# Provenance Management for Data Quality Assessment

Hua Zheng[1,2]
[1]School of Management and Engineering, Nanjing University, Nanjing, Jiang Su, China, 210093
[2] Department of Computer and Information Management, GuangXi University of Finance and Economics, Nanning, GuangXi, China, 530003
Email:gxhuazheng@yahoo.com.cn

Qinghua Zhu[3], Kewen Wu[3]
[3]Department of Information Management, Nanjing University, Nanjing, Jiang Su, China, 210093
Email:qhzhu@nju.edu.cn, kewen-wu@163.com

***Abstract*—The ultimate goal of data quality management (DQM) is to improve the data quality (DQ) to facilitate enterprises decision-making, and the data quality assessment (DQA) is an important aspect in the process of DQM. Existing research in DQA focuses on the establishment of evaluation indicators and quantified methods in specific areas of application, but does not take into account the evolution of the data. For the current context of complex heterogeneous data environment, DQA framework based on provenance and SOA is designed, and provenance management process is focused to describe, finally the provenance model is defined and implemented. Overall, an example described in the paper demonstrates the necessity and feasibility of introducing provenance into DQA.**

***Index Terms*—Data quality management, Data quality assessment, Provenance, SOA, SPARQL**

## I. Introduction

Data is the enterprise critical strategic resource, and reasonably, effectively using the correct data can guide business leaders make the right decisions to enhance the competitiveness of enterprises. Unreasonably using incorrect data (ie, poor DQ) can lead to the failure of decision-making.

A survey of the total data quality management (TDQM) [1] project from Massachusetts Institute of Technology (MIT) is shown: only 35% of the companies trust the own data, only 15% of the companies trust the partner data. In the United States there is cost about 600 billion U.S. dollars annually to ensure DQ, or to compensate for DQ problems caused economic losses. The other survey is shown by An IDC White Paper(2008)—the data integration and data quality in China market：”Because a complex data environment is formed by the background of China's special construction of software, more than 70% of surveyed Chinese enterprises have been built or are building data integration projects, and focus on hot issues of DQ.” It can be seen that the issues of DQ have been begun to attach importance by the enterprises.

Currently, most enterprise information technology projects are not starting from scratch, and they need to use the data that already resides in the enterprise and has the poor quality. In particular, because the current scope of data has been expanding and more widely shared and increasingly diverse forms of data have been, a large, complex, heterogeneous data environment has been formed. Therefore, the analysis of data generation and evolution process, then evaluate the quality and accuracy of data, as well as revise data result that is very important.

In this paper, firstly the background of DQA and basic ideas are introduced; In section 2, the related research works are given, which includes DQA, provenance and SOA; The framework of DQA based on provenance and SOA is designed in section 3; Provenance model is given in section 4; A concrete example is implemented that demonstrates the necessity and feasibility of introducing provenance into DQA in section 5; Finally a brief summary is given.

The contributions of this paper are: (1) establish a scalable and extensible framework of provenance-aware DQA; (2) develop a provenance model based on content.

## II. Related Works

The DQA process has the lack of sufficient attention on enterprise infomationization, and existing technologies and methods have their limitations. The current study more cut from a methodological point of view, the object is the target data itself, without taking into account the evolution of complex environment of data on the quality of the process. Provenance is a current research focus, and how to use provenance technology to achieve DQA is a worthy research direction. Some relevant technologies are summarized in this section.

### *A. Data Quality Assessment*

Q[2] is defined as "suitability for use", and this definition is now widely accepted. Much research in DQM[3] focuses on methodologies, tools and techniques for improving quality, and DQA [4] is an important part of DQM(shown in Fig.1). The Assessment for DQ depends on the individual use of data, and for different people under different circumstances the "suitability for use" is different. DQ is relative, not independent of the


