UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

BENCHMARKING METHODOLOGY
FOR GOOD ENOUGH ANSWERS

Coordinator: Pavel Shvaiko

with contributions from: Fausto Giunchiglia, Alan Bundy, Paolo Besana, Carles Sierra, Frank van Harmelen and Ilya Zaihrayeu

January 2008

Technical Report # DISI-08-003

OpenKnowledge

FP6-027253

Benchmarking methodology for good enough answers[1]

Coordinator: Pavel Shvaiko[1]
with contributions from
Fausto Giunchiglia[1], Alan Bundy[2], Paolo Besana[2], Carles Sierra[3],
Frank van Harmelen[4], Ilya Zaihrayeu[1]

[1] Department of Information and Communication Technology (DIT),
University of Trento, Povo, Trento, Italy
{pavel|fausto|ilya}@dit.unitn.it
[2] The University of Edinburgh, Edinburgh, UK
bundy@inf.ed.ac.uk, p.besana@sms.ed.ac.uk
[3] Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
sierra@iiia.csic.es
[4] Vrije Universiteit Amsterdam, The Netherlands
Frank.van.Harmelen@cs.vu.nl

Report Version: final
Report Preparation Date: 04.01.2008
Classification: deliverable 4.2
Contract Start Date: 1.1.2006 Duration: 36 months
Project Co-ordinator: University of Edinburgh (David Robertson)
Partners: IIIA(CSIC) Barcelona
Vrije Universiteit Amsterdam
University of Edinburgh
KMI, Open University
University of Southampton
University of Trento

[1] The originally planned title of this deliverable as from the project proposal was "Benchmarking methodology for answer quality". However, this new title better reflects current contents of the deliverable and needs of the project, and therefore, is used here.

Abstract

This document discusses (i) a methodology for benchmarking good enough answers as well as (ii) the state of the art in the related areas that address the issues of evaluating quality of query answering in peer-to-peer systems.

1 Introduction

The OpenKnowledge (OK) system is a peer-to-peer (P2P) network of knowledge or service providers. Each computer in the network is a peer which can offer services to other peers. OK is viewed as an infrastructure, where we only provide some core services which are shared by all the peers, while all kinds of application services are to be plugged on top of it. These plug-in applications are called the OK Components (OKCs) [4]. Notice that the OKCs link services to the OK infrastructure and may not actually contain the services themselves.

Interaction between OKCs is a very important part of the OK architecture. By using the Lightweight Coordination Calculus (LCC) [14], developers are able to define the Interaction Models (IMs) that specify the protocol that must be followed in order to offer or use a service. OKCs are the ones in charge of playing the IM roles. Different peers select to play different roles in an IM, which is then run to achieve the goals of these peers.

The combination of an IM and the peers assigned to each of its roles is called a *configuration*. The purpose of the good enough answer (GEA) mechanism [10] is to find good enough configurations, i.e., configurations that achieve the purposes of the peers with a reasonable investment of resources in their construction. It does this with the aid of two heuristic measures, namely: (i) of the matching measure between the IM roles and the peers' capabilities [8] and (ii) of some measure of trust in the peers, based on their historical achievements [10]. The goal of this deliverable is to review state of the art in the related areas and provide a benchmarking methodology for good enough answers.

The rest of the deliverable is structured as follows. Section 2 discusses the related work. Section 3 proposes a methodology for benchmarking good enough answers. Finally, Section 4 summarizes the findings of the deliverable and outlines future work.

2 State of the art

A large body of literature has already been devoted to the benchmarking topic, see, for example, [23, 22, 24, 2, 7], as well as the Text REtrieval Conferences (TREC)[2] [28] and Ontology Alignment Evaluation Initiative[3] [5] being notable examples of the ongoing evaluation campaigns. A benchmark is a well-defined set of tests on which the

[2] `http://trec.nist.gov/`

[3] `http://oaei.ontologymatching.org/`

results of a system or subsystem can be measured [2]. It should enable the measure of the degree of achievement of proposed tasks on a well-defined scale. It should be reproducible and stable, so that it can be used repeatedly for (i) testing the improvement or degradation of a system with certainty and (ii) situating a system among other systems. Building benchmark suites is highly valuable not just for groups of people that participate in planned evaluations but for all the research community, since system designers can make use of these at any time [6].

In the rest of this section we first discuss the related work concerned with query answering in P2P settings (§2.1), which we believe is the most relevant topic with respect to our investigations here [35]. Then, we overview the lessons learned out of §2.1 with respect to good enough answers (§2.2).

