Public Access Web Information Systems: Lessons from the Internet EDGAR Project

Ajit Kambil
Leonard N. Stern School of Business
New York University

Mark Ginsburg
Leonard N. Stern School of Business
New York University

March 1998

Working Paper Series
Stern #IS-98-13

Public Access Web Information Systems: Lessons from the Internet EDGAR Project

© Ajit Kambil and Mark Ginsburg

NYU Stern School of Business

44 West 4th Street, NY NY 10012

akambil@stern.nyu.edu

1.0 Introduction

The Internet and the Web Information Systems (WIS) create new channels for communications between governments, corporations and individuals. Governments collect, generate and disseminate vast amounts of information. In a market driven and democratic society, this information is vital to enhance the trust of citizens in their government and institutions, and critical to individual and organizational decision-making. This paper describes lessons learned from the "EDGAR on the Internet" (EOI) project, an early demonstration web information system (WIS) for disseminating corporate disclosure documents filed with the U.S. Securities and Exchange Commission (SEC). Building on our experience with this system, and the emerging impact of the Internet on the market for disclosure documents, we provide guidelines for governments and information vendors to effectively adopt and adapt to web-enabled innovations for data dissemination.

2.0 EDGAR and the EDGAR on the Internet Initiative

EDGAR is the Electronic Data Gathering, Analysis, and Retrieval system of the SEC. The SEC is responsible for regulating financial issuers and market makers, and mandating various financial disclosures that are critical to investor decision making. These disclosure documents enable investors to receive information in a consistent way, allowing investors to monitor firm activities, assess investment prospects and make comparisons among firms. This leads to

increased investor trust and transparency of financial markets. Developed in 1983, introduced to a pilot group in 1984, and fully phased in on May 6, 1996, the modem-based, non-Internet EDGAR system electronically receives all major corporate disclosure forms, and provides various mechanisms for their dissemination. EDGAR was introduced to improve the public's access to disclosure documents and likewise improve the SEC examiners' review process of the substance of the filing. The advent of the personal computer facilitated the initial rollout. The SEC also wanted to alleviate the often cumbersome process of delivering paper documents to its headquarters (Neville, 1998). Today the EDGAR system annually receives over 60 GB of text data, from 16,000 filing entities (corporations, mutual funds, major shareholders, directors and managers) that file over 300,000 filings. Timely access to this information is critical to investors, and the general public who seek information on corporate activities.

Prior to the EOI Initiative, public access to these vital documents was expensive and inconvenient. Major corporations could purchase an online data feed from Mead Data Central for $150,000 per year. The feed provided the company with virtually instantaneous access to the data through dedicated communications links. Companies not requiring real-time access can purchase a day-delayed tape feed for about $75,000 per year. These data feeds provide the filings data as ASCII text with limited SGML-like tags. Information vendors who purchase day-delayed data typically process it to create value-added reports. Some vendors create databases distributed to corporations and libraries as CD-ROMs for thousands of dollars. Alternatively, information intermediaries in this market provided a fax or overnight or normal mail delivery of paper filings for a fee ranging from $20-$30 per filing, normally on a per-page fee basis. Individuals also access filings in a timely way from the Securities and Exchange Commission public reference rooms located in New York, Washington (two locations), Chicago, and Los Angeles. These reference rooms have a limited number of computer terminals with access to the EDGAR feed, CD-ROMs with well-formatted data, and microfiche versions of the documents. The public accesses reference room resources for free, but has to pay for photocopying or other reproductions of the filings. Furthermore, they were not permitted to download files on a computer. Thus, investors who could not afford timely access to these documents had a substantive information disadvantage over those individuals and firms with such resources.

The Internet EDGAR initiative sought to provide widespread access to these important government documents and demonstrate the feasibility of the Internet as a viable low cost information distribution mechanism for public data access. Conceived by Carl Malamud, a leading Internet researcher, this project was sponsored by a two year National Science Foundation (NSF) grant awarded in November 1993 to New York University's Stern School of Business in collaboration with Malamud's Internet Multicasting Services, a non-profit organization.

