Quality-aware Service-Oriented Data Integration: Requirements, State of the Art and Open Challenges

Schahram Dustdar Reinhard Pichler Vadim Savenkov Hong-Linh Truong

Vienna University of Technology
{dustdar,truong}@infosys.tuwien.ac.at
{pichler,savenkov}@dbai.tuwien.ac.at

ABSTRACT

With a multitude of data sources available online, data consumers might find it hard to select the best combination of sources for their needs. Aspects such as price, licensing, service and data quality play a major role in selecting data sources. We therefore advocate quality-aware data services as a natural data source model for complex data integration tasks and mash-ups. This paper focuses on requirements, state of the art, and the main research challenges on the way to the realization of such services.

1. INTRODUCTION

The problem of transferring data between and answering queries over heterogeneous information systems has been one of the key topics of database research over the past fifteen years. Various approaches have been intensively studied. Perhaps the most popular approach is data federation [9, 34], where many sources are encapsulated in a virtual global database that translates queries against the single mediating schema into queries on the sources [35, 43].In data exchange, the data is actually materialized in the target database; the data sources are no longer needed and, possibly, no longer available for query answering [24]. Peer data management advances these basic approaches, allowing many-to-many, bi-directional mappings between the data schemas of systems participating in data integration. Queries posed against one such system are normally answered by using the local database of this system and by retrieving data from other systems in the network, which in turn may also have to contact other systems [8, 14, 37]. In [28] it was shown that all these three approaches can be considered as special cases of a general information integration framework which we refer to as *data network* in this paper. Such a data network consists of autonomous systems – *data peers* – which may have both data exchange relationships and virtual mappings between each other. Moreover, data exchange constraints as well as virtual mappings may exist locally on each data peer.

The semantics of answering queries in such data networks is now well understood [14, 24, 37, 43]. Further aspects, where a lot of progress has been made in recent research, include performance issues and identifying the barrier between tractability and intractability of query answering [1, 57, 24] and important data exchange tasks [24, 30]; providing and using information on provenance [33, 32]; privacy in a data integration environment [49]; dealing with various forms of uncertainty [7, 23], and query answering in the presence of inconsistencies [2, 19].

However, the needs of many real-world applications are still by far exceeding these current developments. Consider, for instance, the case when an enterprise wants to provide some of the data accumulated in its internal data network to third parties for a fee. With recent introduction of *data marketplaces* like Windows Azure Marketplace or Infochimps, such practice is becoming routine for many companies, with millions of data sources being offered for querying already. Currently, data marketplaces aid their customers by providing a unified interface to a multitude of data sources, including support for certain query languages, online schema browsing tools, and schema documentation and licensing information in legal English. Furthermore, marketplaces offer reliable payment processing, cloud storage facilities for improved querying performance, format conversions, and greatly streamlined step-by-step processes for data publishers and data consumers. With a large number of datasets available in the marketplaces, the need for their comprehensive specification becomes especially apparent for users trying to identify the best data sources matching their requirements. As the example of conventional e-commerce platforms shows, extensive product descriptions (that is, metadata) and quality of the information products is crucial. Sim-

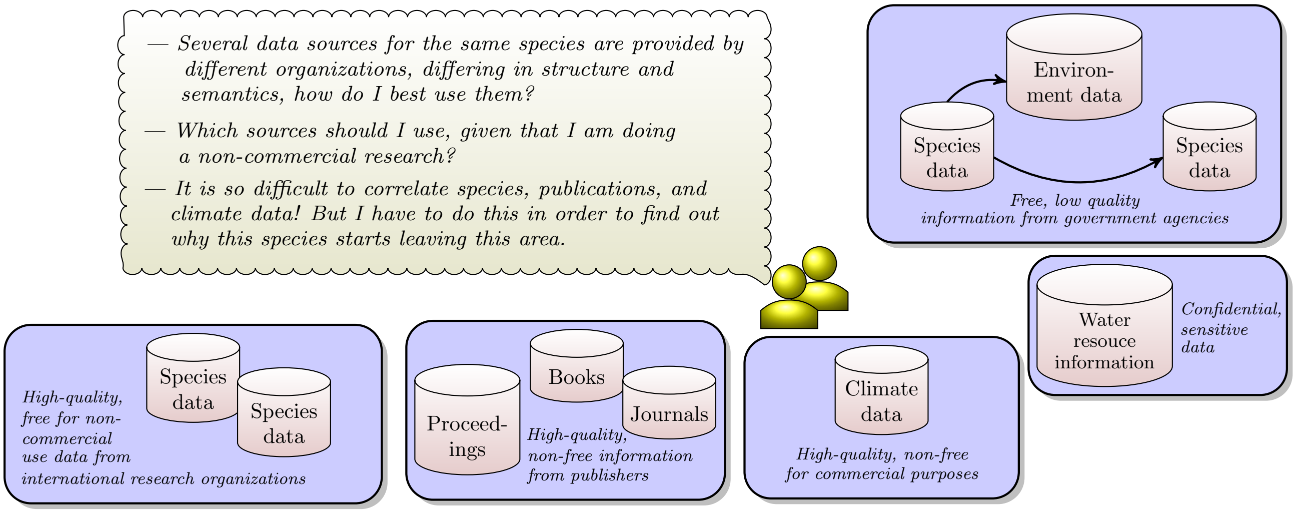
— Several data sources for the same species are provided by different organizations, differing in structure and semantics, how do I best use them?
— Which sources should I use, given that I am doing a non-commercial research?
— It is so difficult to correlate species, publications, and climate data! But I have to do this in order to find out why this species starts leaving this area.
Environ-ment data
Species data
Species data
Free, low quality information from government agencies
Water resouce information
Confidential, sensitive data
High-quality, free for non-commercial use data from international research organizations
Species data
Species data
Proceed-ings
Books
Journals
High-quality, non-free information from publishers
Climate data
High-quality, non-free for commercial purposes