doi:10.4304/jsw.7.8.1905-1910

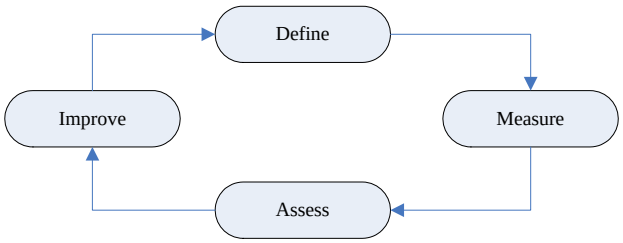


Figure 1. Data quality management process

use of data consumers to be executed. Present method is divided into two categories: qualitative and quantitative strategies. From the various dimensions of the combination of qualitative and quantitative point of view to analyze the "good" or "bad", it is currently the main research of DQA. The current mainstream methods of DQA are summarized in literature[5], including: TDQM、DWQ、TIQM、AIMQ, a total of 13 ways; Yang et al[6] designed a six-dimensional DQA model, and the quality situation of the data system can be assessed by the application of quantitative indicators. It is shown that many scholars, mainly from methodological point of view to research on DQA. Although these assessment methods are not the same, but its scope of application is different, common ground is a combination of subjective and objective to complete the assessment activities.

### *B. Provenance*

Provenance's etymology is the French verb 'provenir', which means to come forth, originate. According to the Oxford English Dictionary, provenance is defined as follows: (i) the fact of coming from some particular source or quarter; origin, derivation. (ii) The history or pedigree of a work of art, manuscript, rare book, etc; a record of the ultimate derivation and passage of an item through its various owners.

The common sense definition consider provenance to be the derivation from a particular source to a specific state of an item. Inspired by previous works [7][8][9][10][11][12][13], we propose the following definition of provenance, which makes explicit the notion of process: the provenance is the process that led to that piece of data.

Data Provenance is a technique for recording the log of transformation process of data to its derived form [7]. It answers the queries about originator of Data, transformation and its path to the derived form. These queries are necessary to be answered to get the trust on data for its use in simulation and experiments.

### *C. Service-Oriented Architecture*

SOA[14](Service-Oriented Architecture) is an abstract model, and represents a specific implementation which does not involve the software infrastructure but direct service-oriented. Enterprise business logic can be achieved at lower cost for rapid reconstruction.

Many open, large-scale systems are typically designed using a service-oriented approach, usually referred to as service-oriented architectural style [15]. In this case, we take services to be components that take inputs and produce outputs. Such services are brought together to solve a given problem such as DQA.

### *D. Combination of the Both*

Currently the Internet, grid and cloud computing and other data intensive applications have led to more complex DQ issues, the level of its quality is difficult to determine by the independent dataset alone. The reasonable assessment of the quality of dataset integrated data provenance by which the process of data generated and the evolution of the specific process can thoroughly be understood can be carry on. After the introduction of provenance, we can know that the tuple come from which data source, and then determine the source of data uncertainty, and finally know the way by which the generation of data to calculate the size of uncertainty.

## III. Architecture Of Provenance-Aware DQA

Provenance in all open distributed systems is necessary to track since data proliferation is too frequent. All network based data driven applications require provenance of data. Realistic DQA must be carry for a specific environment and user, and there is no uniform standard. So in this paper, a framework of DQA is presented that enables the various functions of assessment to dynamically be added and in which the evolution of the original data by data provenance can be technically analyzed, and the appropriate system functionality of assessment can be selected for DQA.

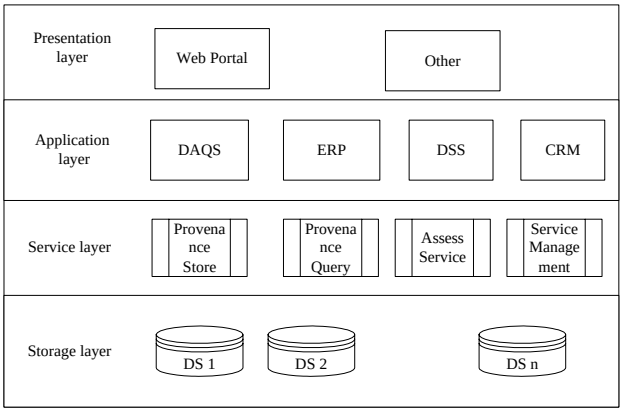


Figure 2. Logical architecture

The framework is divided into logical and physical architecture of two parts.