Beginning in January 1994, IMS provided servers for hosting the data, high-speed connections to the Internet, evolved routines for implementing electronic mail, FTP, gopher, and WWW access to the EDGAR. With funds from the NSF grant, a day-delayed feed tape was acquired from Mead Data Central and processed to parse header information from EDGAR files, to store EDGAR data on server disks, and to construct simple indexes on the data by company name, form type, and date. Early access required users to search the index and submit long filename strings to the FTP or electronic mail service to retrieve filings. The availability of gopher and

the emergence of web browsers greatly enhanced user access allowing a more simplified point-and-click interface. By April 1994, a simplified WWW interface was provided, and by October 1994, the service provided a Wide Area Information Search (WAIS) on key header data within the filings, enabling users to search for companies by their name and locations. In addition to individual access to filings, the FTP system permitted bulk downloads of daily filings, to encourage broad access to the data and innovation. Today, the New York University site and various access programs developed there make up the core of the SEC public access EDGAR service (http://www.sec.gov/edgarhp.htm). This widely publicized initiative has successfully improved access to these documents.

Since the beginning of the project in 1994, over twenty million files have been transferred to the public with daily access to the SEC repository continuing to grow. In August 1997, the SEC site transferred over 400,000 filings daily making it one of the most visited sites on the Internet. This illustrates the success of this method for disseminating disclosure documents. During a regular business day, the less well known NYU Internet site (http://edgar.stern.nyu.edu) receives between 20,000 and 40,000 visitors who generate requests to the SEC and a redundant server provided by NYU's sponsors. Log files show that 70% of visitors to the NYU site come from corporate or access provider sites, and 30% from educational, government and non-profits. 10% of the latter group visited from foreign countries. This illustrates the wide and timely distribution of the data enabled by the Internet regardless of geographical location. An early survey of users by the NYU team showed that users had widely different backgrounds and motivations for using EDGAR. In addition to corporate and private investors, users included chief financial officers and employees tracking corporate and competitor disclosures, unions tracking executive compensation and company actions, researchers, potential employees and others seeking detailed information on companies. Recently, many business students and accountants have relied on Edgar's vast database to do research on industries and companies that would have in the past have required much more time to compile.

During the NSF grant period and a subsequent grant from Disclosure Inc., the NYU team focused on developing a number of routines that allowed value-added user access to filings. Based on our analysis of user needs and server requirements, we undertook four types of value addition: the design of "restrictive" and customized search tools, the extraction of key data elements from filings, the provision of auxiliary data to assist in retrieval from the filing database and the development of client-side tools. This enabled more EDGAR value addition at the client rather than the server. We built these applications based on suggestions from our initial focus group meetings and ongoing electronic mail feedback from users. Applications were developed based upon constraints determined by the availability of technology, skilled staff and costs. Illustrative examples and the rationale for these types of WIS value addition applications are given below

Specialized Search Tools: Restrictive Interfaces and Access

Through the web forms feature, WIS systems are especially suited to creating "restricted" and "customized" interfaces to databases that best suit the need of specific user groups, providing greater convenience in access for users. We quickly found that more sophisticated EDGAR users wanted more specialized web forms through which they could specify criteria for accessing

filings. This was preferred to the WAIS interface that required users to develop complex Boolean queries when they needed specific documents. In contrast to the WAIS search, drop-down list boxes on web forms guide user choices and query formulation. The simplest form of such "restrictiveness" (Silver 1991) is the selection of a specific form type for a specific company, within a limited date range. Another example of specialized access is the custom web form to report on all five percent or greater ownership of a filing company. This is considered as a potential signal to an upcoming acquisition of a target company. The filing, a Schedule 13D or a Schedule 13G, can be parsed to display the names of the potential acquirer and target. We expect users will continue to demand more specialized and customized interfaces. We also expect emerging tools supporting WIS development will enable systems that adapt the interface to best suit the preferences and level of sophistication of individual users.

