Clinical Data Wrangling using Ontological Realism and Referent Tracking

Werner Ceusters
Department of Biomedical Informatics
University at Buffalo
Buffalo, NY 14203, USA
Email: ceusters@buffalo.edu

Chiun Yu Hsu
Neuroscience Program
Medicine and Biomedical Sciences
University at Buffalo, NY 14260, USA
Email: chiunhsu@buffalo.edu

Barry Smith
Department of Philosophy
University at Buffalo
Buffalo, NY 14203, USA
Email: phismith@buffalo.edu

Abstract* — Ontological realism aims at the development of high quality ontologies that faithfully represent what is general in reality and to use these ontologies to render heterogeneous data collections comparable. To achieve this second goal for clinical research datasets presupposes not merely (1) that the requisite ontologies already exist, but also (2) that the datasets in question are faithful to reality in the dual sense that (a) they denote only particulars and relationships between particulars that do in fact exist and (b) they do this in terms of the types and type-level relationships described in these ontologies. While much attention has been devoted to (1), work on (2), which is the topic of this paper, is comparatively rare. Using Referent Tracking as basis, we describe a technical data wrangling strategy which consists in creating for each dataset a template that, when applied to each particular record in the dataset, leads to the generation of a collection of Referent Tracking Tuples (RTT) built out of unique identifiers for the entities described by means of the data items in the record. The proposed strategy is based on (i) the distinction between data and what data are about, and (ii) the explicit descriptions of portions of reality which RTTs provide and which range not only over the particulars described by data items in a dataset, but also over these data items themselves. This last feature allows us to describe particulars that are only implicitly referred to by the dataset; to provide information about correspondences between data items in a dataset; and to assert which data items are unjustifiably or redundantly present in or absent from the dataset. The approach has been tested on a dataset collected from patients seeking treatment for orofacial pain at two German universities and made available for the NIDCR-funded OPMQoL project.

Keywords—referent tracking, data wrangling, ontological realism

I. Introduction

One goal of ontology-based research is the integration of information residing in heterogeneous data collections in the hope that by running queries over the resultant combined data collections we will be able to answer questions that would otherwise remain unanswered [1]. Such integration can be achieved through different paradigms, including: *mediation* [2], *federation* [3], *data warehousing* [4], and, most recently, the *Ontology-Based Data Access* (OBDA) paradigm [5], which is distinguished by the fact that it keeps the data sources and conceptual layer of an information system separate and independent.

To be effective, all such paradigms require ontology-based mappings ranging not only over the database schemas but also over the data types by means of which the data are stored [6]. Research in OBDA revealed that successful information integration requires much more detail than is standardly provided: it requires also suitable mechanisms for mapping individual data *values* – rather than merely data *fields* – to corresponding instances of ontology classes – for example to patients in a clinical study. This in turn requires the specification of how identifiers for such instances can be generated from such data values in order to enable creation of an ABox suitable for answering queries relating to such instances [7]. Such specification, we believe, may well be a critical issue in the context of clinical research datasets, where (as we shall discover below) data values do not always denote what is suggested by the variable or fieldname under which they appear.

Suppose, for example, that in the record of some patient the variable *phenotypic gender* is associated with a value of either '0' or '1' – meaning 'male' or 'female,' respectively. It is then safe to create an ABox statement to the effect that this patient's phenotypic gender is an instance of the corresponding ontology class. If no data value is found, however, then it should not be assumed that the patient in question does not have a phenotypic gender. If, on the other hand a value of '2' – documented as meaning 'unknown' – is found, then this should not lead to an ABox assertion to the effect that the given patient's phenotypic gender is an instance of a special kind which is neither male nor female. The value 'unknown' provides information not about the patient, but rather about the data we have about the patient.

The problem we face in creating data value to ontology mappings from clinical research data repositories is that the information needed for such mappings is not explicitly represented in the datasets. Rather, it is scattered through various data dictionaries and instruction manuals (relating for example on how to extract and process data from responses to standardized questionnaires).

