Viewpoints on Emergent Semantics

Philippe Cudré-Mauroux[1,⋆], Karl Aberer[1] (*editors*),
Alia I. Abdelmoty[2], Tiziana Catarci[3], Ernesto Damiani[4],
Arantxa Illaramendi[5], Mustafa Jarrar[6], Robert Meersman[6],
Erich J. Neuhold[7], Christine Parent[1], Kai-Uwe Sattler[8],
Monica Scannapieco[3], Stefano Spaccapietra[1],
Peter Spyns[6], and Guy De Tré[9]

[1] EPFL, Switzerland
Tel.: +41-21-693 6787
philippe.cudre-mauroux@epfl.ch
[2] Cardiff University, UK
[3] University of Rome La Sapienza, Italy
[4] University of Milan, Italy
[5] University of the Basque Country, Spain
[6] Vrije Universiteit Brussel, Belgium
[7] Fraunhofer IPSI, Germany
[8] Technical University Ilmenau, Germany
[9] Ghent University, Belgium

Abstract. We introduce a novel view on how to deal with the problems of semantic interoperability in distributed systems. This view is based on the concept of emergent semantics, which sees both the representation of semantics and the discovery of the proper interpretation of symbols as the result of a self-organizing process performed by distributed agents exchanging symbols and having utilities dependent on the proper interpretation of the symbols. This is a complex systems perspective on the problem of dealing with semantics. We highlight some of the distinctive features of our vision and point out preliminary examples of its application.

1 Introduction

In this paper, we introduce a novel view on how to deal with the problems of semantic interoperability in distributed information systems. This view is based on the concept of emergent semantics, which sees both the representation of semantics and the discovery of the proper interpretation of symbols as the result of a self-organizing process performed by distributed agents exchanging symbols and having utilities dependent on the proper interpretation of the symbols. This is a complex systems perspective on the problem of dealing with semantics.

We first introduce a step by step reasoning underlying the concept of emergent semantics in Section 2. In the subsequent chapters, our goal is to identify current works that manifest the ideas of emergent semantics more concretely, within

⋆ Corresponding author.

S. Spaccapietra et al. (Eds.): Journal on Data Semantics VI, LNCS 4090, pp. 1–27, 2006.

the scope of today's research in areas such as distributed database systems, the Semantic Web, peer-to-peer computing or agent-based systems. Also, we identify when possible potential starting points for future applications of the concept.

This paper results from extensive discussions that have been taking place within the IFIP WG 2.6. on databases over the last two years. Initial ideas resulting from these discussions have been published in earlier invited publications [3,5]. With this article, we intend to move the discussion one step further by connecting the general concept to concrete ongoing research efforts and existing technologies.

2 The Notion of Emergent Semantics

2.1 The Notion of Semantics

Despite its wide usage in many contexts, the notion of semantics lacks a precise definition. As a least common denominator, we can characterize semantics as a relationship or mapping established between a syntactic structure and some domain. The syntactic structure is a set of symbols that can be combined following specific rules. The possible domains these symbols are related through semantics can vary widely.

Observation 1: The semantics of a syntactic structure is a relationship between a syntactic structure and some domain.

In mathematical logic, a semantic interpretation for a formal language is specified by defining mappings from the syntactic constructs of the language to an appropriate mathematical model. Denotational semantics applies this idea to programming languages. Natural language semantics classically concerns a triadic structure comprising a *symbol* (how some idea is expressed), an *idea* (what is abstracted from reality) and a *referent* (the particular object in reality) [64].

2.2 Semantics in Information Systems

Programs, database schemas, models, ontologies are unconscious artifacts and have no capacity (yet?) to refer to reality. However, software agents have various mechanisms at their disposal for establishing relationships between internal and local symbols and external meaning.

In many cases, humans are responsible for providing software agents with their initial semantics. In the simplest case, natural language vocabulary is used for the local symbols while the associated relationship with the corresponding explanation or definition of the notion concerned is very often left implicit. The hidden assumption is that meaning exchange is achieved through human cognition [80]. This can lead to communication errors as natural language is not free of ambiguity. In addition, it might happen that in a local community of practice symbols acquire an additional meaning depending on the context, which is not propagated as the exact definition is not explicitly provided.

In the setting where humans provide semantics, relationships among symbols, such as constraints in relational databases are means to express semantics. Again, the assumption is that meaning exchange is achieved through human cognition, e.g., during requirement analyses and testing, suffering some of the same problems as with the use of natural language symbols.

In order to rectify some of the problems related to the implicit representation of semantics relying on human cognition, some have proposed the approach of using an explicit reference system for relating sets of symbols in a software system. Ontologies serve this purpose: an ontology vocabulary consists in principle of formal, explicit but partial definitions of the intended meaning for a domain of discourse [34,35]. In addition, formal constraints (e.g., on the mandatoriness or cardinality of relationships between concepts) are added to reduce the fuzziness of the informal definitions. Specific formal languages (e.g., OWL) allow to define complex notions and support inferencing capabilities (generative capacity).

Observation 2: Explicitly represented semantics of a syntactic structure in an information system consists of a relationship between this syntactic structure and some generally agreed-upon syntactic structure. Thus, the semantics is represented itself by a syntactic structure.

2.3 Semantics in Distributed Systems

In a distributed environment of information agents such as in the Semantic Web or peer-to-peer systems, the aim is to have the agents interoperate irrespective of the source of their initial semantics. To that aim, an agent has to map its vocabulary (carrying the meaning as initially defined in its *base* ontology) to the vocabulary of other agents with which it wants to interoperate. In this way, a relationship of the agents' symbols to the domain consisting of other agents' symbols is established. This relationship may be considered as another form of semantics, independent of the initial semantics of the symbols.

Assuming that autonomous software agents have acquired their semantics through relationships to other agents and that agents interact without human intervention, the original *human assigned* semantics would loose its relevance; from an agent's perspective, *new* semantics would then result from the relationships to its environment. We view this as a novel way of providing semantics to symbols of autonomous agents relative to the symbols of other agents they are interacting with. Typically, this type of semantic representation is distributed such that no agent holds a complete representation of a generally agreed-upon semantics.

Observation 3: Explicitly represented semantics of an agent in a system of distributed agents can be represented through the (distributed) ensemble of relationships to other agents' syntactic structures.

2.4 Processes Creating Semantics

With the classical notion of semantics in information systems, the process of generating semantic interpretations, e.g., the generation of ontologies which

reflect shared semantics, is somewhat left outside the operation of the information systems proper. The process is assumed to rely on social interactions among humans, possibly supported in their collaborative effort by some computational and communicational tools.

Viewing semantics of information agents as a relationship to other agents allows us to internalize the discovery process of those relationships to their operation. We abandon the idea of a preexisting outside agency for forming semantic agreements, but see those as a result of the interaction of autonomous, self-interested agents. This is in line with the concept of expressing semantics through internal relationships in a distributed system. By this approach, we aim at consolidating the local semantics of autonomous information agents (respectively information systems) into a global semantics that results from a continuous interaction of the agents. The structures emerging from these continuous interactions provide meaning to the local symbols. We consider semantics constructed incrementally in this way as *emergent semantics*.

From a global perspective, considering a society of autonomous agents as one system, we observe that the agents form a complex, self-referential, dynamic system. It is well-accepted and known from many examples that such systems result (often) in global states, which cannot be properly characterized at the level of local components. This phenomenon is frequently characterized by the notion of *self-organization*. Thus, emergent semantics is not only a local phenomenon, where agents obtain interpretations locally through adaptive interactions with other agents, but also a global phenomenon where a society of agents agree on a common, global state as a representation of the current *semantic agreement* among the agents. This view of semantics as the emergence of a distributed structure from a dynamic process – or more specifically as an equilibrium state of such a process – is in-line with the generally accepted definitions of emergence and emergent structures in the complex systems literature.

Observation 4: Emergent semantics refers to distributed, emergent structures for representing semantics in a distributed information system and results from a dynamic process.

