# Automated geographic information fusion and ontology alignment

Matt Duckham[1] and Mike Worboys[2]

[1] Department of Geomatics, University of Melbourne, Australia 3010
mduckham@unimelb.edu.au
[2] NCGIA and Department of Spatial Information Science and Engineering,
University of Maine, ME 04469, USA worboys@spatial.maine.edu

## 1 Introduction

*Geographic information fusion* is the process of integrating geographic information from diverse sources to produce new information with added value, reliability, or usefulness (cf. [14, 67]). Geographic information fusion is an important function of interoperable and web-based GIS. Increased reliance on distributed web-based access to geographic information is correspondingly increasing the need to efficiently and rapidly fuse geographic information from multiple sources.

The overriding problem facing any geographic information fusion system is *semantic heterogeneity*, where the concepts and categories used in different geographic information sources have incompatible meanings. Most of today's geographic information fusion techniques are fundamentally dependent on human domain expertise. This chapter examines the foundations of *automated* geographic information fusion using *inductive inference*. Inductive inference concerns reasoning from specific cases to general rules. In the context of geographic information fusion, inductive inference can be used to infer semantic relationships between categories of geographic entities (general rules) from the spatial relationships between sets of specific entities. However, inductive inference is inherently unreliable, especially in the presence of uncertainty. Consequently, managing reliability is a key hurdle facing any automated fusion system based on inductive inference, especially in the domain of geographic information where uncertainty is endemic.

This chapter develops a model of automated geographic information fusion based on inductive inference. Central to this model are techniques by which unreliable inferences and data can be accommodated. The key contributions of this chapter are to:

- define the foundations of automated geographic information fusion using inductive inference;

- explore some of the limitations of automated geographic information fusion, inherent to inductive inference; and
- indicate initial techniques to adapt the automated fusion process to operate in the presence of imperfect and uncertain geographic information.

Following a brief motivational example (section 1.1), section 2 presents a review of the relevant literature. Section 3 then sets out the foundations of inductive inference for automated geographic information fusion. The limitations resulting from the unreliability of the inductive reasoning process are set out in section 4, while section 5 addresses the management of uncertainty in the input geographic data. Finally, section 6 concludes the chapter with a look at future research.

### 1.1 Motivational example

Data on the structural characteristics of buildings is often important to decision-makers as part of an emergency response effort. It is not unusual for several different agencies to collect such data for the same geographic region and to make it available online. These agencies may use heterogeneous definitions or may even produce semistructured data without separate or fixed definitions (see chapter **??**). For example, the category "Reinforced concrete building" in spatial database $A$ may not have the same meaning as the category "Non-wooden building" in spatial database $B$ (this example is taken from a study of the 1995 Kobe earthquake [64]). Current geographic information fusion techniques rely on the generation of a manual specification of the semantic relationships between different categories by a human domain expert. Such manual techniques can be slow, unreliable, and do not scale easily to web-based information fusion scenarios.

However, if all the *instances* of buildings categorized as "Reinforced concrete building" in spatial database $A$ are categorized as "Non-wooden building" in spatial database $B$, then this provides evidence that the *category* "Reinforced concrete building" is a subcategory of "Non-wooden building." Although this example is highly simplified, it does support the central intuition behind using inductive inference for geographic information fusion: that analysis of spatial relationships can be used to infer semantic relationships. It is important to note that this inference process does not necessarily require an understanding of the meaning of "Non-wooden building" or "Reinforced concrete building," and hence can be applied in the context of automated reasoning systems.

## 2 Background

The semantics of an information source may be described using an *ontology* (defined as "an explicit specification of a conceptualization," [32]). The task of fusing information compiled using different ontologies is a classical problem in information science (e.g., [69]), and continues to be a highly active research issue within many topics, including databases [40, 42, 57]; interoperability [56, 65]; the semantic web [6, 18]; medical information systems [28, 55]; knowledge representation [10]; data warehousing [68, 73]; and, of course, geographic information fusion (section 2.3).

The term "schema" is not identical to "ontology," but the two terms are often used near-interchangeably. A schema is a formally (or otherwise precisely) defined taxonomy. Thus, a schema is an ontology in the sense of [32]. However, the term "ontology" encompasses a broad spectrum of specification methods, from schemas at one extreme through to general logical systems that can be used to define and reason about sophisticated relationships and constraints between elements within a taxonomy [47]. From this point onwards the term "ontology" is preferred because this term covers both schemas and more sophisticated types of ontologies. However, it should be noted that the ontologies in this chapter are simply schemas.

A critical step in the fusion process is to fuse the ontologies for the different information sources. A variety of closely related terms are used in the literature to refer to aspects of this task, including:

- integration and alignment;
- merging and matching; and
- transformation and mapping.

To further confuse the issue, almost all of these terms may appear in the literature combined with any one of "ontology," "schema," or "semantic" (e.g., "ontology alignment," "schema matching," and "semantic integration"). The choice of which precise terms are adopted by particular researchers is often more a matter of preference and domain than a strict difference in definitions. However, a clear distinction is usually made between process of identifying the relationships between corresponding elements in two heterogeneous ontologies (termed alignment/matching/mapping); and the process of constructing a single combined ontology based these identified relationships (termed integration/merging/transformation) [45, 51]. For consistency, in this chapter we use the terms "(ontology) integration" and "(ontology) alignment" to distinguish these two concepts.

### 2.1 Ontology integration and mediators

The concept of a *mediator*, a software system that can assist humans in integrating heterogeneous information sources, was first explicitly described by Gio Wiederhold [69]. Based on his general vision, dozens of different mediation systems have been proposed and developed over the past decade (for a full survey of mediation systems see [66]). For example, TSIMMIS was one of the earliest mediator systems to be researched. The core idea behind TSIMMIS was to mark up information sources with standardized tags, which included labels describing the semantics of each data item [12, 30]. Rather than use unstructured tagging to describe information sources, subsequent mediators, like SIMS [1], OBSERVER [48], InfoSleuth [3], and OntoSeek [34], use predefined domain ontologies as the basis for integration. More recently, a suite of web-based languages and mediation technologies have emerged around the topic of the semantic web (e.g., [31]). The primary focus of all these systems is efficient *integration* of information across multiple information sources with heterogeneous ontologies. Ontology alignment is a prerequisite for such systems to operate, but the question of how the alignment semantics are generated is not explicitly addressed by these systems. Chapter **??** contains further information on mediator systems, in particular on the distinction between local-as-view (LAV) and global-as-view (GAV) approaches to mediation.

