Paper 180-27

Empowering Better Decisions with the use of Metadata

Michelle Ryals, SAS Institute Inc., Cary, North Carolina

ABSTRACT

Is your company using metadata? Are you aware of the advantages metadata can bring to your organization? If you are not, you should be. As your business grows, so does the data contained within. Metadata is becoming more essential than ever for those who wish to stay ahead of their data, and thus, ahead of their competition. Speed is of essence and metadata is the driver.

This paper will give you a better understanding of what metadata is and how it can help you answer the where, how and why about your data. Metadata acts as a key enabler for intelligence and gives you information critical to succeed in understanding and managing your data. The result is the best possible utilization of your data resources.

INTRODUCTION

What is metadata? There are a number of definitions, but it has been described in the most basic terms as “data about data”. Metadata is information surrounding the data resources in an organization. In addition, it contains descriptive Information about the structure and content of data and the applications that process and manipulate it. The clear advantage metadata gives is its ability to make sense of your data. If SAS® allows you *The Power to Know™*, then metadata allows you *The Power to Understand.*

Metadata gives you faster, more focused search and retrieval ability. If you store information about your data you will have easier access and be able to get answers to direct questions quickly.

BREAK IT DOWN

Metadata is typically broken down into two levels:

- Technical
- Business

Technical metadata supports the development, maintenance and management of an Information Technology Environment. This type of metadata is concrete and normally answers the where and how. Some examples of technical metadata include physical storage structures, server systems, installed applications, and data manipulation processes. Technical metadata will allow you to answer questions such as:

- Who created this data?
- When was it created?
- Where is it located?
- How does it interact with other data stores?
- How is it used in reporting?
- When was it last updated?
- Who updated it last?

Business metadata makes the data and services in the environment easier to understand and use. Though it is less concrete than technical, it allows the business analyst to make sound decisions based on the data and normally answers the why. Some examples of business metadata include data classification, presentation definitions, and business meaning and usage. Business metadata will allow you to answer questions such as:

- How is the data created?
- How often is the report generated?
- How is a change in data captured?
- How are the pieces of data related to one another?
- How are the rules defined?
- How are the values defined?
- What do the values mean?

WHY DO I CARE?

One thing we can all agree on is the fact that the amount of data is not starting to slow down and level out. It seems the definition of advancement involves the ability to digest data at a faster rate. This means the amount of data in organizations is growing at a rapid pace, as well as the number of systems producing the data. Likewise, the number of users who need to access and understand the data is growing. This increase can only create more complexity and less understanding.

How useful is your data if it is not understood? The data resources in your organization can be one of your most valuable assets. Without metadata, these assets can go under utilized because they cannot be found, accessed, or understood. Metadata is the answer to this problem and will allow you to fully utilize these resources. One way this is accomplished is by providing integration in a world where many data sources are talking different languages. Communication is the key. Figure A illustrates how many different data stores could use a common metadata model, which many clients could then query.

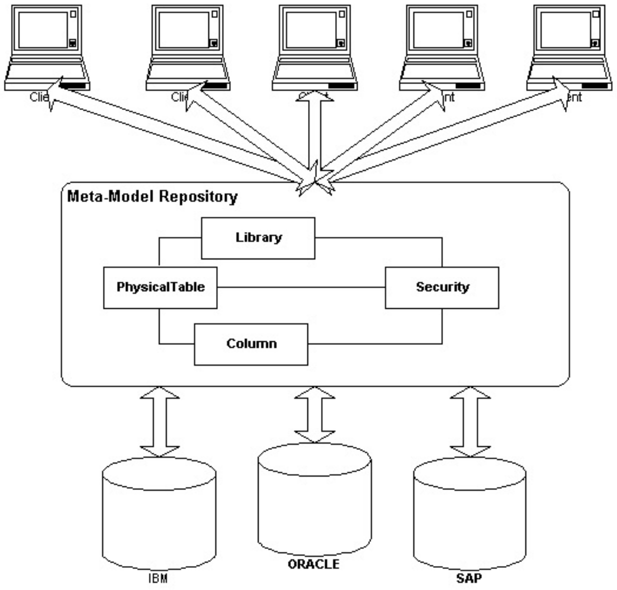


Figure A

Having the ability to query your data through the means of metadata will allow you to answer more direct questions quicker. You will have the ability to know where your reports are stored, when the most recent changes occurred and how they were calculated. You will be able to facilitate impact analysis, which helps explain how data relates with other data throughout the system.

