A Data Quality Management Maturity Model

Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park

Many previous studies of data quality have focused on the realization and evaluation of both data value quality and data service quality. These studies revealed that poor data value quality and poor data service quality were caused by poor data structure. In this study we focus on metadata management, namely, data structure quality and introduce the data quality management maturity model as a preferred maturity model. We empirically show that data quality improves as data management matures.

Keywords: Data quality, metadata, maturity model, standard data, data value quality, data service quality, data structure quality, data management, process quality.

I. Introduction

Quality plays an important role as one of the powerful competition advantages for those companies that run businesses in the information society [1]. Data quality of an information system is regarded as the most important factor because it is the basis of an information system [2]-[10]. Low quality data brings several negative effects to business users through the loss of customer satisfaction, high running costs, inefficient decision making processes, and performance [5], [8], [10], [11]. These shortcomings of low quality data affect not only corporate competitiveness but also have negative effects on the organizational culture, such as a demoralization of employees and a trend of mutual distrust within an organization. There have been many studies on data quality to solve such problems [2], [11], [12].

However, most previous studies have considered data value quality and service quality as the main data factors, and evaluated the quality level based on them [13], [14].

Meanwhile, we are focusing on data structure quality and its management. Structurally, disordered data will give rise to the wrong data value and data service. To manage and evaluate structural quality, it is essential to manage metadata [13].

In this study, we will define metadata [15], [16] to maintain high quality data and will introduce a data quality management maturity model based on the capability maturity model discussed in [17]-[21] to manage metadata.

This model is to be used as a management tool for data structure quality, whereas previous studies were applied to the evaluation and management of quality. Our data quality management maturity model can be applied to business fields to appraise their levels of data quality management and to acquire better quality that will make companies more competitive.

Manuscript received Mar. 30, 2005; revised Jan. 16, 2006.

Kyung-Seok Ryu (phone: + 82 42 860 5143, email: ksryu@etri.re.kr) is with IT Services Research Division, ETRI, Daejeon, Korea.

Joo-Seok Park (email: jspark@khu.ac.kr) is with Department of Business Administration, Kyung Hee University, Seoul, Korea.

Jae-Hong Park (email: jhp1@andrew.cmu.edu) is with Department of Statistics, Stanford University, Palo Alto, USA.

II. Data Quality Architecture

1. Definition of the Data Quality Domains

Early studies of data quality focused on data status quality and data service quality. Later, the importance shifted from data status quality to data structure quality [5], [22]. More recently, integrated data quality management studies have included a data management process as well as data value, service, and structure.

Previous studies evaluated only data quality. But integrated data quality management includes not only the evaluation but also the maintenance and improvement of quality.

Figure 1 shows that data quality factors can be divided into three domains: status quality factors (data value and data service), data structure factors, and data management process quality factors.

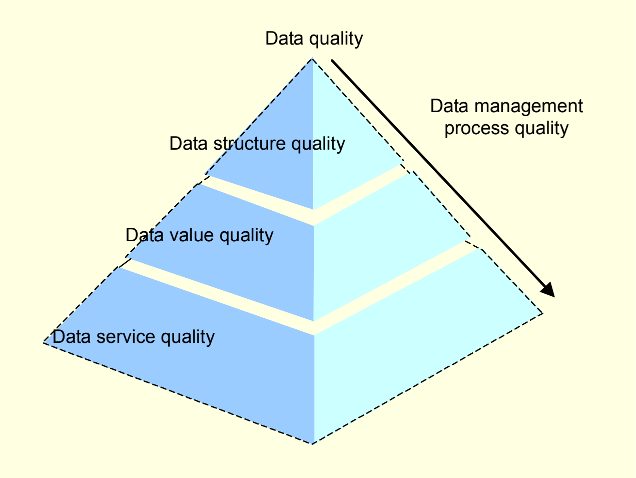


Fig. 1. Data quality domains.

2. Definition of Data Quality

Figure 2 shows that data quality evaluation and management should be maintained while considering depth and width at the same time. The depth of data quality means independent data quality, which includes accuracy, completeness, up-to-dateness, and searching ability.

The width of data quality represents integrated data quality under a discrete information system and database environment. The width of data quality includes data structure quality in regard to corporate integration and consistency.

Previous studies have focused mainly on the depth of data quality. However, under the discrete and complicated information system environment, it is hard to evaluate data quality objectively without considering the width of data.

For example, although department A has accurate and complete data of a product, department B might manage its data as a separate code and domain. In such a case, data should be manipulated to be evaluated and managed in width.

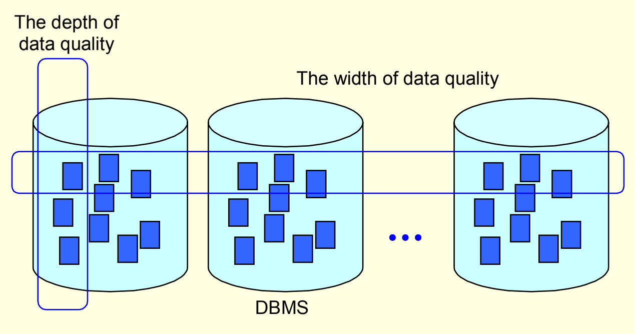


Fig. 2. Data quality with respect to depth and width.

To enhance the depth of data quality, data quality management should be processed with regard to corporate integration and consistency [23], [24]. This enterprise data quality management ultimately brings up data structure quality management through data standardization and data architecture management. We will discuss this in detail for the data quality management maturity model.

3. Definition of Metadata Architecture

Table 1 shows the categories of metadata to be managed for data structure quality management in the respect of an information system structure.

Standard data information defines the common standard of the structure and domain for a corporation. Through standardization, companies will develop data, improve the data structure quality level, and finally enhance corporate data quality.

For this process, logical data objects should be managed first. These are analyzed at the stage of logical modeling of data

Table 1. Classification of metadata in the respect of layer structure.

Category	Definition	Example
Standard data information	Logical information of common data factors (business object) which are defined at the enterprise level	Information of common data for the enterprise data model: entity information, attribute information, domain information, etc.
Physical database information	System catalogue information of the physical database systems which are generated at each base systems	Physical DBMS management information: database names, table names, view names, column name, etc.
Mapping information of standard data and physical database	Information which manages the relationship between standard data and physical database	Mediator information: source and target mapping information, transformation rules, etc.

modeling [25]-[29].

Among these logical data objects, some data factors are selected as the enterprise standard and managed as metadata information. These logical data are transformed to physical metadata information to be used as a physical database table [25]-[29].

The last stage is to manage the mapping information of how a data object is related to a discrete information system. In this stage, a data object has the metadata of the logical and physical stages. There are two aspects in the management of mapping information. One is the relationship information with a newly developed system. The other is the transformation/matching information of the existing legacy system [26], [27].

We classify metadata into three categories: logical metadata information, physical metadata information, and mapping metadata information [29].

