Towards a Similarity-Based Identity Assumption Service for Historical Places

Krzysztof Janowicz

Institute for Geoinformatics
University of Muenster, Germany
janowicz@uni-muenster.de

Abstract. Acquisition and semantic annotation of data are fundamental tasks within the domain of cultural heritage. With the increasing amount of available data and ad hoc cross linking between their providers and users (e.g. through web services), data integration and knowledge refinement becomes even more important. To integrate information from several sources it has to be guaranteed that objects of discourse (which may be artifacts, events, persons, places or periods) refer to the same real world phenomena within all involved data sources. Local (database) identifiers however only disambiguate internal data, but fail in establishing connections to/between external data, while global identifiers can only partially solve this problem. Software assistants should support users in establishing such connections by delivering identity assumptions, i.e. by estimating whether examined data actually concerns the same real word phenomenon. This paper points out how similarity measures can act as groundwork for such assistants by introducing a similarity-based identity assumption assistant for historical places to support scholars in establishing links between distributed historical knowledge.

1 Introduction & Motivation

This section describes both the motivation for writing this paper as well as an insight into the idea of and need for identity assumption services within the domain of cultural heritage.

1.1 Motivation

Within the last years similarity measurements gained credence as method for information retrieval and integration within the GIScience community. The research however is mostly focused on inter-concept measurements using several (incompatible and proprietary) knowledge representation formats. Therefore only few real world applications such as a similarity enabled pedestrian navigation service developed by Raubal [1] are discussed in the literature until now. Additionally most conceptualizations published on the web use more or less standardized (description) logic based formalization languages which results in a gap between available theories

and available ontologies. Moreover it is claimed that similarity measurements are closer to the human way of thinking than logic based reasoning services such as subsumption reasoning and therefore deliver more adequate results. Nevertheless there is no elaborate study supporting this theory (at least for GIScience information retrieval scenarios).

The main motivation underlying this paper is to give an insight into how similarity measurements can help in solving real world problems and to show (instead of contrasting both approaches) how similarity theories can interact with existing and well established reasoning services. The presented use case is therefore chosen in a way that it benefits from both, rigid logic based reasoning for knowledge extraction and discovery and flexible measurement theories to return ranked similarity assumptions back to the system user. The paper focuses on spatiotemporal relations, but also takes thematic and referential relations into account for similarity measurements.

The theory presented in this paper is focused on instance rather than concept similarity. It is adapted to fit the identity assumption use case and hence the chosen top-level ontology and knowledge representation format. First steps toward a full grown similarity measurement framework for high expressive description logics (aiming to close the mentioned gap), currently developed at the Muenster Semantic Interoperability Lab (MUSIL), are discussed in [2].

1.2 Introduction

The domain of cultural heritage is very heterogeneous in a sense that the themes or exhibits that museums and related institutions are concerned with range from history of science through all kinds of art and historical documents up to biodiversity. Accordingly the number and type of preserved exhibits rang from millions of collected organisms to a small number of valuable paintings.

Creating and maintaining metadata about historical facts and exhibits gets increasingly important for scholars and curators to structure, manage, and query their own data. As long as metadata is used for internal workflows only (such as the preparation of an exhibition), each institution may develop and maintain their own schema and representation format; however to refine and enrich the own knowledge base or to answer complex scientific questions, interchange with external sources is needed. To support these tasks the Committee on Documentation (CIDOC) provides a well established and standardized core ontology (called CIDOC CRM) [3] intended to annotate heterogeneous cultural heritage information to make it available in a machine readable (RDF) and reasonable way for knowledge integration, mediation and interchange. The vision is to make all annotated datasets available through web (or grid) services to enable automatic metadata harvesting [4] and to form a shared network of interlinked historical information. The CIDOC CRM ontology can be regarded as the underlying semantic level that provides meaning within the intended cultural heritage data infrastructure (which can bee seen analogous to an SDI) by delivering a common metadata schema.

To make use of external data sources, however, a common language is not enough. Moreover it has to be guaranteed that the collected metadata refers to the same real

world phenomenon (which could be a historical place, person, event or object) as the local datasets. Global authorities (such as the Alexandria Digital Library Gazetteer Server [5]) provide unique identifiers and annotated datasets for some common kinds of real world phenomena. Scholars can refer to these global identifiers in addition to (or instead of) their local identifiers and therefore reduce maintenance effort and redundancy on one hand and to enable data interchange on the other. If compared datasets refer to the same global identifier and one decides to trust the global authority as well as the external party that linked their dataset to the specific identifier, it can be assumed that the same real world phenomenon is meant.

Nevertheless until now most datasets do not refer to global authorities and scholars have to decide as the case arises if the harvested information is relevant for the own. This is for several reasons: First, our knowledge about historical places, which are of primary interest within this paper, is often vague and incomplete. Moreover the referring place names are ambiguous and may change during history (however Gazetteer services can be used to disambiguate common place names). The same is true for the geopolitical units the historical place belongs to. Imagine the Turkish city Istanbul, which was founded as Byzantium as part of the Greek Empire; conquered by the Persian Empire; renamed as Nova Roma and Constantinople (called Tsargrad by ancient Slavics) as the second capital of the Roman Empire; later acted as capital of the Ottoman Empire and finally lost the capital status and was renamed to Istanbul in the early 20th century as part of the Turkish Republic. While both the Alexandria Gazetteer and the Getty Thesaurus of Geographic Names [6] contain more than ten alternative (historical) names for Istanbul, the names themselves differ (e.g. Nova Roma is missing in the ADL Gazetteer while Stambul is missing in Getty). Moreover all entries, independent from the historical context connected with the certain place name, refer to the same geopolitical hierarchy (i.e. as part of Turkey). The impact of these shortcomings for the Gazetteer feature types used within the presented similarity measure is discussed later on. In addition to these problems many places are only referred to within historical documents by their role in certain historical events (such as the place where Admiral Nelson died after the Battle of Trafalgar [3] or the spot where a new species was found during an expedition). Such places are not necessarily referred to by spatial relations to other entities or even coordinates. Finally, the most significant reason why global identifiers provided by Gazetteers can only partially solve the problem of identity, is that using Gazetteers to determine whether two datasets refer to the same real world place, presumes that all involved institutions have manually annotated millions of local datasets beforehand, which is not the case until now.