Figure 1: Challenges faced by the user when using biodiversity data for researching species

ilarly to Amazon customers checking other users' reviews when considering a purchase, customers of data marketplaces will rely on the quality metrics and community ratings of the information products offered to them. Moreovoer, mash-up applications will require most of this metadata, along with licensing and pricing terms to be machine readable. Such requirements might translate into complex measures for assessing, e.g., data quality at the data publisher side. Thus, a means for automatic generation and maintenance of metadata from a data network at hand will be greatly beneficial for both data publishers and data consumers.

In this paper we take a *quality-aware service-oriented approach* to data integration and consider Data Services (DSs) built on top of conventional databases or data networks. The main task of DSs is to provide online access to its data *and metadata* while meeting certain quality standards, known as service contracts. We believe that data quality must be a core of service contracts, along with further service-related quality criteria, data license and pricing [67]. We will discuss the challenges associated with building DSs and align those challenges with the state-of-the art in research in information integration and service-oriented architectures.

2. A CASE FOR DATA SERVICES

To elaborate possible requirements for quality-aware data services outlined in the introduction, let us consider the case of biodiversity data integration, illustrated in Figure 1. As described in [69], biodiversity data have several properties which strongly impact data integration: (i) Biodiversity data can have complex structure, describe a large number of species and observations, with broadly varying degrees of quality. (ii) Many organizations are involved in providing and continuously updating the data. Data sources of different origin vary in semantics and formats. (iii) Similar and/or complementary data is provided by different organizations. There are many data owners who can set different data access policies. (iv) Intellectual property rights (IPRs) or other disclosure concerns may apply, for example, to data acquired by commercial research.

Biodiversity data is available from the Global Biodiversity Information Facility (`www.gbif.net`). Depending on specific research goals, it may need to be enhanced, e.g., with climate or water resource information.

Assume that a user wants to determine why a specific type of species starts to disappear from a particular area. She probably needs to access and integrate the data from multiple sources scattered over the Internet (see Figure 1). These sources can be free for non-commercial purposes or completely free, and sensitive or associated with some complex IPRs. Similar data can be offered by different providers, with varying structure and quality. For the user, it can be difficult to select the optimal combination of data sources for a task at hand.

This scenario poses several research questions. To aid the user in her job, one would need to know, e.g. how to support the discovery of data services based on quality and licensing aspects, and how to support the composition of the data from different quality-aware data services?

From the data publisher point of view, the related questions are: How to choose the quality, licensing and pricing models for the data service? How to align these models with the underlying data network, and how to perform the necessary measurements? Is it possible to optimize the data network (which is most likely in use for quite a few enterprise tasks) for data publishing, so that the service contracts are observed?

We will discuss some of the research issues associated with the implementation of the above scenario in the next section.

3. ENABLING DATA SERVICES

One of the most general architectural models for data integration considered in the literature is a *peer data network*, in which nodes are independent databases (data peers) and links between nodes represent schema mappings or other specifications of data dependencies. Our goal is to provide an infrastructure for publishing the data from a data network, through a thoroughly specified interface, which we have introduced before as *data service*. In this section, we subdivide such an infrastructure into three conceptual levels:

- *Data Network Level*, concerned with the data network management tasks, such as querying and updating data and metadata (e.g., data lineage) and collecting the quality and performance metrics.
- *Service Interface Level*, concerned with the presentation of the distributed enterprise data to the user through a single service interface, including formal specification of the service, data and quality models.
- *Meta-Service Level*, performing discovery and integration of multiple services (possibly, other data services).

