Social Annotation of Semantically Heterogeneous Knowledge

Matthias Nickles[1], Tina Froehner[2], and Gerhard Weiss[1]

[1] Artificial Intelligence/Cognition Group, Department of Computer Science,
Technical University Munich (TUM)
D-85748 Garching b. München, Germany, {nickles,weissg}@cs.tum.edu
[2] Research Center Knowledge Management (RCKM),
University of Applied Sciences Cologne
D-50678 Köln, Germany, tina.froehner@fh-koeln.de

Abstract. An important kind of tacit knowledge in the context of the Semantic Web are the *social communication structures* among heterogeneous knowledge sources and users. Communication structures heavily influence the way knowledge is generated and used, because in a context of distributed and autonomous information sources like in the Semantic Web, knowledge is constituted and adapted pragmatically through possibly conflictive communication processes. As a way to set social structures in relation to distributively acquired knowledge, this work proposes Open Ontologies and Open Knowledge Bases for the annotation of (first-level) knowledge with emergent social meta-data (*social reification*). Whereas traditional approaches to knowledge and ontology integration emphasize the consensus finding among the participants, Open Ontologies and Open Knowledge Bases explicitly model semantical heterogeneity in multiple levels of complexity reduction, and allow the probabilistic weighting of inconsistent knowledge resulting from their assertive weight in their communicative context.

Keywords: Semantic Web, Semantic Knowledge Annotation, Emergent Semantics, Ontologies, Social Data Mining, Computational Autonomy

1 Introduction

The Semantic Web can be seen as the most important effort toward large scale knowledge building and sharing in an open information environment. Decisive for the success of this long-term task is the provision of formalisms and mechanisms for the communication (i.e. symbolic interaction) of a very large number of distributed, autonomous knowledge sources and users. Shared ontologies and knowledge bases play a crucial role in this scenario, since they enable such communication, and knowledge acquisition among autonomous information sources is basically a communicative act.
Traditional approaches to the modeling and acquisition of ontologies and instance knowledge have several shortcomings in this respect as they seldom handle meaning dynamics, they seldom consider knowledge as being contextualized with intentions, processes and effects from the “outside world”, and they usually have no concept for the treatment of semantic heterogeneity (e.g. resulting from contradictions) that does not result

in a loss of information. Whereas approaches like *Emergent Semantics* [1], *Dynamic Ontologies* [2] and semantical ontology merging and alignment have caused significant improvements regarding some of these problems, semantical inconsistencies due to conflicting knowledge sources are almost always still taken for something which either should be avoided, or should be homogenized using, e.g. clustering techniques, or should be filtered out (e.g., using criteria like (dis-)trust or source reputation [5]).In demarcation from such views, it should be recognized, that semantical inconsistencies are not just unfavorable states, but that they are in real-world environments often unpreventable due to stable belief or goal conflicts [3] of knowledge sources, that they can even provide the knowledge user with valuable meta-information about the intentions, goals and social relations among the knowledge sources, and, if they have been made explicit and visible, that they can be prerequisites for a subsequent conflict resolution. In general, in the absence of a normative meaning governance, mechanisms for knowledge integration can only be a preliminary decision about the reasonable modeling of communicated knowledge artifacts, because within a heterogeneous group of autonomous knowledge sources and users, in the end each user can only decide for himself about the relevance and correctness of the given information, which provides a strong argument for the conservation of knowledge heterogeneity while integrating.
With this work we propose *Open Ontologies* and *Open Knowledge Bases* as a general approach to the *social* acquisition and annotation of knowledge for open environments like the Semantic Web (but also, e.g., for open P2P systems and Semantic Grids). It is primarily meant to introduce a fundamentally novel perspective rather than providing technical specifications.

2 Towards a Socially-Aware Semantic Web: Knowledge as a result of controversial mass communication

The Semantic Web has several key characteristics that make the acquisition and representation of knowledge complicate in contrast to closed systems and applications:

Openness Access, number and contributions of information sources are unrestricted for its major part.

Opaqueness of knowledge sources The intentions of knowledge providers are more or less unknown and their trustability and reliability cannot be guaranteed.

Opaqueness of users The impact of a knowledge contribution to the Semantic Web on its users is often hard to predict.

High dynamics and complexity There are very large, heterogeneous and fluctuating amounts of knowledge sources, knowledge contributions and users.