Therefore an identity assumption assistant should support scholars in analyzing the harvested metadata and returning promising datasets - promising in a way that the external datasets *probably* refer to the same real world place addressed by the own data. The identity assumption theory used by such an assistant should be non-rigid in a way that it returns a ranked list of estimations instead of trying to automatically conclude safe predictions out of vague historical data. This paper proposes a similarity-based theory that generates such ranked assumptions by comparing CIDOC CRM annotated information (sets of RDF-Triples) about historical places. The proposed theory will be introduced stepwise and elucidated by the scenario “Battle of Trafalgar”, specifying the places, actors and events that are being compared.

2 Related Work

This section provides a brief overview about existing similarity measures focusing on those related to GIScience and the CIDOC conceptual reference model.

2.1 Similarity Measurements

The notion of distance is central to the idea of similarity measurements as it determines how close certain aspects of compared entities or classes are to each other. From this perspective, research about similarity is concerned with finding and combining distance metrics for kinds of aspects. Depending on the chosen knowledge representation approach these aspects can be: features, dimensions, transformations, and language constructors; virtually everything that is used to describe the compared classes or entities. In contrast to subsumption reasoning, similarity returns the degree of overlap and therefore is usually a function from compared classes or entities to numeric values (normalized to values between 0 and 1).

The idea of similarity measurement is widely applied across cognitive and information science. An overview about different theories (from cognitive science), their benefits and shortcomings is discussed in [7]. MDSM, a feature-based approach for lightweight ontologies, well established in GIScience is introduced in [8] and extended by thematic roles in [9]. Similarity theories based on conceptual spaces [10] are presented in [1, 11]. A hybrid model is discussed in [12]. A similarity theory for semantic web services is introduced in [13]. Measurements for similarity between different ontologies are discussed in [14, 15]. First steps towards a similarity theory for Description Logics are discussed in [2, 16].

2.2 The CIDOC Conceptual Reference Model

The CIDOC conceptual reference model [3] is a top-level ontology specifying the most fundamental concepts common to all fields of the cultural heritage community. To be that generic, CIDOC CRM does not provide conceptualizations for concrete kinds (such as kinds of exhibits), but defines a framework providing the base terminology necessary for annotation of and reasoning within historical information. While the classes and relationships, which are of major interest within the identity assumption theory discussed here, are introduced in the theory section below (see section 4), this section first gives a broad overview about some characteristics and design decisions underlying CIDOC CRM.

The current release of the CIDOC CRM (version 4.2) is structured into a class hierarchy (allowing for multiple inheritance) specifying 84 top-level entity classes and 141 relations (properties) between their instances (some of them also structured hierarchically). By convention, the names of classes always start with an **E** (entity) followed by a unique number, whereas properties are marked by a leading **P** (property), a unique number and (if an inverse property is defined) the letters **F** (forward) or **B** (backward). The properties are specified by restricting their domains and ranges and with regard to the classes by property quantification (cardinality

restrictions). The CIDOC manual however explicitly points out that these quantifications should not be treated as implementation recommendations to allow incomplete information within the knowledge base (difficulties for the similarity measurement arising from this semiformal kind of specification are discussed in section 4). The classes themselves are specified in an informal way as plain text description, except for their super/sub-class relations. Some of them are declared *abstract*, which means that they have no direct instances.

To adapt the generic CIDOC CRM framework to concrete annotation needs, each institution can define extensions (as long as they are consistent with the existing model) to the core model or use the *E55.Type* metaclass. This class, which's instances are in fact classes again, is intended to support the annotation of concrete types within metadata. In other words instances of *E55.Type* are categories, such as *naval engagement* or *war* for the class *E5.Event*, that are not specified within the CIDOC core model, but in external, application specific vocabularies. Difficulties arising from the extensive use of *E55.Type* are discussed in section 4.2.

A (partial) definition of the reference model is available as RDF schema in [3].

3 The Battle of Trafalgar

The scenario introduced within this section will later on be used to demonstrate certain aspects of the similarity-based identity assumption theory.

The Battle of Trafalgar is one of the most significant naval battles within the Napoleonic Wars and the 19th century. The battle took place during the Third Coalition War and prevented Napoleon's Invasion of Britain, establishing Royal Navy's position as the dominating naval power for more than a century.

Napoleon's strategy was to lure the Royal Navy away from the English Channel by attacking colonies in West Indies, then turn the fleet back to Europe, meet up with the allied Spanish fleet and jointly break the blockade at Brest to attack the remaining British fleet protecting the Channel, to establish a save passage for the French invasion troops. The responsible French Admiral de Villeneuve however ignored Napoleons strict order and sailed to the harbor of Cádiz near the Strait of Gibraltar. To permanently avoid a French invasion, the Royal Navy tried to block his fleet there, but instead of breaking out immediately, de Villeneuve hesitated and did not leave Cádiz until he was informed about Napoleons plan to replace him. The Royal Navy (under command of Horatio Nelson) was already waiting for this moment and attacked the disorganized Franco-Spanish fleet at Cape Trafalgar. The resulting battle was a great success for the British fleet because they destroyed or captured most of the enemies' ships without loosing one of their own. Admiral Nelson however was deadly wounded during the battle.

