

8-25-1995

The Edgar Internet Project: Web Application Development Considerations

Mark Ginsburg

New York University, mark@edgar.stern.nyu.edu

Ajit Kambil

New York University, akambil@stern.nyu.edu

Follow this and additional works at: <http://aisel.aisnet.org/amcis1995>

Recommended Citation

Ginsburg, Mark and Kambil, Ajit, "The Edgar Internet Project: Web Application Development Considerations" (1995). *AMCIS 1995 Proceedings*. 164.

<http://aisel.aisnet.org/amcis1995/164>

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISEL). It has been accepted for inclusion in AMCIS 1995 Proceedings by an authorized administrator of AIS Electronic Library (AISEL). For more information, please contact elibrary@aisnet.org.

The Edgar Internet Project: Web Application Development Considerations

Mark Ginsburg, Doctoral Student, New York University Stern School of Business,
Information Systems Department. mark@edgar.stern.nyu.edu

Ajit Kambil, Assistant Professor, New York University Stern School of Business,
Information Systems Department. akambil@stern.nyu.edu

Overview

This paper describes the implementation of World-Wide Web (WWW) access of the SEC EDGAR (Electronic Data Gathering, Archiving and Retrieval) data base. EDGAR is a large, heterogeneous financial data archive that has been available to Internet users since January 1994. It is composed of all forms filed electronically to the Securities and Exchange Commission (SEC) by domestic publicly traded corporations and mutual funds.

We describe WWW design decisions and problems encountered in implementing a public access system to a large database. Our current applications include: an object-oriented mutual fund equity holdings database, a structured full text index search on corporate profiles, and real-time graphical visualization of stock price and mutual fund position changes.

Introduction

The EDGAR database consists of various text files representing over a hundred different form types electronically filed to the SEC by over 20,000 entities. These forms include the: 10-K (annual report), 8-K (Change in Material Status), 10-Q (Quarterly Report), DEF 14A (Proxy), 485APOS and 485BPOS (mutual fund prospectus), and so on. The information in these forms is of vital interest to professional and individual investors, librarians, public advocacy groups, business researchers, journalists, accountants and lawyers. For example, the Management Discussion and Analysis, which is Item 7 of the 10-K is scrutinized by investors as a key source of company management perspectives of prior performance and future trends. Similarly the proxy forms provide information on executive compensation, and corporate governance to different corporate stakeholder. The 13-D forms provide information on 5% or greater acquisitions.

Since the beginning of the Internet Edgar dissemination project in January 1994, the various Edgar forms have been available on the Internet within 2 days of their filing with the SEC. This service is provided as part of a research project on dissemination of large databases by New York University's Stern School (<http://edgar.stern.nyu.edu/>) and Internet Multicasting Service a nonprofit organization in Washington D.C (<http://www.town.hall.org/edgar/>). The Internet Multicasting Service (IMS) stores the

EDGAR data and serves it onto the Internet. NYU provides various access mechanisms, front-ends to, and certain types of analysis on the data.

Since the beginning of the project the IMS data store grown in excess of 25 GB, and the final storage requirement is expected to be greater than 40 GB by the end of 1995 when all publicly traded companies and mutual funds will file electronically with the SEC. To date 6,741 distinct annual reports (Form 10-K) and 5,588 distinct Proxies (Form DEF 14A) are available on the system. Key forms not in the Internet EDGAR archive include Initial Public Offerings (IPOs) prospectuses, and SEC Forms 3, 4, and 5 (officer, or "inside" purchases and sales) which are closely watched by professional investors. These are expected to be submitted via EDGAR sometime in 1996. As the EDGAR filing system only accepts text, the filings also miss any photographic exhibits.

The scale of the EDGAR database is thus orders of magnitude greater than experimental hypermedia databases such as the work of Salton et al. with the Funk and Wagnall encyclopedia [Salton]. Second, the contents of the database change daily requiring new indexes for retrieval. Finally elements and files in the database vary considerably. Whereas an encyclopedia's universe of articles has a relatively low size variance, and is fairly well structured with stable data that changes slowly over time, the EDGAR store can vary from small encoded filings that might be 800 bytes, to a large corporation's annual financial statement (10-K), that typically ranges from 300 KB to 900 KB. Second, the content and types of filings that arrive on the system vary from day to day. Currently there are over 166,000 filings requiring a master index at NYU for fast lookup of filings that is over 22.3 MB. Daily over 20,000 filings are transferred from the Internet Multicasting Site. Figure 1 shows representative NYU EDGAR Web server statistics from June 25 to June 30, 1995 It gets between 12,000 and 14,000 accesses per weekday. Including weekends, our average in June 1995 was 9,573 accesses per day. International usage, not shown here, accounts for between 5% and 10% of all accesses monthly.

