing an adjuvant therapy has a relation *is_after* with an individual of the same class describing a primary therapy (mastectomy or breast-conserving surgery in the case of breast cancer).

Peculiar to the DOLCE-based approach is the notion of *quality*. Qualities are entities such as shape, height, weight, which characterize the features of the different items in ontology. Adopting this paradigm, in the therapies ontology we defined a hierarchy of quality concepts that characterize the therapies according to various dimensions. For example, a therapy that involves the use of antineoplastic drugs absorbed into the bloodstream and distributed to all parts of the body has, at the same time, pharmacological quality, systemic quality and, if used to shrink the tumor before the primary treatment, neoadjuvant quality. The diagram in Fig. 2 depicts the hierarchy of main qualities. Rather than establishing multiple *IS_A* links between a therapy and several superclasses (multiple hierarchy), we added existential restrictions that act along the property *has_quality* with fillers of subclasses of *Quality* (Fig. 2). By attaching several qualities to the same therapy or treatment, many possible combinations can be defined avoiding the entanglement of multiple hierarchies.

The most important quality classes for oncological therapies are *Therapy_role* and *Therapy_modality*, whose class diagrams are shown in Figs. 3 and 4, respectively. *Therapy_role* represents the roles that an oncological therapy can play in the treatment process of a cancer patient (adjuvant, neoadjuvant, primary, etc.), whereas *Therapy_modality* represents the different forms of cancer therapy: mainly pharmaceutical, surgical and radiation.

3.3. Temporal features

Many classes in the therapies ontology are described or defined by means of properties incorporating a well-defined temporal relationship between the concepts. For example a concurrent therapy is a non-primary therapy executed at the same time as a primary intervention. To use our ontology to reason on therapies data extracted from the EPR, the reasoning must be based on these temporal relations. OWL and SWRL technologies; however, have very limited support for the modeling of temporal information. There are no standard high-level mechanisms to consistently represent and reason with temporal information. This restricts the complexity of temporal information that can be represented and reduces the possibilities for automated reasoning using temporal information. The temporal ontology developed at Stanford [25] encodes a valid time temporal model along with the definition of a SWRL library of methods that implement the Allen's interval based temporal operations [26]. The core class of the model is *temporal:Entity*, whose subclass *temporal:ValidTime* represents the time or times during which the associated information is held to be true. The latter class has subclasses *temporal:ValidPeriod* and *temporal:ValidInstant*, which represent intervals and instants, respectively. Built-ins defined in the SWRL library can be used to write rules to check the temporal relation between two valid times associated to an entity. In addition, an SWRL-based query language [27] provides the support to write temporal queries on ontology.

Fig. 2. The main qualities of the ontology of therapies. According to DOLCE foundational ontology, the Stanford temporal concepts (the class *temporal:Entity* and its subclasses) are imported as subclasses of *Quality*, so that that the temporal information is a quality attached to an information entity. *Therapy_role* and *Therapy_modality* classes, used in the SWRL rules, are detailed in Figs. 3 and 4.

In order to attach temporal information to therapy descriptions (i.e. the time interval during which the therapy was performed) and to reason with it, we imported the Stanford temporal ontology in our therapies ontology and made the core class *temporal:Entity* a subclass of *Quality* (Fig. 2). Then to use the time classes in our ontology we added to the class *Therapy_description* a user-defined property *has_temporal_entity* with range *temporal:ValidPeriod*. Consequently, all instances of this class can use the properties *temporal:hasStartTime*, *temporal:hasFinishTime* and *temporal:hasGranularity* associated with the interval, in order to consistently record the temporal information associated with the description of clinical findings and activities.

3.4. Adding SWRL rules

To attach meta-information to therapy and treatment description class instances (e.g., of being primary, adjuvant, neoadjuvant) depending on their relations with other events and conditions, we had to exploit the additional expressivity provided by SWRL. In addition to the temporal querying capabilities allowed by the SWRL library defined in the temporal ontology, we needed to express knowledge like “absence of new stages” as in Rule 1 (see below), which are difficult to express in OWL because of open world assumption. For this, we used the core SQWRL built-ins to implement these functionalities, specifically SQWRL sets operators. To this end, we wrote a set of SWRL rules operating on the individuals of the populated ontology,

We built the OWL part of the therapies ontology (classes and properties) to represent therapy-related concepts common to the entire oncologic domain. Given the complexity and extent of the domain, however, this domain-wide approach would require the writing of a set of SWRL rules for each particular condition (tumor type, tumor staging, etc.). Since the ontology was created for the OncoCure project, we focused the work on breast cancer, and wrote the rules for this restricted sub-domain. Using our approach, however, the analysis can be extended to other sub-domains.

