Documents Data Miner 2©: Demonstration of a Pilot Project: http://govdoc.wichita.edu/ddm2

Nan Myers
Wichita State University Libraries
Wichita, KS

Development and System Overview

What is Documents Data Miner 2?

Documents Data Miner 2 (DDM2) has been under development as a pilot project practically since the announcement of Documents Data Miner itself in April of 1998. In 1997, the Federal Depository Library Program began offering files of its shipping lists on the Federal Bulletin Board, but did not make them available in any searchable format.

Then, beginning with a file for December 1998, GPO Cataloging records were made available at the Federal Bulletin Board in a file called SPCMOCAT. Files of GPO's Cataloging Branch output have been posted monthly ever since, but as one long data stream requiring conversion to the MARC format. These two additional pieces of the Federal depository workload puzzle made it seem possible to provide, in one online location, a more complete Library Management System for United States Federal documents.

Documents Data Miner 2, then, is:

- A complete Library Management System for U.S. Government documents, under development as a pilot project.
- A national-level utility.
- A Web-based data mining tool.
- A multi-format tool, allowing for the processing and deployment of Federal government information both in physical and Web-based formats.
- A site with unique capabilities, such as OPAC and shelf listing functions, exportable USMARC records, and searchable categories of P/URLs.
- A utility developed as a Library/IT collaboration by the same development team as Documents Data Miner.
- A utility based on Documents Data Miner (©1998) and including all the modules of DDM itself.

Development of Documents Data Miner (DDM):

Documents Data Miner began in 1995 at the Wichita State University Libraries' Technical Services Department as an in-house relational database in Paradox, designed to support management of Federal depository library collections. Preliminary design was accomplished through a partnership between Nan Myers, Assistant Professor and Government Documents Cataloger; John Williams, Head of Acquisitions; and graduate students of the University's departments of Electrical Engineering, Decision Sciences and Computer Science.

The initial prototype for the data mining function was written by Dr. Xumin Nie, formerly Professor, WSU Computer Science Department. (For a more complete discussion of the initial database, see: Myers, Nan. "GPRD - Institutional and Statewide Benefits of an Internet-Accessible Relational Database." *Proceedings of the 6th Annual Federal Depository Library Conference, April 14-17, 1997*. Also published online at: <http://www.access.gpo.gov/su_docs/dpos/97pro.html>.

In 1997, we moved the relational database to the Internet, on server space leased from the National Institute for Aviation Research (NIAR) on the WSU campus. SQL server database implementation, query algorithms, and Web database publication were developed by two of NIAR's staff: John Ellis, Senior Database Analyst, and Dr. John Hutchinson, Professor of Mathematics and Statistics. The new utility was named Documents Data Miner, or DDM.

Documents Data Miner was built on official sources of data from the Government Printing Office (GPO) files at the Federal Bulletin Board. At this point, DDM provided an Internet-accessible relational database for the use of the government documents community (1350 Federal depository libraries and the Government Printing Office).

DDM became an official partnership site of the GPO in April 1998 and was announced at the Federal Depository Library Conference. (For a more complete overview of DDM, see: Myers, Nan. "Collection Management Using the Documents Data Miner." Ellis, John. "Architecture and Functionality of Documents Data Miner." Hartman, Cathy. "Documents Data Miner: A Resource for Collection Development and Management." *Proceedings of the 7th Annual Federal Depository Library Conference, April 20-23, 1998*. Also published online at: <http://www.access.gpo.gov/su_docs/dpos/98pro13.html>.

There are five databases in Documents Data Miner:

- List of Classes
- Government Authors Files
- Item Lister's Current Item Number Selection Profiles for Depository Libraries
- Federal Depository Libraries Directory
- Inactive or Discontinued Items List

The DDM Development Goals were as follows:

- Searchable List of Classes
- Searchable Inactive/Discontinued List
- Union Lists which could be associated with the List of Classes
- Collection Profiling Tools
- Directory and E-mail Access
- Easy Export of Tables and Query Results
- Mirroring and Security/User Profiling
- Open System Follow-Ons

The last development goal, "open system follow-ons," provided the basis for the prototype version of Documents Data Miner 2.

