# Generating explicit data repositories from clinical research datasets using Referent Tracking: challenges and strategies

Werner Ceusters[1,*] Chiun yu Hsu[2]

[1] Institute for Healthcare Informatics, University at Buffalo, 923 Main Street, Buffalo NY, 14214, USA
[2] Neuroscience Program, School of Medicine and Biomedical Sciences, University at Buffalo, USA

## ABSTRACT

The goal of Ontological Realism is the development of high quality ontologies that faithfully represent what is general in reality with the further goal to use these ontologies to render heterogeneous data collections comparable. The latter would be for clinical research datasets an easy task if (1) the required ontologies do exist, and (2) the data sets were equally faithful to reality, i.e. if their data items would denote exclusively particulars - and relationships amongst them - in terms of the types and type-level relationships described in these ontologies. While much attention is currently devoted to (1), there is not much work on (2), which is the topic of this paper. Using Referent Tracking as a basis, we describe a technical solution which is to create for each data set a template that when applied to a particular record in the dataset leads to the generation of a collection of Referent Tracking Tuples (RTT) for what is described in the record by means of data items.

The approach proposed is based on (1) the distinction between data and what data are about, and (2) the explicit way in which RTTs can describe portions of reality involving not only particulars described by data items in a dataset, but also the data items themselves. This allows to describe particulars that are implicitly referred to, to provide information about correspondences between data items in a dataset, to assert which data items are unjustifiably or redundantly present or absent, and to provide detailed statistics about the occurrence in specific data sets of each of these issues. The approach is successfully tested on a dataset collected from patients seeking treatment for orofacial pain at the Department of Prosthodontics, Martin Luther University, and the Department of Prosthodontics and Materials Sciences, University of Leipzig, Germany and made available for the NIDCR-funded OPMQoL project. We conclude that a collection of RTTs generated through the application of this method can yield a maximally explicit and self-explanatory representation of the portion of reality described by the corresponding dataset, modulo issues for which further research is required.

## 1 INTRODUCTION

One goal for using ontologies is the integration of information residing in heterogeneous data collections in the hope that queries run over such combined data collections can answer questions that would remain unanswered otherwise (Haas, 2007). Ontology-based information integration can be achieved through different paradigms such as, for example, *mediation* (Marenco, Wang, & Nadkarni, 2009), *federation* (Sim et al., 2012), *data warehousing* (Baumbach, Brinkrolf, Czaja, Rahmann, & Tauch, 2006) or the newer *Ontology-Based Data Access* (OBDA) paradigm (Rodriguez-Muro & Calvanese, 2011) which keeps the data sources and the conceptual layer of an information system separate and independent. It is well recognized that all paradigms require not only some form of ontology-based mapping between the database schemas but also of the semantics of the data as well as of the data types by means of which they are stored (Kohler, Philippi, & Lange, 2003). Research in OBDA, thereby inspired by the Semantic Database paradigm of the eighties, made it clear that for information integration to work adequately much more detail is required and that defining suitable mechanisms for mapping individual data *values* - rather than merely data *fields* - to corresponding instances of ontology classes, and specifying how instance identifiers can be built or resolved starting from data values in order to build a suitable ABox for answering a specific question, are of not less importance (Poggi et al., 2008). The latter, so we believe, may well be a critical issue in the context of clinical research datasets as data values do not always denote what is suggested by the variable or fieldname under which they appear.

If, for example, in a patient's record for the variable *phenotypic gender* a value of either '1' or '2' - meaning resp. 'male' or 'female' - is found, it is safe to create an ABox statement that asserts that patient's phenotypic gender to be an instance of the corresponding ontology class. If however no data value is found, then it should not be assumed that that patient does not have a phenotypic gender at all. If at the other hand a value of '3' - documented as being 'unknown' - is found, it should not lead to an ABox assertion stating that that patient's phenotypic gender is an instance of a special kind which is not male or female.

The problem for making data value to ontology mappings from clinical research data repositories is that the information required to do so adequately is not explicitly represented in the datasets, but scattered over data dictionaries, guidelines for obtaining data through standardized questionnaires and how to process them, and so forth. Explicit representation is the main driver for Referent Tracking (RT), a methodology which is based on Ontological Realism (Smith & Ceusters, 2010), and for which an algorithm has been described to recover ambiguous and implicit information from highly structured EHRs (Rudnicki, Ceusters, Manzoor, & Smith, 2007). The questions we address here are:

[*] To whom correspondence should be addressed: ceusters@buffalo.edu

(1) to what extent can a similar algorithm be used for clinical research data collections,

(2) what kind of ambiguous and implicit information can one expect to encounter,

(3) is it useful to set limits on the type and amount of implicit information to render explicit, and

(4) is it possible to use RTTs in combination with appropriate ontologies to provide a complete and explicit representation of clinical research datasets including all constraints and provisions documented in data dictionaries or other data-related sources ?

