TITLE: "THE BENEFITS OF A DATA TAXONOMY"

by Tim Bryce
Managing Director
M. Bryce & Associates (MBA)
P.O. Box 1637
Palm Harbor, FL 34682-1637
United States
Tel: 727/786-4567
E-Mail: timb001@attglobal.net
WWW: http://www.phmainstreet.com/mba/
Since 1971: *"Software for the finest computer - the Mind"*

"You must first plant the seeds in order to harvest the crop."
- Bryce's Law

INTRODUCTION

In "PRIDE" Special Subject Bulletin #30 ("*Stagnation in Data Administration*" - June 27, 2005) I mentioned the concept of a "Data Taxonomy" as used to classify data so it can be standardized, shared and re-used in multiple systems. This is a concept we first introduced with the advent of the "PRIDE"-Data Base Engineering Methodology (DBEM) in 1987.

As I mentioned in the bulletin, the standardization of data definitions is a major problem in I.T. departments around the world. Instead of defining the characteristics of a data element one time and re-using it over and over again, most companies re-define data with each application. Consequently, inconsistent results begin to emerge (sometimes called "Dirty Data"). For example, we know of one state government who conservatively estimated "Net Pay" was defined over 100 different ways in their organization. Not only does this lead to inconsistencies and erroneous information, it also inhibits the implementation of change. Does anybody remember the Y2K problem just five years ago when teams of developers tracked through voluminous program libraries to find and correct dates? Had date-related data elements been properly defined and cataloged, this would never have been a problem.

There are three reasons for redundant data definitions:

1. The lack of an effective tool to define and cross-reference data elements. This was the intent of the "Data Dictionary" which was later referred to as "Encyclopedia" or "Repository" (in "PRIDE" we call it the "Information Resource Manager"). Today, there are numerous interpretations of the Data Dictionary, all providing basic support for cataloging data elements and showing where each element is used in records, files, and programs. If such tools are currently available, why do we still have a problem? See #2.

2. Companies lack the foresight or will to standardize on data definitions. You may recall my telling of the story from years ago when India had a serious problem with famine. To help solve the problem, the Americans sent tons of seed-grain to India for planting. Instead of planting and harvesting the grain, the Indians ate the seeds. You cannot harvest what you do not plant. The same is true in defining data. The real benefits are long term in nature and requires an up-front investment in time required to properly define data elements. But once the data has been properly defined, this intelligence can be used over and over again in as many systems as you can imagine. The problems of data sharing and systems integration as mentioned above are eliminated; even better, application development time is reduced as data definitions are re-used.

The only problem here is that it requires management vision and commitment to its implementation. The reality, however, is most companies are short-sighted and content with defining data over and over again with each application.

3. The third reason is that people simply do not know how to properly define data elements. Most application developers only look at it through the programmer's eyes and rarely consider data beyond its program label.

This is where we come in.

CLASSIFICATION

Sharing and re-using data doesn't happen by accident. There has to be a premeditated and concerted effort introduced. In other words, data must be defined in a consistent manner making data sharing not only feasible, but a natural part of the development process. To do so, management needs to create a standardized and methodical approach for defining data elements and enforcing its use on a corporate basis. Fortunately, there are some simple techniques to help in this regard.

The management of any resource requires the development of a classification system. Financial resources are typically arranged according to a chart of accounts; material and human resources are categorized by type. In science, everything from chemical elements to the animal kingdom are organized according to a class struc-

(continued on page 2)

(continued from page 1)

ture. There obviously is a purpose to uniquely identify common elements; to provide for the ability to distinguish one from another, and eliminate redundancy. In all instances, classification is based on the inherent characteristics of the component.

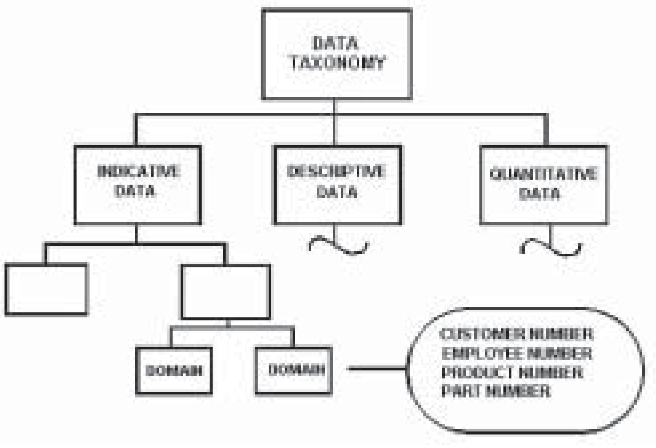
To classify data elements, we must have an appreciation of data's logical and physical properties. "Logical" properties refer to the business purpose of the data and includes such things as a dictionary-like definition, along with its "source" (for "primary" values, where it originates from in the company; for "generated" values, the other data elements used in its calculation); and "type" (how used for Indicative, Descriptive, or Quantitative purposes). "Physical" properties refers to how data is to be recorded, stored and presented to the user, and includes such things as programming labels, length, validation/editing rules, etc. Understand this, a data element has only one logical definition but can have multiple physical expressions; e.g., how dates and currencies are expressed, or different program labels for COBOL, C++, etc. (more on this shortly).