2.1 Query answering in P2P

Preliminaries. An important aspect of query answering is the *quality of answers* to user queries in a P2P system, which is fundamental to the quality of service provided by the system.

The notion of *data quality* is well understood in the context of centralized information systems (ISs) [20, 32, 29, 30]. Quality assessment in these approaches relies on the assumptions that data can be accessed and evaluated centrally, and that any piece of data in the system can be retrieved upon a user request. Therefore, the *quality of query answers* in these systems can be directly evaluated by assessing the *quality of the data*. Moreover, these assumptions give rise to some standard data quality dimensions, such as correctness and completeness, which can be objectively evaluated given the global view of the data in the system.

In a P2P IS, the quality of query answers depends not only on data quality of particular local information source, but also on the *quality of the correspondences* [34, 6] relating the local information sources. In fact, the semantics of a query can be distorted and/or data loss and misinterpretation can take place if some correspondences are imperfect. Thus, for instance, the quality of query results may degrade due to their incorrect and/or incomplete translation when propagating from one peer to another.

Standard data quality metrics cannot be directly applied in the P2P settings due to its decentralized, dynamic and subjective nature. In fact, in a P2P IS query results are often incomplete (under the classical interpretation of the notion of completeness [29]) due to the fact that not all the data in the P2P IS are accessible to the peer where a user query is submitted. Furthermore, it is hard to evaluate the correctness of a query result due to the fact that each peer maintains its *subjective* view on the data. In fact, the correctness of data[4] is referred to the extent in which the representation of some part of the real world, as it is inferred from the data stored in the IS, coincides with the *user's view* of this part of the world [29]. Thus, the same data can be considered as correct by one peer and as incorrect by another.

[4] In the data quality literature correctness is often called accuracy [29].

Another data quality metric, *consistency*, is defined in [21] as *the degree to which the values of the attributes of an instance of a schema element satisfy the specific set of semantic rules defined on the schema element*. The problem with this metric in the P2P settings is that different peers may define semantic rules on data differently, and, therefore, the same data can be considered as consistent by one peer and as inconsistent by another.

Similar arguments can be made about other data quality metrics, such as reliability, timeliness, relevance, currency, availability, and others [31]. Thus, the standard data quality metrics have to be *reconsidered* in order to reflect the distinct characteristics of P2P ISs. Particularly, there are (at least) three kinds of unpredictable runtime factors, peculiar to P2P ISs, which influence the answer to a given user query, and, therefore, which also influence the quality of query answers [11]:

- *Network (dependent) variance:* the P2P IS changes over time. Peers may change the data of their information sources, change their ontologies, redefine correspondences, finally, new peers may join the P2P IS and existing ones may leave it. Therefore, the same query submitted to the same peer but at *different times* may yield different answers and of different quality.

- *Peer (dependent) variance:* peers define correspondences differently from other peers. Therefore, the same query submitted at the same time but by *different peers* will result in different query propagation graphs. Consequently, the query results may be different and of different quality.

- *Query (dependent) variance: different queries* submitted to the same peer and at the same time may result in different query propagation graphs, and, therefore, may produce different results and of different quality.

Network dependent variance forces us to assume that data quality metrics must be evaluated in the context of the current state of the P2P IS. Peer dependent variance forces us to assume that they must be evaluated in the context of a particular peer. Finally, query dependent variance implies that data quality is also sensitive to the context of each particular query. As it can be seen, the three variances make it very challenging to develop a unified evaluation methodology for measuring the quality of query results returned in a P2P IS.

Related works. The problem that standard metrics for measuring quality of answers, such as correctness and completeness, cannot be directly applied in the P2P settings has been recognized by the research community [11, 17, 13, 19, 15, 33, 25]. However, very little has been done so far for the development of methodologies to assess the quality of query answers in heterogeneous schema-based P2P systems [35].

As a step from centralized to distributed but still integrated systems, the works in [18, 21, 17] discuss data quality issues related to query answering against a global schema. Particularly, [18] discusses how to answer a user query using only a subset of relevant relational sources based on predefined data quality criteria. The work reported in [21] discusses data quality issues in the context of cooperative information systems, and it improves on the work in [18] by allowing the association of meta-data to the quality values, which allows it to improve the quality of query answers. Finally, [17] discusses how the data quality criteria shift, when going from database integration solutions to the integration of autonomous information sources. Apart from this, the work presented in [17] shows that in a large scale and autonomous environment, such as the web, users cannot expect correct and complete query answers, but they accept incomplete and partially incorrect answers.