The set of specific issues that would arise in practice is largely determined by such factors as the data model used by data peers, expressiveness of mappings and ETL operations used to transfer the data, the nature and volume of the published information, required query capabilities, and countless further considerations. However, as we point out in the remainder of this section, already catering for a basic data service functionality with relational databases as peers leads to a number of open issues and research opportunities.

3.1 Managing metadata in data networks

To enable the data service interface, the data network must be able to to store, maintain, and provide access to the metadata describing the data quality, licensing and pricing properties. This functionality is by no means straightforward, as we will argue shortly. For example, during updates or caching, both data and metadata changes must be propagated through the network. For many classes of metadata, already the basic rules of propagation through the schema mappings have to be defined. For instance, what does it mean to propagate the data quality descriptors. Even a comparatively simple completeness metric depends crucially on the semantics of the schema mappings, namely if the closed- or open-world assumption is taken [60]. Here, we discuss such questions for several types of metadata relevant for the functionality of data services.

Provenance, often also called "lineage", is the key type of metadata on which the derivation of other metrics in data networks depends. Usually, one distinguishes the "why-" and "where-" provenance [13], where the former gives a comprehensive justification for creation of a data item, while the latter references the origins of the values used in a derived data item. A unified way of treatment of annotations propagated through data dependencies was presented in [33], with the examples of handling incompleteness, probabilities, why-provenance, and mulitplicities of tuples. No similar work considering the unified treatment of metadata relevant for data services (i.e., related to quality, licensing or pricing) is known to us.

Data quality (DQ) was identified as an intrinsic issue for information systems in the early 90-s [63]. Since then, it has become a subject of a huge body of publications (see [5] for a survey). The main problem addressed in research is the construction of frameworks and methodologies for assessment and continuous improvements of the enterprise DQ. For data services, the following issues play a crucial role: (i) measurement and propagation of DQ through the data network, (ii) extraction of a concise characterization of DQ as part of the data service description, (iii) quality-aware query answering, and (iv) maintenance of the DQ model associated with the data network (e.g., accommodating updates).

The typical DQ dimensions considered by the majority of authors are *accuracy, completeness, timeliness* and *consistency* [55, 5]. Each dimension can be represented by one or several quality metrics. The metrics and further quality dimensions heavily depend on the application domain, or *context*: the quality of data is often identified with its "fitness for use" [59]. Formalizing the notion of context in data quality is an interesting challenge, which only recently received attention [10]. However, even within a fixed context, assessing data quality is commonly recognized as a highly non-trivial problem, often requiring human intervention (in the form of expert evaluation or rating by the community of data users [54]). Such aspects as heterogeneity of the data add to the intricacy of the issue [4]. Still, for some quality metrics, unsupervised measurement algorithms have been studied in the literature [3, 29] and approved in practice [46]. This line of research and industrial usage, supported by tools from the major vendors of enterprise data management software, has gained a lot of momentum in recent years.

For data services offering *information products* composed from disparate enterprise data, an ability to *derive* the quality metrics of the product based on the quality metrics of the data sources is desired, instead of performing the assessment anew. The foundation for such a derivation has been laid by Wang et al. in [72], where computing the DQ metrics of the basic SQL operators' results is considered, and further extended by Sarkar et al. [51, 52]. For data networks, with multiple data flows and expressive schema mappings between the peers, the problem of deriving DQ metrics becomes increasingly complex, leaving an ample space for future research.

One of the dimensions often identified with data quality is the existence of object duplicates, which do not agree on all attribute values. Deduplication, or *record linkage*, is especially important for cases where data is distributed over several sources (which is our assumption in this paper). In a simple case, duplicates may result in violations of primary key constraints (and then correlate with the consistency DQ dimension). This case is often studied in the database literature, under the tile *consistent query answering.* For an inconsistent database state, a minimal set of updates (a minimal *repair*) is constructed, and *certain answers* – present in the query results under each repair – are computed.

However, there may be cases where conflicting duplicates do not lead to constraint violations: for instance, if inaccurate key values are present. Moreover, defining repairs through simple operations of, e.g. tuple deletion can lead to undesirable information loss. More sophisticated and domain-dependent record linkage techniques are then used, giving rise to the *data fusion* process [11]. Object linkage algorithms are often complex and can hardly be captured by simple logical formalisms like the language of database constraints.

