Referent Tracking and its Applications

Werner Ceusters
CoE in Bioinformatics & Life Sciences
701 Ellicott Street
Buffalo, NY 14203
(1) 716 881 8971
ceusters@buffalo.edu

Barry Smith
Department of Philosophy,
University at Buffalo
Buffalo, NY 14260
(1) 716 645 2444
phismith@buffalo.edu

ABSTRACT

Referent tracking (RT) is a new paradigm, based on unique identification, for representing and keeping track of particulars. It was first introduced to support the entry and retrieval of data in electronic health records (EHRs). Its purpose is to avoid the ambiguity that arises when statements in an EHR refer to lesions, disorders, and other entities on the side of the patient exclusively by means of compound descriptions utilizing general terms such as 'pimple on nose' or 'small left breast tumor'. In this paper, we describe the theoretical foundations of the RT paradigm and show how it is being applied to the solution of problems of ambiguous identification in the fields of digital rights management, corporate memories and decision algorithms.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *Representations (procedural and rule-based).*

General Terms

Management, Documentation, Design, Standardization.

Keywords

Referent Tracking, Basic Formal Ontology, Referential Semantics, Knowledge Management

1. INTRODUCTION

In 1979, the late William Kent, author of *Data and Reality*, defended a view according to which integers should be employed to uniquely identify entities in the real world and to serve as their surrogates in databases: '*If everything we dealt with in a database had single, unique, simple names, then we would have no need for domain rules on joins (nor would we have to distinguish two kinds of join)*' [31]. Where the traditional join operation in relational databases connects two tuples if specified fields in the tuples contain *the same symbol*, Kent's remark refers to a proposal for a new type of join that would relate two tuples if the specified fields refer to *the same entity*. In 2003 Kent introduced in this spirit the notion of a *Globally Unique and Singular Identifier* (GUSI) and defined it as a computational surrogate that can be placed in one-to-one correspondence with the things they denote, thereby satisfying the principles of (1) *globality*: a GUSI is recognized throughout the universe, (2) *uniqueness*: different things cannot be denoted by the same GUSI, and (3) *singularity*: a thing cannot be denoted by more than one GUSI. [32].

1.1 Unique Identifiers

While Kent himself saw this paradigm as unfeasible for both theoretical and pragmatic reasons, a number of approaches which come close to his ideal have emerged in recent years.

Microsoft introduced in the early nineties the *Globally Unique Identifier* paradigm (GUID) which implements UUIDs (*Universally Unique IDs*) as defined by the Open Software Foundation in its Distributed Computing Environment specification [56]. UUIDs have recently been standardized through ISO/IEC 9834-8:2004, which specifies format and generation rules that enable users to produce every 100 nanoseconds 128-bit identifiers which are either guaranteed to be, or have a high probability of being, globally unique [30].

In 1998, the International DOI Foundation was created to support the development and promotion of the DOI system [40] resting on the notion of a *Digital Object Identifier* (DOI). A DOI is a single, unambiguous and persistent string that references a single entity and that is generated on the basis of a consistent syntactic frame (a '*numbering scheme*' as defined in the NISO standard ANSI/NISO Z39.84) in a form suitable for use in an automated system [51]. The DOI system is a specific implementation of the Uniform Resource Identifier paradigm advanced by W3C [6] supplemented by management policies for use in the domain of Digital Rights Management.

An enormous boost to the use of unique identification has been given by the wide adoption of Radio Frequency Identification (RFID) technology, which initiated in its turn what may be the next wave of hype in information technology: *The Internet of Things* [41].

1.2 Entity Descriptions

Introducing global unique identification is indeed a first and much needed step towards bringing some clarity to our understanding of what the descriptions in knowledge management systems and in what is called the '*Semantic Web*' are actually *about*. There are several reasons for the current lack of clarity. One is the overemphasis on syntactic regimentation and the false claims, for instance made in the early days of XML but still prevailing today among non-expert professionals, to the effect that such regimentation provides the needed sort of referential semantics as a byproduct [36, 39]. We can, certainly, make legacy electronic documents more easily accessible by manually or semi-automatically annotating documents with tags that reformulate words or relevant phrases in a document in a more structured and standardized manner (e.g. by tagging all occurrences of the words *car*, *van*, *bus*, etc. with the compound *motor vehicle*), or by using meta-tags that add additional context

WWW 2007, May 8--12, 2007, Banff, Canada.

to phrases or paragraphs (e.g. *important*, *motivation*, *ignore*, etc.). Such tags enable retrieval of documents or document sections on the basis of queries issued by users with specific information needs. But they only add more syntax; they do not contribute in any way to providing some formal reference to the *entities in reality* with which they might be associated.

A second reason is the blind, yet unwarranted, trust in the suitability of Description Logics (DL) as a vehicle for making unambiguous descriptions about entities in some domain of discourse [13]. DLs can do no more than guarantee consistent reasoning according to the descriptions and definitions provided to them. But if the latter fall short of correspondence to the reality that they are designed to represent, then even the most powerful DL will do very little to help resolve such problems.

Finally, there is the dominant view that ontologies designed to allow software agents to understand how the entities in a given domain are structured and in what relationships they stand to each other should be organized around '*concepts*' rather than around those entities themselves. This view, rather than solving problems of ambiguity, introduces additional ones [42, 46].