### *A. Logical Architecture*

The assessment functions based on SOA will be abstracted in the form of service, and then the framework of DQA that can adapt to all kinds of requirement for assessment in a different environment is created. The logical architecture of the framework shown in Fig.2, is divided into four layers (storage layer, service layer, application layer, presentation layer). Its core is the service layer, in which all the functional components are packaged into the form of web service, here including the assessment of service components, provenance storage components, provenance query components, service management components. All of the services will be

integrated by the ESB (Enterprise Service Bus). This architecture provides an extensible and scalable framework in which services and modules can be added and updated without disrupting exiting functionality and is particularly well-suited to provenance recording at a coarse-grained level in that recorded provenance. There is no limitation for the architecture to support fine grain provenance.

*B. Physical Architecture*

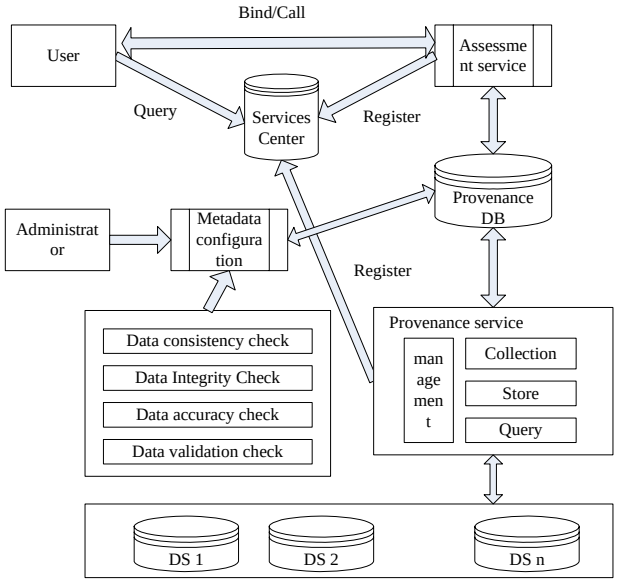


Figure 3. Physical architecture

The physical architecture of DQA system based on provenance and SOA is shown in Fig.3, and its core modules are as follows: (1) assessment module: a variety of assessment components are packaged into the form of web service, which can be dynamically added and set to suit the different assessment requirements.(2)provenance service module: responsible for the collection, store, query and management functions of data provenance, data can be traceable that is to be achieved. The specific process of DQA can be divided into three parts: (1) metadata configuration. Data quality metrics are applied to determine data quality levels; (2) provenance services. The differences between the target systems are analyzed, and then the recommendations for the elimination of differences are created. Then provenance services are called to collect, store and query provenance data. (3) Alignment and harmonization requirements for each relevant data elements are assessed. The results of DQA indicators are interpreted, and translated into business terms. User query the registry center for the required assessment services, and then call the services to evaluate quality of the target data source by using pre-established evaluation criteria. Detailed reports, charts and summary are created to describe DQ levels and provide recommendations.

*C. Provenance Process Design*

One of the core services is provenance service, so the design of provenance service process is very important. It concerns the data content and its evolution, and its core is the traceability of the data. In summary, the lifecycle of provenance service is composed of five different phases: plan process, data consistency process, record process, request process, and implementation process (shown in Fig.4) .

(1)Plan process

It is a prerequisite for the implementation of provenance management, and the main actors are all traceable participants of data management. The main function of this sub-process is to determine what include the distribution of traceable data, collection, sharing and preservation methods, and input, the link between internal processes and the management of output.

(2)Data consistency process

It determines that objects and data exchange approaches between objects. This process includes allocation for the object identified, the physical location of the distribution of identification, and data exchange procedures.

(3)Record process

It determines the distribution of traceable object identity, application and acquisition mode, and the collection, sharing and preservation methods of the tracing data in the entire management process.

