Emergent Semantics Systems*

Karl Aberer[1], Tiziana Catarci[2], Philippe Cudré-Mauroux[1]**, Tharam Dillon[3], Stephan Grimm[4], Mohand-Said Hacid[5], Arantza Illarramendi[6], Mustafa Jarrar[7], Vipul Kashyap[8], Massimo Mecella[2], Eduardo Mena[9], Erich J. Neuhold[10], Aris M. Ouksel[11]***, Thomas Risse[10], Monica Scannapieco[2], Fèlix Saltor[12], Luca de Santis[2], Stefano Spaccapietra[1], Steffen Staab[4], Rudi Studer[4], and Olga De Troyer[7]

[1] Swiss Federal Institute of Technology (EPFL), Switzerland
[2] Univ. of Roma La Sapienza, Italy
[3] Univ. of Technology, Sydney, Australia
[4] Univ. of Karlsruhe, Germany
[5] Univ. of Lyon 1, France
[6] Univ. of the Basque Country, Spain
[7] Vrije University of Brussels, Belgium
[8] National Library of Medicine, USA
[9] Univ. of Zaragoza, Spain
[10] Fraunhofer IPSI, Germany
[11] Univ. of Illinois at Chicago, USA
[12] Univ. Politècnica de Catalunya, Spain

Abstract. With new standards like RDF or OWL paving the way for the much anticipated Semantic Web, a new breed of very large scale semantic systems is about to appear. Traditional semantic reconciliation techniques, dependent upon shared vocabularies or global ontologies, cannot be used in such open and dynamic environments. Instead, new heuristics based on emerging properties and local consensuses have to be exploited in order to foster semantic interoperability in the large. In this paper, we outline the main differences between traditional semantic reconciliation methods and these new heuristics. Also, we characterize the resulting *emergent semantics* systems and provide a couple of hints vis-à-vis their potential applications.

1 Introduction

Global economics needs global information. The time is over when enterprises were centralized and all the information needed to operate an enterprise was

* The work presented in this paper reflects the current status of a collaborative effort initiated by the IFIP 2.6 Working Group on Databases. It was partly carried out as part of the European project KnowlegeWeb No 507482. A previous version of this work was published in the proceedings of DASFAA04.
** Corresponding author. Phone: +41-21-693 6787.
E-mail address: Philippe.Cudre-Mauroux@epfl.ch
*** The research of this author is partially supported by NSF grant IIS-0326284

stored in the enterprise database. Nowadays, all major economic players have decentralized organizational structures, with multiple units acting in parallel and with significant autonomy. Their information systems have to handle a variety of information sources, from proprietary ones to information publicly available in web services worldwide. Grasping relevant information wherever it may be and exchanging information with all potential partners has become an essential challenge for enterprise survival. Shortly stated, information sharing, rather than information processing, is IT's primary goal in the 21st century. Not that it is a new concern. It has been there since data has been made processable by a computer. What is (relatively) new is the focus on semantics, which takes the issue far beyond the syntactic functionality provided by exchange standards or standard formatting à la XML. The reason that makes semantics re-emerge so strongly is that now information has to be sharable in an open environment, where interacting agents do not necessarily share a common understanding of the world at hand, as used to be the case in traditional enterprise information systems.

Lack of common background generates the need for explicit guidance in understanding the exact meaning of the data, i.e., its semantics. Hence the current uprising of research in ontologies, for instance. Ontologies are the most recent form of data dictionaries whose purpose is to explain how concepts and terms relevant to a given domain should be understood. However, ontologies are not the panacea for data integration [1]. Consider a simple example from traditional data management: an enterprise database will most likely contain data about employees, and every user will be expected to understand the concept of "an employee". Yet a closer look at the concept reveals a number of possible ambiguities, including whether specific types of personnel (e.g., students in their summer jobs, trainees, visitors) have to be considered as employees. Without an agreement between the interacting units as to the correct mapping between these concepts, interpretation may not be decidable.

Ontologies are forms of "a-priori" agreements on concepts, and therefore, their use is insufficient in ad-hoc and dynamic situations where the interacting parties did not anticipate all the interpretations and where "on-the-fly" integration must be performed [2]. In fact, the commensurability of knowledge and the desirability of developing efficient solutions for the open environment preclude an approach which realistically captures the space of interpretations in a finite structure. Semantic errors compound even intuitively well understood concepts. In the absence of complete definitions, elicitation of explicit and goal-driven contextual information is required for disambiguation. In human conversations, the context may be implicit, elicited through a dialogue between the interlocutors, or gathered from additional information sources. The new computing environment in the Internet demands similar capabilities. Increasingly, information systems are represented by agents in their interactions with other autonomous systems. These agents must therefore be capable of building the context within which "on-the-fly" integration could occur. What ought then be the appropriate mechanisms and tools that agents must possess to accomplish the task of resolving

semantic conflicts in a dynamically changing environment, such as the Internet and the Web?

The above discussion serves as a motivation for the general principles enunciated thereafter which could drive the development of the next generation of semantic reconciliation methods. The rest of this paper is organized as follows: We first take a look back at classical data integration techniques in Section 2 and summarize the rationales and principles of the new *emergent semantics* trend in Section 3. Section 4 gives some details on an important aspect of emergent semantics systems, namely *self organization*. Opportunities and challenges related to emergent semantics systems are outlined in Section 5 and 6. Finally, we present three case studies in Section 7 before concluding.

2 Classical Information Integration

The need to integrate heterogeneous information sources is not new; For decades, researchers have been working on building data integration systems providing uniform query interfaces to a multitude of data sources, thereby freeing the user from the tedious task of interacting and combining data from the individual sources. Given a user query that is formulated in the query interface (also called the mediated schema), these systems use a set of semantic mappings to translate the query into queries over source schemas, then execute the queries and combines the data returned from the sources, to produce the desired answers to the user. Numerous research activities have been conducted along those lines, both in the artificial intelligence and database communities. Much progress has been made in terms of developing conceptual and algorithmic frameworks; query optimization; constructing semi-automatic tools for schema matching, wrapper construction, and object matching; and fielding data integration systems on the Internet.

2.1 Information Integration from a Database Perspective

The motivation for data integration from a database perspective is old and reflects the activities from the 90s, when various databases were integrated. Most of the databases integration systems described in the literature (see, e.g., [3–7]) are based on a unified view of data, called mediated or global schema, and on a software module, called mediator that collects and combines data extracted from the sources, according to the structure of the mediated schema. The existing mediator-based information integration systems can be distinguished according to: 1) the type of mappings between the mediated schema and the schemas of the sources (Global As View versus Local As View), and 2) the languages (therefore, the expressivity) used for modeling the mediated schema and the source descriptions.

2.2 Global As View versus Local As View

According to [8], information integration systems can be related to two main approaches for modeling inter-schemas correspondence: Global As View (GAV) and Local As View (LAV). The GAV approach has been the first one to be proposed and comes from the Federated Databases world. The mediated schema is defined in function of the schemas of the sources to integrate, i.e., each relation of the mediated schema is defined as a view on the relations of the sources schemas. The advantage of this approach is the simplicity of query reformulation which simply consists of replacing each atom of the query by its definition in terms of the relations of the sources schemas. Its drawback is its lack of flexibility with respect to the addition or deletion of sources to the mediator: adding (or deleting) a source to the mediator may affect the definitions of all the relations of the mediated schema. The LAV approach is dual and has opposite advantages and drawbacks. It consists of describing the contents of the sources in function of the mediated schema. In such an approach, adding a new source is quite straightforward because each source is described independently of each other. The price to pay for this flexibility is the difficulty of the query answering processing which requires a more complex process of rewriting queries using views (see [9] and [10] for more details on the problem of answering queries using extensions of views).

2.3 Relational versus object-based mediated schema

The selection of the language used to modeling the mediated schema and the source descriptions is a very important aspect: the expressitivy of such a language will restrict the kind of semantic relationships that can be described among data elements. We can distinguish between those approaches providing a relational view of data sources and those systems using an approach based on objects. The most representative information integration systems of the relational approach are: Razor [11], Internet Softbot [12], Infomaster [13] and Information Manifold [14]. They all follow a LAV approach. The Razor and Internet Softbot systems use datalog (without recursion) for modeling the mediated schema, the views describing the sources contents and the users queries. Infomaster and Information Manifold are based on extensions of datalog. Infomaster exploits integrity constraints in addition of datalog rules. Information Manifold extends datalog by allowing that some predicates used in the rules are concepts defined by using description logics constructors. The most representative information integration systems of the object-based approach are: TSIMMIS [6], SIMS [15, 16], OBSERVER [17] and MOMIS [18]. TSIMMIS is based on the object-oriented language OEM for describing the mediated schema and the views, and on the OEM-QL query language. It follows a GAV approach. The SIMS and OBSERVER systems use a description logic for modeling the mediated schema, the views and the queries. SIMS follows the LAV approach while OBSERVER follows the GAV. The MOMIS system is based on the use of a very expressive description logic (ODL-I3) for describing the schemas of the sources to integrate. It follows the GAV approach.

