Applying a Metadata Framework to Improve Data Quality

Victoria Youngohc Yoon
University of Maryland, Baltimore, USA

Peter Aiken
Defense Information Systems Agency, USA

Tor Guimaraes
Tennessee Technological University, USA

INTRODUCTION

The importance of a company-wide framework for managing data resources has been recognized (Gunter, 2001; Lee, 2003, 2004; Madnick, Wang & Xian, 2003, 2004; Sawhney, 2001; Shankaranarayan, Ziad & Wang, 2003). It is considered a major component of information resources management (Guimaraes, 1988). Many organizations are discovering that imperfect data in information systems negatively affect their business operations and can be extremely costly (Brown, 2001; Keizer, 2004). The expanded data life cycle model proposed here enables us to identify links between cycle phases and data quality engineering dimensions. Expanding the data life cycle model and the dimensions of data quality will enable organizations to more effectively implement the inter- as well as intra-system use of their data resources, as well as better coordinate the development and application of their data quality engineering methods.

BACKGROUND

The proposed model has a number of inputs/outputs distributed throughout eight phases: metadata creation, metadata structuring, metadata refinement, data creation, data utilization, data assessment, data refinement, and data manipulation. Each of these phases is described next in more detail.

Two possible cycle "starting points" are shown bolded in Figure 1. The first starting point is applicable to new systems where there exists no data to be migrated, and/or converted from existing system(s). In these instances, the model cycle begins with metadata creation and proceeds counter- clockwise around the cycle. However, according to a survey of CIOs by Deloitte & Touche (1996), an average of more than 90% of organizational legacy systems were scheduled to be replaced by 2001. Legacy systems continue to impede business effectiveness for many. Only 6% of insurers plan to replace their systems in the next two years, while 29% plan to replace their legacy systems within four years (Chordas, 2004). In a survey of 115 business-technology professionals conducted by Optimize Research, 42% called their legacy IT a barrier to innovation and flexibility, while 8% said that it bogs down business and is expensive to maintain (Whiting, 2003). Thus, it is more likely that an organization's legacy data will become the major data asset to be managed. In these cases where data already exist, structural data quality re-engineering becomes necessary, and the cycle begins with data assessment. Next, each cycle phase is described in more detail.

