**Proceedings of the 6th Annual
Federal Depository Library Conference**

April 14-17, 1997

Managing the Depository Database: Some Opportunities with Shared Technology

Part I: GPRD-Institutional and Statewide Benefits of an Internet Accessible Relational Database

Nan Myers, Wichita State University
Wichita, KS

Good afternoon. Today I am sharing a session titled "Managing the Federal Depository Database: Some Opportunities with Shared Technology." My co-presenter is Tom Tyler, of the University of Denver. I am Nan Myers, Government Documents Librarian and Cataloger, at Wichita State University in Wichita, Kansas.

Part I:

GPRD - Institutional and Statewide Benefits of an Internet-Accessible Relational Database

My portion of the presentation, "GPRD - Institutional and Statewide Benefits of an Internet-Accessible Relational Database," showcases a relational database called "jeopardy," which we are developing at Wichita State as a tool for processing Federal documents. The acronym GPRD stands for Government Documents Processing Relational Database. The logo is a modified GPO eagle.

For the past eighteen months, I have worked with a colleague at Wichita State, John Williams, who is Acquisitions Manager, on the development of GPRD. During that time, GPRD has evolved from flat files, or a simple data warehouse, into a complete relational database. Now, GPRD is an information management system with data mining programs, which provide a platform for report generation and a decision-support system. In addition, GPRD will be both LAN and Internet accessible.

In Part II of this presentation, Tom Tyler will discuss his BDLD (Basic Depository Library Documents) Web site and its relationship to depository libraries and the GPO. The common theme of the two parts of our presentation is "shared technology" and the formation of institutional relationships.

I think we can agree that "sharing" is one of our best survival tactics in the documents environment today. The electronic transition in Government publishing has impacted us all significantly. It has produced substantial change without a well-planned infrastructure. For several years our needs for state-of-the-art technology for processing and control of documents have been running ahead of standardized product development. Since most depository libraries appear to be searching for similar solutions, and duplicative efforts waste time and money, shared technology can be a salvation.

In fact, the documents community has a strong track record for sharing, especially with regard to reference access and bibliographic tools. Now, the rapid development of electronic versions of Federal Depository Library Program publications, such as the List of Classes, has encouraged depository site development of net-based processing utilities (such as GPRD) and adaptation of electronic documents for re-publication on the World Wide Web, through value-added Web sites such as BDLD. In addition, the GPO is making strides with its partnership program development concept. For example, the Shipping List Site, which assists us with shipping list processing and label production, is at a Web site supported by UT/Arlington and SUNY/Buffalo.

To set the stage for GPRD, I will provide an overview of Government documents at Wichita State University (WSU), which has been a United States Federal Government documents depository since 1903. Wichita State is a metropolitan university, situated in the largest city in Kansas, and it has a student body of 14,000. WSU's Ablah Library is a moderately large research library and selects about 60% of the Government Printing Office output. I have listed total holdings on the slide for your information. The library has also been a full Patent and Trademark Depository since 1991.

In 1994, the Wichita State University Library embarked on a plan to increase access to the Government Documents Collection. Elements of the plan:

1) Load over 200,000 records for post-1976 imprints into LUIS, the library's on-line public access catalog, and provide for ongoing maintenance (June 1994);

2) Eliminate duplicative services between departments by mainstreaming Government documents into the automated materials flow of Technical Services (April 1995);

3) Eliminate labor-intensive tasks by developing in-house technology (fall 1995 to present); and

4) Develop a two-librarian model for documents, which would allow for excellence in both public access and technical services management of the collection. (began November 1996)

My program today focuses on item number three: "Eliminate labor-intensive tasks by developing in-house technology." When we moved documents processing to Technical Services, we realized that our Library Management System, NOTIS, did not provide a sufficient platform to manage all the record-keeping tasks mandated by the Superintendent of Documents, as we made this transition from a paper-based system to an electronic management system.