Of course our knowledge about the Battle of Trafalgar is very detailed and historically well documented; nevertheless the scenario satisfies our requirements as the following questions show: Which spatial relation holds between the naval battleground and the terrestrial cape of Trafalgar? Nelson was wounded during the battle, but did he die during or after it? As similarity measures the degree of overlap between assertions, we expect it to handle ambiguities arising from different

perspectives or ontological modeling decisions. The cape itself is located at a strategically prominent position at the Strait of Gibraltar and is therefore relevant for the European history (Carthaginian Empire, Roman Empire, Napoleonic Wars) as well as the African (Muslim Iberia). Hence the name and the geopolitical assignment to states, empires and provinces has changed over time (note that most gazetteers do not contain the Arabic name). Finally ships, which are of major interest for the Battle of Trafalgar, were frequently renamed (sometimes even several times within one year). Moreover the old names were reused for other ships during the same period. Therefore one cannot conclude from two datasets describing the participation of a ship, referenced by its name, within several historical events, that this particular ship actually sailed from one event to the other. In addition, three ships with similar names were involved in the Battle of Trafalgar called (H.M.S.) Neptune respectively Neptuno.

To measure similarity, all metadata concerning the battle itself and all entities linked to it have to be compared for overlap. To keep the scenario focused and concise, we limit the scenario to the Cape of Trafalgar and some selected, mostly spatiotemporal, relations to other places, events and actors.

4 Similarity-Based Identity Assumption Theory (SIAT)

Places referred to in historical sources (represented in CIDOC CRM as instances of *E53.Place*) probably refer to the identical real world place if they are related through the same or similar relationships to other instances, which themselves again refer to identical real world places, events, actors, or objects. These relations to other instances are annotated as RDF-triples. The more common triples two instances share, the more similar they are and the higher is the probability that both point to the same real world place. However, instances within a knowledge base always represent the approximated and partial knowledge an authority or museum has about a real world phenomenon. Hence, even if two instances share all triples, identity cannot be guaranteed. Therefore the identity assumption assistant delivers estimations rather than assertions. In other words, measuring similarity between real world places means to develop (or apply) distance metrics for their descriptions and to determine their overlap.

This section stepwise introduces the components forming the similarity theory and explains how they are jointly used for identity assumptions. Distance weightings (rephrased to similarity measures) for neighborhoods and hierarchies are discussed as well as inference rules generating new triples out of existing ones.

4.1 Recursive Similarity Function

This section elucidates how the similarity framework compares CIDOC CRM instances by recursively comparing their descriptions (RDF-triples) for overlap, while the concrete distance measures, prototypical expansion rules as well as the identity assumption itself are defined in later sections (see 4.2 and 4.3) .

Similarity between Predicate-Object-Tuples

As CIDOC CRM annotated metadata is represented by RDF-triples, these triples have to be compared by similarity measurement. Each triple consists of three components: the described resource itself (called subject), a relation (called predicate) and another resource (called object) linked to the first one by the chosen predicate. Note that the object itself may be the subject of another triple again. In other words a CIDOC CRM instance (subject) is described by its relations to other CIDOC CRM instances. Two RDF-triples are similar if their similar subjects are related by similar predicates to similar objects. Subject similarity however measures the overlap between all (predicate, object)-tuples describing the compared subjects.

Equation E1 defines the similarity (sim_t) for such tuples as the product of the predicate (p_1 and p_2) and object (o_1 and o_2) similarities. While the similarity between predicates (sim_p) is determined in notion of hierarchical or neighborhood distance (see section 4.2) the similarity between the involved objects is just the subject similarity (sim_s); reflecting the fact that those objects are again described by sets of (predicate, object)-tuples. Consequently, similarity does not only depend on the similarity of the referred objects, but also on the kind of this reference.

$$sim_t((p_1,o_1),(p_2,o_2)) = sim_p(p_1,p_2) * sim_s(o_1,o_2) \quad \textbf{(E1)}$$

If new triples are generated out of existing ones by inference rules (see section 4.2) and similarity between instances is measured by measuring similarity to other instances, then a maximum search depth has to be specified. This search depth determines the maximum number of expansions (using inference rules) and recursion steps before the measurement terminates. On the one side a low search depth decreases computing time, but on the other side also the expressiveness of the measurement. If the maximum search depth is reached, only the Resource-URIs (values of *RDF:about*) are compared (see section 4.2). However, this is rather a theoretical problem than of practical relevance for most local (cultural heritage) knowledge bases, because they focus on the description of their local exhibits and information and use additional resources/entities only as a kind of reference point.

In terms of the Battle of Trafalgar scenario the knowledge bases would contain all locally available knowledge about the cape (the subject) such as the events that took place there as well as the actors and objects participated in these events and the broader geopolitical units (in other words the objects of interest). The second level knowledge, however, would not again be described in such detail but more generic, while the objects (third level knowledge) involved in those descriptions may be only referenced to by global identifiers. Therefore the local knowledge base would not store all historical knowledge about Spain or even Europe.

According to Equation E1 the similarity derived from comparing the RDF-triples R1 and R2 about *Cape Trafalgar*, is the product of the similarity sim_p between *P89F.falls_within* and *P121.overlaps_with* and the similarity sim_s between *E53.Place(Province Cádiz) and E53.Place(Cádiz)*.

P89F.falls_within(E53.Place(Cape Trafalgar), E53.Place(Province Cádiz)) **(R1)**

P121.overlaps_with(E53.Place(Cape Trafalgar), E53.Place(Cádiz)) **(R2)**

Even if the compared representations of *Province Cádiz* and *Cádiz* would be equal, the overall similarity for the compared RDF-triples is decreased by the fact, that they describe different spatial relation between *Cape Trafalgar* and (*Province*) *Cádiz*.

Similarity between Subjects

Real world phenomena are not represented within knowledge bases as single RDF-triple, but as sets of them. Therefore the similarity between all RDF-triples that contain the intended instances as subject or object, have to bet taken into account. To avoid loops during comparison (see Equation E1) all predicates used in given and inferred triples have to be aligned in search direction before similarity is measured. In case of asymmetric relations this means that they have to be replaced with their counterparts. All triples with interchanged subjects and objects are removed from the set of compared triples and are therefore not taken into account for the similarity measurement if an "inverse" triple already exists.