DDM2 Development Goals are:

- Provide export of USMARC records from GPO Cataloging (12/98 - present)
- Searchable shipping lists.
- National shelf-listing capability, recording items shipped to depositories from the GPO.
- Development of a Web-based national public access catalog to government information.
- Explore the potential to offer server space to libraries using DDM2 as their OPAC (online public access catalog).

Documents Data Miner 2 System Overview:
In this section, I will be discussing:

- Design parameters
- Current attributes
- Current data statistics
- How we add value
- Recent enhancements
- Future enhancements

The current design parameters for DDM2 are as follows:

- Use only GPO data
- Use GPO data "as is"

- Use low cost servers
- Use only Web-based clients
- Design for both Netscape and IE browsers
- Use off-the-shelf development software

GPO data is drawn from the following files at the Federal Bulletin Board at <http://fedbbs.access.gpo.gov/liblist.html>

- Profile.dbf (depository directory)
- Listclass.txt (list of classes)
- Inactlst.txt (inactive/discontinued list)
- Loga.txt (agencies)
- Ff-ddis or cd-ddis (union list)
- HMCMmmdd.txt (marc records)
- Shiplist.dbf (shipping lists)

The current attributes of DDM2 are:

- 2 dual Pentium (733 Xeon) servers
- Windows 2000 Server sp1
- Microsoft Sql Server 7.0 & 2000
- Microsoft IIS 5.0 with ASP
- 1.5 Gigabyte of storage
- University supplied network
- GPO data
- WebTrends analysis software
- Netscape and IE clients
- Maintained by 1 analyst
- A work in progress

Development software included:

- Microsoft Interdev 6.0
- Sql Server 2000
- Vbscript & Javascript (server and client)j
- ADO 2.6
- Microsoft VB 6.0 (loading updates)
- A few special libraries (aspmail)

Current Data Statistics for DDM2 as of October 12, 2001:

● Depository active	1,313
● Depository inactive	63
● List of classes active	8,534
● List of classes inactive	10,447
● Unionlist active	2,557,650
● Unionlist inactive	559,974
● Shipping lists	6,278
● Shelflist items	122,899
● MARC records	50,056
● MARC tags	1,446,708
● MARC URLs	14,215

Value added by the modules of DDM2 (which include those in DDM):

- Cost savings (USMARC records).
- Time and staff efficiencies — streamlining of workload.
- Highly customizable tool.
- Only national union catalog for Federal depository libraries.
- Searchability of FDLP databases made available for the first time.
- Federal data from various databases made available in a common format.
- Depository library data downloadable in a common format for the first time, and exportable to Excel.
- Relational integrity to FDLP databases provided for the first time.
- Provides a common data store with date-tagging.
- Union list, classlist, shipping lists can be filtered by depository.
- Virtual catalog for government information available.
- Built with off-the-shelf tools and databases, making it easy to modify.

Recent Enhancements:

- Searching by Subagencies
  - From TOOLS page
  - Requires IE 4.0 or higher
- Full-text indexing of MARC records
  - Title Key Word

Projected Enhancements to the DDM2 Pilot Project (would require grant funding):

- Virtual Catalogs—Leased server space for depository library catalogs, with automated holdings and predictive check-in.

- XML Web pages — Dynamic data

- XML exports (MARC records and DDM data)

- Sessions without cookies

- Labeling

- Resolution of fugitive government Web documents—

  - Spidering the Web from the P/URLs in GPO MARC records; developing an indexing system; developing software for creation of temporary cataloging records.

Issues of Cost Recovery

Partners in Documents Data Miner 2:

The two current partners for DDM2 are Wichita State University Libraries and University Computing and Telecommunications. Partnership with the Government Printing Office is pending. After the announcement of Documents Data Miner in April 1998, the political scenario at the WSU Libraries changed when the Dean of Libraries retired in the summer of 1998 and a year's search concluded in the hiring of a new Dean in the summer of 1999. In addition, John Ellis moved from NIAR to a position as Web Applications Manager for University Computing in January 2000. There was then a period of transition and education, negotiation with University Computing, and seeking of direction from the University's Office of Research Administration and Legal Counsel.