1.3 Towards a Solution

These three false beliefs continue to enjoy wide acceptance as foundational requirements for the Semantic Web approach to the creation of the knowledge management system of the future. Yet we believe that they each contribute to a potentially fateful inability of the Semantic Web to do justice to the way in which our data and information refers to entities in reality – and to the associated phenomena of identification [54]. Promoters of the Semantic Web conceive everything through the spectacles of the Uniform Resource Identifier (URI), with all its associated problems. [9] for instance proposes a solution to these problems that focuses on keeping track of provenance, i.e. of how names and identifiers come to be assigned to entities. In the work described here, we direct our efforts towards the complementary issue of keeping track of the entities themselves on the basis of what we have called Referent Tracking (RT), a paradigm rooted in the solid foundations offered by an approach to ontology based on philosophical realism. We first summarize the theories underlying RT presented in earlier papers [12, 18], and then outline how the approach has allowed us to uncover inadequacies in less rigorous approaches to entity identification in domains such as electronic health record management, digital rights management, corporate memory systems and algorithmic treatment optimization.

2. BASIC FORMAL ONTOLOGY

Basic Formal Ontology (BFO) is a framework that is designed to serve as basis for the creation of high-quality shared ontologies especially in the domain of natural science [24]. It holds (1) that reality and its constituents exist independently of our (linguistic, conceptual, theoretical, cultural) representations thereof; (2) that our theories and classifications can be subject to revision; (3) that there exists a plurality of alternative but equally legitimate perspectives on reality, and (4) that these alternative views are not reducible to any single basic view.

BFO subdivides reality according to a number of basic dichotomies. **First**, it distinguishes *particulars* from *universals*; the former are entities such as Werner Ceusters, the first author of this paper; the latter are entities such as *person*, which have the former as their instances. Both universals and instances are restricted to what exists (or has existed) in reality, and are thus different from classes and instances as referred to in ontologies adhering to a concept-based view [42]. From the BFO perspective, the view advocated in [34] that '*individual instances are the most specific concepts in an ontology*' rests on a confusion. This confusion supports in turn a recommendation according to which '*deciding whether a particular concept is a class in an ontology or an individual instance depends on what the potential applications of the ontology are*'. The implementation of such a recommendation would cripple the ability of ontology to realize its goal of integrating information derived from heterogeneous sources.

Second, BFO distinguishes, within the realm of particulars, between *continuants* and *occurrents*. Continuants are those entities that endure continuously through a period of time while undergoing changes of various sorts. Occurrents *are* such changes: they are entities which unfold in time through their successive temporal parts or phases, otherwise called 'processes,' 'actions', 'events,' 'changes.' The difference between occurrents and continuants is crucial, and any ontology neglecting this distinction is not capable of dealing with changes of entities over time in an adequate way. While, for instance, a continuant particular may instantiate different universals at different times (the first author of this paper was once an instance of *child*, later an instance of *adult*; his societal role was once an instance of *student*, now of *professor*), occurrents cannot undergo such changes because occurrents *are* changes.

Third, there is the distinction between *dependent* and *independent* entities, where each dependent entity is defined as being such that it cannot exist without some independent entity which is its bearer. All occurrents are dependent in this sense on the continuants which participate in them. Thus the process of signing a contract cannot exist without some person who signs. But there are also dependent continuants, for example the contract itself, which cannot exist without contracting organizations or persons. Persons themselves, in contrast, are from the very first moment of their existence independent. Certainly they may require the services of their parents; they will require food, oxygen, and so forth; but they are not dependent on these things in the ontological sense that is relevant to us here.

Fourth, there is the distinction between *fiat* and *bona fide* entities, which is based on the opposition between bona fide (or physical) and fiat boundaries, the latter being exemplified especially by those boundaries – such as the boundary of Utah, or of the 20th century – which are introduced via human demarcation [48]. Fiat boundaries are overwhelmingly present in the realm of social entities, where they delineate for example markets, parcels of real estate, postal districts, and where they serve in establishing what is an employee, what is a taxpayer, what is an able-bodied person, and so forth.

Relations. BFO also distinguishes three major families of relations between the entities just sketched: (1) <p, p>–relations, obtaining between particular and particular (for example: Werner Ceusters being Director of the Ontology Research Group); (2) <p, u>-relations, obtaining between particular and universal (for example: Werner Ceusters being an instance of the universal *person*); and (3) <u, u>-relations, obtaining between universal and universal (for example: *person* being a subkind of *cognitive being*) [45]. The importance of this distinction is exemplified by the fact that relationships such as parthood have distinct properties at the particular and at the universal levels, and that ignoring these distinctions has led to a number of erroneous representations of relations in Description Logic-based approaches to ontology development [21].

3. GRANULAR PARTITION THEORY

Granular Partition Theory is a framework for understanding the ways in which, when cataloguing, classifying, mapping or inventorizing a certain *portion of reality* (POR), human beings and other cognitive agents divide up or partition this reality at one or more levels of granularity [8]. The resultant partitions are composed of *units* (analogous to the cells in a grid), which may be organized into larger sub-partitions in a modular fashion, and the theory provides a formal account of the different ways in which such modules can correspond, or fail to correspond, to the entities in reality towards which they are directed. The theory takes account for example of the degree to which a partition represents the mereological structure of the domain onto which it is projected, and also of the degree of completeness with which a partition represents this domain. Drawing on this framework, we have proposed a calculus for use in quality assurance of complex representations created for clinical or research purposes in the context of both ontology evolution [14] and ontology mapping [11]. The calculus is based on a distinction between three levels [47]: (1) the level of reality, (2) the cognitive representations of this reality, and (3) the publicly accessible concretizations of these representations in artifacts of various sorts, of which ontologies and documents are specific examples. The representations on levels 2 and 3 are partitions in the sense of Granular Partition Theory. Thus they are composed in hierarchical fashion out of modular sub-representations built ultimately out of smallest modules called *representational units*, whereby: (1) each module is assumed to be veridical, i.e. to conform to some relevant POR on the basis of our best current understanding (which may, of course, be based on errors); (2) distinct modules may correspond to the same POR by presenting different though still veridical views or perspectives of this reality, for instance one and the same event may be described both as an *event of buying* and as an *event of selling*; and (3) the modules included in a given representation are determined by the purpose which the representation is intended to serve.