As far as *quality-driven querying* of data is concerned, the main reference so far remains the work of Naumann [50], who proposed a framework for taking quality-related metrics into account when answering queries against web sources (i.e., in a data federation setting). More recent advances on quality-aware query answering can be found in [10, 6]. The quality metadata in XML are considered in [61, 47], where a decentralized architecture of "quality brokers", or quality-aware query mediators, is proposed. Such an architecture is well suited for data networks. However, so far only the simplest mappings between the sources and the global schema have been considered [50, 61]. Taking more expressive mappings into account, including those between data peers, can be seen as a reasonable next goal for quality-aware data querying.

The study of maintenance of data quality metrics has been merely initiated [45]. Here again, the data federation model rather than a data network with complex patterns of update propagation is considered (see also Section 3.3).

Licensing metadata, to the best of our knowledge, has received only little attention in database research. One of the most important developments is the Open Data license [48], which allows one to license a published dataset under various conditions. An overview of further real world data contracts including licensing terms can be found in [67]. However, such questions as combining the differently licensed data in the same information product and checking for licensing conflicts still wait for further exploration.

Pricing in conjunction with query processing is a typical instrument in query optimization: all resources needed to evaluate the query are generalized as monetary cost (per byte transferred, per disk read operation etc.). A total cost of a query plan is then computed. Many cloud services nowadays implement such cost models literally by charging users for actual resource consumption. Such approaches are not the ultimate answer for data services though, since this kind of pricing is content- and quality-independent. Query optimization under per-tuple fees or fees depending on DQ can be seen as viable directions for research. For instance, one can ask for best quality-dependent query plans for a given price range, or optimize data pricing for a set of typical user queries.

Quality of service (QoS) describes the properties of the data service, which are not specific for the content it serves. Typical QoS aspects are reliability, availability, security, and performance. These properties are an important part of the data service specification as well, and have to be aligned with capabilities of the underlying data network. For instance, the reliability and availability of the data service depend on the queries that can be answered at a given time instant, which in turn depends on the availability of data peers.

For data service discovery and selection (see Section 3.4) the underlying data network must be able to characterize its ability to support data services in terms of a schema or ontology of the data as well as the above mentioned quality and licensing criteria. Such a characterization has to be given in principle (if the design of the data network supports a specified service) as well as for the current state of the network characterized by quality metrics (e.g., logical consistency and data availability).

Data network models are typically built using logical formalisms [28, 14, 37]. To support data services, such models must take metadata as part of their basic concepts: For example, consistency of sources equally depends on the data and on the metadata (e.g., sources might disagree on the license of a data item). Moreover, the sources must adhere to the same definition of quality metrics.

Further extensions of the foundations are needed as well. One extension concerns the type of mappings between the data peers. In contrast to distributed databases, there is an independent application behind each data peer, employing processing mechanisms that are generally hard to formalize declaratively. Therefore, the notion of uncertain mappings (based on [23]) with possible adaptations has to be studied to approximate the data processing rules. Such mappings are important both as a means of expressing the dependencies between the local and global schemas, but also as a way of simulating the application logic at the data peers. Another issue is concerned with the data format heterogeneity in the data network: The data with varying levels of structure (as addressed in [36]) could be integrated (in particular, semi-structured and even unstructured data) provided that the underlying data peers are able to formulate and support a certain level of data and service quality.

3.2 Publishing a data service

A data service must provide a comprehensive specification of the information products it offers. This specification will then allow data consumers to make informed decisions when searching for and comparing data sources for their applications.

Currently, QoS models for Web services are well developed and various techniques and tools support engineers modeling QoS for Web services [42, 58, 73]. However, the data-specific aspects of service description have not yet received proper support. In [66], various data-related aspects of service specifications – called *data concerns* – are identified and analyzed, and then in [65, 71] a service engineering process for publishing those concerns is developed.

Given the considerable effort spent on data quality from the database perspective, as outlined in the previous section, there is a lack of integration between DQ metrics and parameters commonly used to specify data services. In fact, no standard model of data quality that could serve as a basis for the data service specification is available so far. Furthermore, since the data quality notions are often domain-specific, actually a hierarchy of quality models will be desirable, whereby the general quality metrics like completeness can be extended by or mapped to the specialized ones (like "number of OCR failures" for an electronic library), allowing the customer to better compare similar services and still have compatible service descriptions for the case of data mash-up. Similarly, existing service licensing [26] and service level agreements (SLAs), see e.g, [39], are mainly based on "operational" QoS models. Some data licensing models exist [21], but are neither standard nor formalized to allow automatic processing, and thus cannot be used in the service model. Typical *data rights* that license models have to support include derivation, collection, reproduction, attribution and restriction to noncommercial use [67].