At that time, we initially identified three needs:

> **First, we needed to warehouse map holdings**. The map collection had never been shelflisted and was not profiled for records. For ten years, maps had been listed on ledger sheets at receipt. Prior to that, map records were non-existent.
>
> **Second, we wanted to track unresolved claims and rainchecks** electronically, but outside of the NOTIS LSER system for periodicals. Most of our documents periodicals are checked in and barcoded through LSER; but we do not use the LSER claim module, due to the unpredictability of receipt for documents.
>
> **And third, we hoped to provide easy storage and retrieval of item file information.**

Therefore, GPRD was initially developed to archive Government documents processing information, with a check-in module for maps. The information was warehoused into a series of flat file databases in Paradox 7.0 for Windows 95 application.

However, we quickly perceived the potential of Paradox, not only to manage cumbersome tasks which had previously been handled manually, but also to provide Technical Services and Government Documents with a complete relational database. GPRD now provides a platform for report generation and a decision-support system for collection development. Let me explain how the evolution, which is ongoing, has taken place to date.

Overview

GPRD initially consisted of information in tables, or files, most of which were imported from legacy systems--our own NOTIS database and others. Development continued such that information in these tables (essentially a data warehouse), could be associated using a query generator (Paradox's Query by Example), to achieve a discovery-driven data mining system.

The terms "discovery-driven" and "data mining" are buzz words in the relational database field, but I think they are very descriptive. "Data mining" is the process of extracting valid and comprehensible information from large databases and using it to make crucial business decisions. The term "discovery driven" refers to the quest for unknown information or patterns which can be used for day-to-day tactical decisions, as well as for long-term strategic planning and forecasting. I have provided a list of suggested introductory reading on this topic among my handouts.

The **platform** for GPRD is Borland/Corel's Paradox 7 for Windows 95. Our library's administration has provided its employees with technical support for the Borland/Corel Office Professional 7 Suite, which includes 32 bit versions of Paradox, WordPerfect, Quattro Pro, Netscape Navigator, Presentation and GroupWise. All applications are Windows 95/NT compatible, and all software is Internet-enabled. Two **computer languages** have been used for the programming of GPRD: Paradox Application Language to create the forms (or views) from the tables, and Borland's C++ Compiler for Windows 95 for the data mining features. In the future, the database will be mounted on a Novell NetWare 4.1 LAN. **Hardware requirements** for the database are at least a 486/66 PC with a large hard drive and 16 megs of RAM. **Costs** for hardware and software should be under $2,000 with academic discounts.

Functionality

The database itself consists of nine tables which are very simply designed. As you view the tables, you will notice that the primary key for the database, the common denominator of all the tables, is the Item Number **and** SuDocs Number pair. A compound key was selected as an unique identifier.

I'm going to summarize the components of each table, but only toggle over into the first one in GPRD, due to time constraints. The forms which I will show you later in my presentation contain the information in the tables and are more interesting.

The main item table contains:

1. Main Item Table: Item Number/SuDocs Number

Title and Agency

Active and/or Selected

2. Format Table: Item Number/SuDocs Number

Format Type: P, MF, E

Format has its own table, because format constitutes a many to one relationship. That is, many formats can be attributed to one item number/SuDocs stem pair.

3. Library Table: Item Number/SuDocs Number

Library ID

(This table provides a State of Kansas Union Listing for Current Items selected. You will notice that the proper names of the libraries are displayed, rather than their depository numbers or their NUC symbols, which is the data available in files from the GPO and BDLD, respectively. We wrote a program to convert this information to allow quick recognition by the user.)

This is our largest table, containing 62,000 item records, because individual item records are in the database for each institution which selects that record. In other words, an item record could have as many as 18 duplicates in GPRD.

4. Maps Table: Item Number/SuDocs Number

Map Number and Date

State/Region

City/Area

Inspected

Copies

The Map Table is our only use of this database for bibliographic check-in. GPRD could be adapted for this purpose, however.

5. Rain/Claim Table: Item Number/SuDocs Number

Shipping List Number

SuDocs Extension/Piece

Title

Claim or Raincheck

Record Date

Result

6. Use Table: Item Number/SuDocs Number

Month/Year of Use

Total of Use for Month/Year

7. Subject Table: Item Number/SuDocs Number

LC Subject Headings

8. Note Field Table: Item Number/SuDocs Number

Notes from the Discontinued List

This table is new. It was added to accomodate the field for notes, which appears in the Discontinued List, but not in the List of Classes.