Relevant portions of reality can include not only physical things (buildings, physical goods) but also mental acts and states (feelings of pain, states of desire or fear) and entities of many other types, including social roles and relations.

4. REFERENT TRACKING

Referent tracking (RT) is a new approach to the handling of data about real world entities introduced in [18]. It is designed to allow instances in reality to serve as benchmark for the correctness of the ontologies used to describe them. The RT paradigm has been developed thus far to support the entry and retrieval of data in the Electronic Health Record (EHR), where its purpose is to avoid the problems which arise when statements in an EHR refer to disorders, lesions and other entities on the side of the patient by means of logically complex descriptive phrases such as ‘the fracture in the leg of patient X’ or ‘the tumor in the lung of patient Y’. These problems arise because the phrases in question employ generic terms in ways which may fail to identify the relevant instances unambiguously. (John may have multiple fractures in his leg; or he may have fractured his leg twice at different times in his life.) Referent tracking

Table 1: Abstract syntax and semantics of information templates in a referent tracking system

<table>
<tr><th>Template Name</th><th>Abstract Syntax</th></tr>
<tr><th colspan="2">Description</th></tr>
<tr><td><b>A</b></td><td>$A_i = < IUI_p, IUI_a, t_{ap}>$</td></tr>
<tr><td colspan="2">Captures the assignment of a IUI to a particular where
• IUI_p is the IUI of the particular in question,
• IUI_a is the IUI of the author of the assignment act, and
• t_{ap} is a time-stamp indicating when the assignment was made.</td></tr>
<tr><td><b>PtoP</b></td><td>$R_i = <IUI_a, t_a, r, o, P, t_r>$</td></tr>
<tr><td colspan="2">Description of a relationship between particulars, where
• IUI_a is the IUI of the author of the assertion to the effect that the relationship referred to by r holds between the particulars referred to by the IUIs listed in P,
• t_a is a time-stamp indicating when the assertion was made,
• *r* is the designation in o of the relationship obtaining between the particulars referred to in P,
• *o* is the ID of the ontology from which r is taken,
• *P* is an ordered list of IUIs referring to the particulars between which r obtains, and
• t_r is a time-stamp representing the time at which the relationship was observed to obtain.</td></tr>
<tr><td><b>PtoU</b></td><td>$U_i = <IUI_a, t_a, inst, o, IUI_p, u, t_r>$</td></tr>
<tr><td colspan="2">Description of an instantiation, where
• IUI_a is the IUI of the author of the assertion to the effect that IUI_p <b>inst</b> u,
• t_a is a time-stamp indicating when the assertion was made,
• <b>inst</b> is the designation in o of the relationship of instantiation,
• *o* is the ID of the ontology from which <b>inst</b> and u are taken,
• IUI_p is the IUI referring to the particular whose inst relationship with u is asserted,
• *u* is the designation of the class in o with which IUI_p enjoys the <b>inst</b> relationship, and
• t_r is a time-stamp representing the time at which the relationship was observed to obtain.</td></tr>
<tr><td><b>PtoCo</b></td><td>$Co_i = <IUI_a, t_a, cbs, IUI_p, co, t_r>$</td></tr>
<tr><td colspan="2">Annotating a particular with a code from a concept-based system, where
• IUI_a better to use a single letter instead of ‘*IUI*’ here I think -- also I am now really confused about what your rule is for use of italics and non-italics e.g. in the case of ‘t’) is the IUI of the author asserting that terms associated to co may be used to describe p,
• t_a is a time-stamp indicating when the assertion was made,
• *cbs* is the ID of the concept-based system from which co is taken,
• IUI_p is the IUI referring to the particular which the author associates with co,
• *co* is the concept-code in the concept-system referred to by cbs which the author associates with IUI_p, and
• t_r is a time-stamp representing a time at which the author considers the association appropriate.</td></tr>
<tr><td><b>PtoU$^-$</b></td><td>$U_i = <IUI_a, t_a, r, o, IUI_p, u, t_r>$</td></tr>
<tr><td colspan="2">The particular referred to by IUI_a asserts at time t_a that the relation *r* of ontology *o* does not obtain at time t_r between the particular referred to by IUI_p and any of the instances of the class *u* at time t_r</td></tr>
<tr><td><b>PtoN</b></td><td>$N_i=< IUI_a, t_a, nt_j, n_i, IUI_p, t_r>$</td></tr>
<tr><td colspan="2">The particular referred to by IUI_a asserts at time t_a that n_i is the name of the nametype nt_j assigned to the particular referred to by IUI_p at t_r.</td></tr>
<tr><td><b>Meta-template</b></td><td>$D_i = <IUI_d, X_i, t_d>$</td></tr>
<tr><td colspan="2">Publication of a description of a portion of reality in the RTS where IUI_d is the IUI of the entity registering X_i in the system, X_i is the information-unit in question (in the form of any other template above), and t_d is a reference to the time the registration was carried out.</td></tr>
</table>

avoids such ambiguities by introducing unique identifiers, called IUIs – Instance Unique Identifiers (pronounced you-eye) – for each numerically distinct entity that exists in reality and that is referred to in statements in a record. In the currently still dominant paradigms, the items uniquely identified for EHR purposes are restricted to entities such as patients, care providers, buildings, machines and so forth. The referent tracking paradigm expands this list to include also fractures, polyps, seizures and a vast variety of other clinically salient real-world instances in all the categories distinguished by BFO.

