Integrating Environmental Information through Data Warehousing with Data Quality Assurance

Yu-Chi Chu[1], Chea-Yuan Young[1], and Chen-Chau Yang[2]

Abstract

This paper presents a comprehensive data warehouse framework for integrating environmental information. The proposed framework may serve as a platform and provide appropriate mechanisms for retrieving and gathering environmental data from various sources. For ensuring the quality of data in the environmental data warehouse, we introduce a four-phase process to ensure data quality while performing information integration. Overall data quality conditions can be identified and relevant information can be provided for determining whether the data meet "fit to use" criteria and whether they need to be improved. Accordingly, users may filter the data retrieved from the warehouse based upon various quality requirements.

1. Introduction

Environmental protection has become an important public issue throughout the world in recent years. For making sensible, justifiable, and legally correct decisions to protect our earth, both government agencies and private sectors need detailed information regarding the current state of the environment and ongoing developments (Kashyap 1999). However, environment information systems are often created parallel and independently. This inevitably leads to very heterogeneous structures and content of these systems. Currently it is very difficult to share environmental data since the information typically resides on geographically disparate and heterogeneous databases (systems). These systems often do not facilitate access by secondary users and frustrate attempts to draw data together to form a more comprehensive understanding of environmental conditions and actions. Therefore, there is a major demand for appropriate systems and adequate tools to provide integrated information for managing the issues of environmental protection.

There are a number of challenges that come up when trying to integrate environmental information from the various sources. We illustrate these challenges by using the following example.

Example 1: We consider that Environmental Protection Agency (EPA) has a number

[1] Bureau of Environmental Monitoring and Data Processing, Environmental Protection Administration, Taipei, Taiwan, Republic of China. e-mail: {ycchu, cyyoung}@sun.epa.gov.tw

[2] Dept. of Electronic Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, Republic of China. e-mail: ccyang@et.ntust.edu.tw

of offices. For the specific environmental interest such as air quality, waste management and so on, each office might maintain a database of the potential pollution site (factories, hospitals, etc.) EPA wants to create an integrated information system containing the data of all offices. The system will help EPA's users to locate the environmental risk from an integrated viewpoint. It also can be used by corporate experts to assess environmental problem and support decision making.

However, the offices in EPA do not all use the same database schema. For example, one office might store factory in the relations shown as follows:

```
        Factory(facId, name, address, ...)
        Permit(facId, permitNo, description, ...)
```

while another office might use the schema that looks like:

```
        Plant(serialNo, plant-name, plant-location,...)
        OpRecords(recordNo, serialNo, date, status,...)
```

Based on the observation on the above example, we may divide the challenges of environmental information integration in three aspects:

1. **Autonomy**. Each information source maintains data in various format and coverage because most of the sources are intended to preserve their autonomy. Therefore, some sources may contain good latitude/longitude data to represent location information, while others may contain only unformatted address information.
2. **Heterogeneity.** Heterogeneity may occur at various levels and for various reasons. Different sources may use different names for the same kinds of entities; even worse, they may use the same names for different kinds of entities. We notice in Example 1 that not only the structure of the schema is different, but the names of schema and the names of attributes.
3. **Data quality.** One of the most important tasks of information integration is the selection of good data sources. Clearly, information sources in Example 1 store data of varying since they are implemented by difference offices at different time. Hence, the result of the integration process is directly influenced by data quality.

To address these three issues, standards communities, professional societies, as well as industry associations have developed standards to facilitate sharing and integration. However, the standards themselves are part of the problem. The first requirement is to develop standards for relating standards (Sowa 2000).

2. Warehousing Environmental Data

In recent years, data warehouse systems have attracted a great deal of interest in both academic and industrial communities. In the typical data warehouse architecture, the data subject to analysis is integrated from multiple sources; both internal and external, and selected information is extracted in advance and stored in a repository. Generally, the information stored in the warehouse can be structured and organized in a form that makes it easy to use for applications. A data warehouse can therefore be seen as a set of *materialized views* defined over the remote sources, and warehoused data is usually used for decision making, rather than for operations. The data warehouse can be dedicated to supporting a large subset of the environmental management functions that are largely oriented toward long-term record keeping, data aggregation and summarization, and information dissemination (Ponniah 2001).