Highly controversial Several domains of web knowledge are highly controversial, e.g. in regard to politics, culture and product assessments by consumers. It seems to be extremely unlikely that such fundamentally divergent world views can be homogenized even in regard to general ontological concepts in the foreseeable future. Thus, semantic inconsistency is a reality knowledge management must cope with.

No authoritative background knowledge Decentralized structures and different background knowledge lead to a high diversity of individual knowledge.

Missing process knowledge Currently, the representation of machine accessible knowledge focusses on "knowledge end-products", not on the representation of processes that generate, modify or use knowledge.

These issues have in common that they rise mainly from the *autonomy* and *proactivity* of knowledge sources and users, being black- or gray-box actors with more or less opaque goals they pursue asserting or forming their individual world views. The way such autonomous entities (conceptually captured in the notation of *information agents* in this work) exchange information is *communication*. Although truly intelligent information agents are not expected to be widely spread on the internet in the foreseeable future, web knowledge can already be considered as communicative, because it is generated in order to influence its recipients and its intentionality and reliability is often unknown. This is even true if knowledge is communicated indirectly, tacitly or asynchronously using e.g. static web sites. Web knowledge is also contextualized with other web knowledge, and it can be agreed as well as denied by other knowledge facets (respectively their sources). Therefore, it appears to be reasonable to consider the Semantic Web as a very large, heterogeneous and hybrid system of interacting information agents (including humans), where information provided by humans and computationally generated knowledge co-exist. Due to the highly distributed character and the heterogeneity of this partially "wild grown" multiagent system, besides agreed protocols and formalisms, shared ontologies and knowledge bases are expected to be extremely useful to enable and improve mutual understanding and interactivity. Because knowledge on the Semantic Web is not only required in order to improve communication, but, maybe even more important, is an emergent outcome and constituent of communication, the key properties of communication need to be taken into account when it comes to building such ontologies and knowledge bases. Thus, viewing the Semantic Web as a system of directly or indirectly communicating information agents, we propose a communication-oriented paradigm, which has several implications for the retrieval and modeling of distributed knowledge. Most important, knowledge management for the Semantic Web needs to cope with the fact that the meaning of information on the web can never be determined for sure in general, might change, and might be constituted from the possibly conflicting opinions of large sets of knowledge sources.
The primary goal of Open Ontologies and Open Knowledge Bases is to make the knowledge contributions of large, fluctuating and possibly conflicting sets of autonomous sources usable in a computational sense, i.e. to provide computationally accessible meta-data to the users even if such socially accumulated knowledge is inconsistent or unreliable (especially in the absence of trustability). For this purpose, the *social layer of knowledge* on the web needs to be found and made explicit by means of semantic annotation to the web users. In particular, the technical openness of shared knowledge like ontologies and the comparability of distributed, local knowledge needs to be improved, knowledge artifacts need to be interpretable as parts of *communication processes* (with induced relationships like assertion, agreement, contradiction, request, revision, specialization, generalization...), and the complexity of socially accumulated knowledge needs to be reduced *without* the need to come to a consent among the participants and with as less loss of information about social heterogeneity as possible.
Largely neglecting these aspects, most of the current efforts in order to build the Se-

mantic Web concentrate on the specification of languages and tools for the modeling of agreed, homogeneous knowledge, and research is just beginning to take into consideration phenomena like the social (i.e. communicative) impact of resource descriptions, conflicting opinions, information biased by e.g. competing commercial or political interests, and inconsistent or intentionally incorrect information. Bringing information (e.g. via web sites or web services) into the web is in fact a social act, and the relationship between informational artifacts on the web is communicative (i.e. specifying, agreeing, contradicting...). This can of course produce intentional and unavoidable inconsistencies (e.g. company interests versus customer interests or various conceptualizations due to differences in culture). If these are ignored, or filtered out, ranked/recommended or homogenized too early (e.g. applying trust), important information for the user or the application might be lost. In order to make this important information available, we propose the following:

- Knowledge facets on the web like meta-data annotating web pages must be seen as *subjective belief assertions* of rational intelligent black-box agents (artificial agents as well as human users). They are created with certain intentions which are more or less hidden and are situated within action processes in order to make the successful assertion of this particular "truth" more likely (with advertisement as the most usual case, but also e.g. user recommendations regarding products and political statements, and even lexicon entries).
- Knowledge heterogeneity needs to be made *explicit*. Since knowledge sources are more or less opaque with hidden belief and goals, the need for instruments that enable the comparison of different standpoints becomes more important for knowledge users.
- Knowledge heterogeneity needs to be *explained*. Publication of knowledge on the web is an assertive act that is embedded within a pragmatical context of reasons and implications. In fact, the meaning of knowledge cannot be determined without considering this pragmatical context [8].
- The representation of web knowledge has to comprise *uncertainty* on the social level. Knowledge assertions uttered from black- or gray-box agents are basically more or less unreliable, and they might be misleading. One way to ensure reliability is the establishment of trust relationships. But to establish trust, one has to accumulate experiences and weigh different opinions. In addition, heterogeneous knowledge contributions of large numbers of agents need to be generalized using stochastical methods in order to reduce their complexity and to make practical use of them (e.g. to derive average opinions). From the viewpoint of a knowledge consumer, even though someone cannot say how things "are" in reality, a knowledge base must provide an approximate value for her decision finding.
  Whereas it is already widely agreed that the statements of human individuals can only be transferred to machine understandability with a more or less degree of uncertainty, the need for the use of probabilistic and approximate representation formalisms in order to model collectively constituted knowledge on the web is still largely neglected.

Figure 1 shows the semantical levels proposed by Tim Berners-Lee for the structure of the forthcoming Semantic Web, with extensions (red/light gray font) we recommend

for some aspects of this concept in response to the mentioned issues. In particular, it appears to be inevitable to us to provide formalisms and calculi that explicitly consider semantically heterogeneous meta-data like resource descriptions and ontologies created from the contributions of multiple sources that compete for the assertion of their individual "truths" and interests. Of course, the Semantic Web is already open, but for a broad acceptance and to provide value to its users, we strongly suppose that communicative (i.e. social) relationships among closed "islands" of knowledge like contradiction or agreement need to be made explicit formally and technically as part of the layers of a "socially-aware" Semantic Web, using a concept called *social reification* (cf. next section). In this regard, the empirical derivation and stochastic modeling of open meta-data seems inevitable if the set of knowledge sources is either very large, or fluctuates, or generates indefinite information.

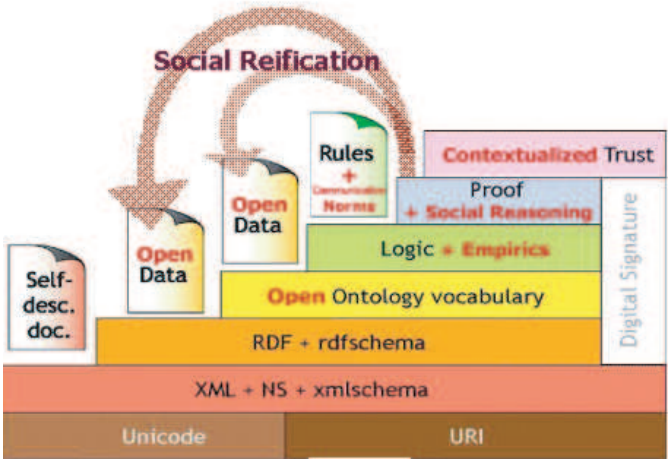

Fig. 1. A socially-aware Semantic Web

3 Open Ontologies and Open Knowledge Bases

3.1 Characteristics

Formal ontologies and knowledge bases are traditionally defined as agreed descriptions of certain domains which serve as common ground for distributed tasks like knowledge exchange, modeling and user information. This understanding leads to difficulties if the informational input these media are build from is likely to be intentionally inconsistent, and there either does not yet exist enough meta-knowledge like trust to identify and filter out "inappropriate" or "wrong" data a priori, or there does not even exist a concept of global inappropriateness or correctness at all. On the other hand, sound and agreed ontologies are doubtless an inevitable prerequisite for efficient knowledge creation, representation and exchange, whereby we consider implicit and emerged ontologies and schemata (e.g. in the context of semi-structured data modeling) to be

