Documents Data Miner 2©
Search Strategies & Statistics

Nan Myers
Wichita State University

Federal Depository Library Conference
Washington, DC * October 25, 2006

Development Background

Documents Data Miner 2©

http://govdoc.wichita.edu/ddm2

- Library management system for U.S. government documents.
- Assists in processing, cataloging and bibliographic control
- Web-based data mining tool
- Based on Documents Data Miner© (1998)
- Developed as Library/IT collaboration

DDM© Background: Development & partnership

- Began in 1996 at Wichita State University Libraries' Technical Services Department as a relational database in Paradox.
  - – Nan Myers, Government Documents Cataloger
  - – John Williams, Manager of Acquisitions
- Designed to support collection development and to provide union lists.

Development & partnership, continued

- Moved to Internet in 1997
  - – Partnered with NIAR at WSU
  - – John Ellis, Sr. Database Analyst, provided SQL Server database implementation, query algorithms, and Web database publication.
  - – Project conception and management supplied by Nan Myers and John Williams.

Development & partnership, continued

- Built on official sources of data from the Government Printing Office files at the Federal Bulletin Board File Libraries
- Announced as a partnership site with the GPO in April 1998

DDM© development goals

- 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
- Mirroring & Security/User Profiling
- Open System Follow-ons

DDM2© development goals

- Searchable Shipping Lists
- National shelf listing capability, recording items shipped to depositories from GPO
- Searchable Superseded List
- Provide export of USMARC records from GPO Cataloging (available at FBB 12/98 on)
- Identify subset of records with URLs

--continued

DDM2© development goals

- Full-Text Indexing of GPO MARC Records
- Develop online national public access catalog to government information, which can be adapted for individual libraries
- Provide bulk export of GPO MARC Records.

Databases in DDM2

Includes everything in DDM ...

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

Databases in DDM2 PLUS …

- 2002 Superseded List
- GPO Shipping Lists
- Shelf Lists
- GPO MARC Records
  - – Records with URLs (subset of MARC records)

Development & partnership of DDM2©

- DDM2© was announced in the Fall of 2001 as a pilot project.
- Collaboration between WSU Libraries and the University Computing Center.
- GPO/WSU Partnership arrangements for DDM2© are in process.
- Development team = Nan Myers, John Williams (Head/Acquisitions), & John Ellis, programmer, now Manager of Internet Applications for University Computing.

What was new in DDM2© in 2003

- Login no longer required
- Full-text indexing in MARC and URL Locators, and Catalog
- Improved Excel formatting (XML)
  - Requires Excel 2000 w/SP3 or higher
  - Otherwise, use the CSV button
- Upgraded server to SQL Server 2000

What’s new in 2006 & What’s coming up

- MARC Record Download
  - Feature on the Tools Page
  - Allows batching records for download
    - By depository number
    - By SuDoc number
    - By Item number
    - By date
    - Can select URL records only
- Coming soon – move DDM2 to another, faster webserver. DDM will go away.

Training

Getting started with DDM2©

- Tools
  - Session configuration*
    - Limit to your depository
    - Selecting to view selections of nearby depositories (i.e., set Union List)
  - Set records per page
  - Exports & Downloads
  - Export last query into Excel or CSV
  - Reports
  - Agency/Sub-agency List (IE only)
  - MARC Record Download

Modules

- List of Classes
- Inactive/Discontinued
- Superseded List (2002)
- Shipping Lists
- Shelf Lists
- MARC Locator
- Url Locator
- Catalog

List of Classes

- Search by:
  - Agency – can search all, or limit via drop down box.
    - Note current total active items for each agency
  - Item Number – full item (exactly), or partial number
  - SuDoc Stem – can search full or partial. Wildcard searches using %
  - Title – exact or words from a title. Automatic left/right truncation.
  - Format – use drop down box
  - Status – active, inactive/discontinued or all

Possible uses

- Annual update cycle
  - Download agency file and review
  - Look at who is also selecting nearby
- Collection analysis
  - Limiting by format to evaluate what types of material you’re selecting
  - Review collection by specific “titles” eg. Posters, Maps, General Publications

Inactive/Discontinued List

- Originally an important separate module
- Now can be accessed from the LOC module as well

2002 Superseded List

- A monograph publication, searchable just like the LIST OF CLASSES (on a “string”).
- Searchable by:
  - Agency name
  - Item Number
  - SuDoc Number
  - Title
- Query return = Agency Name, SuDoc, Item Number, Title, Instructions, Regional Note, and filter by profile.