To date, these types of metadata have not been combined in a single service model [71], and there exist no specifications to support the publishing and discovery of data services based on such information. This calls for a new research focus on the development of publishing and discovery of quality aspects of data services. The data publishing not only requires an appropriate description of the semantics of the data (e.g. a schema or an ontology), but also a more general specification reflecting the data quality, QoS, and data service licensing. In addition to languages and specifications for modeling these types of information together, we also need a scalable and extensible framework to manage the lifecycle of this information. Existing SOA techniques can address many aspects in the modelling, publishing and management of data services. In our view, existing QoS and service licensing models can be extended with data quality metrics. Then, quality-related metadata can be linked with other types of service information, for instance, integrated into the Web services information model [64].

3.3 Optimizing the design and operation of the data network

To provide a certain quality of service with minimal cost, various kinds of optimization – both at design time and at run time – are required. This leads to the following issues.

The first issue is *network optimization*. Network topologies are characterized by the data stored on each data peer and the inter-peer mappings. Hence, the optimality of a network topology and also the equivalence of several network topologies depend on the quality parameters of the service. Various kinds of optimization of the design of a data network have to be considered – taking the formalization of QoS in a data network into account. The concrete values of the QoS metrics are influenced by the design of the network topology, which heavily depends on data aspects like data allocation, data replication, and inter-peer mappings. Thus, a number of optimization problems naturally arises from the QoS metrics like, e.g., minimizing the cost of the data network while guaranteeing a certain QoS level. Another interesting research question concerns equivalence between network topologies (understood, e.g., as a set of schema mappings between peers, together with local data constraints), with respect to certain quality metrics. This is in spirit of the recent work by Fagin et al. [25], where several notions of equivalence between schema mappings have been proposed, in particular, relative to a given class of queries. A study of replication in data networks [62], but taking QoS into account is also a promising research direction.

Second, *query optimization* is an important issue in all data management systems – no matter whether they are considered in the context of data services or not. In a data network, traditional approaches to query optimization (see, e.g., [31, 41]) have to be extended so as to take the inter-peer and intra-peer mappings into account. Two important sub-problems that deserve further study are query routing and load distribution (see, e.g., [56]).

Finally, the *optimization of updates*, i.e., update, insert, or delete operations in a data network is an important issue. The modification of data at one peer normally leads to the violation of the inter- and intra-peer mappings, which are re-enforced by performing the chase procedure. Update optimization in a data network has two important facets. First, minimizing the propagation time or minimizing the impact on the other data peers is critical. Second, there may exist many possible states of the data network such that all constraints are fulfilled again after the update. In this case, an important optimization problem is concerned with finding the "smallest" one, which is usually referred to as the core. In [28], core computation has been extended from data exchange to a data network. In addition, core computation methods should be made incremental, i.e., be able to "locally repair" the core after an update, rather than recompute it from scratch.

Updating the data network brings about the problem of transaction support. A discussion of transactions in a service-oriented distributed environments can be found in [68]. In the context of data services, the mapping from the inter-service transaction semantics to the underlying data network transactions must be investigated.

3.4 Selecting and mixing data from quality-aware data services

With published quality- and licensing-related metadata, the data consumers will have a chance to decide if the data from a given data service fits their intended usage scenario. This is in clear contrast with a situation where only descriptions of syntax (format) and semantics (ontology) of the data are available.

Rich data service specifications are especially important when discovery, comparison, and selection of data sources are computer assisted. Users should be able to automatize such tasks as comparing and checking compatibility of data service contracts, or determining trade-offs between quality, licenses and costs associated with information products from a certain set of data services.

To support the composition of data sources on the Internet, in particular in the collaborative Web 2.0 context, many tools have been developed [22, 38]. However, existing techniques mainly focus on selecting data sources based on data structures and on dealing with syntax and semantics of the data [22, 40, 53]. In the area of Web data mash-ups, a need of considering data quality has been understood [17], and the study of DQ models initiated [16]. More comprehensive data mash-up models, including licensing and pricing in conjunction with data and service quality, are yet to be developed.

Similarly, most of the contemporary service selection and combination techniques are built around the QoS, cost, and the semantics of service operations [58, 73], rather than content-related aspects. This is true, in particular, for concepts such as ad-hoc flows [70]. For service mash-ups [44], a need for data quality support has been recently realized. In particular, the Mashup Services System [12] takes information quality metrics into account when generating and managing service mash-ups.

Service selection techniques currently do not deal with the compatibility between different license models [27] when integrating data from different services. A recent work has supported the evaluation of service contracts, but its support of data-related concerns is limited [20].