2.5 Assumptions for Enabling Emergent Semantics

The possibility to realize such an interaction process among autonomous and self-interested agents relies on a set of assumptions, each of which is quite natural in the context of distributed and autonomously operating software. First, the agents have to be able to relate their local symbols to each other. This is nothing else than the requirement of being able to communicate at a syntactic level. Then, the agents have to be able to measure the quality of the outcome of an interaction with another agent. Usually, such quality measures are encoded representations of utility measures of (human) users of the software agents. Finally, the agents have to be capable of adapting their relationships to other agents as a reaction to the measurable outcomes of earlier interactions. This corresponds to providing a certain level of autonomy to the agents in order to adapt their behavior, including their relationships to other agents, in response to earlier actions.

Observation 5: Emergent semantics is likely to occur in distributed information systems since the underlying assumptions are frequently and naturally satisfied.

2.6 Introducing Pragmatics

The careful reader will have noticed that by requiring the capability to qualitatively measure the outcomes of actions, we have introduced at this point a further dimension into the discussion, the dimension of *pragmatics*. Without pragmatics, it would be impossible to guide the process of constructing semantics during interactions with other agents. We are thus adopting a semiotic approach, jointly considering the dimensions of syntax, semantics and pragmatics. Syntax is required for agents to interact with their environment, namely other agents, semantics is required to formally describe the intended meaning of vocabularies, and in this context pragmatics provides the decision mechanisms to guide future actions based on the current interpretation of the agents state.

Observation 6: Pragmatics realized through self-interested agents that can measure the quality of the semantic interpretation of their syntactic structures in terms of their utility is an inherent prerequisite for emergent semantics.

In the following, we discuss some of the consequences we can derive from introducing the general concept of emergent semantics. These concern functional properties of emergent semantics, the potential of emergent semantics to better address hard problems of semantic interoperability, and questions related to applicability and acceptance of emergent semantics systems.

Semantic Interoperability in Information Systems. Relating information systems created independently has a long history in computer science. Section 3 illustrates how techniques drawn from distributed databases and peer data management systems can be relevant in an emergent semantics scenario. Section 4 revisits classical ontology-based systems in a similar context.

Uncertainty. Dealing with semantics and pragmatics implies the ability to quantify or measure properties of an agent's state in order to support decision making. In the case of emergent semantics, these measures are related to the proper interpretation of the agent's semantic structure. The better we understand the meaning of symbols and the more we remove uncertainty from their interpretation, the more beneficial the use of the symbols will be. Emergent semantics is based on incrementally reducing the uncertainty of symbols through exchanging information with other agents. In many cases, it will therefore be necessary to have the ability to represent uncertainty about symbols. Therefore, formalisms for representing uncertain data are an essential ingredient for emergent semantics systems.

We discuss in Section 5 which formal approaches exist for this purpose, and to what extent they are already in use in existing systems taking an emergent semantics approach.

Social Dimension. Emergent semantics systems are inherently social systems consisting of self-interested agents. Many issues relevant in artificial or natural social systems are relevant in emergent semantics systems. For example, the problem of privacy, i.e., protecting one's own information from others, leads to the inherent problem of having conflicting goals. By not revealing information, an agent can obtain an advantage in decision making whereas by revealing information it might improve the interpretation of other symbols and thus increase its utility. Also, information and the trustworthiness of agents play a role for assessing the extent to which information received from other agents is relevant for improving semantic interpretations, that is to reducing the uncertainty on the semantics of symbols.

We discuss in Section 6 current approaches in these two areas and in which ways they relate to emergent semantics.

Applicability of Emergent Semantics. The observation that emergent semantics results from a self-organizing process has some interesting consequences on the stability of emergent semantics structures. It is well-known that self-referential dynamic systems may exhibit stable states. Even if the state space of a dynamic systems is continuous, the space of stable states is discrete (Eigenstates) and stable states can be reached from many different initial states. Thus, the structure of the dynamic system implies specific states, corresponding to emergent semantics structures that we can interpret as the socially stable mutual interpretations of local symbols of autonomous agents.

This opens interesting perspectives and promises to address some of the inherently hard problems of classical ways of providing semantics in information systems. It is well known that ontologies are inherently unstable and ontology evolution is a constant challenge. Here, emergent semantics provides a natural solution as its definition is based on a process of finding stable agreements; constant evolution is part of the model and stable states, provided they exist, are autonomously detected. On the more speculative side, we see a further potential for emergent semantics. On one hand, the syntactic structure of ontologies (and other logic-based languages) is identical for local agents and for global semantic agreements. On the other hand, the available state space for processes generating emergent semantics structures might be more complexly structured and holds the potential to express semantics in a non-standard, more expressive way.

In Section 7 we outline some application areas where we expect the emergent semantics concept to be most applicable or where we can already find steps leading to solutions based on ideas related to emergent semantics.

3 Semantics in Distributed Database Systems

Observation 3 expresses semantics as a distributed ensemble of relationships to syntactic structures. Today, many distributed information systems can be characterized in a similar way, due to the existence of many interrelated data sources accessible over the Internet. Examples of such systems are among others information integration systems, data sharing and exchange applications, catalogs in e-business, and data annotation systems for scientific data. At a very

abstract level, we can see all these systems as distributed systems of interconnected nodes where nodes represent data sources.

The most well-known example of this class of systems is the mediator-wrapper architecture [85]: a mediator defining the global schema and providing facilities for answering queries on this schema is linked to all data sources which are encapsulated by wrappers. A more advanced case is a Peer Data Management System (PDMS) where the peers (nodes) represent data sources providing query answering functionalities [4,38]. Here, each peer is linked to some *neighbor* peers. The difference to the first case is that the PDMS approach does not require a dedicated centralized mediator node – instead, each peer can both ask and reformulate queries.

In both cases, the links between nodes are *semantic* links representing mappings. A *mapping* explains the meaning of an element (schema element or data value) of a given node A in terms of concepts or elements of node B, which we assume have a known meaning (at least from B's point of view). Though mappings are primary used for query rewriting on heterogeneous schemas, they can also be seen as a way to capture semantics. Basically, we can distinguish two different ways of representing mappings:

Direct mapping: a schema element of node A is mapped onto one or more elements of B. Usually, these mappings are expressed as view definitions. Here, different approaches exist [50]. In the global-as-view (GAV) approach, the integrating schema is defined as a view on the local schema. In contrast, in the local-as-view (LAV) approach, the local schemas are expressed on the global schema defined by the integration node. The combination of both solutions, the GLAV approach, combines the expressive power and allows a more flexible mapping definition. For all these kinds of views, appropriate rewriting techniques exist, e.g., query unfolding for GAV or the bucket algorithm and the MiniCon algorithm for LAV [37].

Indirect mapping: here, a common conceptualization $\mathcal{C}$, i.e., a taxonomy or an ontology, is shared by all nodes. The meaning of the elements of each node is defined in terms of concepts from $\mathcal{C}$, e.g., by annotating (linking) the elements with the concepts [78]. Based on these links one can either infer direct mappings between the nodes or simply asking queries on the conceptual level. This approach is conceptually related to the lexical approach described in Section 4.

As observed above (Observation 4), emergent semantics refers to a dynamic process. Distributed data management applications as introduced above are not static: new nodes are added or deleted and mappings have to be adjusted due to schema changes. Thus, the system evolves in a a distributed dynamic process and new semantic structures are created implicitly or explicitly. So, the question arises if and how we can feedback this new knowledge into the system. The most obvious approach is repeating the initial steps of creating mappings by hand or using schema matching techniques. A more interesting approach, closer to emergent semantics concerns, is to do this incrementally and in a (semi)automatic way. For this purpose, we distinguish in the following three kinds of system dynamics and discuss their recent developments.

3.1 Link Improvement

Mappings used for query reformulation and result translation are often not exact due to several reasons, e.g., because some concepts are not supported by a source or because of wrong decisions during mapping design. Such inaccuracies result in information loss during query answering, i.e., incomplete results or irrelevant data. This might occur both at schema level (missing attributes) as well as at data level (missing data). In order to improve a mapping we have first to assess the mapping quality. For this purpose, several quality criteria can be used, e.g., extensional and intensional completeness and relevance. The quality indicators are not only useful to choose the best source for a given query but also to try to adapt the mapping.