such ontologies too. Of course, ontological heterogeneity can be overcome by means of techniques like the renaming of inconsistent concepts, and in general, inconsistent knowledge can be made consistent providing appropriate *truth contexts* [10]. However, such solutions often generate redundancy instead of an informational benefit for the knowledge users, or lead to difficulties finding other than trivial annotations like "In the belief of agent x, the following is true:...". OO&OKB aim at the solution for this dilemma by embedding conceptual knowledge facets gained from a heterogeneous set of self-interested autonomous knowledge sources (e.g. information agents or humans) within contextual information about their communicative (i.e. social) origin, impact, and relationships (e.g., contradiction, approval, revision or specification) to other communicated knowledge facets (which can be communicated by means of formal communication languages, but also be derived from, e.g., structured, semi-structured or natural language documents) and their sources. Doing so, in OO&OKB, knowledge as it can be found in conventional knowledge or ontology bases, is *lifted* to the social level and thus to a level where the sources and the users of the ontology are likely to achieve an agreement with the *social assessments* of possibly inconsistent and uncertain facts (e.g., if $agent_1$ contradicts $agent_2$, both usually agree that they do so!). The judgement of assessed facts is then a subsequent task based on rich social knowledge instead of binary distinctions like to trust or not to trust particular agents. *OO&OKB are thus dynamic communication media which receive their content from the communication of multiple autonomous information sources and users, and provide a dynamic representation of socially annotated heterogeneous knowledge.*

Communication is here not so much to be understood as the exchange of symbols with a fixed meaning, but the other way round as a means to generate supra-individual meaning from interrelated interactions among black- or gray-box agents (i.e., agents with more or less unknown internal states, cognition and goals). The practical consequences arising from this are that OO&OKB need to be continuously adapted to new information, and the processes of creation, contextualization and interpretation of knowledge are integral aspects of OO&OKB themselves. In addition, communication among multiple agents likely requires mechanisms for the generalization of emergent meaning, since otherwise the complexity would grow too large due to the sheer number of individual knowledge contributions. Generalization is also a way to make OO&OKB look like homogeneous ontologies or knowledge bases if necessary, because at its highest level, generalization causes semantical homogenization among contradicting knowledge sources. Summing it up, Open Ontologies and Open Knowledge Bases have the following characteristics:

Openness No (or as few as possible) initial assumptions are made regarding the benevolence, trustworthiness, relevance, informedness and cooperativeness of its sources. Nevertheless, information about e.g. (dis-)trust and knowledge (un-)reliability is likely derivable from Open Ontologies and Open Knowledge Bases, since these are special cases of social structures.

Dynamical derivation from communication OO&OKB are emergent from and evolving with ongoing communication (e.g. agent interaction, but also asynchronous, indirect or tacit communication e.g. via the semantically interrelated contents of web sites) of knowledge sources and knowledge users to assert (deny, specify...) information and to express and specify informational needs and expectations. Social

background knowledge (existing social structures like laws) can be included in the derivation process.

Explicitness and social annotation of semantical heterogeneity OO&OKB *maintain* semantical inconsistencies arising from contradictions and conflicts, and contain (consistent) annotations of (conceptual or instance) knowledge with meta-information about its *social meaning* within the course of communication.
This concept is related to context logic [10], but in contrast does not aim for the provision of logical truth contexts. Rather, social annotations state the sound social meaning of subjective statements without judging them as true or false.

Multiple, probabilistically modeled levels of social generalization They allow multiple, application-dependant levels of generalization of social concepts (like the generalization of single information agents as *agent roles* or groups, allowing to derive "average" or shared group opinions from the communications of multiple knowledge sources), weighting the degree of inconsistency and the degree of details of the annotating meta-information (cf. section 4). Generalization can also help to overcome privacy issues by averaging individual information contributions.

3.2 Social Reification

OO&OKB contain as first-order objects knowledge facets that have the form 1st-level knowledge ← 2nd-level knowledge, where 1st-level knowledge partially describes a domain concept in the same way as within usual ontologies (or instances of such concepts, respectively, for Open Knowledge Bases), but probably in an inconsistent way regarding other 1st-level knowledge in the same ontology. Since Open Ontologies are primarily an abstract meta-concept build upon conventional approaches for the representation of conceptual knowledge, we do not constrain or specify the sort of concrete entities that are to be "wrapped" within an Open Ontology (Open Knowledge Base) or at the content level of agent messages, like first-order logical statements, classes or frames. For the same reason, we do also not make any assumptions relating to ontology domains or concrete areas of application here. In contrast to 1st-level knowledge, 2nd-level knowledge (also called *social knowledge*) depicts the social context of 1st-level knowledge, the latter taken as generated from a communication act of an autonomous source of knowledge. This kind of annotation of 1st-level knowledge with 2nd-level knowledge we call *social reification*. A quite trivial kind of social reification is *quoting* (e.g., 'Sue says: "..."'), but in general, all kind of information which describes how and to what effect certain data is produced within a process of communication can be informally understood as 2nd-level knowledge (and, of course, we can apply social reification recursively, i.e. annotate 2nd-level knowledge with 3rd-level knowledge as in 'Sue says: 'Tom says: "..."' and so on). The most elementary forms of such social meta-data are considered agent speech act types like assertion, denial or query, inducing relations among single communication like 'Sue contradicts Tom's statement saying "..."' and rich 2nd-level knowledge types such as knowledge source and user profiles and even complex social systems like organizations. In an empirical communication model [8] symbolic communicative acts gain their semantics from their expected effect on the subsequent trajectory of communications, which can be learned empirically from past