9. Library Key Table: Depository numbers from the State of Kansas with their proper names.

Original Implementation

In developing GPRD, we consciously incorporated as much extant information as possible. In addition to our original "three needs"-- for maps, claims, and item file--we included the State of Kansas Union Listing for current items, LC subject headings for our current items, our GPO and Marcive profiles, and external use data. Check-out data for documents at Ablah Library is unautomated and has only been maintained since 1990. This was the first time we have been able to assess this data.

The initial work, done in the fall of 1995, pre-dated the ASCII text files available online today from the Government Printing Office.

• At that time, we acquired an Item File from Margaret Mooney at UC/Riverside on diskettes for around $100.

• Then, we imported the State of Kansas Union Listing distributed to us on diskettes by the Documents Librarian at Washburn Law Library.

• Our Marcive profile was another diskette import,

• but our GPO profile was "pre-Item Lister" and was originally keyed in by a student.

• Other data keyed in included:

un-automated check-out data for the use file,

information in the map file

• Finally, to obtain an initial sample of about 400 LC subject headings from our Government document bibliographic records, a programmer from our computing center ran extract programs written in COBOL from our NOTIS system.

Utilization of the Data

We then worked to identify and utilize information hidden in these tables. The goal was to organize the information in ways that enable decision making. Requirements for this phase of the development included:

1) Integrating the captured data into task-specific VIEWS or FORMS (rather than tables) for Online Transaction Processing.

2) Extracting or mining the information contained in the integrated data using queries for Online Analysis Processing.

3) To achieve this, software had to be written. The **design** of the database was accomplished in the spring of 1996 to our specifications by a team of four graduate students directed by Professor Sunderraman of our Computer Science faculty. At Wichita State, we have been fortunate to build cooperative relationships between departments. It is another example of sharing for survival.

Task-Specific Forms

GPRD currently has six task-specific forms or views: the LIBRARY FORM, the FORMAT FORM, the MAP FORM, the RAINCLAIM FORM, the SUBJECT FORM, and the USE FORM.

I'm now going to toggle over to the database and show you the forms.

1. We call the enhanced Item File the LIBRARY FORM. The LIBRARY FORM includes and can access:

    1) Item Number

    2) SuDocs Stem

    3) Publishing Agency

    4) Title

    5) Whether the title is active in the List of Classes

    6) Whether WSU selects that title

    7) Format Information

    8) LC subject headings for that title

    9) The Kansas Union Listing for the title

    10) Usage count for the title by month and year

    11) Claims and rainchecks notations

2. The FORMAT FORM shows:

    1) Item Number

    2) SuDocs Stem

3) Format selected by our Depository

3. The MAPS FORM provides:

1) Item Number

2) SuDocs Number

3) Map Number

4) State

5) City and Region

6) Date of Map

7) Whether we have inspected the map

8) Number of copies

9) Information Space: There is information for Technical Services use on the lower portion of the screen, showing the agency name, all titles connected to this Item and SuDocs stem, whether the item is active and whether it is selected.

4. The RAINCLAIM FORM provides a date-linked log of all pieces missing from shipping lists and claimed or rainchecked. Fields include:

1) Item Number

2) SuDocs Stem

3) Shipping List Number

4) Extension of the SuDocs Number

5) Action Date

6) Whether it is a Raincheck or a Claim

7) Title

The SUBJECT FORM includes:

1) Item Number

2) SuDocs Stem

3) Title

4) Subject(s) for that title

The USE FORM shows:

1) Item Number

2) SuDocs Stem

3) Agency

4) Title

5) Date of the Use (month and year)

(All days default to the first of the month)

6) Number of Uses

Building On the Work of Others

I have already discussed the initial building blocks of our data warehouse. These were a good beginning; but in the past six months, we have replaced the data in these tables via Tom Tyler's BDLD, which is a value-added Web site, or network--often called a VAN.

BDLD as a VAN for GPRD

Last fall, I phoned Tom Tyler to compliment him on his very useful and attractive Web site, the BDLD. I was interested in his Union Listing for the State of Colorado, which Tom had extracted from a tape of depository library profiles he purchased from the GPO.