Logical metadata information includes data element (attribute), primary word (entity), and data model. The logical attribute information includes naming rules for terms, attributing primary words, and the relationship between data factors and primary words [28], [29].

Management of physical metadata information includes table name, column name, schema, and so on [28].

Management of mapping metadata information includes database names, table names, and transformation rules [28]-[30].

With these three categories, data structure quality can be synchronized and managed through the whole organization.

4. Extensive Definition of Data Quality Architecture

In the previous chapters, we defined quality domains, data quality [15], [23], [24], [31]-[40], and metadata architecture [25]-[30]. Figure 3 shows the extended data quality architecture. The definition of data quality should be extended from an independent point of view to a corporation's integrated data quality [15], [32], [33]. This definition could reflect the depth and width of the data. Data structure quality gives rise to data status quality such as data value and service quality. Hence, managed and evaluated data structure quality brings up management and maintenance of status quality.

In particular, data structure quality should be managed and evaluated to maintain the depth of the whole entrepreneurial point of view.

Data structure quality should be extended from each information system level to that of an enterprise level.

Therefore, metadata quality should be maintained through organizational integration along with the data standard and data architecture.

Data management process quality is basically the quality of the data structure management. The data management process will improve data status quality such as data value and service quality.

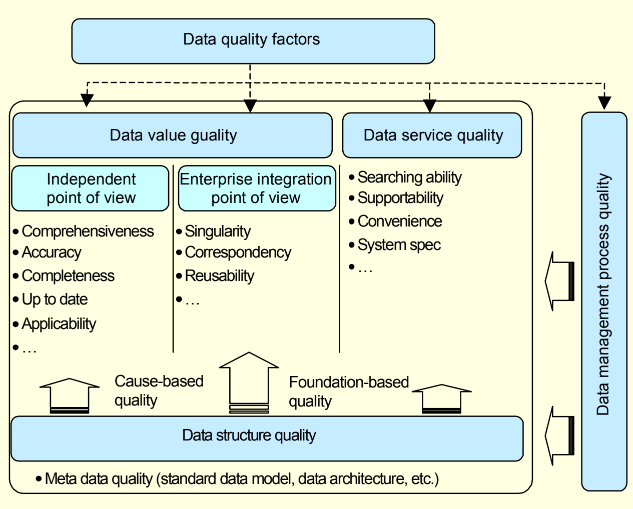


Fig. 3. Extended data quality architecture model.

To enhance total data quality, the data structure quality level should be raised and therefore data management process quality should be upgraded.

Data structure management factors should be derived according to each maturity process. Enterprises can evaluate their data quality management level and upgrade data quality to a more advanced level through the definition of maturity processes.

Therefore, this definition of maturity can overcome the shortcomings of data quality evaluation.

III. Data Quality Management Maturity Model

1. Definition

In the previous chapter, we showed that data quality evaluation and management should be defined from many points of view such as total corporate integration management [15], [23], [24], [31]-[40], data structure quality management, and management maturity stages.

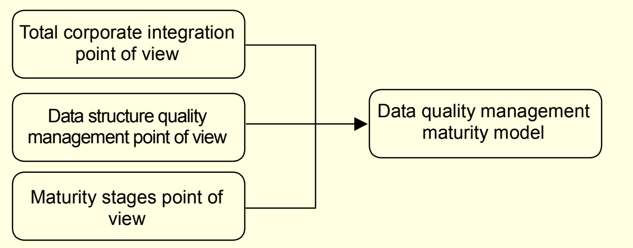

Total corporate integration point of view
Data structure quality management point of view
Maturity stages point of view
Data quality management maturity model

Fig. 4. Three viewpoints reflected in the data quality management maturity model.

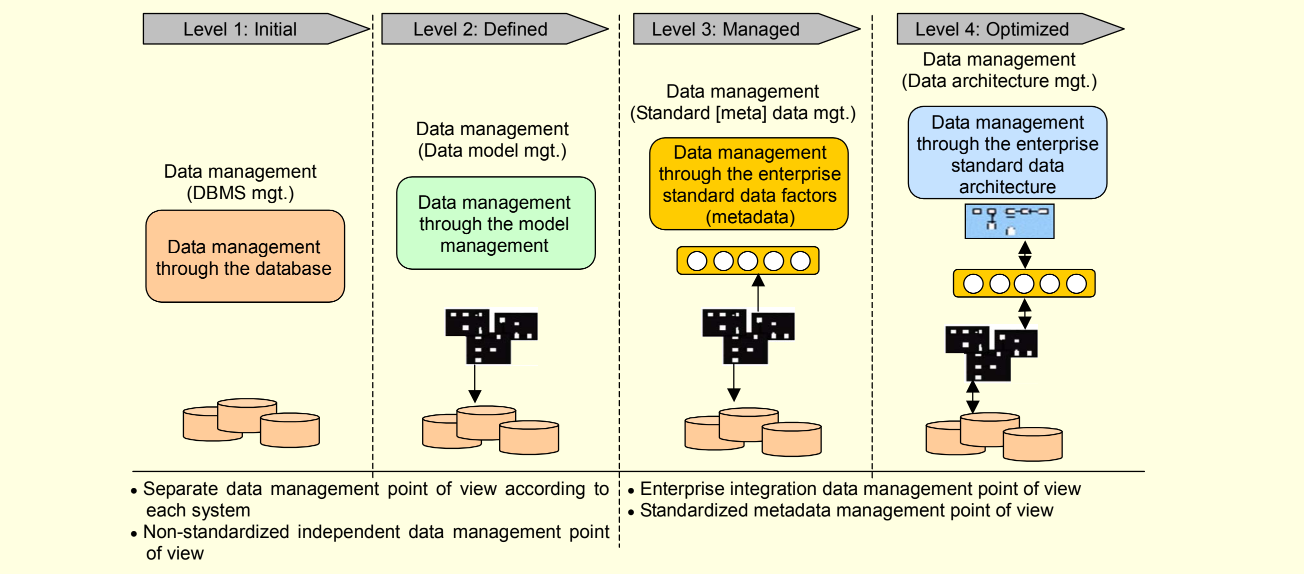

Level 1: Initial
Level 2: Defined
Level 3: Managed
Level 4: Optimized
Data management (DBMS mgt.)
Data management through the database
Data management (Data model mgt.)
Data management through the model management
Data management (Standard [meta] data mgt.)
Data management through the enterprise standard data factors (metadata)
Data management (Data architecture mgt.)
Data management through the enterprise standard data architecture
• Separate data management point of view according to each system
• Non-standardized independent data management point of view
• Enterprise integration data management point of view
• Standardized metadata management point of view

Fig. 5. Data quality management maturity model.

Here, we introduce the data quality management maturity model that reflects the above points of view as shown in Fig. 4. To reach full corporate integration, it is necessary for the whole organization to manage the metadata of the standardized data and data architecture [25]-[30].

Figure 5 shows the data management maturity model that is used as the capability maturity model of software process evaluation [17]-[21].

