Service-oriented Data Quality Engineering and Data Publishing in the Cloud

Marco Comerio*, Hong-Linh Truong†, Carlo Batini*, Schahram Dustdar†
* *Department of Informatics, Systems and Communication*
University of Milano - Bicocca, Italy
{*comerio,batini*}@*disco.unimib.it*
†*Distributed Systems Group, Vienna University of Technology*
{*truong,dustdar*}@*infosys.tuwien.ac.at*

***Abstract*—Traditional data quality engineering techniques, often used and deployed within a single enterprise environment, are inadequate to cope with the rapid change of data, with a multitude of quality degrees, to be used in contemporary business models. The emerging cloud computing paradigm could potentially offer high-quality, composable data and techniques, under the Software-as-a-Service (SaaS), Data-as-a-Service (DaaS) and crowdsourcing models, for data quality engineering and data publishing. However, so far how to utilize the potential of cloud computing models for data quality engineering has not been discussed. In this paper, we analyze requirements of data quality engineering and quality-aware data publishing processes in the cloud and we provide a conceptual architecture utilizing and supporting the SaaS, DaaS and crowdsourcing models for the realization of such processes.**

***Keywords*-Data Quality Engineering; Cloud Computing; Software-as-a-Service; Data-as-a-Service; Crowdsourcing**

I. Introduction

The quality of data in business process is of paramount importance as low quality of data will hamper enterprise's business severely [1], [2]. With the current way of doing business involving several networked enterprises and customers, ensuring high-quality data for business processes is even more important and challenging due to two main issues. First, very often the data used in business processes are collected from the Web (e.g., online product information and customer activities) and the quality of such data is hard to ensure [3], [4]. Second, several types of data are related to global enterprises and consumers (e.g., emails, addresses, credit balances) that are dynamically changed. These issues cause: (i) the knowledge used by traditional data quality engineering techniques will be outdated quickly, as such knowledge is normally built by individual providers and maintained locally, and (ii) these techniques are not able to deal with the complexity of data, as they typically support only limited types of data.

Potentially, the emerging cloud computing paradigm [5] could offer powerful solutions to deal with the data quality issues mentioned above. Cloud computing describes new consumption and delivery models for IT services based on the Internet, such as the Software-as-a-Service (SaaS) [6], Data-as-a-Service (DaaS) [7], [8], and crowdsourcing [9], that could potentially solve the limited knowledge and isolated techniques for data quality engineering. First, data quality engineering tools can be provided under SaaS, data/information/knowledge [1] for data quality improvement can be provided under DaaS, and the crowdsourcing model can enable community collaboration in data quality engineering activities. Second, such SaaS, DaaS, and crowdsourcing services can be composed to create powerful solutions for data quality engineering.

However, so far how to utilize the potential of cloud computing models for data quality engineering has not been discussed. Past research has investigated the benefits and the drawbacks of the SaaS model [10], [11], but has not considered the data quality engineering domain. Only few papers [7], [8] have focused on the trend of publishing data under the DaaS model and on the role of DaaS in data quality engineering processes. The possibility to perform a data quality engineering process in the cloud on data from heterogeneous sources (e.g., databases and Web services) characterized by different quality levels should be deeply investigated. By moving from data sources in an organization to DaaS in the cloud, traditional data engineering activities, such as data profiling, data cleansing, data integration and data enrichment must be re-formulated according to the vision of a service community in the cloud.

In this paper, we analyze data quality engineering (DQE) and quality-aware data publishing (QDP) processes in the cloud computing. We contribute (i) a list of requirements to enhance the basic data profiling, data cleansing, and data enrichment activities using SaaS, DaaS and crowdsourcing models, and (ii) a cloud-based conceptual architecture and its data concerns for supporting these activities.

The rest of this paper is organized as follows: Section II presents the background of DQE and QDP. Section III shows a cloud-based conceptual architecture for data quality engineering and publishing. Related works are described in Section IV. We conclude the paper and outline our future work in Section V.