[18] sets forth the conditions for assigning a IUI to a particular, and describes the templates according to which some portions of reality are to be represented in an RT implementation. An additional template for dealing with what in healthcare is known as "negative clinical findings", is introduced in [12]. Note that RT is free from the erroneous *assumption of inherent classification* adhered to in many database design circles according to which entities can be referred to only as instances of pre-specified classes [35]. Thus it is possible to relate particulars to other particulars, and thus do useful inferencing, even where we do not specify of what universals these particulars are instances.

Finally, we have proposed an outline template for registering names by which a particular is referred to in reality (e.g. "John" as first name for the particular John). This template will be expanded along the lines described in [9] in such a way as to allow temporal aspects to be taken into account. The current set of templates is shown in Table 1. The templates are to be interpreted as constituting an abstract syntax; it is left to the developers of an RTS to implement the specifications in the most optimal way given the constraints of the environment in which the system has to operate.

4.1 RT Implementations

A system that implements the RT paradigm (called an RTS) should offer at least three services: (1) generation of unique identifiers to be used as IUIs, (2) management of the IUIs generated, and (3) provision of access to the IUIs stored.

As to (1), the schemes for generating unique strings described in section 1.1 can be used unproblematically. If RTS services would be offered by an entity external to a specific organization, then it may be beneficial for this entity not only to register IUIs but also to certify the uniqueness of the strings to be used within a given IUI-repository and to guarantee that the assignments claimed to have been made by given authors were indeed made by those authors. This can be compared to the services offered by trusted third parties in private key management for asymmetrical encryption purposes [4].

Service (2) involves what we refer to as the *IUI-repository*, whose purpose is to keep track of the identifiers assigned to already existing entities, or reserved for entities that are expected to come into existence in the future. It will do this in such a way that (i) each IUI represents exactly one particular, and (ii) no particular is referred to by more than one IUI. These two requirements are not always easy to fulfill, since both depend on the ability and willingness of users to provide accurate information. This, however, introduces no problems different in principle from those already faced by the users of existing systems when called upon to provide information of a non-trivial and occasionally sensitive sort about individuals.

Service (3), here called the referent tracking database (RTDB), should provide access to the information entered into a given knowledge management system about the particulars referred to in the IUI repository. Where the latter is an inventory of concrete entities that have been acknowledged to exist, and, consequently, of what IDs to use if one wants to refer to them, the RTDB is an inventory of descriptions of features of and interrelations between these entities and of the ways in which they change in the course of time. The RTDB, too, does not need to be set up as a single central database but can rely on any paradigm for distributed storage.

A prototype implementation of an RTS is available through SourceForge under an Open Source license. It is designed in such a way that it can be used as a server application as well as a Java library. As a server, the system runs as a standalone application inside an apache tomcat HTTP web server at port 8080 [50] and it can communicate simultaneously with multiple EHR clients running at remote locations. The server is intended to be hosted by a health institute which serves as the hub for other health institutes (clients). The hosting health institute is responsible for taking care of the administration and privacy issues of the shared information stored at the server. The prototype itself is implemented to serve as a centralized registry system, but the addressing scheme of the identifiers can accommodate distributed implementations.

5. CASE STUDIES

5.1 Electronic Health Records

In [18] we sketched how the referent tracking paradigm might be implemented in the healthcare environment, particularly in relation to clinical record-keeping. The key idea is to do full justice to the *what it is on the side of the patient* that is documented in an EHR, an issue that is severely neglected in prevailing approaches to clinical record keeping, where the (billable) *actions of health practitioners* take center-stage. The need for unique identification of clinically salient entities in a patient's documentation was however recognized already very early on in the history of medical informatics. The central idea of Weed's *Problem Oriented Medical Record* (POMR) is to organize all medical data around a problem list, thereby assigning each individual problem a unique ID [55]. Unfortunately Weed proposes to apply the IUI methodology *only* to problems, and thus not to the various particulars that cause or are symptomatic for them, or are involved in their diagnosis or therapy. The same holds of the problem-based approach of Barrows and Johnson, which suffers further from an ambiguity in its treatment of unique IDs, which sometimes seem to refer to problems themselves and sometimes to statements about such problems [3]. The argument often used in favor of a POMR is that it makes it possible to track a problem such as *chest pain* over time as it evolves into a problem of *angina*, from there into a problem of *myocardial infarction*, of *CABG* (Coronary Artery Bypass Graft), and so forth. However, we consider it wrong to use the labels '*chest pain*', '*angina'*, '*myocardial infarction*', and so on to denote some one enduring thing defined by POMR as 'the problem'. Rather, these labels refer to very different kinds of particular entities that appear and disappear in the unfolding of the *history* of the problem, all of them related in various ways to another particular by which the problem is caused, namely the underlying disorder. Hence we argue that an adequate POMR should embrace also unique identifiers for particulars of all of these latter types.