(4)Request process

This process determines the request and response mode of the traceability, and any participant may propose traceability request, in which is included by the tracing requests, receiving tracing requests, responding tracing requests, receiving replies and other steps.

(5)Implementation process

This process enables the DQA module based on the actual requirements to trace the provenance object by using the default process of traceability, in which by Participants the traceability is implemented for the provenance objects, and meet the assessment requirements.

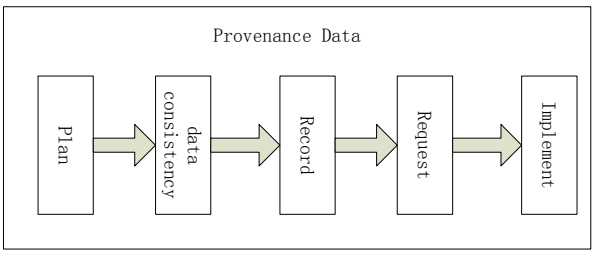


Figure 4. Provenance service process

## Ⅳ. PROVENANCE MANAGEMENT TECHNIQUES

In order to collect and query the provenance data in provenance management module, firstly the semantic model which can describe provenance is given, and the corresponding process design is presented. Detailed exposition is given below.

*A. Semantic Model of Provenance*

According to the literature [16], we studied and

implemented the storage and query of provenance. These include some important related concepts, and the definitions are given.

Definition 1.1 a namespace is countable infinite set, which can be recorded as NS. All string sets can be viewed as a namespace.

Definition 1.2 a named value is denoted by (name, value-list), name$\in$NS, value-list is a data list, which can be empty.

Definition 1.3 ID (identifier) can be viewed as a named value.

Definition 1.4 a application environment is denoted by （NS,CLK,ADDR,DICT）. NS is a namespace, CLK is the system recognizing time, ADDR is the system recognizing address, DICT $\subseteq$ NS， is a pre-defined character set. It is used to represent minimum individual of data and provide a level of interaction.

Definition 1.5 A annotation is defined as a named value. Each named value can be viewed as an entry, which can be an extension way used to describe the data.

Definition 1.6 a static entity record can be defined as(entity-id, entity-address, entity-type, annotation, snapshot-time, record-id)，entity-id can be viewed as identifier, entity-address $\subseteq$ ADDR, entity-type $\subseteq$ DICT, annotation can be defined from the preceding definition, snapshot-time $\subseteq$ CLK, record-id $\subseteq$ NS. Entity information of some aspects at some point will be captured by that object.

Definition 1.7 activity type can be defined as(type-name, incoming-list, outgoing-list, annotation)， type-name is an identifier, incoming-list is an orderly named value (the named value came from each column is called input queue), outgoing-list is also an orderly named value (the named value came from each column is called output queue).

Definition 1.8 an activity record can be defined as(activity-id, activity-type, activity-span, annotation, record-id),activity-span=(start,end),start and end $\subseteq$ CLK.

Based on the above definition, we give a description of provenance entities and their relationships. A provenance entity can be viewed as a static entity record or a activity record, which can be defined as(type, entity)， type $\subseteq$ {S,A}. Suppose type=S, then the provenance entity is a static entity record; suppose type=A，then the provenance entity is a activity record. a provenance relationship is a relationship between the two provenance entities, which can be defined as(causal-entity, consequential-entity, role, annotation, relationship-id), causal-entity and consequential-entity both are provenance entity, role $\subseteq$ DICT，relationship-id $\subseteq$ NS.

There are two types of provenance entities: static entity records and activity records. A provenance entity can be accessed, so we can define the storage of the provenance entity as a physical object that can be accessed (called PES). Provenance relationship is actually the relationship between the two provenance entities, and we define a provenance relationship store here (called PRS) as a storage object of provenance relationship. Therefore, the provenance storage is actually to achieve storage of provenance entities and their relations.