Hence, we envisage a proliferation of new techniques and algorithms for composition of data services (be it a mash-up Web application for end users, or a new data service) taking the full spectrum of data concerns and QoS aspects into account. Such algorithms and techniques should extend current service contract compatibility evaluation [20], syntax mismatching [40], and QoS-aware and semantic workflow composition [18, 15] with data quality and contract aspects. Moreover, unlike current techniques which require user interaction for selection and maping of the data sources [53], mixing data services should be possible with minimum involvement (or no involvement at all) of the end user.

4. CONCLUSION

We have presented the concept of data services which takes a quality-aware service-oriented view on data integration, and identified the main research challenges on the way to its realization.

The main benefit of this approach comes from the cross-fertilization of the database and distributed systems fields: The peer data management concept [8, 14, 28, 37] was a first step in this direction by making P2P technology available to data management and vice versa. We consider combining service-oriented computing techniques with database technology as an important second step. This will lead to a better understanding and interesting extensions of the frameworks and methods on both sides: On the one hand, SOA key concepts like quality, licensing, service selection, and service composition must be extended so as to take data aspects into account. On the other hand, the concept of data networks must be significantly enhanced by integrating quality considerations into it and by developing new methods of data network optimization with respect to the various quality criteria.

Acknowledgements. This work has been funded by the Vienna Science and Technology Fund (WWTF) through project ICT08-032. Savenkov was supported by the Erasmus Mundus External Co-operation Window Programme of the European Union.

5. REFERENCES

[1] S. Abiteboul and O. M. Duschka. Complexity of answering queries using materialized views. In *Proc. ACM PODS '98*, pages 254–263.

[2] P. Andritsos, A. Fuxman, and R. J. Miller. Clean answers over dirty databases: A probabilistic approach. In *Proc. IEEE ICDE '06*.

[3] D. P. Ballou, I. N. Chengalur-Smith, and R. Y. Wang. Sample-based quality estimation of query results in relational database environments. *IEEE Trans. Knowl. Data Eng.*, 18(5):639–650, 2006.

[4] C. Batini, D. Barone, F. Cabitza, and S. Grega. A Data Quality Methodology for Heterogeneous Data. *Int. J. of Database Management Syst.*, 3(1):60–79, Feb. 2011.

[5] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52, 2009.

[6] D. Beneventano and R. C. N. Mbinkeu. Quality-Driven Query Processing Techniques in the MOMIS Integration System Advances in Databases and Information Systems. In *Proc. ADBIS '10*, pages 46–57. Springer, 2011.

[7] O. Benjelloun, A. D. Sarma, A. Y. Halevy, M. Theobald, and J. Widom. Databases with uncertainty and lineage. *VLDB J.*, 17(2):243–264, 2008.

[8] P. A. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. Mylopoulos, L. Serafini, and I. Zaihrayeu. Data management for peer-to-peer computing: A vision. In *Proc. WebDB '02*, pages 89–94, 2002.

[9] P. A. Bernstein and L. M. Haas. Information integration in the enterprise. *Commun. ACM*, 51(9):72–79, 2008.

[10] L. E. Bertossi, F. Rizzolo, and L. Jiang. Data quality is context dependent. In *Proc. BIRTE '10*, pages 52–67. Springer, 2011.

[11] J. Bleiholder and F. Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1–41, jan 2009.

[12] A. Bouguettaya, S. Nepal, W. Sherchan, X. Zhou, J. Wu, S. Chen, D. Liu, L. Li, H. Wang, and X. Liu. End-to-end service support for mashups. *IEEE Trans. Serv. Comput.*, 3:250–263, July 2010.

[13] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In *Proc. ICDT '01*, pages 316–330. Springer, 2001.

[14] D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati. Logical foundations of peer-to-peer data integration. In *Proc. ACM PODS '04*, pages 241–251.

[15] G. Canfora, M. D. Penta, R. Esposito, and M. L. Villani. An approach for QoS-aware service composition based on genetic algorithms. In *Proc. GECCO '05*, pages 1069–1075. ACM, 2005.

[16] C. Cappiello, F. Daniel, A. Koschmider, M. Matera, and M. Picozzi. A quality model for mashups. In *Proc. ICWE '11*, pages 137–151. Springer, 2011.

[17] C. Cappiello, F. Daniel, M. Matera, and C. Pautasso. Information quality in mashups. *IEEE Internet Computing*, 14(4):14–22, 2010.

[18] J. Cardoso and A. Sheth. Semantic e-workflow composition. *Journal of Intelligent Information Systems*, 21(3):191–225, 2003.

[19] J. Chomicki. Consistent query answering: The first ten years. In *Proc. SUM '08*, pages 1–3. Springer, 2008.