2.4 Information Integration from Knowledge Representation Perspective

The schematic and semantic heterogeneity is one of the difficulties in the integration of heterogeneous information sources. Normally, the information in every information source is stored with regards to its users' requirements, disregarding access from other sites or their integration. Two critical factors for the design and maintenance of applications requiring Information Integration are conceptual modeling of the domain, and reasoning support over the conceptual representation. It has been demonstrated that knowledge representation and reasoning techniques can play an important role for both of these factors. Two relevant works that follow the knowledge representation approach are from Calvanese et al. [19] and Rousset and Reynaud [20].

2.5 Information Integration and the World Wide Web

With the progress in global interconnectivity, the scale of the problem has changed from a few databases to an open and dynamic environment of millions of heterogeneous information resources. Current keyword-based approaches are usually found to provide a poor quality of result. However, the key challenges to be faced are at the semantic level, where people would increasingly expect the information systems to help them not at the data level, but at the information, and increasingly knowledge levels which call for semantic interoperability. In essence, we need an approach that reduces the problem of knowing the contents and structure of many information resources to the problem of knowing the contents of easily-understood, domain-specific ontologies, which a user familiar with the domain is likely to know or understand easily. Recent papers focused on some of the issues characterizing data integration over the Web. For example [21] identifies the problem of similarity matching among XML data. The proposed algorithm is able to find the commonalities and differences which give rise to a numerical rank of the structural similarity. [22] investigates the heterogeneity problem of information sources from the query answering point of view. To handle semantics inconsistencies between the same attributes used at different sites, task ontologies are used as a communication bridge between them. Information Retrieval is also a well established domain [23] that now has a new field of application: Web-based IR. [24] isolates four different approaches of this problem in the Web context: (1) human or manual indexing, (2) automatic indexing using classical IR techniques, (3) intelligent or agent-based indexing such as using Crawlers and Robots and (4) meta-data, RDF (Resource Description Framework) and annotation-based indexing. Meta information (and hence semantic Web) can also be important in the prospect of information integration: [25] proposes web servers export specific meta-data archives describing their content. In order to offer better processing and integration of information, a unified representation for Web resources (data and services) is becoming a necessity. The use of ontologies to provide the means to machines of understanding the data they are manipulating is increasing. With the emergence of Semantic Web [26],

the study of ontologies and their uses has increased, since they provide a shared and common understanding of a domain that can be communicated between people and application systems.

2.6 The future of data integration

To date, all the integration technologies have been offered by various independent vendors and products. Even if the information technology organizations can appropriately match the right integration technology to the problems related to information integration, it still has the problem of how many skilled, specialist resources are needed to implement these technologies and integration scenarios, as well as how many different vendors must be contracted. Each of these separate technologies has its own user interface to the developer, brings its own development environment (often graphical), has its own meta data repository to document the interface, its own security framework, and its own management framework. Developing intelligent tools for the integration of information extracted from multiple heterogeneous sources is a challenging issue to effectively exploit the numerous and dynamic sources available on-line in global information systems.

3 The Emergence of Emergent Semantics

Nowadays, several research areas such as peer-to-peer data management, information agents, Semantic Web or Web data mining and retrieval all address the problem of semantic interoperability in widely distributed information systems with large numbers of agents[1] [27, 28] using remarkably similar ideas. Global information is seen as highly evolutionary: documents of already existing sources may be updated, added or deleted; new sources and services may appear and some may disappear (definitively or not). Semantic interoperability is viewed as an emergent phenomenon constructed incrementally, and its state at any given point in time depends on the frequency, the quality and the efficiency with which negotiations can be conducted to reach agreements on common interpretations within the context of a given task. We refer to this type of semantic interoperability as "emergent semantics". In the following we outline what we believe are the key characteristics underlying this concept.

3.1 Principle 1: Agreements as a Semantic Handshake Protocol

Meaningful exchanges can only occur on the basis of mutually accepted propositions [29]. The set of mutual beliefs constitutes the "agreement" or "consensus" between the interacting agents. It is the semantic handshake upon which shared

[1] the term "agents" refers to both humans through computed-mediated communication and to artificial surrogates acting as information and/or service consumers and producers. The term "peers" is used as a synonym.

emerging and dynamic ontologies can be established and exchange context can be constructed. In practice, the agreement can be over the real-world meaning of some model, as it is typically assumed in conceptual modeling, on schema mappings, on consistent data usage or on any other meta-data information relevant to the task at hand. The strength of the agreement will depend on the strength of the accepted propositions, their quality and trustworthiness.

3.2 Principle 2: Dynamic agreements emerge from negotiations

Information exchange between agents is necessary to negotiate new agreements or to verify preexisting ones. This is a recognition that the information environment is dynamic, and thus, assumptions must be constantly validated. Agreements evolve as agents learn more about each other and as interests broaden or become more focused. Interaction is required to identify and resolve semantic conflicts, to negotiate and establish consensus on the data interpretation, and to verify whether a consensus leads to the expected actions. Communication can be realized in terms of explicit message exchanges or implicitly by reference to distributed information resources.

Note that in our context, agreements are no longer "static". Unlike ontological commitments, these agreements are likely to change dynamically as the network of information parties evolves. Also, agreements do not need to be "binary", i.e., it is not the case that either there is or isn't consensus about the meaning of a particular term. In fact there might be fuzzy notions of consensus, such as weak, strong, etc. which may have a bearing on the semantics. Finally, agreements do not necessarily result from negotiations of "equals". In general it is assumed that (in the committee approach) all the people contributing to creation of ontology/enterprise model have equal expertise. This is definitely not the case in an emergent semantics scenario, where different people from a variety of backgrounds will be contributing to the negotiations, and thus to the agreements generated.

3.3 Principle 3: Agreements emerge from local interactions

The principles stated so far are analogous to those formulated for introducing the concept of ontological commitments [30], except that "emergent semantics" assumes that commitments are dynamic and are established incrementally. The key challenge for emergent semantics remains scalability. The complexity of "emergent semantics" and communication costs preclude the option for an agent to seek agreements simultaneously with a large number of other agents. The combinatorial nature of such an endeavor will limit the viability of the approach in distributed environment. Thus, pragmatics dictate that "emergent semantics" be kept local to reduce communication costs and that global agreements are obtained through aggregations of local agreements. As a result, even if agents are only aware of a small fraction of a network directly, they will nevertheless be able to interoperate over the whole network indirectly by exploiting aggregate

information. This raises the immediate question on how to technically perform aggregation and inference of new agreements.

3.4 Principle 4: Agreements are dynamic and self-referential approximations

Making an appeal to context in resolving semantic conflicts is a recognition that traditional schema or conceptual analysis leave open several possible interpretations of a mapping between the information sources of two interacting agents. However, the problem with context in general is that the space of possibilities is very rich, and that it has no well defined boundary. Since agreements rely on the context of interaction, their boundaries are also fuzzy. The way out of this conundrum may lie in the fact that we view "emergent semantics" as an incremental and goal or query-directed process which sufficiently constrains the space of possibilities.

Two interacting agents may achieve an agreement in one application and fail in another even if the set of identified semantic conflicts are basically the same. Interpretations may depend on the context. In turn, agreements are dynamic. Local consensus will be influenced by the existing context of existing global agreement, thus the process of establishing agreements is self-referential.

3.5 Principle 5: Agreements induce semantic self-organization

Considering the dynamics and self-referential nature of emergent semantics, it is not far-fetched to view it as the result of a self-organization process. Self-organization is a principle that has been studied in many disciplines, in particular physics, biology, and cybernetics for a long time, and has been attracting substantial attention in computer science as well (see Section 4. Informally, self-organization can be characterized by a complete distribution of control (which corresponds to complete decentralization) and by the restriction to local interactions, information and decisions. Global structures can then emerge from such local interactions.