The explicit representation that is pursued by the Referent Tracking (RT) methodology is based on Ontological Realism as described in [8], and on the thesis that explicit representation can best be achieved by generating unique identifiers to all instances of ontology classes which are described – whether explicitly and implicitly – in our data. In [9] we described an algorithm to achieve explicit representation of this sort from highly structured electronic health record (EHR) data. The research questions we address here are:

(1) to what extent can a similar algorithm be used for clinical research data collections, for instance to provide information both about particulars that are implicitly referred to and about correspondences between data-items in a data set,

(2) what kinds of ambiguous and implicit information can one expect to encounter in such data collections,

(3) is it useful to set limits on the types and amounts of implicit information that we will render explicit, and

(4) is it possible to use the referent tracking methodology in combination with appropriate ontologies to provide a complete and explicit representation of clinical research datasets that will take account of the constraints and provisions typically documented in data dictionaries and other data-related sources, for instance to describe which data items are unjustifiably and redundantly present or absent?

Our hypothesis is that, even where it is not possible to provide a completely accurate RT representation of the entities in reality described by a given body of data, identifying the types of challenges to such representation would itself yield a useful resource for avoiding similar problems in future clinical research studies.

II. Materials

The work described below is part of the NIDCR-funded project *Ontology for Pain-related Mental Health and Quality of Life* (OPMQoL) which involves the integration of five datasets which – although collected independently – cover similar sorts of information about patients who experienced one or other form of orofacial pain [10]. All datasets are made available as spreadsheet tables (from here on referred to as 'source tables'). Each row in the body of each such table is a collection of data items obtained from a single patient; each column is a collection of data items resulting from some specific type of observation. If a header row is present, its cells indicate what sorts of observations are reported on in the respective columns.

The de-identified dataset used for the work described here – from here on referred to as the 'study set' – was collected from 390 patients seeking treatment for orofacial pain [11]. Inclusion criteria were that patients had at least one diagnosis according to the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) [12]. The study set comes with a variable (n=161) codebook and a technical report explaining certain dependencies and implicit assumptions [13].

III. Methods

A. Referent Tracking

RT is designed to yield data repositories whose content can be expressed as a collection of Referent Tracking Tuples (RTT) [14]. An RTT is an assertion about a particular, i.e. an entity in reality that exists in space and time [15]. Each RTT follows a semi-formal syntax which is close to the one used for instance-level relationships in the definitions of the Relation Ontology [16]. Ignoring here certain housekeeping parameters we can assert that RTT assertions about *continuants* (entities such as patients, hospitals, teeth, jaws which endure through time, as contrasted with *occurrents* or *processes*), are of the form *'x p-rel y t-rel t',* where:

- '*x*' is the (ideally) singular and globally unique instance identifier (IUI) denoting the particular described,
- '*y*' is either: (1) a IUI denoting another particular or: (2) a representational unit drawn from either a realism-based ontology or a concept-based terminology,
- '*p-rel*' expresses a relationship obtaining between the referents of *x* and *y*,
- '*t*' denotes a particular temporal region, and
- '*t-rel*' expresses the relationship obtaining between the temporal region denoted by *t* and the temporal region during which *p-rel* obtains between *x* and *y*.

RTT assertions that do not mention a continuant have the form '*x p-rel y*,' where '*x*', '*p-rel*' and '*y*' are otherwise treated in the way described above.

RT aims to do away with the ambiguity in assertions such as '*John has a benign duodenal polyp*'. This assertion tells us that there exists *some* instance of a given type, but not *which one in particular*. This ambiguity is preserved in John's EHR, where diagnostic codes drawn from some terminology or ontology are used to assert existence in John at some time t_1 of polyps of a given type. The consequence is that, when a later assertion is added to John's EHR to the effect that he has a malignant duodenal polyp, the data provides no basis for inferences concerning whether it is the very same polyp as the one referred to at t_1 that has turned malignant or some other polyp appearing at some later time t_2 [14]. This ambiguity disappears when we represent the first-described situation using the following RTTs:

- #1 part-of #2 at t_1 (1)
- #1 instance-of benign duodenal polyp at t_1 (2)
- #1 instance-of malignant duodenal polyp at t_1 (3)

where '#1' denotes the polyp and '#2' John. The alternative situation, would be represented by using distinct IUIs for each polyp as follows, where '#3' denotes a second polyp:

- #1 part-of #2 at t_1 (4)
- #3 part-of #2 at t_2 (5)

L	Var	IT	REF	Min	Max	Val	IUI(L)	IUI(P)	P-Type	P-Rel	P-Targ	Trel	Time
1		IM	patient_study_record					#psrec-	DATASET-RECORD			at	t
2	id	LV	patient_identifier				#pidL-	#pid-	DENOTATOR	**denotes**	#pat-	at	t
3	id	IM	patient				#patL-	#pat-	PATIENT			at	t
4	sex	CV	gender				#patgL-	#patg-	GENDER	**inheres-in**	#pat-	at	t
5	sex	CV	male			0		#patg-	MALE-GENDER	**inheres-in**	#pat-	at	t
6	sex	CV	female			1		#patg-	FEMALE-GENDER	**inheres-in**	#pat-	at	t
7	sex	UA	sex	BLANK	BLANK			#patgL-	UNDERSPEC-ICE			at	t
8	q3	CV	no_pain_in_ lower_face			0	#q3L0-	#pat-		**lacks-pcp**	PAIN	at	#tq3-
9	q3	CV	pain_in_ lower_face			1	#q3L1-	#pq3-	PAIN	**participant**	#pat-	at	#tq3-
10	q3	IM	in_the_past_month					#tq3-	MONTH-PERIOD				
11	q3	IM	lower_face					#patlf-	LOWER-FACE	**part-of**	#pat-	at	t
12	q3	IM	time_of_q3_concretization					#cq3-	TIME-PERIOD	**after**	#tq3-		
13	q3	RP	an_8_gcps_1	0	0	0	#q3L-	#q3L-		**co-ref-with**	#q3L0-	at	t
14	q3	UP	an_8_gcps_1	1	10	0	#q3L-	#q3L-	DISINFORMATION			at	t
15	q3	UA	an_8_gcps_1	BLANK	BLANK	1	#q3L-	#q3L-	UNDERSPEC-ICE			at	t
16	q3	JA	an_8_gcps_1	BLANK	BLANK	0	#q3L-	#q3L-	J-BLANK-ICE			at	t

Table 1: Simplified template for data expansion of the variables ('Var') 'id', 'sex' and 'q3' of the original dataset ignoring time-related information. **Legend:** 'L' = Line number in this table; 'IT' = Information Type (possible values being 'LV' = Literal Value, 'CV' = Coded Value, 'UA' = Unjustified Absence, 'IM' = IMplicit reference, 'RP' = Redundant Presence (RP), 'UA' = Unjustified Absence, 'JA' = Justified Absence); 'REF' = Reference; 'Min' = lowest possible value for variable; 'Max' = highest possible value for variable; 'Val' = possible value for variable; 'IUI(L)' = prefix for generating an IUI proxy for the information content entity which refers to the corresponding value for the variable under 'Var' for the patient being processed; IUI(P) = prefix for generating an IUI proxy for whatever is denoted by this information content entity; P-Type = ontological type of the entities denoted by instantiated IUI(P)s; P-Rel = relation between the entity denoted by an instantiated IUI(P) and the entity denoted by an instantiated P-Targ; 'Trel' - temporal relation; 'Time' - temporal period during which P-rel holds. Only entries relevant to the discussion in this paper are shown. See discussion section for other details.

- #1 instance-of benign duodenal polyp at t_1 (6)
- #3 instance-of malignant duodenal polyp at t_2. (7)

A further goal of RT is to make explicit all the implicit assumptions that need to be taken into account to interpret given data correctly. Some of these assumptions result from the use of broken information models or from practices such as registering ICD-9-CM code 659.7 – '*Abnormality in fetal heart rate or rhythm*' – in the diagnosis field of the *mother's* EHR. The RT method is most effective when its principles are applied at the time of data collection and registration, though as shown in [17] post-hoc translations are also possible.