[1] In this paper we do not distinguish between data, information and knowledge.

II. Background of Data Quality Engineering and Publishing

A. Data Quality Engineering Phases

The migration of information systems from a monolithic to a network-based structure, where the potential data sources that organizations can use have dramatically been increased in size and scope, makes the issue of data quality more complex and controversial [12], [13]. The literature provides a wide range of techniques (e.g., record linkage and similarity measures) to assess and improve the quality of data. In the last decade, the research has focused on defining *data quality methodologies* [14], [15], [16], [17] that help to select, customize, and apply data quality techniques. A data quality methodology provides a set of guidelines that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data.

As defined in [12], in the most general case, a data quality methodology is composed of three phases: (i) *State reconstruction*, which is aimed at collecting contextual information on organizational processes, services and data in order to guide the DQE; (ii) *Assessment*, which measures the quality of data along relevant quality dimensions; (iii) *Improvement*, which concerns steps, strategies, and techniques for reaching new data quality targets.

The State reconstruction phase is optional if the Assessment phase can be based on existing documentation. Since methodologies typically make this assumption, we will not further discuss this phase. The Assessment and Improvement phases are composed of several sub-phases. In this paper, we focus only on the basic sub-phases (see Figure 1) that are traditionally performed by means of common DQE tools: *Data Profiling* for the Assessment phase and *Data Quality Improvement* for the Improvement phase.

Data Profiling analyzes the structure, relationships and content of existing data sources to create an accurate picture of the state of the data. Data profiling helps in planning the best ways to correct or reconcile data assets.

Data Quality Improvement selects and configures the most effective and efficient strategies and the corresponding techniques to improve data quality. Different data quality methodologies propose different types of strategies to reach the goal. In this paper we propose to divide these strategies into three different categories: (i) *data cleansing* strategies, which support the utilization of techniques to correct errors, standardize information, and validate data. Typical data cleansing techniques are *standardization/normalization* (i.e., replace or complement non-standard data values with corresponding values that comply with the standard) and *record linkage* (i.e., identify that data representations in multiple tables that might refer to the same real-world object). Data cleansing strategies are proposed by the Comprehensive methodology for Data Quality management (CDQ) [14] and by the Total Information Quality Management (TIQM) methodology [15]; (ii) *data integration* strategies, which support the linking of data elements about the same item available in different data sources and the consolidation of these data elements into a single view. The main goals of data integration consists in identifying duplicates, merging dissimilar data, purging redundant data and linking data sets. A typical data integration technique is *data and schema integration* (i.e., define a unified view of the data provided by heterogeneous data sources solving technological, schema and instance-level heterogeneities). Data integration strategies are proposed by CDQ and by the Datawarehouse Quality Methodology (DQM) [16]; (iii) *data enrichment* strategies, which support the incorporation of additional data into existing records. Typical data enrichment techniques are *data verification* (i.e., complete data with information from external data/service) and *data validation* (i.e., verify available data comparing to a reference database). Data enrichment strategies are proposed by the DataFlux methodology [17].

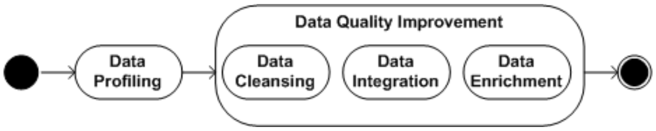

Figure 1. Basic sub-phases for data quality engineering

B. DaaS and SaaS for Data Quality Engineering and Publishing in the Cloud

Several existing tools/services can be used to profile, clean, integrate and enrich data inside an organization. An example is the DataFlux DfPower Studio [18] that supports data quality engineering activities allowing data providers [2] to build complex management workflows. However, most tools/services have not utilized benefits of the cloud environment where several external DaaS and SaaS can be utilized for complex DQE processes as well as crowd computing principles can be applied for improving the sharing and quality of these processes. Therefore, the basic DQE sub-phases described in Figure 1 should be revised according to the DaaS and SaaS principles.