The theoretical foundations of most mediator systems are similarly focused more on ontology integration than ontology alignment (e.g., [33, 62]). Formal concept analysis (FCA, [29]) is a widely used technique, dating back more than 20 years, for structuring and integrating ontologies based on *concept lattices* (a special type of ordering relation on categories). Although the integration of concept lattices has a precise formal definition and can be automated, human domain expertise is required to identify salient attributes and categories within a domain and their inter-relationships. Another widely used technique for representing and reasoning about heterogeneous ontologies is description logics. Description logics are decidable, tractable fragments of first order predicate calculus and form the basis of many mediator systems and components, including SIMS, OBSERVER, and OWL (Web ontology language, one of the primary standards for the semantic web). Description logics are especially useful in the context of ontology integration because they provide several important reasoning services, most notably a *subsumption* service which can classify the relationships between categories and derive a complete and consistent integrated ontology (see [11, 19] for more information on description logics and their reasoning services). However, while description logics can offer efficient formally-founded reasoning about ontologies, defining the relationships and rules that connect elements of different ontologies remains a primarily human activity.

In summary, much of the existing research into mediators and ontology integration does not address the issue of ontology alignment directly. Instead, such research typically assumes that the semantic relationships between ontological elements will be established using user interaction, predefined mappings, pre-existing top-level ontologies, or existing lexical correspondences [66]. Building these mappings is assumed to require an understanding of the underlying concepts, and so is at root a human activity.

### 2.2 Automating ontology alignment

Some researchers have turned their attention to creating semi- or fully-automated ontology alignment systems (see [54] for an overview). Much of this research adopts an *intensional* approach: it aims to analyze the definitions (intensions) of the concepts and categories used in the input information sources. Intensional techniques usually analyze heterogeneous ontologies to identify lexical similarities (e.g., PROMPT [52], Active Atlas [61]), structural similarities (e.g., DIKE [53], ONION [50]), or some combination of these (e.g., CUPID [45], FCA-Merge [60]).

There are two main drawbacks of adopting a purely intensional approach to ontology alignment. First, how concepts are defined is not necessarily the same as how they are used. As an analogy, people who learn a to speak language from a dictionary (definitions) often have very different speech patterns from native speakers, who also learn from example. Only by looking at *extensional* information (specific instances in data) is it possible to begin to determine how concepts are actually used. Second, extensional information forms a rich source of examples that can be used as the basis for automated pattern recognition techniques.

Recognizing the importance of instance-level information, an increasing number of researchers have turned to extensional approaches, including:

- The SemInt system clusters patterns in instance-level information, and uses these clusters to train a neural network to identify intensional relationships [43, 44].
- Doan and collaborators [17, 18] and the Autoplex system [5] use Bayesian machine learning techniques on instance-level information to identify intensional relationships.
- The Clio [49] and iMAP [16] systems search for filters that relate sets of instance-level information within a database. These filters are then used to infer intensional relationships.
- He and Chang make use of patterns of co-occurrence of related attributes for Web pages [36]. The positive correlation between related attributes, along with an expected negative correlation between synonyms, is used to automatically infer semantic mappings between attributes within a domain.

Fundamentally, all these extensional approaches apply different forms of inductive inference: they use the structure and patterns in instance-level information to infer semantic relationships. An inherent limitation of using inductive inference is that it is unreliable. In many of the extensional approaches outlined above unreliability is combated using probabilistic techniques (such as Bayesian probability). We return to the topic of reasoning reliability in section 4.

### 2.3 Geographic information fusion

Research into geographic information fusion mirrors the more general approaches to information fusion cited above. Fonseca and coauthors have published a series of papers on so-called *ontology-driven GIS* [25–27]. This work aims to augment conventional GIS with formal representations of geographic ontologies, leading to tools that enable improved ontology-based information integration. A wide variety of related work has addressed the issue of geographic information integration from a similar perspective (e.g., [2,7,15,58,63]). In common with the research presented in section 2.1, such research focuses on the integration itself, but assumes the semantic relationships between different ontologies are already known.

A relatively small amount of work has begun to provide tools for geographic ontology alignment. Most of this work adopts an intensional approach. Kavouras, Kokla, and coauthors use FCA as the basis for their approach to geographic ontology alignment [37,38,41]. Manoah et al. apply the intensional machine learning techniques discussed in section 2.2 to geographic data [46]. Duckham and Worboys have investigated using description logics [20] and a formal algebraic approach [71] to ontology alignment. Because of the diversity of geographic terms and concepts, this work is at best semi-automated, and still require human domain experts at critical stages in the alignment process.

To our knowledge, [22] is the only research in the geographic domain that adopts an extensional approach to automating information fusion (although it is not the only work to acknowledge the importance of extensions in the representation of geographic knowledge, e.g., [8, 9]). Geographic information is a richly structured and voluminous source of instances upon which to base inductive reasoning processes, more so than many other types of information source. In this respect it is well suited to extensional approached to automated information fusion. However, the problems of unreliable inference introduced in section 2.2 are exacerbated in the geographic domain because uncertainty is an endemic feature of geographic information. Applying an unreliable reasoning process to uncertain data could potentially produce information that is degraded to the point of being meaningless. Consequently, following a closer look at using induction as a basis for automated geographic information fusion in section 3, we turn to the issues of unreliability in the reasoning process (section 4) and reasoning under uncertainty (section 5).

## 3 ROSETTA: Automated extensional geographic information fusion

At the core of an extensional approach to automated geographic information fusion is the process of inferring semantic relationships from spatial relationships. As already discussed, this process is an example of inductive inference: reasoning from specific cases to general rules. As an analogy, archaeologists were able to decipher the meaning of ancient Egyptian hieroglyphs following the discovery of the Rosetta Stone, a 2nd century tablet that contained the same official decree in both Egyptian (hieroglyphs and text) and Greek (text). Before the discovery and subsequent analysis of the Rosetta Stone, all attempts to decipher hieroglyphs were unsuccessful and Egyptian hieroglyphics were considered to be merely primitive picture writing. Only by comparing examples (extensions) of the Greek text with Egyptian text and hieroglyphs on the Rosetta Stone were archaeologists able to correctly infer a "dictionary" (intensions) for translation between these different information sources. In a similar way, the extensional approach to geographic information fusion constructs a shared "dictionary" for translating between the ontologies of the different information sources, based on the relationship between the spatial extents of the categories used in those information sources. In the remainder of this chapter we use the term "ROSETTA" to refer to the extensional approach to automating geographic information fusion.

To illustrate, Figure 1 contains a much-simplified example of a ROSETTA-based fusion. In Figure 1, each data set comprises an extensional component (the mapped spatial data) and an intensional component (the ontology for that spatial data). On the left-hand side of Figure 1, the intension for data set $A$ contains the categories Forest and Built-up area, while the extension contains two regions, one of each category. Similarly, on the right-hand side of Figure 1, data set $B$ contains the intensions Woodland and Urban along with a map of the spatial extensions of the Woodland and Urban categories. The geographic extents of data sets $A$ and $B$ are identical (i.e., the data sets cover the same area). Thus, Figure 1 might represent the situation where two different environmental agencies have both mapped the same geographic region using different ontologies.