B. *Methodology applied*

The work reported here involved the following steps:

(1) cross-checking the study set with the variable codebook and technical report for appropriate coding of values, field names, and field descriptions,

(2) annotating the dataset with appropriate descriptions,

(3) building an executable template that makes explicit, for each of the data values, how their referents must be analyzed in RT terms; this is achieved by applying the following data expansion algorithm [9]:

a. identify all the possible particulars that are explicitly referred to by a specific data value when applied to a specific patient;

b. determine for each particular identified under (3a) whether it is a dependent or independent entity [8];

c. if a given particular is a dependent continuant, identify the independent continuant on which it depends; if an entity is an occurrent, identify the continuants which participate in it;

d. repeat steps (3b) and (3c) as required;

(4) selecting from appropriate realism-based ontologies the representational units that denote universals or defined classes whose instances or members are either directly referred to in the dataset or implicitly referred to as discovered through application of the algorithm described in (3);

(5) implementing an algorithm that uses outputs from (3) and (4) to generate for each patient described in the dataset a collection of RTTs that provides a realism-based representation of that patient's situation;

(6) generating statistics needed to answer the research questions described in the INTRODUCTION, above.

IV. RESULTS

Research questions (1) and (4) are answered by our development of a technical approach which enables the creation for each dataset of a template which, when applied to a particular record in the dataset, yields a corresponding collection of RTTs. Part of the approach is captured in Table 1, which shows a simplified version of some sample lines (indexed under 'L') as they appear in the template produced at step (3) (under METHODS, above) for the variables 'id', 'sex' and 'q3'. What the template lines encode is determined by the

	Template			Patients		
	Av. (SD)	Min	Max	Av. (SD)	Min	Max
CV	3.57 (2.27)	0	11	0.82 (0.38)	0	1
IM	2.79 (1.43)	0	6	2.69 (1.46)	0	6
UA	0.16 (1.02)	0	12	0.01 (0.09)	0	10
JA	0.16 (1.02)	0	12	0.04 (0.34)	0	12
RP	0.13 (0.98)	0	12	0.01 (0.10)	0	11
UP	0.13 (0.98)	0	12	0.00 (0.01)	0	5

Table 2. Occurrence of Record Types (see Table 1) per variable (n=161) in the study set for the template (left block) and per patient (n=390) after application of the template (right block).

information type (IT), the detailed semantics of which is described in section V. Common to all information types is that part of the template that appears to the left of the dashed vertical line in Table 1. This specifies the conditions which must be satisfied if RTTs are to be generated on the basis of the information provided to the right of this line.

Table 2 answers research questions (3) and (4) by providing statistics relating to the lines from out of which the data translation template for the study set is composed, on the extent to which each of these lines were in fact applied to the patient population described in the study set. The table shows, for instance, that *unjustified absences* and *presences* were encountered, albeit in a small percentage of cases, and that on average for each variable and for each patient roughly 3 *implicit particulars* needed to be accounted for. It shows that the increase in the size of the dataset resulting from applying this methodology is, for the Halle-Leipzig dataset, roughly 300%, and also that the quality of this dataset (measured in terms of UA, RP and UP) is quite good.

V. Discussion

Our vision is that the Big Data repositories of the future should be maximally explicit and maximally self-explanatory. By ‘maximally explicit’, we mean that each such repository should contain explicit reference to any and all the entities, including their interrelationships, that must exist for an assertion encoded in the repository to be a faithful representation of the corresponding part of reality. By ‘maximally self-explanatory’ we mean that the data in the repository should be presented in such a way that a researcher seeking to query the repository does not need to concern himself with any idiosyncrasies of and between datasets, or codes or formats, that were combined or used to build the repository. A strategy to achieve this is to submit to such a repository only individual datasets which are themselves maximally explicit and self-explanatory.

Our approach is based on the – to us – obvious distinction between data and what data are about. It then takes advantage of the fact that RTTs can be used to describe in explicit fashion not merely the portions of reality described by data items in a dataset, but also these data items themselves. This allows us to describe explicitly even those particulars that are only implicitly referred to in a dataset by *generating* suitable unique identifiers. It also allows us to provide information about correspondences (such as co-reference) between data items in a dataset, and also to assert which data items are redundant, or unjustifiably absent, and so forth.

A. Explicit data items

The study set contains some explicit data items which are about particulars on the side of the patient such as gender, facial pains experienced, clicking noises heard when opening their mouths, and so forth. Referent Tracking requires each of these particulars to be assigned an IUI; Ontological Realism tells us that each one of them is instance of at least one universal. What universals these particulars are instances of is typically only very indirectly represented in the study set.