Currently, several DaaS and SaaS are available over the Internet to profile, clean, integrate, enrich and publish data. According to [8], DaaS can be categorized into (i) *Read-only DaaS* which only provide data based on existing data sources, and (2) *CRUD DaaS* which are not simply a "Data marketplace" but allow the DaaS provider to create, retrieve,

[2] In this paper, we adopt the following terminologies: *Data provider* is the actor who wants to improve her data; The *DaaS provider* is the actor who offers her data as a service; The *SaaS provider* is the actor who offers her software as a service.

update and delete data. CRUD DaaS can be infrastructure-based services which typically just provide a storage capability and it is up to DaaS providers to define their own data schema and/or to publish their data. Examples of existing DaaS and SaaS, and their capabilities, that could be used for DQE activities are shown in Table I.

System	DaaS	SaaS	Capabilities
Strikeiron[19]	X (Read-only)		clean, verify and validate data.
Jigsaw [20]	X (Read-only)		clean, verify and validate business contact.
PostcodeAnywhere [21]	X (Read-only)		capture, clean, validate and enrich business data.
TheWebService [22]	X (CRUD)		publish and share data.
Caspio [23]	X (CRUD)		publish and share data.
The *Trillium Software Quality* [24]		X	clean and standardize data.
Uniserv Data Quality Solution [25]		X	profile and clean data.
Adeptia Integration Solution [26]		X	integrate data.

Table I
EXAMPLES OF DAAS AND SAAS AVAILABLE OVER THE INTERNET.

Currently, DaaS and SaaS are mainly used in isolation and are not composed into the whole life-cycle of DQE processes. There are a wide range of techniques (e.g., [27]) for service composition, but only few papers (e.g., [28], [29]) propose a SOA-based solution to perform DQE. Moreover, these solutions cover only a particular aspect (i.e., data quality assessment [28] and data provenance [29]) of the DQE process. Therefore, to the best of our knowledge, no comprehensive solution exists for composing DaaS and SaaS to perform DQE&QDP in the cloud.

In our vision, DQE&QDP solution providers and data providers could get several advantages by means of DaaS and SaaS composition in the cloud. A DQE&QDP solution provider can offer its solutions to be developed in a private or public cloud. Using such cloud-based solutions, data providers can perform data engineering and publishing activities for their data sources (e.g., databases and documents) managed by means of different enterprise software. When using cloud-based public/private DQE&QDP solutions, data providers can also utilize available DaaS and SaaS in the public cloud for solving their data engineering and publishing requirements which may not be fulfilled by existing DQE&QDP tools. Moreover, data providers can publish their data under DaaS thus becoming a DaaS provider and, following the SaaS principle, they can also publish their software to be used by other data providers in the cloud becoming also a SaaS provider.

III. TOWARD A CLOUD-BASED CONCEPTUAL ARCHITECTURE FOR DATA QUALITY ENGINEERING AND PUBLISHING

A. Move to Cloud and Requirements

To take into account the advantage of cloud computing and crowd computing, basic sub-phases for DQE and QDP can be conducted by using suitable SaaS, DaaS and crowdsourcing services. In order to reach this objective, the following requirements must be considered:

1) Moving DQE Tools to SaaS: such tools should be available as services in the cloud. Service engineering techniques [30] to perform the migration of legacy tools into services for the cloud must be applied.

2) Classifying DaaS, SaaS and Crowdsourcing Services: DaaS, SaaS and crowdsourcing services available in the cloud for DQE&QDP should be classified and associated with activities in DQE&QDP phases. In this view, existing service classification (e.g., UNSPSC [31]) should be enhanced to support the specification of DQE&QDP specific metadata reflecting data quality (DQ), quality of service (QoS), data/service usage, and data/service licensing [8].

3) Composing DaaS, SaaS and Crowdsourcing Services: existing service selection and composition techniques [27] must be extended for SaaS, DaaS, and crowdsourcing services used in DQE. Contemporary techniques focus mainly on quality of service (QoS) and pricing. To make them applicable in DQE processes, they will need to cover data quality engineering specific aspects such as data quality (DQ), data/service usage, and data/service licensing information. A systematic approach [32] to the description of such information needs to be developed together with data quality-aware service selection and composition techniques.

