Matching Formal and Informal Geospatial Ontologies

Heshan Du, Natasha Alechina, Mike Jackson and Glen Hart

Abstract The rapid development of crowd-sourcing or volunteered geographic information both challenges and provides opportunities to authoritative geospatial information. Matching geospatial ontologies is an essential element to realizing the synergistic use of disparate geospatial information. We propose a new semi-automatic method to match formal and informal real life geospatial ontologies, at both terminology level and instance level, ensuring that overall information is logically coherent and consistent. Disparate geospatial ontologies are matched by finding a consistent and coherent set of mapping axioms with respect to them. Disjointness axioms are generated in order to facilitate detection of errors. In contrast to other existing methods, disjointness axioms are seen as assumptions, which can be retracted during the overall process. We produce candidates for retraction automatically, but the ultimate decision is taken by domain experts. Geometry matching, lexical matching and cardinality checking are combined when matching geospatial individuals (spatial features).

1 Introduction

In recent years, the emergence and development of crowd-sourcing or volunteered geographic information has challenged and also provided opportunities to the traditional model of geospatial data collection, storage and updates. Allowing amateurs to collect geospatial data helps lower the cost, capture richer user-based information and reflect real world changes more quickly. At the same time it may

H. Du (✉) · N. Alechina · M. Jackson
The University of Nottingham, Nottingham, UK
e-mail: psxhd1@nottingham.ac.uk

G. Hart
Ordnance Survey of Great Britain, Southampton, UK

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_9,

also dilute information quality, such as completeness, consistency and accuracy (Jackson et al. 2010). It is desirable to use volunteered and authoritative geospatial information as complements to each other, taking the best of both.

Ontology refers to an explicit specification of a shared conceptualization (Gruber 1993) and plays an important role in establishing shared formal vocabularies. A spatial individual has a certain and verifiable location and a meaningful label, which together distinguish itself from others. Geospatial ontologies describe conceptual hierarchies and interrelations of terminologies in the domain of geospatial science, which are used to describe facts (classifications, relations, attributions and locations) about spatial individuals. Compared to other ontologies, geospatial ontologies have some special properties. Firstly, many geospatial terminologies are commonly used in daily life and their meanings vary in different contexts. For example, "College" may refer to an institution within a university in one ontology, whilst meaning a secondary school in another. In addition, geospatial ontologies often do not have a huge number of classes as ontologies in several other subject areas (for example, biomedicine) do, but may represent many real world spatial individuals, whose locations, at least in theory, can be verified. For example, *Space*, a large-scale geospatial ontology constructed using WordNet, GeoNames and Thesaurus of Geographical Names, contains 845 classes and 6, 907, 417 individuals (Giunchiglia et al. 2012). Since geospatial ontologies for authoritative and volunteered data sets are developed independently, matching geospatial ontologies is an essential step to use them synergistically.

We propose a new semi-automatic method for matching geospatial ontologies, at both terminology level and instance level. Geographic information quality includes several aspects, viz., completeness, logical consistency, positional accuracy, thematic accuracy, temporal quality and usability (International Organization for Standardization 2011). We focus on logical consistency, ensuring that, after adding a mapping, overall information is logically coherent and consistent, without any contradictions. We assume that the TBoxes[1] of geospatial ontologies are not very large, but contain concepts which are more ambiguous, compared to, for example, biomedical ontologies. The matching process is reduced to the problem of finding a coherent and consistent set of assumptions (including disjointness axioms, equivalence and inclusion relations between concepts from different ontologies, and "sameAs" and "partOf" relations between spatial instances) with respect to input ontologies. Unlike a premise, an assumption is believed by default, but can be retracted later if found to be not reasonable later. Disjointness axioms are generated in order to facilitate detection of errors or contradictions. Geospatial individuals are matched using location, lexical and classification information.

The rest of the chapter is organized as follows. Section 2 reviews related work on ontology matching and geospatial data integration. We describe our new method in Sect. 3, and evaluate it in Sect. 4. Finally, Sect. 5 provides conclusions.

[1] Definitions of concepts and roles.

2 Related Research

Ontology matching is the task of finding a mapping, i.e. a set of correspondences, between entities from different ontologies (Euzenat and Shvaiko 2007). It includes two main levels, the terminology level and instance level. Many ontology matching methods and systems have been developed in recent years (Euzenat and Shvaiko 2007; Shvaiko and Euzenat 2012). Most of them are based on lexical and structural analysis and similarity measurements. However, mappings generated by these methods often contain logical contradictions. Logical reasoning is employed for either mapping generation or verification in some systems, including the early logic-based attempts, such as CtxMatch (Bouquet et al. 2003) and its extension S-Match (Giunchiglia et al. 2004), and more recently, ASMOV (Jean-Mary et al. 2010) which verifies mappings against five specified inconsistent patterns, KOSIMap (Reul and Pan 2010) based on description logic coherence checking assuming the disjointness of siblings, ContentMap (Jimenez-Ruiz et al. 2009) which computes new entailments from initial mappings generated by other systems, LogMap (Jimenez-Ruiz and Grau 2011) and CODI (Niepert et al. 2010).