interactions (although we recommend empirical semantics to disregard mentalistic details which are unknown for autonomous agents and allow for the handling of uncertain meanings, the usage of such a semantics is not required to define an Open Ontology or an Open Knowledge Base). Because meaning is contextualized by the situation (history) of the respective act occurrence, in general 2nd-level knowledge describes communication processes (this applies even to simple quotations: In Sue says: "...", "Sue" is in fact just an abbreviation for the pragmatic impact utterances from Sue are expected to have. This concept is not meant to be a replacement for the usage of e.g. first-order predicate logic for Web reasoning, but instead as a completion which could be introduced gradually. E.g., the Resource Description Framework *RDF(S)* and *Notation3* already have elementary reification capabilities, which could be used for elementary social annotations (e.g. collective rating of RDF statements) as described in [6, 7], but would require an appropriate specification of this kind of usage. In the following, we will outline a more ambitious approach to this issue.

4 Derivation of Open Ontologies and Open Knowledge Bases

Open Ontologies and Open Knowledge Bases need to be learned from the observation of communication processes. The technical requirements for this learning process are:

- information agents or other knowledge sources (e.g. peers in a P2P network, or passive resources like web documents) able to communicate and query 1st-level knowledge facets. In case of software agents, this can be done by means of a formal agent communication language (since OO&OKB do not require agent cooperativeness, speech act performatives used for collaboration like negotiation are not required, although they would be useful).
- a facility for the acquisition of OO&OKB from the observation of above communications, e.g., a dedicated middle agent within the infrastructures of the respective application, called a *semantics observer* (cf. figure 2).
- optionally, a pre-defined content of the Open Ontology or Open Knowledge Base, in order to speed up the learning process of the semantics observer, and to avoid the bootstrapping problem known from e.g. recommender systems, or to set static social structures like norms
- a facility for the low-level storage and querying of persistent knowledge (e.g., a database management system).
- optionally, a facility for the social reasoning upon the 2nd-level knowledge within the Open Ontology or Open Knowledge Base. respectively (to deduce new facts like "Sue is likely to contradict or specify Toms information", but also to derive trust relationships among the participants subsequently. Here, known techniques as described in e.g. [5] can be used).

The acquisition of OO&OKB comprises the following main tasks, which have to be performed in a loop as a continuous, incremental learning process for the whole period of agent communication (please find details in [9]).

1. Observation of communication. In addition, implicit or tacit communication might needs to be made explicit beforehand.