Metadata can offer a common, consistent and secure representation of your data. It is easily accessible and compatible with existing and evolving standards, such as Object Management Group's (OMG) Common Warehouse Metamodel (CWM).

MAKING BUSINESS DECISIONS

Do not categorize metadata as something only useful as Information Technology. Metadata is excellent at empowering the best and most accurate business decisions. Any detective can solve the case if metadata has been defined properly and a typical scenario will help to prove this point. Therefore, let us begin an investigation.

Imagine you are an executive of a successful multi-national enterprise with regional offices in a number of cities throughout the United States. In order to stay on top of things you always begin your day reviewing the latest sales reports. One day in particular you notice what appears to be a huge error in sales reporting for the southeast region. Figure B indicates sales are unusually low in the month of January, as compared to previous months.

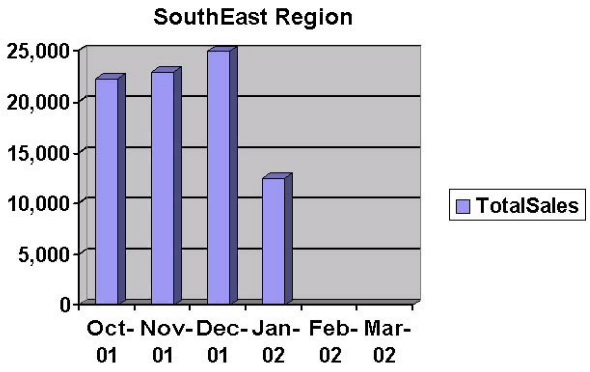
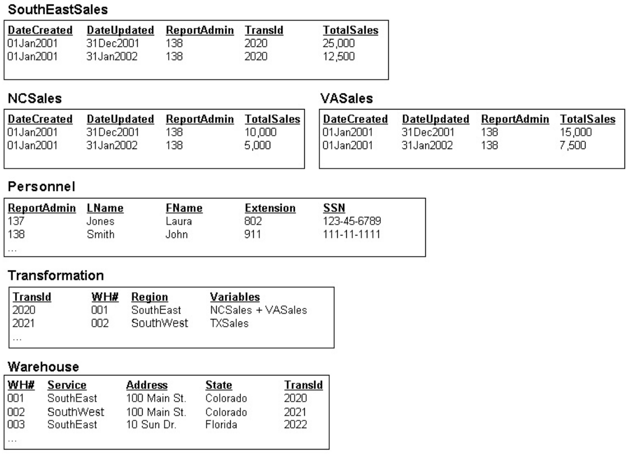


Figure B

Your shock leads you to ask yourself: “Can this be true?” What could have happened between last month and this month to cause a dramatic decrease in sales for this region? Metadata will help you to answer this question. Not only can metadata give you the answer, but it can also help you make better business decisions by allowing you to see the “big picture”. This is the true advantage of being able to understand and manage your data.

THE KEY TO ANY INVESTIGATION

There are a number of metadata tables that contain information relevant to this investigation. Namely, we will look at the *SouthEastSales, Transformation, NCSales, VASales, Warehouse* and *Personnel* metadata tables. Figure C describes these tables in detail.



SouthEastSales

DateCreated	DateUpdated	ReportAdmin	TransId	TotalSales
01Jan2001	31Dec2001	138	2020	25,000
01Jan2001	31Jan2002	138	2020	12,500

NCSales

DateCreated	DateUpdated	ReportAdmin	TotalSales
01Jan2001	31Dec2001	138	10,000
01Jan2001	31Jan2002	138	5,000

VASales

DateCreated	DateUpdated	ReportAdmin	TotalSales
01Jan2001	31Dec2001	138	15,000
01Jan2001	31Jan2002	138	7,500

Personnel

ReportAdmin	LName	FName	Extension	SSN
137	Jones	Laura	802	123-45-6789
138	Smith	John	911	111-11-1111
...				

Transformation

TransId	WH#	Region	Variables
2020	001	SouthEast	NCSales + VASales
2021	002	SouthWest	TXSales
...			