Shipping List Searching

- Only searchable depository shipping list utility.
- Search by:
  - Shipping List Number
  - Title
  - Fiscal Year and Month
  - Shipping Year and Month
  - Item Number
  - SuDoc Number
  - Category (All or filter for Paper, MF, Electronic, Separates.
  - Depository Filter (Note: this eliminates shipping lists with item numbers not selected by the depository.

Shipping List Services

- Searchable from January 1997 to present
- PDF Files – Current Shipping lists to pdf versions from the FDLP Desktop.
- MARC Records – Linked to MARC LOCATOR module. Download individual records or bulk.

How do we match MARC records to the Shipping Lists? Three steps ...

- Monographs
  - Exact match: item number + sudoc number
- Serials
  - Run for exact match:
item number + sudoc stem
  - Run what’s left for exact match:
item number + sudoc stem + wildcard

The Shelf List:
Derives from shipping lists

- 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 SuDoc class system.
- Currently holds data elements for 182,728 individually shipped pieces (10/2006)

Using the Shelflist Function

1. Use Depository Selection & Directory
2. Click on your Depository number
3. Search for group of documents
4. Click on “Shelflist” when present in resulting table.

MARC LOCATOR

- Warehouses MARC records created by GPO Cataloging from monthly files posted at the FBB (began 12/98).
- Warehouses MARC records created by GPO Cataloging 1990-11/98 loaded as batch file on 10/02.
- Total MARC records on 10-19-06 = 242,273.

Searching the MARC Locator

- Full-text indexing environment
- Available fields:
  - OCLC number
  - Item or SuDoc numbers
  - Agency (from 1xx fields)
  - Title
  - Title Key Words
  - Subject (from 6xx fields)
    - Formats

Searching MARC Locator, cont.

- When searching full text on title or subject use “and” “or” or “near” as operators between words
- Phrase search using quotes also possible
- Help is very detailed – might help, but probably not.

Query Return Provides:

- Title
- Item Number
- SuDoc Number
- Hotlinked PURLs
- OCLC number
- Access to the MARC view of record
- GPO timestamp
- Option to download the record into your OPAC
- If search is done on agency, agency name appears

URL LOCATOR

- Subset of MARC Locator
- Restricted to records with 856 field for hotlinking to Web resources
- Searchable in same multiple fields as MARC Locator records
- Query return provides same data as MARC Locator records

DDM2 Catalog

- Public access catalog to government information resources
- Both public and staff views
- Could serve as an individual library’s government information catalog
  - Possible to filter against profile

Catalog public view includes:

- Title
- Author
- Publication
- Description
- Subject Headings
- Hotlinks from PURLs
- Call Number
- OCLC Number
- MARC Revision Date
  - Last update from GPO Cataloging

Subject headings

- Can be cut and pasted into a box at the bottom of the record.
- Clicking on “search” provides an index of all records with the same subject heading.

Staff view includes, in addition to MARC data:

- OCLC number
- Whether record is monograph or serial
- MARC revision date
  - Date of last GPO update
- DDM2© revision date
  - Date loaded into DDM2©

Using Documents Data Miner 2 in the Annual Selection Update Cycle

See “Hints for Using Documents Data Miner in the Annual Selection Update Cycle” from the University of Wisconsin/Madison online at:

http://www.library.wisc.edu/guides/govdocs/federal/usingddm.htm

To create a list of all the adds and deletes to the *List of Classes* since June 1, 2005

1. Select TOOLS from the top menu bar of the DDM2 homepage.
2. Scroll to the REPORTS heading.
3. For CSV version, click on CHANGES to the right of "CSV download of classlist..."
4. For Excel (XP or 2003) version, click on CHANGES SPREADSHEET to the right of "Excel (XP or 2003) download of classlist..."

