Surfing the Data:  
Using the World Wide Web 
to Make Every Client of 
the Data Warehouse 
a Data Expert

Charlie J. Klumpp, Financial Management Analyst  
Wm. Clifton Ivy, Systems Analyst

Purdue University  
West Lafayette  
Indiana

As Data Warehouses greatly simplify data access, the people directly using the data expand far beyond the traditional “data reporting experts”. These new data users need to learn about the data, and they need an authoritative source for information about the data and its proper use.

Metadata (data about data) need to be:
- Easily accessible from each user’s PC, *without interrupting* tasks in progress
- Presented in the user’s own terminology
- Complete enough to answer the majority of data questions
- Linked so well that both novice and experienced users can quickly find what they need

Holding metadata centrally makes it easy to maintain, ensuring that everyone sees the same current information. The hypertext capabilities of the web allow users to hot-key from indices to definitions to illustrations.

This paper illustrates Purdue University’s metadata structure, the data within that structure, and how source data are stored to optimize usability *and* maintainability.
When the project charter was drafted to create Purdue’s management decision support system, the focus was the University’s #1 Client Request:

Provide more management data and make it easy to use

Analysis of Purdue’s management data needs revealed:

- Most people did not know what data were (or could be) available
- The data in the source files did not match the reporting needs (key data were missing or granularity not right)
- Data information was too broad and complex -- it took too long to “master the data”
- Mainframe reporting tools did not meet user needs; conversely, customers were comfortable with PC tools

Even in the initial design stage it was apparent that a quality management decision support system (DSS) would be comprised of two complementary elements:

- Quality data
- Quality metadata (information about the data).

Everyone agreed that management data need to be easy to use. Many potential users of management data were unfamiliar with the client-server concept -- but they certainly could relate to the client-service concept!

The DSS design team held a series of information-gathering sessions where clients were encouraged to identify their information needs and to describe how they would use management information. It turned out that to clients “easy to use” didn’t necessarily mean particular software features -- “easy to use” meant not having to learn another application -- not having to remember another password -- not having to deal with another icon...or window...or manual.

Participants in the Project

The ten-person DSS design team was drawn from technical experts (Network and ORACLE specialists, data analysts, programmers) and business analysts from the client side of the house (selected for their experience and knowledge of the data). All but one of these people were committed full-time to the DSS project.

The initial target audience for management data was the academic area business managers and their support staff, and the managers and clerks from Purdue’s central business office. Five business managers from the academic areas and seven data experts from various central offices each spent 5-20 hours per week helping identify, define and test the data and metadata. Weekly working meetings kept all participants updated and provided a forum for resolution of questions and problems.

The participant mix reflected the proposed client base: Some experienced experts and some data novices. Most importantly, it was (and continues to be) the users of the data who determine what is good, what is usable -- what is easy.

Design Considerations

The team kept coming back to the issues identified in the customer needs survey: The data need to work for clients who “know what they need to know” (but not where to find the answers), as well as for clients who are already very experienced data users.
Three key factors predicated the metadata design:

- The metadata need to be accessible to all staff and accessible at all times, including while the data are being queried
- The metadata need to be quickly and easily maintained by non-technical staff
- Good resource information stored in many places is not very usable; staff need one place to go for business rules, business facts and all related business information

The selected alternative was to place the metadata on the World Wide Web (WWW). LAN (local area network) connections to the University’s fiber-optic network allow the information to be easily accessed from any work station on any of Purdue’s campuses. Holding the metadata centrally makes it easy to maintain, and ensures that everyone is simultaneously seeing the same current information. Best of all for the clients, all their reference data are accessed one way, using one icon. (The web browser icon functions as the desktop equivalent of the public library “Reference Section” sign -- it identifies one spot where they can find all sorts of useful facts, definitions, and how-to’s.)

**Determining the Metadata Structure**

In part the metadata structure is defined by the database it describes. Purdue’s Decision Support System is a multi-dimensional database of fact and dimension tables which are presented to the clients in logical data models.

Clients need to know what data models are available, and what data elements are contained in each table in the model. They also need to know which data tables can and should be used together to answer management questions.

To accommodate these needs Purdue's DSS metadata are organized to fit the way clients think about and use the data. The “doorway” to the metadata is a DSS home page. This home page is the directory of all information about using DSS, from how to request access, training materials and sample queries, to a list of the DSS steering team. This page also has links to the top-level DSS metadata indexes.

Due to space and printing limitations, the sample web pages may not be fully readable. The web pages are available at “http://www.adpc.purdue.edu/DSS/”.

The indexes are the Tables Index, the Fields Index by Common Name, and the Fields Index by Oracle Name.
Tables Index

This index describes the two types of tables used in multidimensional modeling, Fact tables and Dimension tables, and lists the tables.

Each entry contains the table name, description, and a status alert if the table is being modified or is still being developed. The table name is a link to more detailed information about the table.