Level 1. The initial data management stage is the early management stage of data structure quality through the rules defined in the database system catalogue.

Level 2. Defined data management is the data management stage through the logical data model and physical data model. This is the stage defining the data of database design and management through the logical and physical data model design. When the data structure is modified or remodeled, it should refer to the data model. The modification is to be reflected in the data model and finally returned as a new input to the database.

Level 3. Managed data management is data management through data standardization. From this stage on, data standardization is going on under the enterprise integration, which is different from Level 1 and Level 2. This stage is the management of metadata that selects all corporate data and standardizes various attributes, schema, domain, data model, and so on. Level 3 enables sharing and reuses of standardized data through standardization of the metadata. It also integrates the base information system units.

Level 4. Optimized data management is data management through data architecture management. This stage is to define the enterprise standard architecture model, which is the optimized data management stage to manage the data, data model, and data relationship on the basis of the defined enterprise standard architecture model.

Figure 6 shows the data quality achievement domain through the data quality management maturity stages. Because data quality management is running under the separate information system at Level 1 and Level 2, these levels satisfy the depth of independent data quality. We can say data quality management is better at Level 2 than Level 1 because the former maintains the logical data model, while the latter simply manages the data by the physical system catalogue of a database management system.

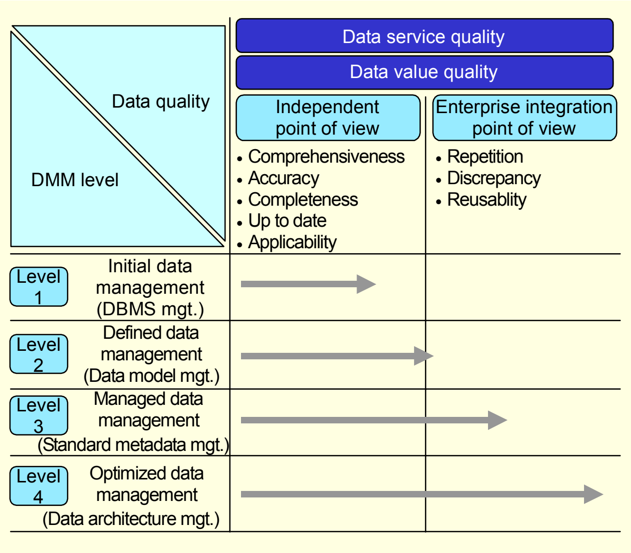

Data quality
DMM level
Data service quality
Data value quality
Independent point of view
Enterprise integration point of view
• Comprehensiveness
• Accuracy
• Completeness
• Up to date
• Applicability
• Repetition
• Discrepancy
• Reusablity
Level 1
Initial data management (DBMS mgt.)
Level 2
Defined data management (Data model mgt.)
Level 3
Managed data management (Standard metadata mgt.)
Level 4
Optimized data management (Data architecture mgt.)

Fig. 6. Data quality achievement domain according to the maturity stages.

At Level 1, it is hard to recognize the essential data structure for physical performance, while Level 2 manages the database based on the logical database model so data can be prevented from deforming although data is added or changed.

Level 3 and Level 4 satisfy the enterprise integrated data quality as well as independent data quality, which is possible because data management is based on sharing and reusing data through the whole corporate data standardization at these levels.

Enterprise data quality management is more intensive at Level 4 than at Level 3 because the former includes the data architecture while the latter only standardizes the data.

At Level 3, standardization of isolated data might occur if the data architecture concept is lacking.

This isolated standardization will be discussed in the following section.

2. Issues and Solutions at Each Maturity Stage

In this section, we explain the issues at each maturity stage of the proposed data quality management model. Then, we suggest some solutions to the issues.

A. Issues and Solutions of Data Quality Management through a Database Management System at Level 1

Level 1 is the early management stage to define data structure quality through the database system catalogue and to manage the data as a physical table. Thus, data is maintained by the definition of physical attributes and the reference integrity of tables.

However, there is a possibility of de-normalization of data as a table. That is to say, the data might be explained and integrated according to various physical database systems.

Data are objects explaining the corporate business, separate from the physical system. However, the data might be modified or distorted during physical system processing. In the early stage of system build-up, the data can easily be recognized by administrators.

But as time goes on, the conceptual data may lose its original features because the modified/distorted physical table is managed instead of the original conceptual model.

Second, there is a possibility a company could lose its business rules, which might become worse as time goes on. This can happen even when the business is logical because management is based on a physical system.

To solve these issues, the database should be generated and managed through both the physical data model and logical data model at the same time. The logical data model explains the corporate business conceptually, while the physical data model systemizes the logical data.

We suggest that combined management of logical data and physical data models. Through this combination, data deformation can be protected in the case of addition, integration, and modification. It also helps the data sustain a certain level of quality.

B. Issues and Solutions of Data Quality Management through Data Model at Level 2

Level 2 is the data management stage through logical data and physical data models. At this stage, corporate business objects are explained as a logical data model. Later, these objects are transformed to a physical data model with respect to the physical system, database, and program language. A new database is designed and built up according to this physical data model. Combined management of the logical and physical models enables one to trace the original feature of the modified/distorted physical table. It also prevents the system from deforming as it refers to the logical data model when there is any addition, modification, or business rule change.

However, enterprise integration is still weak at this stage because data has been separately designed and built up by the unit information system. It has difficulty in data relationship and integration with other organizations, departments, and information systems.

Also, there is no corporate consistency in the rule name due to the lack of a standardized conventional name. That may result in different names for the same data with respect to the organizations/departments, tasks, and systems. This confusion will occur in the definition of the same data, value domain, data type, and representation styles.

In summary, due to a lack of enterprise mechanism of acquisition, storage, and search capability, it is impossible to reuse and share the data. This phenomenon is worse at Level 1 than at Level 2.

To solve these issues, it is necessary to select the data to be standardized and to standardize the data attributes. Among those attributes are data definition, value domain, data types, and standardization methods. That is to say, essential common data should be centrally standardized for the whole corporation. After standardization, the whole company should share and make use of the integrated and related data.

C. Issues and Solutions of Data Quality Management through Metadata at Level 3

Level 3 is data management through enterprise data standardization. At this stage, data standardization is going on under enterprise integration. The standardization of metadata is defined and centrally controlled. Each department develops new systems using data that have been standardized when the new system was developed. Existing legacy systems use standardized data thorough migration or mapping.

This enterprise data standardization enables data integration of each department and its reuse. Data quality can be maintained and improved by technically and functionally standardized data and the standardization process. But some problems still exist, which are to be solved at the Level 4 data architecture management level.

1) Issues on Isolated Standardization

Most standardization uses the bottom-up style, which infers standard data factors from physical schema. For example, data are extracted from table specifications, the system catalogue of the database, or a data model.