The fused data set is contained within the center of Figure 1. Because all locations that are categorized as Built-up area in data set $A$ are categorized as Urban in data set $B$, we have inferred in our integrated taxonomy that the category Built-up area is a sub-category of the category Urban. Similarly, because all locations that are categorized as Woodland in data set B are categorized as Forest in data set $A$, Forest subsumes Woodland in the integrated ontology. A new category Forest & Urban has been created to represent those regions that are categorized as Forest in data set $A$ and Urban in data set $B$. In other words, although there exists no subsumption relationship between Forest and Urban, we have inferred that these categories overlap, on the grounds that their extensions overlap. Note that although highly simplified, the process illustrated by Figure 1 is more than a simple overlay. The data sets have been fused, in the sense that we have gained (a small amount of) new information about the relationships between the categories represented in each of the input data sets.

**Fig. 1.** Simplified example of inductive reasoning for automated geographic information fusion

### 3.1 Computational approaches

Although the discussion above provides an informal description of a ROSETTA system, recent work by the authors does provide a formal basis for such as system [22]. In this section we provide a brief overview and synthesis of some the central ideas and results of this work.

#### Extensional form

A key concept in the development of a ROSETTA system is to consider categories in geographic ontologies in their extensional form. The extensional form of a category is the set of all instances of that category. For example, one way to describe what is meant by the category "Car" is to refer to the set of all objects that we call cars. Subcategories, such as "Tan-colored Chevrolet Lumina," will contain only a subset of those objects. Using the extensional form of a category makes explicit the link between extensional and intensional information, enabling an automated computational system to manipulate categories without any requirement to understand the semantics of that category.

#### Reasoning system

An initially attractive route to realizing a ROSETTA system, such as described informally above, is to formalize the rules required for the inductive inference process, and then implement those rules within an automated reasoning system. We might start by representing a taxonomy as a partially ordered set $(C, \leq)$, where $C$ is a set of categories and $\leq$ is the ordering (subsumption relationships) on those categories. Now, a *geographic data set* can be represented as a set $S$ that is a partition of a region of space; a *taxonomy* $(C, \leq)$; and a *function* $e : C \rightarrow 2^S$ that defines which spatial regions are labeled with which categories ($2^S$ is the power set of $S$). Thus, $e$ associates each category in the taxonomy with a unique set of elements from the partition of space $S$. We call $e$ an extension function because it provides the extensional form of each category within the context of its data set.

To illustrate, for data set $A$ in Figure 1 the taxonomy $(C_A, \leq_A)$ is represented by hierarchy of categories; the partition of space $S_A$ is represented by the map itself, comprised of jointly exhaustive and pairwise disjoint regions; and the extension function $e_A$ is represented by the labels on both the taxonomy and the map (i.e., for each category we can identify on the map the set of locations that are labeled as that category).

From this basis, it is possible to start to define simple first-order logical rules that embody our inductive inference process. For two data sets $G_1 = \langle S_1, (C_1, \leq_1), e_1 \rangle$ and $G_2 = \langle S_2, (C_2, \leq_2), e_2 \rangle$ we wish to construct the fused data set $G_f = \langle S_f, (C_f, \leq_f), e_f \rangle$. We might specify as a first rule:

$$\text{for all } x \in C_1 \text{ and } y \in C_2 \begin{cases} \text{if } e_1(x) \subseteq e_2(y) & \text{then } x \leq_f y \\ \text{if } e_2(y) \subseteq e_1(x) & \text{then } y \leq_f x \end{cases}$$

In other words, where the spatial extent of a category $a$ contains the spatial extent of a category $b$, we infer that $a$ is a subcategory of $b$ in our fused taxonomy. Similarly, we could formulate further rules dealing with more of the possibilities for spatial relationships between the extensional forms of two categories, such as:

$$\text{for all } x \in C_1 \text{ and } y \in C_2 \text{ if } e_1(x) \cap e_2(y) = \varnothing \text{ then } x \not\leq_f y \text{ and } y \not\leq_f x$$

In plain language, the rule above states that if the extensions of two categories $x$ and $y$ are disjoint then we infer that the categories themselves are incomparable. A further obvious rule, suggested by the category Forest & Urban in figure 1, is to create a new category corresponding to two overlapping category extents as follows:

$$\begin{aligned} \text{for all } x \in C_1 \text{ and } y \in C_2 \text{ if } & e_1(x) \cap e_2(y) \neq \varnothing \\ & \text{and } e_1(x) \not\subseteq e_2(y) \text{ and } e_2(x) \not\subseteq e_1(y) \\ & \text{then } x \cap y \in C_f \text{ and } x \cap y \leq_f x \text{ and } x \cap y \leq_f y \end{aligned}$$

The rule above creates a new category $x \cap y$ in the fused taxonomy that lies at the intersection of categories $x$ and $y$. For two data sets $G_1$ and $G_2$, the conclusions from such rules form an ordering relation that relates categories in the two source taxonomies. Together with those source ordering relations, this enables the derivation of a new fused partial order in $G_f$ that defines the subsumption relationships between categories within the different taxonomies of $G_1$ and $G_2$.

Once formalized, these rules can be implemented within an automated reasoning system. Indeed, early versions of our ROSETTA system adopted this approach, using the RACER description logic engine [35] for automated reasoning. An advantage of using description logics for this purpose is that any inconsistencies between the chosen rules can be automatically detected, using the consistency and satisfiability services provided by any description logic. Having generated the ontology alignment, the spatial data itself can then be automatically fused based on the standard geographic information integration techniques (i.e., overlay the two spatial data sets, and assign to each fused region the category in the fused partial order that lies at the intersection of the two source categories for the fused region).

### Algebraic system

The reasoning system approach described above provides an important step on the road to practical automated geographic information fusion systems. However, it has at least two important shortcomings.

First, a partial order is a rather too general a structure for describing a geographic ontology. For each pair of input categories we need to be able to identify a unique category in our fused taxonomy that corresponds to the fusion of those input categories. Using partial orders, it may not be possible to guarantee that such a unique fused category exists, since a pair of elements in a partial order may have multiple incomparable least upper and greatest lower bounds. A more appropriate structure is a lattice, which as we have already seen is commonly used in formal approaches to ontological information [29, 37, 60]. A lattice is a special type of partial order, where all subsets of elements have a unique least upper bound and a unique greatest lower bound in the lattice. The simplified taxonomies in Figures 1 and subsequent figures can be represented as lattices[1].