An important operation on the provenance is the query, so we have developed a query model by which provenance entities and relationships can be operated. The query model is constituted by PES, PRS, and various queries operators. With comparing the general data query, the difference is that the results of a provenance query include not only the structure of complex content, but also includes the structure of complex relationship. The query of provenance information is mainly completed by defining operators of these two types of objects. TO query PES (type as a query parameter) as an example for a description: Suppose S is a object of PES, type=t，use S and t as input, the operator can be defined as：

$$\sigma_{type}(S,t) = \sigma(S, \odot type = t) \quad (1)$$

TO query PRS (relationship-id as a query parameter) as an example for a description: Suppose S is a object of PRS, relationship-id=R， use S and R as input, the operator can be defined as：

$$\sigma_{rel-id}(S,R) = \sigma\left(S, \in \left(\odot \text{relationship - id}, \text{R}\right)\right) \quad (2)$$

Here, $\{\sigma, \odot\}$ is an access sequence to retrieve. You can use the operators to retrieve the records of all activities of PES and PRS.

It is known that the semantic model is constituted by provenance entities and their relationships from the above definitions. By defining the semantic model, the provenance information of all the static and dynamic elements in application system can be described by a flexible way.

### *B. Implementation of Provenance Store*

The provenance information of a static entity record maintained in XML file contains {entity-ID, entity-address, entity-type, annotation, snapshot-time, record-id}. The example XML file generated by local node is as follows:

```
<provenance information>
  < entity-ID >111</ entity-ID >
  < entity-address >B</ entity-address >
  < entity-type >2</ entity-type >
  < annotation >(Input,1)</ annotation >
  < snapshot-time >083012142010</ snapshot-time >
  < record-id >R08</ record-id >
  </provenance information>
```

In this example 111 is the dataset id existed in address B at time 08:30 am dated 14th Dec, 2010. B is the id of one of nodes who is the user of this dataset. It refers to the evolving physical existence, while the latter one refers to a virtual representation (as a record) of some aspects of that entity at a particular moment. The algorithm working in archive generates tree by merging these XML files on the basis of data access and timestamp of same entity-ID. The provenance can be tracked by using simple tree traversal algorithm from leaf to root.

### *C. Implementation of Provenance Query*

Provenance queries conducted using SPARQL Protocol and RDF Query Language [17].

In the case of provenance data, queries are translated into patterns understood by which in turn queries the semantic repository to return the results. To understand this procedure, let us consider an example: Given a dataset(say D), a user wants to obtain details of all the analysis already applied on this dataset and the parameters associated with each analysis. A corresponding SPARQL query can be defined as:

Select ?datasetName, ?processName, ? parameter Where
{{?data < datasetName > ?datasetName}
{?process < entity-type > ?processName}
{?process <hasParameter> ?parameter}
{?data < entity-ID > ?process}
{?data<hasAnnotation > Annotation (D)}}

The example described above, though being simplified to facilitate the understanding of the query structures, demonstrates the feasibility of a full spectrum of provenance support by the implementation.

### *D. Experiment and Results*

The experiments were conducted using Oracle10g DBMS on a Windows Xp with 1.6 GHz Intel processors and 1GB of main memory.

The four datasets with increasing size of data used to evaluate. The size of DS1, DS2, DS3, DS4 are 1.4G, 13.2G, 28.2G, 57.2G. The SPARQL query patterns corresponding to the example provenance queries represent varying levels of query patterns complexity in terms of total size of data. The four queries were executed against the four datasets, DS1 to DS4. Fig.5 shows that the response time of the query increases with increasing size of data.

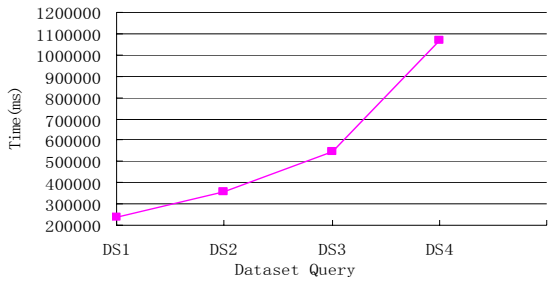


Figure 5. The response time for provenance query with increasing size of dataset

The set of results demonstrate: (1) it is acceptable in the actual operation that query response time increases with the increasing amount of data. (2) The need for effective optimization techniques to enable practical use of the query.