Warehouse

WH#	Service	Address	State	TransId
001	SouthEast	100 Main St.	Colorado	2020
002	SouthWest	100 Main St.	Colorado	2021
003	SouthEast	10 Sun Dr.	Florida	2022
...				

Figure C

Metadata is carefully defined to provide the smartest information, which will allow you the make the most out of understanding your data. For example, *DateCreated* and *DateUpdated* attributes are established to answer when reports were defined and last generated, identifiers are established to create uniqueness throughout the metadata, and descriptions are included to define data items such as units of measure.

There are two questions you decide to ask to determine the accuracy of this report: “When was the report created?” and “Who is responsible for creating the report?” Maybe the report was generated several days or weeks ago. Depending on the reporting schedule, this information could be obsolete. Perhaps the data lineage information reveals that more than one table is involved and an error has occurred in one of these tables. There is also a chance that an error occurred during the actual generation of the report.

Understanding the data lineage can be the key to surfacing and solving business problems. Figure D illustrates the data lineage for the *SouthEastSales* metadata table. In this case the “SouthEast Region” report in Figure B was created using metadata captured in the *SouthEastSales* table. The *SouthEastSales* table calculates its *TotalSales* value by obtaining the formula from the Transformation table, which combines the *TotalSales* reported in the *NCSales* and *VASales* tables.

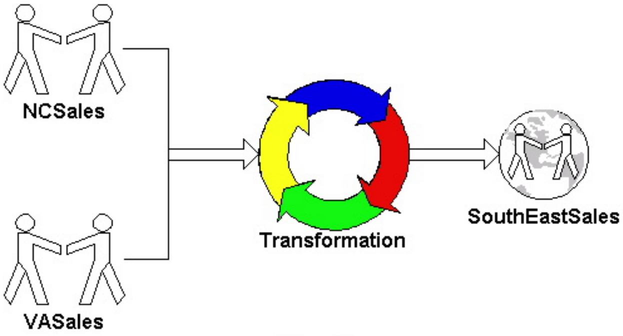
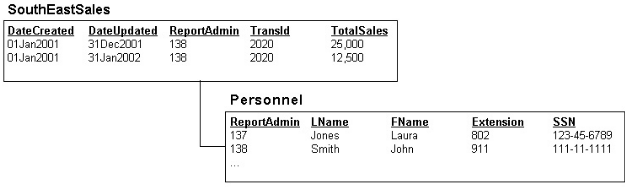


Figure D

As the metadata is studied in detail, you determine the report was generated within the last few days. You were able to determine this from the *DateUpdated* attribute in the *SouthEastSales* table. Since there are no obvious problems with the date, you attempt to answer the second question required to determine the report accuracy: “Who is responsible for creating this report?” The *ReportAdmin* attribute contains an employee number for reference. Figure E is a technical view of how the tables relate to one another. A simple query to surface *Extension*, where *ReportAdmin* = 138 leads to an employee named John Smith at extension 911. Time to give John a call.



SouthEastSales

DateCreated	DateUpdated	ReportAdmin	TransId	TotalSales
01Jan2001	31Dec2001	138	2020	25,000
01Jan2001	31Jan2002	138	2020	12,500

Personnel

ReportAdmin	LName	FName	Extension	SSN
137	Jones	Laura	802	123-45-6789
138	Smith	John	911	111-11-1111
...				

Figure E

THE PLOT THICKENS

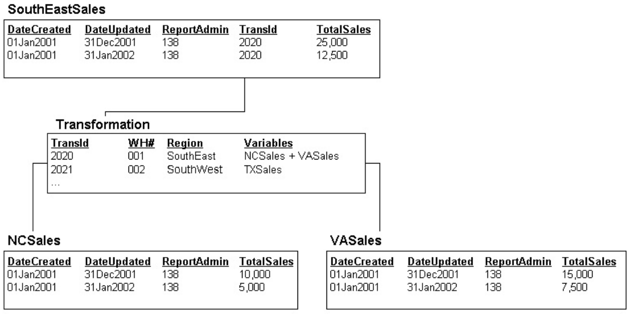
You can bet John Smith will want to quickly find out why these numbers reported for the southeast region have changed so drastically. There are many things that could have happened and it may be there is not a problem at all.

The first question John decides to ask: “Is the report accurate?” The investigation so far has proven the numbers to be recent, but this does not prove accuracy. Since our executive is not into micromanaging he is not aware of the nuts and bolts of the report. He relies on his administrators for this.