The hypothesis is that even if it would not be possible to provide a completely accurate RT representation of the part of reality described by the data, identifying the type of challenges itself would yield a useful resource to avoid similar problems in future clinical research studies.

## 2 MATERIALS

Part of the NIDCR-funded OPMQoL project - Ontology for Pain-related Mental Health and Quality of Life - involves the integration of five datasets of which the data have been collected independently from each other, yet cover similar sorts of information about patients which experienced one or other form of orofacial pain (Ceusters, 2012). All datasets are made available as spreadsheet tables (from here on called 'source tables'). Each row of such a source table - except the header row in case one is present - is a collection of data items obtained from a single patient and each column a collection of data items resulting from some specific type of observation. If a header row is present, the cells of it indicate what sort of observations are reported on in the respective columns.

The de-identified dataset used for the work described here - we call this set from here on the 'study set' -was collected from 390 patients seeking treatment for orofacial pain at the Department of Prosthodontics, Martin Luther University, and the Department of Prosthodontics and Materials Sciences, University of Leipzig, Germany (John, Reißmann, Schierz, & Wassell, 2007). Inclusion criteria were that patients had at least one diagnosis according to the German version of the Research Diagnostic Criteria for Temporomandibular Disorders (Dworkin & LeResche, 1992). The study set comes with a variable (n=161) codebook and a technical report explaining certain dependencies and implicit assumptions related to the RDC/TMD part of the dataset (Mancl, Whitney, & Zhu, 1999).

## 3 METHODS

### 3.1 Referent Tracking

RT is designed to build data repositories of which the content can be expressed as a collection of Referent Tracking Tuples (RTT) (Ceusters & Smith, 2006). An RTT is an assertion about a particular, i.e. an entity in reality that carries identity (Ceusters & Manzoor, 2010). 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 (Smith et al., 2005).

The core - leaving out house keeping parameters - of assertions about relationships in which a continuant is involved 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) an 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 exact, i.e. maximal, temporal region during which p-rel obtains between x and y. Whereas in (Smith et al., 2005) '*t-rel t*' is restricted to '*at t*' in the meaning of '*obtains at least during t, perhaps also at other times*', RTTs can deal with all temporal relationships defined in (European Committee for Standardization, 2005).

The syntax of RTTs that do not mention a continuant is of the form '*x p-rel y*' where '*x*', '*p-rel*' and '*y*' carry the same meaning as just sketched.

One goal of RT is to do away with the ambiguity in assertions such as '*John has a benign duodenal polyp*' or the equivalent thereof by registering in John's electronic health record (EHR) the corresponding diagnostic code drawn from a terminology or ontology. The ambiguity is that such assertions refer to the existence of *some* instance of a given type, but not to *which one in particular*. As a consequence, a later assertion in John's EHR to the effect that he has a malignant duodenal polyp does not allow inferences to be drawn whether it is the very same polyp that turned malignant or another one (Ceusters & Smith, 2006). The ambiguity disappears by representing the former situation using the following RTTs:

- #1 part-of #2 at t1 (1)
- #1 instance-of benign duodenal polyp at t1 (2)
- #1 instance-of malignant duodenal polyp at t1 (3)

where '#1' denotes the polyp and '#2' John. The latter situation would be represented by using distinct IUIs for each of the polyps as follows:

- #1 part-of #2 at t1 (4)
- #3 part-of #2 at t2 (5)

| RN | Var | RT | 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- | | **corresp-w** | #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' ignoring time-related information. Legend: 'RN' - template Record Number, 'RT' - Record Type, 'REF' - Reference, 'Min' - lowest possible value for variable, 'Max' - highest possible value for variable, 'Val' - possible value for variable, 'IUI(L)' - IUI template for the IUIs of information content entities of which data items in the data set are concretizations, IUI(P) - IUI template for the entity about which one or two RTT-templates are formulated by means of the last three elements of the record, 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.

- #1 instance-of benign duodenal polyp at t1 (6)
- #3 instance-of malignant duodenal polyp at t2. (7)

Another goal of RT is to make explicit all implicit assumptions that need to be taken into account to interpret data correctly, some of which resulting from crippled information models or 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 thus most effective when its principles are applied at the time of data collection and registration although post-hoc translations are possible (Hogan, Garimalla, Tariq, & Ceusters, 2011).