The heterogeneity of the EDGAR data, and the minimal tagging of elements within filings, poses interesting challenges in the construction of WWW based dissemination and retrieval system. This paper identifies and addresses some of the problems encountered in such large scale dissemination.

Figure 1: NYU EDGAR Site WWW Access Statistics June 25, 1995 to June 30, 1995

| Domain | No. of Requests (total) | Megabytes Transferred (Total) | Range of Daily Requests, M-F | Range of Megabytes Transferred Daily, M-F |
|----------------|-------------------------|-------------------------------|------------------------------|---|
| US Commercial | 31,938 | 329.3 | 7,000 - 9,000 | 72 - 90 |
| US Educational | 7,453 | 52.9 | 2,000 - 3,000 | 20 - 30 |
| US Government | 947 | 9.5 | 500 - 1,500 | 5 - 8 |
| US Military | 299 | 2.4 | 250 - 1,000 | 3 - 4 |
| Network | 7,965 | 53.0 | 2,000 - 3,000 | 14 - 20 |
| Non-Profit | 739 | 5.0 | 500 - 1,000 | 5 - 6 |

User Applications

Access to Filings using Perl Gateway Applications

Using WWW Common Gateway Interface (CGI) tools, it is a simple matter to provide an attractive front-end for point and click forms retrieval. We identified a number of application categories based on interviews of EDGAR users.

For example, we provide a Schedule 13D application, where a user enters a company (e.g. Gabelli) in WWW form. The result is all companies in which Gabelli has acquired a 5% or greater ownership position. Another application searches for all prospectuses filed by a company. These applications are fairly straightforward: the user inputs a few variables, then we extract records from a master index file and provide a clickable answer which points to a filing. The last step, which takes place when the user clicks on the highlighted filing, is that it provides the native FTP service to the IMS server (town.hall.org located in Washington, D.C.). The filing is then transferred to the user's WWW browser for viewing.

These simple applications all use Perl (Practical Extraction and Report Language) as a scripting language. Perl routines are easy to read and maintain and PERL easily supports binary searches on sorted data files. We make use of public domain Perl code which has been developed to support CGI programming. We do this in part to enable other sites who wish to implement EDGAR front ends to easily port our code to their sites. Eventually this will reduce server loads from processing queries at our site.

An interesting problem we faced in constructing useful WWW applications for users is that of maintaining state. The underlying httpd protocol is inherently stateless; that is, client connection to server is dropped after one cycle of request and response. Hence we either have to generate applications that ask the user to specify all relevant query parameters at once (e.g. the ticker symbol, form type, date range and number of hits they wish to see) or work around the problem of maintaining state to create more adaptive user interfaces that return partial results to help the user construct a query. For example if the user wants a 10K reports but only knows the first few letters of the ticker or company name the user's partial query should return the potential list of companies they can choose from. On clicking on the company name, the programs should automatically retrieve the 10-K. As we do the ticker to company resolution, and then the company index lookup to identify specific 10K filings we have to maintain state across several scripts and keep the initial query parameters. To overcome limitations of the stateless protocol we use hidden CGI forms variables, or pass arguments to the secondary scripts with special CGI environmental variables such as PATH_INFO and QUERY_STRING.

Indexing Documents for Retrieval

In addition to retrieving specific filings, users also want to search the contents of the filings. Full text indexing using inverted indices is expensive. For example, freeWAIS and even commercial versions of WAIS, create large index files. Our experiments showed the ratio of index file size to source document size was approximately 1:1. We estimate full text indexing of proxies and annual reports will require more the 12 GB of storage.

To reduce indexing requirements while providing users with useful company information we created a corporate profile of key elements from various filings. We profile key information from the annual reports (10-K), proxies (DEF 14A), current events (8-K), stock and bond registrations (S-3), and acquisitions (SC 13D). To construct the profile we developed a set of Perl routines that extracted key information and organized it for a test bed of 200 large corporations, such as Exxon, Deere, etc.