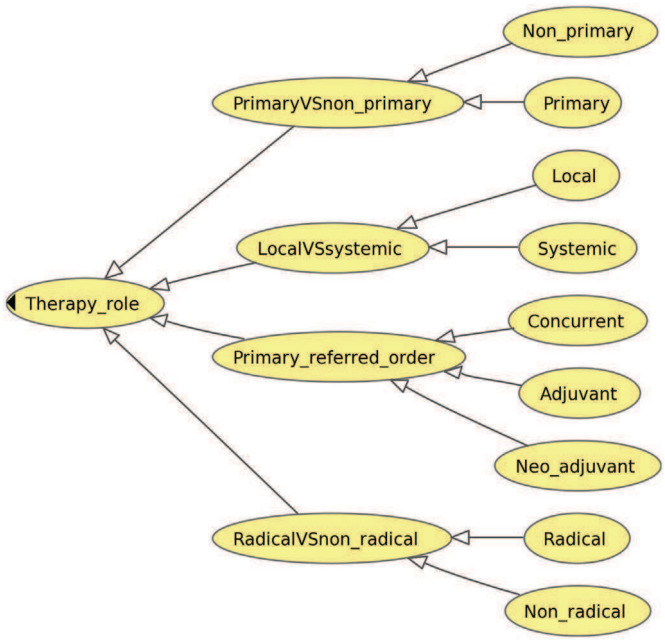

Fig. 3. The class diagram of *Therapy_role* quality, that represents the roles that an oncological therapy can play in the treatment process of a cancer patient.

We give here an example of the rules to express the knowledge that mastectomy is one primary action for a non-metastatic (M0) operable breast cancer that does not have recurrences or that has recurrences after the surgical intervention (progression).

When expressing disjunction it is convenient to break the rule in two: one for the absence of new stages (Rule 1), the second for the case in which onset of metastasis occurs after the surgical intervention (Rule 2). In Rule 1 we need to use set operators of Semantic Query-Enhanced Web Rule Language (SQWRL), the SWRL-based language for querying OWL, to implement Negation as failure, not supported by SWRL. In the examples, *has_new_stage* (*?mm*,*?rec*) is the property relating a malignant neoplasm with a local or distant recurrence of the cancer.

The user-defined property *has_temporal_entity* adds the temporal dimension defined in the temporal ontology (time interval) to a class of the therapies ontology. Entities prefixed by *temporal* are properties and built-ins defined in the temporal ontology, whose meaning is evident from the name. The effect of the consequent of each rule is to link an individual of *Therapy_description* satisfying the rule with that of *Primary* through the property *has_quality*. Then in Rule 3 we express the knowledge that chemotherapy given before a primary therapy (mastectomy) is a neoadjuvant therapy. As before, the consequent links an individual of *Therapy_description* to the individual of *Neoadjuvant*.

Fig. 4. The class diagram of *Therapy modality* quality. The entities under *Pharmaceutical* class are specific qualities pertaining to drug-based cancer therapies.

Using the SWRL editor plug-in in Protégé we have defined 36 rules for the entire set of meta-information that can label breast cancer therapy description individuals.

4. The ontology in the clinical practice

In this section we sketch out how the ontology developed can be practically employed and useful in the clinical practice. Its first aim is to remove ambiguities in therapy definition, but it can also be employed to check the consistency of data entered in electronic patient records. Last but not least, we envisage its usage for preventing clinical errors.