Francis Heylighen characterized self-organizations as follows: "The basic mechanism underlying self-organization is the noise-driven variation which explores different regions in a system's state space until it enters an attractor." In the case of emergent semantics, the state space consists of all local communication states reached in consensus building. The attractor is obtained when agents locally reach acceptable agreements that are as consistent as possible with the information they receive. The attractor actually embodies what we call the global semantic agreement. The noise-driven variation results from randomness of interactions induced by environmental influence (e.g., network connectivity) and autonomous decisions.

3.6 A canonical and well-known example

We illustrate the principles of emergent semantics by referring to one particularly successful example of emergent semantics, namely link-based ranking as used in

Google [31]. A global semantic agreement is obtained for a simple property of Web documents, namely their "general importance". The local communication is performed by Web document authors implicitly by referring to other Web documents through hyperlinks. The global agreement is determined by using the PageRank algorithm.[13] It provides a ranking of all Web documents. This ranking is approximate, as surely not all authors would agree on it, it is dynamic as the Web evolves, and it is self-referential as the impact of a link is derived from the importance of its Web document. The self-referential nature of link-based ranking actually leads to problems when link spammers exploit their knowledge on the Google ranking method in order to influence the rankings in their interests.

We can compare Google's approach to the one taken by Web directories, such as Yahoo. In web directories the decision on importance of Web documents with respect to some ontological concept (the directory) is taken globally, manually and centrally. This clearly limits the scalability of the approach.

Other approaches similar to Web document ranking are currently appearing in other Web applications. For example, several works on trust and reputation mechanisms relies on similar principles as Google's ranking approach. A practical application of reputation-based trust management is found with Ebay. More complex tasks, such as Web document classification and clustering based on emergent semantics principles are reported in the literature.

3.7 Extending the scope of emergent semantics

A next natural step beyond ranking-based methods ignoring the structure of the content would be to apply the principle of emergent semantics to obtain interpretations for structured data. The Semantic Web is currently laying foundations for the use of semantically richer data on the Web, mainly through the use of ontologies for meta-data provisioning. The effort of establishing semantic agreement is largely related to the development of shared ontologies. The question we pose is whether principles of emergent semantics could be a solution for obtaining semantic agreement in the Semantic Web with its richer data models in a more scalable fashion.

One possible avenue of how this might be achieved is currently being opened in the area of peer-to-peer data management, where local schema mappings are introduced in order to enable semantic interoperability. We may see such local schema mappings as the local communication mechanisms for establishing consensus on the interpretation of data. Once such infrastructures are in place, the principles of emergent semantics become directly applicable. Relying on local consensus, automated methods may then be employed in order to infer more expressive and accurate global semantic agreements.

[13] The fact that the ranking is computed on a central server is from the perspective of establishing a semantic agreement just an implementation issue.

4 Self Organizing Systems in Nature and Computer Science

As stated above, self organization is an essential property of emergent semantics systems. A self organizing system essentially consists of a system that evolves towards displaying global system behaviours and structures that are more than an aggregation of the properties of its component parts. Such systems generally have complex dynamic characteristics that allow them to evolve from a given state towards attractors, which exhibit stable patterns in structure and behaviour. There can be more than one attractor, in a given self organizing system each having its region of attraction. An important element of these self organizing systems is that there is no external influence or central controller that leads to these patterns. Rather these patterns are arrived at through interactions between components such that these components only have local information, knowledge or local rules. The collection of information arising from local rules and knowledge leads to the emergent properties of the global system as a whole.

Several examples can be found in science and nature of self organizing systems. A few examples are discussed below: Two examples from Physics are [32] a) magnetization and b) Bernard Rolls. In magnetization, spins (the equivalent of tiny magnets) are randomly changing orientation at high temperatures. At a lowered temperature these align themselves to reinforce each other's magnetization leading to a magnetized bar. Bernard rolls are circular movements of liquid flows, which result from the heated liquid moving from a hot bottom of the container to the top whilst the cooler liquid from the surface moves down setting up a circular pattern. In each case the particles are reacting locally without an external or central controller but their interactions lead to the stable patterns at the global system level. Examples from biology or chemistry [33] include the stripped patterns in Zebras, Fish and the ocular dominance columns of the brain. These patterns are produced due to the individual responses of the cells to local conditions and the response of the neighbouring cells. Again there is no central controller involved and the patterns are an emergent property of the collection of cells. An example from nature consists of a flock of birds that flies in a certain formation. These birds develop into this formation and preserve it despite several changes in direction and environmental factors. Again there is no leader or controller that marshals the birds into these formations

There are several different examples of self organizing systems in computer science and related disciplines. Specifically one can distinguish: a) Self Organizing Neural Networks such as Kohonen Nets and Edelman Nets b) Hopfield Neural Nets and Boltzmann Nets c) Particle Swarms d) Evolutionary Computation e) Cellular Automata f) Peer to Peer Networking. The self-organizing feature map, also known as Kohonen network [34], was developed by Teuvo Kohonen. The Kohonen network has the ability to find clusters in the data as well as structure and to perform an ordered or topology-preserving mapping, thus revealing existing similarities in the inputs. The topology preserved with this network need not correspond to a physical arrangement; it can correspond to a statistical feature of the input set. In a typical Kohonen network, units are

arranged in a two-dimensional grid. However, it is possible to use one or more dimensions. This grid of units is usually referred to as a Kohonen layer. All units in the input layer are fully connected to the units in the Kohonen layer of Feature Map. Feedback is restricted to lateral interconnections with immediately neighbouring units in the Kohonen layer. Each link between an input and a Kohonen layer node has an associate weight. The net input into each neuron in the Kohonen layer is equal to the weighted sum of the inputs. Learning proceeds by modifying these weights from an assumed initial distribution with the presentation of each input pattern vector. A central aspect of a Kohonen network is that it uses competitive learning. As learning involves adjustment of weight vectors, the importance of determining the winner resides in the fact that only the neurons within a small region around the winner are allowed to learn this particular input pattern. There are two types of Hopfield networks [35], the first is a discrete output, stochastic network whilst the second has a deterministic continuous form. A key element of the Hopfield network is feedback. This essentially means that the weighted output from every neuron in the network is fed back to the input of each neuron. Another important element is the network's updating technique, i.e., for a modified input, when does the network change its output? In the discrete stochastic network, each neuron samples its input at random times. Furthermore the sampling times of each neuron are independent of every other neuron, i.e., the system is asynchronous. The motion of the state of the system with N neurons in state space describes the computation being performed. Any model must describe evolution of the state with time. Hopfield proposed a model with stochastic evolution. In analogy to spin glass models, Hopfield introduced the generalized energy function. Therefore the iteration must lead to stable states. The energy function can be visualized as a surface. The stable states correspond to the local minima on this surface. Each stable state can be considered as an attractor with its own basin of attraction. A Hopfield net can converge to a local minimum that is not an optimal solution. A process described by metaphor from metal annealing can be applied using a net algorithm similar to Hopfield's to encourage convergence to a global minimum. The network, a Boltzmann machine, has a probabilistic update rule which sometimes allows jumps into higher energy states rather than lower ones as a strategy to escape local minima. The thermal noise added to the network is initially high and it is slowly decreased to encourage thermal equilibrium in the net.

Particle Swarm [36] and evolutionary computation are both adaptive optimization techniques. Particle Swarm is inspired by bird flocking and fish schooling phenomena whilst evolutionary algorithms are inspired by genetic evolution. Particle Swarm and Evolutionary Computation both start from an initial randomly generated population, have a fitness function which represents closeness to the optimal volume and carries out selection from the current population based on the best fitness value until a stopping criteria is met. In evolutionary computation and genetic algorithms, these operators namely reproduction, crossover and mutation are utilized to generate new candidates for the population. Note that the whole retained population is used to search for the optimum. In contrast

in particle swarms each particle's velocity is updated by the best historical value and the current global best value.

Cellular automata consist of a two dimensional arrangement of cells. Each cell interacts with cells in proximity and has transition rates that determine its state. Camazine et al. [33] have shown how cellular automata can be used to model stripped and mottled patterns that appear in animals.

Peer to Peer networks and their self organizing properties are discussed later in this paper.

5 Opportunities

Despite the specificities mentioned in Section 3, Emergent Semantics can still benefit from the heuristics and results of many different research fields. We detail below how dynamics in decentralized settings, data mining or lexical resources can all in their own ways help building Emergent Semantics systems.