Search Tools: Full text search on documents

A number of full text search engines are freely available on the WWW. However, the EDGAR database is very large, in excess of 60GB. Given the costs of storage and the thirty to fifty percent increases in storage required for full text indexing, we initially only provided full text search on a company annual report to shareholders (Form 10K) using the Excite search engine. As storage costs fell, we implemented a modified version of Cornell's SMART system as a search engine for a further subset of popular SEC filings. This is being used to research the effectiveness of search engines, that in contrast to relevance feedback techniques implement learning methods to permanently modify indexes for full text search. As costs fall and full text indexing software is increasingly bundled with web servers (e.g., Microsoft Information Server and Index Server) we expect the entire SEC database to be fully indexed. Future WIS systems for public access to data will have minimum expectation of full text indexing.

Automatic Data Extraction to Create Value Added Reports:

Initially, we parsed some simple data out of filings such as five percent ownership by potential acquirers along with the name of the target companies to create value-added reports for users. Next we undertook efforts to extract a wider range of data, such as detailed balance sheet and income statement elements, "Executive Compensation", "Management's Discussion and Analysis of Financial Condition and Results of Operations", and other key text sections from different filings to build a composite corporate profile. Another value-added report created by extracting data from filings focused on mutual fund equity holdings extracted from Schedule 13F-E filings. A small number of fund families such as Fidelity initially used the EDGAR system to file these equity holdings electronically. From these filings, we extracted data to a relational database, and then developed a CGI interface to convert the user's input into standard SQL query language to report, quarter by quarter, a given fund's equity holding. Furthermore, we grouped the equity holdings into major industry sectors using the Zacks Investment Services sector classification scheme. To represent the data as a graph at run-time, we have made use of proof-of-concept freeware Gnuplot and a higher-end commercial product, XrtGraph to display charts of mutual fund equity holdings grouped by Zacks investment category. However, further extraction efforts were limited given a number challenges which we explore in greater detail in the next section.

Adding Value through Aggregation or Addition of Data:

One way of adding value to data is to combine the EDGAR data with other data sets to improve access or reports. One such addition we undertook was to implement a ticker symbol search. Investors prefer to access documents using the exchange ticker symbol. However, the developer of the mapping usually copyrights the mapping of ticker symbols to the SEC file names. Thus we have accumulated a list of 8,000 ticker to SEC Central Index Key (CIKs are the unique ten-digit identifier of filers assigned by the SEC) mappings and implemented routines that search the web and automatically map most companies to SEC names at low creation and maintenance cost. This permits a public domain mapping to allow other sites to offer such a search.

The ability to hyperlink also simplifies the virtual aggregation of data from multiple sources. Many web sites hyperlink to the SEC or NYU sites to complement and add value to company or other data they offer to users.

Migrating Value Addition to Client Software: Given increasing demands on our server, we felt we could dramatically reduce server loads and search and retrieval times on the latest filings by shifting more of the search function to the client. In 1996, we implemented a Java applet that represents a single entry point to all common NYU retrieval tools and allows the regular EDGAR user to search the latest filings without having to make search requests of the server. It features a clickable scrolling list with company names. It takes the user directly to the SEC, or a backup data server, to retrieve the filing. Figure 1 shows an example of the Java applet with the user entering "DEAN" (for all companies starting with "DEAN", for example DEAN WITTER) and selecting an additional criterion of "Form Type 10-K (Annual Report)". The answer is returned in the simple visual metaphor of the top horizontal scrolling list; the user need only click on the item desired in the scroll list to start an FTP of the full textual filing.

