

Emergent Semantics Principles and Issues^{*}

Karl Aberer¹, Philippe Cudré-Mauroux¹ and Aris M. Ouksel² (*editors*)
Tiziana Catarci³, Mohand-Said Hacid⁴, Arantza Illarramendi⁵, Vipul
Kashyap⁶, Massimo Mecella³, Eduardo Mena⁷, Erich J. Neuhold⁸, Olga De
Troyer⁹, Thomas Risse⁸, Monica Scannapieco³, Fèlix Saltor¹⁰, Luca de Santis³,
Stefano Spaccapietra¹, Steffen Staab¹¹, and Rudi Studer¹¹

¹ Swiss Federal Institute of Technology (EPFL), Switzerland

² Univ. of Illinois at Chicago, USA

³ Univ. of Roma 1, Italy

⁴ Univ. of Lyon 1, France

⁵ Univ. of the Basque Country, Spain

⁶ National Library of Medicine, USA

⁷ Univ. of Zaragoza, Spain

⁸ Fraunhofer IPSI, Germany

⁹ Vrije Univ. of Brussel, Belgium

¹⁰ Univ. Politècnica de Catalunya, Spain

¹¹ Univ. of Karlsruhe, Germany

Abstract. Information and communication infrastructures underwent a rapid and extreme decentralization process over the past decade: From a world of statically and partially connected central servers rose an intricate web of millions of information sources loosely connecting one to another. Today, we expect to witness the extension of this revolution with the wide adoption of meta-data standards like RDF or OWL underpinning the creation of a semantic web. Again, we hope for global properties to emerge from a multiplicity of pair-wise, local interactions, resulting eventually in a self-stabilizing semantic infrastructure. This paper represents an effort to summarize the conditions under which this revolution would take place as well as an attempt to underline its main properties, limitations and possible 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 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

^{*} The work presented in this paper reflects the current status of a collaborative effort initiated by the IFIP 2.6 Working Group on Data Semantics.

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 next section summarizes the rationale and

principles of this new semantic trend. We then present some related opportunities and challenges in Sec. 3 before delving into some technical issues. Finally, we go over three short case studies in Sec. 5 and conclude in Sec. 6.

2 The Emergence of Emergent Semantics

Information systems have since long been characterized by a multitude of autonomous, heterogeneous information repositories. The problem of how to provide transparent access to heterogeneous information sources while maintaining their autonomy already appeared decades ago. Information integration systems typically provide a uniform query interface to a collection of distributed and heterogeneous information sources, giving users or other agents the illusion that they query a centralized and homogeneous information system. As such, they are considered as mediation systems between users and multiple data sources which can be syntactically or semantically heterogeneous while being related to the same domain. The existing mediator-based information systems can be distinguished according to: (1) the type of mappings between the mediated schemas and the schemas of the sources: There exist basically two approaches for such mappings, the Global As View (GAV) and the Local As View (LAV). The Global As View approach describes the global schemas as a view over all local schemas, whereas the local as view approach describes each local schema as a view over the global schemas; (2) the languages used for modelling the mediated schemas and the source descriptions and (3) the expressivity of the mediated schemas.

Independently of this main thread, several research areas, including peer-to-peer data management, information agents, the Semantic Web and Web data mining, have progressively converged in the last decade to remarkably similar ideas on how to address the problem of semantic interoperability in widely distributed information systems with large numbers of agents¹ [3, 4]. 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”.

2.1 Principle 1: Agreements as a Semantic Handshake Protocol

Meaningful exchanges can only occur on the basis of mutually accepted propositions [5]. 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.

2.2 Principle 2: 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.

2.3 Key Principle 3: Agreements emerge from local interactions

The principles stated so far are analogous to those formulated for introducing the concept of ontological commitments [6], 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.

2.4 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.

2.5 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. 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.

2.6 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.

3 Opportunities – Challenges

Semantics do not emerge from purely random settings, but rather from environments exhibiting specific, well-known properties. We detail below some important opportunities and challenges related to emergent semantic systems in probabilistic and decentralized contexts.

Locality is often referred to 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 [7] 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 [8]. 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 [9] and decentralized peer-to-peer [10, 11] 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 [12] 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 [13] 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 [14], 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 [15] or make use of gossiping and local translation mappings to incrementally foster interoperability in the large [16].

4 Technical Issues for Emergent Semantics

In this section we discuss structures, functions and architectures for emergent semantic systems. This preliminary overview of technical issues and the current state of the art is organized along the categories of representation model, meta-data, local consensus construction, derivation of global agreements, and physical implementation.

4.1 Representational Model

There is the need to commit to a common representational model. The trade-off among different models is one between expressive power and efficiency. Whereas for the relational data model efficient implementations exist, implementations of semantically richer models supporting reasoning over schemas, such as OWL (Web Ontology Language) [17], are far from supporting scalable applications. Currently, semi-structured data models like RDF [18] seem to provide a good middle ground for supporting flexibility, richer semantics and efficient implementations.