5.1 Dynamics in Decentralized Settings

Semantics do not emerge from purely random settings, but rather from environments exhibiting specific, well-known properties. Locality has been referred to (Section 3) as an essential component of emergent systems. Semantic networks – as many social or natural networks – build up from large numbers of purely local, pair-wise interactions. *Scale-free* networks [37] have been designed specifically for studying systems resulting from such a construction process. These networks differ from random networks in the sense that they first start from a small nucleus of nodes, and expand then with the arrival of new nodes that join the network following some preferential attachment law. We can expect semantic networks to expand following a similar process, where new nodes connect to already existing nodes because of some semantic affinity. Results from *scale-free* graph theory range from network formation to statistical connectivity and could be directly applied to model the shaping of semantic networks as well as to highlight some of their essential attributes, like scalability which is one of the inherent properties of such graphs.

Also, locality may be seen as a real opportunity to leverage investments while establishing semantic interoperability. This is important both in cases where communication used to establish semantic agreement requires human intervention or when it is automated. When human intervention is required, it is instrumental to minimize it, as human attention is one of the scarcest resources today [38]. On the other hand, automated methods to locally establish semantic interoperability (e.g., schema matching or natural language translations) are computationally very intensive and would directly benefit from decentralization and from localized view on global agreements.

The fact that no central component is at hand for coordinating the various interactions in the semantic system imposes some autonomous behaviors on its constituents. Autonomy has been studied in bio-inspired [39] and decentralized

peer-to-peer [40, 41] approaches, which are particularly good at decomposing large or complex problems otherwise hard to tackle using standard centralized solutions. Autonomy also directly refers to intelligent and multi-agent systems [42] in general, where coordination and distributed problem planning/solving are tackled using distributed artificial intelligent techniques.

Randomness clearly induces a certain loss of efficiency but leads to a higher failure resilience and robustness of the system. This relates to the dynamics of decentralized environments and to the fact that a large fraction of nodes may be faulty or off-line at any given point of time in such settings. Built-in load-balancing and replication algorithms [43] usually handle the problem from a data-availability point of view, while overall connectivity is typically not at stake, as long as a reasonable fraction of preferred (i.e., highly connected, cf. above) nodes still function properly in the system.

Naturally, locality, autonomy and randomness may all be seen as harmful to different degrees to the global integrity and completeness of the system. Even if algorithms have been devised for taking care of data availability and integrity in highly dynamic environments [44], global semantic integrity in heterogeneous environments remains for the time being a challenging research problem. The lack of any agreed-upon global schema or ontology makes it very difficult for the participating parties to reach a global consensus on semantic data. Initial approaches rely on some pre-defined corpus of terms serving as an initial context for defining new concepts [45] or make use of gossiping and local translation mappings to incrementally foster interoperability in the large [46, 47].

5.2 Data Mining

Once some initial level of semantic agreement has been reached, individual entities can make use of data mining techniques to refine the agreements. *Data Mining for Emergent Semantics* aims at enhancing semantic interoperability by exploiting:

1. content data (of texts, multimedia, relational table);
2. structural data (e.g. links between texts, coordination between multimedia objects, multirelational structures, network data);
3. usage data (e.g. usage of texts, video, data).

In addition, to overcome data sparseness, which is often a problem for achieving semantic descriptions from data mining, there is the possibility to

4 *actively collect data* (also cf. Active Learning [48]).

For instance, data mining based on Web resources to achieve emergent semantics uses globally available Web data and structures to define new local semantics. Blueprints for this paradigm are found in works such as the following:

1. Web Content Mining: Some researchers use explicit, linguistically motivated natural-language descriptions to propose semantic relationships ([49–52]).

2. Web Structure Mining: In [53, 54], the Web structure itself is used to determine a focus for harvesting data. Thus, specialized semantic relationships, such as recommendations coming from a particular Web community can be derived.
3. Web Usage Mining: [55]
4. Active Learning: Others use the Web to cope with data sparseness problems in tasks that require statistics about possible semantic relationships ([56–59]).

Currently, people work on combinations, e.g. content and structure mining [60] or content mining and active learning [61].

Particularly relevant work in this area of global approaches of emergent semantics are the areas of 'traditional' Ontology Learning — mostly, though not only, from texts (see [51, 62]).

Other techniques for bilateral semantic alignment can also be used. The basic idea of bilateral semantic interoperation is to generate an alignment between two different semantic descriptions based on a number of heuristics (see in particular the survey [63]). These generations can be based on identity of lexical labels, agreements on common semantic structures, indirect mapping via thesauri or overlap of extensional descriptions such as found through machine learning. Multi-lateral consensus tries to generalize from bilateral semantic interoperation. Some of the basic ideas here include the composition of individual bilateral agreements — be it by forwarding through gossiping [46, 47] or by more centrally directed algorithms [64–66].

5.3 The influence of usability perspectives on locally axiomatized semantics

We use the term semantic axiomatization in order to refer to a formal description accounting for the intended meaning of a vocabulary, represented in a machine-processable manner[14]. Same semantics can be axiomatized in different ways. This usually reflects different usability perspectives, such as granularity, scope, representation primitives and constructs, reasoning and computational scenarios, and so forth. In other words, local semantic axiomatizations are substantially influenced by "what the semantic is being axiomatized for" and "how it will be used". Bylander and Chandrasekaran argued in [68] 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." We believe that establishing formal semantic interoperability among different *local* semantic axiomatizations mostly fails due to the diversity of usability perspectives, although all axiomatizations might *intuitively* agree at the domain/knowledge level[15].

[14] This definition is derived from Guarino's definition of the term ontology as found in [67]

[15] See [69] for the definition of "knowledge level"

Intuitive definitions and agreements about the intended meaning of certain vocabularies are implicit assumptions shared among human cognitive agents. Such informal definitions and agreements can be found in lexical resources (e.g., dictionaries, lexicons, glossaries, lexical databases). Linking or grounding the vocabulary used in local axiomatizations with terms found in lexical resources can help achieving basic semantic interoperability between different axiomatizations. For example, by using (euro)WordNet synsets [70] as a shared vocabulary space, autonomous semantic axiomatizations will be able to interoperate at least freely from language ambiguity and multilingualism.

Using lexical resources as shared vocabulary spaces could be seen as an attachment law of emergent semantics networks; or, it could be advised in case of failure or uncertain semantic interoperations. The basic (or maybe the only) requirement for a lexical resource to be used as such, is that it should provide (1) a discrimination of word/term meaning(s) (2) in a machine-referable manner. Lexical resources that only list vocabularies and their similarities are irrelevant to our purposes. Semantic or linguistic relationships between word forms (such as hyponymy, meronymy, and synonymy) could be significant but not essential. Our basic target is to enable emergent semantics networks to communalize word/term senses, which are largely independent of usability perspectives.

In comparison of using lexical resources with the use of axiomatized domain theories (i.e., ontologies), building adequate ontologies is difficult and very expensive, while many reliable and comprehensive lexical resources are available. Further, lexical resources are generally easier to extend than ontologies.

As a related work, Jarrar et al proposed in [71, 72] an ontology engineering approach that uses the notion of "ontology base" as a controlled vocabulary space shared between application axiomatizations. An ontology base is intended to capture context-specific domain vocabularies, i.e., lexical rendering of domain concepts.

6 Threats and Limitations

In this section, we investigate characteristics and problems of emergent semantics systems from two distinct points of view. We first illustrate which are the threats and limitations strictly inherent to emergent semantics systems. Then, we describe issues related to decentralized and peer-to-peer architectures and how those issues can influence emergent semantics systems. Table 1 summarizes the results of our analysis.

6.1 Emergent Semantics Systems: Threats and Limitations

Representational Model First, there is the need to commit to a particular representational model, i.e., a relational data model like the relational algebra, a semi-structured data model like Lore [73], a semi-structured data model like

	Threats and Limitations
ES Systems	Representational Model Common Upper Ontology & Extensibility Data Integration & Querying Provenance Information & Trust Incompleteness Consensus Derivation
P2P Infrastructures	Degree of Centralization Degree of Inter-Peer Coupling Data Availability & Updates Anonymous Entities

Table 1. Emergent semantics systems and P2P Infrastructures: threats and limitations

RDF [74] with its schema language RDFS [75], or a full-blown knowledge representation language like OWL (Web Ontology Language) [76]. The trade-off along these lines is one between expressiveness and efficiency. While on the one end the relational algebra is a model for which highly efficient systems exist, it will hardly be sufficient to prescribe semantic definitions. At the other end, OWL allows for comprehensive definitions, including e.g. cardinalities and arbitrary Boolean expressions for defining classes — but currently there is no system on the horizon that efficiently handles more than several dozen of tuples at the instance level. Furthermore, there are currently no algorithms that would infer complex constraints from observed data with reasonable accuracy. Thus, representational models like RDF(S) currently appear to constitute the appropriate paradigm for defining some semantics as well as handling reasonably sized data stores.