Second, developing ad hoc fusion rules, such as illustrated above, may not always lead to an associative and commutative fusion system. Thus, we could add further rules to our reasoning system that would result in different fusion products, depending on what order we input data into the system. This is clearly undesirable. To be well-formed we would expect a fusion process to produce a unique fusion product for a set of inputs irrespective of the order in which they are fused. As a parallel, GIS would be considerably less useful if the overlay operator were defined in such a way that the order in which source data sets were overlaid affected the output results of the overlay operation.

An important result of [22] is to formalize geographic information fusion in such a way that:

1. the taxonomy associated with a data set can be represented as a lattice;
2. the fusion process is represented as an associative and commutative binary operator; and
3. the fusion process is closed, in the sense that the fusion product is itself a valid geographic data set that can be used in subsequent fusion operations.

Formally, [22] shows that the geographic data sets (represented as a partition of space, a lattice, and an extension function) combined with the fusion operator form a fusion algebra with the properties of a commutative semigroup (closed, associative, commutative). The reader is referred to [22] for more detail on this topic; the remainder of this chapter turns to issues of reliability and uncertainty rather than formalization of fusion systems.

## 4 Reliability

The ROSETTA system outlined above is simple, effective, and has a clear theoretical basis. However, in developing practical automated geographic information fusion systems, there are two main issues that must be addressed: unreliability and uncertainty in the fusion process. In this section we first examine the issue of the unreliability of inductive inference.

[1] Strictly, the taxonomies in the figures in this chapter are shown as join semi-lattices, but any finite join semi-lattice can be trivially transformed into a lattice with the addition of a bottom element.

### 4.1 Deductive validity

An inherent limitation of the extensional approach to geographic information fusion is that inductive inference is not deductively valid. In general, an inference is said to be deductively valid if, given that all the premises are true, then the conclusion must be true also. Using inductive inference, it is entirely possible to formulate deductively invalid inferences. For example, given the premise that all the birds I have ever seen can fly, I might inductively infer the conclusion that all birds can fly. Clearly, this conclusion is not necessarily valid, even though the premise may be. A similar problem can occur with a ROSETTA system. In the example in Figure 1, it might be that if we had used data sets with greater spatial extents, we would have discovered a region of built-up area in data set $A$ that overlapped a region of Woodland in data set $B$. In this case, the inductive inference procedure would have different premises, leading to a different fused data set and ontology. Figure 2 illustrates this situation. Note also that the resulting data set no longer contains new information about the semantic relationships between the different categories. In this case the "fusion" has degraded to a simple overlay. We return to this issue later on in this chapter (section 5.4).

**Fig. 2.** Unrepresentative spatial extents (dotted line) may lead to invalid inferences

The primary guard against deductive invalidity is to ensure that the data sets to be fused are large enough to contain a representative range of the possible spatial relationships between the different categories represented in the data sets. Thus, a feature of ROSETTA systems is that they are "data-hungry," in the sense that we expect the fusion process to become more reliable the more data we can feed into the process. Small fragments of data sets will tend to yield integrated ontologies that embody chance, rather than real semantic, relationships.

By way of analogy, when the Rosetta Stone was discovered, almost half the text on the artifact was damaged in some way (even missing in the case of hieroglyphs). More extensive damage would have further reduced the availability of corresponding words upon which to base lexicographic inferences. With fewer examples of correspondences between the different languages, any process of deciphering would be more likely to lead to incorrect inferences.

### 4.2 Semantic and spatial extents

An underlying assumption of the extensional approach to geographic information fusion is that the thematic domains for the input data sets are semantically related. In our example in Figure 1, both input data sets concerned land cover. Similarly, in earlier examples, we considered the fusion of data sets that concerned the structural characteristics of buildings. Using a spatial metaphor, we can say that for information fusion to take place we expect the semantic extents of two information sources to overlap.

Returning our analogy, it was only because the Rosetta Stone contained three copies of the same decree in different languages that the attempt to derive a meaningful Egyptian-Greek dictionary was successful. The direct correspondence was known about because it is explicitly stated in the Greek version of the text. If, instead, the different versions of the text on the Rosetta Stone had contained different decrees, then the Stone's usefulness as an aid to understanding hieroglyphs would have been severely limited.

In the context of geographic information fusion, there may still be some benefit to applying automated inductive inference to data sets that are topically unrelated. Although the results of such a process would not constitute geographic information fusion according to the original definition of the term, the process may be useful as a data mining technique for discovering relationships between semantically unrelated information sources. For example, Figure 3 illustrates the fusion process applied to semantically unrelated land cover and socioeconomic data sets. The relationships generated between categories in the input data sets are not subsumption relationships (it would not be true to say that Woodland is a subcategory of Low income), but might provide useful as summarizations of the semantic relationships embedded in the data set.

**Fig. 3.** "Fusion" of semantically unrelated data sets

It may also be important to consider the spatial extents of the information sources. In the simple automated information fusion systems discussed in this chapter, the inference process is driven by direct spatial coincidence. Thus, only those locations that are represented in both information sources to be fused provide premises for the inductive inference process. However, current research is also investigating the possibility of using other types of spatial relationships, such as proximity or topology, to drive inductive inferences about spatial data that is not necessarily coincident.

## 5 Uncertainty

Geographic information is inherently imperfect, leading to uncertainty about the real features represented in a geographic data set. Imperfection is often represented and quantified using spatial data quality elements and standards (chapter **??**). However, there are many different spatial data quality elements that have been proposed in standards and the research literature. Three fundamental types of imperfection are commonly identified in the literature: *inaccuracy*, *imprecision*, and *vagueness* [21,70,72]. In this section we look at the effects of each of these types of imperfection in turn, followed by an overview of ongoing research into ways to regulate uncertainty in a ROSETTA system.

### 5.1 Inaccuracy

Inaccuracy in geographic information concerns a lack of correspondence between information and the actual state of affairs in the physical world. In a ROSETTA system, inaccuracy degrades the reliability of the inductive inference process, potentially leading to semantic relationships being inferred between categories that are, in reality, unrelated. Conversely, inaccuracy may lead to a failure to identify semantic relationships between categories that are, in reality, related. For example, suppose that in our land cover data set $B$ part of the Urban region has been misclassified as Woodland such that it overlaps the Built-up area in data set $A$. In turn, this might lead to the incorrect inference that Woodland and Built-up area are semantically overlapping (Figure 4). Note that the inaccuracy has again produced a fused ontology that is not particularly informative, in the sense that we have gained no new information about the relationships between the categories in the input data sets (we could have achieved the same results using a simple overlay).