Some of organizations have been employing data warehousing approach to integrate environmental data. For example, the Environmental Protection Agency of the United States (USEPA) has developed a data warehouse known as Envirofacts (Envirofacts). It combines data from a number of the USEPA's databases. The Envirofacts data warehouse can map sites that handle potentially dangerous chemicals and identify their harmful substances. The data stored in Envirofacts are automatically updated monthly or quarterly from the isolated databases. USEPA are currently providing direct SQL access to data stored in Envirofacts via web pages. The database schema and appropriate connection information of Envirofacts are also provided in web pages. The system not only made it quicker for many people to procure the data they needed, but it also greatly reduced the workload for the USEPA's information staff. Following are some examples of how Envirofacts is being used: (Laudon 2000, Fabris 1998)

- An environmental lawyer can access Envirofacts to keep track of clients or potential clients. For instance, she could look up a metal fabricating plant and find out what kinds solvents it uses and the compliance issues it faces.
- A newspaper reporter can use Envirofacts to research an article about the level of carcinogenic chemical emissions in his community.
- EPA personnel from the various offices can share data more easily. For example, water specialists might use a web browser to analyze how a municipal dump contributed to the pollution of a nearby river.

However, there are some problems that should be addressed in data warehousing (Ponniah 2001, Kimball 2002). If the data in the warehouse do not meet quality characteristics required to support decisions, the data warehouse effort will be blamed for the shortcomings. Poor-quality data will lead either to wrong decisions being made, or knowledge workers losing confidence in the data warehouse. Although some companies recently have become aware of the importance of high-quality data and some straightforward approaches have been proposed for data quality management, there is a definite need for comprehensive approaches to improve data quality.

Figure 1 gives an overview of the framework we proposed for integrating environmental information. In general, the framework can be divided three aspects: (1) new data model, (2) data extraction, transformation, integration, and loading}, and (3) data quality management. In this paper, we focus on the third aspect, i.e. we will investigate how to ensure data quality when performing information integration. The research topics will involve the mechanism for data quality assurance and the methods for analyzing data quality in data warehouse environments. Furthermore, two technologies need to be involved in proposed framework: efficient data loading and incremental update maintenance. When the data is imported to a data warehouse or replicated in a server, its contents must be maintained when sources change in order to preserve its consistency with respect to the base data. Concerning the issues of maintenance, we adopt the following approaches that have been proposed in (Garcia-Molina 2000) depending on circumstances.

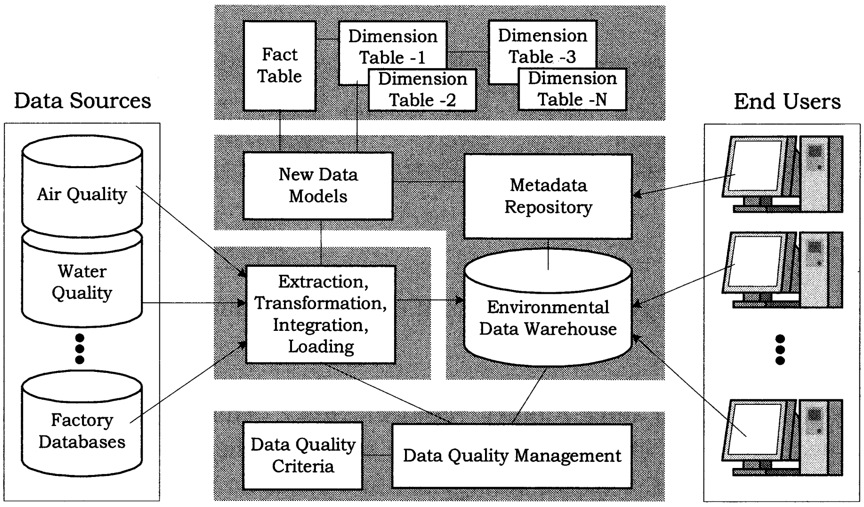

Figure 1: The framework of environmental data warehouse

1. The warehouse is periodically and entirely reconstructed from the current data in the sources. The main drawback of this approach is the requirement of shutting down the warehouse so queries cannot be issued while the warehouse is being reconstructed. For environmental protection applications, this approach might cause that the data, such as air quality and water quality, in the warehouse become seriously out of date.
2. The warehouse is maintained by computing the incremental updates to the view based upon the updates to the sources. This approach can involve smaller amounts of data, which is very important when the warehouse is large. The drawback is that the process of calculating the incremental updates is complex, compared with algorithms that simply construct the warehouse from scratch.
3. The warehouse is changed immediately, in response to each change of the sources. This approach requires too much communication and processing to be practical for all but small warehouses whose underlying sources change slowly.

3. Assuring Quality of Data in the Warehouse

3.1 Data quality management

The data quality in the warehouse is determined not just by a single process; all the processes that take place in the warehouse environment may affect it. Thus, quality considerations have accompanied data warehouse research from the beginning.

We propose a systematic approach for data quality assurance. In this approach, data quality management consists of the following four phases as shown in Figure 2. After accomplishing these activities, we may determine the data items that need to be improved to meet the goal of "fit to use."