4.2 Meta-data

Common vocabularies and agreed-upon measures, both for data and schemas, are an essential constituent of any mechanism for establishing semantic agreement. We identify the following classes of meta-data to that end:

Lexical information: Textual data is frequently part of the bootstrapping mechanism when establishing local consensus on the meaning of data objects. Thus lexica like WordNet [19] supporting the reuse of existing lexical information and semantic relationships among text entities are central. As lexica are dynamic and multiple lexica may be used in common, the conceptual structures underlying lexica themselves require agreed-upon representation, as proposed in [20].

Trust and quality information: Evaluating the degree of consensus requires measures. These measures may refer to the assessment of trust into an information providing agent and to the assessment of the perceived quality of information received. Trust is typically based on the “consensus” of people’s opinions about each other. The e-bay rating system is an example of this. However, it doesn’t take into account the trustworthiness and the “expertise” of the raters. Various recent works investigated mechanisms to establish consensus on trust taking into account reputation of referrals [21]. The quality of information is either dependent on the opinions of people or upon the applications consuming that information. In the former case, consensus computation will be a crucial component in determining the quality of information [22]. Provenance information [23] allows to relate information to its source, and thus to its trust rating. Siebes and van Harmelen[24] and Tempich et al. [25] have shown how provenance information can be used for agreeing on semantics.

4.3 Local Consensus Construction

Local consensus building requires to relate information sources using different representations. For structured data, this problem has been studied extensively in the context of schema matching. [26] provides an overview on automatic schema matching approaches. Many of those apply machine learning techniques [27–29]. Frequently shared ontologies are used to facilitate schema matching. For example, in OBSERVER [30] each information source maintains an ontology, expressed in description logics, to associate semantics with the information stored and to process distributed queries. [31] uses machine-generated ontologies extracted from web forms to integrate web services. In [1] a probabilistic framework for reasoning with assertions on schema relationships is introduced. Thus this approach deals with the problem of having possibly contradictory knowledge on schema relationships.

Local schema mappings are the basis of several recent proposals for P2P data management. The Piazza system [32] proposes an architecture and mapping languages for connecting XML or RDF data sources. [33] proposes an architecture for managing distributed relational databases in a P2P environment. Edutella [34] is a recent approach applying the P2P architectural principles to build a semantically interoperable information system for the educational domain based on semantic Web standards.

In summary, there exists a rapidly growing number of approaches to support automated construction of local consensus based on schema matching and the use of the resulting schema mappings.

4.4 Building Global Agreements

Link-based ranking as performed by Web search engines may be seen as a simple global agreement derivation method. Variations of PageRank have been investigated, such as HITS [35]. Information extracted globally from such web minings can support the disambiguation of specific local semantic relationships, i.e., the

building of local consensus [36–39]. Web content mining extends the scope of Web mining beyond link analysis and uses explicit, linguistically motivated natural-language descriptions to propose semantic relationships [40–43]. The Web structure itself can be used to determine a focus for harvesting data [44, 45] and to classify and cluster data on the Web. Examples of combinations of several of these techniques are given in [46, 47]. “Traditional” ontology learning is an area which aims at extracting ontologies from, mostly, text documents [48–53].

In the context of P2P systems, approaches for achieving *multilateral consensus* among peers managing structured data have recently been introduced. In [16] participating agents incrementally develop a global agreement in an evolutionary and completely decentralized, probabilistic process based on local schema mappings in a P2P network. [54] proposes the use of consensus analysis as a tool for extracting controlled vocabularies and domain ontologies for the semantic Web.

4.5 Physical Implementation

Any mechanism for establishing semantic agreements grounds in the ability of locating resources in a network. Since, for attaining scalability, we aim at decentralization at the logical level, the same property should hold for the physical implementation. Recently, substantial progress has been achieved on efficient, decentralized resource location in the area of P2P systems. For a comprehensive overview see, for example, [55]. Roughly, we can distinguish among unstructured P2P systems [56], based on gossiping techniques, hierarchical P2P systems, with designated superpeers responsible for routing [57], and structure P2P systems based on some variation of distributed hash tables (DHTs), combining efficient search and maintenance while avoiding centralized components [58, 10]. As soon as logical identifiers bear semantics, load balancing becomes an issue as semantically meaningful identifiers are not necessarily uniformly distributed [59, 60].

P2P networks are logical overlay networks over a physical infrastructure. Their maintenance is closely related to the problem of identification. Being able to relate possibly changing physical identifiers to stable logical identifiers is central, in particular if the logical identifiers bear semantics. Several approaches aim at maintaining a stable and consistent P2P network [58, 61, 62], but at the cost of (unnecessarily) changing the logical reference structure of the network. [63] is an approach that allows to maintain logical neighborhoods in the presence of physical network changes.

