Proceedings of the 7th Annual
Federal Depository Library Conference

April 20 - 23, 1998

Architecture and Functionality of Documents Data Miner

John Ellis, National Institute of Aviation Research, Wichita State University
Wichita, KS

Actually, I had no notes when I did my presentation of the technical side of the DDM. Most developers are like new fathers or mothers who need only an audience to expound for hours about their latest creation. I will, however, use the presentation slides and try to recreate some of the ideas that came to mind as they were flashed on the screen.

[slide 1]

Wichita State University

National Institute for Aviation Research

John M. Ellis

A.J. McCormack

The discussion of the technical side of the DDM will center on these key points:

- Design Parameters
- Current Attributes
- What It Cost
- Some Statistics
- Future Enhancements
- KISS

[slide 2]

GPO Data

- Profile.dbf (txt)
- listclass.txt
- appendix.doc
- unionl (ff-ddis or cd-ddis)

- fedbbs.access.gpo.gov

Early on we decided on these parameters:

Use only GPO data and when external data was added we tried to find FIPS data to augment what the GPO supplied. For instance, longitude and latitude data for the institutions were supplied mostly from USGS databases.

[slide 3]

Design Parameters

- Use only GPO data
- Use low cost wintel database server
- Use low cost wintel Web server
- Use only Web based clients
- Target Netscape and IE browsers
- Make it flexible

Use low cost wintel (Windows OS and Intel cpu) database servers and Microsoft Operating systems. Our experience with Intel based servers has proven them to be reliable and extremely effective if configured correctly. Our servers use dual cpu's, dual SCSI disk controllers, mirrored disk systems exceeding 20 Gb and a minimum of 256 Mb of memory. These typically sell in the $8-12,000 range and Microsoft offers extremely deep discounts on their OS products for educational institutions. You can do it even cheaper than this. The demo that we did at the conference was LIVE and it was all on a Pentium 133 laptop running a Microsoft Access 7.0 database on a CD, Microsoft personal Web server and IE 4.0 for the browser. It wasn't fast, but it did the show and cost about $3,000. This wasn't a watered down demo with a couple of depositories loaded up. This was the full database with all 3 million item lister elements included.

Use low cost wintel Web server. There are several to choose from but the IIS 4.0 server from Microsoft comes bundled with NT server 4.0 and is therefore free. It also is extremely powerful and easy to use. There are about 4 sets of ASP (active server pages) which are the core components of the DDM. The Microsoft paradigm for serving dynamic data on Web pages is both elegant and simple. It takes very little (but finely tuned) code to produce the pages that pop up when those queries are submitted.

Use only Web based clients. There is a plethora of products out there that will serve quite well as a front-end client. The big problem is, " How do you distribute these programs out to the masses?" The Web browser has made that problem moot. Everyone has a browser on his or her desk in the form of Netscape or Internet Explorer. For the developer this is as good as it gets. All users have a client program on their computer that acts pretty much the same no matter what you throw at it, and when they want a new client, they go bug Netscape or Microsoft, not the developer.

Target Netscape and IE browsers. We were very careful to make sure that our programs would work equally well on either Netscape 3.0 or better and IE 3.0 or better. The bottom line was that the client had to accept cookies and had to haveJavaScript enabled.

[slide 4]

Keep It Simple

- Microsoft VB 5.0
- Microsoft Access 97
- Microsoft Interdev
- No C++
- No special libraries
- No custom CGI's
- Built by 2 developers

Make it flexible and keep it simple. You can design any system to death. Or you can use flexible tools to throw something together and see what you like, and don't like. The DDM was throw together from pieces of previous projects and data downloaded from the GPO--in 2 days. I did the prototyping and Nan Myers did the critical evaluation of the directions we should pursue. I'm not sure that either one of us understands quite what the other does for a living, but between us we collectively built something that had never existed before.

Programmers or Web masters can't do a project like this on their own. It is very necessary that the client be very involved in the development process to keep the developer from wandering off track. The programming for this project was completed using only Microsoft VB 5.0, Microsoft Access 97 and Microsoft Interdev. There is no C++, no special libraries and no custom CGI's. It was built by 2 developers.