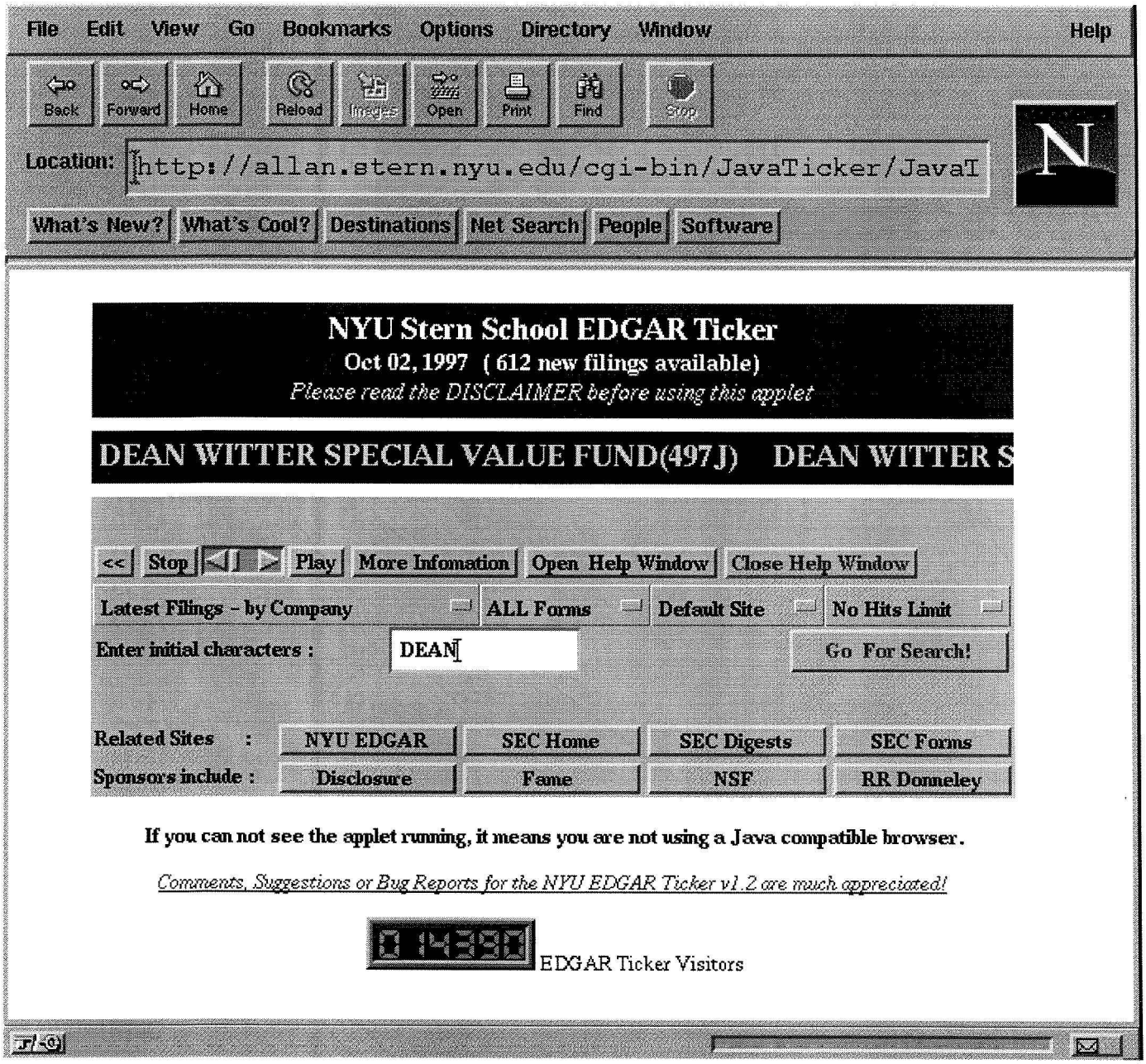

Figure 1: Java Applet Interface

When the user launches the Java applet, it retrieves an index of the daily filings allowing the user to sort it by various criteria such as company name, form type, and date filed and move directly to the relevant object. The applet interface also uses "cookies" to store information on the client side in order to save preferences, such as a portfolio of ticker symbols. In this case, the filings the user specifies appear first on the screen similar to a notification system. If the client does not support "cookies", we allow an alternate selection of server-side preference storage (in a flat file format, not a database in this case). These features allow users to customize their interface, as discussed earlier, to suit their preferences. This flexibility of providing server- and

client-side mechanisms for storing preferences is important because this is a public-access database with a wide variety of client technology platforms.

The Java model was chosen because we envision not only processing simple search at the client side but also adding hypertext formatting and data extraction features at the client side. However, the original Java security model did not permit us to manipulate filings retrieved from a server, other than that from which the applet was downloaded, nor did it allow us to store processed files on a user's local drive. To implement greater value-added processing at the client side, and increase the efficiency of downloading the applet we are now evaluating the migration of the applet feature to a Marimba and Castanet environment. This will allow us to push the index out instead of the whole applet and create features for processing the files downloaded on the client side.