5 Cases studies

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.

5.1 Service Discovery

The discovery of services is the most important functionality in distributed and service-oriented environments like web services or GRID computing. The problem is to discover services responding to user requirements. Standards like UDDI or WSDL support description of services and discovery functionalities from a syntactic perspective. But the major problem remains: the semantics of the service description. Often, the same services are described by users and service providers in different ways. The obvious approach is standardization. The problem with standardization is that it does not usually cover all requirements a priori and thus service providers are tempted to introduce new types of service descriptions. As services and their descriptions are evolving quickly responding to market changes, it is in general impossible to keep up with all requirements in time. A decentralized discovery service exploiting emergent semantics approaches to extend the standards in a controlled way and distribute the changes among the peers might be an adequate solution. With such an approach, peer could “learn” about new descriptions and mappings incrementally departing from existing standards used for bootstrapping the process.

5.2 Digital Libraries

With the growing availability of cyber-infrastructures like GRID, Peer-to-Peer and Web Services, the access to digital documents and all types of multimedia objects stored in digital libraries (DL) becomes easier. However, the common problem is that most DLs are using different data schemas as well as different classification systems. Hence users have to generate for each library they are accessing a new mapping between the schemas resp. classifications. As long as a fixed set of data sources is used, static mappings are a straightforward solution with a reasonable effort. But users require a more flexible selection of sources and like to have integrated query facilities among several DLs. Hence a DL will be able distribute a query among its neighbours. This implies for the schemas and classifications a more dynamic handling of the mappings. With an emergent semantics approach, anyone querying a DL would generate a mapping between his own and some other library schema or classification. The mappings would be sent together with the query to the library, which would then distribute them among other neighbour libraries. In this way, every DL could learn about new mappings, which they could use later on. The construction of mappings would not be completely automated, but the distribution, reuse and composition could be performed automatically and thus effectively used would be made of investments into establishing local mappings by domain experts.

5.3 Scientific Collaboration

Semantic reconciliation is crucial in scientific collaboration. Let us 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. For example, query “Where do the sub-models agree on soil moisture at the beginning of the season?” will depend on the model used and its context assumptions, including at least the spatial context, the attributes’ context, and the temporal context. These same observations apply to 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.

6 Conclusions

The preceding work results from a larger collaborative effort initiated about one year ago by the IFIP 2.6 Working Group on Data Semantics. The project has since then evolved to include external contributions. This work is still in progress, and we would welcome remarks as well as any kind of feedback 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. K. Aberer, Ph. Cudre-Mauroux, and M. Hauswirth. A framework for semantic gossiping. *SIGMOD Record*, 31(4), 2002.
4. A. M. Ouksel and C. Naiman. Coordinating context building in heterogeneous information systems. *Journal of Intelligent Information Systems*, 3,1:151–183.
5. A. M. Ouksel. *A Framework for a Scalable Agent Architecture of Cooperating Heterogeneous Knowledge Sources*. Springer Verlag, 1999.

6. 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.
7. R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2001.
8. M. Goldhaber. The attention economy and the net. In *First Monday, Vol 2, No 4*, 1997.
9. 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*.
10. K. Aberer. P-Grid: A self-organizing access structure for P2P information systems. *Lecture Notes in Computer Science*, 2172:179–185, 2001.
11. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content addressable network. In *Proceedings of ACM SIGCOMM 2001*, 2001.
12. G. Weiss (ed.). *Multiagent Systems*. MIT Press, 2000.
13. 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.
14. 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.
15. R. McCool and R.V. Guha. Tap, building the semantic web.
16. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. The Chatty Web: Emergent Semantics Through Gossiping. In *International World Wide Web Conference (WWW)*, 2003.
17. Owl web ontology language reference, 2003. W3C Candidate Recommendation 18 August 2003. <http://www.w3.org/TR/owl-ref/>.
18. Resource description framework (rdf) model and syntax specification, 1999. W3C Recommendation 22 February 1999. <http://www.w3.org/RDF/>.
19. G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
20. 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.
21. S. Kamvar, M. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *International World Wide Web Conference (WWW)*, pages 640–651, 2003.
22. M. Scannapieco, V. Mirabella, M. Mecella, and C. Batini. Data Quality in e-Business Applications. In *Web Services, E-Business, and the Semantic Web (WES)*, 2002.
23. 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.
24. 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.
25. C. Tempich, S. Staab, and A. Wranik. REMINDIN’: Semantic query routing in peer-to-peer networks based on social metaphors, 2003. submitted for publication.
26. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.

