**Proceedings of the 7th Annual
Federal Depository Library Conference**

April 20 - 23, 1998

Collection Management Using the Documents Data Miner

Nan Myers, Wichita State University
Wichita, KS

Introduction

Thank you for attending our demonstration of the Documents Data Miner (DDM). I'm Nan Myers, Government Documents Librarian at Wichita State University. With me today are John Ellis, Senior Database Analyst with the National Institute for Aviation Research, and Cathy Hartman, Documents Librarian at the University of North Texas. John and I have been working on the Documents Data Miner since August 1997. Included in our effort was the development of the GPO partnership which we hope will facilitate better access to FDLP administrative information.

I will set the stage for our presentation with a little background on the Data Miner and then demonstrate the features available at the site. John will explain the technical development of the utility. And, Cathy has been generous enough to take time to critique the site over the past couple of months. She will address several collection development scenarios which employ the Data Miner in decision support.

[slide 1]

Timeline: Fall 1995 thru Spring 1998

- Benchmarks:

1. Records Tapeload (1994)
2. Mainstreaming Documents (1995)
3. Collection Decisions: RDBMS (1996)
4. GPO Partnership: Web Database (1997/98)

- Acquisition of a Repertoire of Techniques

- Development of Interactive Professional Relationships

- Open Systems Follow-On's

To briefly place the Data Miner in a research and political context at Wichita State, I prepared a slide which shows a time line overview of progress towards this Web-based tool for managing the collection development tasks we all face in depositories. When I came to WSU in 1994, my first task was to oversee a tapeload of bibliographic records for the documents collection, both retrospective holdings and currently received items. I was also charged with regularizing cataloging functions in Technical Services. Second, we mainstreamed the documents processing workload in Technical Services. And, third, we began a challenging collection management initiative:

- to downsize the physical collection,

- to emphasize electronic access, and

- to market Government documents.

At this point, what is now the Documents Data Miner began to grow out of operational imperatives at Wichita State. Basically, we needed a decision support system that would allow us to do three things:

1. eliminate labor-intensive tasks associated with documents processing and record keeping;
2. electronically warehouse item card information; and
3. facilitate our vision for Documents collection management.

As space to house the collection was critical, we wanted to intelligently downsize based on usage statistics and holdings of other Kansas depositories. Eventually, we are interested in statewide cooperative collection management.

[slide 2]

Desktop Data Mining

- Fast, Accurate Data Acquisition
- Reliable Data Updating at Regular Intervals
- Simple, Comprehensive Application Platform
- Utilitarian Data Extraction
- Speed
- Reliability: Stamina of the System

In the fall of 1995, we began developing an in-house relational database, which we called GPRD, for Government documents Processing Relational Database. GPRD was an Information Management System for our depository complete with data mining programs. It provided a report generation platform, and it was a decision support system.

We are all familiar with data mining, unfortunately. When our credit card providers sell our purchasing history to marketers and we receive unwanted third class mail and salespeople call during dinner, we know we're being mined. A glance at the slide behind me will indicate the characteristics of desktop data mining.

[slide 3]

Federal Documents Metadata

- List of Classes
- Inactive and Discontinued Items
- Selection Profiles of Depository Libraries
- Other Data as May Arise

The application of this same methodology to Federal Documents metadata (the equivalent to a purchasing history) is focused, for us, upon electronic versions of the List of Classes, inactive and discontinued items, and selection profiles of depository libraries. In the future, we may add other databases to the DDM.

We began prototyping our mining operation last spring from Tom Tyler's BDLD Web site using C++ programs run on a Unix server. From the BDLD we mined the List of Classes, the Kansas Union List (which Tom had put up for us) and inactive/discontinued information. The mined data was housed in Paradox 7.0 and was extracted using Structured Query Language.

This work confirmed the viability of our techniques and, because of the cooperation and interest of faculty and students in our Computer Science and EE Departments, the costs were affordable.

To me, GPRD pointed to a national need for a standardized set of electronic tools for working on documents. Our primary goal by last summer was to meet the need of our community for a powerful search engine which could associate various databases of administrative information from the FDLP.