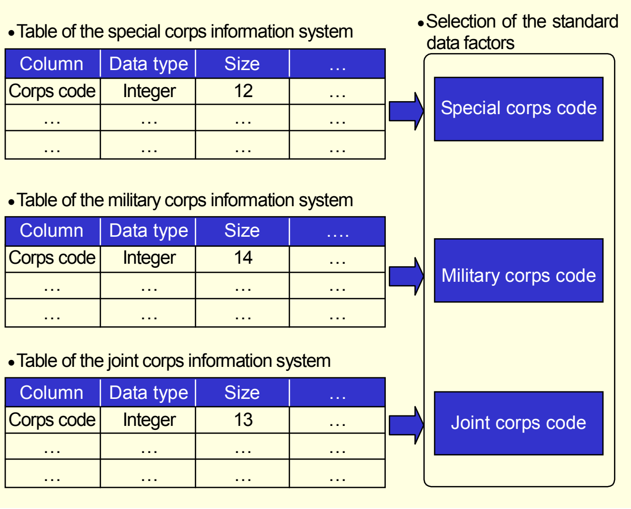
This bottom-up standardization results in isolated data standardization due to it selecting the standard data factors from the physical schema (database, table specifications). So it lacks the logical and structural data definition from an enterprise point of view. At this level, data standardization may be achieved in unit information systems. However, this is not complete data sharing and data standardization from a whole enterprise point of view.

Figure 7 shows an example of selecting the data to be standardized from the viewpoint of the unit information system without considering the enterprise data system. A data factor (corps code) is separately standardized to the standard data factor (special corps code, military corps code, joint corps code) according to unit information systems.

However, a corps code is a data factor that has the data system shown in Fig. 8.

Without a whole enterprise analysis of the data system, standardization from a unit information system might cause isolated standard data to lose relationship and integration.

As shown from the above example, a parent entity (corps) is classified into various corps types, and a primary key (corps code) is succeeded to each corps type. If each code is separately used to standardize the code system, schema, and value domains without considering the whole data system, corporate data integration is impossible. This is caused by bottom-up standardization selection from unit systems that are lacking in top-down data system analysis.



• Table of the special corps information system

Column	Data type	Size	…
Corps code	Integer	12	…
…	…	…	…
…	…	…	…

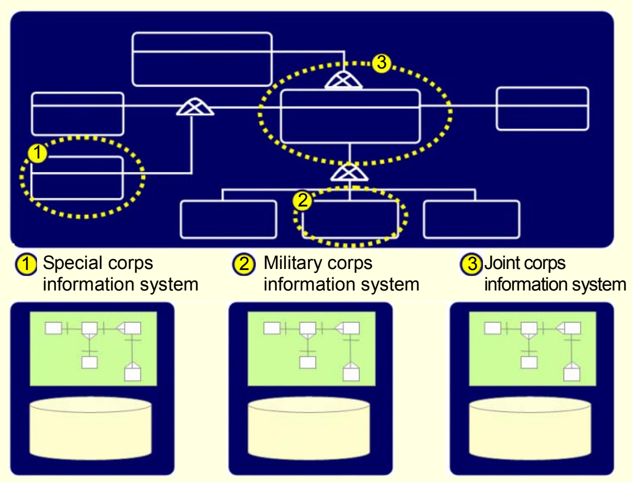
• Table of the military corps information system

Column	Data type	Size	….
Corps code	Integer	14	…
…	…	…	…
…	…	…	…

• Table of the joint corps information system

Column	Data type	Size	…
Corps code	Integer	13	…
…	…	…	…
…	…	…	…

Fig. 7. Standardization from a unit system view point.



1
2
3
1 Special corps information system
2 Military corps information system
3 Joint corps information system

Fig. 8. Data system of the corps code data.

To solve the isolated data standardization, it is necessary for a company to understand and manage the relationship between each unit's information system and whole corporate data architecture. With total data management, a company can achieve real data standardization, enabling it to share and integrate enterprise data.

2) Issues of Vague Identity of Standard Data

One of the problems in data standardization is to assign the entity to the standard data factor and to manage the relationship with unit information systems.

However, it is difficult to respond to certain questions such as, "which entity of the data model does include standard data factors," "if the standard data factor belongs to many entities, which is the source entity of the standard data factor," and "how is the standard data factor mapped to each unit's information systems?"

This could happen because bottom-up standard data factors lack the combined management ability of the standard data factors and data model.

An enterprise standard architecture analysis should be done to clearly specify the identity of standard data factors. This model helps to clarify the relationship between the standard data factors, data model, and unit information systems.

Figure 9 shows how the enterprise standard data architecture assigns the relationship between logical and physical data models. It also shows the clarification between standard data

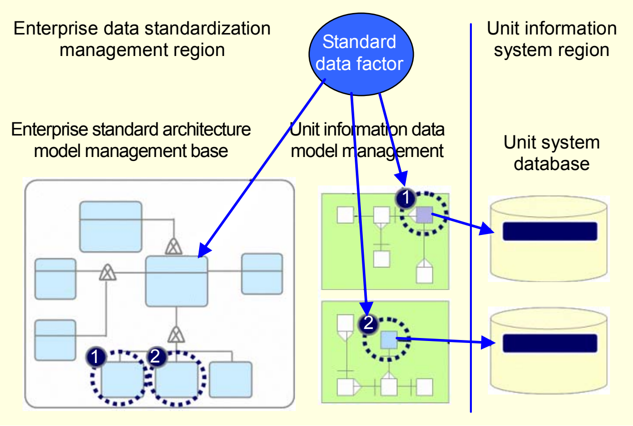

Enterprise data standardization management region
Standard data factor
Unit information system region
Enterprise standard architecture model management base
Unit information data model management
Unit system database
1
2
1
2

Fig. 9. Data factor relationship based on standard data architecture.

factors and the data model. It is very important to specify the exact relationship because it helps to mange the data change and to realize which standard data factor is used under which system and condition. However, bottom-up data standardization has many shortcomings in defining the relationship.

3) Administrator Vagueness of the Standard Data

One of the problems in data standardization is how to arrange the administrator for each item of data. It is difficult to determine who would make the decision for data standardization and who would maintain the database. Table 2 shows the classification of the data according to their sharing. As shown in Table 2, shared data and universal data are generated by various departments. Without a clear settlement of the generation/management tasks, it is very hard to standardize and manage data. This is because data standardization should be made from the data generator's point of view. Recalling the standardization problem of a corps code, there might be a dispute between the standardization schemes. One is to follow a particular corps code; the other is to accept every corps' standardization schemes.

Table 2. Classification of the data according to the sharing.

Category	Details
Data that are common to all departments	*Shared data* : generated by a particular department, but shared and managed by various departments
	Universal data : generated by uncertain departments, referred by whole organizations as their universality
Data that appear to a single department	*Unique data* : generated and managed by a particular department, unique within the organization

There are two solutions to the problem of administrator vagueness of data generation and management. First, data factors should be structurally identified on the basis of standard data architecture. Systematic data analysis helps to determine the original source of the data, and confirm the viewpoint of data classification.