Common Upper Ontology & Extensibility Emergent semantics systems can make use of text mining and existing lexical information to incrementally come up with a consensus on the data they share (cf. Section 5). However, some common understanding is usually necessary to bootstrap the process, thus the need to agree on some upper-ontology (e.g., [77]).

Furthermore, once mining has yielded new conceptual structures, the results should be added in an appropriate way to the existing background information for later re-use. This second step requires extensible representations. In addition, to counter the need for integrating multiple mining results, the second step requires an agreement on how the conceptual structures are stored (according to the representational paradigm) and how the very particular lexicon structures are named (according to the upper ontology or a vocabulary of meta metadata for emergent semantics systems).

For instance, WordNet [70] allows for one concept to be referred to by several lexical terms (e.g. the lexical terms 'hard' and 'difficult' may refer to one concept) and it allows for on lexical term to refer to multiple concepts (e.g. 'hard' may refer to the concepts for 'difficult' and for 'non-soft'). There exist first considerations

to provide a data model to this end (cf. [78]), but no final conclusion exists yet in this matter.

Data Integration & Querying In an emergent semantics system, different entities may have different knowledge levels on other parties schemas and mappings. The problem of defining how to answer to a query posed on the schema of a specific entity arises.

One approach to solve this problem is the one used in traditional data integration systems [79]: a global schema is constructed starting from the schemas locally exported by the different data sources (see also Section 2). The assumptions in the case of the emergent semantics paradigm are completely different: no global schema is a-priori constructed in order to make the system work, instead it is an inherent function of the system to construct a global knowledge "dynamically" while working. In this case, mappings cannot be defined with respect to a global schema, therefore the research problem of mappings definition and resolution arises. Semantic Gossiping [46, 47] could be a promising approach to reach semantic interoperability in a network of semantically heterogeneous parties.

Provenance Information & Trust Provenance information may be important in order to cluster or categorize data according to where they came from. Results could be that particular quality/trust ratings are given for particular provenances or that semantic structures are treated individually based on where they came from. Such pieces of information are particularly difficult to gather and verify in open and dynamic environments such as emergent semantics systems. Ehrig et al. [80] present a quite specific metadata model based on RDF(S) to this end. Siebes and van Harmelen[81] and Tempich et al.[82] are examples on how to exploit such models for negotiating meanings and routing semantic queries, respectively.

It is also on provenance information that one can build trust mechanisms or ratings for the various entities in the system. Also, mechanisms should be developed in order to check mappings and results received from other peers; Misbehaving peers could populate the community with erroneous mappings or bogus schemas and could answer queries with fake data. Such situations must be detected and actions must be taken to exclude malicious peers and remove fake data from the system. Coming up with good heuristics for solving these issues is especially complex given the dynamics of emergent semantics infrastructures (cf. also Section 6.2 below).

Incompleteness Incompleteness in an emergent semantics context is related to the impossibility to obtain all the information available in the system due to a lack of knowledge on the information that peers commonly share and to a lack of global semantic interoperability. In a traditional data integration system, with a global schema summarizing all the available information, the incompleteness

problem does not occur in these terms: It is usually known *a priori* which pieces of information can be provided by whom. The absence of complete indexes on resources in emergent semantics systems and the presence of replicated copies of the same semantic information, could cause system inefficiency. Therefore, on one hand we have to assure a high level of completeness in information searching; on the other hand, it is also desirable to avoid network request flooding. The adoption of specific semantic query models need to be investigated in order to consider possible tradeoff among search strategies, optimal request load balancing and system robustness to failures.

Consensus Derivation Related to the Incompleteness problem presented above, Consensus Derivation can be considered as a key component when deriving semantics that emerge from the interactions of people and from the various messages generated to express their opinions. Given a set of observations by a set of people, requirements on a consensus computation scheme could be:

- The ability to compute the consensus semantics or reality based on an analysis and aggregation of the individual events observed.
- Based on the computed reality, estimate the individual expertise of the people involved based on how close their opinions corresponded to the central reality.
- Update the consensus and associated expertise estimates whenever current observations change and new observations are added to the mix.

Work done in cultural anthropology and approaches such as Delphi methods and Repertory Grids need to be explored to come up with effective algorithms for consensus derivation. Besides semantics, consensus computation might also have some impact on other issues such as trust, quality and assessment of satisfaction. All these issues can in turn influence the computation of new consensus, thus outlining once more the self-referential property of agreements in emergent semantics systems.

6.2 Peer-to-Peer Systems as Infrastructure for Emergent Semantics Systems: Threats and Limitations

As stated above (Section 5), we expect emergent semantics properties to appear in large-scale, decentralized and dynamic environments. Thus, P2P systems represent a natural infrastructure on which to base emergent semantics systems. By *P2P*, we do not only consider the well-known file-sharing applications, but also all the access structures and distributed systems where participating nodes can be both clients and servers. In other words, all nodes provide access to some of the resources they own, enabling a basic form of interoperability. Below, we illustrate some peculiarities and issues of P2P infrastructures and we analyze how they could influence emergent semantics systems.

Degree of Centralization A first architectural problem in P2P systems is related to the *degree of decentralization*: decentralized, centralized and hierarchical models are all possible [83]. The topology of centralized systems causes the well-known problems of bottlenecks and single points of failure. On the other hand, fully decentralized systems are difficult to implement and their performances are relatively low. This is also proven by the fact that many P2P systems are built with an hybrid approach (such as Napster, KaZaA, or eDonkey). Also, P2P softwares should not require any significant set up or configuration of either networks or devices [84]. Though much progress has been made in designing P2P systems, such constraints still complicate the implementation of "actual" emergent semantics systems.

Degree of Inter-Peer Coupling The degree of inter-peer coupling takes into account how much *tight* can a peer interaction be. For example, with systems such as Kazaa, the interaction is not tight since users only search for data and establish temporary connections. On the other hand, with distributed workflow systems, each node can have significantly more sophisticated and longer interactions with other nodes, thus originating tighter interactions. We can expect some applications of emergent semantics systems to require such tight interactions, thus the necessity for the system infrastructure to support various inter-peer coupling models.

Data Availability & Updates Even if an efficient indexing mechanism is developed, in many cases data can be unavailable, simply because the peers storing such data are offline or unreachable. In order to achieve better data availability, peers should replicate their own data in the community. The replication could be controlled by the originator or the data could be replicated through gossiping mechanisms by other peers. Very popular data might need to be highly replicated. It is also possible to exploit standard fault tolerance techniques, such as software replication [85] in order to enhance data availability and reliability.

Introducing replication makes updates more complicated, because of the necessity to update replicas as well. Some approaches already exist that work under probabilistic guaranties [44].

Anonymous Entities Anonymity is often associated to P2P systems, because of their open nature and the lack of any central authority. Anonymity can be defined with respect to a communicating pair in the P2P system. Specifically, three kinds of anonymity are possible: *sender anonymity*, which hides the sender's identity; *receiver anonymity*, which hides the receiver's identity; and *mutual anonymity*, in which the identities of the sender and the receiver are hidden to each other and to other peers.

From a system perspective, the major drawback of peer anonymity is the limitation in implementing security controls in upper layers. In an emergent semantics context, retrieving the peers' identity might be essential to enable trust mechanisms or counter malicious attacks (see above Section 6.1).

7 Examples/Cases studies evaluation

In this section we present three possible application scenarios for the concept of emergent semantics. The case of Service Discovery shows how emergent semantics could help to improve data freshness and quality of the discovery process. The second example from the digital library area indicates in which way emergent semantics can support the integrated access on heterogeneous libraries. Elicitation of interpretation semantics in scientific collaborations is presented in the last example.

7.1 Semantic Service Discovery

Semantic Web Services combine Web Services technology with machine-understandable meta data annotation emerging in Semantic Web research. Just as the WWW moves towards offering dynamic content and Web Services instead of static content alone, part of the Semantic Web vision is to establish a network of semantically annotated services. In such a network agents are able to combine the functionality of several Web Services in order to achieve complex high-level goals in an automated way without human intervention.

This fully automated scenario involves discovery, composition and execution of Web Services [86] and requires formal descriptions of service semantics for software agents to reason about. Discovery includes the task of locating Web Services that provide certain capabilities and fulfil the constraints specified by the requestor. Composition comprises the combination of several services to a more complex one [87]. Execution involves the invocation of an identified service by an agent including proper message exchange with the service's interface [88].