Next the similarities sim_t (see Equation E1) between all (predicate, object)-tuples derived from the RDF-triples describing the local subject (called source) and those describing the compared-to subject (called target) have to be measured. In the following the resulting set of similarities (which is the Cartesian product of the sets of all RDF-triples from the source and the target subject with respect to their similarities) is stepwise processed so that the triples $((p_s,o_s), (p_t,o_t), sim_t)$ with the maximum similarity value for sim_t are saved for further processing and all triples containing either the involved source or target tuple are removed from the set of similarities.

The similarity between two compared subjects sim_s is just the normalized (to [0,1]) sum of these selected similarities (see Equation E2). Note that similarities involving the comparison of predicates between which a meaningful notion of distance cannot be defined (see section 4.2), as well as 'unused' tuples (if source and target are described by a different amount of RDF-triples), are not taken into account and are therefore not element of C (which moreover additionally decreases computing time).

$$sim_s(S_s, S_t) = \frac{1}{|C|}\sum_{i \in C} sim_{t_i} \quad \text{; where C:=\{i| i is selected } sim_t \text{ similarity value}\} \qquad \textbf{(E2)}$$

In terms of the Battle of Trafalgar scenario, if the source *Cape Trafalgar* is described by the RDF-triples R1, R3, R5 and the target cape is described by R2 and R4, the similarity between the source and target cape is:

$sim_s(S_s,S_t) = ½*(sim_t((p_{R1},o_{R1}),(p_{R2},o_{R2})) + sim_t((p_{R3},o_{R3}),(p_{R4},o_{R4})))$

P8B.witnessed(E53.Place(Cape Trafalgar), E7.Activity(Battle of Trafalgar)) (**R3**)

P8F.took_place_at(E7.Activity(Battle of Trafalgar), E53.Place(Cape Trafalgar)) (**R4**)

P53B.is_former_or_current_location_of(E53.Place(Cape Trafalgar), E24.Physical_Man_Made_Thing(HMS Victory)) (**R5**)

The set of selected similarities $C:=\{sim_t((p_{R1},o_{R1}),(p_{R2},o_{R2})); sim_t((p_{R3},o_{R3}),(p_{R4},o_{R4}))\}$ used for the computation of $sim_s(S_s,S_t)$ is derived from the Cartesian product of all potential similarities, however no other combinations are possible because distance cannot be measured between spatial and temporal predicates. Note that for

demonstration purpose R4 was specified with P8F (forward) and has to be switched to its inverse predicate P8B (backward) before measurement. To avoid back-pointing references, the RDF-triples R3 and R4 are not used later on within the comparisons of the *Battle of Trafalgar* instances.

The RDF-triple R5 does not influence the similarity between the compared instances. Insofar, and in contrast to models such as MDSM that define similarity as the ratio between common and distinguishing features, the similarity theory underlying SIAT can be regarded as a so called common elements approach [7] however it (in contrast to MDSM) supports partial matches.

4.2 Similarity Measures and their Application within SIAT

While the previous section describes the similarity framework as such, this section introduces the underlying similarity measures derived from converting distance weightings for predicates, types and identifiers.

Similarity within Hierarchies and Neighborhoods

The notion of distance (see also [17]) is used within SIAT as generic quantification for the relatedness between universals (predicates or types) arranged within neighborhoods or hierarchies. The similarity weightings discussed here are therefore inverse distance (dissimilarity) measures.

In contrast to theories assuming a constant distance within subsumption hierarchies, SIAT proposes a variable weighting depending on the hierarchy depth. This reflects the fact that abstract universals are less similar to each other than concrete ones, because those already share all features of their ancestors.

$$hsw(u_1,u_2) = \frac{depth(lub(u_1,u_2))}{depth(lub(u_1,u_2)) + edge\_distance(u_1,u_2)} \quad \textbf{(E3)}$$

In Equation E3 *hsw* is defined as the ratio of the hierarchical depth level of the least upper bound (*lub*) of the compared universals (u_1 ands u_2) and the sum of this depth and the edge distance between the those universals. The edge distance is the shortest path, in other words it is the number of edges to be passed from u_1 to u_2. This depth-weighted similarity is only applicable for subsumption hierarchies and used within SIAT to calculate the distance between hierarchically ordered CIDOC CRM predicates and types such as those derived from the ADL Gazetteer feature type hierarchy and the WordNet taxonomy.

The weighting *nsw*, specified in Equation E4, is used to calculate similarity between spatial and temporal CIDOC CRM predicates. Their graphs describe neighboring state changes (either in space or in time) instead of hierarchically ordered predicates with shared features and therefore one can not argue for a depth weighted measure.

$$nsw(u_1,u_2) = \frac{max\_distance - edge\_distance(u_1,u_2)}{max\_distance} \quad \textbf{(E4)}$$

In Equation E4 *max_distance* is defined as the maximal edge distance (longest path) within the neighborhood graph. With increasing graph depth the distance between adjacent nodes (here predicates) decreases, but is independent of the relative position of the nodes within the graph.

Topological Distance and Spatial Reasoning

In CIDOC CRM the class *E53.Place* represents extents in space in the pure sense of physics, independent of temporal and contextual constraints [3]. Scholars create instances of *E53.Place* within their documents to refer to a certain spatial extent on the surface of the earth that is of interest for some reason at a given time during history. The real world extent referred to is present over time while the point of interest may change its spatial disposition or even (temporally) disappear. The places can be identified by instances of *E44.Place_Appellation* which themselves are not necessarily stable over time and therefore may refer to several places. Moreover, historical knowledge is vague and incomplete and therefore the spatial extent described by an instance of *E53.Place* is not known exactly, but instead determined through its relation to other places within the SIAT approach. The points of interest are represented in CIDOC CRM by instance of *E18.Physical Thing* and its subclasses such as E27.Site.