The strategy for translating explicit data items into RTTs is covered by the Literal Value (LV) and Coded Value (CV) records in the template (Table 1). Template lines of either type have under ‘REF’ the label obtained or constructed from the relevant data dictionary or other supporting information associated with the code value. The template shows, for example, that if, for a patient in the study set, the value for the variable ‘sex’ is ‘0’ (L5), then the gender of this patient is described as ‘male.’ This can be translated in RT terms into a assertion that the given patient’s gender is an instance of the universal *male gender* (or, in case gender does not qualify as a universal [18], that it is a member of the defined class ‘male gender’ – we will ignore this distinction in the remainder of this paper).

The IUIs assigned through application of our method are in reality very large numbers generated by an RT system to ensure the needed high probability of uniqueness. For the sake of readability, however, we provide simple abbreviations to stand in for these IUIs. We also leave out full specification of time-related information (which would be needed, for example, to deal with cases where a patient’s gender changes from one time to the next), and certain housekeeping details required by syntactically and semantically correct RTTs [15]).

To see how IUI assignment works, now, we will suppose that, while processing the study set on the basis of the template illustrated in Table 1, the IUI *#pat-1* is assigned to the first patient described and that *#patg-1* is assigned to his gender. Then the following collection of assertions would be generated as part of a faithful RT-like representation of the corresponding portion of reality (POR) on the basis of lines L3 and L5 of the template:

- *#pat-1* **instance-of** PATIENT **at** *t* (8)
- *#patg-1* **instance-of** MALE-GENDER **at** *t* (9)
- *#patg-1* **inheres-in** *#pat-1* **at** *t* (10)

Of course, the study set, too, is a particular, and so also are the data items from out of which it is built. According to the Information Artifact Ontology (IAO) the study set and its parts are particular concretizations of particular information content entities (ICEs). Thus the ‘0’ in a particular position of the spreadsheet on your screen indicating that *#pat-1*’s gender is male could be assigned an IUI, as also could the corresponding bits on the hard drive of your laptop which bring it about that your spreadsheet software causes the laptop to display the '0' in that position. In addition, also the ICEs here concretized can be assigned IUIs of their own. For example in L1 of the template the IUI *#psrec-1* is assigned to the ICE that is concretized on your screen as a row of the patient's record, and in L4 *#patgL-1* is assigned to the ICE whose concretizations inform us what the gender of *#pat-1* is. Since referent tracking implementations also assign IUIs to RTTs, *#RTT-patg-1-RN5a* would be assigned to the ICE of which assertion (9) which is generated by L5 is a concretization. On this basis, now, the following assertions can be added:

- *#patgL-1* **component-of** *#psrec-1* **at** t (11)
- *#RTT-patg-1-RN5a* **instance-of** RTT **at** t (12)
- *#patgL-1* **co-ref-with** *#RTT-patg-1-RN5a* **at** t (13)
- *#patgL-1* **instance-of** DATA-ITEM **at** t (14)
- *#patgL-1* **is-about** *#patg-1* **at** t (15)
- *#psrec-1* **instance-of** DATASET-RECORD **at** t (16)

Assertions of types (11) and (14) are generated whenever an IUI(L) – here *#patgL-1 – is* for the first time generated while processing the data for a specific patient. Assertions of type (15) are generated wherever IUI(L) and IUI(P) values co-occur in a template line. Assertions of types (12) and (13) are generated for all template lines in which there is both (1) a value for P-Rel and (2) a condition expressed in the left part of Table 1 that is satisfied by a data item in the original dataset. Assertion (16) expresses the assertional content of L1. The **co-ref-with** relationship – short for 'co-referential-with' – used in (13) holds between two ICEs whenever concretizations thereof describe the same portion of reality (POR). Both ICEs then (in harmony with talk of a 'correspondence theory of truth') enjoy a **corresponds-to** relationship with the same POR. Where the assertions (8) to (10) describe parts of first-order reality, (11) to (14) describe the second-order entities that have some sort of aboutness relation with these first-order items. Assertion (15) provides the link between the two.