DATA TAXONOMY

A Data Taxonomy is simply a hierarchical structure separating data into specific classes of data based on common characteristics. The taxonomy represents a convenient way to classify data to prove it is unique and without redundancy. This includes both primary and generated data elements.

CLASSIFYING DATA

The objective is to eliminate redundancies and promote sharing/integration



DOMAIN - Elements with similar characteristics

The lowest level in the classification hierarchy represents what is commonly referred to as the "domain" of a collection of data elements, one or more, with common characteristics. For example, "text" related data elements would be in one domain, "weights" in another, "percentages" in another, "monetary values" in another, etc.

The domain also defines the standard physical characteristics and values the data may assume. For example, we could establish that all "location" values are alphanumeric, left justified, with blank fill and void characters. In other words, data elements such as "Address," "City," and "Country" should assume these physical characteristics for consistency. If a data element does not have the standard logical and physical characteristics, it must belong to another "domain."

In the situation where a data element has only one logical definition, but multiple physical definitions, its primary physical definition must first conform to the Domain standards before it can be deviated from in an application record. In other words, the primary physical representation of "Unit Cost" is expressed as an eight character numeric to conform to the "currency" domain. However, in one application, a user desires the data element be expressed as a ten character numeric. It is the same logical data element with just another form of physical expression.

With a classification system in place, data elements can then be uniquely and consistently defined. When this is done, we then have a basis for checking data redundancy. Also, when a data element has been properly specified in this manner, it becomes rather simple to locate it in other applications.

GUIDANCE SYSTEM

To expedite data definition, developers should be provided a "Guidance System" to prompt them through the proper classification of a data element. This can be used to either define a new data element or validate the integrity of an existing data definition. The "Guidance System" follows the hierarchy of the Data Taxonomy which records the characteristics of the data element until it finds its domain. The result is a uniquely defined data element suitable for sharing and use in multiple systems. At this point, the data definition should be locked to prohibit changes from occurring either accidentally of intentionally. For those of you considering the purchase of a data dictionary/repository, this is a highly desirable feature.

(continued on page 3)

(continued from page 2)

ENFORCEMENT

Classifying data as described herein represents a discipline which can be performed voluntarily by developers. However, safeguards should be added to enforce proper usage. A couple of suggestions come to mind: First, data definitions should be reviewed and approved by a neutral party. Whereas system developers will be charged with identifying the need for data elements, the Data Resource Management department should inspect and approve all data definitions. Second, get system developers out of the data base design business and leave this to the Data Resource Management department. After all, developers will only do what is necessary for their specific application and not necessarily what is best for the company overall. To enforce this, all file structures should come from the Data Resource Management department and nowhere else. As an example, years ago we enforced such a policy over programmers by controlling the COBOL copybooks.

With an enforceable discipline in place, your chances for success have increased radically.

CONCLUSION

Classifying data helps to fulfill one of the the major objectives of Data Resource Management: to eliminate redundancy and promote the re-use of data in systems. The initial investment in documenting data elements pales in comparison to the long-term benefits derived from the effort. For example, integrated systems assures consistent results ("Clean Data") and simplifies maintenance and implementing changes; and, ultimately leads to reduced time in systems development. But make no mistake, the benefits of classifying data are long term in nature, not short term.

But why stop at data elements? Why not classify and re-use all information resources and put an end to the redundancy issue once and for all? I can build a compelling argument for classifying records, files, inputs, outputs, programs, modules, business processes, etc. From this perspective, a "Data Taxonomy" should be superseded by a "Resource Taxonomy" which considers all information resources, not just data. But who am I kidding? This will only work if management wakes up and has the foresight to develop a long-range plan to manage information resources. Unfortunately, most will continue to think on a short-term basis and continue to eat the seeds.

END

"PRIDE" Special Subject Bulletins can be found at:

http://www.phmainstreet.com/mba/mbass.htm

They are also available through the "PRIDE Methodologies for IRM Discussion Group" at:

http://groups.yahoo.com/group/mbapride/

You are welcome to join this group if you are so inclined.

"PRIDE" is the registered trademark of M. Bryce & Associates (MBA) and can be found on the Internet at:

http://www.phmainstreet.com/mba/pride/pride.htm