Topological Distance

The CIDOC conceptual reference model distinguishes the following spatial relations between places (see [3] for disambiguation): *P88.consists_of (forms_part_of), P89.falls_within (contains), P121.overlaps_with* and *P122.borders_with***.** To measure similarity by comparing the relations to other places, a topological distance between the CIDOC CRM predicates has to be defined. To achieve this, the predicates are mapped (see Fig.1) to those defined in the Closest-Topological-Relationship-Graph [18]. The similarity (n*sw*) is applied to the graph to generate the weightings needed to calculate sim_p for (predicate, object)-tuples involving spatial predicates (see section 4.1). Relations that are topologically close have a higher similarity value and therefore more impact on the similarity of the places they refer to. These places are again compared by taking into account their spatial relations to other places and so on.

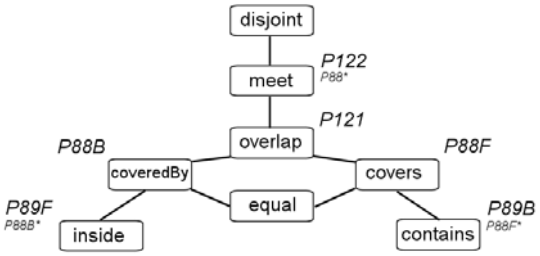

Fig. 1: The CICOC CRM spatial relations within the Closest-Topological-Relationship-Graph

The CIDOC conceptual reference model does not specify relation for *disjoint* and *equal* and therefore no mapping is possible. The properties *P88F.consists_of* and *P88B.forms_part_of* describe the fact that a place can be subdivided into one or more

constituents (which are places themselves again), and implies spatial as well as contextual containment. This kind of composition cannot be clearly mapped to one of the topological relationships within the graph and therefore may be assigned to *covers/coveredBy* and *meet* (or even *inside/contains*) as well. SIAT maps the *P88* relations to *covers/coveredBy* to refer to the idea of common boundaries (purely spatial) as well as the aspect of containment (spatial and contextual). However this decision is debatable and should be answered by empirical findings within complex real world applications (see section 5).

With regard to the Battle of Trafalgar scenario the similarity between the tuples derived from R1 and R2 about Cape Trafalgar is:
$sim_t(R1, R2)=0.5 * sim_s(E53.Place(Province Cádiz), E53.Place(Cádiz))$

Spatial Reasoning
Besides topological neighborhood, spatial inference is used within SIAT to increase the available amount of place information used for the similarity measurement (see section 4.1). These inferences are drawn from simplified reasoning rules based on the spatial relations introduced in CIDOC CRM and their combination. Each applied spatial reasoning rule generates one or more new RDF-triples. Two rules are illustrated here representative for spatial inference rules in general.

$$\text{AND}(P89F(E53(x), E53(y)), P89F(E53(y), E53(z))) \rightarrow P89F(E53(x), E53(z)) \quad (\textbf{S1F})$$

$$\text{AND}(P89B(E53(x), E53(y)), P89B(E53(y), E53(z))) \rightarrow P89B(E53(x), E53(z)) \quad (\textbf{S1B})$$

Rule S1F and S1B generate new triples based on the transitivity of *P89*.

$$\text{AND}(P121(E53(x), E53(y)), P89F(E53(y), E53(z))) \rightarrow P121(E53(x), E53(z)) \quad (\textbf{S2})$$

Rule S2 infers from the triples (x overlaps_with y) and (y falls_within z) a new triple stating that x overlaps with z. In same cases x could also fall within z, but this could not be concluded for sure.

Temporal Distance and Temporal Reasoning
Besides the relationship to other places, historical events can act as reference points for identity assumptions. If two instances of *E53.Place* are related in a similar way to a certain event, they probably refer to the same real world place. However this implies that the identity of the event can be assumed out of its representation in the database. In CIDOC CRM *E2.Temporal_Entity* is defined as the abstract root class of all perdurants. Its direct subclass *E2.Period* describes (historical) periods as well as all kinds of events (*E5.Event*) which are further distinguished into instances of *E7.Activity*, *E63.Beginning_of_Existence* or *E64.End_of_Existence* (see [3] for further subtypes and disambiguation). All periods are related to at least one place (*E53.Place)* by the *P7.took_place_at (witnessed)* relation.

Temporal Distance
The following relations between temporal entities are distinguished (and related to Allen's temporal logic [19]) within the CIDOC CRM: *P114.is_equal_in_time_to*,

P115.finishes (is_finished_by), P116.starts (is_started_by), P117.occurs_during (includes), P118.overlaps_in_time_with (is_overlapped_in_time_by), P119.meets_in_time_with (is_met_in_time_by), and P120.occures_before (occurs_after).

To determine the distance between these predicates, we use Freksa's conceptual neighborhood [20]. As our knowledge about historical periods is incomplete and often no crisp starting and ending points are defined, we assume that the 'temporal location' of events [20] is more or less fixed, while their duration varies. Therefore the C-Neighbor structure is chosen as model for temporal distance within SIAT (see Fig. 2).

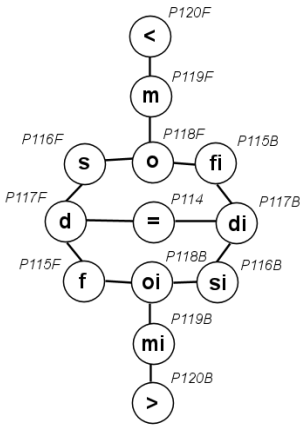

Fig. 2: The CICOC CRM temporal relations within the Freksa's [20] C-Neighborhood

In addition to this purely temporal relation, CIDOC CRM also introduces spatio-temporal relations such as *P9.consists_of (forms_part_of), P10.falls_within (contains)* and *P132.overlaps_with* between periods. These properties cannot be integrated into the temporal conceptual neighborhood because they do not only assume a temporal relation between the periods, but also one between the places at which the periods took place [3].One may argue that these properties can be split up into their temporal and special aspects and then be integrated into the according graphs; however this is not possible for *P132.overlaps_with* because the overlapping relationship within the C-Neighborhood is not symmetric. Therefore the integration of this additional predicates is not discussed here in detail, but left for future work.