Fields Index by Common Name

This index lists the fields in alphabetic sequence by the common name of the field. Each entry contains the common name, the name of the table in which the field is found, a description of the field, and a status alert if the field is being modified or is still being developed.

The field name is a link to more detailed information about the field. There are letter pointers to allow the client to instantly jump to any point in the alphabetic listing.
Fields Index by Oracle Name

This index lists fields by the Oracle (database) name of each field. These names are usually abbreviated versions of the common name, and this is the name displayed by the Brio query tool.

The entries in this index are like those in the common name index, with the Oracle name substituted for the common name.

The Oracle name links to detailed information for the field.

The indexes provide clients with the ease of an encyclopedia index -- without the hassles of juggling a stack of heavy reference books. The World Wide Web is ideal for just such a presentation. With a single mouse click clients can jump to a data definition, drill down for greater detail, or jump back to check for alternative data sources.

Surfing the Data

Clients typically approach the data through the metadata. A client asked to check into strangely varying account balances might go to the index of tables, and display the information about the Account Balance table. They would see:

Table Detail

This page shows all the information stored for a table, including description, definition, typical business uses of the table, location of summarized or more detailed data, refresh rate, table size, the staff person to contact with questions or comments, and security information.

Links give a list of fields, external references (a picture of the model), and other related tables.

Links are also on every page to jump to the table index, both field indexes, and the home page.

The client might then click on the “Field list for table: ACCOUNT BALANCE” link to see what data fields are available:
Fields for a table

This list shows the fields in the sequence they are displayed by the Brio query tool. In some cases, this is alphabetic sequence, but many “most used” fields were moved to the top of the lists to make them easier to find.

The list shows the field’s common name, description, and a status alert. The field name is a link to detailed information about the field.

Creating and running a simple query on the Account Balance table shows that the balances on a few specific accounts do vary in a regular manner. Returning to the metadata and the table detail for Account Balance, the client clicks on the “Supporting Dimensions” link to see what other tables contain related information.

Related Tables

This page lists the other tables in DSS that contain logically related data. The page is tailored by type of table (fact or dimension).

For each related table, the list shows the table name, description, and status alert. The table name is a link to detailed information for that table.

This information is also available graphically by clicking on the “External References” link.
External References for Tables

This page provides a list of links to other documents that help describe a table.

The first use for this has been to link to a graphical picture of the data model for this table.

Data Model

This is a typical diagram. It shows the data model for the Account Balance fact table. The fact table is in the center, and the supporting dimension tables are arrayed around it. The field names shown are Account Balance fields used to join to the dimension tables.

Checking the related tables, the client wonders if the accounts whose balances vary regularly might have common account characteristics. Looking at the fields in the Account Number table, the client notices the “Pay Method” field, and decides to see what information it contains.
Detailed Field Information

This page shows all information stored for a field, including description, definition, oracle name, originating document, whether the field is indexed, format, and source file.

This information is tailored by table, to accommodate differences when the same field means different things in different tables.

Links are provided to go to the list of all fields in this table, the detailed information for the table, and the common links to the indexes and the home page.

Seeing that this field describes the terms in which a sponsor transfers funds to the University, and noting a link to the list of payment methods, the client clicks on the link.

Ad Hoc links

A link can be inserted into virtually any text description field within the metadata, and that link can be to any URL supported by the browser.

In many cases, these are to hypertext documents that describe code values in plain English. Since these documents may be “linked-to” from various places, they contain instructions to use browser capability to return to the earlier page.

The client adds the Pay Method field from the Account Number table to the simple query that was created earlier. Running the query again shows that all of the accounts in question have payment methods that indicate quarterly payments, and quarterly payments are the reason that the balances of these accounts to vary on a regular basis.
Deciding where to Store Metadata

A database query tool selection process was conducted concurrent with the DSS project. Part of that process examined the capability of each tool to store and display metadata. None of the tools reviewed provided as broad a definition of metadata as it was evident was needed, and all of them restricted access to the metadata to their tool. For these reasons, other storage and display options were examined.

The selected alternative placed the metadata on the World Wide Web (WWW). Separating the metadata from the query tool allows maximum flexibility and ease of use. The metadata can be viewed, paged, scrolled, and linked to answer client’s questions -- **and can be used along with or separate from any query tool**. LAN (local area network) connections to the University’s fiber-optic network allow the information to be easily accessed from any workstation on any of Purdue’s campuses. Holding the metadata centrally makes it easy to maintain, and ensures that everyone is seeing the same current information.

The wisdom of this choice became even more apparent as the use of the metadata became broader than just support of DSS queries. The metadata quickly became a definitive source of information for all data applications. Never before had source file data elements been so clearly and completely defined, nor had data definitions been accessible to such a broad client base.

Purdue is on its third generation of providing metadata on the web. All three embody the same design concepts, but as newer tools became available we were able to better realize the design vision. The key consideration in moving from one generation to the next was to make it simpler to maintain and present the metadata.