4.1. Use of the ontology

An ontology is essentially a model to capture the precise meaning of a therapy term and remove ambiguities. Additionally, our ontology can be practically used in combination with an EPR for several automated tasks. In general, the set of data regarding therapies entered by an oncologist in the EPR varies depending on whether the oncologist is responsible for that therapy. For surgical interventions, an activity performed outside the oncological ward, the medical or radiation oncologist may input only the operation type name (e.g., mastectomy or quadrantectomy) and its date. For the medical oncologist, the case of pharmacotherapy prescription is more complex. The need for a pharmacological treatment for a patient in a certain stage of her disease is firstly decided during an oncological visit. Then the oncologists inputs in the EPR the coordinated succession of pharmacotherapies, namely the name of drugs or combination of drugs of each therapy, doses, modality of administration, number of cycles (e.g. Anthracycline and Cyclophosphamide, "AC", for 4 cycles followed by Tamoxifen for 5 years), and the time scheduled for the administration of each therapy. The therapy administration sessions should follow the programmed timeline. The periods of administration may vary. The consequence of adverse clinical events that result form pharmacological toxicity or disease progression can result in suspension, modification or interruption of one part or of the entire treatment. The parameters of the enacted therapy, including the start and end times, are in any case registered at each administration session or block of sessions. These data constitute the treatment history of that patient. In addition to these 'low level' data, an oncologist may label therapies with information at a higher level of abstraction: e.g., intent of treatments, role of therapies, etc., if the EPR has room for this.

In clinical practice, the knowledge contained in the therapies ontology can be used by a software system to reconstruct this kind of 'high-level' information that is missing or not explicitly stored in an EPR to favor system interoperability. It is often the case that this kind of information is required by other systems to perform specific tasks. For example, a DSS based on the breast cancer guideline, as shown above, requires the assertion of the drugs administered as adjuvant therapy. Moreover, the labels automatically assigned by the system to therapies and treatments can be used to determine the coherence of this 'high level' information in relation to the data entered in the EPR by physicians at compilation time, in order to check the EPR database for errors. Errors in the EPR may in fact (i) invalidate the conclusions of a Decision Support System relying on this information, or (ii) hinder the automatic extraction and statistical analysis of data for research and epidemiology.

Lastly, the ontology-based system presented here can be used in real time to avoid typing and medical errors during the data input by controlling the coherence between "low-level" and "high-level" information. A system that immediately signals errors in data input, complemented with error explanation, could also provide benefit for oncology trainees and medicine students.

4.2. An experiment in breast cancer domain

To demonstrate that our ontology can be effectively used for the tasks listed above, it is necessary to assess the reliability of the ontology-based labeling process. To this end, we conducted an experiment aimed at labeling with high-level information the therapies and treatments in the oncologic EPR being used in the Medical Oncology Unit of the regional hospital of our province.

The oncological EPR was built in close collaboration with the oncologists at the point of care; hence it has room for the high-level information related to the role of the therapies (primary, radical, adjuvant etc.), inputted by the oncologists at the time of treatment decision. Hence, the idea was to use the high-level information provided by the oncologists to validate the automatic labeling process using the therapies ontology.

To this end we implemented the following steps:

i. Ontology population with class instances created from the EPR DB data;
ii. Execution of the SWRL rules by a WRL engine (Jess) to infer the therapy and treatment classification from the relations between the individuals;
iii. Comparison of the automatic classification with the labels assigned by the oncologists to therapy events at compilation time and determination of the causes of the discordances.

4.2.1. Ontology population

The first step of the process is the population of the ontology with the individuals representing the information related to each patient (diseases, diseases characteristics, administered therapies, etc.)

We populated the ontology with an instance for each leaf quality class (e.g. one individual is created under the class *Chemotherapeutic*, one under *Mastectomy_quality*, etc.). This can appear odd at a first glance, but these individuals are necessary to write SWRL rules maintaining the DOLCE approach of attaching qualities to concepts, as we will seen in the next section, since SWRL rules cannot be used to add new instances to an ontology since the language does not allow that variables not defined in the antecendent are used in the consequent: e.g., if all the therapies were mastectomies and in populating the ontology we wanted to create a new individual of type *Mastectomy_quality* for every individual of type *Therapy_description*, a rule like *Therapy_description*(*?thd*)→*Mastectomy_quality*(*?mq*) cannot be written.

Secondly, we populated the ontology creating instances representing the relevant patient data from the legacy EPR. To this end we created views in the EPR representing the subset of data of interest, and then a custom Java component mapped the views to individuals and properties in the ontology using the Protégé OWL API.

Unfortunately, the domain conceptualization process showed that the database schema was flawed by the semantic confusion between therapies and treatments discussed above. The latter are not explicitly represented in the DB as the medical oncologist assigns the intent to therapies (palliative, curative therapy) rather than to treatments (cure, palliation). This is not a problem however, since the set of therapies planned in the same visit can be grouped in a treatment for the specific cancer stage.