To generate a list of items added to your own depository profile in the past 12 months (or # you want to track)

1. Click on DEPOSITORY LIBRARY & SELECTION on the DDM2 homepage.
2. Type your depository number in the DEPOSITORY NUMBER box.
3. Click on SUBMIT.
4. Click on the deository number in the DEP# column.
5. Type the number of months you want to track in the NEW ITEMS IN LAST ____MONTH box.
6. Click on SUBMIT (in the left column).

To create a list of active SuDoc stems for each agency:

1. Click on LIST OF CLASSES on the DDM2 homepage.
2. Select a field from the AGENCY drop down menu.
3. Click on SUBMIT (in the left column)

To create a list of EL titles:

1. Click on LIST OF CLASSES on the DDM2 homepage.
2. Select ELECTRONIC LIBRARY from the FORMAT drop down menu.
3. Click on SUBMIT (in the left column).

To create a list of titles your library currently selects:

1. Click on DEPOSITORY LIBRARY & SELECTION on the DDM2 homepage.
2. Type your depository number in the DEPOSITORY NUMBER box.
3. Click on SUBMIT.
4. Click on the depository number in the DEP# column.
5. Select ACTIVE FOR [DEPOSITORY #] from the STATUS drop down menu.
6. Click on SUBMIT (in the left column).

Additional Options

Note: You can also create lists of items your library does not select, or has dropped, by selecting different options from the STATUS drop down menu.

A Look at the Statistics

DDM2© Webtrends: Use statistics in 2000

Total hits	179,437
Average per day	1,080
Visitor sessions	10,950
Average per day	65
Average visitor session length	07:02
Unique visitors	3,839
Visitors who visited once	2,446
Visited more than once	1,393

DDM2© Webtrends:
Use statistics in 2003

Total hits	1,495,627
Average per day	4,026
Visitor sessions	41,960
Average per day	115
Average visitor session length	10:55
Unique Visitors	9,292
Visitors who visited once	6,230
Visited more than once	3,062

DDM2© Webtrends:
Use statistics in 2005

Total hits	1,800,568
Average per day	4,933
Visitor sessions	57,369
Average per day	157
Average visitor session length	11:37
Unique visitors	9,287
Visitors who visited once	5,642
Visitors more than once	3,645

About LOC data in DDM2

- We began collecting *List of Classes* data in DDM in 1997.
- The data is “official,” deriving from the GPO’s public files at the Federal Bulletin Board or an ftp file sent to us from the FDLP staff.
- Every data element is date-tagged.

List of Classes Data in DDM2 -- Active Item Numbers

- Active Item Numbers Oct. 2001 = 8534
- Active Item Numbers Oct. 2002 = 8025
- Active Item Numbers Oct. 2003 = 6476

Total Item Number Decline from 2001-2003:

2058 or 24%

BUT ... the item count has been going up!

Total item number count Oct. 2006 = 7384*

[*Note: Count in ItemLister for Oct. 2006 = 7431. GPO refreshes weekly. DDM2 refreshes monthly.]

Meaning of Rising Item Count?

- Addition of more and more electronic only titles, all of which have assigned item numbers.
- There is a correlation between declining shipping lists and increasing item numbers.
- Depositories need to know what is physical in their profiles, what is virtual and what is “both.”