Another example of an EHR regime involving the use of unique identifiers is that proposed by Huff *et al*. [26], who, refreshingly, take "*the real world to consist of objects (or entities)*". They continue by asserting: "*Objects interact with other objects and can be associated with other objects by relationships … When two or more objects interact in the real world, an* '*event*' *is said to have occurred*." Each event, on the Huff approach, receives an explicit identifier called an *event instance ID*, which is used to link it to other events (reflecting the goal of supporting temporal reasoning with patient data). This ID serves as an anchor for describing the event via a frame-representation, where the slots in the frame are name-value tuples such as *event-ID* = "#223", *event-family* = "diagnostic procedures", *procedure-type* = "chest X-ray", etc. Via other unique IDs the framework incorporates also explicit reference to the patient, the physician and even to the radiographic film used in an X-ray image analysis event. Unfortunately, because they concentrate too narrowly on the events themselves [19], Huff and his associates do not allow explicit reference to those entities in reality *which are observed* during events. This is in spite of the fact that the very X-ray report that they analyze contains the sentence: "*Surgical clips are again seen along the right mediastinum and right hilar region*." [26]

Because they have no means to refer directly to those clips, Huff *et al*. must resort to a complex representation with nested and linked event frames in order to simulate such reference, in ways which once again create opaque contexts which severely reduce the degree to which the resultant information can be used to support reasoning, e.g. for purposes of clinical decision support, tracking of surgical items, and the like.

5.2 Digital Rights Management

Digital Rights Management covers the description, identification, trading, protection, monitoring and tracking of all forms of rights over both tangible and intangible assets, including management of relationships between rights holders in a digital environment. The Digital Object Identifier (DOI) system provides a framework for the persistent identification of artistic and other types of content in its broadest interpretation. Although the system has been very well designed to manage object identifiers, some important questions related to the assignment of identifiers are left open.

In [16] we demonstrated the usefulness of the RT paradigm by showing how it was able to bring to light inconsistencies in the DOI models and how such inconsistencies would be avoided through use of an RTS. The main problem with the DOI approach turned out to be its dependence on the *<Indecs> Framework* [40], which is itself based on the first version of ISO 11179 [29]. The latter restricts an identifier to '*a language independent unique identifier of* ***a data element*** *within a registration authority*'. Each data element is itself such that it relates to an '*object'*, which is in turn defined, in the usual ISO parlance, as '*any part of the* ***conceivable*** *or* ***perceivable*** *world*', including not only existing things but also, for example, unicorns.

For <Indecs>, in consequence, identifiers relate not to entities in reality (such as Werner Ceusters) but rather to pieces of data (such as Werner Ceusters' name). And because, according to ISO, an object need not exist in order to have data 'associated' with it, the result, when <Indecs> is used as the basis for a system of *object identifiers*, is an abundance of confusions (analyzed in our [16]). Some examples:

- '*The <Indecs> model elaborates a logical and semantic framework for describing entities, their attributes and, where appropriate, values of each. Entities, attributes and values are referred to as types of metadata elements'*
- '*a thing must be both thought about or perceived and identified before it exists in a metadata framework'*
- '*all metadata relationships are either events in themselves, or rely on events to establish them'*
- *'nothing exists in any useful sense until it is identified'*.

The orientation of the underlying <Indecs> Framework towards particular, identity-bearing entities in the real world, rather than to generic or conceptual entities, exhibits a clear understanding of what is at stake in facing the challenge of object reference and identification. Unfortunately however the framework itself provides no clear ontological underpinning to support this understanding. We therefore argue that, by subjecting <Indecs> to a deep ontological analysis based on philosophical realism, and by adjusting its data dictionary accordingly, we can make the system fit better the requirements of the Semantic Web.

5.3 Corporate Memories in Enterprises

Another area where appropriate identification is of utmost importance is in corporate memory (CM) systems designed to keep track of the history and evolution of an enterprise with the goal of using lessons learned from past experiences to enhance performance in the future. Well designed CMs should contain data about both the enterprise and the environment in which it operates [33, 37, 52]. Enterprise Ontologies can play an important role in this context as a means of organizing and standardizing the meta-tags used for annotating documents in such a way as to create more powerful CM applications that would work over corporate networks linking together multiple heterogeneous systems [20]. But also in this area, our analysis revealed the existence of much unclarity concerning object reference and identification [15].

The ACORD insurance industry Data Dictionary, for example, which is used to assist in automating business interactions between insurers and clients [1], defines a building as '*a construction that normally has a roof and walls*'. 'Air conditioning', however, it defines as '*information necessary to describe a given type of air conditioning in a building.*' Consistency in providing definitions would dictate that either all entries involve ***information about*** something in reality, or that they denote ***that something in reality*** itself. ACORD, however, provides a problematic mishmash, in which *buildings* would contain *information about air conditioning* as parts.

The same confusion is found in [23]. The latter correctly argues that the Enterprise [49] and TOVE [22] ontologies do not emphasize the distinction between things and their changes on the one hand and conceptual entities on the other, drawing hereby on the work of Bunge [10] and specifically on its application in the Bunge-Wand-Weber model in the domain of information systems for accounting [53]. This analysis led them to develop the PSIM (***P***articipative ***S***imulation environment for ***I***ntegral ***M***anufacturing renewal) Ontology, which was inspired also by earlier work conducted in the European Research Project CIMOSA [2] and by Peircean Semiotics [25]. The result, however, is not without its own dramatic mysteries and misinterpretations. Thus we read that the PSIM Ontology distinguishes the three main categories of: Activity, Object and Information (element), whereby an 'Information (element)' is defined as: '*a characteristic of either an object or activity or information, which is used to constrain directly or indirectly the involvement of an object in an activity*' [23]. PSIM classifies as information elements not only '*the time needed to perform an activity*' and '*how an activity has to be performed*', but also '*how the enterprise is organised*', '*the way the responsibilities are distributed among the enterprise*', and even '*the weight of a piece of material*'. Weight, for RT, however, is a dependent continuant that depends on the material object of which it is the weight, and this independently of whether or not a cognitive being has any sort of information about the matter.

