Dealing with Mistakes in a Referent Tracking System

Werner CEUSTERS
Center of Excellence in Bioinformatics & Life Sciences, Buffalo, NY

1 Introduction

Referent Tracking (RT) is a paradigm introduced in 2005 intended to provide a means of ensuring unambiguous reference to the particulars in reality that are mentioned in statements of given sorts [1]. Central to this paradigm, which was conceived originally in the context of work on electronic health records, is the use of globally unique singular identifiers – called IUIs (for Unique Instance Identifiers) – that stand proxy for the entities in reality to which they refer. For an identifier (ID) to be a IUI, it must refer to one and only one particular, and this tight connection between the particular and its IUI must be asserted by an author in an RT system (RTS) [2]. One purpose of the RTS is to give agents who wish to make statements about entities in reality a means to retrieve IUIs for particulars to which identifiers have already been assigned, and to create IUIs in other cases. Another purpose is to provide an efficient way to store data about particulars in terms of their relations to other particulars and to the universals which they instantiate [3]. The RT paradigm is associated with developments such as the LSID initiative and the so-called 'web of things', but offers a number of advantages brought about by the strict ontological principles under which unique identification is achieved.

2 Coping with change

The real world is subject to constant change, and so also is our knowledge thereof. To keep track of these two sets of changes, an RTS requires that any assertion concerning a relationship between entities is associated with an index for the time period during which the relationship obtains, for the time at which the assertion is made, and for the author of the assertion. In [4], we proposed a methodology for realism-based ontology versioning and evolution that makes explicit whether changes in a new version are due to changes in (1) reality (ΔE), (2) the ontology authors' understanding thereof (ΔB), (3) relevance for inclusion of representations (ΔRv), or (4) corrections of mistakes (R- or R$\neg$). The same methodology can be applied to the RTS because the main difference with a realism-based ontology is that the former is used to store information about particulars, and the latter information about universals. But because of this difference, the RTS will undergo many more changes and this, when used on a large scale, on a constant basis.

3 Types of mistakes

3.1 Mistakes in RTS entries

An entry in the RTS is erroneous if it either violates the principles of referent tracking or fails to mirror the reality to which the RTS is intended to refer.

In either case the entry contains an ID which is believed to be a IUI, but in reality is not. This applies either when the entry does not refer to anything existing (now or in the past), or when it contains a non-singular or non-unique reference. In the case of non-singular reference the same particular is represented in the RTS as two or more numerically distinct entities. In the case of non- unique reference at least two numerically distinct entities are represented as being only one.

Entries in the RTS come in various flavors. One type of entry, called A-tuples (for 'assignments'), are used to assert the existence of some particular in reality at some time, and to differentiate this particular from other particulars by assigning it a unique singular ID. Hence, A-tuples can be qualified as being in error for one or other of the following reasons:

- A1: the ID does not refer
- A2: the ID refers to two (or more) distinct particulars
- A3: the ID is not the only ID in the RTS that refers to this particular
- A4: the ID does not refer to the intended particular.

A second type of entry, called PtoU-tuples, relates a particular to a universal the reference to which is drawn from some external ontology. Tuples of this type can also be in error for several reasons, including:

- U1: the relationship between the particular referred to by the IUI and the universal in question does not hold during the stated time period,
- U2: the ID for the universal does not refer to the intended universal or it refers to no universal at all.

Where the ID for the particular is subject to an A-type error, the following additional PtoU-errors may occur:

- U3: there is an A1 error in the corresponding A-tuple: the PtoU-tuple is nonsensical
- U4: the ID is subject to a mistake of type A2 and for at least one of the particulars referred to by it, the stated relationship does not hold,
- U5: the ID is subject of a mistake of type A3, and the particular referred to by the ID is not an instance of the universal during the stated time period,
- U6: similar to U5, but involving a type A4 mistake.

Four further types of mistakes are such that reality is mirrored by the PtoU-tuple in question, but what is mirrored is either not what was intended, or is irrelevant:

- U7: the ID is subject of a mistake of type A2, but for ***all*** particulars referred to by it, the stated relationship holds,
- U8: the ID is subject of a mistake of type A3, but the particular referred to by the ID is an instance of the universal during the stated time period,
- U9: similar to U5, but involving a type A4 mistake,
- U10: there is no A-type of mistake but the stated relationship is irrelevant.

Finally, entries expressed through PtoP-tuples relate particulars to each other and may involve many further sorts of mistakes (P1, P2, …), depending on whether one or both IDs involve an A-type of mistake, and whether the relationship in question holds.

3.2 Mistakes of omission

In [4] it is argued that an ontology should contain representational units for all universals that are relevant for the purpose for which the ontology is built. The same principle holds in the context of an RTS: all portions of reality that are relevant for the purpose for which the RTS is maintained should be represented by means of corresponding tuples. If not, the following mistakes occur, in both cases leading to the absence of an A-tuple:

- A–1: the existence of a relevant particular is not acknowledged,
- A–2: the relevance of a particular for the purpose of the RTS is not acknowledged.

Similarly, we recognize two further types of errors involving universals or particulars:

- U–1 / P–1: the existence of a relevant relationship between a particular and some other entity is not acknowledged,
- U–2 / P–2: the relevance of a relevant relationship between a particular and some other entity is not acknowledged.

Whereas mistakes of omission may occur independently of other mistakes, some mistakes of type A and type U will automatically bring in their wake mistakes of other types: Thus for

example, mistakes of type U6 and U9 will be automatically associated with a mistake of type U–1.