Second, a functional analysis of data should be done to ascertain the administrator of data generation and management. For example, a data flow diagram should be used to recognize which data are generated and maintained through the external agent. That is to say, two aspects are needed at the same time to clarify the subject of data standardization through the recognition of the data generation/management administrator, structural analysis of a data system, and functional analysis of data generation and flow.

4) Difficulties in Change Management of Standard Data

A corporation constantly needs a data management strategy for new business rules set to new goals and strategy under a new environment. However, the existing data standardization stage has difficulty in the change management of standard data. This shortcoming has been brought about by the absence of standard data architecture management.

Figure 10 shows the difficulties in change management of

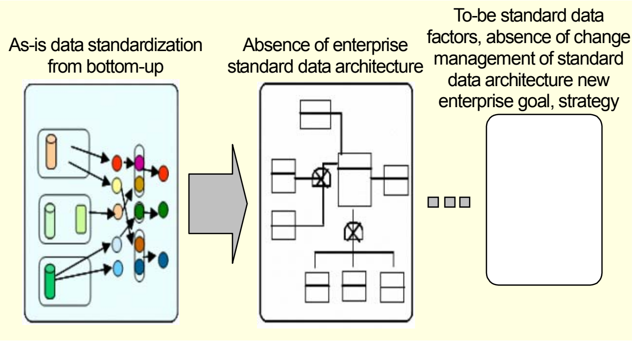

As-is data standardization from bottom-up
Absence of enterprise standard data architecture
To-be standard data factors, absence of change management of standard data architecture new enterprise goal, strategy

Fig. 10. Difficulties in change management of standard data.

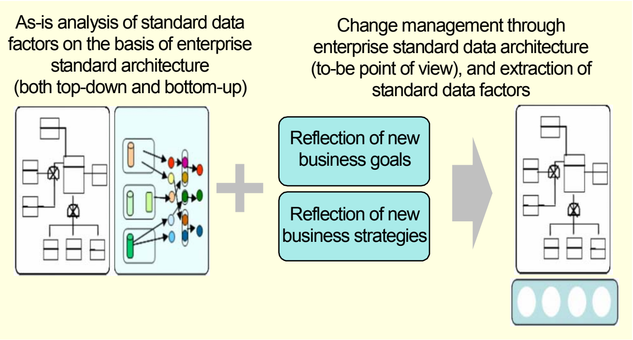

As-is analysis of standard data factors on the basis of enterprise standard architecture (both top-down and bottom-up)
Change management through enterprise standard data architecture (to-be point of view), and extraction of standard data factors
Reflection of new business goals
Reflection of new business strategies

Fig. 11. Change management on the basis of the enterprise standard data architecture.

standard data. Change management standardizes only the bottom-up phenomenal data, and because the enterprise system image is not managed, it is difficult to manage the changing data environment.

Figure 11 shows data standardization on the basis of the enterprise standard data architecture. Change management of standard data should be done for a corporation accepting a new business strategy based on the enterprise standard data architecture. With management of the enterprise standard data architecture, companies can manipulate the proper corporate change in the case of data change.

D. Switching Strategy to Level 4

True improvement of data structure quality through standard data factor management can be achieved on the basis of the enterprise standard data architecture. Standard factor management should be concerned with and maintain the relationship between the enterprise standard data architecture model, standard data factors, and data model of unit information systems.

At Level 4 management of the standard data architecture, data structure factors are systematically managed. That is, Level 4 could solve the issues of data standardization raised at Level 3.

To set the enterprise data architecture, it is necessary to apply two methods of analysis. One is the existing bottom-up analysis on the basis of reverse engineering. The other is the top-down analysis based on enterprise standard data. Additionally, a functional data analysis should be done to identify the administrator of data generation/management. Next, the architecture should be expanded to the planned standard data architecture through analysis of new business goals and strategies.

The top-down and bottom-up approaches for data management should be considered at the same time to achieve the standardization and enterprise standard data architecture. With the bottom-up approach, we can collect data factors to be standardized while a data structural analysis should be done to the collected data factors for the top-down approach.

These two-way approaches should be done not only with data structural analysis but also with a functional analysis. As described earlier, it is difficult to standardize and manage data without a clear specification of the data generation/management administrator. Companies should select the standard data system and data factors through those two approaches from structural/functional points of view. They should expand their data management systems to the planned data management systems that fit the new business environment.

IV. Empirical Study of Data Quality Management Maturity Model

1. Empirical Study Model and Method

In the previous section, we proposed the data quality management maturity model as a preferred model. Issues at each stage were resolved by shifting to higher stages. We verified the data quality management maturity model by theory and case study.

In this section, the data quality management maturity model will be proved empirically using the study model. Figure 12 shows the empirical study model that has been used to verify the model.

We measured the degree of data quality change for companies belonging to each stage of data management. We empirically checked whether data quality becomes higher as the data management stage is raised. For this study, we have done two stage surveys to assess the data management level and data quality change.

As shown in Table 3, we determined the data management maturity level of each company (six companies in finance and public fields) by interviews with their CIOs. The interviews consisted of questions asking whether a company had performed the essential items at each maturity stage. As a result of the first interview, we classified the companies into two groups: three companies in Group A (Level 1 -> Level 2 at present), and three companies in Group B (Level 2 -> Level 3 at present).

Most of the domestic companies are at the Level 1 or Level 2 stage. Some conglomerate companies are shifting to the Level 3 stage. However, very small numbers of domestic companies are at the stage of Level 4, and therefore we could not include those companies which had switched from Level 3 to Level 4.

The second survey was with employees of the IT department asking about data quality change with respect to maturity change. The employees of Group A were asked to compare the data quality before and after data model management,

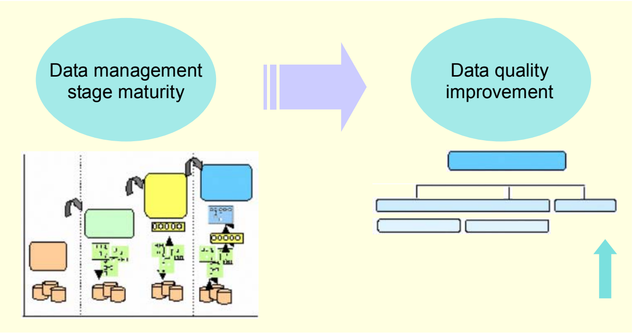


Fig. 12. Empirical study model for the data quality management maturity model.

Table 3. Summary of the survey.

Category	Group A: companies at maturity level 2 (shifted from maturity level 1 to maturity level 2)	Group B: companies at maturity level 3 (shifted from maturity level 2 to maturity level 3)
Number of companies	3	3
Number of respondents	46	73
Survey method	Criteria: data management through the data model management Survey: the data quality level before and after the criteria	Criteria: data management through the metadata management Survey: the data quality level before and after the criteria
Measurement method	5-point measurement	
Number of questions	Response to 16 questions before and after the criteria	

while the employees of Group B were asked to compare the data quality before and after metadata (standard data) management. The questionnaire was composed of 16 questions asking about data quality regardless of the company's maturity stage. The total number of respondents was 119 persons, 46 of whom came from Group A (three companies at Level 2), and the other 73 came from Group B (three companies at Level 3). With the data collected from the survey, we completed a reliability test, factor analysis, and paired sample t-test.

