# Effective Use and Management of Metadata

Cathy Phipps, SAS Institute Inc., Cary, NC
Jim Davis, SAS Institute Inc., Cary, NC

## ABSTRACT

Metadata: you know you have it (somewhere), you know you should be using it better, but how? The SAS/Warehouse Administrator™ software in Version 6 and the Common Metadata Repository for the SAS system can help answer that question.

The V6 SAS/Warehouse Administrator software provides the tools you need to define your data warehouse and to build a single store of metadata about your operational and data warehouse environment. Its read and write API opens up your metadata to other applications for exploitation.

The metadata story expands quite a bit in Version 7 with the introduction of the SAS CMR software, a common metadata facility. SAS/Warehouse Administrator will soon be one of several SAS products using SAS CMR, which means that metadata can be shared by SAS/Warehouse Administrator, SAS/EIS® software, SQL Query Window and other SAS products without metadata exchange or import/export.

## INTRODUCTION

Managing metadata is not only a key component of a well architected IT infrastructure, but it also represents a valuable road map that can provide an organization's decision support users with a strong competitive advantage. For years companies have been accumulating massive amounts of transaction data to support their operational systems. Only recently have these organizations begun to realize the benefits associated with reorganizing that data into useful information.

For example, a mail order company enters thousands of orders each day into their computer systems. Those orders then trigger appropriate shipping and billing activities. While these systems keep the operation running, they do very little to assist the organization with the analysis necessary to keep them ahead in a highly competitive market place. These systems were optimized to get data in, not to get information out.

The same data that are used to support the process of handling individual orders, hold enormous potential as the company searches for new and better ways to manage their products. For example, the transaction data might be summarized by product or product category. Trends may then emerge that help pinpoint reasons for a product's acceptance or rejection in the market place. The data warehouse stores information in a state optimized for this type of decision support activity.

Data warehousing is the process of extracting data from diverse operational environments and reorganizing that data into useful information. The information about how the data are moved from the operational to the warehousing environments is contained in its metadata. Effective use and management of this metadata allows the business analyst to turn his attention away from finding where the data are buried in the organization to exploiting this new information store to help make timely decisions that move the company forward.

But metadata is more than just a roadmap for extracting and transforming operational data. Metadata is used by your decision support applications to help you understand and analyze your business data. Integrating metadata at all levels of your IT organization into a common metadata facility, such as the SAS CMR software, will provide the framework for centralized metadata definition and administration. The benefits will be reflected in decisions based on consistent, up-to-date information.

## METADATA: WHAT IS IT?

Simply put, metadata is data about data. At a high level, metadata can be defined as either technical metadata or business metadata. Technical metadata are typically used by IT personnel in support of development and data management activities. Business metadata are used by the end user community to help them locate and identify the data, or information, necessary to support their decision support activities.

In his book entitled *Managing the Data Warehouse*, Bill Inmon states, "The heart of the architected environment is the data warehouse; at its nerve center is metadata. Without metadata, the data warehouse and its associated components in the architected environment are merely disjointed components working independently and with separate goals. In order to achieve harmony and unity across the different components of the architected environment, there must be a well-defined and disciplined approach to metadata."

If we divide the warehousing process into three basic steps, we can illustrate the type of information that the metadata store maintains. Information includes, but is not limited to:

1. *Definition of Operational Data*
   - Location of the operational data
   - Table and column definition
   - Owner and administrator
   - Descriptions
2. *Definition of the Target Warehouse Store*
   - Location of the target store
   - Table and column definition
   - Owner and administrator
   - Descriptions
3. *Transformation of Data*
   - Data mapping definition
   - Derived definitions
   - Business rule descriptions
   - Dependencies
   - Compute platform
   - Load frequency

## SAS/WAREHOUSE ADMINISTRATOR

SAS Institute has long been a dominant supplier of the tools necessary to facilitate the warehousing process. SAS data access technology provides the means to extract data from diverse sources. SAS software is platform independent and provides robust transformation capability. SAS/CONNECT® software provides the pieces necessary to move data between platforms.

The difficulty lies in managing the warehousing process across a diverse IT environment. SAS/Warehouse Administrator software provides a flexible framework for effective warehouse management through a metadata-driven architecture. SAS/Warehouse Administrator provides a central point of control for managing the movement of data from the operational environment into the data warehouse through the use of a graphical user interface. All activity associated with the creation and management of the data warehouse is managed via the metadata layer maintained by SAS/Warehouse Administrator.

## COMMUNICATING WITH METADATA

The metadata maintained by SAS/Warehouse Administrator is open and available to applications outside the warehouse administration environment. A metadata application program interface (API) is a set of tools that enable users to write applications that access metadata. Using the metadata API, you can write SCL applications that read, add, or update the metadata store.