The profile construction has a historical phase, where existing filings from the testbed are automatically FTP'd from the town.hall.org site and then fed into a filing type-specific Perl parser. This phase semi-automatically constructs an HTML profile, with a master Table of Contents file for each companies. The different profile components for each company are crosslinked, and links are also made to the original filings. Next, there is an automatic update to the profile contents and links when new eligible filings are detected in the daily index feed NYU receives from the IMS.

Next we enhanced the profile subsystem to include a WAIS-like keyword search. We used a public domain package, freeWAIS-sf (structured field freeWAIS) which permits us to define and search fields within textual documents, delimited by regular expressions. Since the profiles are between 500 and 700K per corporation, full text indexing of profiles requires far less storage in comparison to full text indexing of all key documents. In addition to fielded search freeWAIS-sf, has better support for phonetic ("sounds like something") keyword searches. However, we are examining packages such as "glimpse" from the University of Arizona. Glimpse provides an interesting alternative, since its indices are much smaller (2-3% for a small index; 20-30% for a large index to speed retrieval). Furthermore, users can specify error tolerances in a glimpse query.

Mutual Fund Visualization Project

As users' Internet access mechanisms improve we are trying to take advantage of new browser and access capabilities to enhance the value-added provided by the Internet EDGAR archive. Using the quarterly Schedule 13F-E Filings, we constructed a small time series database of major mutual fund equity holdings. Using Illustra, a commercial object-oriented relational database management system, in conjunction with customized front-end Perl scripts on the World Wide Web we now allow end-users to track equity sectors on a quarter by quarter basis. For example, questions such as "show me all equities whose share prices declined in the Mining Sector for Fidelity Magellan fund between first quarter, 1994, and first quarter, 1995" can now be quickly answered. Furthermore, the fund positions are also reported so it is now possible to compare stock price changes with fund position changes. After the database engine answers the original user query, an option is provided to generate dynamically two images, corresponding to the fund position and stock price changes for the ticker symbols in question. The latter function is provided by XrtGraph, a commercial Motif graphics software package. We pipe the tabular data into XrtGraph; generate an xwd (X-window formatted) image, then convert it into a gif for the Web by the freely available netpbm image manipulation library. We used the Standard Industrial Classification (SIC) code in our preliminary sector grouping, but found existing SIC mappings to be incomplete. To overcome

limitations of the SIC classification we are experimenting with commercial sector ("major code") and industry ("minor code") classifications furnished by Zacks Investment Research.

There are several important issues in database and real-time graphics routines: (i) we need to be able to fulfill a request in a reasonable time (database tuning and well-chosen image utilities), (ii) we need to guarantee integrity so multiple simultaneous users do not corrupt query and image results (use of unique temporary files), and (iii) we need to analyze network traffic patterns to and from our server to overcome bottlenecks that slow down the delivery of graphics.

Figure 2 shows server access log data for the week following the mutual fund interface announcement on our home page. Since *sector.html* was the only application mentioned explicitly on our home page, and the others were only reachable from *sector.html*, it is instructive to see how the World-Wide Web encourages user exploration and browsing of material on previously unannounced hypertext links.

Figure 2: Mutual Fund Interface Usage 6/23/95 to 6/30/95

| Application | Number of Accesses | Comments | Industry Classification Scheme | Link Mechanism |
|-------------------------|--------------------|---|--------------------------------|--------------------------|
| sector.html | 234 | Industrial sector stock price and fund position changes | Zacks | from NYU Edgar home page |
| industry.html | 15 | Industrial sub-category price and fund position changes | Zacks | only from sector.html |
| fund_sector.html | 43 | SIC Sector stock price and fund position changes | SIC | only from sector.html |

Concluding Comments

WWW applications development is a moving target as server and browser capabilities expand. The massive and complex EDGAR archive is an excellent testing ground for novel Web applications as well as the ability of existing tools to scale to meet the challenges of disseminating a large data set. We will continue to research: database, graphical visualization, indexing, and user interface design issues to improve the dissemination of corporate and fund filings on the Internet.

Acknowledgments.

This project supported by NSF Grant No. 9319331, *Internet Access to Large Government Data Archives*.