LogMap is a logic-based ontology matching tool, designed for large-scale biomedical ontologies. It employs lexical and structural methods to compute an initial mapping. LogMap iterates two main steps. In Step 1, unsatisfiable classes will be detected using propositional Horn representation and satisfiability checking, and be repaired using a greedy diagnosis algorithm. However, the propositional Horn satisfiability checking is sound but incomplete, and the underlying semantics is restricted to propositional logic, and thus cannot guarantee the coherence of the mapping between more expressive ontologies. In Step 2, new mapping relations will be generated based on the similarity of classes related to established correspondences. Only newly discovered correspondences can be eliminated in the repair step, whilst correspondences found in earlier iterations are seen as established or valid. In other words, each mapping relation will be checked once, against the available information at that time, which, however, cannot guarantee its correctness when new information is discovered later.

CODI is a probabilistic logical alignment system based on Markov logic (Domingos et al. 2008). It transforms the matching problem to a maximum-a posterior optimization problem subject to cardinality constraints, coherence constraints and stability constraints. The GUROBI optimizer (Gurobi Optimization Inc 2012) is employed to solve the optimization problems. CODI reduces incoherence during the alignment process for the first time, compared to all other existing methods repairing alignments afterwards. CODI is based on the rationale of finding the most likely mapping by maximizing the sum of similarity-weighted probabilities for potential correspondences. However, during the optimization process, some valid correspondences can be thrown away.

It is a central problem within the context of Linked Data to identify correspondences between instances from different sources. Wolger et al. (2011) provide a summary of the existing data interlinking methods. Most of them are based on

lexical methods, such as string matching and word relation matching, machine learning and natural language processing techniques. Only within some systems, such as L2R (Sais et al. 2007), KnoFuss (Nikolov et al. 2007) and RDF-AI (Scharffe et al. 2009), consistency is checked.

In addition, there is some recent work on debugging and repairing ontology mappings (Meilicke and Stuckenschmidt 2009; Qi et al. 2009; Wang and Xu 2008), which is still at an early stage. However, to the best of our knowledge, all of them, as well as the ontology matching or data interlinking systems described above, treat disjointness axioms as premises, rather than retractable assumptions, and none of them have addressed the special properties of geospatial information.

In geospatial information science, several data conflation methods have been developed for matching or integrating geospatial vector data, mainly based on the similarities of geometries or topological relations, as well as attributes, if available. Most of them focus on conflating road vector data. However, few of these techniques check and ensure the logical coherence and consistency of integrated information (Du et al. 2012). In addition, several ontology-driven methods have been developed for integrating geospatial terminologies. Most of them are based on similarity measures or a predefined top-level ontology. Logical reasoning is only employed when formal ontologies commit to a same top-level ontology (Buccella et al. 2009). However, when ontologies are developed independently, the common top-level ontology is not usually available. Additionally, there exist some other methods (Volz and Walter 2004; Jain et al. 2010), following the bottom-up approach to linking geospatial schemas or ontologies, inferring terminology correspondences from instances correspondences. Though this works well when instance data is representative and overlapping, it uses a very strong form of induction from particular to the universal, which leads to lack of correctness and completeness (Bouquet 2007). Therefore, more research is required to fill in the gap, exploring logic-based approaches to matching geospatial ontologies.

3 Method

We propose a new semi-automated method for matching geospatial ontologies. Initial mappings between concepts and between individuals are generated using lexical matching and geometry matching. Logical coherence and consistency is ensured by automatically generating sets of assumptions responsible for incoherence or inconsistency using description logic reasoner Pellet (Sirin et al. 2007), and asking domain experts to decide which assumptions from these sets should be removed to restore coherence and consistency. Due to limited space, we recommend Baader et al. (2007) for the basic notions of description logic.

Definition 1 (*Ontology*) An ontology O has a TBox which contains knowledge at the conceptual level, and an ABox which describes facts about individuals using terminologies described in the TBox.

Definition 2 (*Coherence*) An ontology O is coherent if there is no class which only admits an empty interpretation. Otherwise, it is incoherent.

Definition 3 (*Consistency*) An ontology O is consistent if

- there exists no individual name a can be shown to belong to a concept C and to its negation, C;
- there exists no individual names a, b can be shown to belong to a role R and its negation, R;

Otherwise, O is inconsistent.

This method matches ontologies from the terminology level to the instance level. It includes four main steps. Since the original ontologies are often lightweight, disjointness axioms are generated in *Step 1*, to facilitate detection of incoherencies and inconsistencies. Ontology TBoxes are matched in *Step 2*. In *Step 3*, we match ABoxes of geospatial ontologies using location and lexical information. The whole ontologies are matched in *Step 4*. Mapping relations are represented as axioms in standard description logics, making use of existing and highly optimized reasoning techniques, for example Pellet (Sirin et al. 2007). Differing from other existing methods, this method treats generated disjointness axioms and the mappings between different ontologies, as *assumptions*, rather than premises. Users are allowed to retract or enable existing assumptions, and add new assumptions, during the matching process.

Definition 4 (*Premise and Assumption*) A premise is believed all the time, whilst an assumption is believed by default, but may be retracted later.

To represent and reason with two ontologies O^i and O^j, where i, j are their names, as well as the mapping M between them, as if they all belong to one super ontology ($O^i \cup O^j \cup M$), we label all atomic concepts, roles and individual names in each ontology by the name of the ontology. An atomic concept C and an individual name a from ontology i are represented as i: C and i: a respectively.

Definition 5 (*Union of Ontologies*) The union of ontologies O^i and O^j, represented as ($O^i \cup O^j$), is an ontology containing all axioms in O^i and O^j.