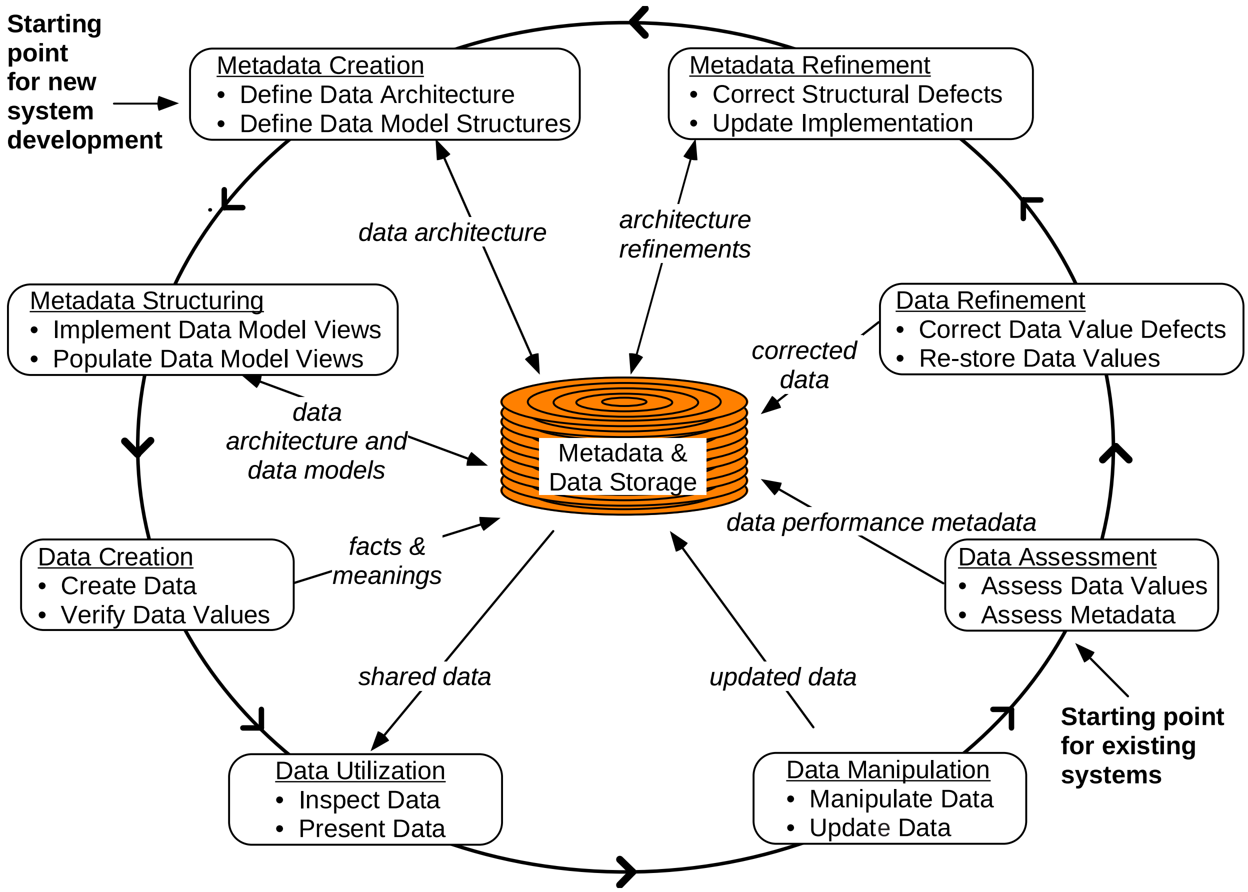
Metadata Creation

When the requirements dictate that users interact with multiple systems across functional area boundaries, a formal organizational data architecture is required to coordinate data quality engineering efforts. While all organizations have data architectures, only formally specified architectures can be formally managed. This phase typically corresponds to increasing awareness of data as an organizational asset. The architectural metadata created and evolved consist of the organizational data architecture structure definitions and specific associations among individual system data models.

Metadata Structuring

This phase focuses on developing a framework guiding the organizational data architecture implementation as it populates data models in the next phase. Metadata creation is followed by the development of a data model structure. Data models must also be evolved. The term

Figure 1. Newly proposed eight phases of extended data life cycle model with metadata sources and uses



"structuring" indicates the iterative development process that occurs as the organizational data architecture structure developed during the previous phase is populated with metadata. Defining data model structures permits organizations to understand the categories of data that comprise its data models. The process consists of populating the data architecture with data models describing the various specific systems. Each data model corresponds to one physical occurrence. In addition, when physically implemented, logical model components can be physically implemented by multiple systems, accessing common DESs. The process of defining data models as components extends the organizational data architecture comprehensiveness. Metadata structuring is complete when all entities can be associated with specific model components. Perfect model metadata occurs when a correct data model exists for each physical system, and each physical system component is associated with one and only one common organizational data architecture component.

Metadata Refinement

At various points, portions of some metadata can be determined imperfect. Architecture refinement implements an iterative approach to refining the existing metadata-based concepts, correcting factual errors, and evolving the structure to a more perfect state. This usually occurs in response to data assessment activities.

Data Creation

Data creation occurs when data values are captured from some external source and stored in systems. Data sources can range from a point of sale terminal, to EDI, to floppy disk exchange. Data creation is the most popular focus of data quality engineering efforts. These are commonly implemented as edit masking, range checking, or other forms of validation. Data value quality efforts are aimed at perfecting data values as they are captured and before they are stored or re-stored in the database.

Related Content