The read API can be used to extract information from the metadata store to make it available to other applications. For example, metadata can be extracted and published for users other than those responsible for the day-to-day maintenance of the warehouse. HTML pages can be generated on a company intranet for both IT and end users. IT can benefit in the form of web-based process documentation. Information such as table and column descriptions, location of data, data dependencies, and source code can be made available to the technical community on their intranet. The end user can access information regarding the source of their information, as well as identify the owner and administrator. This information can be critical to an end user who is about to make a major business decision based on the accuracy and timeliness of data in the data warehouse.

The write API can be used to populate the metadata from sources outside the SAS environment. For example, often an organization has their data defined in case tools, data dictionaries, or modeling tools. Rather than using SAS/Warehouse Administrator as a data entry mechanism for tens or even hundreds of tables, the metadata write API can be utilized to populate structural information from the organization's existing tool into the metadata store.

## METADATA: A VALUABLE CORPORATE ASSET

By now, the merits of a metadata store should be clear. Metadata provides a single point of control for managing, maintaining and documenting an organization's data warehousing environment. Metadata also represents a new, valuable corporate asset in itself.

By defining the data warehouse with SAS/Warehouse Administrator, the organization has built a single store of metadata that contains information on the content and structure of the operational and data warehouse environments. This new information store can be fully exploited.

Applications can be built that allow a user to use the metadata store to navigate the enterprise, determine the relevance of certain information to a particular task at hand, and surface that information on the end users desktop for further analysis. For example, data can be made readily available to a wide audience via an intranet. End users with web browsers can access the metadata via a SAS/IntrNet™ server, search for relevant information, and download it to their desktop to be further exploited with their tool of choice.

## MOVING BEYOND THE DATA WAREHOUSE

As we've discussed, metadata is critical to the successful creation and management of the data warehouse. But the process of moving data out of operational systems into a decision support environment is only half the battle. The applications used by our decision support communities are driven by metadata as well. Report writers, OLAP solutions, query generators, EIS applications are all driven by metadata. These tools need to know where the data reside, how the data are to be accessed, what sorts of hierarchies are available, just to name a few examples. Historically these applications have had their own independent metadata stores. The next generation of SAS software will introduce an open metadata repository that manages the warehousing process and provides a single point of control for a wide variety of end user tools.

In such an integrated environment, the business analyst can browse the metadata to see what, for example, hierarchies are available. With the integrated metadata, the analyst can not only see about the hierarchies, but can use the metadata to see how the hierarchies were derived, from which operational data sources they were derived and what business rules were applied to the data. In essence, you can determine the source of any information available and how the data have been manipulated.

## VERSION 7: METADATA INTEGRATION

The SAS system is prepared to address the issue of metadata interoperability, both for SAS products and with products sold by other vendors, with the release of Version 7 and the introduction of the Common Metadata Repository for the SAS system, or SAS CMR.

### Common Metadata Repository

The SAS CMR software provides a common metadata facility for SAS applications. This means that metadata you enter for one product is available to other SAS software products without conversion or import/export. For the initial release of Version 7, these products will include SAS/EIS software, SAS/MDDB™ software, SAS External File Interface and the SQL Query Window. The SAS/Warehouse Administrator will support SAS CMR in its release 2.0. Other products will support SAS CMR in upcoming releases.

SAS CMR is a general-purpose metadata repository, providing metadata-related services to products of the SAS System. It is

object-oriented in nature, with different types of metadata defined as different classes with class-specific attributes. The metadata are persisted as SAS data sets with a collection of such data sets stored in a SAS data library as one "repository". The SAS MultiEngine Architecture™ allows you to persist your repository in an alternate DBMS system as well. There can be more than one repository on a system, all managed by a "repository manager".

### Interoperability with Other Vendors

In addition to the seamless interoperability between products of the SAS system, SAS CMR also allows the import of metadata from other vendors, as well as the export of its metadata to other tools. The SAS/Warehouse Administrator in Version 6.12 and SAS CMR in Version 7 are compliant with MDIS 1.0, which is the Metadata Interchange Specification produced by the Metadata Coalition.

MDIS-compliant software, including the SAS system, is able to interoperate with the Microsoft Repository via the free translation software available to Coalition members. This translation "bridge" allows metadata from the SAS system to be exported via the MDIS interchange format and imported into the Microsoft Repository, and vice versa.

An even closer integration with Microsoft Repository meta-models is under investigation.

## "HUB AND SPOKE" FRAMEWORK

SAS CMR supports a "hub and spoke" metadata architecture. That is, there is a server-based "hub" of centralized metadata control, but multiple "spokes", or local copies, for exploitation of the metadata with various client-based tools.