3.1 Matching Terminologies

A terminology mapping is a set of correspondences between concepts from different ontologies. A terminology correspondence is represented in one of the two basic forms:

$$B^i \sqsubseteq C^j \tag{1}$$

$$B^i \sqsupseteq C^j \tag{2}$$

where B, C denote concepts.[2] The relation (1) states that the concept B from the ontology i is more specific than or equivalent to the concept C from the ontology j. The relation (2) states that the concept B from the ontology i is more general than or equivalent to the concept C from the ontology j. The equivalence relation (3) holds if and only if (1) and (2) both hold.

$$B^i \equiv C^j \tag{3}$$

It states that the concept B from the ontology i and the concept C from the ontology j are equivalent.

A disjointness axiom states that two or more concepts are pairwise disjoint, having no common element. For example, *Person* and *Place* are disjoint, which can be represented as $Person \sqsubseteq \neg\ Place$, where $\neg$ denotes negation. Disjointness axioms in ontologies play an important role in debugging ontology mappings. However, within original geospatial ontologies, disjointness axioms are not always available or sufficient. Adding disjointness axioms manually, especially for large ontologies, is time-consuming and error-prone. Many existing systems employ more automatic approaches, either assuming the disjointness of siblings (e.g. Reul and Pan 2010), or employing machine learning techniques to detect disjointness (e.g. Meilicke et al. 2008a). After disjointness axioms are generated by whatever means, all existing ontology matching or debugging methods, to the best of our knowledge, use them as premises, though the input disjointness axioms can be insufficient or too restrictive. Differing from these methods, we use generated disjointness axioms as assumptions, and ensure the assumption set is coherent.

Definition 6 (*Coherence of an Assumption Set*) An assumption set A_s is incoherent with respect to an ontology O, if $O \cup A_s$ is incoherent, but O is coherent. Otherwise, it is coherent with respect to an ontology O.

When some incoherence is introduced by assumptions, minimal incoherent assumption sets (MIA) will be computed. The notion of MIA is defined by extending the minimal conflict set defined for mappings (Meilicke et al. 2008b) to this context.

Definition 7 (*Minimal Incoherent Assumption Set*) Given a set of assumptions A_s, a set $C \subseteq A_s$ is a minimal incoherent assumption set (MIA) iff C is incoherent and each $C' \subset C$ is coherent.

A minimal incoherent assumption set can be fixed by removing any axiom from it. When a MIA contains more than one element, one needs to decide which axiom to remove. Most of the existing methods remove the one either with the lowest confidence value or which is the least relevant. However, there is no consensus with respect to the measure of the degree of confidence or relevance. In several cases, confidence values or relevance degrees might be unavailable or difficult to

[2] When B, C denote atomic concepts, $B^i = i{:}\ B$, $C^j = j{:}\ C$.

compute or compare. Rather than relying on them, we allow domain experts to make ultimate decisions.

Algorithm 1 is designed to generate a coherent assumption set (CAS) with respect to an ontology.[3] The set of minimal incoherent assumption sets will be visualized clearly (Line 5). Domain experts are employed to take repair actions (Line 6). Currently, a repair action can be retracting an assumption axiom. Users are allowed to take several repair actions at one time.

ALGORITHM 1: CAS
Input: O: a coherent ontology
A_s: an assumption set for O
Output: A_{cs} : a coherent assumption set with respect to O. $A_{cs} \subseteq A_s$.
1. $A_{cs} \coloneqq A_s$;
2. $O_{tmp} := O \cup A_{cs}$;
3. **while** O_{tmp} is incoherent **do**
4. $S_{mia} := MIA(O_{tmp})$;
5. *visualization*(S_{mia});
6. * *repair*(O_{tmp}, S_{mia});
7. *update*(A_{cs});
8. **end while**
9. **return** A_{cs}

Step 1: Generating coherent disjointness assumption sets (CDAS). For each coherent ontology TBox,[4] as shown in *Algorithm 2*, we generate disjointness axioms as assumptions for sibling classes and refine them by applying *Algorithm 1*.

ALGORITHM 2: CDAS
Input: T: a *coherent* ontology TBox
Output: D_{cs}: a coherent disjointness assumption set with respect to T.
1. $D_s \coloneqq disjointnessOfSiblings(T)$;
2. $D_{cs} \coloneqq CAS(T, D_s)$;
3. **return** D_{cs}

Step 2: Matching terminologies (*Algorithm 3*). Currently, an initial terminology mapping is generated by using a very simple lexical matching method, i.e. stating equivalence of atomic concepts with identical names (Line 1). *Definition 6* can be extended from one ontology O to two ontologies T_1 and T_2, given that the union of two ontologies, $T_1 \cup T_2$, is an ontology. A coherent disjointness assumption set for TBoxes (union of CDAS for each TBox) and an initial terminology mapping form an initial assumption set, from which, a coherent assumption set with respect to T_1

[3] In an algorithm, lines marked with * may require manual intervention.

[4] An ontology only with a TBox.

and T_2 is calculated by applying *Algorithm 1*. An assumption in a minimal incoherent assumption set can be a disjointness axiom or a terminology correspondence axiom. Domain experts are consulted to decide which assumption(s) to retract when incoherence arises.