A first approach for determining information loss was proposed by Mena et al. [57] in the context of a ontological mediator. In this work, information loss is defined for the intensional level as the terminological difference between a query and its translation. A difference exists if concepts which are referenced in the query are not subsumed by concepts used in the translated query. At the extensional level, the Information Retrieval measures *precision* and *recall* are used and are computed based on the size of the extensions of the queried concepts. A related approach is presented in [6]. Here, several similarity measures for queries and their translations are introduced. At the intensional level, *syntactic* similarity deals with attributes used in a query, which are lost after transformation. Whereas this measure ignores the semantics of attributes, *semantic* similarity measures take this into account using two mechanisms. First, cycles in the network and therefore in the mappings are exploited to detect implicit semantic agreements. The second mechanism is based on an analysis of the query results and therefore addresses the extensional level. Another measure is described in [7] which analyzes to which extent functional dependencies or other integrity constraints are preserved after translation.

Based on mapping quality measures, we can decide if an improvement is necessary. Basically, we could simply create a new mapping and asses its quality. This ranking of candidate mappings is an important step in schema matching and the search techniques used in these approaches can be applied directly (see also Section 5). An alternative solution is an incremental adaptation. Several approaches have been proposed for this problem, e.g., [82]. However, they are primary intended for schema evolution. Hence, the adaption process is triggered by predefined schema evolution primitives.

3.2 Deriving New Links

Very often in an environment with direct mappings, one needs to follow several links, thus to compose series of mappings, in order to query a distant database. The problem of *mapping composition* can be described as follows: given two mappings $M_{A\rightarrow B}$ and $M_{B\rightarrow C}$ for three data sources A, B, C, the goal is to derive a new but equivalent mapping $M_{A\rightarrow C}$, i.e., a mapping that produces for all queries the same answers as the mappings $M_{A\rightarrow B}$ and $M_{B\rightarrow C}$. A first approach addressing this problem was described by Madhavan and Halevy [55]. This algorithm is based on so-called query rewrite graphs (QRG) encoding the mapping

formulas in the composition. In [87] another composition approach is proposed, which addresses mapping adaptations when schemas evolve. The idea is to to consider schema evolution itself as a mapping and – instead of performing a list of incremental adaptations for each schema change – to derive a composition of mappings which allows to obtain the adapted mapping through query rewriting.

Mapping composition addresses mainly the problem of deriving a shortcut for a sequence of mappings. However, if several alternative paths exist, there are still two questions: *(i)* which pair of nodes should be linked directly and *(ii)* which path among a set of candidates should be chosen? The latter can be treated as the shortest path problem in graphs where the weights of edges correspond to the quality of the represented mapping. The first question is related to the case of adding a new node. Here, we have to decide to which member node a link should be established. Under the assumption that mapping quality is the primary measure to be taken into account, this can be seen as a subproblem off clustering where we try to create direct links between nodes which are semantically close. Hence, standard (hierarchical) clustering algorithms (e.g., [11]) or dedicated decentralized approaches, e.g., as proposed in [71], can be applied.

3.3 Adding New Nodes

Adding a new data source to the system might introduce new concepts as long as they can be related to existing elements. Thus, the main task is to define a mapping between the new node and a node already participating in the system. This requires two steps: first to select an appropriate participant and second to match the schemas of the two nodes in order to derive a mapping. The first step can be supported by semantic clustering approaches described above, or by graph-theoretic heuristics assessing the connectivity of the semantic network (percolation theory) [25]. For the second step, several matching algorithms have been proposed in the literature (see [74] for a comprehensive survey). Finally, the new mapping can be further refined as already discussed.

4 Semantic Interoperability Through Linguistic Resources in Ontological Systems

4.1 On Usability Perspectives

Ontologies can be seen as *semantic axiomatizations*, that is, formal descriptions accounting for the intended meaning of a vocabulary [36]. As noted in Section 2, however, these descriptions are usually neither complete nor unequivocal [66]. Same semantics can be axiomatized in different ways, which usually reflect different *usability perspectives*, such as granularity, scope boundaries, representation primitives and constructs (i.e., epistemology), purpose/application/context, reasoning or computational scenarios. In other words, local semantic axiomatizations are substantially influenced by usability perspectives and application requirements at hand. In the problem solving research community, such an issue is called the *interaction problem*. Bylander and Chandrasekaran argued in [21] that "representing

knowledge for the purpose of solving some problem is strongly affected by the nature of the problem and the inference strategy to be applied to the problem".

As undisputed and standard ontologies are only available for a few, specific domains today, this argument leads to a fundamental challenge in ontological systems: establishing formal semantic interoperability among different *local* semantic axiomatizations fails mostly due to the diversity of usability perspectives, although all axiomatizations might intuitively agree at the domain/knowledge level (See [63] for the definition of *knowledge level*). In other words, in most cases semantic interoperability might not be achieved between two agents because their semantics are formalized in different ways, rather than because these systems do not agree on the factual/intuitive meaning in reality (also called *ontological semantics*).

Some advocate the use of *ontology alignments* (see [40] for a recent survey) to tackle this problem. Ontology alignments usually consist of formal descriptions accounting for the relationships between heterogeneous ontologies. Analogously to the *Peer Data Management Systems* paradigm described in the preceding section, these alignments create semantically interoperable networks by linking pairs of related ontologies directly or indirectly. In the following, we propose a different, complementary approach to overcome semantic heterogeneity based on linguistic resources.

4.2 An Attachment Law for Emergent Semantics

One may wonder whether ontological semantics exists, and/or whether the intuitive meaning of vocabularies can be found, even informally. Intuitive definitions and agreements about the intended meaning of vocabularies are implicit assumptions shared among human cognitive agents. Informal definitions and agreements can be found in linguistic resources (e.g., dictionaries, lexicons, glossaries, lexical databases, etc.) [41]. A linguistic resource renders the intended meaning of a linguistic term – in a gloss – as it is commonly agreed. Such agreements are not rigorous, of course, but are *commonly accepted* meanings. For example, when we use the English word "book", we actually refer to the set of implicit rules that are common to English-speaking people for distinguishing "books" from other objects. Such implicit rules (i.e., meaning) are learnt from the repeated use of word-forms and their referents in the English literature. Usually, lexicographers and lexicon developers investigate the repeated use of a word-form (e.g., based on a comprehensive corpus) to determine its underlying concept(s).

Linking or rooting the vocabulary used in local axiomatizations with concepts found in linguistic resources can help achieving *basic* semantic interoperability between different axiomatizations. For example, by using (euro) WordNet synsets [33] as a shared vocabulary space, autonomous semantic axiomatizations will be able to interoperate at least freely from language ambiguity and multilingualism.

Using linguistic resources as shared vocabulary spaces could be seen as an attachment law of emergent semantic networks; or, it could be advised in case of failures or uncertain semantic interoperations.

Linguistic resources can thus be seen as common, basic elements guiding the distributed semantic agreement process in heterogeneous ontological systems. Notice that for this purpose, not all linguistic resources can be adopted and reused; the basic (or maybe the only) requirement for a linguistic resource to be used as such is that it should provide (1) a discrimination of word meaning(s) (2) in a machine-referable manner. Resources like WordNet provide a machine-readable conceptual system for English words. Lexical resources that only list vocabularies and their similarities or that mix meaning descriptions with morphological issues are irrelevant to our purposes. Semantic or linguistic relationships between word forms (such as hyponymy, meronymy, and synonymy) could be significant but not essential in this regard. Our basic target is to enable emergent semantics networks to communalize a large asset of common word senses (i.e., concepts), independently of usability perspectives.

4.3 Axiomatization Perspectives in Two Existing Approaches

Dogma is an ontology engineering approach (see [42,43]) that allows knowledge to be modeled and represented in a double-articulation manner (domain axiomatization versus application axiomatizations). Dogma uses the notion of *ontology base* as a controlled vocabulary space shared between application axiomatizations. Such axiomatizations are called applications ontological commitments to the ontology base. The ontology base is intended to capture domain vocabularies, i.e., lexical rendering of domain concepts, similar to the knowledge level of a linguistic resource. In this way, Dogma enables different application axiomatizations to coexist and interoperate regardless of the diversity of their usability perspectives.

Similarly, MADS (see [16,68,69]) supports multiple perceptions of the same real world approach, allowing each application/task to perceive and represent real world facts according to its usability perspectives and requirements. This multi-perception approach is motivated by the fact that each application/task perceives and represents the *factual meaning* of a vocabulary according to its usability perspectives and requirements at hand. In other words, applications perceptions are (in most cases) different views of the same semantics. In this approach, a multi-perception and multi-representation database model allows designers to describe all the perceptions in the same database, and users to access either a peculiar perception or several perceptions in the same query. The multi-perception approach has been applied successfully in geographical information systems, where different axiomatizations of the same maps are seen as multiple perceptions of the same semantics.