According to [18, 21], they are the first of a few works which shift the focus of the data quality problem from single centralized systems to distributed integrated systems[5]. To the best of our knowledge, none of the existing works addresses the problem of the quality of query answers in P2P ISs in enough detail. However, some first preliminary works have been reported. For instance, [33] proposes to measure the quality of query results by introducing the measure of (user) satisfaction, as some minimal number of query results (e.g., 10, 50), and time to satisfaction (e.g., 2 sec.), namely time that elapsed from the moment of query submission to the moment when the predefined number of query results is computed. The work in [15] proposes a list of quality dimensions for semantic overlay P2P networks. The list of dimensions include completeness, accuracy, response time, and amount of data. While giving intuitive definitions of the dimensions, due to their application in the P2P settings, the authors do not discuss how these dimensions are different from those defined in the standard data quality literature.

The problem of the quality of correspondences among peer schemas has been addressed in [1, 12, 13]. Specifically, in [1] the authors discuss how the quality of correspondences in a schema-based P2P system can be assessed when some query arrives in a loop to the same peer. The intuition is that the query in the loop should refer to the same attributes in the schema of the peer, as the original query, which initiated the loop. If this is not the case, then some of the correspondences in the loop are incorrect. However, in their analysis, the authors rely on the assumption that each peer's schema is represented with a *single* relational table. Such an assumption cannot be made in the autonomous P2P environment. In [12], it is introduced the idea that correspondences can be differentiated on the base of how well they serve their primary purpose. For instance, among the various possible correspondences from the schema of a product retailer to the schema of a product merchant, those correspondences are better which ultimately "sell" more products. The authors of [13] introduce some basic quality metrics, such as completeness and relevance, in the context of a relational P2P IS, and show how certain properties of correspondences (mappings), such as the presence or absence of projections and selections, may affect the completeness dimension of the quality.

[5] See [31] for a survey of data quality evaluation approaches for centralized ISs.

2.2 Good enough answers

A good-enough answer is *an answer to a user query which serves its purpose given the amount of effort made in computing it* [11]. There are two key points in this definition: (i) that a query answer should serve its purpose, and (ii) that a query answer should be parametric on the initial effort. Let us discuss them in more detail [35]:

Purpose-driven: users submit queries in a specific context, giving the queries a specific purpose. Since information sources are developed *locally* taking into account their intended use, the ontologies, w.r.t. which a user query is submitted, usually correspond to the query context and serve the query purpose. However, as discussed earlier, different information sources are developed independently and, in most cases, they represent related concepts differently and assume a different context of their use. The latter particularly means that the different information sources can be heterogeneous at the application level. Therefore, when a query is propagated from one peer to another, it is likely to be evaluated in a different context, and to yield query results, which do not necessarily match the original query purpose. Given this, the user may receive an answer from another peer, which only to a certain extent serves the purpose of the query and corresponds to the initial context. Nevertheless, this result may still be useful for the user, and, when combined with results coming from other peers, it can better meet the user's needs.

Effort-driven: the main motivation for peers to join a P2P system consists in their immediate access to a large pool of data sources in return for a relatively little effort of setting up a small set of acquaintances, namely, the peers that a given peer knows about. The same holds for query answering, where the basic criterion becomes whether a query result justifies the effort, which the user made for its computation. The intuition is that users are likely to be willing to invest more in setting up acquaintances with the purpose of getting higher quality answers, and they will be happy with some answer if they set up acquaintances with much less care. In this respect, in P2P settings it is enough that every peer establishes and maintains a small set of acquaintances for all the peers to benefit from collective query answering. Also, users express their queries w.r.t. the ontologies they own, and which they know well, and, therefore, in order to send a global query to the P2P system, only knowledge of the local ontology is required. Finally, the fact that setting up acquaintances is still a cost[6] suggests that peers should set up and change acquaintances *gradually*, reusing the same acquaintances for many user queries, thus "amortizing" their setup costs.

Requirements for being good enough depend on the application domain. For instance, a partially incorrect result in the medical care domain may potentially lead to

[6] Note that the cost of participation in a heterogeneous schema-based P2P system is considered to be higher than the cost of participation in schema-less or homogeneous schema-based P2P systems.

serious consequences, such as improper patient treatment. A partially incorrect and incomplete result may be good enough in the tourism domain as long as it serves the user needs, e.g., it correctly provides important data such as hotel stay costs.