4) Sharing DQE&QDP Workflows: several DQE&QDP phases require the definition of workflows. A workflow is a set of activities (e.g., selecting a data source, parsing the data, verifying data and outputting the data into a new table) to be performed by using SaaS, DaaS and crowdsourcing services. Therefore, DQE&QDP workflow tools should be developed to allow data providers to create a new workflow or reuse and/or enhance existing workflows. Furthermore, data providers should be able to publish their DQE&QDP workflows. This fosters the creation of a community where specific workflows are shared by different users. In this view, existing research results, such as myExperiment [33], can be adapted for the DQE domain.

5) Quality-aware Data Publishing under DaaS: data providers should be able to publish their data for business purposes in the cloud. Existing DaaS for data publishing (e.g., TheWebService and Caspio) should be extended in order to allow DaaS providers to specify data quality (DQ), data usage, and data licensing information.

B. SaaS, DaaS and Crowdsourcing services for DQE

Basic DQE sub-phases in Figure 1 could be performed by means of cloud computing models. In this section, we provide a cloud-based conceptual architecture that enables Data Profiling (DP), Data Cleansing (DC), Data Integration (DI) and Data Enrichment (DE) activities by means of DaaS, SaaS, and crowdsourcing services. For space reasons, we limit the description of our architecture to DQE but our solution is also extensible to QDP.

As shown in Figure 2, our conceptual architecture includes five main building blocks: *DP SaaS*, *DC SaaS*, *DI SaaS*, *DE SaaS/DaaS* and *Crowdsourcing DQE*. These building blocks include several services which can be composed and executed for DP, DC, DI and DE activities.

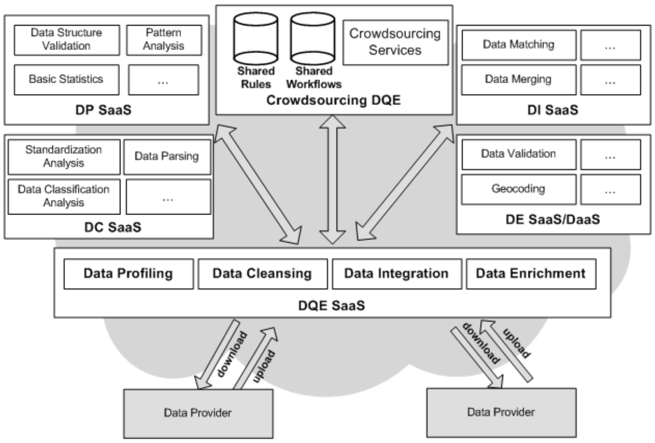

Figure 2. The cloud-based conceptual architecture for DQE.

The basic idea is that a data provider invokes a DQE SaaS able to support DP, DC, DI and DE services. The DQE SaaS can directly offer the services or it can act as a broker that selects and invokes the best SaaS/DaaS over the cloud able to satisfy the data provider request. In the following we detail core service categories in these building blocks.

1) Data Profiling SaaS: are used to examine the structure, relationships and contents of existing data sources. Data profiling SaaS are classified as (i) *Data Structure Validation* services which analyze the content of data by setting validation conditions; (ii) *Pattern Analysis* services which generate an alphanumeric pattern that represents each value in a selected field; (iii) *Basic Statistics* services which display basic statistics (e.g., number of min/max values, percentiles) about selected fields.

2) Data Cleansing SaaS: are used to correct errors, standardize information, and validate data. They are classified as (i) *Data Parsing* services which separate multi-part field values into multiple, single-part fields; (ii) *Standardization Analysis* services which conform data values to a particular standard in order to make similar items the same. For example, an `Address Format Analysis Service` can be used to check whether the data about address meet the required standard for specifying them; (iii) *Data Classification Analysis* services which classify and match data into classes. For example, a `Gender Analysis Service` can be used to determine whether a particular name is feminine, masculine, or unknown.