In the first generation, multiple hypertext documents were created by doing a one-way conversion of existing word-processing documents that described the data. These provided access and redundant display, but were *never* updated due to the difficulty of finding and updating all iterations of repeated data.

The second generation stored the source metadata in Microsoft Word documents, in table form. Microsoft’s Internet Assistant allowed us to create web documents directly in Word, and macros were written to create the various interconnected types of hypertext documents. Six macros were written to create the five different type documents. This solution worked quite well until the list of fields reached nearly 1000, and then we started hitting some Word memory limitations. These macros are a good illustration of a low-tech method of creating interconnected web pages. Copies of the macros are available by contacting either of the presenters.

**Maintaining the Metadata (or is it MEGAdata?)**

The third generation stores the source metadata in Oracle tables, and uses Oracle’s Web Server tools to gather and present the data as the client clicks on a link.

Each of the pages shown above are created by a stored Oracle procedure that combines a query with Hypertext Markup Language (HTML) formatting. The query can include any data from any Oracle table, although these are limited to the tables of metadata information. Formatting can include boilerplate paragraphs in combination with metadata from the Oracle tables. The metadata formatting creates the links to interconnect the pages.
Update the Oracle Metadata Tables

The data in the Oracle tables are maintained by a Microsoft Access update application. MS Access attaches to the Oracle tables through an Open Database Connectivity (ODBC) connection, and uses Access forms to simplify the presentation of the metadata in an updateable format. Audits and edits within the forms enforce data integrity, and make it easy to enter the correct information.

Access to data that are changed by the MS Access update application is as quick as clicking the “reload” button on your browser.

So Much More than Just a Dictionary....

As it turned out, Purdue’s Decision Support System metadata has become much more than just a dictionary or encyclopedia to support management reporting from the DSS database.

The information stored in Purdue’s metadata is every bit as valuable and useful as the management data it describes. DSS metadata can be profitably used by clients who do not even access the DSS system. In many instances, this is the only public-accessible source for complex business rules and plain-english interpretations of cryptic codes. For example, Purdue’s staff position codes are an intimidating melange of numbers and letters which defy interpretation by even seasoned veterans. By using the DSS metadata, clients can quickly and easily discover how to determine a staff member’s rank, voting status or tenure status -- an effort which previously required an extensive apprenticeship in arcane decryption techniques. (The sort of stuff that has never been written down -- or at least not published!)

So what is this batch of information we call Purdue’s DSS metadata? Well, it certainly is a dictionary or encyclopedia of data about data...and of Purdue’s business rules....with dynamic links which allow drilling down for greater detail or exploring related topics.

Perhaps best of all, the DSS metadata have been discovered to be a wonderful training tool. The way we approach teaching Purdue staff what data are available and how to use the data begins with the metadata. Time and again we hear, “I didn’t know we had that information” and “So that’s what that means!” -- oftentimes, from staff who have been trying (with limited success) to use these data for years. The first groups of clients to be trained devised training for all the groups which followed them. In every case, training begins with and is centered around the metadata. We regard this as a strong validation of the metadata concept and design -- and acknowledge that the combined efforts of the DSS team and our very hard-working clients has resulted in a more useful product than we had ever envisioned.

The keystone to the amazingly rapid acceptance and utilization of Purdue’s management decision support data has been accessibility and ease of use. Using the World Wide Web as the delivery mechanism provides staff with a continuously current, homogenous, indexed, linked business resource in the same place and accessible in the same way as their other business office reference tools: Dictionaries, thesari, the Indiana Statutes, federal laws and regulations -- all are accessible using the same easy-to-use tool. Each reference source is a bookmark, and the DSS metadata bookmarks are among the most-utilized references in Purdue’s electronic library.
Purdue’s Administrative Computing Master Plan identifies data access as the University’s #1 management need. In October of 1993, 80 interviews were conducted, involving 116 managers representing all areas of the University. The results of that survey determined which data should be provided, to whom, and how the data were needed. (Access to Data Problem Summary, Wilson, et al., 1993)

Fact tables are tables of recurring business events, such as Accounting Transactions or Payroll Charges. The Data Models listed in Brio are named for the Fact table which is central to that model. Dimension tables are tables of multiple-use definitional information, such as Account Numbers. The same Account Number information is useful whether you are looking at accounting transactions, payroll charges, or account balances. These tables are included on Brio Data Models wherever they apply.

Training prerequisites include mastery of Windows; training in use of BRIO, Purdue’s selected query tool; familiarity with use of the Internet, and ability to utilize file transfer protocols. Training classes for prerequisite skills are offered centrally and by training specialists in the various areas and schools.

Roll out of Purdue’s DSS and associated metadata began in August of 1996. After only three months there were 212 staff trained and operating as active DSS clients, with all training sessions through December fully enrolled. (As of publication date, January enrollment had not been opened.)

Within the first three months of deployment there were active DSS clients in every business and academic area of the university, and on each regional campus.