There is no such notion as response time for good enough answers. Unlike the traditional centralized information systems, where a single (and probably) correct and complete answer is provided to a user query, in P2P systems users receive a continuous sequence of query answers until the query answering is complete. In such settings, the first few results may be already good enough for the user, or a thorough query answering has to be performed before the overall answer becomes good enough.

For an answer to be good enough, it is not absolutely necessary that it satisfies all the constraints specified in the query. In fact, results which are not an exact match, but which are an approximate match to the requirements specified in the query, can still serve the purpose of the user, see, e.g., [3, 26]. For instance, when looking for a flat with a certain number of rooms and of a certain cost, the user may also want to consider flats, whose cost is slightly higher, but which have one more room.

The subjectivity dimension discussed above can be partly modelled with the computational notion of trust [16]. In fact, users tend to consider data and information from trustworthy sources more important than those coming from unknown or untrustworthy ones.

These general ideas discussed above have been further elaborated in the context of the OK system and concrete algorithms for computing good enough answers have been proposed in [9, 10]. In particular, the purpose of the good enough answer mechanism [10] is to find good enough configurations (see §1) that achieve the purposes of the peers with a reasonable investment of resources in their construction. In OK it is done with the aid of two heuristic measures, namely: (i) of the matching measure between the interaction model roles and the peers' capabilities [8] and (ii) of some measure of trust in the peers, based on their historical achievements [10].

3 A benchmarking methodology for GEA

In designing a benchmarking methodology for GEA we have to address the following questions:

What claims or hypotheses are being evaluated? We want to be able to say that the configurations found by the GEA are in some sense good ones. However, since they are not intended to be perfect, but only good enough, there is no absolute measure we can impose.

What is the reference against which the performance of GEA is being compared? As from the previous discussions, we must appeal to human judgement about how good GEAs are. An option would be to run the GEA configurations to completion and then use some automated measure of the results. It would be good if that were possible, but there is the possibility of bias if the success measures were unconsciously influenced by the heuristic measures used to construct the configuration, i.e., circularity may creep into the assessment. This suggests a human double blind comparison. Note that GEA is unlikely to come up with identical configurations to the human produced ones, so we must usually compare *different* configurations.

Is there a control situation to help determine what would constitute an acceptable performance against this reference? Similar to the previous item, we may also have to appeal to human ability, i.e., to provide manually produced good enough configurations.

How can the evaluation be assured to be objective? For example, not unintentionally biased in favor of the GEA. To achieve objectivity we must either automate the comparison of the automatically and manually produced configurations or use a double blind assessment. We will also want to use a statistical test to ensure that any observed differences are significant and not due to noise.

Bearing the pervious analysis in mind, the following benchmarking methodology is proposed:

Step 1. For each task in a set of N, a GEA automatic and a human manual configuration will be produced and compared. N is to be determined in a pilot experiment to be a compromise between statistical accuracy and achievable human resources. In both the control and the experimental situation a time limit should be imposed to ensure that the configurations constructed are good enough rather than perfect. The time limits should be estimated in a pilot experiment.

Step 2. Several human experts, who were not involved in configuration construction, should then compare each pair of automatically and manually produced configurations in a double blind set-up. A 5-point ordinal scale should be sufficient for the comparison. 5 points is generally regarded as providing sufficient variability without introducing spurious accuracy. The comparisons may be based on the configurations themselves, the results of their runs or both. Pilot experiments should be run to determine which is most effective, i.e., which the evaluators report as being most feasible and comfortable for them.

Step 3. The hypothesis to be tested could be one of the following:

(a) *GEA produces statistically significantly better results than human experts.*

(b) *GEA produces results that are statistically indistinguishable from those of human experts.*

To determine which of these hypotheses is to be evaluated some preliminary exploratory analysis is required. For example, by plotting the paired rankings on a scatter-plot, i.e., if, for some task, x is the ranking of the automatic configuration and y is the ranking of the manually produced one then $\langle x, y \rangle$ is a point on a graph, where the axes run from 1 (poor) to 5 (excellent). If (b) were true then we would expect these points to cluster around the 45° line. If (a) were true then most of the points would lie *below* this diagonal line. If most of the points lie *above* the diagonal then we would evaluate (b).