Simplifying, for a patient with breast cancer in the EPR we added the individuals corresponding to the DB entities to the information entity classes. We added to individuals the temporal information registered in the DB through the entities of the

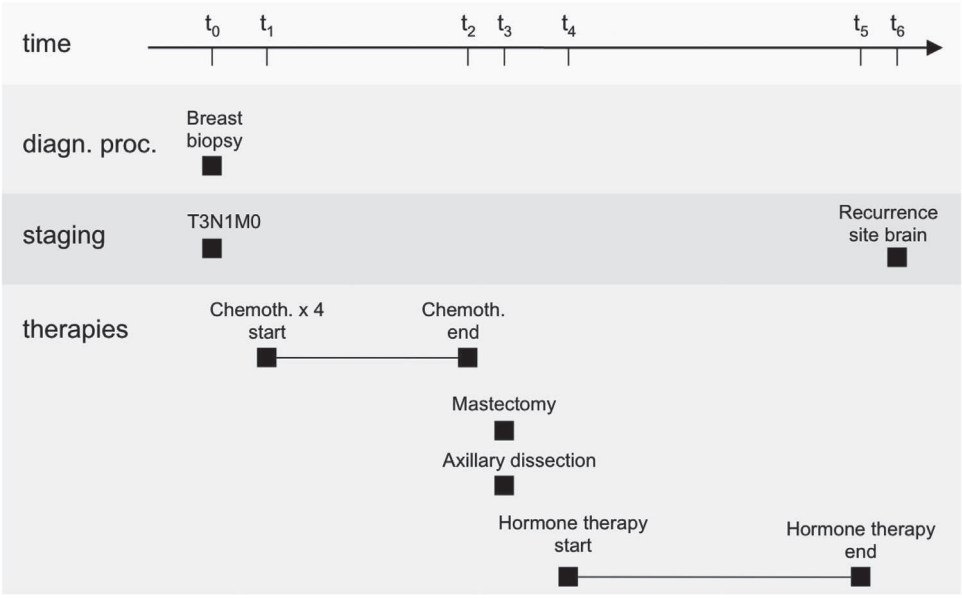

Fig. 5. The set of events representing a patient history in the oncological EPR. As a matter of fact, the initial chemotherapy and the hormone therapy have neoadjuvant and adjuvant role, respectively.

temporal ontology. Then we added the properties between information entity individuals; for example the property *has_location* between the *Malignant_neoplasm* class and an anatomic concept class, as *Breast* or *Brain*[*] used in the example below; and the properties relating them to individuals of the quality classes as required: for example the property *has_quality* between an individual of the class *Therapy_description* and that of the class *Mastectomy_quality* for each mastectomy found in the EPR.

To clarify the approach, let's consider the example of the history of a patient with a diagnosis of operable invasive breast cancer and the set of events shown in Fig. 5. From the information retrieved from the EPR, the ontology is populated as follows (see Fig. 1):

- An individual of the class *Malignant_neoplasm* with start time and end time t_0, which is linked with an individual of the class *M0*[†] and one of the class *Breast* by the properties *has_M_staging* and *has_location*, respectively.
- An individual of *Treatment_description* with start time t_1 and end time t_5. Since this entity is not recorded in the EPR, we inferred it from the group of therapies planned in a visit contextually or after a staging of cancer and performed before the next staging variation registered in the EPR. The start or the first therapy and the end of the last one correspond to the start and end of the treatment, respectively. In the example, the two staging events are the initial staging of cancer (T3N1M0 in the graph) and the onset of metastasis (Recurrence site: Brain). The property *is_applied_for* links this individual with that of *Malignant_neoplasm*.
- An individual of *Therapy_description* with start time t_1 and end time t_2, which is linked to an individual of the quality class *Chemotherapeutic*. Since this therapy is part of the treatment for non-metastatic breast cancer, the property *is_part_of_plan* links this individual with the individual of *Treatment_description*.
- An individual of *Therapy_description* with start time and end time t_3[‡]. An instance of the property *has_quality* links this individual with the individual of the class *Mastectomy_quality*. This individual is linked to the individual of *Treatment_description* by the property *is_part_of_plan*.
- An individual of *Therapy_description* with start time t_4 and end time t_5, which is linked to an individual of the quality class *Hormonal* by the property *has_quality* and to the individual of *Treatment_description* by the property *is_part_of_plan*.
- An individual of the class *Recurrence* with start time and end time t_6, the time of the onset of the new stage, linked to an individual of the class *Brain* through the property *has_location*. The instance of *Malignant_neoplasm* is linked to *Recurrence* instance by the property *has_new_stage*.