Once again, the time frame last fall was "pre-Item Lister," and depository profiles were not yet available from a GPO Web site. In addition, Washburn Law Library had abandoned updating the Kansas Union List, and I was seeking an alternative to manual keying of data. Subsequent discussions yielded a mutual agreement that GPRD could import data from the BDLD. We chose to derive data from the BDLD rather than from the flat files available from the GPO because the BDLD provides uniform data. Tyler has been committed to developing and maintaining Government document computer files, as well as scrubbing that data, for over 15 years.

Features Imported from the BDLD to Date:

1) WSU's List of Item Selections (our profile)

2) Kansas Union List (18 other Kansas depository profiles)

3) List of Classes

4. List of Discontinued Items

Implementation of Those Features at WSU:

• **The BDLD to GPRD Data Migration** (please refer to the Four Step Data Mining handouts):

1) The BDLD Web site (data as we view it)

2) BDLD HTML Source Code (data available for data mining)

3) Bar Delimited ASCII File Sample (data mined from BDLD after conversion by C++ extract program written at WSU)

4) Sample Report Output (Data after import into GPRD)

• We decided to **merge the List of Classes with the List of Discontinued Items** in order to create the largest possible number of entries of post-1976 imprints for our Cataloging Staff. An additional field was created in GPRD to accept notations on titles available in the Discontinued List.

• The **programmed refresh of the GPRD database tables** required program(s) written in Borland C++ for Windows 95 for creation of bar (|) delimited text files. There are three of these files: main item file, format file and library file. These files are extracted from the BDLD Web site, imported into Paradox, and in the process, converted into Paradox tables. We are using a temporary directory for all imports and deriving the information over to the main directory after checking it.

The BDLD **arrays the List of Classes and the Discontinued List**. We re-array the data to suit our database design, in particular 1) the format table and 2) default values on current and discontinued items. For GPRD, the format constitutes a many to one relationship, as I stated earlier in discussing creation of the original tables. That information has to be isolated into its own table. With regard to defaults, in the List of Classes, we created the default value "Yes" in the active field. In the Discontinued List, the default value is "No" in the active field. The defaults are the distinguishing feature once the two lists are merged. In other words, when you look at the list of titles, you can distinguish between active and inactive.

In order to accomplish this programming, John and I once again depended on departmental cooperation at WSU. All of the C++ programming has been done by Professor Xumin Nie of the Computer Science Department. He has made his Web site available for the **Data Miner for GPRD**. The address is: **http://riker.cs.twsu.edu:1081/~nie/GPRD/main.html**.

• Once the above tables are created, two queries are run to create an intelligible Kansas Union List of Depository Holdings. This process is detailed in a handout

titled "Data Mining with GPRD," which I have with me and would be glad to share with anyone interested after the program.

• All of the text files are converted into tables in a temporary directory. When we have verified the accuracy of the files, they are derived over to the permanent project directory.

Report Generation

An extremely useful aspect of GPRD is report generation. As I have stated, GPRD provides a platform that allows all the diverse data in its tables to be quickly extracted and associated. The staff can initiate standard, pre-programmed reports, or they can produce ad hoc reports using Query By Example (QBE).

There are three types of reports we plan to generate regularly:

1. **Map Reports**: Reports can be created for our topographic map holdings by state or by SuDocs stem. These reports can be exported to QuattroPro and laser printed for use at the map cases.

2. **Claims and Raincheck Reports**: Information on shipping list claims and rainchecks is warehoused for periodic retrieval and review. At intervals, we print a report to forward to the Documents Information Librarian, so she can review missing titles. An action box to the right of each title allows her to request: 1) acquiring the title through Interlibrary Loan, 2) purchasing the title, or 3) disregarding the title.

3. **Collection Management Reports**: Because of the space limitations in the Government documents area of Ablah Library, this type of reporting may play a crucial role in collection management decisions. At Wichita State, documents now occupy 86% of their allocated space. We have been told that we should not expect additional space for 15 years; therefore, we must look seriously at zero physical growth.

GPRD can provide:

A) **Subject Analysis Reports**: Several subject analysis reports can be used for decision support.

1. A simple subject analysis of selected/active items can indicate whether those items support the curriculum.

2. The addition of subject headings to use report, will correlate subject distributions for high, low, and no use titles.