5 Imperfect Information in Emergent Semantics

5.1 Representing Imperfection

Emergent semantics processes need ways of representing and assessing imperfection in order to dynamically refine semantic agreements. Imperfection may

be in the form of imprecision, vagueness, uncertainty, incompleteness, inconsistency, *etc.* Traditional database models and data management systems are not equipped to cope effectively with information imperfection. However, emergent semantics systems can benefit from several richer, more flexible database models better equipped to handle imperfections, both at the modeling (design time) level and at the querying (run-time) level. At design time, traditional database models (e.g., the relational model) are enriched with an ability to quantitatively or qualitatively specify imperfection, using tools such as probability theory, Dempster-Shafer theory, fuzzy logic, surprisal, and entropy. At run-time, flexible querying is introduced, defining preferences inside queries [17]. This can be done at two levels, namely intra-query and inter-query. Intra-query preferences allow to express that some values are more adequate than others, whereas inter-query preferences are used to associate different levels of importance with query conditions.

Over the years, several categorical classifications of the different types and sources of imperfect information have been presented. In accordance with the classifications of Bosc and Prade [18], Motro [60], and Parsons [70], imperfect information can be categorized as follows:

Uncertain information: information for which it is not possible to determine whether it is true or false.
Imprecise information: information which is not as specific as it should be.
Vague information: information that include elements (e.g., predicates or quantifiers) that are inherently vague (in the common day-to-day sense of the word cf. [60]).
Inconsistent information: information which contains two or more assertions that cannot hold at the same time.
Incomplete information: information for which some data are missing.

Data management approaches dealing with *uncertainty* include the possibilistic approaches and the probabilistic approaches. With possibilistic approaches, possibility theory [89] is used, where a possibility distribution is used to model the value of an attribute that is known to be uncertain. Each possible value for the attribute is assigned a membership grade that is interpreted as the degree of uncertainty [72]. Furthermore, possibility and necessity measures are attached to each tuple in the result set of a query to express the possibility and necessity of the result to be an answer to a query. Probabilistic approaches are based on probability theory, where each result in the result set of a query is extended with a probability, representing the probability of it belonging to the set [86]. Both approaches have their advantages and disadvantages. Probabilities represent the relative occurrence of an event and therefore provide more information than possibilities. Possibilities, however, are easier to apply because they are not restricted by a stringent normalization condition of probability theory.

Imprecision of data is mostly modeled with fuzzy set theory [88] and its related possibility theory [89]. Fuzzy set theory is a generalization of regular set theory in which it is assumed that there might be elements that only partially belong to a set. Therefore, a so-called membership grade, denoting the extent to

which the element belongs to the fuzzy set, is associated with each element of the universe. Two main approaches can be distinguished when modeling imprecision. First, similarity relations are used to model the extent to which the elements of an attribute domain may be interchanged [20]. Second, possibility distributions [72] are used, having the benefit of being suitable to cope with uncertainty (see above) and *vagueness*.

The treatment of *incomplete* information in databases has been widely addressed in research. A survey that gives an overview of the field is presented in [28]. The most commonly adopted technique is to model missing data with a pseudo-description, called *null*, denoting missing information. A more recent approach, based on possibility theory, [81] provides an explicit distinction between the cases of unknown data and inapplicable data.

5.2 Assessing Imperfection in Emergent Semantics Systems

Pragmatics realized through self-interested agents that can measure the degree of imperfection of semantic interpretations is an inherent prerequisite for emergent semantics (Observation 6). Modeling imperfection, however, is insufficient when it comes to measuring it. Measuring imperfection often involves an iterative process, in which initial assumptions are strengthened or discarded, and initial measures of imperfection are being refined. Such an iterative process may involve bringing together and relating information from several sources. Alternatively, one may attempt accessing a user with well-defined questions that eventually will minimize imperfection. In approaches based on possibility theory, refinement can be done by composing all available fuzzy sets related to the same imperfect data. Hereby, the intersection operators for fuzzy sets (t-norms) can be used as composition operators [89].

Recently, specific approaches emerged for assessing and dealing with imperfection in schema or ontology mappings. OMEN [59] is a probabilistic ontology mapping tool based on Bayesian Networks. Pan et. al [67] introduced ontology mapping based on a probabilistic framework developed for modeling uncertainty on the Semantic Web. Haase et al. [32] surveyed different approaches to handling inconsistency in description logics based ontologies. Corpus-Based Schema Matching [54] shows how a corpus of schemas and mappings can be used to augment the evidence about the schemas being matched. Probabilistic Message Passing [26] creates a probabilistic network to assess mapping qualities and route queries in a peer data management system. In [11], the statistical method Latent Class Analysis (LCA) is used to compute uncertainties of class memberships in an integrated database. The estimation of the completeness criteria in integrated sources is discussed in [62].

Finally, several papers appearing in this special issue deals with the problem of handling imperfect information in semantic applications. In the paper titled "Managing Uncertainty in Schema Matching with Top-K Schema Mappings", uncertainty is refined by a comparison of K schema mappings, each with its own uncertainty measure (modeled as a fuzzy relation over the two schemata). The process yields an improved schema mapping, with higher precision. In

"Intensional Semantics for P2P Data Integration", a new logical framework based on intensional logic is proposed to take into account the incomplete and locally inconsistent information on the Semantic Web. In "f-SWRL: A Fuzzy Extension of SWRL", finally, Pan et al. propose f-SWRL, a highly expressive language for the Semantic Web supporting fuzzy assertions and fuzzy rules.

6 Introduction on Social Aspects of Trust and Privacy

Emergent semantics systems are inherently social systems consisting of self-interested agents. However, while in social networks there is some form of trust among individuals belonging to the same social network, in emergent semantics systems individual peers may have serious concerns about the extent to which they may be unknowingly sharing private or personal information due to a possible inappropriate usage of these information by other peers.

This section mainly deals with the problems of sharing structures or data to enable semantic emergence, when privacy constraints are taken into account and specific agents play the role of trusted-parties whose structures are preferred in the emergence process. Data publishing and exchange are dynamic processes which are required in order for semantics to emerge: whereas private data need to be exchanged, specific protocols should be devised. Trustworthiness it related to the way local agents can build local semantics by selecting some (trustworthy) structures.

6.1 Data Privacy in Data Publishing and Data Exchange

Preserving privacy of information owned by each peer/agent is a major challenge of the emergent semantics paradigm. Peers joining a semantic community have to disclose information in order to bootstrap the agreement process and accept propositions [65]. Nevertheless, peers require privacy guarantees on data they make available to the community, such as the protection of the identities of individuals and entities. A peer can choose different forms for sharing data within the semantic community:

Data Publishing: the peer can publish its own data so that they are available to the whole community.

Data Exchange: the peer can choose to conduct data exchanges with some peers of the community. This means that data querying capabilities must be ensured, and, therefore appropriate data integration strategies (see Section 3) must be adopted in order for the peers to communicate with each other.

In the following, we summarize the current strategies and techniques relevant to privacy preservation in emergent semantics systems.

In data publishing, a major problem is to assess the risk of privacy violation, once properly disclosed data are published. Typically, anonimyzation does not mean zero privacy risk. Therefore, more sophisticated techniques need to be applied for properly dealing with privacy assurance. Among the techniques

proposed in the literature, two major classes can be distinguished, namely: perturbation-based techniques and suppression-based techniques. The former techniques have been deeply investigated in the context of statistical databases [9] and privacy preserving data mining [83]. We focus instead on some recent proposals for suppression-based methods, namely for methods that either suppress single data items in order for privacy to be preserved, or alter elementary data, e.g., by means of attribute domain generalization. K-anonimity [77] is a technique that given a relation T, ensures that each record of T can be indistinctly matched to at least k individuals. It is enforced by considering a subset of T's attributes, called quasi-identifiers, and forcing the values that T's records have on quasi-identifiers to appear with at least k occurrences. A recent technique [49] considers the quantitative evaluation of the privacy risk in case anonymized data are released. In this work, a database is modeled as a sequence of transactions, and the frequency of an item x in the database is the fraction of transactions that contain that item. An hypothetical attacker can have access to similar data and use them in order to breach the privacy of disclosed data. The knowledge of the attacker is modeled as a belief function that represents the guess that the attacker can make on the actual frequencies of items in the database. In [58], the authors provide an analysis of the query-view security problem. Given n views, the problem is to check if the views disclose any information about a given secret query. The query-view security problem is characterized by means of the notion of critical tuple for a query Q, that considers a tuple t critical for Q if there are some instances of the database for which dropping t makes a difference. In [58], the authors demonstrate that a query Q is insecure w.r.t. a set of views if and only if they share some common critical tuples.