John decides there are two questions he should ask in determining accuracy: “How was the value of *TotalSales* in the “SouthEast Region” report generated?” and “Have there been any changes to the report?”. It is important to know how the value of the *TotalSales* attribute in the *SouthEastSales* table was calculated. What tables did the report retrieve information from? Is there an obvious problem with one of the tables?

It could be that one of the reports changed which in turn lead to an inaccurate calculation. This could occur with a change in a unit of measure or the definition of the unit of measure. For example, if a “package” definition has changed from one dozen to one half dozen and we continue to sell the same amount then our “quantity sold” will double. Does this mean our revenue doubles? No, and in fact it will remain the same, but this may not be obvious by glancing at the report. When these definitions change your report can take on a completely different face, which could lead to inaccurate assumptions. Defining descriptions is another example of where metadata can help you make informed business decisions.

Figure D explains the data lineage from a business perspective and Figure F explains it in a technical perspective. This view allows you to understand how the *TotalSales* calculation formula is obtained from the *Variables* attribute in the *Transformation* table. John can follow a query through the metadata and determine that the southeast region consists of sales from Virginia and North Carolina. For example, show me all *Variables* where *Region* = “SouthEast”.



SouthEastSales

DateCreated	DateUpdated	ReportAdmin	TransId	TotalSales
01Jan2001	31Dec2001	138	2020	25,000
01Jan2001	31Jan2002	138	2020	12,500

Transformation

TransId	WH#	Region	Variables
2020	001	SouthEast	NCSales + VASales
2021	002	SouthWest	TXSales
...			

NCSales

DateCreated	DateUpdated	ReportAdmin	TotalSales
01Jan2001	31Dec2001	138	10,000
01Jan2001	31Jan2002	138	5,000

VASales

DateCreated	DateUpdated	ReportAdmin	TotalSales
01Jan2001	31Dec2001	138	15,000
01Jan2001	31Jan2002	138	7,500

Figure F

CULPRIT REVEALED

John has successfully answered his question regarding data lineage: “How was the value of *TotalSales* in the “SouthEast Region” report generated?” His query result included the *NCSales* and *VASales* tables.

A close examination of these tables will answer his second question to determining accuracy: “Have there been any changes to the report?” After comparing the results of the *TotalSales* attribute he discovers that sales from both states were cut in half, with North Carolina at $5,000 and Virginia at $7,500.

The sales numbers are half of what they usually are, but there appears to be no changes in the report that would indicate a problem with the calculations. Does this mean the calculations are accurate? Well, technically yes, but there still must be a reason the numbers are lower.

The next step in the investigation involves an examination of the process involved in gathering these numbers. This is more than simply reading which attributes to add together from which tables, but rather it involves determining what other tables are effected. The business view of the data lineage in Figure G will help John focus on the “big picture” and determine where to turn next.

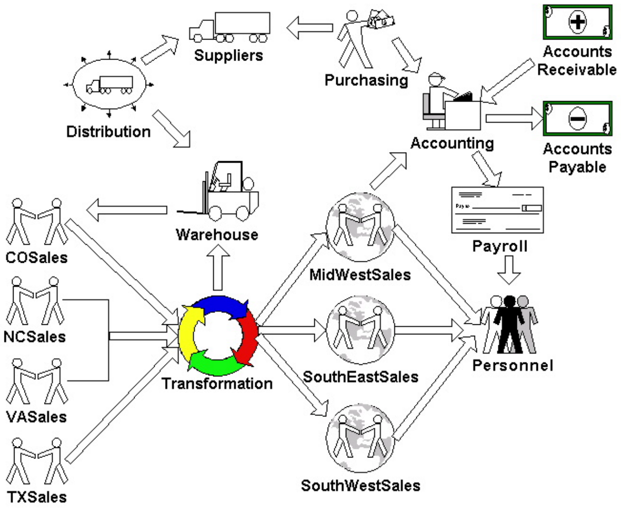
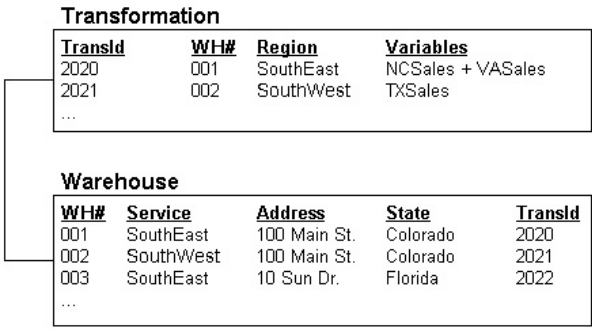


Figure G

There are a variety of spokes defined within the larger scope of the metadata. This allows John the information needed to trace the business problem directly as opposed to combing through tedious data. There is no doubt that John will have a much easier time making sense of this using the metadata that has been defined.