ALGORITHM 3: Matching Terminologies
Input T_1, T_2: *coherent* ontology TBoxes
D_{cs}: a coherent disjointness assumption set with respect to T_1, T_2
Output T_{cs}: a coherent terminology assumption set with respect to T_1, T_2

1. $M_{st} := lexicalMatching(T_1, T_2)$;
2. $T_{cs} := CAS(T_1 \cup T_2, D_{cs} \cup M_{st})$;
3. **return** T_{cs}

3.2 Matching Geospatial Individuals

An instance level mapping is a set of individual correspondences. An individual correspondence is represented in one of the following forms:

$$(i : a, j : b) \in sameAs \tag{4}$$

$$(i : a, j : b) \in partOf \tag{5}$$

where a, b denote individual names. The relation (4) states that the individual name a from the ontology i and the individual name b from the ontology j refer to the same object. The relation (5) states that the individual name a from the ontology i refers to an object which is a part of the object the individual name b from the ontology j refers to.

ALGORITHM 4: Matching Geospatial Individuals
Input A_1, A_2: ontology ABoxes
Output M_{sa} : an initial geospatial instance mapping between A_1, A_2

1. M_{sa}:={};
2. **for each** spatial individual a_1 in A_1 **do**
3. **for each** spatial individual a_2 in A_2 **do**
4. **if** *geo_poss_match*(a_1.*geometry*, a_2.*geometry*)
5. **and** *lex_poss_match*(a_1.*lexicons*, a_2.*lexicons*) **then**
6. add (a_1, a_2) ∈ *sameAs* to M_{sa};
7. **end if**
8. **end for**
9. **end for**
10. *cardinalityChecking* (*Msa*)
11. **return** *Msa*

Step 3: Matching Geospatial Individuals. *Algorithm 4* is designed to match geospatial individuals whose geometries are represented using the same coordinate reference system (CRS). The geometry of a spatial object can be represented in different accuracy levels, granularities or world views in different ontologies. In other words, for the same spatial object, the recorded geometry in ontology *i* may not be exactly the same as the recorded geometry in ontology *j*.

The *geo_poss_match* (Line 4) between two geometries returns true if the geometries are similar enough given a margin of error in representation. Currently, it requires input geometries as polygons. Two polygons are possibly matched if one of them is the smallest polygon containing the characteristic point from the other.[5]

The *lex_poss_match* (Line 5) between two lexical descriptions returns true if the lexicons (meaningful labels indicating identity) are similar enough, tolerating partial differences, for example, a full name and its abbreviation, and recognizing different names for the same location. Currently, it employs a series of basic string matching strategies, such as equivalence, inclusion and abbreviation.

For each pair of spatial individuals a_1 and a_2 from different ontologies, if their geometries are possibly matched (Line 4) and their lexicons are possibly matched (Line 5), then they can be assumed to be the same. We generate a "sameAs" relation linking them and add it to an instance mapping M_{sa} (Line 6).

It is currently assumed that, within a local ontology, a spatial individual has at most one representation. In other words, there are no "sameAs" relations within a local ontology. The cardinality checking (Line 10) revises M_{sa}, a set of "sameAs" relations, ensuring that "sameAs" is one-to-one. If not, we remove them from M_{sa}, and add corresponding "partOf" relations. For example, if M_{sa} contains (i: a, j: b) ∈ *sameAs* and (i: c, j: b) ∈ *sameAs*, we replace them with (i: a, j: b) ∈ *partOf* and (i: c, j: b) ∈ *partOf*. [6]

The geometry matching, lexical matching and cardinality checking complement each other to cope with the following possibilities. Different geospatial individuals may share the same label or the same location in an ontology. In addition, a same geospatial individual may be represented as a whole in one ontology, whilst as several parts of it in another.

---

[5] Individuals from the Ordnance Survey of Great Britain (OSGB) Buildings and Places ontology and the OpenStreetMap ontology (See Sect. 4.2) are spatially linked by finding the smallest OSM polygon containing a point from OSGB address Layer 2. See Fig. 1 for examples. Polygons containing the same red point are linked.

A more sophisticated geometry matching method for generating spatial "sameAs" and "partOf" relations is under development and evaluation.

[6] We are aware that, an individual a in one ontology O^i can be part of an individual b in another ontology O^j, even if there are no other individuals in O^i who can be part of b. This has be considered when designing our new geometry matching method.

Definition 8 (*Consistency of an Assumption Set*) An assumption set A_s is inconsistent with respect to an ontology O, if $O \cup A_s$ is inconsistent, but O is consistent. Otherwise, it is consistent with respect to an ontology O.

Definition 9 (*Minimal Inconsistent Assumption Set*) Given a set of assumption A_s, a set $C \subseteq A_s$ is a minimal inconsistent assumption set (MIA),[7] iff C is inconsistent and each $C' \subset C$ is consistent.

Similarly, a minimal inconsistent assumption set can be fixed by removing one element from it. The algorithm for calculating consistent assumption set (CAS) can be generated from *Algorithm 1*, changing coherence checking to consistency checking. *Definition 8* can be easily extended to deal with two ontologies.

ALGORITHM 5: Matching Geospatial Ontologies

Input $O_1 = (T_1, A_1)$, $O_2 = (T_2, A_2)$: coherent and consistent geospatial ontologies
T_{cs}: a coherent assumption set with respect to T_1, T_2
M_{sa}: an initial geospatial instance mapping between A_1 and A_2.