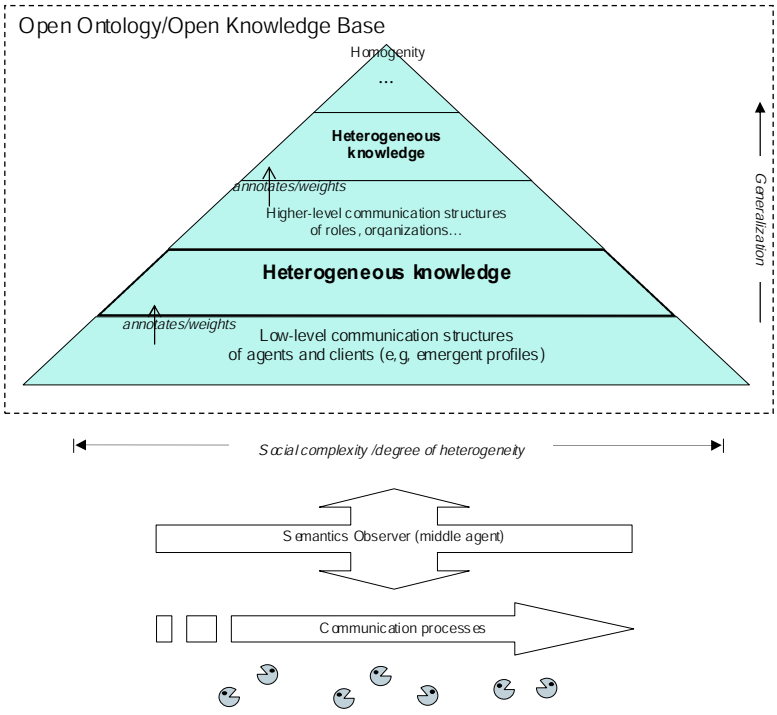

Fig. 2. Emergence and Generalization of Open Ontologies and Open Knowledge Bases

2. Derivation and/or adaptation of 2nd-level knowledge according to the respective semantical model (e.g. empirically)
3. Stochastic generalization of 2nd-level knowledge
4. Social reification and generalization of 1st-level knowledge
5. Alignment with given, obligatory 1st-level knowledge (e.g. a normative top-level ontology) or normative 2nd-level knowledge (e.g. laws preventing certain utterance of certain information), if necessary.

As mentioned earlier, OO&OKB also require the generalization of meaning in order to reduce their complexity (cf. figure 2). Generalization as a task in this sense has two steps: 1) the merging of 2nd-level knowledge, 2) the subsequent merging of related 1st-level knowledge facets. Typically, 1) comprises the merging of similar social processes to interactions patterns, and the combination of multiple similar behaving agents to social groups or social roles. After applying such generalization rules to 2nd-level knowledge, the annotated 1st-level knowledge needs to be merged accordingly. If, for example, multiple agents forming a single social group make inconsistent assertions, within the Open Ontology (Open Knowledge Base) each of these assertions obtains a probabilistic weight expressing the degree of expected approval this assertions gets from the role or group as a whole (calculated, e.g., from the frequency this assertion

has been uttered by different agents within this role or group) [7, 6]. We propose the usefulness of a co-presence of multiple levels of generalization, tailored to the desired levels of heterogeneity of the respective Open Ontology or Open Knowledge Base (cf. figure 2). Of course, the concrete representation and degree of heterogeneity that should be maintained strongly depends from application and user needs.

5 Conclusion

There is an obvious and rapidly growing need for knowledge-based systems capable of running in open environments like the Semantic Web with autonomous knowledge sources and users, given the increasing inter-operability and inter-connectivity among computing platforms. On the one hand, knowledge bases and ontologies should provide a stable ground for user information, agent and user communication and subsequent knowledge modeling, on the other hand, in open environments concept descriptions tend to be semantically inconsistent, they emerges from a possibly very large number of competing subjective beliefs and goals, and a priori there might be no such thing as a commonly agreed “truth” (in the “real world”, not even a discursive trend towards such a thing can be assumed). To cope with these two contradictory aspects must be a core concern of the communication-oriented paradigm of knowledge modeling and management, and is the basic motivation underlying the work described here. To this end, we have proposed Open Ontologies and Open Knowledge Bases as a fundamental step towards the modeling and representation of socially-induced knowledge heterogeneity for the Semantic Web.

References

1. A. Maedche, F. Nack, S. Santini, S. Staab, L. Steels. Emergent Semantics. IEEE Intelligent Systems, Trends & Controversies, 17(2), 2002.
2. J. Heflin, J. A. Hendler. Dynamic Ontologies on the Web. Procs. of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, p. 443 - 449, 2000.
3. R. Dieng, H.J. Mueller (Eds.). Conflicts in Artificial Intelligence. Springer, 2000.
4. http://www.w3.org/2001/sw/meetings/tech-200303/social-meaning/
5. J. Golbeck, B. Parsia, J. Hendler. Trust Networks on the Semantic Web.Proceedings of Cooperative Intelligent Agents, 2003.
6. M. Nickles, G. Weiss. A framework for the social description of resources in open environments. Procs. of the Seventh International Workshop on Cooperative Information Agents (CIA, pp. 206-221). LNCS Volume 2782. Springer, 2003.
7. M. Nickles, Towards a Multiagent System for Competitive Website Ratings. Research Report FKI-243-01, Technical University Munich, 2001.
8. M. Nickles, M. Rovatsos, G. Weiss. Empirical-Rational Semantics of Agent Communication. Procs. of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS’04), New York City, 2004.
9. M. Nickles, T. Froehner. Social Reification for the Semantic Web. Research Report FKI-24x-04, Technical University Munich, 2004. To appear.
10. A. Farquhar, A. Dappert, R. Fikes, W. Pratt, Integrating Information Sources using Context Logic. Procs. ofthe AAAI Spring Symposium on Information Gathering from Distributed Heterogeneous Environments, 1995.