5.4 Psychiatric Treatment Optimization

The International Psychopharmacology Algorithm Project (IPAP) is an international initiative set up in 1985 by a team of psychiatrists, psychopharmacologists and algorithm designers in an effort to improve choice of medication in psychiatry [27]. In 1995, the *IPAP Schizophrenia Algorithm* (IPAP-SA) was published by IPAP as a guideline consisting of four schizophrenia treatment algorithms developed, respectively, for the first schizophrenic episode, long-term medication maintenance, schizophrenia complicated by comorbid psychiatric disorders, and schizophrenia complicated by neuroleptic malignant syndrome [5].

In 2006, we analyzed the January 2005 version of this IPAP guideline which was made available on the web. (This has since been replaced by a newer version (v. 20060327 [28]), which however does not differ in substantial ways for the purposes of this discussion.) The algorithm is presented in the form of a flow chart with an established diagnosis of schizophrenia or schizoaffective disorder as its single entry condition and two exit conditions, one suggesting a modification to the patient's current treatment program, the other suggesting unaltered continuation of this program. The on-line version provides some obvious advantages over a traditional journal or textbook publication. It can be accessed immediately through any suitable browser, and new versions become accessible as soon as they are released. Given that the algorithm is currently implemented as a simple flow-chart, however, in which the included hyperlinks serve only human browsing, it still fails to exploit the real power of the computer, which is to perform reasoning automatically. We accordingly investigated the possibility of developing an implementation which could draw on information already available in the patient's electronic health record (EHR) in such a way as to process relevant features of the patient's current condition in light of those criteria which play a role in the corresponding step of the algorithm.

In [17], we reported on our research to carry out the first step of enhancing the present version of the IPAP algorithm along these lines in such a way that it can be used in automatic decision support. To this end it was necessary to identify the minimal set of universals and particulars which must be represented in a referent tracking system in order to allow software agents to carry out real-time monitoring and control activities to optimize the treatment of schizophrenic patients in accordance with IPAP guidelines. The analysis was performed with the goal of demonstrating how the RT approach could be used for upgrading static and inert flow-chart algorithms like IPAP in such a way that they would constitute dynamic application ontologies. It revealed, again, how important it is not just to uniquely identify patients, but also their individual diseases and associated phenomena.

For the execution of the IPAP schizophrenia algorithm, it is mandatory that the patient's disease be an instance of one or other of the universals *schizophrenia* or *schizoaffective disorder*. This entry condition is phrased in the algorithm itself as: '*meet DSM-IV and ICD10 criteria for schizophrenia and schizoaffective disorder*', referring respectively to the Diagnostic and Statistical Manual of Mental Disorders published by the American Psychiatric Association and to the International Classification of Diseases published by WHO. This, unfortunately, poses certain problems. The first is logical in nature: does the patient's established diagnosis need to satisfy the diagnostic criteria of *both* DSM-IV and ICD-10, or is it sufficient that either one or the other be satisfied? This question is important, since there is only a partial concordance between the two, concrete figures for this concordance ranging from 60% to 83% depending on the subtype of schizophrenia [7]. Thus it is possible that a patient's disease has to be classified as schizophrenia according to one system, but that it is not allowed to be so classified by the other.

The second question is ontological in nature: to what extent do the terms ("schizophrenia" or "schizoaffective disorder") used by ICD-10 and DSM-IV represent one, or two, or no universals at all on the side of biomedical reality? Here, too, the referent tracking idea brings certain advantages. We first make what seems to us to be a reasonable assumption to the effect that, if a given body of patient records systematically includes diagnoses of schizophrenia and/or of schizoaffective disorder, then there is *something* to which these terms refer on the side of the corresponding patients. Each such something can be given an IUI – even should it turn out that the something in question is, for example, some different disease. Let us suppose, for example, that we assign #I-9001 to the putative case of schizophrenia diagnosed in John, and that we include this IUI in a referent tracking database that is used while carrying out a variety of different types of diagnostic tests. By analyzing the results of such tests, we may in the long run be led to the conclusion that #I-9001 is in fact a compound of two or more disease particulars (or, in the worst case, that it is an empty ID designating no disease at all) [43]. In this way experience might indeed prove in the course of time that "schizophrenia" itself is a term that has no referent, for example because what had been thought to be a single disease is in fact a compound of several diseases hitherto not cleanly separated – in ways which might then lead to modifications to the IPAP algorithm itself.

6. CONCLUSION

Our case studies indicate that the currently predominant enabling technologies for building knowledge management systems are still too narrowly oriented around the paradigm of *information* modeling, which is a matter of the tracking (or modeling, or representation) of *information*. A referent tracking system, in contrast, tracks entities in reality. The latter can indeed include also pieces of information about entities (for example in the form of images), which are acknowledged as entities in their own right, but it should do this in such a way that first-level entities are never confused with those entities which carry information about them – a confusion of a type which, as we have seen, is endemic on current paradigms. Consider, to take just one illustrative example, the influential paper [38] of Rector *et al.*, which contains assertions such as: '*Every occurrence level statement concerning Jane Smith's Fracture of the Femur is an observation of the corresponding individual*'; whereby: '*The existence* [sic] *of the individual Jane Smith's Fracture of Femur does not imply that Jane Smith has, or has ever had, a fracture of the femur* [sic], *but merely that some observation has been made about Jane Smith regarding a fracture of the femur.*' Such confusions are manifested in a quite peculiarly egregious form in the case described in [44].