The usage of Web Services involves a requesting and a providing party both of which can be either human users or software agents. Automated discovery is a means for the requestor to find potential providers by querying a registry. Providers advertise the capabilities of their services to the registry whereas requestors formulate the goals they want to achieve. For a description of the semantics of goals and capabilities they make use of an ontological vocabulary based on some underlying knowledge representation formalism. Doing so they refer to commonly used domain ontologies that capture general knowledge about the corresponding domain of discourse, as e.g. delivery of products. Discovery then reduces to the task of matching goals and capabilities expressed as ontological descriptions [89].

Besides the actual business-level semantics of a service, aspects of choreographic or compositional semantics can be taken into account as well in the context of discovery. Part of the discovery semantics of a service could, for example, be characterised by its pre and post conditions or by certain parameters being part of the protocol, as e.g. the occurrence of a credit card number in the choreography of the service interface.

A concrete service instance determines all parameters - nothing is left open for the two parties to decide about. In the book-selling example a service instance specifies exactly which book is going to be delivered to which address and which

amount of money has to be paid in which form. In this sense descriptions of goals and capabilities are templates for service instances - they allow several possibilities of how the service can be carried out. For example, the provider of the service decides to accept several payment methods and does not specify in the capability description which one is finally being used. The semantics of discovery matchmaking can be defined in terms of sets of service instances: a goal and a capability match if the sets of service instances they allow intersect. In this case there is at least one possible service instance which they both agree on. This approach is followed in [90].

In the general case discovery does not directly lead to a concrete service instance. Once a service has been discovered its parameters have to be negotiated between the requestor and the provider. The outcome of the discovery is not a service instance to be carried out but just the fact that the two parties can potentially do business with each other [90]. After discovery, negotiation might lead to a concrete service instance but it does not necessarily have to. For example, a book selling service provider advertises that it sells books, which is sufficient for successful discovery involving a requestor who searches for a book selling service in the internet. However, the particular book the requestor is looking for might be out of stock.

Currently, several ontological languages for declarative description of Web Service semantics emerge, such as OWL-S[91] and WSMO [92]. They provide top-level ontologies for Web Services covering the specification of service profile, process control flow, message exchange and mediation. Goals and capabilities can be expressed combining these upper level service ontologies with domain ontologies.

Considering such a top-level ontology for Web Services there are several technical approaches to discovery and appropriate description of service semantics. One of them is to model knowledge about services on the ontological level of concepts and relations and then perform schema matching. An example is given in [93] where description logic reasoning is used. Both, goals and capabilities, are described as description logic concept expressions. To check the intersection of the two concepts for satisfiability reduces matchmaking for discovery to standard description logic inferences. Another idea is to use the stronger subsumption inference and to check whether the goal describes a specialized form of the service advertised by the capability or vice versa. In [89] and [94] modified structural subsumption algorithms are applied in frameworks that support partial matches on a discrete scale. Subsumption in either direction is considered stronger than satisfiability of concept intersection but weaker than an exact match. An alternative approach to schema-level matching is to model knowledge about services on the level of instances. In this case discovery is achieved by querying the extension of a goal concept expression and applying ontology-based information retrieval techniques.

Discovery of Web Services benefits from ontological descriptions of service semantics in that reasoning based on formal semantics can be applied in matchmaking algorithms for goals and capabilities. The knowledge captured in domain

ontologies can potentially be used to derive a match where the facts stated in the goal and capability alone would not be sufficient to do so. Formal semantics helps to derive facts that are not explicitly stated.

Incorporating semantics into Web Services is quite a new field and it is still an open issue how to semantically describe and annotate services in order to properly discover them by appropriate techniques - two aspects that go hand in hand. Application of discovery approaches to concrete case study scenarios have to show which aspects of service semantics have to be exploited and which reasoning techniques have to be applied to yield good solutions.

7.2 Digital Libraries

The growing availability of cyber infrastructures like GRID, Peer-to-Peer and Web Services, will lead to more open and flexible digital library (DL) architectures, e.g. BRICKS [95]. Hence DL will opened to a wider clientele by enabling more cost-effective usage and better tailored DL. Furthermore new types of infrastructures allow dynamic federative models of content and service provision involving a wide range of distributed content and service providers. This has implications for the realization of digital library functionalities mainly rooted in the increased heterogeneity of content, services and metadata. Future distributed DL infrastructures will consists of a large number of loosely coupled DL systems all over the world. Users of these infrastructures will be able to retrieve information from all involved DL. Due to the high degree of distribution, these infrastructures will often omit centralized management systems. Hence no central retrieval service and no central authority, which has a complete system overview, will exist. The decentralization approach poses new challenges to various areas like information retrieval, security, etc.

One major problem in decentralized DL infrastructures is that most DLs are using different data schemas as well as different classification systems. The standard data integration strategies like Global As View resp. Local As View or approaches to define standards for schemas or ontologies works for many specialized applications very well but is problematic in decentralized and highly dynamic environments. DL Nodes may appear and disappear in the system for several reason like network problems, economic problems, etc.. In these environments many local data schemas and ontologies exist. The local DL owners have their own semantic understanding of their data. Due to the diversity of information, reaching a global agreement among all is difficult. By viewing semantics as a form of agreement the emergent semantics approach is to enable the participating data sources to incrementally develop a global agreement in an evolutionary process that solely relies on pair-wise interactions.

The idea behind to use the emergent semantics approach is the assumption that local experts have the best knowledge about their data. Hence they know best the semantic interpretation of the data. Furthermore they have preferred collaboration partners, whose information and semantic interpretation the local experts know quite well. This assumption belongs to all nodes within the DL infrastructure. Hence local experts are able to generate high quality mappings

between their own schema and classifications and those of their partners. Often the mappings are already available from previous collaborations and can be reused in the process.

The mappings are distributed together with the query to the queried library. The queried library will integrate the mapping in their local mapping table and performs the query. In addition the query will be send also other neighbours, which also receive the mapping in this way. The neighbours learn from the received mappings how to interpret the semantics of other DL. They are also able to derive mappings, e.g. the DL A sends a mapping $A \rightarrow B$ to the DL B. DL C knows already the mapping for $B \rightarrow C$ and receives the mapping for $A \rightarrow B$. Hence DL C will be able to derive a mapping $A \rightarrow C$ by using $A \rightarrow B \rightarrow C$.In this way every DL will learn about the new mapping, which can be used later on.

Nevertheless also the method of emergent semantics has prerequisites and limitations. First limitations arise from the mapping itself. So it is not always possible to generate a complete and 100% accurate mapping, e.g. due to missing fields or ambiguous semantics. Furthermore mappings between heterogeneous standards, e.g. between Dublin Core and MPEG7, also leads to problems. Hence a good practice is to restrict the process to a specific domain, e.g. science of art. With this restriction several semantic problems can be avoided as all involved persons have a similar understanding domain and of the semantics. Even if the emergent semantics approach will not solve all interoperability problems in digital libraries, will it be a very useful method to support ad-hoc collaborations.

7.3 Scientific Collaboration

Semantic reconciliation is crucial in scientific collaboration [96]. Consider the case of integrated environmental models. These models represent the consensus understanding of earth systems reached by scientists in the field at some period in time. They are composed of sub-models, which attempt to capture particular environmental systems. For example, ground water models describe subsurface water flow; infiltration models describe the movement of water into soils, and so on. These sub-models alone describe only small parts of the environment, but together they can address questions concerning the environment as a whole. The challenge is to find ways of integrating successfully a subset of these sub-models to deal with a specific goal while preserving the autonomy of the individual models. In other words, integration of sub-models must be goal-driven between peers, and similarly integration of heterogeneous information sources must be query-driven, while also preserving the autonomy of the individual models and/or information sources and services. Each goal and each query may require the elicitation of different interpretations of the models and the information sources and services within specific contexts [28]. The semantics necessary for integration emerges incrementally from the interaction of peers, as additional queries are posed and information sources and services become available. Thus, semantic reconciliation in model integration is an emergent phenomenon.