3) Data Integration SaaS: are used to link data elements about the same item available in different data sources. Data Integration SaaS are classified as (i) *Data Matching* services which create one-to-one mappings between similar fields in different data sources and (ii) *Data Merging* services which merge the data about the same item from all records into a single surviving record.

4) Data Enrichment SaaS/DaaS: are used to incorporate additional external data to add value to existing records. Data enrichment activities are supported by both SaaS and DaaS models. The difference is that a DE DaaS also provides the data (i.e., reference databases) to be used for the enrichment. DE Saas/DaaS are classified as (i) *Data Validation* services which verify, correct, and enhance data. Data validation services can be further classified into domain-independent and domain-specific services. For example `Address Validation Service` and `Phone Validation Service` are domain-independent data validation services for validating addresses and phone numbers, respectively, according to reference databases; (ii) *Geocoding* services which provide mappings between geographic information (e.g., latitude/longitude and ZIP codes). For example, `YahooPlaceFinder` [34] allows the conversion from addresses to geographic coordinates.

5) Crowdsourcing DQE: allows community participation and sharing in DQE processes. Current DQE activities do not support the crowdsourcing model which has demonstrated its several benefits [9]. In order to introduce this model into DQE activities, two aspects must be considered:

- several DQE activities require the definition of workflows. A data provider can decide to develop a new workflow or to use an existing one to perform its activities. The possibility to re-use workflows allows the creation of a community where workflows are shared by different data providers.
- DQE workflows are defined according to rules stored into a central data repository. These rules can be shared by different data providers.

The crowdsourcing of workflows and rules reinforces the vision of DQE processes in the cloud. As shown in Figure 2, the basic idea is that a data provider invokes the DQE SaaS for uploading/downloading workflows and rules. After the creation of a new workflow/rule, the data provider uploads it to the DQE SaaS asking for execution and specifying condition on sharing. The DQE SaaS executes the workflow/rule, returns the results to the data provider and shares the workflow/rules according to her specification. Alternatively, the data provider invokes the DQE SaaS to search and download a shared workflow/rule and to execute it locally or remotely.

C. Open Issues in Cloud-based DQE&QDP

As DQE&QDP are based on cloud SaaS and DaaS, several services and data concerns have to be considered. Generally, these concerns are discussed in SaaS concerns [6] and DaaS concerns [8]. However, in the context of DQE&QDP, they represent open issues. Currently, data sources being profiled and enriched are managed by services within an enterprise system, therefore several concerns, such as security, performance, data compliance, and data retention are governed by enterprise policies and the law of the countries where the enterprise is located. Moving to the cloud, these data sources are either hosted in the cloud or accessed from the cloud, and/or the data of these data sources have to be moved between the data provider's side and the cloud. Therefore, several concerns need to be addressed in DaaS and SaaS descriptions. We envisage the following open issues:

- *the quality of data used in the DE activities*: data provider's data is enriched by DaaS in the cloud, therefore, the quality of data offered by DaaS must guarantee to ensure that the enrichment process yields a better quality. It is crucial that the quality of data provided by DaaS is explicitly modeled into DaaS descriptions.
- *data life-cycle*: when the data to be enriched is sent to DaaS/SaaS or DQE workflows access the data to be enriched, there is a concern about how to manage the life-cycle of the data, such as distribution and disposition. Depending on the size of data to be enriched, the enrichment process might take a long time to finish. If the data is distributed or re-located, the compliance with business confidence rules must be accurately checked.
- *DQE performance*: the question of performance is important in case of large volume of data movement. DQE SaaS providers should develop adaptive algorithms to decide whether the data should be sent to DaaS/SaaS or move DaaS/SaaS close to the data.
- *network and data security/privacy*: the possibility to apply common network and data security/privacy concerns [35] to DQE&QDP DaaS and SaaS must be deeply investigated due to the possible high confidentiality of the data to be managed.