This is not to deny that much valuable work has been invested in information model- and concept system-based tools for knowledge management systems. But we believe that the referent tracking paradigm – and the concomitant clear understanding of the distinction between an entity and the data about an entity which it brings in its wake – must be called in aid to support any application of such tools in mission critical domains such as healthcare (or indeed in any domain where quality of work is considered to be of importance). Referent tracking gives us the means to allow reality itself to serve as benchmark for the correctness of such application, where, on current paradigms, we have only 'concepts' and 'models'.

7. ACKNOWLEDGMENTS

This work has been funded in part by grant 1 U 54 HG004028 from the National Institutes of Health through the NIH Roadmap for Medical Research.

REFERENCES

[1] ACORD Data Dictionary for Insurance Industry, 2005. http://www.acord.org/dataDictionary/dataDictionary.htm

[2] AMICE-Consortium *Open System Architecture for CIM, Research Reports of ESPRIT Project 688*. Springer Verlag, Berlin 1989.

[3] Barrows, R.C. and Johnson, S.B. A data model that captures clinical reasoning about patient problems. Gardner, R.M. ed. *19th Annual Symp Computer Applications in Medical Care*, Hanley & Belfus, Inc., New Orleans, 1995, 402-405.

[4] Bellare, M. and Rogaway, P. The exact security of digital signatures – How to sign with RSA and Rabin. in *Lecture Notes in Computer Science*, Springer, 1996, 399-416.

[5] Bender, K.J. Algorithm Project Provides Guides to Current Knowledge *Psychiatric Times*, 1996.

[6] Berners-Lee, T., Fielding, R. and Masinter, L. Uniform Resource Identifier (URI): Generic Syntax, The Internet Society, 2005.

[7] Bertelsen, A. Schizophrenia and related disorders: experience with current diagnostic systems. *Psychopathology*, *35* (2-3). 89-93.

[8] Bittner, T. and Smith, B. A Theory of Granular Partitions. in Duckham, M., Goodchild, M.F. and Worboy, M.F. eds. *Foundations of Geographic Information Science*, Taylor & Francis Books, London, 2003, 117-151.

[9] Bouquet, P., Stoermer, H., Mancioppi, M. and Giacomuzzi, D. OKKAM: Towards a Solution to the "Identity Crisis" on the Semantic Web. Tummarello, G., Bouquet, P. and Signore, O. eds. *Semantic Web Applications and Perspectives (SWAP 2006)*, Pisa, Italy, 2006.

[10] Bunge, M. *Treatise on Basic Philosophy, Ontology I: The Furniture of the World*. Reidel, Boston, 1977.

[11] Ceusters, W. Towards A Realism-Based Metric for Quality Assurance in Ontology Matching. in Bennett, B. and Fellbaum, C. eds. *Formal Ontology in Information Systems*, IOS Press, Amsterdam, 2006, 321-332.

[12] Ceusters, W., Elkin, P. and Smith, B. Referent Tracking: The Problem of Negative Findings. in Hasman, A., Haux, R., Lei, J.v.d., Clercq, E.D. and Roger-France, F. eds. *Studies in Health Technology and Informatics. Ubiquity: Technologies for Better Health in Aging Societies - Proceedings of MIE2006*, IOS Press, Amsterdam, 2006, 741-746.

[13] Ceusters, W. and Smith, B. Ontology and Medical Terminology: why Descriptions Logics are not enough. *Towards an Electronic Patient Record (TEPR 2003)*, San Antonio, 2003.

[14] Ceusters, W. and Smith, B. A Realism-Based Approach to the Evolution of Biomedical Ontologies. in *Proceedings of AMIA 2006*, 2006, 121-125.

[15] Ceusters, W. and Smith, B. Referent Tracking for Corporate Memories. in Rittgen, P. ed. *Handbook of Ontologies for Business Interaction*, Idea Group Publishing, 2007 (forthcoming).

[16] Ceusters, W. and Smith, B. Referent Tracking for Digital Rights Management. *Forthcoming in International Journal of Metadata, Semantics and Ontologies*.

[17] Ceusters, W. and Smith, B. Referent Tracking for Treatment Optimisation in Schizophrenic Patients. *Journal of Web Semantics - Special issue on semantic web for the life sciences*, *4* (3). 229-236.

[18] Ceusters, W. and Smith, B. Strategies for Referent Tracking in Electronic Health Records. *Journal of Biomedical Informatics*, *39* (3). 362-378.

[19] Coyle, J.F., Rossi-Mori, A. and Huff, S.M. Standards for detailed clinical models as the basis for medical data exchange and decision support. *International Journal of Medical Informatics*, *69* (2-3). 157-174.

[20] Davies, J., Fensel, D. and Harmelen, F.v. (eds.). *Towards the Semantic Web - Ontology-driven Knowledge Management*. John Wiley & Sons, 2002.

[21] Donnelly, M., Bittner, T. and Rosse, C. A formal theory for spatial representation and reasoning in biomedical ontologies. *Artificial Intelligence in Medicine*, *36* (1). 1-27.

[22] Fox, M.S. The TOVE Project: Towards A Common-sense Model of the Enterprise, Enterprise Integration Lab, 1992.