Inactive List Data in DDM2

-- Inactive Item Number/SuDoc Pairs

- Inactive Item Numbers, Oct. 2001 = 10,447
- Inactive Item Numbers, Oct. 2002 = 11,472
- Inactive Item Numbers, Oct. 2003 = 11,705
- Inactive Item Numbers, Oct. 2006 = 13,002

Total Inactive Item Number Increase 2001-2003=11%

Total Inactive Item Number Increase 2001-2006=10%

Shipping List Data (FY2001-FY2006)

- Shipping lists in DDM2
  - 01/01/97 to 10/15/01 6,278
- Shipping lists in DDM2
  - As of 10/17/02 7,258
- Shipping lists in DDM2
  - As of 10/17/03 8,166
- Shipping lists in DDM2
  - As of 10/19/06 10,414

980 Lists added in FY2002
908 Lists added in FY2003

Lists added from FY2004 to FY2006 = 2248

Shelf List Volume in DDM2

- 122,899 individually shipped pieces
  - As of 10/15/01
- 139,503 individually shipped pieces
  - As of 10/17/02
- 154,100 individually shipped pieces
  - As of 10/16/03
- 182,728 individually shipped pieces
  - As of 10/16/06

Volume of Items Shipped Since FY1997

1997:	28,087	*2004:	10,635
1998:	32,499	*2005:	8,393
1999:	27,342	*2006:	7,227
2000:	21,984	*2007:	15
2001:	16,523		
2002:	15,860		
2003:	13,918		

Note: There are also 37 null items

and 123 items from 1912.

Decreasing shipping percentages

- 1998-2003
  - volume of items shipped decreased 60%
- 1998-2006
  - volume of items shipped decreased 78%

[If you think your processing/cataloging workload has been decreasing drastically, you are right!]

2006 Item #'s Not Shipped

- Total current active item #s = 7384
- Of 7384, online only item #s = 3683
- Leaving 3701 item #s that could be shipped against
- Of these, 1897 distinct physical item #s were not shipped against
- GPO shipped against 1804 item #s, or 49% of possible item #s

What is your true profile percentage?

- In the past year, GPO only shipped against 1804 item numbers, which is about 25% of the total current active item number count of 7384.
- In other words, depositories are receiving physical items for only 25% of active item numbers.
- As a ballpark figure, you can project that 25% of your stated profile percentage is your “real” percentage for physical item receipt.

MARC Records Data in DDM2

- MARC Records Total Oct. 2001 = 50,056
- MARC Records Total Oct. 2002 = 60,120
- MARC Records Total Oct. 2002 = 191,584*
- MARC Records Total Oct. 2003 = 206,924
- MARC Records Total Oct. 2006 = 247,273

*131,464 Cataloging records were added to DDM2's database on Oct. 17, 2002 representing GPO MARC records from 1991 to 1998.

MARC Records with PURLs

- Total Oct. 2001 = 14,215
- Total Oct. 2002 = 25,475
- Total Oct. 2003 = 38,565
- Total Oct. 2006 = 63,963

Online-Only Records in DDM2

- Total Oct 2003 = 10,443
- Total Oct 2006 = 63,963

Workload Assessment

- At WSU, workload in cataloging and processing physical items decreased from 1997-2006 by 75%.
  - Departmental statistics validate this.
- However, we must concentrate heavily on cataloging online titles and especially online-only titles to fulfill our mission as a depository library.

Workload Decisions

- From now on, every decision we make is a "Business Decision."
- Do we want to accept <u>all</u> the online item numbers and move towards a 100% goal?
- If so, should we contract with a vendor to push those records to us and use staff time for other projects?
- Could another project be retrospective cataloging for our pre-1976 holdings?

Further reading

Myers, Nan. “Documents Data Miner: Creating a Paradigm Shift in Government Documents Collection Development and Management.” *The Reference Librarian*, v.45 (94) 2006.

[Simultaneously published as The Changing Face of Government Information: Providing Access in the Twenty-First Century.]

Contact Information

Nan Myers

Associate Professor and Librarian for Government Documents, Patents and Trademarks

Wichita State University

1845 Fairmount

Wichita, KS 67260-0068

Voice: 316-978-5130 or
1-800-572-8368

Fax: 316-978-3048

E-mail: nan.myers@wichita.edu