**Fig. 4.** Inaccuracy in input data sets (black region indicates sliver polygon)

We can imagine what might have happened if some of the words on the Rosetta Stone had been incorrectly drafted or inscribed. It is possible that such inaccuracies would lead to incorrect lexicographic inferences, especially in the case of systematic inaccuracies. To guard against inaccuracy, it is important is to ensure the extensions used in the inference process are large enough such that examples of incorrect correspondences due to random inaccuracies will be greatly underrepresented when compared with examples of by correct correspondences. In terms of a ROSETTA system, the situation is a little more complex. However, in principle the possibility of random inaccuracies is another reason why the ROSETTA systems are fundamentally data hungry: the more examples used in the inference process, the more likely it is that these examples will provide a basis for valid inferences.

In addition to spatial inaccuracy, inaccuracies may occasionally occur within the taxonomy itself (e.g., where one category is incorrectly labeled or incorrectly positioned within the taxonomy). Since the taxonomy is central to the fusion process, it is difficult to see how the automated fusion process described here (or indeed any of the fusion systems encountered in this chapter) could hope to effective combat such inaccuracies.

### 5.2 Imprecision

Imprecision, a lack of detail in information, is another intrinsic feature of geographic information. Imprecision leads to *granularity*: the existence of “clumps” or “grains” in the data. The granularity at which geographic phenomena are represented strongly influences what features are observed. Like inaccuracy, heterogeneous levels of granularity degrade the reliability of the inductive inference process. For example, imagine that land cover data set $A$ has been collected at a coarser level of spatial granularity than data set $B$. Then it will be likely that the detailed features found in data set $B$ will simply not be represented in data set $A$ (such as small pockets of Woodland within the predominately Urban area that are represented in data set $B$, but have no correspondent in data set $A$). As a result, a naïve inductive inference process may again incorrectly infer that Woodland and Built-up area are semantically overlapping, as in Figure 5 (similar to the effects of inaccuracy above in Figure 4). As for inaccuracy, the fusion product in Figure 5 is not particularly informative, as it is essentially a simple overlay of the data.

**Fig. 5.** Granularity in input data sets (black regions in fused spatial data shows fine grained “pockets” of Woodland)

It is difficult to say how the efforts to decipher Egyptian hieroglyphics would have fared if the different versions of the official decree on the Rosetta Stone contained different levels of detail about the official declarations. The structure of natural language does not make it easy to automatically infer relationships between texts at different levels of detail. However, the spatial structure of geographic information does make inferences between information sources at different levels of detail more feasible (e.g., section 5.4 and [20]).

In addition to spatial imprecision, it may also be important to also consider the possibility of heterogeneity in taxonomic granularity. In this case, semantic differences that are distinguished apart in the taxonomy for one data set may not be distinguished in the taxonomy for a different data set. For example, the category Woodland is at a coarser level of semantic granularity than the category Broadleaved native woodland. In general, the inductive inference process is able to operate satisfactorily in the presence of taxonomic imprecision: after all granularity is an integral feature of the hierarchical structure of taxonomies themselves.

Nevertheless, geographic information sources are especially interesting in this respect as they often exhibit *contravariant* granularity, where an information source is at a relatively fine spatial granularity but relatively coarse taxonomic granularity when compared with another information source. This situation may occur as a result of the economies of scale for spatial data capture. The high cost of performing large scale spatially detailed data capture tends to ensure that such data is collected in a general purpose form (taxonomically coarse granularity), so as to maximize its utility to the widest possible range of potential uses. Conversely, limited resources mean that spatial data collected for specific application domains (taxonomically fine granularity) tends to be at a spatially coarse granularity. An example of data sets at contravariant granularities is the topographic data collected at by the UK national mapping agency, Ordnance Survey, when compared with the CORINE land cover data set for the UK. Ordnance Survey topographic data is at a much higher spatially granularity that the CORINE data set, being derived from ground survey rather than satellite imagery. Conversely, the CORINE data set is at a much higher taxonomic granularity than Ordnance Survey topographic data, providing more detailed information about the actual land cover categories present at a particular location [20].

### 5.3 Vagueness

Vagueness concerns the existence of borderline cases in information. For example, the category "mountain" is vague, because for any particular mountain we expect there to exist locations which are definitely on the mountain, locations that are definitely not on the mountain, and locations for which is it indeterminate whether on not they are on the mountain. Unlike imprecision and inaccuracy, which may occur independently in both extensional and intensional aspects of the data, vagueness is directly associated with the intensional aspects of the data. In other words, we regard vagueness as a type of imperfection in definition, rather than imperfection in observation (i.e., we adopt an epistemic view of vagueness, leaving to one side for the moment debates about ontic vagueness [39]).

Although vagueness is an intensional phenomenon, vagueness can have an extensional expression in spatial data sets, which typically impose precise spatial boundaries around spatial regions. If, as is often the case in spatial data, the underlying categories are vague (such as the categories "Mountain" or "Forest" [4, 24]) then the actual boundaries imposed will be somewhat arbitrary. The effect of such boundary arbitrariness on a ROSETTA system will be similar to those resulting from inaccuracy: it will degrade the reliability of the inductive inference process, potentially leading to errors of omission and commission in identifying semantic relationships between categories represented in the source data sets.

In order to tackle vagueness, it is first necessary to provide an explicit representation of the existence of vagueness. Typically, this is done by replacing the crisp boundaries for regions used in conventional spatial data with a representation of regions with broad boundaries, such as fuzzy sets [23], rough sets [21], two-stage sets [59], or egg-yolk representations [13]. For example, Figure 6 shows a hypothetical fusion of data sets $A$ and $B$, containing broad boundaries between the regions Built-up area and Forest in data set $A$, and Urban and Woodland in data set $B$.

**Fig. 6.** Fusion of information sources including regions with broad boundaries

The question of exactly how such a fusion operator should be constructed is the topic of current research (hence, unlike previous figures, Figure 6 is a hypothetical fusion product). The structure of the data in Figure 6 is incompatible with the formal structures discussed so far. Either the extensions in Figure 6 contain regions that have no corresponding intensions in the taxonomy (i.e., the unlabeled broad boundaries are themselves separate regions); or from another perspective the extensions do not form a partition of space (i.e., the broad boundaries constitute an overlap between two or more neighboring regions). Thus, the formal mechanisms currently being developed for fusing data containing regions with broad boundaries are generalizations of those formalizations already discussed.

Whatever the formal structures used, the goal is to infer crisp semantic relationships between vague categories based on indeterminate spatial extents. For example, we may be certain that a “Copse” is a sub-category of “Woodland,” even if both categories are vague. In the case of Figure 6, we might devise new inference rules like, those in section 3, that only consider the *core* of the extent of each category (those parts of space that are classified as definitely belonging to the category). Conversely, a weaker inference systems could be developed by allowing semantic relationships to be inferred where the core of one category is is contained within the entirety of another category.