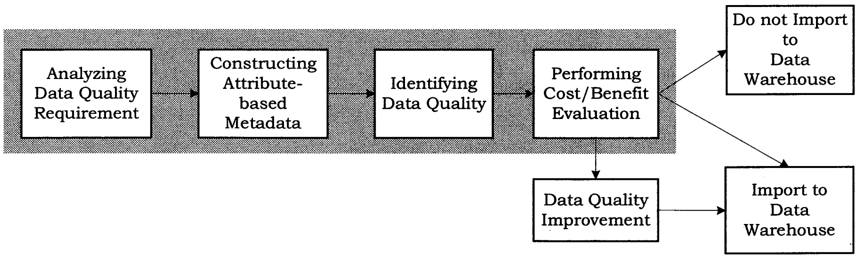
Analyzing Data Quality Requirement
Constructing Attribute-based Metadata
Identifying Data Quality
Performing Cost/Benefit Evaluation
Do not Import to Data Warehouse
Import to Data Warehouse
Data Quality Improvement

Figure 2: The process of data quality management

Phase 1: Analyzing data quality requirements. This phase is similar to the logical design of conventional database systems, wherein the system designers have to figure out semantic ambiguities and syntactic inconsistencies from various sources. Data issues and quality issues should be taken into account during this phase. For determining the type and number of data quality factors to be covered to meet user's needs, the results of this phase will be the specifications for data quality management requirements in the data warehouse.

Phase 2: Constructing attribute-based metadata. Data warehouse systems usually have a multi-dimensional schema to store integrated data from different sources. To ensure the data quality, we add an extra dimension dedicated to the description of data quality for each specific attribute. Moreover, the quality data can be combined with attribute data to simplify the description. Although the description of data quality will cause overhead in storage resources, we believe the benefits of having good data quality can cover the cost of storage space.

Phase 3: Identifying data quality. A data warehouse may support decision making for users at different levels in an organization. The corresponding requirements of data quality are different for each user. This is consistent with the principle of "fit to use." For example, for the quality factor of timeliness, some users may need data collected during last year, while the others may need the data collected during the past decade for detailed analysis. Therefore, we need to identify data that may not fit needs and the causes of the mismatches.

Phase 4: Performing cost/benefit evaluation. Once the unqualified data are identified, we have to find out how to improve the quality of those data. In practice, we need to take cost issues into account to determine the efforts needed to improve the quality of unqualified data. Since it is impractical to achieve a flawless state of data quality, we need to perform a cost and benefit analysis to determine to which level we should improve the quality of unqualified data.

After finishing the above processes, the results may provide the system designer with helpful support to adopt appropriate strategies for data quality assurance. We may import data that meet the requirements of data quality, to the warehouse immediately.

Unqualified data can be divided into two categories. The first portion will be imported to the warehouse after we improve the quality, for the cost of improvement is acceptable. Another portion represents unqualified data that are too expensive to be improved. It is a tradeoff issue whether such data should be imported to the warehouse, and the system designers and users should make decisions with deliberation.

3.2 Cost/benefit evaluation

We propose that the *total cost* of data quality should include the *lost cost* and the *improvement cost*. In accordance with these two issues, we can determine what kinds of data items in the data warehouse should be modified and improved. We define the *lost cost* as the cost caused by poor-quality data. The cost may be expressed in terms of lost funding, lost production, lost assets or legal liability. The *improvement cost* means the cost to improve data quality to a certain level. It is dependent on the number of data quality indicators that need to be improved or modified.

Let poor-quality data exist from $t=t_0$, the improvement activities start at $t=t_1$ and finish at $t=t_2$. Then, the lost cost caused by poor-quality data is $Q(t_2)$, and the improvement cost is $Q(t_2-t_1)+C$, where C is the cost of time-independent issues such as material resources. Moreover, we found that dQ/dt is more convenient for evaluation than $Q(t)$, because dQ/dt represents the proportions of poor-quality data within a time unit, and appropriately stands for the degradation of data quality in the warehouse.

We construct the model based upon the dQ/dt □ t relationship. During t_0□t□t_1, the degradation rate of data quality, denoted as □, will be a linear ratio along with time. As $t=t_1$, the amount of inadequate data become $dQ(t=t_1)/dt = q$.

When t_1□t□t_2, is the time period when inadequate data occurs, and the degradation of data quality is $Q(t_2) = \int_0^{t_2} (dQ/dt)dt \cdot$ Let $\frac{q}{t_2 - t_1} = \lambda x - \beta$, we may compute the lost cost:

$$\mathbf{LC} = C_1 Q(t_2) = C_1 \int_0^{t_2} \frac{dQ}{dt} dt = \frac{1}{2} C_1 q t_1 + \frac{C_1 q^2}{2(\lambda x - \beta)} \qquad (1)$$

Then, we may obtain the improvement cost as follows.

$$\mathbf{IC} = C_2 x(t_2 - t_1) + C_3 x = \frac{C_2 q x}{\lambda x - \beta} + C_3 x \qquad (2)$$