27. T. Milo and S. Zohar. Using schema matching to simplify heterogeneous data translation. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 122–133, 24–27 1998.
28. A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of the eleventh international conference on World Wide Web*, pages 662–673. ACM Press, 2002.
29. J. Berlin and A. Motro. Autoplex: Automated discovery of content for virtual databases. In *CoopIS 2001, Trento, Italy*, 2001.
30. E. Mena, V. Kashyap, A. P. Sheth, and A. Illarramendi. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. *Distributed and Parallel Databases*, 8(2):223–271, 2000.
31. H. M. Jamil G. A. Modica, A. Gal. The use of machine-generated ontologies in dynamic information seeking. In *CoopIS*, pages 433–448, 2001.
32. A. Y. Halevy, Z. G. Ives, P. Mork, and I. Tatarinov. Piazza: Data Management Infrastructure for Semantic Web Applications. In *International World Wide Web Conference (WWW)*, 2003.
33. P. A. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. Mylopoulos, L. Serafini, and I. Zaihrayeu. Data management for peer-to-peer computing: A vision. In *Workshop on the Web and Databases (WebDB)*, 2002.
34. W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palmér, and T. Risch. EDUTELLA: a P2P networking infrastructure based on RDF. In *International World Wide Web Conference (WWW)*, pages 604–615, 2000.
35. J. M. Kleinberg. Hubs, authorities, and communities. *ACM Computing Surveys*, 31(4es), 1999.
36. G. Grefenstette. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99 Translating and the Computer 21*, 1999.
37. 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.
38. F. Keller, M. Lapata, and O. Ourioupina. Using the web to overcome data sparseness. In *Proceedings of EMNLP-02*, pages 230–237, 2002.
39. K. Markert, N. Modjeska, and M. Nissim. Using the web for nominal anaphora resolution. In *EACL Workshop on the Computational Treatment of Anaphora*, 2003.
40. M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.
41. 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.
42. A. Mädche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, March/April 2001.
43. Googlism, 2003. <http://www.googlism.com>.
44. 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.
45. E. J. Glover, K. Tsioutsouluklis, 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.
46. S. Chakrabarti. Data mining for hypertext: a tutorial survey. *ACM SIGKDD Explorations Newsletter*, 1(2):1–11, January 2000.
47. P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web, 2003. Submitted for publication.

48. A. Maedche, G. Neumann, and S. Staab. Bootstrapping an ontology-based information extraction system. In J. Kacprzyk, J. Segovia, P.S. Szczepaniak, and L.A. Zadeh, editors, *Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web*. Springer, 2002.
49. A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers, 2002.
50. G. Bisson, C. Nedellec, and L. Canamero. Designing clustering methods for ontology building - The Mo'K workbench. In *Proceedings of the ECAI Ontology Learning Workshop*, 2000.
51. M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Research and Development in Information Retrieval*, pages 206–213. 1999.
52. P. Cimiano, S.Staab, and J. Tane. Automatic acquisition of taxonomies from text: FCA meets NLP. In *Proceedings of the International Workshop on Adaptive Text Extraction and Mining*, 2003.
53. M. Missikoff, R. Navigli, and P. Velardi. The usable ontology: An environment for building and assessing a domain ontology. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2002.
54. C. Behrens and V. Kashyap. The "Emergent" Semantic Web: A Consensus Approach for Deriving Semantic Knowledge on the Web. In *Semantic Web Working Symposium (SWWS)*, 2001.
55. K. Aberer and M. Hauswirth. P2P Information Systems. In *International Conference on Data Engineering (ICDE)*, 2002.
56. E. Cohen and S. Shenker. Replication strategies in unstructured peer-to-peer networks, 2002. ACM SIGCOMM.
57. B. Yang and H. Garcia-Molina. Designing a Super-peer Network. In *IEEE International Conference on Data Engineering*, 2003.
58. I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A scalable Peer-To-Peer lookup service for internet applications. In *ACM SIGCOMM*, 2001.
59. K. Aberer. Scalable Data Access in P2P Systems Using Unbalanced Search Trees. In *Workshop on Distributed Data and Structures (WDAS)*, 2002.
60. A. Rao, K. Lakshminarayanan, S. Surana, R. Karp, and Ion Stoica. Load Balancing in Structured P2P Systems. In *International Workshop on Peer-to-Peer Systems (IPTPS)*, 2003.
61. L. O. Alima, S. El-Ansary, P. Brand, and S. Haridi. DKS(N, k, f): A Family of Low Communication, Scalable and Fault-Tolerant Infrastructures for P2P Applications. In *International Symposium on Cluster Computing and the Grid (CCGrid)*, 2003.
62. A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. *Lecture Notes in Computer Science*, 2218:329–350, 2001.
63. M. Hauswirth, A. Datta, and K. Aberer. Efficient, self-contained handling of identity in peer-to-peer systems.