### 5.4 Computation with uncertain data

From the discussion above, we can begin to suggest simple mechanisms for incorporating inaccuracy and imprecision into the automated information fusion process (vagueness is the topic of current research). One such mechanism for incorporating inaccuracy and granularity arises from noting that sliver polygons (resulting from inaccuracy) or regions of fine-grained detail (resulting from fine granularity) are expected to make up a relatively small proportion of the entire regions being fused. For example, if the overlap between two regions is smaller than 5% of the total area of either regions, this might constitute evidence that the overlap arises from inaccuracy in the input regions. Similarly, if the overlap between region $A$ and region $B$ is less than 5% of the total area of region $A$ and more than, say, 95% of the total area of region $B$, this might constitute evidence that that the overlap arises from heterogeneous granularity in the data sets (i.e., that region $B$ is at a finer granularity than region $A$).

Consequently, setting thresholds for the proportion of overlap between two extensions of a category provides a basis for detecting spatial relationships that can be attributed to inaccuracy or heterogeneous granularity. Spatial relationships that are attributed to inaccuracy or imprecision then can be omitted from premises for the inductive inference process. The effect of using such an approach is illustrated for our example ROSETTA system in Figure 7, based on Figure 4. Here, the small sliver overlap between the extents of Built-up area and Woodland comprises less than 5% of the total area of these extents. This overlap is omitted from the inductive inference process, leading to a fused taxonomy as for Figure 1. However, in the fused data set, the omitted region (black region) then becomes *unclassifiable* (has no category associated with it).

**Fig. 7.** Sliver polygon, resulting from inaccuracy as in Figure 4, is eliminated from inductive inference process using overlap thresholds

The thresholds can be set arbitrarily, or by a human user. As the thresholds increase, more overlaps are omitted from the inference process, usually leading to more direct subsumption relationships in the fused taxonomy (cf. the taxonomies in Figures 4 and 7). Fused taxonomies containing more direct subsumption relationships are generally more desirable, because they provide more *new* information about the relationships between categories in the source taxonomies (the fused taxonomies in Figures 2, 4, and 5 are degenerate cases that contain no new information that could not have been derived from a simple overlay of the two data sets). Thus, in setting such thresholds, there is a balance to be struck between the quality of extensional and intensional information in the fused data set. Tolerating higher levels of inaccuracy or imprecision generally leads to more useful intensional information, but at the same time lower quality extensional information with more unclassifiable regions. Conversely, tolerating lower levels of inaccuracy or imprecision leads to less useful intensional information, but higher quality extensional information with fewer unclassifiable regions. Current research is investigating techniques for automatically setting the thresholds in such a way as to maximize some overall measure of the usefulness of the fused intensions (e.g., measures of the information content of the fused taxonomy) or quality of the fused extensions (e.g., measures of the area of unclassifiable regions).

## 6 Discussion and conclusions

This chapter has provided the conceptual basis for an extensional approach to automated geographic information fusion. The key innovation in this approach is to infer semantic relationships between those data sets based on their spatial relationships. This process is an example of inductive inference, reasoning from specific cases to general rules. The main obstacles to using inductive inference for automated geographic information fusion are the unreliability of inductive inference and imperfection in both extensional and intensional information. However, this chapter argues that these obstacles are surmountable, and indicates some of the ways they may be overcome.

The approach holds considerable promise for application to web-based environments. The increasing availability of geographic information from web-based sources is only of limited use unless it is accompanied by concomitant ability to combine those information sources in a meaningful way. Non-expert users cannot be expected to do this unaided, so automation is an essential step in extending the usability of web-based GIS into a range of new applications and domains.

However, there are several research issues to be addressed before practical automated geographic information fusion systems become a reality, including.

- *Inclusion of human expert domain knowledge*: Although the ROSETTA approach aims to enable fully automated information fusion, it is also important to allow the inclusion of partial human expert domain knowledge where it already exists, and integrate this knowledge with automatically inferred knowledge. Some initial techniques for dealing with this issue are presented in [22].
- *Integration with existing mediator architectures*: The extensive work on existing mediator architectures and ontology-based GIS (cf. chapter **??**) is complementary to the goals of a ROSETTA system. Future work aims to integrate both in an "intelligent geomediator architecture," which provides the integration capabilities of a mediator with the alignment capabilities of a ROSETTA system.
- *Regions with broad boundaries*: A high-priority goal of current research is to extend the existing formal ROSETTA systems with the ability to operate with vague categories, where the extents of those categories have broad boundaries.
- *Automated thresholding*: In addition to developing new techniques for dealing with imperfection, current research is investigating developing automated thresholds for reasoning in the presence of inaccuracy and imperfection, as discussed in section 5.4.
- *Further spatial relationships*: The inferences discussed in this chapter all concern containment or overlap between extensions of categories. However, given the rich variety of spatial relationships embedded within spatial data, it is expected that many more types of spatial relationships might be useful as a basis for inductive inference, including topological and metric relationships (cf. chapter **??**).
- *Spatially varying alignment*: The approach presented in this chapter aims to infer alignments that are non-spatial, in that they hold for all locations in space. Developing ROSETTA systems that can infer spatially varying ontology alignments (i.e., semantic relationships that hold only in specific regions of geographic space) will potentially provide much greater flexibility in defining future fusion systems.

## Acknowledgments

Matt Duckham is supported by the Australian Research Council under ARC Discovery Grant DP0662906, entitled "Automatic fusion of geoinformation: The intelligent geomediator architecture (iGMA)." Collaboration between Matt Duckham and Mike Worboys has been partially supported by funding from the Australian Academy of Science (AAS) and a University of Melbourne ICR (international collaborative research) grant. Mike Worboys work is supported by the National Science Foundation under NSF grant numbers IIS-0429644 and BCS-0327615. A preliminary version of this chapter was presented to the GISPLANET 2005 conference, Lisbon, Portugal.