Output O_{cs}: a consistent assumption set with respect to O_1 and O_2

1. $O_{cs} := CAS(O_1 \cup O_2, T_{cs} \cup M_{sa})$[8];
2. **return** O_{cs}

Step 4: Matching Geospatial Ontologies (*Algorithm 5*). A coherent assumption set with respect to TBoxes is obtained in *Step 2*. An initial geospatial instance mapping between ABoxes is generated in *Step 3*. The union of them is an assumption set with respect to input ontologies. Applying *CAS*, if overall information is inconsistent, a set of minimal inconsistent assumption sets will be calculated, and visualized appropriately to help domain experts to repair them. These steps iterate until a consistent assumption set is found.

4 Evaluation

The method described above is implemented as a system called GeoMap. Pellet (Sirin et al. 2007) is employed for coherence and consistency checking. Minimal incoherent assumption sets are calculated from explanations for unsatisfiable classes, and minimal inconsistent assumption sets from explanations for inconsistencies.

We evaluate GeoMap using the Ordnance Survey of Great Britain (OSGB) Buildings and Places ontology (Hart et al. 2008) and the OpenStreetMap (OSM) controlled vocabularies (OpenStreetMap 2012). OSGB and OSM are

[7] MIA refers to *minimal incoherent assumption set* when matching terminologies, and refers to *minimal inconsistent assumption set* when matching instances.

[8] *CAS* here refers to the calculation of consistent assumption set.

representatives of authoritative and crowd-sourced geospatial information sources respectively. OSGB is the national topographic mapping agency of Great Britain. OSM is a collaborative project to create a free editable map of the world, relying on volunteers for data collection. Currently, OSM has not established a standard ontology, but maintains a collection of commonly used tags for main map features. An OSM feature ontology is generated automatically from the existing classification of main features. For example, given "Restaurant" is a value under the key "Amenity" in the OSM classification, we formulate this as *OSM*: *Restaurant* $\sqsubseteq$ *OSM*: *Amenity*. Both ontologies are written in the OWL 2 Web Ontology Language (W3C 2009). The OSGB Buildings and Places ontology has 692 classes and 1,230 logical axioms. There are 663 classes and 677 logical axioms in the OSM ontology. Both ontologies, containing no disjointness axioms, are coherent.

4.1 Evaluating Terminology Mapping

Evaluating Step 1. Applying *Algorithm 2*, a coherent disjointness assumption set containing 32,299 pairwise disjointness axioms is generated with respect to the OSGB Building and Places ontology. With respect to the OSM ontology, the coherent disjointness assumption set contains 9,348 pairwise disjointness axioms. A sample of 323 and a sample of 93 are taken randomly from these coherent disjointness assumption sets respectively. Based on manual evaluation, the rates of correctness are 0.951 and 0.892 respectively.

Evaluating Step 2. GeoMap, CODI (Niepert et al. 2010) and LogMap (Jimenez-Ruiz and Grau 2011) are employed to match the OSGB Buildings and Places ontology and the OSM ontology (TBoxes), given the generated coherent disjointness assumption sets. The experiments are performed on an Intel Dual Core 2.00 GHz, 3.00 GB RAM personal computer from command line. The experimental results are summarized in Table 1.

GeoMap time in Table 1 is for generating equivalence relations for same-named classes from different ontologies and checking coherence using Pellet. Total time including human interaction (choosing which assumption(s) to retract, time in average is 105.6 s) is 124.4 s. Based on manual evaluation, the precision rates of GeoMap, CODI and LogMap mappings are 89, 76 and 70 % respectively.

Table 1 GeoMap time

	GeoMap	CODI	LogMap
Time[a]	18.8 s (automatic part)	167.72 s	8.65 s
Output	84	105	91
Precision	0.89	0.76	0.70
Recall[b]	0.71	0.76	0.41

[a] Times are in seconds, averaged over 5 runs

[b] The recalls are calculated based on the ground truth shown in Table 2

Table 2 Equivalence relations provided by domain experts

Ground truth[a]	GeoMap	CODI	LogMap
OSGB: Bank ≡ OSM: Bank	1	1	1
OSGB: Chapel ≡ OSM: Chapel	1	1	0
OSGB: Church ≡ OSM: Church	1	1	0
OSGB: Fire Station ≡ OSM: Fire_Station	1	1	1
OSGB: Hotel ≡ OSM: Hotel	1	1	0
OSGB: House ≡ OSM: House	1	1	1
OSGB: Nursery School ≡ OSM: Kindergarten	0	1	0
OSGB: Library ≡ OSM: Library	1	1	1
OSGB: Market ≡ OSM: Marketplace	0	0	0
OSGB: Museum ≡ OSM: Museum	1	1	1
OSGB: Car Park ≡ OSM: Parking	0	−1	0
OSGB: Police Station ≡ OSM: Police	0	−1	−1
OSGB: Public House ≡ OSM: Pub	0	1	0
OSGB: Restaurant ≡ OSM: Restaurant	1	1	1
OSGB: Shop ≡ OSM: Shop	1	0	0.5
OSGB: Town Hall ≡ OSM: Townhall	1	1	1
OSGB: Warehouse ≡ OSM: Warehouse	1	1	0
Score	12	11	6.5

[a] The ground truth is a small set of equivalence relations provided by domain experts. As future work, we will extend the current ground truth to a larger set, to get a more realistic evaluation

The recalls are calculated based on a small set of "ground truth", i.e. equivalence relations provided by domain experts shown in Table 2. In Table 2, "1" means the mapping contains that relation, "0" means not. "−1" means the mapping contains a "wrong" relation. For example, CODI mapping contains an incorrect relation *OSGB*: *Parking*[9] ≡ *OSM*: *Parking* rather than *OSGB*: *Car Park* ≡ *OSM*: *Parking*. "0.5" means the mapping contains partially the relation. For example, LogMap mapping contains *OSGB*: *Shop* ⊑ *OSM*: *Shop* instead of *OSGB*: *Shop* ≡ *OSM*: *Shop*.