[slide 4]

Documents Web Tools

1. Searchable:

   List of Classes

   Inactive/Discontinued List

2. LOC/Item Lister Merger
3. Regional Union Listing
4. Scalability

I wanted the following:

- A searchable List of Classes.

- Information on Inactive and Discontinued titles at my fingertips.

- Data from the List of Classes associated with the items in our profile from the Item Lister.

- Item selections for the other Kansas depositories and items selected by those depositories within a 100-mile radius of Wichita.

- And, a platform that would grow with the changing documents environment.

Thus, last summer, the Documents Data Miner was born. Moving the techniques we had developed from a desktop to the Web involved much more than differences in scale, however. In order to rapidly implement the development of the Web site, we were fortunate to be able to capitalize on the expertise of colleagues in NIAR, the National Institute for Aviation Research.

[slide 5] **[** Large screen shot - 70 kb gif file **]**

Initially, NIAR was willing to lease space on their server to us for the Data Miner project. However, John became interested in the complexities of prototyping a database for a large body of metadata because of similarities to some other projects he was working on. In exchange for a nominal monthly storage fee, they have contributed time, expertise, hardware and software. If properly billed to us, the Data Miner would have cost in excess of $25,000!

In addition, without the cooperation of the GPO, we would not have had access to all the official data that we wanted. We began to explore a partnership relationship last fall and, at our request, GPO made files available in a custom format for FTP. In addition, they have agreed to publish the Inactive or Discontinued Items list on the Web in the near future.

There were other, technical challenges the Library could not have singly addressed, but John will discuss those in a few minutes.

[slide 6]

Web Data Mining

- Official Source of Data (FDLP)
- Very Large, Fast Server
- Mirroring and Security/User Profiling
- User Friendly Data Extraction
- Sophisticated Platform
- National Utility

Turning to the next slide, now, I would like to add the following about Web data mining:

- Our official source of data was the FDLP.

- Server requirements eventually reached 1 Gigabyte of storage and the requirements for higher memory became critical, because of the size of some of the tables.

- The partnership raised mirroring and backup as central issues and informed our wish to know how our users were working the site.

- The GPO partnership, obviously, was the only relationship that would justify the expense of the site.

[slide 7]

DDM: Development Goals

- Searchable List of Classes
- Searchable Inactive/Discontinued Items
- Collection Profiling
- Build/Downsize w/State/Region Profiling
- Easy Export of Query/Profile Results
- Directory/E-mail access
- In-House Functionality w/Exported Tables
- Open System Follow-On Development

The primary goal of the Data Miner development is selection/deselection decision support. The handouts given to you show eight criteria for meeting that goal. In addition to the first four, we projected that users would want to export the information from profile queries in order to build in-house databases. We further decided to include the information from the depository directory with e-mail access for convenient communication among depositories. And, we designed DDM as an open system, allowing for rapid redesign and modification.

Demonstration

I'm going to move now to a demonstration of the main features of the DDM.

First, requirements for use of the Documents Data Miner are:

1. Netscape or Internet Explorer at 3.0 or higher.
2. Browsers must support frames.
3. Cookies must be turned on.
4. JavaScript must be enabled.

There are two ways to navigate in the DDM. The home screen offers the four key tasks performed by the DDM, and these are also accessible at the frame on the top. The frame also offers an Introduction, Support screen, and Home. We determined to have a frame in the overall design:

- For ease of use
- For quickness of use
- To eliminate having to back out of a series of pages

If you read the Introduction, you will learn that the DDM is a search engine combining files from the latest version of three databases published by the FDLP: The List of Classes, Current Item Number Profiles for Depository Libraries, and the Federal Depositories Library Directory. A fourth title, Inactive or Discontinued Items, is not yet available online from the FDLP. However, we have mined this information from an external source (the BDLD). The latest date that the files have been refreshed on the DDM is built into the frame. Our intention is to update files on the same day they are updated at the FDLP sites, which is usually the first Friday of each month.

Button 1 - List of Classes:

Returning to the home page, I want to start the demo from the List of Classes. From this screen, you can:

- Search the current List of Classes by field,
- Or, search the Inactive/Discontinued items by field,
- Or, merge the searches by choosing "all" at the "Status" box.