## Ⅴ. Case Study

In the absence of provenance recording, a DQA system lacks information about the context, configuration, and parameters used for data synthesis, making it difficult to validate scientific analysis.

We consider illustrating the benefits of provenance-aware DAQ system mention gained through the support of provenance.

Here, the DQ problems of the sugar factory are selected to study. In the infomationization process of sugar factory, a distributed control system (DCS), equipment management systems (EMS), condition monitoring systems (CMS), enterprise resource planning (ERP) systems and so on are respectively constructed. For the different business lines, channels or products categorization, the data is often stored and processed in different ways by using different technologies. Core enterprise information is distributed in multiple vertical systems with multiple copies. The maintenance information of each system is according to their own context, regardless of the context of the whole enterprise, which further exacerbates the inconsistencies in the process of business.

Based on the solution that was discussed in the previous sections, we developed a DQA system(shown in Fig.6) for the sugar factory, in which the data of management and control will be monitored by provenance management, and the problems of DQ caused by misuse and false behavior will be found. The dimension of evaluation can be customized by user. The major algorithm of evaluation is using simple ratio, maximum/minimum operation, weighted average, etc.

This system that is currently implemented with Java v1.6 on top of Oracle 10g provides strong support to enterprise DQM and contributes heavily to control/management integration.

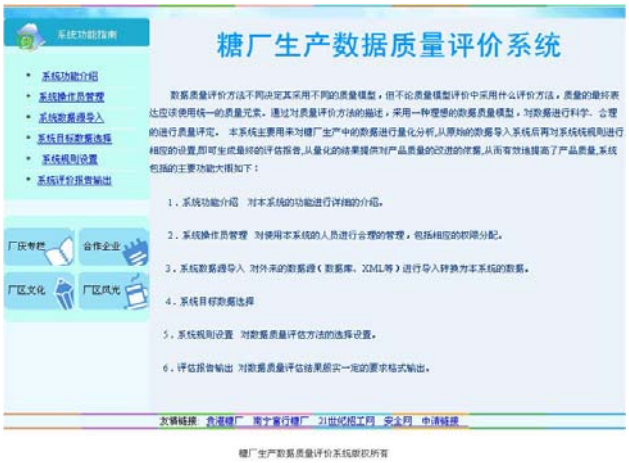


Figure 6. System implementation

## Ⅵ. Conclusions And Future Work

Currently normal data quality evaluation dimension and system is not usable in practical applications. In order to find out the optimized evaluation method, a specified method has to be carefully evaluated and selected in terms of individual application. This paper proposed an evaluation solution frame based on SOA structure, evaluating feasibility of proposed solution and adjusting it dynamically. This evaluation system is provided as a web service. The key point of this paper is not in how to improve or design a evaluation system, however, it focuses on the data detail and evolution of genealogy technology, paying much attention to how these details and evolution process contribute to the data quality

evaluation efficiency. The paper also defines a semantic model, laid a theoretical foundation for the future provenance technology research, providing a theoretical model for provenance data searching. But because of the complexity of the genealogy data, it is not only necessarily to search its content but also needs to understand its evolving process, therefore this searching model is far from perfect.

The future works will be focused on further optimizing the details of this solution, including structural provenance model, automatically collecting and storing provenance data, and how to secure the provenance data, etc.

## ACKNOWLEDGMENT

This work was supported in part by a grant from the Research Foundation of Philosophy and Social Science of GuangXi Province, China (No.08FTQ001), National Social Science Foundation Important Project, China (No.10ATQ004), Ministry of Education of the People's Republic of China, Humanities and Social Sciences project (No.09YJA870014), the Graduate Student Innovation Project of Jiangsu, China (No.CX09B_027R and CX10B_022R), the Key Project of Social Science of Education Department of Jiangsu Province, China (No.2010ZDIXM022).

## REFERENCES

[1] R.Y.Wang, "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, vol.41, no.2,pp.58-63,1998.