Proceeding this way, from the EPR we extracted the clinical histories of 100 breast cancer patients, for a total of 357 therapies.

4.2.2. Rule execution and result comparison

Protégé-OWL can access external DL Implementation Group (DIG) compliant reasoners over HTTP using the DIG language, XML based representation of ontological entities and subclass axioms. The Protégé-OWL distribution provides a bridge for the DIG compliant rule engine Jess [28] accessible from the SWRLTab plug-in or, to speed up the execution, from a Java application using the Protégé APIs. We ran the Jess engine to execute the SWRL rules on the populated ontology obtaining a set of inferred axioms that we put back in the ontology.

After the run, there was some discordance between the classifications inferred by the reasoner and those in the EPR:

- seven treatments, whose therapies were labeled curative in the DB were reclassified palliative;
- one therapy labeled palliative in the DB, was classified as adjuvant;
- the reasoner did not classify 26 therapies.

[*] The anatomic concepts are imported from an anatomy ontology.

[†] T and N are of no interest in this example.

[‡] Surgical interventions are registered in the EPR with the date of execution.

The discordant results were investigating by manually reviewing the corresponding cases in the EPR with an expert oncologist. We ascertained that the seven treatments and the adjuvant therapy were correctly classified by the reasoner, but the oncologists erroneously registered their intents in the EPR.

Regarding the therapies not classified by the automatic reasoning procedure:

- 7 therapies were not classified because the diseases were staged Mx (distant metastasis not ascertainable) at the time of diagnosis and never updated.
- 7 therapies were not classified because the TNM staging data were registered with the wrong date.
- 12 therapies were not classified because other data were missing. For example, the lack of the end date of neoadjuvant chemotherapy administration prevents the firing of Rule 3.

The unsuccessful classification of these therapies reveals the problem of incomplete, missing or wrong data in the EPR, and supports our initial contention regarding the usefulness of ontology use in clinical circumstances.

5. Discussion and future prospects

Medical language is complex, dense to observers and often idiosyncratic in its use. Understanding the exact meaning of concepts can be a difficult task if one is not endowed with medical knowledge. As a matter of fact, one important problem we dealt with during the knowledge acquisition phase of the OncoCure project was the understanding and the disambiguation of therapy-related terms, either cited in the textual treatment guidelines or used by oncologist recounting the cancer treatment process carried out in the ward.

We looked for available knowledge bases of cancer therapies suitable for this task; but could not find any artifact suitable for our purposes, so we decided to build the ontology of therapies based on DOLCE. If the ontology is an accurate reflection of the expert domain knowledge, it should be able to pinpoint poor or absent classification of clinical records and assist with knowledge acquisition.

To test the application of the ontology we conducted an experiment for re-classifying therapies stored in an oncological EPR in use at the point of care according to the meta-categories used by the oncologists in their everyday practice. For temporal reasoning with individuals instanced from the EPR database data we defined additional rules in SWRL and import the Stanford temporal ontology

The results of this experiment demonstrate that a formal ontology disambiguating medical concepts complemented with SWRL rules can be used to (i) implement ontology-base interoperability through the automatic classification of low level therapies data in higher level categories needed by external systems such as decision support tools, and (ii) find inconsistencies in an EPR, due either to actual medical errors (e.g., erroneous labeling of therapies) or to missing, incomplete or not updated information. The latter is rarely a problem for the skilled physician, who can reconstruct the clinical history from the context, but it is a problem for retrospective studies of quality of care, disease control and statistical or epidemiological studies conducted by oncologists, non-oncologists or through automatic extraction tools.

This tool could be used in real time to control the consistency and coherence of data input from the physician during the compilation of the EPR, and also suggest applicable guidelines. Moreover, it can run as a Quality Assurance check in background at the end of a ward working day to signal problems of quality of data entered in the EPR.

Last but not least, our work shows how a therapies ontology, which incorporates domain expert language and concepts, allied with reasoning services can be used to facilitate the system interoperability, by providing the classification of EPR data into categories that another system, e.g., a decision support system, can use.