The reliability test verified the reliability of the survey response.

The factor analysis classified the questions asking about data quality into elements so we could test whether the questionnaire is reasonable and then categorize the data quality factors.

The paired sample t-test proves the difference of data quality with respect to maturity levels. We operated the paired sample t-test to compare the mean of the data quality because we wanted to differentiate the quality before and after the criteria. That is, the questions of the survey were made to contrast the data quality at the present data management level compared with the previous level.

There are three points to compare the mean of data quality according to each data management level. The first is to compare the total mean for 16 questions regarding data quality before and after the criteria. That is, we proved the reasonability of our data quality management maturity model by checking whether there had been any improvement of total data quality in Group A (Level 1 -> Level 2) and Group B (Level 2 -> Level 3).

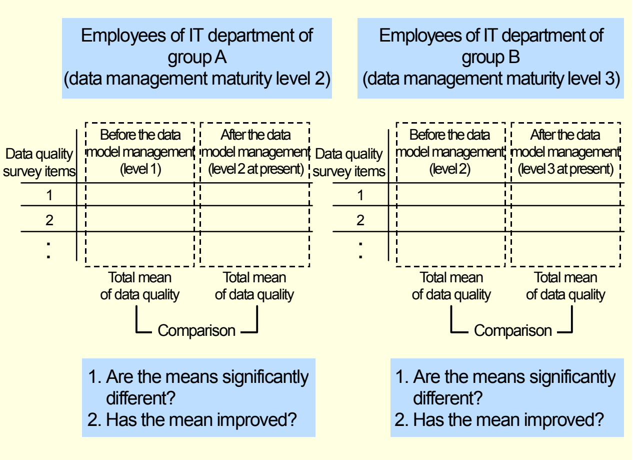

Employees of IT department of group A (data management maturity level 2)
Employees of IT department of group B (data management maturity level 3)
Data quality survey items
Before the data model management (level 1)
After the data model management (level 2 at present)
1
2
:
Total mean of data quality
Total mean of data quality
Comparison
1. Are the means significantly different?
2. Has the mean improved?
Data quality survey items
Before the data model management (level 2)
After the data model management (level 3 at present)
1
2
:
Total mean of data quality
Total mean of data quality
Comparison
1. Are the means significantly different?
2. Has the mean improved?

Fig. 13. Comparison of data quality improvement according to each maturity level.

The second point is to compare the mean of data quality regarding the factors before and after criteria through the factor analysis. That is, by checking whether there had been any improvement of data quality factors in Group A (Level 1 -> Level 2) and Group B (Level 2 -> Level 3), we proved our suggestion that the progress of the data management level brings improvement to every factor.

The third point is to compare the difference in data quality level improvement regarding enterprise integration and the relationships between Group A and Group B. As stated earlier, in the data quality management maturity model, companies at Level 3 or higher might have greater data quality improvement. So we juxtaposed the weight of data quality improvement from the point of enterprise integration at Group A (Level 1 -> Level 2) and Group B (Level 2 -> Level 3). With this comparison, we checked the diffusion effect of data quality improvement that had been proposed by our data quality management maturity model.

2. Results of the Empirical Study

A. Results of the Reliability Test for the Survey Questionnaire

As a result of the reliability test, we have Chronbach's Alpha of 0.918 (before the given criteria) and 0.873 (after the given criteria). These results show that survey items are reliable for data quality.

B. Results of the Factor Analysis for the Survey Questionnaire

We used the factor analysis to extract the 16 items of data quality measurement terms. Next, these items were used to compare data quality before and after the given criteria. Principal component analysis was used to extract the factors that had Eigen values higher than 1. The extracted factors were systemized using the Equamax method to clarify the extracted

Table 4. Results of the factor analysis.

Summary of the questionnaire	Factor 1	Factor 2	Factor 3	Factor 4
	Enterprise integration quality	Accuracy quality	Expression quality	Service quality
Inter-information systems consistency of data format	0.843			
Inter-information systems consistency of code value domain	0.739			
Inter-information systems consistency of data naming	0.695			
Inter-information systems consistency of naming rules	0.628			
Inter-information systems non-overlap of the data	0.598			
Omission of the data value or not		0.787		
Omission of the data value items or not		0.716		
Reliability of the data accuracy		0.555		
Erratum of data or not		0.549		
Objectiveness of the data value		0.511		
Adequacy of the data expression			0.819	
Identity of data expression			0.744	
Suitability of data summary expression			0.502	
Suitability of data authentication and security				0.787
Up to date data				0.677
Convenience of data search				0.570

pattern.

Table 4 shows the results of the rotated factor patterns. We named the factors as shown in the table, which referred to the common concepts of data quality factors from our questionnaire and from previous studies. That is to say, factor 1 is related to the enterprise integration quality, factor 2 related to data accuracy quality, factor 3 to data expression quality, and factor 4 to data service quality.

C. Analysis Results of the Data Quality Difference for Each Maturity Level

The paired sample test was used to verify empirically the data quality management maturity model proposed in our study. We wanted to check our hypothesis of whether data quality level improves as the data management level matures.

Thus, the employees of the IT departments of six companies (three in Group A at Level 2, three in Group B at Level 3) were asked to assess the present data quality of their companies compared with the previous data quality. The criteria dividing the present and previous states were given in the following manner: data management through the data model for Group A companies (Level 1 -> Level 2 at present), and data management through metadata management for Group B companies (Level 2 -> Level 3 at present).

Hypothesis 1. Data quality level improves as data quality management level matures.

- Hypothesis 1-1: Total data quality level improves as data quality management matures.
- Hypothesis 1-2: Enterprise integration data quality improves as data quality management matures.
- Hypothesis 1-3: Data accuracy quality improves as data quality management matures.
- Hypothesis 1-4: Data expression quality improves as data quality management matures.
- Hypothesis 1-5: Data service quality improves as data quality management matures.

Based on responses to the survey, we performed the paired sample test for the hypothesis 1. The reason we applied the paired sample test to check the difference of mean is that we wanted to verify whether the data quality is significantly different before and after the given criteria.

The total data quality of hypothesis 1-1 means the total mean of 16 questions for data quality. All the data quality terms used at hypotheses from 1-2 to 1-5 are four factors drawn from the factor analysis of the 16 questions.

Table 5 shows that the null hypothesis could be rejected at $\alpha = 0.05$ because there is no difference in the mean of Level 1 and Level 2. Since the P-value was 0.000 with the mean of

Table 5. Result of the paired sample test.