[slide 5]

Current Attributes

- Dual Pentium Servers

  1 backup/development server

  1 production Web server

  1 production SQL Server

- Microsoft NT Server 4.0

- Microsoft SQL Server 6.5

  1 development database

  1 production database

- Microsoft IIS 4.0

  1 backup/development server

  1 production server

- 1.2 Gigabyte of storage

  about 600Mb per database

- University supplied T1 network
- GPO data + "discontinued notes"
- WebTrends analysis software
- Can use any ODBC compliant Db
- A work in progress

The last statement if of particular interest. This project will probably never be finished. We constantly see things that need to be corrected, refined, or added. We encourage the user community to make suggestions for enhancements and to notify us when the programs or data appear to be in error.

[slide 6]

What It Cost?

- Data free
- Microsoft Back Office $3000
- MS Visual Studio $1400
- 1 Pentium server $8000
- Db Analyst 40@60/hr. $2400
- Db Programmer 120@50/hr. $6000
- Web programmer 120@40/hr. $4800
- approximate $25600

The items on this slide represent approximate street prices for the Microsoft products that were used to develop and support the DDM project. The majority of the cost is for developers and servers. If you already have this resource on site then the cost is diminished. In reality, at least two servers are needed and three is even better. The Web server should be different than the database server and it is always nice to have a separate development server that is isolated from the production system.

[slide 7]

Current Statistics

- depository active 1,364
- depository inactive 5
- gpoclasses active 9,507
- gpoclasses inactive 8,158
- Unionlist active 2,718,126
- unionlist inactive 225,754

The current stats slide shows the record counts for the depository table (profiles), the classlist table and the unionlist table. Inactive counts are the items that have been dropped from the official monthly data feeds from the GPO. Instead of dropping these items from our tables, we just mark them as inactive.

[slide 8]

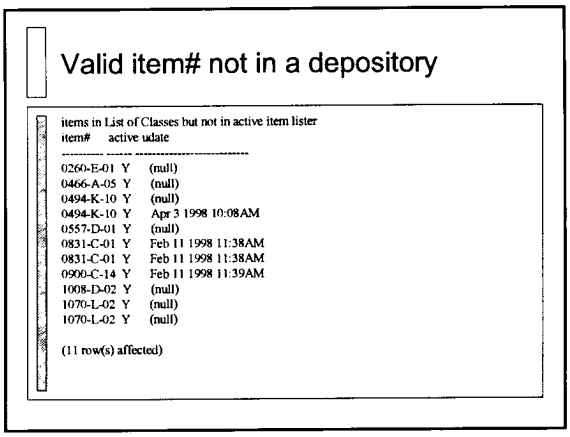
Future Enhancements

- Subagencies
- Export Features (full or profiled)
- Inactive Profiles?

Future enhancements will probably involve designing an export feature to allow users to download all or any part of the database that they wish to maintain on their site. Most of the future features will probably come from requests made by the user community.

Finally, I can offer some sample report output from the administrative side of Documents Data Miner.

[slide 9]



[slide 10]

Orphan Item/Sudoc pairs

status of orphan items found in gpoclasses

orphan items	sudocstem	status	last update
0231-B-23	C 61.40:	inactive	Mar 6 1998 10:48AM
0238-B	C 13.36/7:	inactive	(null)
0378-E-08	D 5.351/3-2:	inactive	Oct 10 1997 4:12PM
0431-I-41	EP 1.8/7:	inactive	Mar 24 1998 2:49PM
0435-E-12	E 3.11/4:	inactive	Oct 15 1997 5:21PM
0449	HE 23.12:	inactive	Nov 10 1997 12:46PM
0486-F-02	HE 20.5301:	inactive	(null)
0507-G-49	HE 20.3161/2:	inactive	Nov 10 1997 12:46PM
0512-A-27	HE 22.25/2:	inactive	(null)
0512-A-27	HE 22.25:	inactive	(null)
0581-E-11	HH 1.46/6-2:	inactive	Mar 6 1998 10:48AM
0581-E-11	HH 1.46/6:	inactive	Mar 6 1998 10:48AM
0768-C-03	L 2.3/5:	inactive	Oct 15 1997 5:23PM
1061-L	Y 3.F 31/25:	inactive	Oct 10 1997 1:26PM

(14 row(s) affected)

[slide 11]

Orphans - Depository count

orphan items	number of depositories
0231-B-23	552
0238-B	217
0378-E-08	429
0431-I-41	247
0435-E-12	220
0449	1100
0486-F-02	394
0507-G-49	633
0512-A-27	315
0581-E-11	481
0768-C-03	1238
1061-L	287