[23] Goossenaerts, J. and Pelletier, C. Ontology and Enterprise Modeling, 2003. http://is.tm.tue.nl/staff/jgoossenaerts/4PublicPdf/PSIM%20book%20ch%205%20Ontol&EM.pdf

[24] Grenon, P., Smith, B. and Goldberg, L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. in Pisanelli, D.M. ed. *Ontologies in Medicine*, IOS Press, Amsterdam, 2004, 20-38.

[25] Hoopes, J. *Peirce ON SIGNS. Writings on Semiotic by Charles Sanders Peirce*. The University of North Carolina Press Chapel Hill and London, 1991.

[26] Huff, S.M., Rocha, R.A., Bray, B.E., Warner, H.R. and Haug, P.J. An event model of medical information representation. *Journal of the American Medical Informatics Association*, *2*. 116-134.

[27] IPAP. About The International Psychopharmacology Project, 2006. http://www.ipap.org/about.php

[28] International Psychopharmacology Algorithm Project. IPAP-Schizophrenia Algorithm Interactive Flowchart 2006. http://www.ipap.org/schiz/schizalg.php?screen=flowchart

[29] International Standards Organisation ISO/IEC 11179-1:1999(E) Information technology -- Specification and standardization of data elements -- Part 1: Framework for the standardization of data elements.

[30] International Standards Organisation ISO/IEC FDIS 9834-8:2004. Information technology – Open Systems Interconnection – Procedures for the operation of OSI Registration Authorities: Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 Object Identifier components.

[31] Kent, W., The Entity Join. in *Fifth International Conference on Very Large Data Bases*, (Rio de Janeiro, Brazil, 1979), Morgan Kaufmann Publishers, 232-238.

[32] Kent, W. The unsolvable identity problem *Extreme Markup Languages 2003*, Montreal, Canada, 2003.

[33] Kühn, O. and Abecker, A. Corporate memories for Knowledge Management in Industrial Practice: Prospects and Challenges. *Journal of Universal Computer Science*, *3* (8). 929-954.

[34] Noy, N.F. and McGuinness, D.L. Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory, 2001.

[35] Parsons, J. and Wand, Y. Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Transactions on Database Systems*, *25* (2). 228-268.

[36] Paskin, N. Digital Object Identifiers for Scientific Data. *Data Science Journal*, *4*. 12-20.

[37] Prasad, M.V.N. and Plaza, E. Corporate Memories as Distributed Case Librairies. in *Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Canada, 1996, 40-41 40-19.

[38] Rector, A.L., Nowlan, W.A., Kay, S., Goble, C.A. and Howkins, T.J. A framework for modelling the electronic medical record. *Methods of Information in Medicine*, *32* (2). 109-119.

[39] Renear, A., Dubin, D., Sperberg-McQueen, C.M. and Huitfeldt, C. XML semantics and digital libraries in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries table of contents*, IEEE Computer Society, 2003, 303 - 305.

[40] Rust, G. and Bide, M. The <Indecs> metadata framework: principles, model and data dictionary. WP1a-006-2.0, 2000.

[41] Shannon, V. Wireless: Creating Internet of 'Things': A scary, but exciting idea International Herald Tribune, Sunday, November 20, 2005.

[42] Smith, B. Beyond concepts: ontology as reality representation. in *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*, IOS Press, Amsterdam, 2004, 73-84.

[43] Smith, B. From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. *Journal of Biomedical Informatics*, *39* (3). 288-298.

[44] Smith, B. and Ceusters, W. HL7 RIM: An Incoherent Standard. in Hasman, A., Haux, R., Lei, J.v.d., Clercq, E.D. and Roger-France, F. eds. *Studies in Health Technology and Informatics. Ubiquity: Technologies for Better Health in Aging Societies - Proceedings of MIE2006*, IOS Press, Amsterdam, 2006, 133-138.

[45] Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L. and Rosse, C. Relations in biomedical ontologies. *Genome Biology*, *6* (5). R46.

[46] Smith, B., Ceusters, W. and Temmerman, R. Wüsteria. in Engelbrecht, R., Geissbuhler, A., Lovis, C. and Mihalas, G. eds. *Connecting Medical Informatics and Bio-Informatics. Medical Informatics Europe 2005*, IOS Press, Amsterdam, 2005, 647-652.

[47] Smith, B., Kusnierczyk, W., Schober, D. and Ceusters, W. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain *KR-MED 2006, Biomedical Ontology in Action.*, Baltimore MD, USA 2006.

[48] Smith, B. and Varzi, A.C. Fiat and Bona Fide Boundaries: Towards on Ontology of Spatially Extended Objects in *Lecture Notes In Computer Science*, Springer Verlag, London, UK, 1997, 103 - 119.

[49] Stader, J. Results of the Enterprise Project *16th Annual Conference of the British Computer Society Specialist Group on Expert Systems* Cambridge, UK, 1996.

[50] The Apache Software Foundation. Apache Tomcat Server, 2006.

[51] The International DOI Foundation. The DOI Handbook (Version 4.4.1, released 5 October 2006). 2006.

[52] Van Heijst, G., Van der Spek, R. and Kruizinga, E. Organizing Corporate Memories *Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Canada, 1996, 42-41 42-17.

[53] Wand, Y., Storey, V. and Weber, R. An Ontological Analysis of the relationship Construct in Conceptual Modeling. *ACM Transactions on Database Systems*, *24* (4). 494-528.

[54] Warren, P. Knowledge management and the semantic web : From scenario to technology. *IEEE intelligent systems*, *21* (1). 53-59.

[55] Weed, L. Medical records that guide and teach. *New England Journal of Medicine*, *278*. 593-600.

[56] Williams, S. and Kindel, C. The component object model: A technical overview, 1994.