One limitation of this work is that the time needed for the processing of the entire set of therapies for 100 patients is quite long (more than 300 min) on an Apple computer running the Mac OsX 10.6 operating system, with a 2 Ghz Intel Dual Core CPU and 4 GB RAM. As a matter of fact, managing data in an OWL ontology using the Protégé APIs is far less efficient than managing data in a relational database. However, this work was a proof-of-concept testing the validity of our approach with a large patient sample. These software routines would be optimized before considering routine use. In real world use, the oncologist would employ the reasoner on single patients to produce an ontologically coherent record, and when undertaking investigations on large number of patients could perform the task off-line, or out of hours.

The oncologists involved in the project need a system for discovering or preventing potential errors in the EPR, since they make a heavy use of EPR data for retrospective studies on the relation between tumor characteristics, therapies and outcomes and showed great interest for the automatic, ontology-based system presented here. Hence, in the future we plan to enforce the integration of the ontology into the EPR management system in order to implement a real time consistency checking mechanism able to prevent errors in data entry.

6. Summary

Cancer is a complex group of diseases that affects a significant portion of the population worldwide. Evidence-based, coordinated care provision can be favored and supported by Information and Communication Technology (ICT) tools: shared electronic folders, guideline-based decision support systems, and workflow management tools.

The integration of these systems in the clinical practice, however, can be restrained by semantic conflicts caused by the use of professional jargon and ambiguous and ill defined terms and concepts. Many people working in biomedical informatics have advocated the use of biomedical ontologies to overcome these problems and build better ICT tools. The semantics of resource content and capability can be described in an unambiguous and machine processable form in ontologies, which provide an application-independent conceptual interface between different systems.

In this paper, we present the implementation and use of an axiomatic ontology of cancer therapies, initially developed to facilitate the interoperability between a guideline-based decision support system for breast cancer and an oncological Electronic Patient Record (EPR) in use in an Oncology Unit. The ontology, written in Web Ontology Language (OWL), is based on the DOLCE top level ontology, from which it inherits the basic distinction between *endurants* and *perdurants* and the notion of *qualities*: entities which characterize the features of the different items in the ontology. The core classes of our ontologies are information entitles, non-physical endurants describing therapies and treatments. We defined a hierarchy of quality concepts that characterize the therapies according to various dimensions through the use of existential restriction, in order to avoid the entanglement of multiple hierarchies.

We applied the ontology in a real clinical setting to classify breast cancer therapies data in the oncological EPR according to the high level categories used by the oncologists. The classification mainly implies temporal relations between therapies. To reason with time we imported the Stanford temporal ontology implementing the Allen temporal relations, making the root temporal entity a quality attached to information entities in our ontology. Moreover, we complemented the OWL ontology with the definition of Semantic Web Rule Language (SWRL) rules that use the built in temporal ontology functions to operate on therapies ontology instances. Firstly the therapies ontology was populated with therapy instances from the EPR database. Then the Jess engine was run to execute the rules on the populated ontology and the inferred classification was written back in the ontology. The classification was compared with that resulting from the labeling assigned by the oncologists in the EPR. Respect to the latter, it resulted that the reasoner misclassified seven treatments and one therapy and was not able of classifying 26 therapies. EPR cases were revised with a domain expert to assess the reasons for discordances. It resulted that all discordances were due to incomplete, wrong or missing information in the EPR. As a matter of fact, besides easing interoperability between systems, our ontology can be used to build a tool to control consistency and coherence of information in an EPR.

Rule 1. Mastectomy performed for a non-metastatic breast cancer that does not have recurrences has a primary role. The rule uses SQWRL set operators to express Negation as failure. *has_new_stage* (*?mm2,?rec*) _is the property that links and individual of the class *Malignant_neoplasm* with one of the class *Recurrence*.

Breast(?br) ∧ Malignant_neoplasm(?mm1) ∧ sqwrl:makeSet(?Smm1,?mm1) ∧ Malignant_neoplasm(?mm2) ∧ has_new_stage(?mm2,?rec) ∧ sqwrl:makeSet(?Smm2,?mm2) ∧ sqwrl:difference(?Smm,?Smm1,?Smm2) ∧ sqwrl:element(?mm,?Smm) ∧ Primary (?pi) ∧ has_location(?mm,?br) ∧ M0(?m0) ∧ has_M_staging(?mm,?m0) ∧ Treatment_description(?trd) ∧ is_applied_for(?trd,?mm) ∧ Therapy_description(?thd) ∧ is_part_of_plan(?thd,?trd) ∧ Mastectomy_quality(?mq) ∧ has_quality(?thd,?mq)→has_quality(?thd,?pi)

Rule 2. Mastectomy has role primary when performed for a non-metastatic breast cancer that recurred after the surgical intervention.