## References

1. Arens Y, Knoblock C, Shen W.-M (1996) Query reformulation for dynamic information integration. *Journal of Intelligent Information Systems*, 6, 2–3, 99–130.
2. Baru C, Gupta A, Ludäscher B, Marciano R, Papakonstantinou Y, Velikhov P, Chu V (1999) XML-based information mediation with MIX. In *SIGMOD '99: Proc. ACM SIGMOD*, ACM Press, 597–599.
3. Bayardo R, Bohrer W, Brice R, Cichocki A, Fowler J, Helal A, Kashyap V, Ksiezyk T, Martin G, Nodine M, Rashid M, Rusinkiewicz M, Shea R, Unnikrishnan C, Unruh A, Woelk D (1997) InfoSleuth: Agent-based semantic integration of information in open and dynamic environments. In *Proc. 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD '97)*, ACM Press, 195–206.
4. Bennett B (2001) What is a forest? On the vagueness of certain geographic concepts. *Topoi*, 20, 2, 189–201.
5. Berlin J, Motro A (2001) Autoplex: Automated discovery of contents for virtual databases. In *Proceedings of COOPIS 2001, Sixth IFCIS International Conference on Cooperative Information Systems*, volume 2172 of *Lecture Notes in Computer Science*, Springer, 108–122.
6. Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. *Scientific American*, 279, 5, 34–43.
7. Boucelma O, Garinet J.-Y, Lacroix Z (2003) The virGIS WFS-based spatial mediation system. In *CIKM '03: Proc. twelfth international Conference on Information and Knowledge Management*, ACM Press, 370–374.
8. Brodaric B, Gahegan M (2001) Learning geoscience categories in situ: Implications for geographic knowledge representation. In *GIS '01: Proc. 9th ACM international symposium on Advances in Geographic Information Systems*, ACM Press, 130–135.
9. Brodaric B, Gahegan M (2002) Distinguishing Instances and Evidence of Geographical Concepts for Geospatial Database Design. In *Geographical Information Science*, Egenhofer M. J, Mark D. M (eds), volume 2478 of *Lecture Notes in Computer Science*, Springer.
10. Calvanese D, De Giacomo G, Lenzerini M, Nardi D, Rosati R (1998) Description logic framework for information integration. In *Proceedings 6th International*

*Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, 2–13.

11. Calvanese D, De Giacomo G, Nardi D, Lenzerini M (2001) Reasoning in expressive description logics. In *Handbook of Automated Reasoning*, Robinson A, Voronkov A (eds), volume 2, Elsevier Science, Amsterdam, chapter 23, 1581–1634.
12. Chawathe S, Garcia-Molina H, Hammer J, Ireland K, Papakonstantinou Y, Ullman U, Widom J (1994) The TSIMMIS project: Integration of heterogeneous information sources. In *Proc. 10th Meeting of the Information Processing Society of Japan*, 7–18.
13. Cohn A, Gotts N (1996) The "egg-yolk" representation of regions with indeterminate boundaries. In *Geographic Objects with Indeterminate Boundaries*, Burrough P, Frank A (eds), Taylor and Francis, 171–188.
14. Dasarathy B (2001) Information fusion—what, where, why, when, and how?. *Information Fusion*, 2, 2, 75–76.
15. Devogele T, Parent C, Spaccapietra S (1998) On spatial database integration. *International Journal of Geographical Information Science*, 4, 1, 335–352.
16. Dhamankar R, Lee Y, Doan A, Halevy A, Domingos P (2004) iMAP: Discovering complex mappings between database schemas. In *SIGMOD Conference 2004*, Weikum G, König A, Deßloch S (eds), ACM Press, 383–394.
17. Doan A, Domingos P, Levy A (2000) Learning source description for data integration. In *WebDB (Informal Proceedings)*, 81–86.
18. Doan A, Madhavan J, Domingos P, Halevy A (2002) Learning to map between ontologies on the semantic web. In *WWW 2002, Proceedings 11th International World Wide Web Conference*, ACM, 662–673.
19. Dononi F, Lenzerini M, Nardi D, Schaerf A (1996) Reasoning in description logics. In *Principles of Knowledge Representation and Reasoning*, Brewka G (ed), CSLI Publications, 193–238.
20. Duckham M, Lingham J, Mason K, Worboys M (2006) Qualitative reasoning about consistency in geographic information. *Information Sciences*, 176, 6, 601–627.
21. Duckham M, Mason K, Stell J, Worboys M (2001) A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems*, 25, 89–103.
22. Duckham M, Worboys M (2005) An algebraic approach to automated geospatial information fusion. *International Journal of Geographic Information Science*, 19, 5, 537–557.
23. Fisher P (1996) Boolean and fuzzy regions. In *Geographic Objects with Indeterminate Boundaries*, Burrough P, Frank A (eds), Taylor and Francis, 87–94.
24. Fisher P, Wood J (1998) What is a mountain? Or the Englishman who went up a Boolean geographical concept but realised it was fuzzy. *Geography*, 83, 3, 247–256.
25. Fonseca F, Davis C, Camara G (2003) Bridging ontologies and conceptual schemas in geographic information integration. *Geoinformatica*, 7, 4, 355–378.
26. Fonseca F, Egenhofer M (1999) Ontology-driven geographic information systems. In *Proceedings Seventh Symposium on Advances in Geographic Information Systems*, Medeiros C. B (ed), 14–19.
27. Fonseca F, Egenhofer M, Agouris P, Câmara G (2002) Using ontologies for integrated geographic information systems. *Transactions in GIS*, 6, 3, 231–257.

28. Gangemi A, Pisanelli D, Steve G (1998) Ontology integration: Experiences with medical terminologies. In *Formal Ontology in Information Systems*, Guarino N (ed), IOS Press, 163–178.
29. Ganter B, Wille R (1999), *Formal Concept Analysis*, Spinger, Berlin.
30. Garcia-Molina H, Papakonstantinou Y, Quass D, Rajaraman A, Sagiv Y, Ullman J, Vassalos V, Widom J (1997) The TSIMMIS approach to mediation: Data models and languages. *Journal of Intelligent Information Systems*, 8, 2, 117–132.
31. Gómez-Pérez A, Corcho O (2002) Ontology languages for the semantic web. *IEEE Intelligent Systems*, 17, 1, 54–60.
32. Gruber T (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 2, 199–220.
33. Guarino N (1998) Formal ontology and information systems. In *Formal Ontology and Information Systems, Proceedings FOIS'98*, IOS Press, 3–15.
34. Guarino N, Masolo C, Vetere G (1999) Ontoseek: Content-based access to the Web. *IEEE Intelligent Systems*, 14, 3, 70–80.
35. Haarslev V, Möller R (2001) Description of the RACER system and its applications. In *Proc. International Description Logics Workshop (DL-2001)*, Goble C, Möller R, Patel-Schneider P (eds). http://CEUR-WS.org/Vol-49.
36. He B, Chang K. C.-C (2006) Automatic Complex Schema Matching across Web Query Interfaces: A Correlation Mining Approach. *ACM Transactions on Database Systems*, 31, 1.
37. Kavouras M, Kokla M (2002) A method for the formalization and integration of geographical categorizations. *International Journal of Geographical Information Science*, 16, 5, 439–453.
38. Kavouras M, Kokla M, Tomai E (2005) Comparing categories among geographic ontologies. *Computers & Geosciences*, 31, 2, 145–154.
39. Keefe R, Smith P (eds) (1996), *Vagueness: A reader*, Keefe R, Smith P (eds), MIT Press, Cambridge, MA.
40. Kim W, Sea J (1992) Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Computer*, 24, 12, 12–18.
41. Kokla M, Kavouras M (2001) Fusion of top-level and geographic domain ontologies based on context formation and complementarity. *International Journal of Geographical Information Science*, 15, 7, 679–687.
42. Lakshmanan L, Sadri F, Subramanian I (1993) On the logical foundations of schema integration and evolution in heterogeneous database systems. In *DOOD '93, Proceedings Third International Conference on Deductive and Object-Oriented Databases*, Ceri S, Tanaka K, Tsur S (eds), volume 760 of *Lecture Notes in Computer Science*, Springer, Berlin, 81–100.
43. Li W.-S, Clifton C (1994) Semantic integration in heterogeneous databases using neural networks. In *VLDB'94,Proceedings 20th International Conference on Very Large Data Bases*, Bocca J. B, Jarke M, Zaniolo C (eds), Morgan Kaufmann, 1–12.
44. Li W.-S, Clifton C (2000) SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. *Data and Knowledge Engineering*, 33, 49–84.
45. Madhavan J, Bernstein P, Rahm E (2001) Generic Schema Matching with Cupid. In *Proc. VLDB*, Apers P, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass R (eds), Morgan Kaufmann.