## 3.2 Methodology applied

The work reported on here involved the following steps:

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

(2) annotating the dataset with appropriate descriptions,

(3) building an executable template that for each of the possible data values explains how the data value's referent must be analyzed in RT terms, thereby applying the following data expansion algorithm (Rudnicki et al., 2007):

  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 in (3b) whether it is a dependent or independent entity (Smith & Ceusters, 2010),

  c. if a 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 of which instances, resp. members, are directly referred to in the dataset or are discovered through the algorithm explained in (3),

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

(6) generating the statistics able to answer the research questions described in the introduction.

# 4 RESULTS

We succeeded in developing a technical solution which is to create for each data set a template that when applied to a particular record in the dataset leads to the generation of a collection of RTTs. Part of the mechanism to do so is captured in Table 1 which shows a simplified version of some records which are part of the template produced at step (3) (see methods section) for the variables 'id', 'sex' and 'q3'. What the records encode is determined by the record type (RT), the ones being displayed being Literal Value (LV), Coded Value (CV), Unjustified Absence (UA), IMplicit reference (IM), Redundant Presence (RP), Unjustified Absence (UA) and Justified Absence (JA). The detailed semantics for each of these template record types is described in the discussion section. Common to all record types is that the left part (left of the dashed vertical in Table1) specifies conditions which when satisfied lead to the generation of RTTs based on the information provided on the right side.

Table 2 provides some statistics on the Record Types out of which the data translation template for the study set is composed, and to what extent these Record Types became applied to the patient population described in the study set.

|  | 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).

## 5 DISCUSSION

Our vision is that 'big data' repositories should be maximally explicit and self-explanatory. By 'maximally explicit', we mean that such a repository should contain explicit reference to any and all entities in reality, including the relationships they enjoy with each other, that must exist for an assertion encoded in the repository to be a faithful representation of the corresponding part of reality. By 'self-explanatory' we mean that the repository contains the data in such a way that a researcher seeking to query the repository does not need to worry about any idiosyncrasies of and between datasets, or codes and formats that were combined to build the repository. A strategy to achieve this is to submit to such a repository individual datasets which are themselves maximally explicit and self-explanatory.

The approach proposed here is based on (1) the - for us - obvious distinction between data and what data are about, and (2) the explicit way in which RTTs can describe portions of reality involving not only particulars described by data items in a dataset, but also the data items themselves. This allows to describe particulars that are implicitly referred to, to provide information about correspondences between data items in a dataset, to assert which data items are unjustifiably or redundantly present or absent, and so forth.

### 5.1 Explicit data items

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

How to translate explicit data items into RTTs is covered by LV- and CV-records in the  template (Table 1). Records of either type have under 'REF' the label obtained - or constructed - from the data set's data dictionary or other supporting information corresponding 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', his gender is described as 'male' what can be translated in RT terms that that patient's gender is an instance of male gender - or a member of the defined class 'male gender' in case gender does not qualify as universal (Ceusters & Smith, 2010), a distinction we will not make any further in the context of this paper. If, while processing the study set on the basis of the template, the IUI *#pat-1*were assigned to the first patient described and *#patg-1* to his gender - IUIs are in reality large numbers generated by an RT system what we avoid doing here for readability - and if the study set is faithful to reality, then - again for readability ignoring both the underspecification of the time-related information and the additional detail required in syntactically and semantically correct RTTs (Ceusters & Manzoor, 2010) - the following collection of assertions would be generated as faithful RT-like representation of the corresponding portion of reality on the basis of RN3 and RN5 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, also the study set itself is a particular, and so are the data items that are part of it. According to the Information Artifact Ontology (IAO) the study set and its parts are particular concretizations of particular information content entities (ICE). 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 well as the corresponding bits on the hard drive of your laptop which are such that the spreadsheet software causes the laptop to display the '0' in that position. In addition, also the ICEs of which the former are concretizations, can be assigned IUIs as shown in Table 1. For example, processing the template would lead to the assignment of *#psrec-1* to the ICE of which we can see concretizations in the form of a row of the patient's record in the study set by means of, for instance, spreadsheet software (RN1) and *#patgL-1* to the ICE of which concretizations inform us what the gender of *#pat-1* is (RN4) . Since referent tracking implementations also assign IUIs to RTTs, *#RTT-patg-1-RN5* would be assigned to the ICE of which assertion (9) which is generated by RN5 is a concretization, after which, amongst others, the following assertions would be added:

- *#patgL-1* **component-of** *#psrec-1* **at** t (11)
- *#RTT-patg-1-RN5* **instance-of** RTT **at** t (12)
- *#patgL-1* **corresponds-with** *#RTT-patg-1-RN5* **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 the sort (11), (14) and (15) are generated for all IUI(L)-IUI(P) co-occurrences in the template, (12) and (13) for all template records in which an RTT template is specified and the conditions on the left are satisfied, and (16) because of RN1. The **corresponds-with** relationship used here holds between an ICE of which an RTT is a concretization and another ICE whenever the former **corresponds-to** - i.e. **is-about** in a way that it mimics the structure of the portion of reality which it is describes (Ceusters & Manzoor, 2010) - the same portion of reality as described by the latter. Whereas the assertions (8) to (10) describe part of first-order reality, the assertions (11) to (14) describe the second-order entities that have some sort of aboutness relation with the first-order ones. Assertion (15) provides the link between the two.

## 5.2 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?*', such that a positive answer would be encoded as '1' and a negative one as '0'. Although some particulars on the side of the patient to whom the question is asked (jaw, temple, past month, etc.) are explicitly referred to in the question, they are not so in either possible response. To achieve our objective, explicit reference to some of them is required, which is achieved by means of IM-records, all of which have under 'REF' a textual reference to an entity - or configuration (Ceusters & Manzoor, 2010) - in reality that must exist for the corresponding 'Var' to make sense. IM-records, in this case RN10, RN11 and RN12 are generated manually as a result of applying the data expansion algorithm of methods step (3). When the template is used to generate assertions about *#pat-1*, a negative answer to question q3 (RN8) would generate an RTT to the effect that the patient lacks participation in an instance of pain - pain is a process (Smith B, Ceusters W, Goldberg LJ, & Ohrbach R., 2011) - by using the lacks-family of relations for negative findings (Ceusters, Elkin, & Smith, 2007). In case of a positive answer, an IUI for the instance is generated and participation of the patient therein asserted. Both answers generate IUIs and corresponding assertions for the patient's lower face, the time when the question was asked and the period of one month prior to asking: these entities do indeed exist whatever the answer.

## 5.3 (Un)justified presence and absence

Template records with types UA, UP, RP, and JA make explicit whether there are missing data, or that there are data that shouldn't be there.

Record RN7 for instance brings about that when for patient *#pat-1* in the study set no value for the variable 'sex' is provided - this is expressed by the appearance of 'BLANK' in the template under both 'Min' and 'Max' - an RTT will be generated that declares the data item *#patgL-1* to be an instance of an underspecified ICE. The latter does not mean the data item is absent, rather that some 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 records in Table 1, for example, describe dependencies between the variables 'q3' 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"?*'. Record RN13 states that when the values for both 'q3' and 'an_8_gcps_1' are '0', the two ICEs of which the coding for the answers are concretizations **correspond-to** the same portion of reality. Record RN16 asserts that if a record in the dataset has a '0' value for the variable q3, and there is no value for the variable 'an_8_gcps_1', then the absence of a value for 'an_8_gcps_1' is justified, which then becomes documented by means of an RTT that asserts the corresponding ICE to be justifiably blank as concretized by, for instance, an empty cell in that part of the spreadsheet. As a last example, record RN14 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 is thus *disinformation* rather than information), which, in this case, is dictated by the coding guidelines for the corresponding pair of questions.

## 5.4 Limitations

To achieve the vision of maximally self-explanatory and explicit data repositories, several issues need further investigation. We need for sure a fully adequate set of relations for the various flavors of aboutness and a better theory of ICE, for instance concerning the various types that exist, how they relate to concretizations and to each other, and so forth. It is however not certain at this point that these issues do have to be dealt with for all cases.

## 6 CONCLUSION

We have presented (the beginnings of) a methodology that allows a clinical research dataset to be translated in a series of Referent Tracking Tuples such that both the portion of reality described by the dataset as well as the dataset itself and how components thereof relate to said portion of reality are represented in a way that mimics the structure of reality. Applying the methodology to a concrete dataset and performing some basic exploratory statistics revealed that the various ways in which data items can relate to what they are about (if anything at all) do indeed occur. A set of RTTs of this sort may in the future perhaps replace the overly complicated exchange information models that are used in message-based paradigms or the ETL (Extract - Transform - Load) analyses and procedures in data warehousing. Although the syntax and semantics of RTTs seems 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 in OWL-based languages. Whereas the former is a job to be done by ontologists, the latter is to be addressed by computer scientists.