Accordingly, adding the lost cost and the improvement cost, we obtain the total cost as follows.

$$\mathbf{TC} = \frac{1}{2} C_1 q t_1 + C_3 x + \frac{C_1 q^2}{2(\lambda x - \beta)} + \frac{C_2 q x}{\lambda x - \beta} \qquad (3)$$

Example 2: We use the Environmental Unified Identification Code System (EUIC)[1] of Taiwan Environmental Protection Administration (TEPA) as a pragmatic example to illustrate the computation of total cost for enhancing data quality. EUIC is an integrated data warehouse system that provides a single point of access to data ex-

[1] http://euic.epa.gov.tw

tracted from four major TEPA databases, namely the Air Pollution Control System, the Water Permit Database, the Hazardous Waste Control System, and the Toxic Release Database. According to the EUIC experience, maintaining the data in EUIC so that it is consistent with the legacy databases is the biggest problem in terms of data quality. Due to limited budget and resources, EUIC managers usually face difficulties in determining how often does EUIC need to synchronize with legacy databases, and how many data items should be synchronized. We believe that Equation 3 may assist EUIC managers to evaluate the total cost for obtaining "fit to use" data quality. Figure 3 describes a simplified computation of total cost using Equation 3. A more detailed description of the data quality management and analysis can be found in (Chu 2000).

q = 0.25, represents that around one-fourth of data items might have quality problems.

λ = 0.6, the average rate of improvements.

β = 0.8, the degradation rate of data quality.

t_1 = 4 months, defined by system managers, and shows that the data in the warehouse should be synchronized with its sources once every four months.

C_1 = 1, coefficient for the ratio of lost cost and data quality.

C_2 = \$3,000, the input man-hour cost.

C_3 = \$500, the cost of material resources such as computer hardware.

$$\begin{aligned}\textbf{TC} &= (1/2)0.25\,(4) + 500x + (\,0.25\,(4)^2 / 2(0.6x - 0.8)) \\ &\quad + (\,3000\,(0.25)x / (0.6x - 0.8)) \\ &= 1/2 + ((300x^2+350x+2) / (0.6x-0.8)) \quad \text{where } x > 4/3\end{aligned}$$

Figure 3: A simplified example for computing the total cost

4. Conclusion

We present a comprehensive data warehouse framework for integrating environmental information. The proposed framework may play as a platform and provide appropriate mechanisms for retrieving and gathering environmental data from various sources. The gathered data might be transformed into the warehouse after appropriate conversion, reconcilement, and summarization. Therefore, queries for environmental information may be issued to the warehouse exactly as to any database and obtain better performance. The integrated data warehouse may serve as an auxiliary machinery to achieve the semantic interoperability among heterogeneous information sources.

Concerning the data quality assurance, we introduce a four-phase process to ensure data quality while performing information integration. The main points of our research in data quality assurance can be summarized as follows. (1) Overall data quality conditions can be identified and relevant information can be provided for determining whether the data meet "fit to use" criteria and whether they need to be improved. (2) Users may filter the data retrieved from the warehouse based upon various quality requirements. On the basis of constraints of cost and timing, we may then figure out what kinds of data should be preferentially modified or improved in order to achieve maximum benefit of data quality.

There is a possible extension of this research that can be undertaken in further research. Recently, knowledge management has becoming one of the key progress factors in organizations. It involves explicit and persistent representation of knowledge of dispersed groups of people in the organization, so as to improve the activities of the organization. We believe that there are important aspects that can be supported or even enabled by integrated data warehouse.

Bibliography

Chu, Y. C., S. S. Yang, and C. C. Yang, (2001): "Enhancing data quality through attribute-based metadata and cost evaluation in data warehouse environments," *Journal of the Chinese Institute of Engineers*, (**24:4**) 497-507.

Envirofacts Data Warehouse. `http://www.epa.gov/enviro/`

Fabris, P., (1998): "A Civilian EPA Action," *CIO Magazine*.

Garcia-Molina, H., J. D. Ullman, and J. Widom, (2000): *Database Systems Implementation*, Prentice-Hall Inc., NJ USA.

Kimball, R., (2002): *The data warehouse toolkit*, 2nd edition, Wiley Publishing, New York, USA.

Kashyap, V., (1999): "Design and creation of ontologies for environmental information retrieval," *12th Workshop on Knowledge Acquisition, Modeling and Management (KAW'99)*.

Laudon, K., and J. Laudon, (2000): *Management Information Systems*, 6th edition, Prentice-Hall Inc., NJ USA.

Ponniah, P., (2001): *Data Warehousing Fundamentals*, Wiley Publishing, New York, USA.

Sowa, J., (2000): *Knowledge Representation: Logical, Philosophical, and Computational Foundation*, Brooks/Cole, Pacific Grove, USA.