In the used scenario, R3 and R4 describe Cape Trafalgar by the fact that it is the place where the Battle of Trafalgar took place. This is the usual way how places are connected to events (and vice versa) in CIDOC CRM. However to determine whether the same battle is meant, the similarity measure has to compare not only the relating predicates (using *nsw*) but also the related-to objects (here temporal entities). If we assume that (beside others) the battle in the source annotation is described by R6 and the one in the target annotation by R7, the similarity is:

sim_t(R6,R7)=0.63*sim_s(E5.Event(Trafalgar Campaign),E7.Activity(Battle of Cape Ortegal))

P117F.occurs_during (E7.Activity(Battle of Trafalgar), E5.Event(Trafalgar Campaign)) (**R6**)

P114. is_finished_by(E5.Activity(Battle of Trafalgar), E5.Activity(Battle of Cape Ortegal)) (**R7**)

Note that the Battle of Cape Ortegal is one of the three battles of the Trafalgar Campaign.

Temporal Reasoning
The same kind of simplified inference rules discussed for spatial reasoning is also applied to generate new triples out of existing information. Two rules are introduced here representative for temporal inference rules in general.

$$\text{AND}(\text{P117F}(\text{E4}(x), \text{E4}(y)), \text{P117F}(\text{E4}(y), \text{E4}(z))) \rightarrow \text{P117F}(\text{E4}(x), \text{E4}(z)) \quad (\textbf{T1F})$$

$$\text{AND}(\text{P117B}(\text{E4}(x), \text{E4}(y)), \text{P117B}(\text{E4}(y), \text{E4}(z))) \rightarrow \text{P117B}(\text{E4}(x), \text{E4}(z)) \quad (\textbf{T1B})$$

Rule T1F and T1B generate new triples based on the transitivity of *P117*.

$$\text{AND}(\text{P132}(\text{E 4}(x), \text{E4}(y)), \text{P10F}(\text{E4}(y), \text{E4}(z))) \rightarrow \text{P121}(\text{E4}(x), \text{E4}(z)) \quad (\textbf{T2})$$

Rule T2 is the spatiotemporal equivalent to S1. Note that one could moreover infer purely temporal and purely spatial relations between the involved periods and places, which are not discussed here in detail.

Referential Relations
Besides spatial and temporal relationships to other places or events, referential information is of major importance for identity assumptions. These appellations include all kind of names, (structured) phrases, codes and marks intended to identify a certain instance of a given class in a known context [3].

Within SIAT these appellations (besides types) act as termination point for the recursive similarity function because concrete appellations are not compared by (predicate, object)-tuples again, but by external similarity (respectively distance) measures established for given kinds of appellations such as distance between spatial coordinates, notions of prototypical distances between postal codes as well as temporal distances between time points or spans and purely syntactical edit-distance measures between names. However this distances have to be normalized (and if necessary transformed and classified) to values between 0 and 1 before their integration into SIAT. For instance in cases of global identifiers, 0 should be returned for different and 1 for exactly the same global identifier because – at least in the context of cultural heritage – no meaningful notion of distance between identifiers could be defined.

The description of these measures for all available kinds of appellations is out of the scope of this paper and therefore not discussed here in detail. This is especially because we assume that only some well known (broader / upper level) entities such as countries or epochs are annotated by (unambiguous) appellations and focus on a notion of place identity based on vague knowledge about their spatial, temporal and thematic alignment within historical knowledge.

In general referential information can be obtained from Gazetteers[1], databases about historical events or actors, text corpora, several kinds of global authorities as

[1] Note that the province Cádiz is represented as point-geometry in the ADL Gazetteer and Getty Thesaurus. The (lat/lon) coordinates between both vary about 50km: whereas those from Getty are the same as the coordinates of the city of Cádiz, the coordinates from ADL point to the center of the province. Comparing these coordinates by spatial distance measures (with Spain as reference) would therefore result in a similarity value interpreted to *nearby* instead of *equal* (even for tolerant equal-buffers).

well as from local knowledge or convention. Within CIDOC CRM it is represented by instances of *E41.Appellation*. Place appellations (*E44.Place_Appellation*) are further distinguished in *E45.Address*, *E46.Section Definition*, *E47.Spatial_Coordinates* and *E48.Place_Name* while *E49.Time_Appellation* (and its subclass *E50.Date*) comprises all kinds of references to time-spans. Both the degree of precision and concrete format of the appellations are not specified or restricted by the CIDOC conceptual reference model. All appellations are related to the referred instances by *P1.is_identified_by (identifies)* and its subrelations.

Note that in CIDOC CRM annotated documents *RDF:about* is used for the concrete value of the appellation (in the sense of an identifier) as well as for a description of the resource itself and therefore SIAT has to interpret RDF:about as predicate (sim_p) and its value by the appropriate external similarity theory (string matching in the worst case) which may have strong influence on the quality of the similarity assessment[2]. However we do not claim that this is indented by the CIDOC model but seems to be usual annotation practice.

Actors and Physical Things

Relationships to prominent actors (*E39.Actor*) or physical things (*E18.Physical_Thing*) can be applied the same way as the relations to events are used to generate assumptions about places. On behalf of hierarchically ordered relations in general, the participation of actors within events is discussed here briefly.

P12.occurred_in_the_presence_of (was_present_at) is the most generic relationship defined between persistent items (*E77.Persistent_Item*) and events within CICOD CRM. Its subproperty *P11.had_participant (participated_in)* restricts the range to actors and describes the active as well as passive participation in an event. To emphasize intentionally (and therefore active) participation in a certain activity (*E7.Activity*), the subproperty *P14.carried out by (performed)* is used. These predicates can be compared to each other using *hsw* for the inter-predicate similarity sim_p. Note that as no root relation is defined in CIDOC CRM, the distance between relations not ordered hierarchically (except those for which a neighbourhood is defined such as temporal and spatial relations) cannot be calculated and therefore sim_p is 0 by definition and the according (predicate, object)-tuple has no influence on the similarity between compared subjects.