Table 1 shows that, LogMap calculates a mapping very quickly, the precision rate of GeoMap mapping is the highest, whilst the recall of CODI mapping is the highest. LogMap is designed for matching large-scale ontologies, especially in biomedical domain, in a reasonable time. As mentioned before, we assume that the TBoxes of geospatial ontologies are not very large, but contain concepts which are more ambiguous, compared to biomedical ontologies. We will focus on precision and recall at the current stage of research.

The precision of GeoMap mapping is high, since domain experts are involved to make ultimate decisions. Consider the following example. In the OSM ontology, several classes, such as *Bicycle*, *Clothes*, *Hardware* and *Kitchen*, are defined as subclasses of *Shop*, indicating what a shop sells. In this context, *OSM*: *Clothes* does not refer to clothes, but a clothes shop. Domain experts retract *OSGB*:

[9] The meanings of OSGB concepts are usually normal. *OSGB*: *Parking* ⊑ *OSGB*: *Purpose*.

Clothes[10] ≡ *OSM*: *Clothes*, whilst, CODI removes *OSGB*: *Shop* ≡ OSM: *Shop* and keeps the existing correspondences of subclasses of *Shop*, optimizing the sum of similarity-weighted probability. LogMap weakens the equivalence relation to *OSGB*: *Shop* ⊑ *OSM*: *Shop*. Though mappings calculated by GeoMap and CODI are always coherent[11] with respect to input ontologies, the results based on optimization may not be reasonable in several cases, especially when informal information exists. Domain experts do not necessarily make the same choices, and an individual domain expert may make different decisions on different occasions.

CODI produces more correct correspondences, such as *OSGB*: *Nursery School* ≡ *OSM*: *Kindergarten* and *OSGB*: *Public House* ≡ *OSM*: *Pub*, owing to its usage of more intelligent lexical matching techniques. However, CODI trades off its precision rate, since it also produces incorrect relations, like *OSGB*: *Race Horse*[12] ≡ *OSM*: *Horse_Racing*. The recall of CODI is better than that of GeoMap or LogMap, however, it is still far from covering all ground truth relations. All three fail to calculate the relation *OSGB*: *Market* ≡ *OSM*: *Marketplace* and miss relations of other types, such as inclusions[13] and overlaps. To improve the recall, more sophisticated lexical matching methods are required, and domain experts are needed, at least at the current stage of development.

The experimental results show that domain experts are indispensable when matching terminologies in order to obtain 100 % precision and recall. Mappings produced by fully automatic methods, such as CODI and LogMap, require final validation by experts, which is difficult and time-consuming. Our method reduces human effort by directing experts to make ultimate decisions during the matching process. As future work, more methods need to be developed to support the manual intervention stage, minimizing human efforts.

4.2 Evaluating Geospatial Instance Mapping

We currently require that, geospatial individuals from different ontologies have polygonal geometries, represented as two dimension vector data. The instance data for OSGB Buildings and Places ontology is extracted from the OSGB Address Layer 2 and the OSGB Topology Layer (Ordnance Survey 2012). The Address Layer 2 is a point layer, containing lexical and classification information for spatial individuals. The Topology Layer is a polygon layer, containing geometries of spatial individuals. These two layers are linked together by finding the smallest Topology Layer polygon containing a point from the Address Layer 2. The OSM instances are from the building layer (containing polygonal geometries, names and

[10] *OSGB: Clothes* refers to "garments worn over the body". It is a secondary concept.

[11] A LogMap mapping may not.

[12] *OSGB*: *Race Horse* ⊑ *OSGB*: *Animal*. It is used to define *OSGB*: *Racing Stables*. .

[13] LogMap weakens some equivalence relations to inclusions, but also does not produce enough.

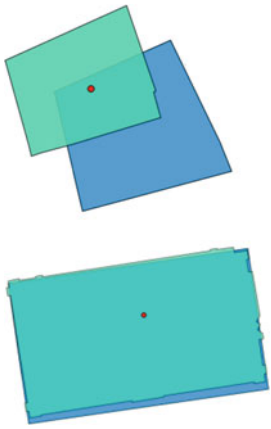

Fig. 1 Prezzo Ristorante (*Up*) and capital one

types) of OSM data, downloaded through Geofabrik (Geofabrik GmbH Karlsruhe 2012) in April, 2012. From the studied area of Nottingham city centre, 713 geospatial individuals are added to OSGB Buildings and Places ontology, 253 geospatial individuals are added to OSM ontology automatically, resulting in two consistent ontologies. Each geospatial individual is classified to a class based on its type information, and has geometry information and lexicon information as its data properties.