Breast(?br) ∧ Malignant_neoplasm(?mm) ∧ has_location(?mm,?br) ∧ M0(?m0) ∧ has_M_staging(?mm,?m0) ∧ Treatment_description(?trd) ∧ is_applied_for(?trd,?mm) ∧ Therapy_description(?thd) ∧ is_part_of_plan(?thd,?trd) ∧ Mastectomy_quality(?mq) ∧ has_quality(?thd,?mq) ∧ has_new_stage(?mm,?rec) ∧ has_temporal_entity(?thd,?vptd) ∧ has_temporal_entity(?rec,?vprec) ∧ temporal:hasStartTime(?vptd,?sttd) ∧ temporal:hasStartTime(?vrect,?strec) ∧ temporal:after(?strec,?sttd) ∧ Primary(?pi)→has_quality(?thd,?pi)

Rule 3. Chemotherapy for a non-metastatic breast cancer administered before a primary therapy (mastectomy) has neoadjuvant role.

Treatment_description(?trd) ∧ Malignant_neoplasm(?mm) ∧ is_applied_for(?trd,?mm) ∧ Therapy_description(?ctd) ∧ is_part_of_plan(?ctd,?trd) ∧ Chemotherapeutic(?chq) ∧ has_quality(?ctd,?chq) ∧ Therapy_description(?ptd) ∧ is_part_of_plan(?ptd,?trd) ∧ Primary(?pi) ∧ has_quality(?ptd,?pi) ∧ has_temporal_entity(?ctd,?vpctd) ∧ has_temporal_entity(?ptd,?vpptd) ∧ temporal:hasFinishTime(?vpctd,?etctd) ∧ temporal:hasStartTime(?vpptd,?stptd) ∧ temporal:before(?etctd,?stptd) ∧ Neo_adjuvant (?nadj)→has_quality(?ctd,?nadj).

Conflict of interest statement

None declared.