In data exchange, proposed techniques investigate how to perform query processing by revealing to the involved parties only a controlled, a-priori defined set of data. More specifically, $S1$ and $S2$ being two data peers, and given a query Q involving data at both peers, privacy preserving query answering ensures that only the result of Q will be learnt by $S1$ and $S2$, without revealing any additional information to either party.

Some of most interesting results in our context regard secure set intersection protocols [61]. Secure set intersection protocols deal with performing intersection between two lists with each party only learning the result of the intersection. In an emergent semantics system, this may be used by two agents to discover which elements they have in common. A work that specifically deals with privacy preserving query answering is Agrawal's work [10] relying on commutative encryption. In [30], aggregation operations are added to the intersection and equijoin operations proposed by Agrawal, and computational costs due to encryption/decription are reduced. In [52], several extension to Agrawal's protocol are proposed, and the notion of secure data ownership certificate is provided, with purpose of attesting the proper ownership of data in a database.

Privacy preservation in both data publishing and data exchanges is a new area that presents several interesting research challenges including: approximate operations, e.g., secure approximate joins and secure record linkage; symmetric

protocols that would be useful for emergent semantics contexts, in which there is no distinction between sender and receiver in data exchanges; schema-level privacy management, in which the rewriting of queries should be performed by taking into account privacy requirements also on schema information.

6.2 Learning Metadata Trustworthiness

On the global Internet, information interchange within distributed communities is mostly self-organizing: as community members interact, useful information is published and exchanged more frequently, soon becoming widespread. Community members often use metadata for creating and spreading their opinions about content, quality, type, creation, and even spatial geo-location of the information items they share. Research has widely acknowledged that sharing metadata within communities makes information discovery easier and may reduce data redundancy; but it is also important to remember that shared metadata are subject to constant scrutiny and debate in the social interaction between community members. Even apparently innocuous assertions on class subsumption (e.g., "*Contemporary Music is a subset of Classical Music*") or instance classification (e.g., "*Mussorgski's "Pictures at-an-Exibition" suite belongs to Contemporary Music*") may turn out to be debatable or plainly wrong according to the prevailing usability perspectives (see Section 4) in the community. In the following, we describe how explicit representation of trust metadata can be a source of emergent semantics. Our discussion is based on a recent research approach [23], which exploits user feedback for adapting metadata to the specific contexts and belief systems where communities operate. The overall effect of a community-wide trust management mechanism can be twofold:

Knowledge Quality Improvement obtained by keeping the community's overall body of knowledge under a continuous evolutionary pressure.

Knowledge Enrichment achieved by generating a layer of metadata expressing the evolution of users' views on each other's assertions. This procedural knowledge can later be queried to monitor the community's collective behavior, and even used to restructure the original metadata.

Trust management in decentralized (P2P) networks was first addressed by Aberer [8]. A complete survey of trust and reputation management systems can be found in [14]. More recently, the research focus shifted to secure algorithms for reputation management in P2P environments, like the `P2PRep` algorithm described in [27]. Unfortunately, the terminology used in the field is not always consistent [14]; for the sake of clarity, we shall use the term *trust* to denote a user p's willingness to rely for some practical purpose on a metadata assertion a stated by another user q (denoted as $T_a(p,q)$). The term *reputation* will be used to quantitatively express p's judgment about q's trustworthiness, denoted by $R(p,q)$ and based on the latest assertion and/or on all metadata q has produced. Indeed, one might be tempted to identify trust and reputation concepts, e.g., by writing $R(p,q) = min_a\{T_a(p,q)\}$. However, in a community-based knowledge sharing scenarios, trust (on an assertion) and reputation (of its source) do not

always coincide. In real-world communities, reputation is only one among the many factors determining mutual trust; at the very least, any model of trust and reputation should take into account *reputation aging*, e.g., by writing $T_a(p, q, t) = R(p, q, t_0)e^{-\beta(t-t_0)}$, for $t > t_0$.

Based on users' behavior, it is possible to generate and publish specific *trust assertions*. For the sake of simplicity, we consider simple assertions of the form $T_a(p, q) = \alpha$, expressing the level of trust α of a peer p in the assertion a put forward by peer q. These assertions are community-specific and provide an interesting example of emergent semantics. For instance, suppose that an assertion a put forward by a user q states that a resource r, a `.mp3` file, belongs to the class of `CountrySongs`. If after downloading r, user p stores it into a local directory named `CountryMusic`, a *trust assertion* $T_a(p, q) = \alpha$ can be automatically generated. Defining the semantics of *trust values* like α in terms of *belief* in assertion a, in terms of a's *relevance* to their purposes, is in itself an open research problem, especially in a non-anonymous scenario. Another open issue is defining the appropriate *trust algebra* for combining trust assertions in order to create a *Web of trust* (an important although preliminary step toward a solution was made in [75]). Here, we simply assume $\alpha \in [0, 1]$. Trust assertions form an independent, evolving metadata layer that can be stored at a central server or at distributed peers. Emergent semantics hidden within the trust metadata layer can be exploited to compute *trusted views* over the original metadata assertions, e.g., by disregarding assertions whose community-wide trust level is below a given threshold.

In this process, individual trust degrees have to be aggregated (in the simplest case, by user and/or by resource). Some approaches [22] use Fuzzy Cognitive Maps (FCM) to model the relevance of the trust inputs before their aggregation, while the REGRET system [76] was an early attempt to use fuzzy concepts for analyzing the impact on trust of social networks in electronic marketplaces. Multi-criteria compensative aggregators like the *Ordered Weighted Average* (OWA) and the *Weighted Ordered Weighted Average* (WOWA)[31] are computationally very efficient and appear to be well suited to the synthesis of peer opinions in decentralized networks [13]. Hybrid approaches including approximate reasoning [79], where aggregated trust assertions are used as inputs to an inference system, look more promising inasmuch they provide a high-level symbolic representation of trust computation as an inference process, potentially supporting full human understanding of trust degree levels.

7 Emergent Semantics Applications

Through the years, organizations and enterprises have developed data and information exchange systems that are now vital for their daily operations. Currently deployed solutions, however, are now facing a major challenge. On today's global information infrastructure, data semantics is more and more context and time-dependent, and cannot be fixed once and for all at design time. Perhaps more importantly, identifying emerging relationships among previously unrelated

information items (e.g., during data exchange) may dramatically change their business value. In this Section, we explore several applications trying to address this challenge.

7.1 Communication of Agent-Based Data Systems

A recent trend has been developed toward enhancing the functionality of data systems by appropriate data agents. A step forward in this scenario consists in offering a real interoperation possibility among agents coming from independently developed data systems, by making minor adaptations on them. By real interoperation, we mean an interoperation based on the semantics of the communications (communication among agents is in general based on the interchange of messages) which takes the matter far beyond the syntactic functionality provided by exchange standards such as the widely spread XML [19] or, more specifically, EDI standards [1] in the area of electronic commerce.

There are two ways in which agent-based data systems can interoperate among themselves. First, through messages that are interchanged among the agents of both systems, and second, using Web Services provided by each data system. We consider here the first way, where agents typically have to be aware in advance of the structure, language and semantics of the messages in order to deal with them. In the following, we sketch an approach based on emergent semantics to relax those constraints, enabling communication (total or partial) for agents coming from different and independently developed systems.