[2] R.Y.Wang, D.M.Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol.12,no.4,pp. 5-33,1996.

[3] A.Evena, G.Shankaranarayananb, P.D.Bergerc, "Evaluating a model for cost-effective data quality management in a real-world CRM setting," *Decision Support Systems* ,vol.50,no.1, pp.152-163,2010.

[4] L.Pipino, Y.Lee, R.Y.Wang, "Data Quality Assessment," *Communications of the ACM*, vol.45, no.5, pp.211-218, 2002.

[5] C.Batini, C.Cappiello, C.Francalanci et al, "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys*, vol.41,no.3, pp.1-52,2009.

[6] Q.Y.Yang, P.Y.Zhao, D.Q.Yang et al, "Research on Data Quality Assessment Methodology,"*Computer Engineering and Applications(in Chinese)*, vol.40,no.9, pp.3-4,15,2004.

[7] P.Buneman, S.Khanna, W.C.Tan, "Why and where: a characterization of data provenance," In Proceedings of the 17th International Conference on Data Engineering, London, UK, April 2001,pp.316-330.

[8] Y.Simmhan, B.Plale, D.Gannon, "A Survey of Data Provenance in e-science," *ACM SIGMOD Record*, vol.34, no.3, pp.31-36, 2005.

[9] Y.Cui, J.Widom, J.L.Wiener, "Tracing the Lineage of View Data in a Warehousing Environment", *ACM Transactions on Database Systems*, Vol.25, No.2, pp.179–227, 2000.

[10] P.Groth, M.Luck, L.Moreau, "Formalising a protocol for recording provenance in grids," In Proceedings of the UK OST e-Science second All Hands Meeting 2004 (AHM'04), Nottingham,UK, September 2004,pp.217-245.

[11] L.Moreau, L.Chen, P.Groth et al, "Logical architecture strawman for provenance systems," Technical report, University of Southampton, 2005.

[12] M.Szomszor, L.Moreau, "Recording and reasoning over data provenance in web and grid services," Lecture Notes in Computer Science, 2003,vol.2888, pp.603–620.

[13] W.C.Tan, "Provenance in Databases: Past, Current, and Future," *IEEE Data Engineering Bulletin*, vol.30, no.4, pp.3-12,2007.

[14] wiki website, http://en.wikipedia.org/wiki/Service-oriented_architecture, Dec 2010.

[15] P.S.Munindar, N.H.Michael, *Service-Oriented Computing: Semantics, Processes, Agents*, John Wiley & Sons Ltd, 2005.

[16] Y.B.Tang, M.Mani, "Butterfly-A Provenance Management System," CS Technical report, Worcester Polytechnic Institute, 2008.

[17] SPARQL Query Language for RDF, W3C Recommendation, http://www.w3.org/TR/rdf-sparql-query/, Dec 2010.

**Hua Zheng**, was born in Nanning, GuangXi, China, in March 25, 1978. He got his Master Degree of Computer Application from Department of computer Science, GuangXi University in 2004. Now, he is a PHD student in School of Management and Engineering, Nanjing University.

He had been working for more than 6 years and currently is a associate professor in the Department of Computer and Information Management at GuangXi University of Finance and Economics and hosted a couple of projects sponsored by GuangXi Province and GuangXi Finance Bureau, published more than 10 academics paper in journals and conference proceedings. His research interests are network management information system, data integration and electronic commerce. Mr. ZHENG is also awarded as excellent youth lecturer of GuangXi University of Finance and Economics.

**Qinghua Zhu**, doctor, doctoral supervisors in Department of Information Management, Nanjing University. He was born in 1963. Now, he has trained several PhD student, and published many papers( SCI, EI). His research interests are network management information system.

**Kewen Wu**, was born in Hubei, China, in March 19, 1985. He got this Master Degree of Management Science and Engineering (e-commerce) from School of Information Management, Wuhan University in 2009. Now, he is a PhD student in Department of Information Management, Nanjing University. His research interests are management information system, human computer interaction and electronic commerce, and more than 8 journal or conference papers have been published.