References

[1] F. Pinciroli, D.M. Pisanelli, The unexpected high practical value of medical ontologies, Comput. Biol. Med. 36 (7–8) (2006) 669–673.
[2] D.M. Pisanelli (Ed.), IOS-Press, Amsterdam, 2004.
[3] K. Munn, B. Smith (Eds.), Ontos Verlag, Frankfurt, 2008.
[4] B. Smith, W. Ceusters, Ontology as the core discipline of biomedical informatics. Legacies of the past and recommendations for the future direction of research, in: G.D. Crnkovic, S. Stuart (Eds.), Computing, Philosophy, and Cognitive Science, Cambridge Scholars Press, Cambridge, 2006.
[5] A.L. Rector, S. Brandt, N. Drummond, M. Horridge, C. Puleston, R. Stevens, Engineering use cases for modular development of ontologies in OWL, Appl. Ontol. 7 (2) (2012) 113–132.
[6] J. Simon, M. Dos Santos, J. Fielding, B. Smith, Formal ontology for natural language processing and the integration of biomedical databases, Int. J. Med. Inform. 75 (3–4) (2006) 224–231.
[7] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support, Yearb. Med. Inform. (2008) 67–79.
[8] A. Kumar, B. Smith, D.M. Pisanelli, A. Gangemi, M. Stefanelli, An ontological framework for the implementation of clinical guidelines in health care organizations, Stud. Health Technol. Inform. 102 (2004) 95–107.
[9] Q. Chong, A. Marwadl, K. Supekar, Y. Lee, Ontology based metadata management in medical domains, J. Res. Prac. Inform. Technol. 35 (2) (2003) 139–153.
[10] H. Min, F.J. Manion, E. Goralczyk, Y.N. Wong, E. Ross, J.R. Beck, Integration of prostate cancer clinical data using an ontology, J. Biomed. Inform. 42 (6) (2009 Dec) 1035–1045.
[11] M. Brochhausen, A.D. Spear, C. Cocos, G. Weiler, L. Martín, A. Anguita, H. Stenzhorn, E. Daskalaki, F. Schera, U. Schwarz, S. Sfakianakis, S. Kiefer, M. Dörr, N. Graf, M. Tsiknakis, The ACGT Master Ontology and its applications —towards an ontology-driven cancer research and management system, J. Biomed. Inform. 44 (1) (2011) 8–25.
[12] D.M. Pisanelli, C. De Lazzari, M. Battaglia, ROME: a reference ontology in medicine, in: H. Fujita, D.M. Pisanelli (Eds.), Proceedings of the 2007 Conference on New Trends in Software Methodologies, Tools and Techniques, IOS Press, Amsterdam, 2007.
[13] Ravi D. Shankar, Susana B. Martins, J.O'Connor Martin, David B. Parrish, Amar K. Das, Towards Semantic Interoperability in a Clinical Trials Management System, in: Isabel Cruz et al. (Eds.) The Semantic Web – ISWC 006, Proceedings of the 5th Internaional Semantic web Conference, ISWC 2006, Athens, GA, USA, November 2006. LNCS 4273, Springer.
[14] P. Johnson, S. Tu, M. Musen, I. Purves, A virtual medical record for guideline-based decision support, in: Proceedings of AMIA Symposium, 2001 pp. 294–298.
[15] S. Quaglini, S. Panzarasa, A. Cavallini, G. Micieli, C. Pernice, M. Stefanelli, Smooth Integration of Decision Support into an Existing Electronic Patient Record, in: Proceedings of 10th Conference on Artificial Intelligence in Medicine (AIME 2005), 2005 pp. 89–93.
[16] I. Román, L. Roa, G. Madinabeitia, A. Millán, Introducing guideline management in the healthcare information system architecture, in: L. Bos, B. Blobel (Eds.), Medical and Care Compunetics 4. Volume 127 of Stud Health Technol Inform, IOS Press, Amsterdam, 2007, pp. 117–124.
[17] C. Eccher, A. Seyfang, A. Ferro, S. Stankevich, S. Miksch, Bridging an Asbru protocol to an existing electronic patient record, in: D. Ria~no, A. ten Teije, S. Miksch, M. Peleg (Eds.), Knowledge Representation for Health-Care: Data, Processes and Guidelines. Revised Selected Paper of the KR4HC 2009 Workshop. Volume 5943 of LNAI, Springer, 2010, pp. 14–25.
[18] National Comprehensive Cancer Network, NCCN Clinical Practice Guidelines in Oncology, Breast Cancer. Version 2., 2011.
[19] A. Fasih, H. Fonge, Z. Cai, J.V. Leyton, I. Tikhomirov, S.J. Done, R.M. Reilly, (111) In-Bn-DTPA-nimotuzumab with/without modification with nuclear translocation sequence (NLS) peptides: an Auger electron-emitting radioimmunotherapeutic agent for EGFR-positive and trastuzumab (Herceptin)-resistant breast cancer, Breast Cancer Res. Treat. (2012) . (Epub ahead of print).
[20] S. de Coronado, M.W. Haber, N. Sioutos, M.S. Tuttle, L.W. Wright, NCI Thesaurus: using science-based terminology to integrate cancer research results, Stud. Health Technol. Inform. 107 (Pt 1) (2004) 33–37.

[21] C. Eccher, A. Seyfang, A. Ferro, S. Miksch, Embedding oncologic protocols into the provision of care: the oncocure project, Stud. Health Technol. Inform. 150 (2009) 663–667.
[22] C. Eccher, A. Ferro, D.M. Pisanelli, An ontology of therapies, in: P. Kostkova (Ed.), eHealth 2009, LNICST, 27, Springer, 2010, pp. 139–146.
[23] Laboratory for Applied Ontology. DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering. Available from: ⟨http://www.loa-cnr.it/DOLCE.html⟩. Last visited, October 18 2012.
[24] K. Hawley, "How Things Persist", Clarendon Press, Oxford, 2001.
[25] M.J. O'Connor, A.K. Das, A lightweight Model for Representing and Reasoning With Temporal Information in Biomedical Ontologies. International Conference on Health Informatics, Valencia, Spain, 2010a.
[26] J.F. Allen, Maintaining knowledge about temporal intervals, Commun. ACM 26 (1983) 11.
[27] M.J. O'Connor, A.K. Das, SQWRL: A Query Language for OWL. OWL: Experiences and Directions (OWLED), Fifth International Workshop, Chantilly, VA, 2009.
[28] Jess, The Rule Engine for the Java platform. Available from: ⟨http://www.jessrules.com/jess/index.shtml⟩. Last visited 30th March 2012.