In our opinion, real data systems interoperation will be possible only if there exists some agreement on the classes of messages used by the agents and the possibility of constructing new kinds of messages by composition or restriction of already known classes. Furthermore, the interpretation of a message should be made on the fly and adapted to the context where it appears. In that scenario, we advocate for a proposal that favors the interoperation among agentized data systems by allowing to send/receive suitable messages to/from agents of another system without requiring the establishment of a common communication pattern in advance. Our proposal (see [15] for details) is used as a basis for automating the detection and resolution of conflicts that arise when dealing with messages interchanged by agents from different systems.

In particular, we have developed a formal ontology we call *CommOnt* (Communication acts Ontology), which is a key element in the proposal and acts as an implicitly shared lexical resource (see Section 4). Agents commit to that ontology if their *observable* actions are consistent with the definitions in the ontology. The main part of CommOnt is constituted by terms related to the messages interchanged by agents representing different data systems. If a data system can deal with a particular class M of messages, then it can also deal with any message of a subclass of M in the CommOnt ontology. We claim that the CommOnt ontology provides interoperability support due to the recognition of communication acts from one language as instances of communication acts in another language. Sometimes, the *translation* will be incomplete, but correctly modeled partial interoperability is a starting point for the emergent agreement

process (see Sections 3 and 5), and is most of the time more preferable to the *not understood* answer given nowadays.

7.2 Self-organizing Hierarchical Structures in Trust-Based Architectures

Current knowledge management systems classify resources of interest within hierarchical structures. In this context, customization and evolution of categories is a major issue, inasmuch there is no unique access structure that suits every community. Traditionally, the approach to this problem involved human attention, valorizing the contribution of each community member in the knowledge creation activity with his daily work [51,84]. As human attention is today considered as one of the scarcest resources, we propose below an approach based on emergent semantics principles to derive hierarchical structures and create customized categories semi-automatically.

We designed an architecture to be deployed in association with existing systems proposed by industrial research groups for *bottom-up* construction of categories. Specific examples of existing systems include the intelligent personal hierarchy for information `iPHI` proposed by BT Exact [56] as well as the KIWI knowledge sharing platform [24], later integrated within the `Verity` knowledge organizer tool by IBM[73]. The idea behind `iPHI` is to auto-configure access to multiple sources of information based on customized categories and fuzzy matching of meta-data structure as well as content. Support for emerging trust enables our architecture to validates existing hierarchies according to the views (*usability perspectives*) of the user community and to discover new categories.

Generally speaking, we introduce a *Trust Layer* including a centralized *Metadata Publication Center* that acts as a Napster-style index, collecting and displaying metadata assertions, possibly in different formats and coming from different sources. Metadata are indexed by the Publication Center and anonymous users interact with them, providing an implicit or explicit evaluation of metadata trustworthiness. Periodically, trust-based evaluations are forwarded by the Publication Center to a *Trust Manager* module, in the form of signed assertions built using the well-known technique of *reification*. This choice allows our system to interact with heterogeneous formats, including Semantic-Web style metadata and XML-based metadata like `iPHI`. In turn, our Trust Manager is composed of two functional sub-modules: the *Trust Evaluator* examines metadata and evaluates their reliability while the *Trust Aggregator* aggregates all inputs coming from the (possibly multiple) trust evaluators. This Trust Layer can manage a large amount of assertions produced by heterogeneous sources, and allows the emergence of metadata complying with specific community views.

7.3 Semantics for the Geospatial Web

Numerous efforts are currently active toward the development of the *Geospatial Semantic Web* (GSW). The GSW, based on a sound spatial data infrastructure

(SDI), aims to enable the discovery, access and utilization of dynamic, global geographic data sets, web resources and services and to allow for their coherent combination and management. Standardized spatial ontologies are at the heart of the GSW and are proposed as means of handling problems of semantic interoperability resulting from the ad-hoc use of geographic data and spatial methods. Specification of such ontologies is the focus of the recently announced Open Geospatial Consortium (OGC) Geospatial Semantic Web Interoperability Experiment [53]. The intention is to develop means of expressing spatial queries in a semantic manner (i.e., with an ontology) and to provide web services to fulfill these queries. An architecture of ontologies is proposed [47], including a base ontology, for capturing the spatial models underlying the geographic information, a geospatial service ontology and domain ontologies. Also, place-name ontologies have been shown to play a central role in supporting the development of a spatially-aware search engines, allowing for geographic information retrieval on the web [44].

The question of which semantics to encode in such ontologies is an active research question [2,29,48]. There are inherent complexities associated with modeling information in the geographic domain, firstly related to the nature of the phenomena themselves, for example, with regards to handling multiple representations and levels of generalization or accommodating levels of error in the geometric locations, and secondly due to the variations in the ways we interpret and use the data (*usability perspectives*), e.g., national, cultural and institutional differences in the description of the data. The problem is non-trivial, as much of the useful semantics of the data are implicit in their inherent spatial structure. In particular, the multiple types of spatial relationships that exist between the geographic phenomena are not normally explicitly derived or coded. In what follows, some examples are given that employ emergent semantics methods for discovering and self-organizing geospatial data.

Automatic extraction of metadata from geographic data sets has been described in [39,46]. However, existing metadata standards facilitate the encoding of only limited semantics of the data, related for example, to the date of creation, geo-referencing system used, total extent, etc. A large amount of useful semantics is implicit and can be interpreted only by the identification of *relationships* between features, and characteristics of features such as their density, distribution, etc. For example, the area designating a city centre on a map can be identified by studying the types of buildings and roads, and their structure and density. Similar studies can distinguish between small towns and large cities, etc. Spatial data mining techniques are proposed in [39] to allow for the automatic extraction of such semantics. One can envision that such a process of semantic discovery and enrichment of metadata to be continuous and dynamic reflecting data updates and evolving geo-ontologies.

Folksonomies have been proposed by Keating and Montoya [45] as a complementary method for metadata enrichment in geoportals. Data mining is used to identify the interesting metadata from the collection of tags, annotations and

comments provided by users. New semantics in the form of new concepts or classification hierarchies or relationships may emerge as a result of this process which can then be reflected back in the underlying ontologies. Geo-semantics discovery of the impreciseness in geographic place names has been demonstrated in the works of Arampatzis et al. [12]. Many place names that are commonly employed within web document and in search queries are vague. For example, terms such as "Midwest" in the US and "Midlands" in the UK have no formal geometric boundary and may be interpreted differently by different people. The method proposed involved soliciting information about the spatial extent of the imprecise region by identifying places that are contained inside it. The assumption is that place names that co-occur in the same web document are related. Hence, web documents are geo-parsed to detect related places, and techniques for isolating places which are likely to be part of the target region are then employed. Boundaries of the contained *crisp* places are derived from the geo-ontology and the new delineated boundary of the imprecise region is added to the geo-ontology. The process is dynamic, as iterative refinement of the boundary of the region may be envisaged when new web resources are found.

7.4 PicShark: Recontextualizing Structured Metadata in a Distributed Photo-Sharing Application

Metadata have long been recognized as an efficient way to help manage data and are today widely used by operating systems, personal information managers or media libraries. The general idea is simple: adding a set of keywords or series of attributes in order to facilitate information categorization and retrieval. What is new is the recent focus on formats that let end-users freely define custom metadata schemas befitting their annotation needs.

More and more applications take advantage of structured metadata to organize large amount of information such as picture collections. The problem we want to tackle lies in the fact that *none* of these applications allows to meaningfully share structured metadata to enable global search capabilities in large scale distributed settings. Exploiting structured metadata in distributed environments is intrinsically difficult, given that the metadata have to be extracted from their original context and integrated, i.e., *recontextualized*, into the distributed infrastructure. In the end, we are confronted with two fundamental hurdles preventing photos annotated with local metadata from being shared:

Local Semantics: The classes and instances introduced by end-users to annotate their photos locally might not make sense on a larger scale, and have to be related to their counterparts in the distributed infrastructure.

Metadata scarceness: Realistically, a (potentially large) fraction of shared photos will not be annotated by the user, leaving some (most) of the related assertions incomplete. This lack of annotation hampers any system relying on annotations to retrieve instances.

