Integrating Environmental Information through Data Warehousing with Data Quality Assurance *

Yu-Chi Chu[1] **Chea-Yuan Young**[2] **Chen-Chau Yang**[3]

[1]Dept. of Information Management
Taipei County Council
Taipei, Taiwan, Republic of China

[2]Environmental Protection Administration
Taipei, Taiwan, Republic of China

[3]Dept. of Electronic Engineering
National Taiwan University of
Science and Technology
Taipei, Taiwan, Republic of China

ABSTRACT

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 [5]. 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.

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 dedi-

*This work was supported in part by Taiwan Environmental Protection Administration, grant#EPA-88-U1L1-03-003.

cated 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 [6].

However, there are some problems that should be addressed in data warehousing [3, 4]. For example, data from different sources may have serious semantic differences, and is likely to contain syntactic inconsistencies. Moreover, the desired data may simply not have been gathered. Therefore, data warehousing projects may not succeed for various reasons, such as poor system architecture or unacceptable query performance, but nothing is more certain to yield failure than lack of concern for the issue of data quality. 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 [1, 2].

On the basis of the consideration of quality assurance (QA), we introduce a two-step process to ensure data quality when performing information integration. We first construct a practical data quality model, then employ the cost/benefit evaluation model to determine the "fit to use" criteria of data quality. Our methodology not only concerns the processes of improving data quality, but also takes into account the cost of data quality improvement. Since metadata plays an important role throughout the life cycle of data warehouses, we adopt an attribute-based metadata model for identifying data quality. By using the information provided by the quality metadata model, we can identify the sources of poor-quality data. Furthermore, a cost/benefit model is used to identify the most fatal poor data and set the priorities for improvement, given limited resources.

We use the Environmental Unified Identification Code System (EUIC)[1] of Taiwan Environmental Protection Administration (TEPA) as a pragmatic example to illustrate the

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

computation of total cost for enhancing data quality. EUIC is an integrated data warehouse system that provides a single point of access to data extracted 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 our approach may assist EUIC managers to evaluate the total cost for obtaining "fit to use" data quality.

References

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

[2] Chu, Y. C., *Integrating Heterogeneous Information Sources through Ontology-Driven Model and Data Quality Analysis.* Ph.D. Dissertation. National Taiwan University of Science and Technology, (2001).

[3] Garcia-Molina, H., W. J. Labio, J. L. Wiener, Y. Zhuge, 1998, "Distributed and parallel computing issues in data warehousing," *Proceedings of the 17th ACM Symposium on Principles of Distributed Computing,* pp. 7-10.

[4] Kimball, R., 1996, *The data warehouse toolkit,* Wiley Publishing, New York, USA.

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

[6] Roberts, D., "Information Technology and ISO 14000," in *Implementing ISO 14000,* T.Tibor and I. Feldman editors, IRWIN Publishing (1997).