IV. Related Work

In the last decade the research has focused on defining *data quality methodologies* [14], [15], [16], [17]. The Comprehensive methodology for Data Quality management (CDQ) [14] supports the selection of the optimal DQE process that maximizes benefits within given budget limits. The Total Information Quality Management (TIQM) methodology [15] focuses on the activities for the integration of operational data sources that are used to create a datawarehouse. The Datawarehouse Quality Methodology (DQM) [16] studies the relationships between quality objectives and design options in data warehousing. Finally, the DataFlux methodology [17] provides support for enterprise data quality and data governance. Even if these data quality methodologies define practical approaches to analyze, improve and control data, the possibility to compose SaaS, DaaS and crowdsourcing services in order to perform DQE processes is not addressed. These methodologies should be revised considering the new aspects introduced by our cloud-based conceptual architecture.

Only few papers [28], [29] propose a SOA-based solution to perform DQE processes. A SOA-based semantic data quality assessment framework which supports automatically searching for proper data quality assessment services which fulfill user requirements is proposed in [28]. A dynamic framework for data provenance (i.e., the origins and routes of data) classification in a SOA system is described in [29]. Respect to our cloud-based architecture, these SOA-based solutions cover only a particular aspect (i.e., data quality assessment [28] and data provenance [29]) of the DQE process and they do not address the possibility to use crowdsourcing rules, workflows and services.

V. Conclusions and Future Work

In this paper, we have analyzed current data engineering and publishing processes supported by various tools/frameworks and the role of cloud computing models on supporting these processes. To utilize the benefits of the DaaS, SaaS and crowdsourcing models for DQE&QDP processes we have presented a set of requirements to extend basic activities in these processes with new activities using SaaS, DaaS, and crowdsourcing services in the cloud. Moreover, we have proposed a conceptual architecture for service-oriented data engineering and publishing in the cloud and we have analyzed open issues of our proposals.

However, our work is just an initial start. We will provide a detailed methodology to extend basic DQE activities according to DaaS, SaaS and crowdsourcing models. Furthermore, we will evaluate the cloud-based conceptual architecture by means of case studies based on real systems.

Acknowledgment

This work is partially supported by the SAS Institute srl (Grant Carlo Grandi) and by the EU under the FP7 Commius project

References

[1] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Commun. ACM*, vol. 41, no. 2, pp. 79–82, 1998.

[2] W. Jung, "A review of research: an investigation of the impact of data quality on decision performance," in *ISICT '04: Proc. of the 2004 Int. Symp. on Information and communication technologies*, 2004, pp. 166–171.

[3] E. Bertino, A. Maurino, and M. Scannapieco, “Guest editors’ introduction: Data quality in the internet era,” *IEEE Internet Computing*, vol. 14, pp. 11–13, 2010.

[4] M. Gertz, M. T. Özsu, G. Saake, and K.-U. Sattler, “Report on the dagstuhl seminar,” *SIGMOD Rec.*, vol. 33, no. 1, pp. 127–132, 2004.

[5] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility,” *Future Gener. Comput. Syst.*, vol. 25, no. 6, pp. 599–616, 2009.

[6] J. Viega, “Cloud computing and the common man,” *Computer*, vol. 42, pp. 106–108, 2009.

[7] A. Dan, R. Johnson, and A. Arsanjani, “Information as a service: Modeling and realization,” in *SDSOA ’07: Proceedings of the International Workshop on Systems Development in SOA Environments*, 2007, pp. 2–15.

[8] H. L. Truong and S. Dustdar, “On analyzing and specifying concerns for data as a service,” in *Proc. of the 4th IEEE Asia-Pacific Services Computing Conference (APSCC 2009)*, 2009, pp. 87–94.

[9] D. C. Brabham, “Crowdsourcing as a model for problem solving,” *Convergence: The International Journal of Research into New Media Technologies*, vol. 14, no. 1, pp. 75–90, February 2008.

