

Appendix 6. What is Metadata?

Metadata is a succinct description of data (literally, data about data). Metadata enables the SCS-DSS to know what data exists and where to find it. Metadata is a kind of index that sits above the data store and keeps track of what kind of data exists, what specific data records exist, and where the specific data records reside. Conversely, metadata constitutes the rules for cleansing data and performing the extract, transform, and load (ETL) process. (During the ETL process, the data is transformed to provide consistency. This is why it is important that the ETL tool is metadata-aware and preferably responsible for managing the central metadata repository.)

To work properly, an SCS-DSS needs a set of standards that agree across functional boundaries. For example, the field called CP_LastName needs to have the same meaning and be structurally identical for collections, locates, parental establishment, or some combination of those.

Figure 6-1 illustrates three ways information can be collected, categorized, and searched. The first example is the once-familiar card catalog found in many libraries. Each of the card catalog's three sections (author, title, and subject) holds metadata (data about data)—pertinent information about the library's books. Which drawer you opened depended on what you already knew about how the card catalog worked and the book you wanted. If you didn't have that basic knowledge, the card catalog (and therefore the library it referenced) was just a bewildering collection. Two people could look up an author and find different genres. Or one book could be referenced on several different cards.

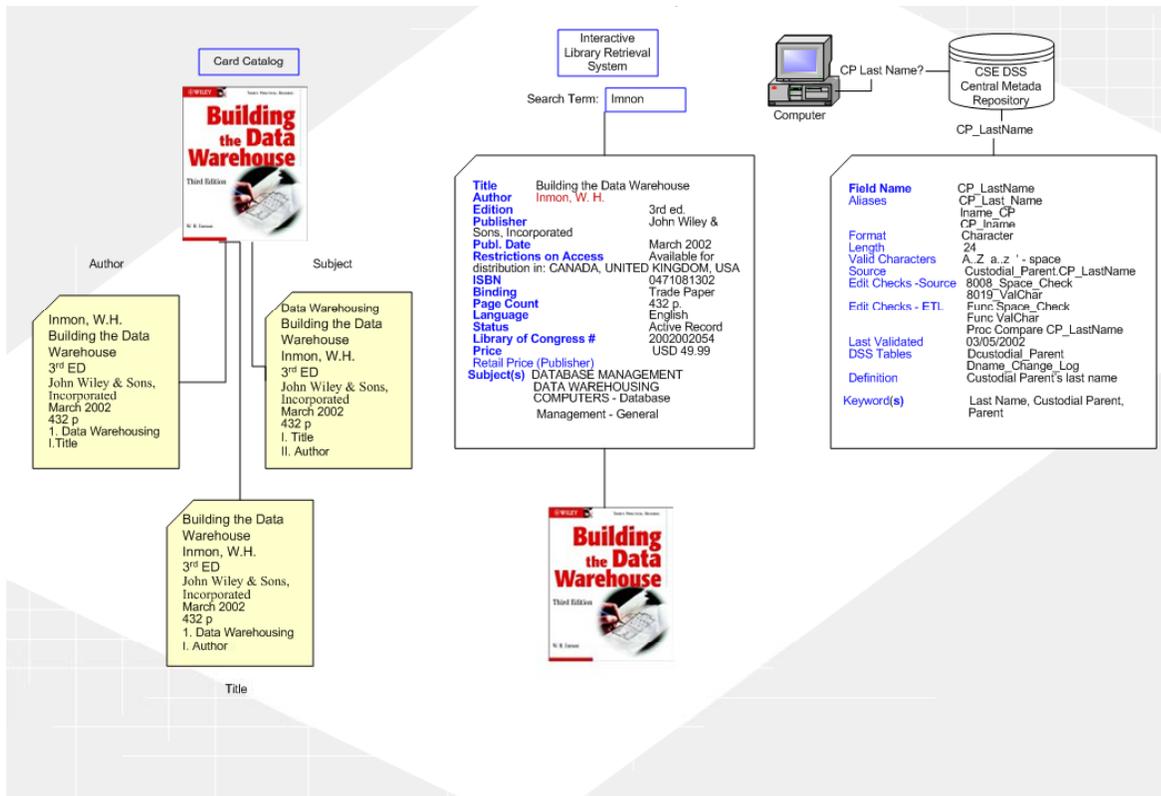


Figure 6-1. Three Metadata Examples

This type of data management in business is what leads to the situation Mr. Jones and Mr. Smith experienced with their very different view of the company's prospects. How were they to know there were two fields called "expected profit," especially if the organization had a poorly designed or nonexistent data dictionary?

The second example is an online book-retrieval system. Book information is categorized by standard identifiers (such as title, author, binding, and page count). These identifiers are called metatags; in **figure 6-1**, they are highlighted in blue.

The final example is a Central Metadata Repository (CMDR), the ideal method for managing large-scale information systems. It provides end users with unambiguous information about the data they want; here, expected profit is always the same, whether you work in finance or marketing. A CMDR provides developers and database administrators with an accurate map of the background and content of each element and tracks new

actions taken on the element. And because good database design ensures that all software elements in the DSS are metadata-aware, system managers can impose greater control and consistency on the entire system and its outputs.

In establishing good metadata management, the practices must be easy to use and draconian at the same time.

Easy to use means that maintaining the elements within the central metadata repository is not onerous, that all the elements of the system are metadata-aware, and that end users know the correct data elements for queries and reports.

Draconian means that, before an element can enter the CMDR, it must have a rock-solid pedigree. Instead of finance and marketing each having a different field called "expected profit," the organization now has one field called "expected profit" containing only one value.