### Administration Tools

SAS CMR and the SAS products which use the common metadata facility provide tools to help you administer your metadata and your repositories. The SAS CMR administrator tool, REPOSMGR, allows you to set up your repository environment, by defining new repositories and setting up default repository managers. You can view the types of relationships between your metadata, and view and define extended attributes to be assigned to your metadata. The SAS/Warehouse Administrator and its metadata API will probably be your primary metadata creation and entry tools at your hub. The Warehouse Administrator also provides a metadata browser.

The SAS CMR also provides copy management tools which allow you to copy one repository or a group of repositories, in effect creating the "spokes" of your hub. The copy can be from a remote machine to your local machine, from your local machine to a remote machine, or within the local machine. When the metadata at your hub are updated, the metadata at the spokes can be refreshed with the copy management tools. This refresh copy can, at your discretion, overwrite any updates made at the spoke or preserve those updates.

### Example Configuration: Centralized Control

Here is how the hub and spoke framework might work for your site if you wish to exercise centralized control over your metadata. All (or the majority) of your metadata would be created and managed by a metadata custodian or administrator group. The SAS/Warehouse Administrator would be the primary metadata entry tool. There may be one repository or several repositories, depending on your data organizational needs, all under one repository manager located on a central server. Update access to these repositories is restricted to authorized persons.

Most of your users might need to exploit your metadata on their PCs, using client tools such as SAS/EIS or SAS/MDDB software. As they need the metadata, they would use the SAS CMR copy management tools to copy the metadata they require from your central server to their PC. For the most part, their access would be read-only. Any updates that they make to the metadata are made only to their local copy.

Meanwhile your metadata custodian may make changes to the metadata on the server. When your PC users wish to see the changes in their local copies, they would refresh their metadata by downloading a new version from the server, using the SAS CMR copy management tools. Any updates that had already been applied to the previous version of the metadata would be reapplied to the new version.

### Example Configuration: Hierarchical Control

In this example, let us assume that you wish to have some metadata which are controlled site-wide and some metadata which are controlled by individual departments. This can be done one of two ways. One way is to have completely separate and autonomous repositories, under the same or different repository managers. One center of repository control would be similar to the centralized metadata administration group in the previous example. Other repository control centers would be custodians responsible for metadata within individual departments, e.g. finance or sales. Your users wishing to exploit metadata from these different repositories would use the SAS CMR copy management tools to select from which repositories their metadata should be drawn. Local copies of a subset of the selected repositories would be created.

Another way to do this would be to have the second-tier metadata administrators download the centralized metadata into repositories managed by their individual departments and create or update the metadata as required by their department. The metadata from the central repository could be refreshed periodically without destroying the department overrides and additions. The PC users would then be able to download the metadata from a single source: the department repository(ies). Thus a "spoke" of one hub can itself be a hub.

### Central vs. Hierarchical Control ?

There are advantages to keeping either central or hierarchical (distributed) control of your metadata. The more central your metadata administration is, the more consistent the definitions will be across your organization. Depending on your IT and business needs, this may or may not be desirable. If you do want to distribute control of your metadata administration, the SAS CMR gives you the option of centralizing some portions of it as well.

## FUTURE ISSUES

Other features of a robust metadata management system will be introduced in future releases. Some of these features may include metadata versioning, security enhancements and an open applications programming interface (API).

### Metadata Versioning

As useful as it is to know exactly what your metadata consists of currently, it is also important to be able to "remember" what it looked like at a particular point in the past. When metadata versioning is added to SAS CMR, you will be able to do just that: reconstruct your metadata as of a particular date and time.

### Security

Metadata security for SAS CMR is achieved through the file system controls provided by your operating system, internal encryption and through use of metadata replication via the hub and spoke architecture. Future releases will introduce additional security features, such as different user types or additional encryption facilities.

### Open API

The current read and write metadata API is SCL-based. This will continue to be supported and enhanced, but future releases will also feature programming interfaces for non-SCL applications, such as web-based or COM applications.

## CONCLUSION

Understanding and fully exploiting your metadata assets is key to getting the most out of your data. The SAS system currently provides a very strong metadata solution in Version 6 in terms of organizing and exploiting your metadata with the SAS/Warehouse Administrator. The key to understanding and exploiting your metadata across your enterprise applications is through a common metadata facility, which is introduced in Version 7 of the SAS system with SAS CMR software. This repository management system, with its "hub and spoke" architecture, allows you to control your metadata centrally and distribute it for exploitation.

## REFERENCES

SAS Institute Inc. (1997), *SAS/Warehouse Administrator Metadata API Reference, Release 1.2*, Cary, NC: SAS Institute Inc.

Lewis, Terry (1997), "SAS/Warehouse Administrator Usage and Enhancements", *Proceedings of the Twenty-Second Annual SAS Users Group International Conference,* Cary, NC: SAS Institute Inc.

Inmon, W.H., Welch, J.D., Glassey, K.L., (1997), *Managing the Data Warehouse*, New York: John Wiley & Sons, Inc.