Step 4. Conduct a statistical test of significance for ordinals, see, for instance, [27].

4 Conclusions and future work

In this deliverable we have discussed the state of the art from the related areas that address the issues of quality assessment for query answering in peer-to-peer systems. We proposed a methodology for benchmarking good enough answers. Specifically, it uses humans to produce the control examples and a different set of humans to assess the results. To avoid bias, a double blind comparison of the control and experimental situation is proposed. Either the configurations or their results or both could be compared. A 5-point ordinal scale should provide appropriate discrimination. The size of the test group needs to be a compromise between providing sufficient statistical accuracy and the investment of human resources in producing the control groups. Various pilot experiments are proposed to set the parameters of the evaluation.

Future work includes building the actual benchmarks within the OK project testbeds, such as emergency response and bioinformatics, and conducting the corresponding case studies in order to evaluate the GEA mechanism on those benchmarks. Finally, we acknowledge the fact that the benchmarking methodology proposed in §3 do not address all the issues of quality evaluation for query answering in P2P raised in §2. However, as we already noted it turns out to be very challenging to develop a unified benchmarking methodology for good enough answers, though since GEAs are configurations the situation may not be so bleak as the general situation discussed in §2. In particular, we can evaluate configurations as to whether the LCC code can be successfully run. A bad enough configuration may just be ill-formed and unrunnable. Also, we may have some criteria to evaluate the result of running such a configuration, e.g., we may end up buying a car which meets our specification. In contrast, in P2P query answering, we may have no objective criteria for assessing how good enough it is. Therefore, we follow an incremental approach here by first implementing the methodology proposed in order to obtain “experimental hooks” that should allow us to further detail it. We plan to report the updates on the benchmarking methodology for GEA together with the forthcoming OK case studies.

References

[1] Karl Aberer, Philippe Cudré-Mauroux, and Manfred Hauswirth. The chatty web: emergent semantics through gossiping. In *Proceedings of the 12th International Conference on World Wide Web (WWW)*, pages 197–206, Budapest (HU), 2003.

[2] Raúl García Castro, Diana Maynard, Doug Foxvog, Holger Wache, and Rafael González-Cabero. Specification of a methodology, general criteria, and benchmark suites for benchmarking ontology tools. Deliverable D2.1.4, Knowledge web NoE, 2004.

[3] Surajit Chaudhuri and Luis Gravano. Evaluating top-k selection queries. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)*, pages 397–410, Edinburgh (UK), 1999.

[4] David Dupplaw, Uladzimir Kharkevich, Spyros Kotoulas, Adrian Perreau de Pinninck, Ronny Siebes, and Chris Walton. *OpenKnowledge Deliverable 2.1: Architecting Open Knowledge*. `http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D2.1a.pdf`, 2006.

[5] Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb, Vojtěch Svátek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2007. In *Proceedings of the International Workshop on Ontology Matching (OM) at the International Semantic Web Conference (ISWC) + Asian Semantic Web Conference (ASWC)*, pages 96–132, 2007.

[6] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.

[7] Raul Garcia-Castro and Asunción Gómez-Pérez. Guidelines for benchmarking the performance of ontology management APIs. In *Proceedings of the 4th International Semantic Web Conference (ISWC)*, pages 277–292, Galway (IE), 2005.

[8] Fausto Giunchiglia, Fiona McNeill, Mikalai Yatskevich, Zharko Alekovski, Alan Bundy, Frank van Harmelen, Spyros Kotoulas, Vanessa Lopez, Marta Sabou, Ronny Siebes, and Annette ten Tejie. *OpenKnowledge Deliverable 4.1: Approximate Semantic Tree Matching in OpenKnowledge*. `http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D4.1.pdf`, 2006.

[9] Fausto Giunchiglia, Fiona McNeill, Mikalai Yatskevich, Carles Sierra, and Jordi Sabater. *OpenKnowledge Deliverable 4.4: Evaluating Good Answers in Open Knowledge*. `http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D4.4.pdf`, 2007.

[10] Fausto Giunchiglia, Carles Sierra, Fiona McNeill, Nardine Osman, and Ronny Siebes. *OpenKnowledge Deliverable 4.5: Good Enough Answer Algorithms*. `http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D4.5.pdf`, 2007.

[11] Fausto Giunchiglia and Ilya Zaihrayeu. Making peer databases interact - a vision for an architecture supporting data coordination. In *Proceedings of the 6th International Workshop on Cooperative Information Agents (CIA)*, pages 18–35, Madrid (ES), 2002.