B. Referencing implicit information

The variable 'q3' in the study set holds responses to the question '*Have you had pain in the face, jaw, temple, in front of the ear or in the ear in the past month?*' A positive answer is encoded as '1,' a negative one as '0'. Although certain particulars on the side of the patient to whom the question is addressed (for example his jaw, temple, the past month, etc.) are explicitly referred to in the question, they are only implicit in admissible responses. To achieve our objective, explicit reference is required, which is achieved by means of IM-records, all of which have under 'REF' a textual reference to an entity – or configuration of entities [15] – that must exist for the corresponding 'Var' to make sense. IM-records – in this case L10, L11 and L12 – are generated manually by applying step (3) of the data expansion algorithm described under METHODS above. When the template is used to generate assertions about *#pat-1*, a negative answer to question q3 (L8) would generate an RTT to the effect that the patient lacks participation in an instance of pain – we view such instances as processes [19] – by using the *lacks*-family of relations for the expression of negative findings [20]. In case of a positive answer, an IUI for the appropriate instance is generated and participation of the patient therein is asserted. Both answers generate IUIs for the patient's lower face, the time when the question was asked, and the period of one month prior to the asking: all of these entities do indeed exist whatever answer is given.

C. (Un)justified presence and absence

Template lines of types UA, UP, RP, and JA make explicit whether there are missing data or data that should not be there.

L7, for instance, brings it about that when, for patient *#pat-1* in the study set, no value for the variable 'sex' is provided – expressed by the appearance of 'BLANK' in the template under both 'Min' and 'Max' – an RTT is generated that declares the data item *#patgL-1* to be an instance of an underspecified ICE. This assertion does not mean that the data item itself is absent; rather it means that certain information is missing.

An absence or presence of a value for some variable may be justified or unjustified depending on the value of some other variable. The last four lines in Table 1, for example, describe dependencies between the variables 'q3' (for which the possible values '1' and '0' mean, respectively, current presence or absence of pain) and 'an_8_gcps_1', the latter containing answers to the question '*How would you rate your facial pain on a 0 to 10 scale at the present time, that is right now, where 0 is "no pain" and 10 is "pain as bad as could be"?*' L13 states that when the values for both 'q3' and 'an_8_gcps_1' are '0', then the two ICEs of which the coding for the answers are concretizations enjoy a **corresponds-to** relation to the same portion of reality.

L16 asserts that, if a record in the dataset has a '0' value for the variable q3, and if there is no value for the variable 'an_8_gcps_1', then the absence of a value for 'an_8_gcps_1' is justified. This is then documented by means of an RTT to the effect that the corresponding ICE is justifiably blank (as concretized by, for instance, an empty cell in that part of the spreadsheet). As a last example, L14 asserts that if the value given for 'an_8_gcps_1' is between 1 and 10 while the value for q3 is 0, then the value for the former is unjustifiably present (the corresponding ICE must thus be classified as *disinformation* – as dictated by the coding guidelines for the corresponding pair of questions).

D. Limitations

To achieve the vision of maximally self-explanatory and explicit data repositories, several issues will need to be addressed. We will need above all a fully adequate set of relations for the various flavors of aboutness and correspondence, and a better theory of ICEs, for instance concerning the various types that exist and how they relate to concretizations and to each other; these issue are currently not addressed in the Information Artifact Ontology or any other realism-based ontology.

VI. Conclusion

We have presented the beginnings of a methodology that allows a clinical research dataset to be translated into a set of of Referent Tracking Tuples that has the following features: not only the portion of reality described by the dataset and the dataset itself are represented in a way that mimics the structure of reality, but so also are the relations between components of this dataset on the one hand and the corresponding portions of reality on the other. Applying the methodology to a concrete dataset and performing some basic exploratory statistics revealed that all of the relations we distinguished between data items and what they are about (if, indeed, they are about anything at all) do indeed occur in our study data. A set of RTTs of this sort may in the future perhaps replace the more complicated exchange information models that are used in message-based paradigms or in the Extract – Transform – Load (ETL) analyses and procedures used in data warehousing. Although the syntax and semantics of RTTs seems to us to be powerful enough to represent what is required, a current limitation is the insufficient development of the Information Artifact Ontology. A second limitation is that not all RTTs can easily be translated into OWL-based languages. Where the former is a job to be done by ontologists, the latter is a task for computer science.