In terms of the Battle of Trafalgar scenario the comparison of the triples R8 and R9 as part of the similarity measurement between the compared battles is calculated as follows and tends to 0:

sim_t(R8,R9)=0.33*sim_s(E21.Person(Nelson), E19.Physical_Object(HMS Victory))

P14F.carried_out_by (E7.Activity(Battle of Trafalgar), E21.Person(Nelson)) (**R8**)

P12B.was_present_at (E19.Physical_Object(HMS Victory),E7.Activity(Battle of Trafalgar)) (**R9**)

[2] In technical terms this involves constructs such as: <crm:E47.Spatial_Coordinates rdf:about="Lat: 36.5333, Long: -6.3000">…(where the type of coordinate system can be specified by E55.Type) or <crm:E69.Death rdf:about="Death of Nelson on the deck of H.M.S. Victory">….

Distance between Types

CIDOC CRM annotated documents make extensive use of the class *E55.Type* and the *P2.has_type (is_type_of)* relation to express all kinds of classifications not further distinguished in the core model. Within SIAT we focus on types of *E4.Event* and *E53.Place*, but other instances of *E55.Type* can be compared accordingly.

The ADL Gazetteer [5] and the Getty Thesaurus [6] do not only deliver unique identifiers to unambiguously link place names to certain geophysical or geopolitical entities, but also deliver feature types for these entities. The ADL feature types are considered here because they are commonly used within the domain of cultural heritage and moreover are organized hierarchically. Since no comparable formal definition is available for the feature types themselves, the introduced *hsw* measure is applied to determine how close two types are related to each other. The ADL thesaurus has no common root element and defines six top types (administrative areas, hydrographic features, land parcels, manmade features, physiographic features, and regions) instead [5]. Therefore *hsw* returns the similarity value 0 for types that do not belong to a common top type, which expresses the fact that these types are fundamentally different.

For the scenario this means that the comparison of types specified in R10 and R11 similarity yields 0, because of a missing common super type for *administrative areas* and *regions*[3].

P2F.has_type (E53.Place(Province Cádiz), E55.Type(administrative areas)) (**R10**)

P2F.has_type (E53.Place(Andalusia), E55.Type(regions)) (**R11**)

To specify and compare types of events the, WordNet [21] hypernym/hyponym hierarchy is chosen and "event" (WordNet database location: {00028105}) is defined as top term. Again *hsw* is chosen as to determine the degree of similarity. The nodes within the WordNet taxonomy are not necessarily single terms but synsets (sets of synonym terms). The edge_distance within a synset is set to 0.

WordNet (and *hsw*) can also be used to compare types of *E70.Thing* and *E39.Actor*. This is, however, not discussed here in detail.

4.3 Identity Assumptions

Similarity is measured between instances, whereas identity is assumed for real world phenomena. A high similarity in general indicates that the compared instances are closely related together in terms of the comparable parts of their descriptions. In SIAT this similarity is primarily measured by comparing the spatial disposition and temporal witness [3] of instances representing real world places. Following the law that set cardinality is decreasing with an increasing amount of membership

[3] The example was chosen intentionally to point out difficulties concerning uncertainty within and between global authorities: While ADL describes Andalusia as *region* and the province Cádiz as *administrative area*; Getty marks Andalusia as *autonomous community* and *first level subdivision* (and therefore as administrative record type) and the province Cádiz as *province* and *second level subdivision.* Note that the ADL Feature Thesaurus also specifies types as *countries, 1st order divisions,* but they are not applied to the examined places.

restrictions, even if different real world places share the same topological relations to other common real world places and even if the same real world event took place at different locations at the same time, it is the more improbable to find such real world places, the more common relationships they need to share. However, to draw this conclusion, two additional parameters are needed: on the one hand the number (n_t) of compared (predicate, object)-tuples has to be taken into account (for sim_s) and on the other hand a measure has to be specified that allows to consider the information value of the returned similarity. Within SIAT this value (s), that could be compared to the notion of variance, is just the difference between $sim_s(S_s,S_t)$ received from E2 and the similarity obtained by taking into account also those tuples that yield 0 because no meaningful distance between them could be specified.
Therefore in fact SIAT does not deliver assumptions about identity, but returns a triple (IA in Equation E5) describing how unlikely the compared instances refer to different places. The term unlikely is used here intentionally instead of improbably to point out the vague nature of such assumptions.

$$IA = \langle sim_s(S_s,S_t),\ n_t,\ s \rangle \qquad \textbf{(E5)}$$

Promising identity assumptions are those where the overall similarity sim_s as well as the number of compared tuples n_t are high and the number of tuples that could not be compared and therefore have no impact on the identity assumption is low (reflected by a small s value). However the last named parameter should not be interpreted in a way that high s values automatically exclude an identity assumption.

In terms of the Battle of Trafalgar scenario, RDF-triples stored in local knowledge bases describe the information about the cape and the battle from the perspective and standard of knowledge of the examined historical sources and therefore may differ in their granularity, perspective and historical reference frame. A document describing the importance of Cape Trafalgar for the history of Spain shares some information about the relation to other places, events and actors with a British view on the Battle of Trafalgar, but contains additional knowledge and focus on other participants.