All discussions about completion of DDM2 have led to the need for an Oracle platform in order to provide a sound national-level utility. Since such an expense could not be absorbed by either University Libraries or University Computing, it became apparent that cost recovery would be required to complete the vision of DDM2. Below is a cost summary of both DDM and DDM2 between 1997 and 2001:

What it Cost:

● Actual costs to date	$10,000
● Unbilled Costs - Fair Market Value 1997-2000 (DDM and DDM2)	$200,000
● Unbilled Costs (DDM2) - Fair Market Value 2000-2001	$75,000
● Projected annual maintenance of current DDM2 design - Fair Market Value	$30,000
● Projected maintenance of future DDM2 design	Cannot speculate

Maintenance requirements even at the most basic level for DDM2 will require daily oversight, as shipping list files are published several times weekly at the Federal Bulletin Board.

Summary of the Documents Data Miner 2 Online Survey:

Cost recovery revenue streams will allow completion of all the modules for DDM2 and provide for the ongoing operation of the site. This revenue could derive from user fees, from a contract with GPO or other vendor, or from grants. In the summer of 2001, it was decided that a survey should be conducted of the over 1300 Federal depository libraries in order to determine whether or not users would be willing to pay modest fees to use DDM2. The DDM2 survey was designed and administered by John Williams, Head of Acquisitions. The information below on the survey results is from his internal reports of August 3 and August 10, 2001.

The survey was announced on the GOVDOC-L discussion list and the DocTech-L discussion list, as well as through a batched direct e-mailing to the depository library addresses in the directory of DDM. The DDM2 Survey was available on the Web from July 13 to August 6. We had 232 responses, a response rate of about 17%. (Late responses up to August 10 boosted the total to 243, or 18.5%.) There was a nice cross-section of the library community from the University of Michigan to UT Austin and from Wellesley to UCLA. There were many small schools that took the time to respond, as well as responses from law, military, and professional libraries.

Survey Introduction— Basically, we wanted to know three things from this survey:

1. Profile the attributes of responding depository libraries: How successful the libraries were in managing their documents collections (8 questions)

2. Were the libraries interested in the features available in the prototype (1 large question)

3. Were the libraries willing to reimburse our costs in making the service available to them and at what level of changes (4 questions)

Profiles — 70% of the responding libraries were medium to large in size, selecting over 45% of available Federal documents. None of the responding libraries had completed the cataloging of their collections. And, finally, all of them had sufficient infrastructure to use all the features that would be available in DDM2.

The summary of results is as follows, based on the 232 responses by August 3, 2001:

1. Question 1 — How frequently do you use the Documents Data Miner?

Greater than Monthly:	57%
Less than Monthly:	43%

2. Questions 2-3 — Do you maintain a paper shelf list? Do you record holdings in an online catalog/database?

Paper and electronic:	57%
Electronic Only:	30%
Neither:	12%

3. Questions 4-5 — Is your Government Documents collection cataloged? Are your Documents accessed through an online catalog/database?

Not online and, perhaps, not cataloged at all:	3%
Partially cataloged but not online:	10%
Partially cataloged in an OPAC	87%

4. Questions 6-8 — Infrastructure (Staffing and Equipment). How many Internet enabled workstations do you have available in your public services area? How many Internet enabled workstations do you have available for staff employed in processing Government Documents? How many staff do you employ in working with Government Documents?

   Responses indicated adequate resources for all responding.

5. Question 9 — What percentage of Government Documents does your institution select?

Greater than 45%:	70%
Less than 45%	30%

6. Questions 10-13 — The Documents Data Miner 2 prototype offers access to all full-text electronic Government Documents in addition to an OPAC profiled to each individual depository. For this service, your depository would expect to pay a base monthly fee of how much? The Documents Data Miner 2 prototype currently offers a shipping list service as well as USMARC record download, either title-by-title or bulk. For this service, your library would expect to pay a fee of how much? The final phase of Documents Data Miner 2 development will involve electronic shelf-listing of Government Documents and interactive updating of each Depository's holdings. For this service, your library would expect to pay a fee of how much? For complete and unlimited access to the Documents Data Miner 2, your library would expect to pay how much per year?