Clicking on this feature brings up a search grid with the following choices:

- **Agency:** You may search "all," or the pop-up box provides the list of agencies with the sum of active item number stems for each agency.
- **Item Number:** You may enter a full item number, which requires exact spacing and punctuation, or you may enter a partial item number.
- **SuDocs Stem:** You may enter a complete or partial SuDocs stem.

  Notice that the function choice (SuDocs stem) is in blue text. Whenever you see blue text, you may click on that and obtain an example or explanation. These are the JavaScript applications that require Java to be enabled. Examples of search

possibilities for SuDocs stems are: C, C 1, or C 1.54. Spacing and punctuation must be exact, but the ending colon is not required.

- **Title Search:** You may search an exact title, or words from a title. Automatic left/right truncation is built in.

- **Format:** A drop down box allows you to choose from: Any Format, Paper, Microfiche, CD-ROM discs, Electronic, and Electronic Library. These are the formats used by the GPO.

- **Status:** You may search for either active, inactive/discontinued, or all.

Some Searches Demonstrated:

1. "Agency" and "Active": To see what is actively published by an agency, such as the FCA (Farm Credit Administration), select that agency and status "active," which is the default at the status box. Then, submit.

The results screen, called "Complete Class List," shows what we requested (FCA and active), then offers a list of results arrayed in SuDocs order. The display gives us the complete information about each item - the SuDocs stem, Item Number, Title, Format, Frequency and Status.

Additional features: If, at this point, you click on the SuDocs stem of an entry, DDM brings up a screen displaying a **list of SuDocs stems assigned to that Item Number**. Or, if you click on the Item Number, you can access the **Union List feature**. I will be discussing the Union List feature in more detail later, but I will mention that if you do not have Union List parameters set at this point, the DDM will take you to the page to set those up.

2. "Agency" and "All": Let's look now at the FCA and all of its Item Numbers, both active and inactive. Whereas we found seven active items for that agency, this search produces 12 records, indicating in the far right status column which are active and which are discontinued.

3. Title Searching: Returning to the List of Classes page, let's try a title search. In a title search, the search engine allows us to input an exact title, or use a word from the title, or make use of the automatic left/right truncation feature. Try a title search on the word "free" and we find titles for A) a Department of the Interior report to Congress about "free-roaming horses," B) the Freedom of Information Act Annual Report, C) the Freedom of Information Case List — both from the Justice Department, and D) a publication from Congress which is printed on "acid free paper."

4. The "And" Function: The Documents Data Miner allows the user to "build" selections. If we want to narrow a search, we may. Suppose we want to determine all the titles issued by the Census Bureau in CD-ROM format, we select the agency (Commerce), and input the SuDocs for the Census Bureau (which is C 3) , and select the format we want (CD ROM).

I find myself using the List of Classes search screen for a variety of purposes. Sometimes I have a mystery item number. Or, I may I want to see all the titles in a certain format -- all the CD-ROM titles, etc. Or, I may want to concentrate on Inactive/Discontinued titles.

Button 2 - Inactive or Discontinued

The Inactive/Discontinued feature is in development. As you know, the publication titled Inactive or Discontinued Items... is not yet available for downloading from the FDLP Administration page. Data available in this section was originally mined from the BDLD site. You may search this data from the List of Classes feature, or from this location.

Search Demonstrated: From this screen, you may search by Item Number, SuDocs stem, or by title.

For example, let's just check to see what is inactive from the U.S. Information Agency. For that search, we can input the partial SuDocs "IA." The results screen provides us with a list of four inactive SuDocs stems for this agency.

Additional Features: At this point, you see the unique feature for the Inactive/Discontinued component: the **NOTES** field. The annotations for the NOTES originated from Shipping Lists, the List of Classes, Surveys, the BDLD title List of Classes - Additions & Changes, and other Depository Library Program sources.

If a note is available, the word "YES" is in the box. A note can be accessed by clicking on the blue text. Clicking on the SuDocs stem provides a note for that SuDocs, and clicking on the Item Number provides all the notes available for the SuDocs stems attached to that item number.

Button 3 - Depository Selection and Directory

Returning to the Home page, the third button is called "Depository Selection and Directory." You will use this point of entry if you want to search your own depository profile. This feature merges profile data from the Item Lister with the List of Classes, allowing the user to view complete item information for the selections in their own profile. It also provides directory information for each depository and e-mail functions for ease in communicating.