5 Discussion and Future Extensions

The theory introduced in this paper compares CIDOC CRM instances by extracting their relations to other instances and recursively comparing both, the relations and the related-to instances. Each resulting tuple from the source instance is compared to the most similar of the target instance whereas each tuple is only used once. The process terminates when all instances (each object of a RDF-triple may be the subject of the next similarity measurement step) are examined and only primitive values are left (primitive in a sense that they are the values of *RDF:about* and could not be further decomposed within the CIDOC framework). The similarity between these primitives, describing all kinds of appellations or types, is determined by using external (not recursive) measures such as distances within type hierarchies or between spatial coordinates. If a high similarity value is obtained from a sufficient number of compared tuples, it is possible, but improbable, to find more than one place that meets the required description (which is independent from the question whether the

annotations and the historical knowledge reflect 'true' information about the compared real world phenomena or not).

Promising results (IA-triples) are reported back to the user for further examination. Whenever a scholar decides to trust a certain assumption, the information provided by the external data source can be used to validate or enrich the local knowledge about the referred real world place. Moreover it becomes possible to establish a persistent link between both data sources used for complex queries. Such queries across multiple databases can provide solutions to scientific questions that could not be answered before.

In contrast to many existing similarity theories the measurement framework presented here focuses on the integration of classical reasoning tasks (to make hidden knowledge explicit and therefore increase the number of comparable tuples) and similarity (to return vague assumptions). Moreover the theory supports partial matches (not possible in models such as MDSM [8]) and integrates spatial, temporal and thematic aspects within one similarity framework.

Nevertheless a lot of work remains to be done and should focus on the application in complex real world scenarios on the one hand and the refinement (supported by empirical finding) of the measurement on the other hand. A fixed set of inference rules should be established for a concrete implementation of the assistant. It has to be examined whether contradicting information expressed in compared RDF-triples should decrease similarity taking into account the vague and incomplete character of historical knowledge. Moreover a theory of trust has to be defined and integrated into the identity assumption theory to indicate how much the user trusts a certain authority. More work is also needed to answer the question how the identity assumptions have to be presented to the user and what kind of additional information is necessary to support scholars in verifying them and their quality. The questions how many compared tuples are sufficient for a precise assumption and whether the similarity of predicates should be weighted differently from the similarity between the related-to objects, cannot be answered beforehand, but only through extensive applications in real world scenarios. Additionally we do not claim that similarity is the only strategy to create identity assumptions from RDF-triples, other approaches have to be examined and integrated into an overall theory.

Acknowledgement

The presented work was inspired and influenced by Martin Doerr who raised the idea of an identity assumption service for historical places as core component of a grid architecture for the domain of cultural heritage. Moreover the author thanks the three anonymous reviewers as well as the MUSIL group for their fruitful comments.

References

1. Raubal, M., *Formalizing Conceptual Spaces*, in *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, A. Varzi and L. Vieu, Editors. 2004, IOS Press: Amsterdam, NL. p. 153-164.
2. Janowicz, K., *SIM-DL: Towards a Similarity Measurement Theory for Description Logics in GIScience*. under review 2006.
3. Crofts, N., et al., *Definition of the CIDOC Conceptual Reference Model (version 4.2)*. 2005.
4. *The Open Archives Initiative Protocol for Metadata Harvesting (Version 2.0): http://www.openarchives.org/OAI/openarchivesprotocol.html*
5. *Alexandria Digital Library Gazetteer: http://middleware.alexandria.ucsb.edu.*
6. *Getty Thesaurus of Geographic Names: http://www.getty.edu/vow/TGNFullDisplay*
7. Goldstone, R. and J. Son, *Similarity*, in *Cambridge Handbook of Thinking and Reasoning*, K. Holyoak and R. Morrison, Editors. 2004, Cambridge University Press
8. Rodríguez, A.M. and M.J. Egenhofer, *Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure.* International Journal of Geographical Information Science, 2004. **18**(3): p. 229-256.
9. Janowicz, K., *Extending Semantic Similarity Measurement by Thematic Roles*, in *First International Conference on GeoSpatial Semantics, GeoS 2005, Mexico City, November 29-30, 2005*. 2005, Springer Verlag: Berlin. p. 137-152.
10. Gärdenfors, P., *Conceptual Spaces - The Geometry of Thought*. 2000, Cambridge, MA: Bradford Books, MIT Press.
11. Schwering, A. and M. Raubal, *Measuring Semantic Similarity between Geospatial Conceptual Regions*, in *First International Conference on GeoSpatial Semantics, GeoS 2005, Mexico City, November 29-30, 2005*. 2005, Springer-Verlag: Berlin. p. 90-106.
12. Schwering, A. *Hybrid Model for Semantic Similarity Measurement*. *4th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE05)*. 2005. Agia Napa, Cyprus: Springer Verlag: Berlin. p.1449-1465.
13. Hau, J., W. Lee, and J. Darlington. *A Semantic Similarity Measure for Semantic Web Services*. in *Web Service Semantics Workshop 2005 at WWW2005*. 2005. Japan.
14. Maedche, A. and S. Staab. *Measuring Similarity between Ontologies*. in *European Conference on Knowledge Acquisition and Management (EKAW)*. 2002. Spain.
15. Ehrig, M., et al. *Similarity for Ontologies - A Comprehensive Framework*. in *13th European Conference on Information Systems*. 2005. Germany.
16. Borgida, A., T.J. Walsh, and H. Hirsh. *Towards Measuring Similarity in Description Logics*. in *International Workshop on Description Logics (DL2005)*. 2005. Scotland.
17. Rada, R., et al., *Development and Application of a Metric on Semantic Nets.* IEEE Transaction on Systems, Man, and Cybernetics, 1989. **19**(1): p. 17-30.
18. Egenhofer, M. and K. Al-Taha, *Reasoning About Gradual Changes of Topological Relationships*, in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, A. Frank, I. Campari, and U. Formentini, Editors. 1992, Springer Verlag: Berlin. p. 196-219.
19. Allen, *Maintaining Knowledge about Temporal Intervals.* Communications of the ACM, 1983. **26**: p. 832-843.
20. Freksa, C., *Temporal Reasoning Based on Semi-Intervals.* Artificial Intelligence, 1992. **54**(1): p. 199-227.
21. WordNet: *http://wordnet.princeton.edu/*.