		Mean	t	Freedom	Significant statistics (both ends)
Companies at data management maturity level 2 (data management through data model)	Level 1: Total data quality Level 2: Total data quality	-1.00	-12.0	45	0.000
	Level 1: Enterprise integration data quality Level 2: Enterprise integration data quality	-1.11	-10.7	45	0.000
	Level 1: Data accuracy quality Level 2: Data accuracy quality	-0.90	-10.0	45	0.000
	Level 1: Data expression quality Level 2: Data expression quality	-0.97	-8.6	45	0.000
	Level 1: Data service quality Level 2: Data service quality	-1.02	-10.2	45	0.000
Companies at data management maturity level 3 (data management through meta data management)	Level 2: Total data quality Level 3: Total data quality	-1.48	-19.0	72	0.000
	Level 2: Enterprise integration data quality Level 3: Enterprise integration data quality	-1.70	-18.0	72	0.000
	Level 2: Data accuracy quality Level 3: Data accuracy quality	-1.26	-13.5	72	0.000
	Level 2: Data expression quality Level 3: Data expression quality	-1.56	-17.6	72	0.000
	Level 2: Data service quality Level 3: Data service quality	-1.40	-14.2	72	0.000

total data quality of Level 1 and Level 2, we can say that there is a difference in the total data quality of Level 1 and Level 2.

Table 6 shows that the mean of total data quality at Level 1 is 2.59, while Level 2 has 3.60 as its mean. This implies that total data quality improves as the data quality matures from Level 1 to Level 2.

Likewise, the null hypothesis can be rejected at α = 0.05, because there is no difference in the mean of Level 2 and Level 3. Since the P-value was 0.000 with the mean of total data quality of Level 2 and Level 3, we can say that there is a difference in the total data quality of Level 1 and Level 2. The mean of total data quality at Level 2 is 2.54, while Level 3 has 4.02. This implies that total data quality improves as the data quality matures from Level 2 to Level 3.

From the above result, hypothesis 1-1, it was accepted that total data quality level improves as data quality management matures. For the other hypotheses from 1-2 to 1-5, all of the P-values (mean of quality difference between Level 1 and Level 2 / Level 2 and Level 3) were 0.000.

This implies that all of the data quality factors (entrepreneurial integration data quality, data accuracy quality, data expression quality, and data service quality) are different with respect to data quality management maturity.

Table 6 shows that every mean of the data quality factors at the upper level is higher than at the lower level.

On the basis of those two results shown in Table 5 and Table 6, we can accept every hypothesis, 1-2, 1-3, 1-4, and 1-5.

Hypothesis 2. The growth of enterprise integration data quality is greater for Level 2 -> Level 3 than Level 1 -> Level 2.

Our data quality management maturity model argues that enterprise integration data quality considerably improves from Level 2 to Level 3. The paired sample test of Level 1 -> Level 2 shows that the degree of improvement for enterprise integration data quality is 1.11. The degree of improvement for data service quality is 0.97 (the second highest, 0.04 lower than enterprise integration data quality), and that of data accuracy quality is 0.90 (the minimum, 0.2 lower than enterprise integration data quality).

On the other hand, for Level 2 -> Level 3, the degree of improvement for enterprise integration data quality is 1.70. The degree of improvement for data expression quality is 1.56 (the second highest, 0.14 lower than the enterprise integration data quality), and that of data accuracy quality is 1.26 (the minimum, 0.44 lower than the entrepreneurial integration data quality). As a result, we can accept Hypothesis 2 and say that the growth of enterprise integration data quality is greater for Level 2 -> Level 3 than Level 1 -> Level 2 compared to other quality factors.

Table 6. Paired sample statistics.

			Mean	Standard deviation	N	Mean of standard error
Companies at data management maturity level 2 (data management through data model)	Total data quality	Level 1	2.59	0.46	46	0.067
		Level 2	3.60	0.37	46	0.054
	Enterprise integration data quality	Level 1	2.38	0.47	46	0.069
		Level 2	3.49	0.55	46	0.081
	Data accuracy quality	Level 1	2.74	0.58	46	0.085
		Level 2	3.65	0.45	46	0.066
	Data expression quality	Level 1	2.62	0.67	46	0.099
		Level 2	3.62	0.46	46	0.068
	Data service quality	Level 1	2.63	0.59	46	0.087
		Level 2	3.66	0.49	46	0.072
Companies at data management maturity level 3 (data management through meta data management)	Total data quality	Level 2	2.54	0.60	73	0.070
		Level 3	4.02	0.35	73	0.040
	Enterprise integration data quality	Level 2	2.32	0.63	73	0.073
		Level 3	4.02	0.51	73	0.060
	Data accuracy quality	Level 2	2.69	0.76	73	0.089
		Level 3	3.95	0.41	73	0.048
	Data expression quality	Level 2	2.51	0.70	73	0.082
		Level 3	4.07	0.46	73	0.054
	Data service quality	Level 2	2.67	0.69	73	0.080
		Level 3	4.07	0.45	73	0.052

With the above empirical study, we verified the effectiveness of the proposed data quality management maturity model.

V. Conclusions

In this study, we proposed a data quality management maturity model that could be applied to evaluate and manage data quality for an enterprise.

Our model has the following implications. First, it shows the essential lists for companies planning to realize their present state of data quality management and to develop it to a higher stage, whereas previous studies have concentrated on the final evaluation.

Second, our study presents structural quality management stages through case studies. This sets the basis for a realistic evaluation of data quality management, compared to the theoretical method of existing phenomenal quality evaluation.

Third, issues and solutions for standardization stages are given to set data quality management through data standardization. This data standardization is going on with some domestic and conglomerate corporations and in the public field.

Fourth, our study introduces a futuristic viewpoint of data quality management.

Last, but most importantly, the meaning of our study is that the data quality management maturity model presents a macroscopic view of data quality management, which plays a major role for companies in recognizing their present state and in setting business goals for the next stage.

References

[1] Peter F. Drucker, “Playing in the Information-Based ‘Orchestra’,” *The Wall Street Journal*, June 4, 1985.

[2] L. English, *Improving Data Warehouse and Business Information Quality*, John Wiley & Sons, Inc., 1999.

[3] Amir Parssian, Sumit Sarkar, and Varghese S. Jacob, *Assessing Data Quality For Information Products*, School of Management University of Texas at Dallas, U.S.A, 1997.

[4] Ken Orr, “Data Quality and Systems,” *Communications of the ACM*, vol. 41, no. 2, Feb. 1998, pp. 66-71.

[5] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, “Data Quality Assessment,” *Communications of the ACM*, vol. 45, no. 4, Apr. 2002, pp. 211-218.

[6] Diane M. Strong, Yang W. Lee, Richard Y. Wang, "Data Quality in Context," *Communications of the ACM*, vol. 40, no. 5, May 1997, pp. 103-110.

[7] John A. Hoxmeier, *Database Quality Dimensions*, Computer Information Systems Department, College of Business, Colorado State University, Fort Collins, *http://www.sbaer.uca.edu/Research/1999/WDSI/99wds578.htm.*

[8] John A. Hoxmeier, *A Framework for Assessing Database Quality*, Computer Information Systems Department, College of Business, Colorado State University, Fort Collins.