Greater than $250.00 per year:	45%
Less than $250.00 per year:	55%

7. Question 9B — Your library would like to see the Documents Data Miner 2 developed with which features?

   - Shipping List Services and the URL Locator were requested by over 75% of responding libraries.
   - 45% of those surveyed want the MARC records utility and shelf-listing.
   - Between 30-40% want a Public Access Catalog and spine labeling.

Implications of responses — Regional depositories, which select 100% of Federal Documents, have the greatest logistical problems and, probably, the largest un-cataloged collections. Small depositories, selecting less than 30% of Federal Documents, have the greatest need for an organized catalog and processing utility.

Demonstration of Documents Data Miner 2 Prototype http://govdoc.wichita.edu/ddm2

Even though still under development, the prototype Documents Data Miner 2 offers unique and highly useable services, including one of the most-requested utilities by the depository community — searchable shipping lists. DDM2 offers the following features — [a demonstration of each feature followed]:

1. Shipping List Services

   - Searchable shipping lists dating from 1997. May be accessed by:
     - Title
     - Month
     - Year (both fiscal year and calendar year)
     - Item Number
     - SuDoc Number
     - Format (electronic, microfiche, paper, separates)
     - Depository Number (sets a filter on the shipping lists)
     - Searches may combine access points.
   - Profiling: Shipping lists may be profiled (or filtered) by depository.
   - Accessed lists offer:
     - Individual MARC record download, or
     - Bulk download of either all monograph records or all serial records associated with the list.
   - Currently warehouses over 6,400 GPO shipping lists.

- Potential to develop labeling function, which would complete the workload range associated with shipping list processing.

2. GPO MARC Records — the “MARC LOCATOR”

All MARC records created by GPO Cataloging Division from December 1998 to present (currently over 51,000 records). Records are searchable using:

- OCLC number
- Item or SuDoc numbers
- Agency (from 1xx fields)
- Title
- Title key words
- Subject (from 6xx fields)
- Formats

Accessed records may be viewed in DDM2's public view, MARC view, downloaded to the user’s PC for import into local databases, or accessed via the Web for records containing P/URLs. Records may be downloaded into OPACs individually or batched. Records may be tagged by depository profiles.

There is the potential to house all GPO MARC records from 1976 to the present: over 350,000 records. In addition, there is the potential to include retrospective cataloging project records from various institutions. All this would require further development.

3. Virtual Catalog — the “URL LOCATOR” (under development)

The URL Locator is a subset of the MARC Locator described above. The URL Locator is restricted to records with the 856 field for hotlinking to Web resources. At present, there are close to 15,000 URLs or PURLs in the records of DDM2, although not that many records because many records contain more than one 856 field. The URL Locator is searchable in multiple fields, like the MARC Locator. Records may be tagged by depository profiles; however, many library may wish to download all GPO MARC records for online titles.

4. Shelf Listing

- This module ties the individual pieces on the shipping lists to the MARC records and offers the only existing automated shelf listing of multi-part titles and the General Publications classes of the SuDocs class system.

- Currently, there are data elements for 124,000 individually shipped pieces in the Shelf Listing module of DDM2.

- About 85% of Federal documents are serial publications. In 2000, the GPO shipped 12 million pieces to 1346 depository libraries representing 26,000 titles.

5. Catalog (under development)

- This feature is designed to serve as an individual library’s catalog, and can be filtered to the depository profile.

- It offers a public view in addition to the MARC view of the record.

- It would potentially offer server space to libraries wishing to use DDM2 as their OPAC for their holdings, but this feature still requires development.

6. Login:

- Unlike DDM, DDM2 requires users to log in before using the utility. The Login is the user’s e-mail address in conjunction with a user selected password.

- New users must set up accounts. Access to the new account page is located on the Login page.

- A successful account setup allows immediate use of the DDM2.

The depository community is encouraged to use Documents Data Miner 2 and to communicate any problems or ideas for improvement to the developers. For additional information or to provide feedback, please contact:

Nan Myers
Associate Professor and Librarian for Government
Documents, Patents & Trademarks
Wichita State University Libraries
316-978-5130
800-572-8368
nan.myers@wichita.edu