Consider, for example, query "Where do the sub-models agree on soil moisture at the beginning of the season?". The answer will depend on the models used and their context assumptions, which in this case include at least the spatial context (where), the attributes' context (soil moisture), and the temporal context (at the beginning of the growing season) [97]. These same observations about the role of context in model integration process recur in other scientific domains. Integration may be triggered by the activity of a scientist exploring the Internet and the web for models or services related to a specific real-time experiment. The models are likely to have been developed autonomously. Model autonomy must be preserved, as models in natural sciences have deeply rooted assumptions to allow representation of processes that may only be partially understood. If the underlying assumptions are not always completely specified, which is usually the case, then the integration of sub-models may result in semantic errors. In geographic applications or satellite-based information systems, for example, a variety of semantic errors resulting from model integration have been investigated [98, 99]. Their occurrences are generally pegged to the lack of unified theories of space, time and accuracy [100]. However, this cannot be the only reason, as ontologies, constructed on the basis of the potential theories, will not feasibly be able to capture every possible application context of model integration. Thus, reliance on ontologies alone will be insufficient to entirely resolve the problem of semantic conflict [1].

Semantic analysis will require that models be able to self-evaluate to determine the level of violation of their own underlying assumptions with respect to an expected behavior defined within an application context. This analysis is performed within an application context guided by ontologies from cognate fields. In the environmental example, the application context consists of information such as field measurements, remotely sensed imagery, and maps. The collected information is interpreted within ontologies from cognate fields such as meteorology, geology, soil science, and ecology. They contribute contextual information about the properties of a natural environment and their aggregation, scale and resolution of observations, and generalization. Consider the example of the RHESSysd system [101]. One spatial aggregation is land unit, which represents a scale that captures long-range spatial variability. Ground water is stored at this level, but a refined semantic analysis may require disaggregating a land unit into its constituent components at lower resolution, such as the land patch. The disaggregation will be performed on the basis of assumed relations between water movement and landscape position. Observe that the ontology knowledge used for disaggregation are assumptions about water movements. These assumptions, which are simply process models, may be only ephemeral estimations subject to reevaluation as new results in the field are obtained. Collectively, all the knowledge gained represents the application context.

Users of environmental databases and models may have had no role in the development of the information sources, but nevertheless, need to use them. End-users contribute contextual information in the form of queries to the information services. The concept of context elicitation [28, 29] is a process, which allows

incremental extraction of relevant information to the query from the information sources and services. There is need for a notion of semantic distance to measure the compatibility of queries with the elicited information.

Queries of scientific literature may not be simple searches for information in a given topic. At the frontier of scientific discovery, investigators may wish to assess untested scientific hypotheses, or to uncover hereto unknown relations between two lines of inquiry. The semantics necessary to validate (or invalidate) these hypotheses may not be readily available. They are constructed incrementally. They form the context elicited from the information sources and services through an interactive (and often non-monotonic) semantic reconciliation process, which incrementally refines the evidence gathered at each stage.

In summary, context in scientific collaboration is elicited from the application context through an incremental query-directed semantic reconciliation process. It is thus emergent. A semantic distance measure is necessary to continuously measure at any state the semantic compatibility between this context and the user query. The challenge in the area is the development of scalable convergent context elicitation algorithms or heuristics.

8 Conclusions

The preceding work results from a large collaborative effort initiated more than one year ago by the IFIP 2.6 Working Group on Databases. The project has since then evolved to include external contributions as well. The field of Emergent Semantics is still clearly in its infancy, and we would welcome remarks as well as any kind of feedback based on this material.

References

1. 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.
2. A. M. Ouksel. In-context peer-to-peer information filtering on the web. *SIGMOD Record*, 32,3, 2003.
3. M. Vincini S. Bergamaschi, S. Castano and D. Beneventano. Semantic integration of heterogeneous information sources. *Data and Knowledge Engineering*, 36(3):215–249.
4. S. E. Madnick C. H. Goh, S. Bressan and M. D. Siegel. Context interchange: New features and formalisms for the intelligent integration of information. *ACM Trans. on Information Systems*, 17(3):270–293.
5. Y. Vassiliou M. Jarke, M. Lenzerini and editors P. Vassiliadis. *Fundamentals of Data Warehouses*. Springer, 1999.
6. H. Garcia-Molina Y. Papakonstantinou and J. Widom. Object exchange across heterogeneous information sources. In *International Conference on Data Engineering (ICDE)*, 1995.
7. J. Widom (ed.). Special issue on materialized views and data warehousing. *IEEE Bull. on Data Engineering*, 18(2), 1995.

8. J. D. Ullman. Information integration using logical views. In *International Conference on Database Theory (ICDT)*, pages 19–40, 1997.
9. G. De Giacomo D. Calvanese and M. Lenzerini. Answering queries using views in description logics. In *Proceedings of AAAI*, 2000.
10. A. Halevy. Answering queries using views: a survey. *VLDB Journal*, 10(4), 2001.
11. M. Friedman and D. S. Weld. Efficiently executing information-gathering plans. In *International Joint Conference on Artificial Intelligence*, 1997.
12. O. Etzioni and D. Weld. A softbot-based interface to the internet. *Communications of the ACM*, 37(7):72–76, 1994.
13. A. M. Keller M. Genesereth and O. M. Duschka. Infomaster: an information integration system. In *ACM SIGMOD International Conference on Management of Data*, 1997.
14. A. Levy J. Ordille and A. Rajaraman. Querying heteregeneous information sources using source descriptions. In *International Conference on Very Large Data Bases (VLDB)*, 1996.
15. A. Tate (editor). *Advanced Planning Technology*. AAAI Press, 1996.
16. Y. Arens and C. A. Knoblock. Sims: Retrieving and integrating information from multiple sources. In *ACM SIGMOD International Conference on Management of Data*, 1993.
17. E. Mena, A. Illarramendi, V. Kashyap, and A. Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *International journal on Distributed And Parallel Databases (DAPD)*, 8(2):223–271, 2000.
18. S. Castano A. Corni R. Guidetti G. Malvezzi M. Melchiori D. Beneventano, S. Bergamaschi and M. Vincini. Information integration: The momis project demonstration. In *International Conference on Very Large Data Bases (VLDB)*, 2000.
19. D. Calvanese, G. De Giacomo, , M. Lenzerini, D. Nardi, and R. Rosati. Knowledge representation approach to information integration. In *AAAI workshop on AI and Information Integration*, pages 58–65, 1998.
20. M. Rousset and C. Reynaud. Knowledge representation for information integration. *Information Systems*, 29(1), 2004.
21. G. Guerrini E. Bertino and M. Mesiti. A matching algorithm for measuring the structural similarity between an xml document and a dtd and its applications. *Information Systems*, 29(1):23–46, 2004.
22. Z. W. Ras and A. Dardzinska. Ontology-based distributed autonomous knowledge systems. *Information Systems*, 29(1):47–58, 2004.
23. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
24. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.
25. H. Garcia-Molina O. Brandman, J. Cho and N. Shivakumar. Crawler-friendly web servers. In *Workshop on Performance and Architecture of Web Servers (PAWS)*, 2000.
26. A. Vakali E. Terzi and M. S. Hacid. Knowledge representation, ontologies and semantic web. In *Asia-Pacific Web Conference (APWeb)*, 2003.
27. K. Aberer, Ph. Cudre-Mauroux, and M. Hauswirth. A framework for semantic gossiping. *SIGMOD Record*, 31(4), 2002.
28. A. M. Ouksel and C. Naiman. Coordinating context building in heterogeneous information systems. *Journal of Intelligent Information Systems*, 3,1:151–183, 1994.

29. A. M. Ouksel. *A Framework for a Scalable Agent Architecture of Cooperating Heterogeneous Knowledge Sources*. Springer Verlag, 1999.
30. T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928.
31. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *International World Wide Web Conference (WWW)*, 1998.
32. F. Heylighen. Principia cybernetica web. http://pespmc1.vub.ac.be.
33. S. Camazine et al. *Self Organization in Biological Systems*. Princeton University Press, 2001.
34. T. Kohonen. *Self Organising and Associative Memory*. Springer Verlag, 1989.
35. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. of Sciences*, 79:2554–2558, 1982.
36. J. Kennedy, R. Eberhart, and Y. Shi. *Swarm Intelligence*. Morgan Kaufmann Academic Press, 2001.
37. R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2001.
38. M. Goldhaber. The attention economy and the net. In *First Monday, Vol 2, No 4*, 1997.
39. A. Martinoli and F. Mondada. Probabilistic modelling of a bio-inspired collective experiment with real robots. In *Proceeding of the Third International Symposium on Distributed Autonomous Robotic Systems*.
40. K. Aberer. P-Grid: A self-organizing access structure for P2P information systems. *Lecture Notes in Computer Science*, 2172:179–185, 2001.
41. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content addressable network. In *Proceedings of ACM SIGCOMM 2001*, 2001.
42. G. Weiss (ed.). *Multiagent Systems*. MIT Press, 2000.
43. K. Aberer, A. Datta, and M. Hauswirth. The quest for balancing peer load in structured peer-to-peer systems. Technical report ic/2003/32, EPFL, 2003.
44. A. Datta, M. Hauswirth, and K. Aberer. Updates in highly unreliable, replicated peer-to-peer systems. In *Proceedings of the 23rd International Conference on Distributed Computing Systems, ICDCS2003*, Providence, Rhode Island, USA, 2003.
45. R. McCool and R. V. Guha. Tap, building the semantic web. http://tap.stanford.edu/.
46. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. The Chatty Web: Emergent Semantics Through Gossiping. In *International World Wide Web Conference (WWW)*, 2003.
47. 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), 2003.
48. V. S. Iyengar, C. Apte, and T. Zhang. Active learning using adaptive resampling. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–98. ACM Press, 2000.
49. M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.
50. E. Charniak and M. Berland. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 57–64, 1999.
51. A. Mädche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, March/April 2001.