The search parameters at "Depository Selection" are designed to let you search in several ways. You may:

- Enter the depository number,
- Enter an institution or library name — or a partial name,
- Search for depositories in a certain city,
- Request depositories for an entire state,

- Or, search by type of library, such as Community College Libraries, Academic Law Libraries, or State Libraries.

There is a pop-up table for states and types of libraries.

Search Demonstrated: Let's do a search for a depository number. I'll enter 0204A, which is Wichita State University's depository number, and click on submit. This presents us with a screen that allows three functions:

- Click on the Depository Name for directory information. All the fields available in the GPO database are displayed: names, addresses, phone numbers, depository type, size, year designated as a depository, and congressional district.

- Click on the E-Mail Address to send a message to the depository librarian at that institution. (Since this data is supplied to the GPO by individual depositories, you may occasionally encounter an outdated e-mail or a blank box if a library has not kept the GPO informed of up-to-date information. Corrections should be sent to the GPO.)

- Click on the Depository Number to search the profile.

Depository Profile Search: Clicking on the Depository Number 0204A brings up a screen titled "Selections for 0204A, Wichita State University, Ablah Library, Wichita, KS." The search grid is similar to that of the List of Classes, with fields for Agency, Item Number, SuDocs Stem, Title, Format and Status. However, queries will only produce results for items selected by the depository you are searching.

Additional Feature: The Agency drop-down box provides a summary of item number totals by agency for this depository. For example, Wichita State selects 230 items from the Agriculture Department and 1,347 from the Commerce Department.

I frequently want to know if our depository has selected a certain item number, so I'll do a search on 0546-D to find out if we do select that. The results screen brings up 21 titles from the General Accounting Office which are attached to that item number. If I wanted to consider de-selecting a specific title here, I would need to be sure that I could let go of all 21.

In addition, I might be better able to make a decision to de-select if I knew which other depositories in my region or state were also selecting this item number. Clicking on the Item Number in the display of the 21 titles will provide me with union list information, after I have set an appropriate filter. If I click on the Item Number without initially setting that filter, Documents Data Miner immediately sends me to the Union List Profile.

Button 4 - Union List Profile

At this point, I'll describe the Union List Feature, which is available at the last button on the Home page, as well as on the frame. Or, as I just mentioned, you will be sent there if you

are requesting union list information, but have not already set up a filter. So, let's go ahead and set that filter. Our choices are to filter by:

- State
- Region
- Or, by Distance from a Depository

Let's go ahead and choose State, which for us is Kansas, and submit. If you are setting this before beginning your work, hitting "submit" will return you to the Home page, where you can enter "Depository Selection." If you have been sent to the Union List Profile while already in "Depository Selection," Data Miner sends you right back to your search grid after you set up a filter.

If you recall, our initial query was for Item Number 0546-D, which produced a total of 21 GAO titles. Checking to see how many depositories in Kansas select that Item Number, we find a total of eight which do. Decisions to retain or deselect the item number could then be based on use of the items, depth of coverage in our area, and availability of reliable courier service. In Kansas, we have an excellent courier service, with potential 24-hour turnaround.

Now, suppose I want to refine my union list further and see how many depository libraries within 100 miles of Wichita State select this item, I return to the Union List Profile and change the filter to 100 miles radius at the "Distance from a Depository" option. Hitting "submit" returns me to my search grid, and I once again click on the item number. This time I find that only one other depository within a 100 mile radius of Wichita selects this item, and that depository is in Oklahoma.

The Union List feature is designed to assist depositories with building and/or downsizing their collections. The addition of a Gazetteer to the tables of the Data Miner allows the user to customize a group of depository profiles and extract selection information on item numbers. For libraries which already have consortial collecting agreements, I hope this utility will very helpful. For those who WANT cooperative collection development, but thought it was too much work, I hope this tool will get you started.

Cathy will talk more about the uses of the Union List function for regional collecting in her presentation, so this concludes my portion of the presentation. I'll turn this over to John Ellis, who will tell you about the architecture and functionality of the Documents Data Miner.