[12] Alon Halevy. Why your data won't mix. *ACM Queue*, 3(8):50–58, 2005.

[13] Ralf Heese, Sven Herschel, Felix Naumann, and Armin Roth. Self-extending peer data management. In *Proceedings of the 11th Conference Datenbanksysteme in Business, Technologie und Web (BTW)*, pages 165–174, Karlsruhe (DE), 2005.

[14] Sindhu Joseph, Adrian Perreau de Pinninck, Dave Robertson, Carles Sierra, and Chris Walton. *OpenKnowledge Deliverable 1.1: Interaction Model Language Definition*. `http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D1.1.pdf`, 2006.

[15] Alexander Löser, Felix Naumann, Wolf Siberski, Wolfgang Nejdl, and Uwe Thaden. Semantic overlay clusters within super-peer networks. In *Proceedings of the Workshop on Databases, Information Systems, and Peer-to-Peer Computing (DBISP2P)*, pages 33–47, Berlin (DE), 2003.

[16] Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Mathematics and Computer Science, University of Stirling, Scotland (UK), 1994.

[17] Felix Naumann. From databases to information systems - information quality makes the difference. In *Proceedings of the 6th International Conference on Information Quality (IQ)*, pages 244–260, Cambridge (MA US), 2001.

[18] Felix Naumann, Ulf Leser, and Johann Christoph Freytag. Quality-driven integration of heterogenous information systems. In *Proceedings of the 25th International Conference on Very Large Databases (VLDB)*, pages 447–458, Edinburgh (UK), 1999.

[19] Beng Chin Ooi, Yanfeng Shu, and Kian-Lee Tan. Relational data sharing in peer-based data management systems. *SIGMOD Record*, 32(3):59–64, 2003.

[20] Thomas C. Redman. *Data Quality for the Information Age*. Artech House, Inc., Norwood (MA US), 1997. Foreword by - A. Blanton Godfrey.

[21] Monica Scannapieco, Antonino Virgillito, Carlo Marchetti, Massimo Mecella, and Roberto Baldoni. The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, 29(7):551–582, 2004.

[22] Dave Sill. comp.benchmarks frequently asked questions. `http://pages.cs.wisc.edu/~thomas/comp.benchmarks.FAQ.html`, 1996.

[23] Terrence D. Sole and Gary Bist. Benchmarking in technical information. *IEEE Transactions on Professional Communication*, 38(2):77–82, 1995.

[24] Michael J. Spendolini. *The Benchmarking Book.* AMACOM, New York (NY US), 2nd edition, 2003.

[25] Steffen Staab and Heiner Stuckenschmidt, editors. *Semantic web and peer-to-peer.* Springer, Heidelberg (DE), 2006.

[26] Heiner Stuckenschmidt, Fausto Giunchiglia, and Frank van Harmelen. Query processing in ontology-based peer-to-peer systems. In Valentina Tamma, Stephen Cranefield, Timothy W. Finin, and Steven Willmott, editors, *Ontologies for Agents: Theory and Experiences*, Whitestein Series in Software Agent Technologies. Birkhäuser, 2005.

[27] Statistical Consulting Group UCLA: Academic Technology Services. Introduction to SAS. `http://www.ats.ucla.edu/stat/stata/whatstat/default.htm`, accessed December 20, 2007.

[28] Ellen M. Voorhees. Overview of TREC 2006. In *Proceedings of the 15th Text REtrieval Conference (TREC)*, pages 1–16, 2006.

[29] Yair Wand and Richard Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[30] Richard Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65, 1998.

[31] Richard Wang, Veda Storey, and Christopher Firth. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):623–640, 1995.

[32] Richard Wang and Diane Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, 1996.

[33] Beverly Yang and Hector Garcia-Molina. Efficient search in peer-to-peer networks. In *Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS)*, pages 5–14, Vienna (AU), 2002.

[34] Mikalai Yatskevich, Fausto Giunchiglia, Fiona McNeill, and Pavel Shvaiko. *OpenKnowledge Deliverable 3.3: A methodology for ontology matching quality evaluation*. `http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D3.3.pdf`, 2006.

[35] Ilya Zaihrayeu. *Towards Peer-to-Peer Information Management Systems*. PhD thesis, International Doctorate School in Information and Communication Technology, University of Trento, Trento (IT), March 2006.