[9] Giri Kumar Tayi and Donald P. Ballou, "Examining Data Quality," *Communications of the ACM*, vol. 41, no. 2, Feb. 1998, pp. 54-57.

[10] Beverly K. Kahn, Diane M. Strong, and Richard Y. Wing, "Information Quality Benchmark," *Communications of the ACM*, vol. 45, no. 4, Feb. 2002, pp. 184-192.

[11] B. Kahn, D. Strong, and R. Wang, "Information Quality Benchmarks: Product and Service Performance," *Communications of the ACM*, vol. 45, Apr. 2003, pp. 184-192.

[12] Yair Wand and Richard Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, vol. 39, 1996, pp. 86-69.

[13] Richard Y. Wang, "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, vol. 41, no. 2, 1998, pp. 58-65.

[14] Yair Wand and Richard Y. Wang, "Anchoring, Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, vol. 39, no. 11, 1996, pp. 90-95.

[15] G. Shankaranarayan, M. Ziad, and R. Wang, "Managing Data Quality in Dynamic Decision Environments : An Information Product Approach," *J. Database Management*, vol. 14, no. 4, 2003, pp. 14-32.

[16] Storey and Wang, "A Framework for the Analysis of Data Quality Research," *IEEE Trans. on Knowledge and Data Engineering*, vol. 7, no. 4, 1995, pp. 623-640.

[17] M. Paulk, "Using the Software CMM with Good Judgment," *ASQ Software Quality Professional*, vol. 1, no. 3, June 1999, pp. 19-29

[18] M. Paulk, "Practices of High Maturity Organizations," *SEPG Conference*, 1999, pp. 8-11.

[19] M. Paulk, C. Weber, B. Curtis, and M. Chrissis, *A High Maturity Example: Space Shuttle Onboard Software, in the Capability Maturity Model: Guidelines for Improving Software Process,* Addison-Wesley, 1994.

[20] James Herbseb, "Software Quality and the Capability Maturity Model," *Communications of ACM*, vol. 40, no. 6, June 1997, pp. 30-40.

[21] H. Saiedian and R. Kuzara, "SEI Capability Maturity Model's Impact on Contractors," *IEEE Computer*, vol. 28, Issue 1, Jan. 1995, pp. 16-26.

[22] P. Howard, *Data Quality Products: an Evaluation and Comparison*, Bloor Research, 2004.

[23] C. Hsu, G. Babin, M. Bouziane, W. Cheung, L. Rattner, and L. Yee, "Metadatabase Modeling for Enterprise Information Integration," *J. Systems Integration*, vol. 2, no. 1, Feb. 1992, pp. 5-37.

[24] S. Lujan-Mora and M. Palomar, "Reducing Inconsistency in Integrating Data from Different Sources," *Proc. 2001 Int'l Symp. Database Engineering & Applications*, 2001, pp. 209-218.

[25] Fred R. Mcfadden, Jeffrey A. Hoffer, and Mary B. Prescott, *Modern Database Management*, Addison Wesley Longman, 1999, pp. 6-7, 537-538, 559-560.

[26] James Martin, *Information Engineering, Book II-Planning and Analysis*, Prentice-Hall, 1989, pp. 147-50.

[27] Clive Finkelstein, *Information Engineering*, Addison-Wesley, 1992, pp. 598.

[28] David C. Hay, *Data Model Patterns*, Dorset House, 1996, pp. 65, 201, 226, 233, 254-55.

[29] G. Lawrence Sanders, *Data Modeling*, Boyd & Fraser, 1995, pp. 99, 103-104.

[30] Michael C. Reingruber and William W. Gregory, *Data Modeling Handbook-A Best-Practice Approach to Building Quality Data Models*, Wiley-QED,1994, pp. 293, 334.

[31] T. Redman, *Data Quality for the Information Age*, Artech House, 1996.

[32] M. Garcia-Solaco, F. Saltor, and M. Castellanos, "A Structure Based Schema Integration Methodology," *Proc. 11th Int'l Conf. Data Engineering*, 1995, pp. 505-512.

[33] J. Lee, K. Sian, and S. Hong, "Enterprise Integration with ERP and EAI," *Communications of the ACM*, vol. 46, 2003, pp. 54-60.

[34] T. Ksiezyk, G. Martin, and Qing Jia InfoSleuth, "Agent-Based System for Data Integration and Analysis," *Proc. 25th Annual Int'l Computer Software and Applications Conf. (COMPSAC)*, 2001, pp. 474-476.

[35] Phan. M. Dung, "Integrating Data from Possibly Inconsistent Databases," Proc. *Cooperative Information System*, 1996, pp. 58-65.

[36] M. Rajinikanth, G. Jakobson, C. Lafond, W. Papp, and G. PietetskyShapiro, "Multiple Database Integration in CALIDA: Design and Implementation," *Proc. First Int'l Conf. Systems Integration*, 1990, pp. 378-384.

[37] R. Elmasri and S.B. Navathe, *Fundamentals of Database Systems*, Addison-Wesley, 2000.

[38] J.D. Gascoigne Rui, A. Hodgson, and C.M. Sumpter, "Automating Information Transfer in Manufacturing Systems," *Computer-Aided Engineering Journal*, vol. 5, no. 3, 1988, pp. 113-121.

[39] E. Schallehn, K.U. Sattler, and G. Saake, "Extensible and Similaritybased Grouping for Data Integration," *Proc. Int. Conf. Data Engineering*, 2002, pp. 277.

[40] S. Jing, Z. Bin,W. Guoren, S. Baoyan, Y. Ge, and Z. Huaiyuan, “An Object-Wrapping Technique for Integrating Non-Traditional Database Systems,” *Proc. of the IEEE Int’l Conf. Intelligent Processing Systems*, vol. 2, 1997, pp. 1576-1580.

Kyung-Seok Ryu is a Researcher in the IT Services Research Division of Electronics and Telecommunications Research Institute (ETRI). He received the BA degree in business administration and the MA degree in management from Kyung Hee University, Seoul, Korea. He is mainly interested in business strategies, corporate strategies, marketing strategies, database modeling, balanced scorecard (BSC), customer relationship management (CRM), IT evaluation, strategic information systems, scenario analysis, and others.

Joo-Seok Park is a Professor of management information systems at the Business School of Kyung Hee University, Seoul, Korea. He received the BA degree in industrial engineering from Seoul National University, the MS degree in industrial engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea, and received the PhD degree in management information systems from UC Berkeley, CA, USA. He is mainly interested in database modeling, IT evaluation, strategic information systems, finance in mobile, CRM, and others.

Jae-Hong Park is an MS student in statistics at Stanford University, Palo Alto, USA. He received the BA degree in business administration from Kyung Hee University, Seoul, Korea and the MA degree in information system management from Carnegie Mellon University, Pittsburgh, USA. He is interested in the economics of information, e-commerce, pricing at e-commerce and internet, data mining, and others.