Evaluating Step 3 and Step 4. When matching geospatial individuals, geometry matching (*geo_poss_match*) is necessary since two different spatial objects may have the same name. For example, *OSGB*: *1000002308426350* refers to a restaurant called 'PREZZO RISTORANTE'. So does *OSGB*: *1000002309000257*. However, they are actually different restaurants which are distant from each other. Without any geometry checking, the existing data interlinking tools, for example KnoFuss (Nikolov et al. 2007), will map them to the same spatial object *OSM*: *116824670*. In *Step 3*, only *OSGB*: *1000002309000257* and *OSM*: *116824670* are linked (Fig. 1), since their polygons contain the same point from the OSGB Address Layer 2. Lexical matching (*lex_poss_match*) is necessary since two different objects may share the same location. For example, *OSGB*: *1000002308427059* refers to an *OSGB*: *Clinic*, labelled as 'N E M S PLATFORM ONE PRACTICE', while *OSGB*: *1000002308427060* refers to a general commercial company, labelled as 'CAPITAL ONE (EUROPE) PLC', in the same building. Without lexical matching, both will be mapped to *OSM*: *17505332*, labelled as 'CAPITAL ONE', based on geometry similarity (Fig. 1). Cardinality checking is necessary since the same object may be represented as a whole in one ontology, whist as several parts in the other. For example, *OSGB*: *1000002308430942* refers to a NatWest bank in the Victoria Centre. *OSGB*: *1000002308429872* refers to Millies Cookies, a bakery in the Victoria Centre. Without cardinality checking, both will be 'sameAs', rather than 'partOf', *OSM*: *16469518*, the Victoria Centre (Fig. 2).

In *Step 4*, domain experts are consulted to make decisions to repair inconsistencies. For example, *OSGB*: *1000002308476718* refers to an *OSGB*:

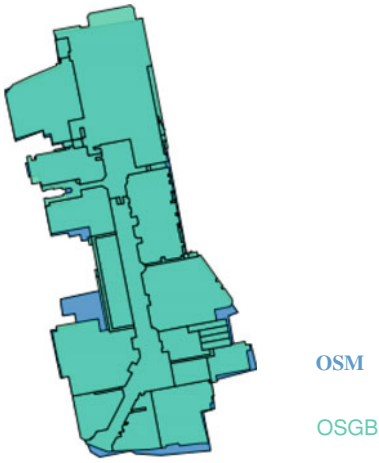

Fig. 2 Victoria centre

HealthCentre labelled as 'SNEINTON HEALTH CENTRE'. *OSM*: *62134030* refers to an *OSM*: *Clinic* labelled also as 'SNEINTON HEALTH CENTRE'. Their geometries are very similar. However, the existence of the following assumptions leads to inconsistency.

$$(OSGB : 1000002308476718, \\ OSM : 62134030) \in sameAs \tag{6}$$

$$OSGB : Clinic \equiv OSM : Clinic \tag{7}$$

$$OSGB : Clinic \sqsubseteq OSGB : HealthCentre \tag{8}$$

Domain experts are consulted to decide which assumption(s) to retract. To keep the individual correspondence (6), it is reasonable to retract (8) or weaken (7) to *OSGB*: *Clinic* $\sqsubseteq$ *OSM*: *Clinic*. This differs from all other methods, which use (8) as a premise, which is not retractable.

Based on manual evaluation, more than 95 percent of the output 139 individual correspondences (37 "sameAs" and 102 "partOf") are reasonable.

Though the initial experimental results seem promising, we are aware that there is still a long way to go before being able to apply this method into practice. Firstly, the "semantic gap" that exists between databases and their corresponding ontologies makes it difficult to populate all individuals from databases to ontologies automatically. For example, "Bar", "POBox" and "Cafe" are individual types in the OSGB database, but are not defined as concepts in the OSGB Buildings and Places ontology. Additionally, some lexical and classification information in OSM data set might be missing, in which case, only geometry matching can be applied. Furthermore, though several geometry matching and lexical matching techniques have been developed, almost none of them ensure overall correctness and consistency of results. As future work, we will explore new ways to make full use of geometry, lexical and classification information for

matching geospatial individuals, aimed for overall correctness and consistency, and minimized human effort.

5 Conclusion

In conclusion, we propose a new semi-automatic method to match disparate geospatial ontologies, guaranteeing the coherence and consistency of overall information. Differing from other existing methods, disjointness axioms and mappings are seen as assumptions, which can be retracted later if found to be too restrictive or inappropriate. A series of algorithms are designed to match disparate ontologies from terminology level to instance level by calculating a coherent and consistent assumption set with respect to them. Geometry matching, lexical matching and logical consistency checking are combined for matching geospatial individuals. The initial experiments show promising results, which indicate that, when matching geospatial ontologies, using geometry or location information helps and domain experts are indispensable. As future work, we plan to develop more sophisticated matching methods, aimed at obtaining 100 % precision and recall, and minimizing human effort.