52. Googlism, 2003. http://www.googlism.com.
53. G.W. Flake, S. Lawrence, C.L. Giles, and F.M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3):66 –70, March 2002.
54. E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the eleventh international conference on World Wide Web*, pages 562–569. ACM Press, 2002.
55. M. Spiliopoulou. Web usage mining for web site evaluation. *Commun. ACM*, 43(8):127–134, 2000.
56. G. Grefenstette. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99 Translating and the Computer 21*, 1999.
57. E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching Very Large Ontologies using the WWW. In *Workshop on Ontology Construction of the ECAI*, 2000.
58. F. Keller, M. Lapata, and O. Ourioupina. Using the web to overcome data sparseness. In *Proceedings of EMNLP-02*, pages 230–237, 2002.
59. K. Markert, N. Modjeska, and M. Nissim. Using the web for nominal anaphora resolution. In *EACL Workshop on the Computational Treatment of Anaphora*, 2003.
60. S. Chakrabarti. Data mining for hypertext: a tutorial survey. *ACM SIGKDD Explorations Newsletter*, 1(2):1–11, January 2000.
61. P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web, 2003. Submitted for publication.
62. A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers, 2002.
63. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.
64. C. Behrens and V. Kashyap. The "emergent" semantic web: A consensus approach for deriving semantic knowledge on the web. In *proceedings of the International of the Semantic Web Working Symposium (SWWS), Stanford University, California, USA 2001*, 2001.
65. E. Cohen, A. Fiat, and H. Kaplan. Associative search in peer to peer networks: Harnessing latent semantics. In *IEEE INFOCOM*, 2003.
66. D. Beneventano, S. Bergamaschi, F. Guerra, and M. Vincini. Building an integrated ontology within sewasie system. In *Semantic Web and Data Bases (SWDB) Workshop*, 2003.
67. N. Guarino. Formal ontology in information systems. In *International Conference On Formal Ontology In Information Systems (FOIS)*, 1998.
68. 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*, 1:65–77, 1988.
69. Newell A. The knowledge level. *Artificial Intelligence*, 18(1), 1982.
70. G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
71. M. Jarrar, J. Demey, and R. Meersman. On using conceptual data modeling for ontology engineering. *Journal on Data Semantics*, LNCS, Vol. 2800, 2003.
72. M. Jarrar and R. Meersman. Formal ontology engineering in the dogma approach. In *International Conference on Ontologies, Databases and Applications of Semantics (ODBase)*, 2002.
73. S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The lorel query language for semistructured data. *Int. J. on Digital Libraries*, 1(1):68–88, 1997.

74. Resource description framework (rdf) model and syntax specification, 1999. W3C Recommendation 22 February 1999. http://www.w3.org/RDF/.
75. Rdf vocabulary description language 1.0: Rdf schema, 2003. W3C Working Draft 10 October 2003. http://www.w3.org/RDF/.
76. Owl web ontology language reference, 2003. W3C Candidate Recommendation 18 August 2003. http://www.w3.org/TR/owl-ref/.
77. A. Pease and I. Niles. Ieee standard upper ontology: A progress report. *Knowledge Engineering Review, Special Issue on Ontologies and Agents*, 17:65–70.
78. E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. Kaon - towards a large scale semantic web. In *Proc. of EC-Web 2002*, LNCS, pages 304–313. Springer, 2002.
79. M. Lenzerini. Data Integration: A Theoretical Perspective. In *21st ACM Symposium on Principles of Database Systems (PODS 2002)*.
80. M. Ehrig, P. Haase, F. van Harmelen, R. Siebes, S. Staab, H. Stuckenschmidt, R. Studer, and C. Tempich. The swap data and metadata model for semantics-based peer-to-peer systems. In *Proceedings of MATES-2003. First German Conference on Multiagent Technologies. Erfurt, Germany, September 22-25*, LNAI, pages 144–155. Springer, 2003.
81. R. Siebes and F. van Harmelen. Ranking agent statements for building evolving ontologies. In *Proceedings of the AAAI-02 workshop on meaning negotiation, Alberta, Canada,July 28 2002*, 2002.
82. C. Tempich, S. Staab, and A. Wranik. REMINDIN': Semantic query routing in peer-to-peer networks based on social metaphors. In *International World Wide Web Conference (WWW), New York, USA*, 2004.
83. K. Aberer and M. Hauswirth. Peer-to-peer information systems: Concepts and models, state-of-the-art, and future systems. In *18th International Conference on Data Engineering (ICDE), San Jose, California, 2002.*
84. D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, and Z. Xu. Peer-to-peer computing. Technical report, HPL, 2002.
85. R. Guerraoui and A. Schiper. Software-based replication for fault tolerance. *IEEE Computer Journal*, 30(4):68–74, 1997.
86. Sheila A. McIlraith, Tran Cao Son, and Honglei Zeng. Semantic web services. *IEEE Intelligent Systems*, 16(2):46–53, 2001.
87. D. Berardi, D. Calvanese, G. DeGiacomo, M. Lenzerini, and M. Mecella. E-service composition by description logics based reasoning. In *Proceedings of the International Workshop on Description Logics*, 2003.
88. A. Eberhart. Ad-hoc invocation of semantic web services. In *Proceedings of the IEEE International Conference on Web Services*, San Diego, 2004.
89. M. Paolucci, T. Kawmura, T. Payne, and K. Sycara. Semantic matching of web services capabilities. In *First Int. Semantic Web Conf.*, 2002.
90. D. Trastour, C. Bartolini, and C. Preist. Semantic web support for the business-to-business e-commerce lifecycle. In *Proceedings of the eleventh international conference on World Wide Web*, pages 89–98. ACM Press, 2002.
91. *OWL Service Coalition. OWL-S: Semantic Markup for Web Services, November 2003.* http://www.daml.org/services/owl-s/1.0/.
92. *Web Service Modeling Ontology.* http://wsmo.org/.
93. L. Li and I. Horrocks. A software framework for matchmaking based on semantic web technology. In *Proceedings of the twelfth international conference on World Wide Web*, pages 331–339. ACM Press, 2003.

94. T. Di Noia, E. Di Sciascio, F. M. Donini, and M. Mongiello. A system for principled matchmaking in an electronic marketplace. In *Proceedings of the twelfth international conference on World Wide Web*, pages 321–330. ACM Press, 2003.
95. *Bricks - Building Resources for Integrated Cultural Knowledge Services, EU-IST 507457*. http://www.brickscommunity.org/.
96. A. M. Ouksel. Emergent semantics and in-context peer-to-peer information filtering and model calibration on the web. In *Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida*, 2003.
97. A. M. Ouksel and D. S. Mackay. Environmental modeling within a scalable multi-agent architecture for semantic cooperation amongst autonomous and heterogeneous information sources. In *NSF Proposal 1004213-9978386. (PI: A. M. Ouksel). UIC-IDS-CRIM/Tech-Report 99-07*, 1999.
98. V. B. Robinson and A. U. Frank. About different kinds of uncertainty in collections of spatial data. In *Auto-Carto 7, American Society for Phtogrammetry and Remote Sensing and American Congress on Surveying and Mapping*, Falls Church, VA, 1985.
99. M. F. Worboys and S. M. Deen. Semantic heterogeneity in distributed geographic databases. *ACM SIGMOD Record*, 20(4), 1991.
100. G. C. Roman. Formal specification of geographic data processing requirements. *IEEE Transactions on Knowledge and Data Engineering*, 2(4), 1990.
101. D. S. Mackay. Semantic integration of environmental models for application to global information and decision-making. *ACM SIGMOD Record*, 28(1), 1999.