PicShark is a distributed, peer-to-peer system taking advantage of structured metadata to meaningfully share annotated pictures in very large scale decentralized environments. It provides a solution to both of the aforementioned problems in a self-organizing context where information entropy (in terms of missing metadata and ontological heterogeneity) is gradually alleviated through user interaction. PicShark indexes photos, low-level features extracted from the photos, metadata and schemas in a distributed index structure. The system then tries to find correspondences between pictures, metadata and schemas in order to relate instances and schemas (through mappings, see Section 3), and to *propagate* metadata from one photo to other related photos. Queries are forwarded dynamically using Semantic Gossiping [7], and schema mappings self-organize through Probabilistic Message Massing [26]. The overall system can be seen as a decentralized emergent semantics application, where computationally expensive operations are confined to the edge of the network and global processes rely on a distributed hash table to ensure graceful scalability.

8 Conclusions

With the rapid emergence of social applications on the Web, self-organization principles have once again proven their practicability and scalability: through Technorati Ranking, Flickr Interestingness or del.icio.us recommendations, an ever-increasing portion of the Web self-organizes around end-users semantic input. The Semantic Web, with its rich heritage in logic, has so far little benefitted from this trend. In this paper, we advocate a more decentralized, user-driven and imperfect (in terms of soundness and completeness) Web of semantics that self-organizes dynamically. We tried to highlight some of the distinctive features of our vision as well as point out existing examples of its application.

One of the important remaining issues we did not tackle in this paper is the necessary human trust that has to be given to the resulting emergent semantics structure. Interpretations of precise formal structures, when they are concerned with real world models, remain incomplete and ambiguous. The very rich and varying experience of human beings allows many interpretations of formal models and as a consequence acceptance of such models is usually only achieved after extensive human experimentation and interpretation. Companies like Google or eBay already have to face similar problems today, but this issue gets even more sensitive in an emergent semantics scenario where data organization, data description and data manipulation all depend on semi-automatically generated, self-organizing structures.

Acknowledgment

We would like to thank Avigdor Gal for his insightful comments and suggestions about this work.

References

1. United nations directories for electronic data interchange for administration, commerce and transport. http://www.unece.org/trade/untdid/.
2. A. I. Abdelmoty, P.D. Smart, C.B. Jones, G. Fu, and D. Finch. A critical evaluation of ontology languages for geographic information retrieval on the internet. *Journal of Visual Languages and Computing*, 16(4):331–358, 2005.
3. K. Aberer, T. Catarci, P. Cudré-Mauroux, T. Dillon, S. Grimm, M. Hacid, A. Illarramendi, M. Jarrar, V. Kashyap, M. Mecella, E. Mena, E. J. Neuhold, A. M. Ouksel, T. Risse, M. Scannapieco, F. Saltor, L. de Santis, S. Spaccapietra, S. Staab, R. Studer, and O. De Troyer. Emergent Semantics Systems. In *International Conference on Semantics of a Networked World (ICSNW)*, 2004.
4. K. Aberer and P. Cudré-Mauroux. Semantic Overlay Networks. In *International Conference on Very Large Databases (VLDB)*, 2005.
5. K. Aberer, P. Cudré-Mauroux, and A. M. Ouksel (Eds.). Emergent Semantics Principles and Issues. In *International Conference on Database Systems for Advanced Applications (DASFAA)*, 2004.
6. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. Start making sense: The Chatty Web approach for global semantic agreements. *Journal of Web Semantics*, 1(1):89–114, 2003.
7. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. The chatty web: emergent semantics through gossiping. In *WWW 2003*, pages 197–206, 2003.
8. K. Aberer and Z. Despotovic. Managing trust in a p2p information systems. In *Intl. Conf. on Information and Knowledge Management (CIKM)*, 2001.
9. N.R. Adam and J.C. Wortmann. Security control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4), 1989.
10. R. Agrawal, A. Evfimievski, and R.Srikant. A formal analysis of information disclosure in data exchange. In *Proc. of SIGMOD*, 2003.
11. E. Altareva and S. Conrad. Statistical Analysis as Methodological Framework for Data(base) Integration. In *ER 2003*, pages 17–30, 2003.
12. A. Arampatzis, M. Kreveld, C.B. Jones, S. Vaid, P. Clough, H. Joho, M. Sanderson, M. Benkert, and A. Wolff. Web-based delineation of imprecise regions. In *SIGIR Workshop on Geographic Information Retrieval*, 2004.
13. R. Aringhieri, E. Damiani, S. De Capitani Di Vimercati, S. Paraboschi, and P. Samarati. Fuzzy techniques for trust and reputation management in anonymous peer-to-peer systems. *Journal of the American Society for Information, Science and Technology*, 1(1), 2006.
14. J. Audun, I. Roslan, and C.A. Boyd. Survey of trust and reputation systems for online service provision. *Decision Support Systems*, To appear.
15. M. I. Bagüés, J. Bermúdez, A. Illarramendi, A. Tablado, and A. Goñi. Semantic interoperation among data systems at a communication level. *Journal on Data Semantics V*, 2006.
16. S. Balley, C. Parent, and S. Spaccapietra. Modeling geographic data with multiple representations. *International Journal of Geographic Information Systems*, 18(4):329–354, 2004.
17. P. Bosc, D. Kraft, and F. Petry. Fuzzy sets in database and information systems: status and opportunities. *Fuzzy Sets and Systems*, 153(3):418–426, 2005.
18. P. Bosc and H. Prade. An introduction to fuzzy set and possibility theory based approaches to the treatment of uncertainty and imprecision in database management systems. In *Workshop on Uncertainty Management in Information Systems: From Needs to Solutions*, Catalina, California, 1993.

19. T. Bray, J.Paoli, C.M. Sperberg-McQueen, E. Maler, and F. Yergeau. Extensible markup language (xml) 1.0. http://www.w3.org/TR/2004/REC-xml-20040204.
20. B.P. Buckles and F. Petry. Generalised database and information systems. In J.C. Bezdek, editor, *Analysis of fuzzy Information*. CRC Press, 1987.
21. T. Bylander and B. Chandrasekaran. Generic tasks in knowledge-based reasoning: The right level of abstraction for knowledge acquisition. *Knowledge Acquisition for Knowledge Based Systems*, 1, 1988.
22. C. Castelfranchi, R. Falcone, and G. Pezzulo. Trust in information sources as a source for trust: a fuzzy approach. In *International Joint Conference on Autonomous Agents and Multiagent systems (AAMAS)*, 2003.
23. P. Ceravolo, E. Damiani, and M. Viviani. *Soft Computing for Information Retrieval on the Web*, chapter Adding a Trust Layer to Semantic Web Metadata. Elsevier, 2006.
24. A. Corallo, E. Damiani, and G. Elia. An ontology-based knowledge management system enabling regional innovation. In *Eurasia-ICT Workshop on E-Learning Platforms Technologies*, 2002.
25. P. Cudré-Mauroux and K. Aberer. A Necessary Condition For Semantic Interoperability in the Large. In *Ontologies, DataBases, and Applications of Semantics for Large Scale Information Systems (ODBASE)*, 2004.
26. P. Cudré-Mauroux, K. Aberer, and A. Feher. Probabilistic Message Passing in Peer Data Management Systems. In *International Conference on Data Engineering (ICDE)*, 2006.
27. E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati. Managing and sharing servents' reputations in p2p systems. *IEEE Trans. Knowl. Data Eng.*, 15(4):840–853, 2003.
28. C.E. Dyreson. A bibliography on uncertainty management in information systems. In A. Motro and P. Smets, editors, *Uncertainty Management in Information Systems: From Needs to Solutions*. Kluwer Academic Publishers, Boston, MA, 1997.
29. M. Egenhofer. Towards the semantic geospatial web. In *Proceedings of ACM-GIS*, pages 1–4, 2002.
30. F. Emekci, D. Agrawal, A. El Abbadi, and A. Gulbeden. Privacy preserving query processing using third parties. In *Proc. ICDE*, 2006.
31. J. Fodor, J. L. Marichal, and M. Roubens. Characterization of the ordered weighted averaging operators. *IEEE Trans. on Fuzzy Systems*, 3(2):236–240, 1995.
32. A Framework for Handling Inconsistency in Changing Ontologies. P. Haase and F. van Harmelen and Z. Huang and H. Stuckenschmidt and Y. Sure. In *International Semantic Web Conference (ISWC)*, 2005.
33. M. George, R. Beckwithand C. Fellbaum, C. Gross, and K. Miller. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.
34. T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 6(2):199–221, 1993.
35. N. Guarino. Formal ontologies and information systems. In Nicola Guarino, editor, *Proceedings of FOIS '98*, pages 3 – 15. IOS Press, 1998.
36. N. Guarino. Formal ontology in information systems. In *Proceedings of FOIS*, pages 3–15, 1998.
37. A. Y. Halevy. Answering queries using views: A survey. *VLDB Journal*, 10(4):270–294, 2001.
38. A. Y. Halevy, Z. G. Ives, J. Madhavan, P. Mork, D. Suciu, and I. Tatarinov. The Piazza Peer Data Management System. *IEEE Trans. Knowl. Data Eng.*, 16(7):787–798, 2004.