4 Dealing with mistakes

To mirror at any given point in time what is believed by the authors of given assertions to be the case in reality at that time, and what was believed to be the case at any earlier point in time, entries in the RTS are never deleted. Rather, the corresponding entities acquire annotations to the effect that they did not mirror reality during the period when they were believed to do so, and possibly also linked to new tuples that function as corrections.

Because of the way an RTS is implemented, no additional templates are required: it suffices to modify the structure of the D-templates as defined in [2] from

$$D_i = <IUI_d, T_i, t_d>$$

where IUI_d is the IUI of the entity registering the tuple T_i – T symbolizing here any tuple – in the system and t_d a reference to the time the registration is carried out, to

$$D_i = <IUI_d, IUI_{Ti}, t, E, C, S>.$$

This change involves RTS entries becoming assigned IUIs of their own which in the restructured D-template is symbolized by IUI_{Ti}. The other components of the D-template are:

- IUI_d: the IUI of the entity annotating IUI_{Ti} by means of the D_i entry,
- E: either the symbol 'I' (for insertion) or any of the error type symbols as categorized in section 3,
- C: a symbol for the applicable reason for change as discussed in section 2,
- t: the time the tuple denoted by IUI_{Ti} is inserted or 'retired', and
- S: a list of IUIs denoting the tuples, if any, that replace the retired one.

We use the Bloodsworth case [5] to demonstrate the principles. In July 1984, 9-year-old Dawn Hamilton was raped and murdered. In August 1984, Kimberly Ruffner was imprisoned for another rape and attempted murder. A composite sketch of the perpetrator in the Hamilton case was shown on the local news. Two anonymous callers advised that Kirk Bloodsworth looked like the composite. Bloodsworth was convicted of the murder. In 1993, new forensic tests discovered semen on Hamilton's underpants. DNA tests proved it was not Bloodsworth's. In September 2003, the DNA sample recovered from Hamilton's underpants was identified as that of Ruffner.

For the sake of conciseness, we describe in Table 1 only verbally what a few relevant IUIs denote, rather than working with a complete ontology. Further relevant tuples not listed in Table 1 are the A-tuples representing the assignment of the IUIs to the corresponding first order particulars, and the D-tuples that go along with them.

Table 2 displays chronologically some of the D- and A-tuples – ignoring their authors – that would result from tracking the particulars. It provides a nice insight into how the RTS changes over time, and how the error correction mechanism goes hand in hand with the representation of changes in reality, our understanding thereof, and changes of relevance. The correction introduced here is the insertion of the D-tuple to which IUI-109 is assigned: this tuple retires PtoP-tuple IUI-9 which contained a Px type of mistake.

IUI-1: Dawn Hamilton	IUI-2: Dawn Hamilton's rape
IUI-3: Composite sketch of Hamilton's rapist	IUI-4: The August 1984 rape
IUI-5: Kimberly Ruffner	IUI-6: Kirk Bloodsworth
IUI-7: the PtoP-tuple representing that IUI-5 committed IUI-4	IUI-8: the PtoP-tuple representing that IUI-6 resembles IUI-3
IUI-9: the PtoP-tuple representing that IUI-6 committed IUI-2	IUI-10: Portion of DNA in Hamilton's underpants
IUI-11: Portion of Bloodsworth's DNA	IUI-12: Portion of Ruffner's DNA
IUI-13: the PtoP-tuple representing that IUI-11 is dissimilar to IUI-10	IUI-14: the PtoP-tuple representing that IUI-5 committed IUI-2

Table 1: Some relevant particulars and their associated IUIs in the Bloodsworth case.

Tuple Type	Tuple IUI	Tuple
A	IUI-101	< IUI-1, – , 1975>
D	IUI-102	< –, IUI-101, July 1984, I, ΔRv, {}>
A	IUI-103	<IUI-2, – , July 1984>
D	IUI-104	< –, IUI-103, July 1984, I, ΔE, {}>
A	IUI-105	<IUI-3, – , August 1984>
D	IUI-106	< –, IUI-105, August 1984, I, ΔE, {}>
A	IUI-107	<IUI-6, – , 1961>
D	IUI-108	< –, IUI-107, 1985, I, ΔB, {}>
D	IUI-109	**< –, IUI-9, 1993, Px, ΔB, {}>**
D	IUI-110	< –, IUI-14, September 2003, I, ΔB, {}>

Table 2: some of the D- and A-tuples – ignoring their authors – that would result from tracking the particulars listed in Table 1

The Bloodsworth case could be represented in many other ways, for instance by assigning a IUI to 'the rapist of Dawn Hamilton' before it is known who that is. In that case, an A3 type of mistake would have to be corrected. The choice of representation is not something that is restricted by the RTS, but rather by the ontologies and the theories upon which they are built.

5 References

1 Ceusters W. Smith B. Tracking Referents in Electronic Health Records. In: Engelbrecht R. et al. (eds.) Medical Informatics Europe, IOS Press, Amsterdam, 2005;:71-76.

2 Ceusters W, Smith B. Strategies for Referent Tracking in Electronic Health Records. J Biomed Inform. 2006 Jun;39(3):362-78.

4 Ceusters W, Smith B. A Realism-Based Approach to the Evolution of Biomedical Ontologies. Proceedings of AMIA 2006, Washington DC, 2006;:121-125.

5 Levine S. Ex-Death Row Inmate Hears Hoped-for Words: We Found Killer. Washington Post, September 6, 2003; Page A01.