3. Agency distributions can be aggregated within subjects and use.

B) **Usage Reports**: GPRD contains external use data from January 1990 through December 1996. Titles which have circulated can be reported in

SuDocs number order or in ascending or descending order by use tools. Use data can also be queried by format.

C) **Kansas Depository Library Reports**: Knowledge of the collections of other Kansas depositories is helpful for interlibrary loan support. We can also make demographic analyses of the distribution of titles through our state. A title in our collection with no demonstrated use, which is available in 14 other depositories in the state might be a good candidate for weeding or de-selection. In addition, the ability to associate data on statewide holdings could underwrite cooperative collection development in Kansas at a time when many of our colleagues are feeling the stresses of space, costs, staffing and accountability.

Conclusion

In summary, GPRD allows us to:

1) Reduce labor-intensive tasks in a period of reduced staffing.

2) Permits us to accomplish inspection requirements.

3) Frees us to address issues of public access in an electronic era.

4) Allows for pro-active rather than re-active decision making.

5) And, can facilitate cooperative statewide depository goal-setting.

In conclusion, I should mention that GPRD is an ongoing research project. In the near future, we plan to:

1) Link GPRD to the Internet, so that other depository libraries in Kansas can access it from a Library home page.

2) Network GPRD to the LAN so that all library employees can get to it.

3) Develop a project to study internal use of documents.

On the one hand, GPRD is a very specific product, tailored to the needs of one university. On the other hand, it represents a multitude of generic needs, and the concepts used in GPRD's development can be duplicated anywhere.

The development of GPRD represents a proactive stance during a period of uncertainty and change in the history of Federal documents. I believe that as Wichita State and other depository libraries in the same position share their resources with the GPO, a synergy will be achieved. Eventually, the GPO will provide a standardized set of tools for working on documents in the 21st century. However, the need for ingenuity in the field will not disappear. And with that thought, I will turn the program over to Tom Tyler.

Define terminology:

First, for those who are not familiar with the concept of a relational database, let me define some of the terms I will be using. A **database** is an organized collection of information or data. If you have an online address book, you have data organized about people into specific categories or tables, such as names, addresses and phone numbers.

Now, if you also have a birthday book that contains birthdays of your family and friends, and maybe some information about their clothing sizes and favorite colors, you have a second database. With two databases and two tables, you have the beginning of a **database system.**

Some database systems look at only one **table** at a time. These are called **flat file systems.** If you use this kind of system, the terms **table** and **database** mean the same thing. Using the example of the address book and the birthday book, you would be able to look at one at a time, or see the names and addresses in one book and the names and birthdays in another book. You would not be able to combine selected information from both tables.

However, with a **relational database**, such as Paradox, you can extract specific information from each table and assemble it in a meaningful way. Using the address and birthday books example again, if you wanted to see a list that includes a friend's name, address and birthday, in a relational database, you could link the address and birthday tables by identifying a common field ("Name"). Then, you would be able to combine the kinds of information you want to see from both tables.

Paradox keeps the tables in a database separate, but understands there is a relationship between them. In a relational database like Paradox, the term database means all your tables and all their relationships. Paradox requires the use of a **primary index** (or **primary key**) used in each of the tables to provide the common link.

When you begin to use a relational database, you probably will not know all of the potential relationships among the tables. But, eventually, you will discover them. You will probably have to dig for them, which is referred to as **data mining**. Data mining is the process of extracting valid, comprehensible and previously unknown information from large databases and using it to make crucial business decisions. Other terms related to this analogy to mining are: **discovery based data mining**, which refers to seeking unknown but valid information or patterns; **data exploration**, digging through large amount of historical detailed data;and **drill-down technology**, which suggests searching deeply into the information.

Another much-used term is **data warehouse**, which is simply a place to store data. The implication in warehousing is that you are managing the database in some way, probably for some kind of decision-support. A **data warehousing system** provides a complete end-to-end solution for delivering information to users. You will be able to process the stored data, transforming it into business information. You can obtain the data from your internal

systems, from external information providers, and most recently, from Web servers. The resulting information you acquire can be instrumental in making tactical decisions about day-to-day business operations and also for strategic decision-making involving longer-term planning and forecasting.