39. F. Heinzle and M. Sester. Derivation of implicit information from spatial data sets with data mining. In *20th Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2004.
40. J. Euzenat et al. State of the art on current alignment techniques. In *KnowledgeWeb Deliverable 2.2.3, http://knowledgeweb.semanticweb.org.*
41. M. Jarrar. Towards the notion of gloss, and the adoption of linguistic recourses in formal ontology engineering. In *Global Wordnet Conference (GWC)*, 2006.
42. M. Jarrar, J. Demey, and R. Meersman. On using conceptual data modeling for ontology engineering. *Journal on Data Semantics (Special issue on Best papers from the ER, ODBASE, and COOPIS 2002 Conferences)*, LNCS 2519:185–207, 2002.
43. M. Jarrar and R. Meersman. Formal ontology engineering in the dogma approach. In *International Conference on Ontologies, Databases and Applications of Semantics (ODBase)*, pages 1238–1254, 2002.
44. C.B. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science: Third International Conference, (GIScience'04)*, volume LNCS 3234, pages 125–139, 2004.
45. T. Keating and A Montoya. Folksonomy extends geospatial taxonomy. *Directions Magazine*, 2005.
46. E. Klien and M. Lutz. The role of spatial relations in automating the semantic annotation of geodata. In *COSIT*, pages 133–148, 2005.
47. D. Kolas, J. Hebeler, and M. Dean. Geospatial semantic web: Architecture of ontologies. In *GeoSpatial Semantics: First International Conference*, volume LNCS 3799, pages 183–194, 2005.
48. W. Kuhn. Geospatial semantics: Why, of what and how. *Journal on Data Semantics III*, LNCS 3534:1–24, 2005.
49. L.V.S. Lakshmanan, R.T. Ng, and G. Ramesh. To do or not to do: the dilemma of disclosing anonymized data. In *Proc. of SIGMOD*, 2005.
50. M. Lenzerini. Data Integration: A Theoretical Perspective. In *PODS 2002*, pages 233–246, 2002.
51. E. Lesser and K. Everest. Using communities of practice to manage intellectual capital. *Ivey Business Journal*, pages 37–41, March/April 2000.
52. Y. Li, J. D. Tygar, and J.M. Hellerstein. Private matching. Intel Research, IRB-TR-04-005, 2004.
53. J. Lieberman, T. Pehle, and M. Dean. Semantic evolution of geospatial web services. In *W3C Workshop on Frameworks for Semantics in Web Services*, 2005.
54. J. Madhavan, Ph. A. Bernstein, A. Doan, and A. Y. Halevy. Corpus-based Schema Matching. In *International Conference on Data Engineering (ICDE)*, 2005.
55. J. Madhavan and A. Y. Halevy. Composing Mappings Among Data Sources. In *VLDB 2003*, pages 572–583, 2003.
56. T. P. Martin and B. Azvine. Acquisition of soft taxonomies for intelligent personal hierarchies and the soft semantic web. *BT Technology Journal*, 21(4):113–122, 2003.
57. E. Mena, V. Kashyap, A. Illarramendi, and A. P. Sheth. Imprecise Answers in Distributed Environments: Estimation of Information Loss for Multi-Ontology Based Query Processing. *Int. J. Cooperative Inf. Syst.*, 9(4):403–425, 2000.
58. G. Miklau and D. Suciu. A formal analysis of information disclosure in data exchange. In *Proc. of SIGMOD*, 2004.
59. P. Mitra, N. F. Noy, and A. R. Jaiswal. OMEN: A Probabilistic Ontology Mapping Tool. In *International Semantic Web Conference (ISWC)*, 2005.

60. A. Motro. Management of uncertainty in database systems. In W. Kim, editor, *Modern Database Systems, The object model, interoperability and beyond*. Addison-Wesley, Reading, Massachusetts, 1995.
61. M. Naor and B. Pinkas. Oblivious transfer and polynomial evaluation. In *Proc. of the 31th ACM Symposium on Theory of Computing*, 1999.
62. F. Naumann, C. Freytag, and U. Leser. Completeness of integrated information sources. *Inf. Syst.*, 29(7):583–615, 2004.
63. A. Newell. The knowledge level. *Artificial Intelligence*, 18(1), 1982.
64. C.K. Ogden and I.A. Richards. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Routledge & Kegan Paul Ltd., London, 10 edition, 1923.
65. A. M. Ouksel. *A Framework for a Scalable Agent Architecture of Cooperating Heterogeneous Knowledge Sources*. Springer Verlag, 1999.
66. A. M. Ouksel and I. Ahmed. Ontologies are not the panacea in data integration: A flexible coordinator for context construction. *Journal of Distributed and Parallel Databases*, 7,1, 1999.
67. R. Pan, Z. Ding, Y. Yu, and Y. Peng. A Bayesian Network Approach to Ontology Mapping. In *International Semantic Web Conference (ISWC)*, 2005.
68. C. Parent, S. Spaccapietra, and E. Zimanyi. *Conceptual Design for Traditional and Spatio-Temporal Applications – The MADS Approach*. Springer, 2005.
69. C. Parent, S. Spaccapietra, and E. Zimanyi. The murmur project: Modeling and querying multi-representation spatio-temporal databases. *Information Systems*, 2005.
70. S. Parsons. Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, 8(3):353–372, 1996.
71. Y. Petrakis and E. Pitoura. On Constructing Small Worlds in Unstructured Peer-to-Peer Systems. In *EDBT Workshops 2004*, pages 415–424, 2004.
72. H. Prade and C. Testemale. Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries. *Information Sciences*, 34:115–143, 1984.
73. P. Raghavan. Structured and unstructured search in enterprises: Verity. *IEEE Data Engineering Bulletin*, 4(6), 2001.
74. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.
75. M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the Second International Semantic Web Conference (ISWC 03)*, 2003.
76. J. Sabater and C. Sierra. Reputation and social network analysis in multi-agent systems. In *International Joint Conference on Autonomous Agents and Multiagent systems (AAMAS)*, 2002.
77. P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proc. of PODS*, 1998.
78. K.-U. Sattler, I. Geist, and E. Schallehn. Concept-based querying in mediator systems. *VLDB Journal*, 14(1):97–111, 2005.
79. S. Schmidt, R. Steele, T. S. Dillon, and E. Chang. Building a fuzzy trust network in unsupervised multi-agent environments. In *OTM Workshops*, 2005.
80. P. Spyns and J. De Bo. Ontologies: a revamped cross-disciplinary buzzword or a truly promising interdisciplinary research topic? *Linguistica Antverpiensia - NS*, (3):279 – 292, 2004.

81. G. De Trè, R. De Caluwe, and H. Prade. The ansi/x3/sparc dbms framework: Report of the study group on data base management system. *Information Systems*, 3, 1978.
82. Y. Velegrakis, R. J. Miller, and L. Popa. Mapping Adaptation under Evolving Schemas. In *VLDB 2003*, pages 584–595, 2003.
83. V. Verykios, E.Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and A.K. Elmagarmi. State of the art on privacy preserving data mining. *Sigmod Record*, 33(1), 2004.
84. E. Wenger. Communities of practice: The key to knowledge strategy. *Knowledge Directions*, 6(4):48–64, 1999.
85. G. Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3):38–49, 1992.
86. S.K.M. Wong, Y. Xiang, and X. Nie. Representation of bayesian networks as relational databases. In *International Conference on Information Processing and Management of Uncertainty*, pages 159–165, Paris, France, 1994.
87. C. Yu and L. Popa. Semantic Adaptation of Schema Mappings when Schemas Evolve. In *proc. of VLDB 2005*, pages 1006–1017, 2005.
88. L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
89. L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.