[10] D. Durkee, “Why cloud computing will never be free,” *Commun. ACM*, vol. 53, no. 5, pp. 62–69, 2010.

[11] D. Owens, “Securing elasticity in the cloud,” *Commun. ACM*, vol. 53, no. 6, pp. 46–51, 2010.

[12] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, 2009.

[13] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, “Overview and framework for data and information quality research,” *J. Data and Information Quality*, vol. 1, no. 1, pp. 1–22, 2009.

[14] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., 2006.

[15] M. A. Jeusfeld, C. Quix, and M. Jarke, “Design and analysis of quality information for data warehouses,” in *In proc of International Conference on Conceptual Modeling (ER ’98)*, 1998, pp. 349–362.

[16] L. P. English, *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley and Sons, Inc., 1999.

[17] D. Teachey, “Five steps to more valuable enterprise data,” in *DataFlux White Paper*, 2006.

[18] http://www.sas.com/data-quality/df-power-studio/index.html, last access: 05 September 2010.

[19] “Strikeiron,” http://www.strikeiron.com, last access: 05 September 2010.

[20] “Jigsaw,” http://www.jigsaw.com/, last access: 05 September 2010.

[21] “Postcodeanywhere,” http://www.postcodeanywhere.co.uk, last access: 05 September 2010.

[22] “Thewebservice,” http://www.thewebservice.com/, last access: 05 September 2010.

[23] “Caspio,” http://www.caspio.com/, last access: 05 September 2010.

[24] “The trillium software quality,” http://trilliumsoftware.com/home/products/page.aspx?id=112, last access: 05 September 2010.

[25] “Uniserv data quality solution,” http://www.uniserv.com/en/data-quality-service/software-as-a-service.php, last access: 05 September 2010.

[26] “Adeptia integration solution,” http://www.adeptia.com/solutions/saas_integration.html, last access: 05 September 2010.

[27] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam, and Q. Z. Sheng, “Quality driven web services composition,” in *WWW ’03: Proceedings of the 12th international conference on World Wide Web*. New York, NY, USA: ACM, 2003, pp. 411–421.

[28] Y. Zhou, S. Hanß, M. Cornils, C. Hahn, S. Niepage, and T. Schrader, “A soa-based data quality assessment framework in a medical science center,” in *Proc. of the 14th International Conference on Information Quality (ICIQ 2009)*, Potsdam, Germany, 2009, pp. 149–160.

[29] W.-T. Tsai, X. Wei, D. Zhang, R. Paul, Y. Chen, and J.-Y. Chung, “A new soa data-provenance framework,” *Autonomous Decentralized Systems, International Symposium on*, vol. 0, pp. 105–112, 2007.

[30] G. Canfora, A. R. Fasolino, G. Frattolillo, and P. Tramontana, “A wrapping approach for migrating legacy system interactive functionalities to service oriented architectures,” *J. Syst. Softw.*, vol. 81, no. 4, pp. 463–480, 2008.

[31] http://www.unspsc.org/, last access: 05 September 2010.

[32] M. Comerio, H.-L. Truong, F. De Paoli, and S. Dustdar, “Evaluating contract compatibility for service composition in the seco2 framework.” in *Proc. of ICSOC/ServiceWave ’09*, Stockholm, Sweden, November 23-27 2009, pp. 221–236.

[33] D. De Roure, C. Goble, S. Aleksejevs, S. Bechhofer, J. Bhagat, D. Cruickshank, P. Fisher, D. Hull, D. Michaelides, D. Newman, R. Procter, Y. Lin, and M. Poschen, “Towards open science: the myexperiment approach,” *Concurrency and Computation: Practice and Experience*, vol. IN PRESS.

[34] “The yahooplacefinder,” http://developer.yahoo.com/geo/placefinder/, last access: 05 September 2010.

[35] D. Lin and A. Squicciarini, “Data protection models for service provisioning in the cloud,” in *SACMAT ’10: Proceeding of the 15th ACM symposium on Access control models and technologies*, 2010, pp. 183–192.