[20] M. Comerio, H.-L. Truong, F. D. Paoli, and S. Dustdar. Evaluating contract compatibility for service composition in the SeCO$_2$ framework. In *Proc. ICSOC/ServiceWave '09*, pages 221–236. Springer, 2009.

[21] Committee on Licensing Geographic Data and Services, National Research Council. *Licensing Geographic Data and Services*. The National Academies Press, 2004.

[22] G. Di Lorenzo, H. Hacid, H.-y. Paik, and B. Benatallah. Data integration in mashups. *SIGMOD Record*, 38(1):59–66, 2009.

[23] X. L. Dong, A. Y. Halevy, and C. Yu. Data integration with uncertainty. In *Proc. VLDB '07*, pages 687–698. ACM, 2007.

[24] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, 2005.

[25] R. Fagin, P. G. Kolaitis, A. Nash, and L. Popa. Towards a theory of schema-mapping optimization. In *Proc. ACM PODS '08*, pages 33–42.

[26] G. R. Gangadharan and V. D'Andrea. Licensing services: Formal analysis and implementation. In *Proc. ICSOC '06*, pages 365–377. Springer, 2006.

[27] G. R. Gangadharan, H. L. Truong, M. Treiber, V. D'Andrea, S. Dustdar, R. Iannella, and M. Weiss. Consumer-specified service license selection and composition. In *Proc. IEEE ICCBSS '08*, pages 194–203.

[28] G. D. Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. On reconciling data exchange, data integration, and peer data management. In *Proc. ACM PODS '07*, pages 133–142.

[29] L. Golab, F. Korn, and D. Srivastava. Discovering pattern tableaux for data quality analysis: a case study. In *QDB '11.*

[30] G. Gottlob and A. Nash. Efficient core computation in data exchange. *J. ACM*, 55(2):9:1–9:49, 2008.

[31] G. Graefe. Query evaluation techniques for large databases. *ACM Comput. Surv.*, 25(2):73–170, 1993.

[32] T. J. Green, G. Karvounarakis, Z. G. Ives, and V. Tannen. Update exchange with mappings and provenance. In *Proc. VLDB '07*, pages 675–686. ACM, 2007.

[33] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *Proc. ACM PODS '07*, pages 31–40.

[34] L. M. Haas. Beauty and the beast: The theory and practice of information integration. In *Proc. ICDT '07*, pages 28–43. Springer, 2007.

[35] A. Y. Halevy. Answering queries using views: A survey. *VLDB J.*, 10(4):270–294, 2001.

[36] A. Y. Halevy, M. J. Franklin, and D. Maier. Principles of dataspace systems. In *Proc. ACM PODS '06*, pages 1–9.

[37] A. Y. Halevy, Z. G. Ives, D. Suciu, and I. Tatarinov. Schema mediation in peer data management systems. In *IEEE ICDE '03*, pages 505–516.

[38] V. Hoyer and M. Fischer. Market overview of enterprise mashup tools. In *Proc. ICSOC '08*, pages 708–721. Springer, 2008.

[39] A. Keller and H. Ludwig. The WSLA framework: Specifying and monitoring service level agreements for web services. *J. Network and Systems Management*, 11(1):57–81, 2003.

[40] W. Kongdenfha, H. R. Motahari-Nezhad, B. Benatallah, F. Casati, and R. Saint-Paul. Mismatch patterns and adaptation aspects: A foundation for rapid development of web service adapters. *IEEE Trans. Serv. Comp.*, 2(2):94–107, 2009.

[41] D. Kossmann. The state of the art in distributed query processing. *ACM Comput. Surv.*, 32(4):422–469, 2000.

[42] K. Lee, J. Jeon, W. Lee, S.-H. Jeong, and S.-W. P. (eds.). QoS for web services: Requirements and possible approaches, Nov. 2003. W3C Technical Report.

[43] M. Lenzerini. Data integration: A theoretical perspective. In *Proc. ACM PODS '02*, pages 233–246.

[44] X. Liu, Y. Hui, W. Sun, and H. Liang. Towards service composition based on mashup. In *Proc. IEEE SCW '07*, pages 332–339.

[45] A. Marotta. Managing source quality changes in a data integration system. In *Proc. CAiSE '06 Doct. Consortium*. CEUR-WS.org, 2006.

[46] E. Masuoka, D. Roy, R. Wolfe, J. Morisette, S. Sinno, M. Teague, N. Saleous, S. Devadiga, C. O. Justice, and J. Nickeson. MODIS land data products: Generation, quality assurance and validation land remote sensing and global environmental change. volume 11 of *Remote Sensing and Digital Image Processing*, chapter 22, pages 509–531. Springer, 2011.