Acknowledgements

This work was funded in part by grant 1R01DE021917-01A1 from the National Institute of Dental and Craniofacial Research (NIDCR). The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCR or the NIH.

References

[1] Haas L. Beauty and the Beast: The Theory and Practice of Information Integration. In: Schwentick T, Suciu D, editors. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag 2007. p. 28-43.

[2] Marenco L, Wang R, Nadkarni P. Automated Database Mediation Using Ontological Metadata Mappings. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):723-37.

[3] Sim I, Carini S, Tu SW, Detwiler LT, Brinkley J, Mollah SA, et al. Ontology-Based Federated Data Access to Human Studies Information. In:AMIA Annu Symp Proc 2012. Chicago IL2012. p. 856-65.

[4] Baumbach J, Brinkrolf K, Czaja LF, Rahmann S, Tauch A. CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. BMC genomics. 2006;7:24.

[5] Rodriguez-Muro M, Calvanese D. Dependencies: Making Ontology Based Data Access Work In Practice. . Proc of the 5th Alberto Mendelzon Int Workshop on Foundations of Data Management (AMW 2011) 2011.

[6] Kohler J, Philippi S, Lange M. SEMEDA: ontology based semantic integration of biological databases. Bioinformatics. 2003 Dec 12;19(18):2420-7.

[7] Poggi A, Lembo D, Calvanese D, Giacomo GD, Lenzerini M, Rosati R. Linking data to ontologies. In: Spaccapietra S, editor. Journal on data semantics X. Heidelberg: Springer-Verlag; 2008. p. 133-73.

[8] Smith B, Ceusters W. Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies. Applied Ontology. 2010;5(3-4):139-88.

[9] Rudnicki R, Ceusters W, Manzoor S, Smith B. What Particulars are Referred to in EHR Data? A Case Study in Integrating Referent Tracking into an Electronic Health Record Application. In: Teich JM, Suermondt J, C H, editors. American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy. Chicago, IL2007. p. 630-4.

[10] Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. Stud Health Technol Inform. 2012;180:68-72.

[11] John MT, Reißmann D, Schierz O, Wassell RW. Oral health-related quality of life in patients with temporomandibular disorders. Journal of Orofacial Pain. 2007;21(1):46-54.

[12] Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications. Journal of Craniomandibular Disorders. 1992;6(4):301-55.

[13] Mancl L, Whitney C, Zhu X. A SAS computer program to evaluate the research diagnostic criteria for classification of temporomandibular disorders: University of Washington1999 June 3.

[14] Ceusters W, Smith B. Strategies for Referent Tracking in Electronic Health Records. Journal of Biomedical Informatics. 2006 June;39(3):362-78.

[15] Ceusters W, Manzoor S. How to track Absolutely Everything? In: Obrst L, Janssen T, Ceusters W, editors. Ontologies and Semantic Technologies for the Intelligence Community Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press; 2010. p. 13-36.

[16] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biology. 2005;6(5):R46.

[17] Hogan WR, Garimalla S, Tariq S, Ceusters W. Representing Local Identifiers in a Referent-Tracking System. In: Smith B, editor. Proceedings of the International Conference on Biomedical Ontology. Buffalo NY2011. p. 252-4.

[18] Ceusters W, Smith B. A Unified Framework for Biomedical Terminologies and Ontologies. In: Safran C, Marin H, Reti S, editors. Proceedings of the 13th World Congress on Medical and Health Informatics (Medinfo 2010), Cape Town, South Africa, 12-15 September 2010. Amsterdam: IOS Press; 2010. p. 1050-4.

[19] Smith B, Ceusters W, Goldberg LJ, Ohrbach R. Towards an Ontology of Pain. In: Okada M, editor. Proceedings of the Conference on Logic and Ontology. Tokyo: Keio University Press; 2011. p. 23-32.

[20] Ceusters W, Elkin P, Smith B. Negative Findings in Electronic Health Records and Biomedical Ontologies: A Realist Approach. International Journal of Medical Informatics. 2007 March;76:326-33.