46. Manoah S, Boucelma O, Lassoued Y (2004) Schema Matching in GIS. In *AIMSA 2004*, Bussler C, Fensel D (eds), volume 3192 of *Lecture Notes in Computer Science*, Springer, 500–509.
47. McGuinness D (2003) Ontologies for information fusion. In *Proc. 6th International Conference of Information Fusion*, volume 1, 650–657.
48. Mena E, Illarramendi A, Kashyap V, Sheth A. P (2000) OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies. *Distributed Parallel Databases*, 8, 2, 223–271.
49. Miller R, Haas L, Hernandez M (2000) Schema Mapping as Query Discovery. In *Proc. VLDB*.
50. Mitra P, Wiederhold G, Kersten M (2000) A Graph-Oriented Model for Articulation of Ontology Interdependencies. In *Advances in Database Technology (EDBT)*, Zaniolo C, Lockemann P. C, Scholl M. H, Grust T (eds), volume 1777 of *Lecture Notes in Computer Science*, Springer, 86–100.
51. Noy N, Musen M (1999) An Algorithm for Merging and Aligning Ontologies: Automation and Tool Support. In *Proc. Workshop on Ontology Management at the 16th National Conference on Artificial Intelligence (AAAI-99)*.
52. Noy N, Musen M (2003) The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59, 6, 983–1024.
53. Palopoli L, Rosaci D, Terracina G, Ursino D (2005) A graph-based approach for extracting terminological properties from information sources with heterogeneous formats. *Knowledge and Information Systems*, 8, 4, 462–497.
54. Rahm E, Bernstein P (2001) A survey of approaches to automatic schema matching. *The VLDB Journal*, 10, 334–350.
55. Rosse C, Mejino J (2003) A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36, 6, 478–500.
56. Sheth A (1999) Interoperability and spatial information theory. In *Interoperating Geographic Information Systems*, Goodchild M, Egenhofer M, Fegeas R, Kottman C (eds), Kluwer, Dordrecht, Netherlands, chapter 2, 5–29.
57. Sheth A, Kashyap V (1993) So far (schematically) yet so near (semantically). In *DS-5, Proceedings IFIP Database Semantics Conference on Interoperable Database Systems*, Hsiao D, Neuhold E, Sacks-Davis R (eds), volume 25 of *IFIP Transactions*, North-Holland, 283–312.
58. Spaccapietra S, Parent C, Dupont Y (1992) Model independent assertions for integration of heterogeneous schemas. *VLDB Journal*, 1, 1, 81–126.
59. Stell J, Worboys M (1997) The algebraic structure of sets of regions. In *Spatial Information Theory, International Conference COSIT'97*, Hirtle S, Frank A (eds), number 1329 in Lecture Notes in Computer Science, Springer, 163–174.
60. Stumme G, Maedche A (2001) FCA-Merge: Bottom-up merging of ontologies. In *Proc. 17th International Conference on Artificial Intelligence (IJCAI '01)*, 225–230.
61. Tejada S, Knoblock C, Minton S (2001) Learning object identification rules for information integration. *Information Systems*, 26, 607–633.
62. Tzitzikas Y, Spyratos N, Constantopoulos P (2001) Mediators over ontology-based infomation sources. In *Proceedings WISE 1*, 31–40.

63. Uitermark H, Oosterom P, Mars N, Molenaar M (1999) Ontology-based geographic data set integration. In *Proc. International Workshop on Spatio-Temporal Database Management (STDBM'99)*, Böhlen M, Jensen C, Scholl M (eds), volume 1678 of *Lecture Notes in Computer Science*, Springer, Berlin, 60–78.
64. Umemura K, Murao O, Yamazaki F (2000) Development of GIS-based building damage database for the 1995 Kobe earthquake. In *Proc. 21st Asian Conference on Remote Sensing (ACRS)*, volume 1, 389–394.
65. Vckovski A (1998), *Interoperable and Distributed Processing in GIS*, Taylor & Francis, London.
66. Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Hübner S (2001) Ontology-based integration of information—A survey of existing approaches. In *IJCAI-01 Workshop: Ontologies and information sharing*, Stuckenschmidt H (ed), 108–117.
67. Wald L (1999) Definitions and terms of reference in data fusion. In *International Archives of Photogrammetry and Remote Sensing*, Baltsavias E, Csatho B, Hahn M, Koch B, Sieber A, Wald L, Wang D (eds), volume 32, 2–6.
68. Widom J (1995) Research problems in data warehousing. In *Proceedings 4th International Conference on Information and Knowledge Management (CIKM).*
69. Wiederhold G (1992) Mediators in the architecture of future information systems. *IEEE Computer*, 25, 3, 38–49.
70. Worboys M. F, Clementini E (2001) Integration of imperfect spatial information. *Journal of Visual Languages and Computing*, 12, 61–80.
71. Worboys M. F, Duckham M (2002) Integrating spatio-thematic information. In *Geographic Information Science*, Egenhofer M, Mark D (eds), volume 2478 of *Lecture Notes in Computer Science*, Springer, Belin, 346–361.
72. Worboys M. F, Duckham M (2004), *GIS: A Computing Perspective*, 2nd edition, CRC Press, Boca Raton, FL.
73. Zhou G, Hull R, King R, Franchitti J.-C (1995) Data integration and warehousing using H2O. *IEEE Data Engineering Bulletin*, 18, 2, 29–40.

# Index