[47] D. Milano, M. Scannapieco, and T. Catarci. A peer-to-peer service supporting data quality: Design and implementation issues. In *Proc. ICSNW '04*, pages 321–322. Springer, 2004.

[48] P. Miller, R. Styles, and T. Heath. Open data commons, a license for open data. In *Proc. LDOW '08*. CEUR-WS.org, 2008.

[49] A. Nash and A. Deutsch. Privacy in GLAV information integration. In *Proc. ICDT '07*, pages 89–103. Springer, 2007.

[50] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*. Springer, 2002.

[51] A. Parssian, S. Sarkar, and V. S. Jacob. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. *Management Science*, 50(7):967–982, 2004.

[52] A. Parssian, S. Sarkar, and V. S. Jacob. Impact of the Union and Difference Operations on the Quality of Information Products. *Inf. Syst. Research*, 20(1):99–120, 2009.

[53] D. L. Phuoc, A. Polleres, M. Hauswirth, G. Tummarello, and C. Morbidoni. Rapid prototyping of semantic mash-ups through semantic web pipes. In *Proc. WWW '09*. ACM, 2009.

[54] R. Pichler, V. Savenkov, S. Skritek, and H. L. Truong. Uncertain databases in collaborative data management. In *Proc. MUD '10*, CTIT Workshop Proc. Series. Univ. Twente, 2010.

[55] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Commun. ACM*, 45:211–218, 2002.

[56] T. Pitoura and P. Triantafillou. Load distribution fairness in P2P data management systems. In *IEEE ICDE '07*, pages 396–405.

[57] R. Pottinger and A. Y. Halevy. Minicon: A scalable algorithm for answering queries using views. *VLDB J.*, 10(2-3):182–198, 2001.

[58] S. Ran. A model for web services discovery with QoS. *SIGecom Exch.*, 4(1):1–10, 2003.

[59] T. C. Redman. Second-generation data quality systems, chapter 34 of *Juran's quality handbook*, J.M. Juran and G.A.Blanton, eds. 1999.

[60] M. Scannapieco and C. Batini. Completeness in the relational model: a comprehensive framework. In *Proc. IQ '04*, pages 333–345. MIT, 2004.

[61] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni. The Da IS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Inf. Syst.*, 29(7):551–582, 2004.

[62] M. Sozio, T. Neumann, and G. Weikum. Near-optimal dynamic replication in unstructured peer-to-peer networks. In *Proc. ACM PODS '08*, pages 281–290.

[63] D. M. Strong, S. E. Madnick, T. Redman, A. Segev, and R. Y. Wang. Data quality: A critical research issue for the 1990s and beyond. In *Proc. ICIS '94*, pages 500–501. Assoc. for Inf. Systems, 2004.

[64] M. Treiber, H.-L. Truong, and S. Dustdar. Semf — service evolution management framework. In *Special session on Quality and Service-Oriented Applications, SEAA '08*, pages 329–336. IEEE, 2008.

[65] H. L. Truong and S. Dustdar. On evaluating and publishing data concerns for data as a service. In *Proc. IEEE APSCC '10*, pages 363–370.

[66] H.-L. Truong and S. Dustdar. On analyzing and specifying concerns for data as a service. In *Proc. IEEE APSCC '09*, pages 87–94, 2009.

[67] H.-L. Truong, G. Gangadharan, M. Comerio, S. Dustdar, and F. D. Paoli. On analyzing and developing data contracts in cloud-based data marketplaces. In *Proc. IEEE APSCC '11*, pages 174–181, 2011.

[68] C. Türker, K. Haller, C. Schuler, and H.-J. Schek. How can we support grid transactions? towards peer-to-peer transaction processing. In *Proc. CIDR '05*, pages 174–185.

[69] O. Unal and H. Afsarmanesh. A final report on distributed information management requirement analysis. Technical report, Uiv. Amsterdam, Inf. Dept., 2004.

[70] M. Voorhoeve and W. M. P. van der Aalst. Ad-hoc workflow: Problems and solutions. In *Proc. DEXA Workshop '97*, pages 36–40. IEEE Comp. Soc. Press, 1997.

[71] Q. H. Vu, T.-V. Pham, H.-L. Truong, S. Dustdar, and R. Asal. On analyzing and developing data contracts in cloud-based data marketplaces. In *Proc. IEEE AINA '12*. to appear.

[72] R. Y. Wang, M. Ready, and H. B. Kon. Toward quality data: An attribute-based approach. *Decision Support Systems*, 13:349–372, 1995.

[73] X. Wang, T. Vitvar, M. Kerrigan, and I. Toma. A QoS-aware selection model for semantic web services. In *Proc. ICSOC '06*, pages 390–401. Springer, 2006.