References

Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) (2007) The description logic handbook. Cambridge University Press, Cambridge

Bouquet P (2007) Contexts and ontologies in schema matching. Context and ontology representation and reasoning. Roskilde University, Denmark

Bouquet P, Serafini L, Zanobini S (2003) Semantic coordination: a new approach and an application. International semantic web conference, pp 130–145

Buccella A, Cechich A, Fillottrani P (2009) Ontology-driven geographic information integration: a survey of current approaches. Comput Geosci 35:710–723

Domingos P, Lowd D, Kok S, Poon H, Richardson M, Singla P (2008) Just add weights: markov logic for the semantic web. Uncertainty reasoning for the semantic web I, ISWC International Workshops, URSW 2005–2007, Revised Selected and Invited Papers, 2008, pp 1–25

Du H, Anand S, Alechina N, Morley J, Hart G, Leibovici D, Jackson M, Ware M (2012) Geospatial information integration for authoritative and crowd sourced road vector data. Transactions in GIS, Blackwell Publishing Ltd, 2012, 16, 455–476

Euzenat J, Shvaiko P (2007) Ontology matching. Springer, Berlin

Geofabrik GmbH Karlsruhe: Geofabrik (2012) http://www.geofabrik.de

Giunchiglia F, Dutta B, Maltese V, Farazi F (2012) A facet-based methodology for the construction of a large-scale geospatial ontology. J Data Semant 1:57–73 Springer

Giunchiglia F, Shvaiko P, Yatskevich M (2004) S-Match: an algorithm and an implementation of semantic matching. European semantic web conference (ESWC), pp 61–75

Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquisition 5:199–220

Gurobi Optimization Inc (2012) Gurobi optimizer reference manual. http://www.gurobi.com

Hart G, Dolbear C, Kovacs K, Guy A (2008) Ordnance survey ontologies. http://www.ordnancesurvey.co.uk/oswebsite/ontology

International Organization for Standardization (2011) ISO/DIS 19157: Geographic information—Data quality

Jackson MJ, Rahemtulla H, Morley J (2010) The synergistic use of authenticated and crowd-sourced data for emergency response. The 2nd international workshop on validation of geo-information products for crisis management (VALgEO). Ispra, Italy, pp 91–99, 11–13 Oct 2010. Available online: http://globesec.jrc.ec.europa.eu/workshops/valgeo-2010/proceedings

Jain P, Hitzler P, Sheth AP, Verma K, Yeh PZ (2010) Ontology alignment for linked open data. Int Semant Web Conf 1:402–417

Jean-Mary YR, Shironoshita EP, Kabuka MR (2010) ASMOV: results for OAEI 2010. The 5th international workshop on ontology matching (OM-2010)

Jiménez-Ruiz E, Grau BC (2011) LogMap: logic-based and scalable ontology matching. Int Semant Web Conf 1:273–288

Jiménez-Ruiz E, Grau BC, Horrocks I, Llavori RB (2009) Ontology integration using mappings: towards getting the right logical consequences. The 6th european semantic web conference (ESWC), pp 173–187

Meilicke C, Stuckenschmidt H (2009) An efficient method for computing alignment diagnoses. In: Third international conference on web reasoning and rule systems, pp 182–196

Meilicke C, Stuckenschmidt H, Tamilin A (2008a) Reasoning support for mapping revision. J Logic Comput 19:807–829

Meilicke C, Völker J, Stuckenschmidt H (2008b) Learning disjointness for debugging mappings between lightweight ontologies. Proceedings of the 16th international conference on knowledge engineering: practice and patterns, Springer, Berlin, pp 93–108

Niepert M, Meilicke C, Stuckenschmidt H (2010) A probabilistic-logical framework for ontology matching. American association for artificial intelligence for ontology matching. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, Georgia, AAAI Press

Nikolov A, Uren V, Motta E (2007) KnoFuss: a comprehensive architecture for knowledge fusion. The 4th international conference on knowledge capture, ACM, NY, pp 185–186

OpenStreetMap (2012) The Free Wiki World Map. http://www.openstreetmap.org

Ordnance Survey (2012) Ordnance Survey. http://www.ordnancesurvey.co.uk/oswebsite

Qi G, Ji Q, Haase P (2009) A conflict-based operator for mapping revision: theory and implementation. In: Proceedings of the 8th *international semantic web conference*. ISWC '09, Springer, Berlin, Heidelberg, pp 521–536

Reul Q, Pan JZ (2010) KOSIMap: use of description logic reasoning to align heterogeneous ontologies. The 23rd international workshop on description logics (DL 2010)

Sais F, Pernelle N, Rousset MC (2007) L2R: A logical method for reference reconciliation. In: AAAI conference on artificial intelligence. pp 329–334

Scharffe F, Liu Y, Zhou C (2009) RDF-AI: an architecture for RDF datasets matching, fusion and interlink. IJCAI 2009 workshop on Identity, Reference and Knowledge Representation (IR-KR)

Shvaiko P, Euzenat J (2012) Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering

Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y (2007) Pellet: a practical OWL-DL reasoner. Web semantics: science, services and agents on the World Wide Web, Elsevier Science Publishers B. V., vol 5, pp 51–53

Volz S, Walter V (2004) Linking different geospatial databases by explicit relations. International society for photogrammetry and remote sensing (ISPRS) congress, communication vol IV, pp 152–157

W3C (2009) OWL 2 Web Ontology Language. http://www.w3.org/TR/owl2-overview

Wang P, Xu B (2008) Debugging ontology mappings: a static approach. Comput Artif Intell 27(1):21–36

Wolger S, Siorpaes K, Bürger T, Simperl E, Thaler S, Hofer C (2011) A survey on data interlinking methods. Semantic Technology Institute (STI) Innsbruck, University of Innsbruck. Available online: http://www.insemtives.eu/publications/A_Survey_on_Data_Interlinking_Methods.pdf