## ACKNOWLEDGEMENTS

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

## REFERENCES

Baumbach, J., Brinkrolf, K., Czaja, L.F., Rahmann, S., & Tauch, A. (2006). Coryneregnet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. BMC Genomics, 7, 24. doi: 10.1186/1471-2164-7-24

Ceusters, W. (2012). An information artifact ontology perspective on data collections and associated representational artifacts. Stud Health Technol Inform, 180, 68-72.

Ceusters, W., Elkin, P., & Smith, B. (2007). Negative findings in electronic health records and biomedical ontologies: A realist approach. International Journal of Medical Informatics, 76, 326-333.

Ceusters, W., & Manzoor, S. (2010). How to track absolutely everything? In L. Obrst, T. Janssen & W. Ceusters (Eds.), Ontologies and semantic technologies for the intelligence community. Frontiers in artificial intelligence and applications. (pp. 13-36). Amsterdam: IOS Press.

Ceusters, W., & Smith, B. (2006). Strategies for referent tracking in electronic health records. Journal of Biomedical Informatics, 39(3), 362-378.

Ceusters, W., & Smith, B. (2010). A unified framework for biomedical terminologies and ontologies. In C. Safran, H. Marin & S. Reti (Eds.), Proceedings of the 13th world congress on medical and health informatics (medinfo 2010), cape town, south africa, 12-15 september 2010 (pp. 1050-1054). Amsterdam: IOS Press.

Dworkin, S.F., & LeResche, L. (1992). Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications. Journal of Craniomandibular Disorders, 6(4), 301-355.

European Committee for Standardization. (2005). En 12388:2005. Health informatics - time standards for healthcare specific problems.

Haas, L. (2007). Beauty and the beast: The theory and practice of information integration. In T. Schwentick & D. Suciu (Eds.), Lecture notes in computer science (Vol. 4353, pp. 28-43). Berlin, Heidelberg: Springer-Verlag

Hogan, W.R., Garimalla, S., Tariq, S., & Ceusters, W. (2011). Representing local identifiers in a referent-tracking system. In B. Smith (Ed.), Proceedings of the international conference on biomedical ontology (pp. 252-254). Buffalo NY.

John, M.T., Reißmann, D., Schierz, O., & Wassell, R.W. (2007). Oral health-related quality of life in patients with temporomandibular disorders. Journal of Orofacial Pain, 21(1), 46-54.

Kohler, J., Philippi, S., & Lange, M. (2003). Semeda: Ontology based semantic integration of biological databases. Bioinformatics, 19(18), 2420-2427.

Mancl, L., Whitney, C., & Zhu, X. (1999). A sas computer program to evaluate the research diagnostic criteria for classification of temporomandibular disorders Technical Report Series (Vol. 9401, pp. 44): University of Washington.

Marenco, L., Wang, R., & Nadkarni, P. (2009). Automated database mediation using ontological metadata mappings. J Am Med Inform Assoc, 16(5), 723-737.

Poggi, A., Lembo, D., Calvanese, D., Giacomo, G.D., Lenzerini, M., & Rosati, R. (2008). Linking data to ontologies. In S. Spaccapietra (Ed.), Journal on data semantics x (pp. 133-173). Heidelberg: Springer-Verlag.

Rodriguez-Muro, M., & Calvanese, D. (2011). Dependencies: Making ontology based data access work in practice. Proc. of the 5th Alberto Mendelzon Int. Workshop on Foundations of Data Management (AMW 2011). CEUR Electronic Workshop Proceedings, http://ceur-ws.org/ (Vol. 749).

Rudnicki, R., Ceusters, W., Manzoor, S., & Smith, B. (2007). 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 & H. C (Eds.), American medical informatics association 2007 annual symposium proceedings, biomedical and health informatics: From foundations to applications to policy (pp. 630-634). Chicago, IL.

Sim, I., Carini, S., Tu, S.W., Detwiler, L.T., Brinkley, J., Mollah, S.A., . . . Huser, V. (2012). Ontology-based federated data access to human studies information. In. (Ed.), Amia annu symp proc 2012 (pp. 856-865). Chicago IL.

Smith B, Ceusters W, Goldberg LJ, & Ohrbach R. (2011). Towards an ontology of pain. In M. Okada (Ed.), Proceedings of the conference on logic and ontology (pp. 23-32). Tokyo: Keio University Press.

Smith, B., & Ceusters, W. (2010). Ontological realism as a methodology for coordinated evolution of scientific ontologies. Applied Ontology, 5(3-4), 139-188.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., . . . Rosse, C. (2005). Relations in biomedical ontologies. Genome Biology, 6(5), R46.