As John studies the data lineage, he decides to start at the root closest to his recent study. There is a spoke directly connected to the Transformation table named Warehouse. Figure H takes a closer look at the *Transformation* table. This table helps us connect the dots because it brings tables together for a final analysis.



Transformation

TransId	WH#	Region	Variables
2020	001	SouthEast	NCSales + VASales
2021	002	SouthWest	TXSales
...			

Warehouse

WH#	Service	Address	State	TransId
001	SouthEast	100 Main St.	Colorado	2020
002	SouthWest	100 Main St.	Colorado	2021
003	SouthEast	10 Sun Dr.	Florida	2022
...				

Figure H

Notice the *WH#* attribute in the *Transformation* table. This tells us where our supplies are coming from for the *Region*. If we follow this link to the *Warehouse* table, we can see that *WH#* 001 corresponds to the warehouse located in Colorado. This distance may not be too unusual depending on the total number of warehouses this corporation owns, but John happens to know that there is a warehouse located in Florida because he issued a query to retrieve all *WH#* where *Service* = “SouthEast”.

THE VERDICT

John decides to call the warehouse in Colorado to see if he can find out any additional information. As a result, he finds out that Colorado ships every two weeks. During the second week of January they had a terrible blizzard, which prevented them from shipping until the end of the month. Therefore, the supplies were cut in half, along with sales. This explains the numbers, but what about the business decision in the warehouse shipping location.

Obviously John does not call the shots around here, but he does think it seems strange to have a warehouse shipping products from Colorado when they could be coming from Florida. After taking his findings to his manager, he learns that the records were never updated when the new warehouse was opened in Florida just last summer.

THE TECHNICAL DECISION

As an executive, you used the technical metadata first to determine when the *SouthEast Region* report was generated. After determining the report was created within an expected time frame, you then used the technical metadata to locate John, the person responsible for administration of the report.

John investigated the technical data lineage information and used it to determine when the report was generated, when it was last updated, and how the numbers were combined to produce a *TotalSales* indicator. After careful examination he determined that the numbers reported for *TotalSales* in the *SouthEastSales* table were, in fact, correct. The use of metadata allowed him to bypass the tedious job of combing through data and instead produce faster, more focused search and retrieval.

THE BUSINESS DECISION

The business side of the data lineage allowed John to see how the data related to each other. This information led to a “big picture” view of the metadata. In turn a real business problem was surfaced that could be addressed immediately. Metadata enabled the business decision to update the *Transformation* table *WH#* attribute with 003, corresponding to the Florida warehouse, for the southeast region. Therefore, the Florida warehouse would replace the Colorado warehouse and begin shipping supplies to the North Carolina and Virginia areas. The result saves the company quite a bit of money related to shipping costs and simply makes better sense. By the way, John got a raise.

CONCLUSION

Metadata is “data about data” and is an essential component to any business that wants to be successful. The only way a company can stay ahead of its competition is by working smarter with their data. Working smarter can be accomplished with better understanding and manageability, allowing you to make sense of your data. An increase in data does not always benefit your company because it is useless if someone in your company cannot connect the dots and create a big picture. Metadata is a tool that will allow you to augment and generate results that can help to determine the root of a problem.

How fast would you like to get results from your data? Speed is of essence and metadata gives you the ability to make faster, more focused search and retrievals. Instead of having to comb through data tables, you will have the ability to generate reports from metadata created for data sources such as OLAP cubes, tables, and data mining models.

Today companies are bombarded by data coming from all different directions. Metadata allows a universal language, which puts an end to confusion so you can concentrate on your business, instead of your data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Michelle Ryals
SAS Institute Inc.
100 SAS Campus Drive
Cary, North Carolina 27513
(919) 531-5671
michelle.ryals@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.