Our model of extending the client side applications. in contrast to the server side applications is predicated on the belief that government or university funding for servers as well as application development and maintenance resources for public access will lag user demands. By building smarter client programs, we can reduce server and communication loads and costs to government providers. This is consistent with emerging uses of Java and push technologies in WIS to efficiently distribute processes to the best processing location. Java and new push technologies provide new capabilities to WIS designers for both customizing the user interface at the client side and distributing value-added applications. They also provide a low cost way for entrepreneurs to create software and add value to the data without having to invest in expensive server resources.

A number of challenges were encountered in implementing the EDGAR on the Internet initiative and adding value to the EDGAR data set. Below we examine some of these challenges.

3 Challenges to WIS Value Addition: Lessons from Internet EDGAR

EDGAR on the Internet is a specific type of WIS. It is not a "de novo WIS" application – a new web information system developed from scratch to take advantage of the WWW (e.g. Onsale.com or Amazon.com). Neither is it a "ported WIS" converting an existing C or C++ application to a Java applet for WWW delivery and multi-platform access. Instead, the Internet EDGAR represents a web interface to a "legacy application". The central challenge to adding value to EDGAR was interfacing to this legacy application with its prior limitations and porting its data into the WWW environment. Limitations of the existing legacy application such as the lack of consistency and structure of filings, and the inadequate identification and tagging of broadly useful data objects posed major challenges to the effective automatic data extraction and value addition. As listed below, many of the underlying causes of this problem arise from prior policy decisions and choices made in the implementation of prior systems. A major cause of the irregularity of the tagging is due to the proprietary nature of the typesetting systems used by the key family of EDGAR filing agents in the creation of the EDGAR documents. EDGAR tagging is different at Bowne & Co., R.R. Donnelley and Merrill Corp. Had the SEC forced this small group of key agents to standardize their tagging, many of the problems would have been lessened (Risser, 1998). Designers of WIS systems that interface to legacy systems within organizations or to government data systems are likely to encounter the types of problems illustrated below by

examples encountered in porting the EDGAR data set to the Internet.

Incomplete Data:

Incomplete information or data within legacy applications can limit the ability of WIS designers to add value to data. For example incomplete information in files and among filing types limited the ability to automatically extract key information and provide comparable reports to users of the Internet EDGAR project. For example, many of the most established corporate filers are given the option to provide an electronic filing where key portions are "incorporated by reference" to previously filed documents. This means users have to purchase or acquire a paper copy of the incorporated document directly from the company or a vendor of filings. In addition to sections, filings often miss vital header data in terms of the company standard industrial classification (SIC) code or other information which data vendors or users must correct at the point of dissemination or use. Beyond omitting elements within filings, firms may also elect to submit some filing on magnetic tape or paper. For example, the Schedule 13Fs on quarterly mutual fund equity holdings is filed by magnetic tape. Magnetic tape filings do not appear in the Internet EDGAR archive. However, if the company opts to file them via regular EDGAR then the form is labeled a "13F-E" and is available to Internet Edgar users. This interesting filing is thus not well represented at present in the Internet Edgar project: in 1994 through 1996, we received 209, 237 and 224 Schedule 13F-E's respectively despite the large number of mutual fund families. These problems occur because the SEC provides wide latitude in how companies can file disclosures; the electronic filings are incomplete and are not checked. For WIS applications with automatic value addition to legacy data to be effective, the legacy systems have to assure that the data reported and collected is both consistent and complete.

Poor Identification of Data Objects:

Legacy systems may have poor data definitions and systems of identifying data objects. This can create problems for automatic extraction and value addition to data. For example, the tags which issuers apply to SEC filings are limited and quite irregular. The SEC has limited the number or items to be tagged and currently allows a wide variety of tags, for example to delimit a financial accounting data table. Previously, a key SEC priority was to create incentives and reduce the efforts of firms to comply with EDGAR-based filing (modem filing versus tape or paper). Instituting or enforcing a strict tagging requirement on the incoming filings was viewed as increasing the burdens of filers. Filers have thus been treated very leniently in terms of what constitutes an acceptable filing format. Consequently, until the EDGAR systems is re-engineered with a fuller set of semantic tags and a more rigorous structure, it is a difficult technical process to automatically and consistently capture specific data items of interest. Currently, the filing transmission process makes use of EDGARLink, a software package developed and maintained by the SEC, to send documents from the filing entities to the SEC via modem. Its error-checking is minimal; however the error-handling routines naturally can evolve as the tags become more precise. For example, it is a difficult parsing job to provide a consistent mechanism to cut and paste numbers directly out of a table for download directly to a spreadsheet. For example, there is no consistent application of column tags; hence the number 1991 is ambiguous: it might be a value associated with a line item or it might be a year in a column heading. Even when a parsing software application is run against an original filings, the

danger is introduced of creating errors in an extracted filing where none existed originally. Key contextual information such as a related footnote could be missed. Effective automatic value addition to information in a WIS environment requires clear and consistent standards for tagging and identifying data of interest to users.

Reporting Errors:

Web information systems connected to the Internet can rapidly disseminate information worldwide. However, this can also lead to widespread dissemination of erroneous data, diminishing the satisfaction of users. For example, data quality issues also proved to be a stumbling block to timely extraction of data and user satisfaction with the Internet Edgar service. With numeric data, it is quite common that a company's initial filing will incorrectly report accounting data of interest. In this case the company has the responsibility of filing an amendment in a timely manner (for example, a form 10-K/A amends a 10-K). To put matters into concrete terms, consider that as of August 25, 1997 the SEC had received 15,146 10-K annual reports and 4,923 10-K/A amendments. Similarly, 68,005 10-Q quarterly reports have been received and 5,978 10-Q/A amendments. Another interesting fact about initial filings is that they are often inaccurate and misleading. That is why the filings are reviewed by SEC examiners. In many instances, /A filings are demanded by the SEC because of a misleading statement or omissions in the original filing. So an interesting question, which merits further exploration, is what happens when a trader trades upon pre-effective information contained in an EDGAR filing (Risser, 1998)?

Presently no routines automatically check the consistency of financial information in a document before they are disseminated. Thus, it is important for users and value adders to monitor for data amendments and check for data prior to its publication and dissemination in a web information system.

Terminology and Nomenclature Issues:

Web information systems allow inexpensive publication of information, often reaching new user groups who may use different terminology or methods to link related data. For example, a problem encountered by users of EDGAR is the differences between the SEC company conformed name (CCN) and company's common name in the marketplace. This often results in a "no hits found" response to investors' queries. Constructing mapping of common names to CCNs is costly and there are no public domain versions of such a list. Similarly, there wasn't a freely available mapping between ticker symbols and company conformed names the preferred access key for investors.

In addition to naming of files, the legal language of the filings often confounds the layman. The original filing definitions stem from The Securities Act of 1933 and the Securities Exchange Act of 1934 and the motivation for and meaning of the myriad of filing type codes and disclosure are best explained in specialized legal or investment handbooks. Individual investors thus often find these filings impenetrable, and they are often unable to identify or adequately understand the content of filings that would answer their questions about a company. For example, if a company recently announced a major factory closing, a savvy Edgar user would look at the 8-K

(current report filing) whereas a less successful technique would be to wait for the next full quarterly accounting statement in a 10-Q. Thus consistent nomenclatures and file types combined with better user education will enhance the user accessibility and comprehension of disclosures

The Challenge of Technological Innovation

Rapidly changing technical environments generate tremendous product variety and uncertainty about appropriate technical solutions to solve a problem. In a public access environment they can generate different dissemination alternatives and support requirements. For example early in the project we had to support Lynx, Mosaic, Netscape, Spry and other browsers that varied in access protocols to web-forms and responses to CGI scripts. As products standardize, support and variation in offerings can be reduced.

Early on in the project when WWW-to-database integration tools were non-existent and storage very expensive, we implemented binary searches on the company name in a flat file structure (which is our choice on our two Unix servers). The binary search is a very fast and efficient way to maintain a simple index and has scaled well. However, as new technology became available we were able to migrate to a relational database (on an NT server which runs an MS-SQL database) to offer more efficient retrieval on multiple index fields. In contrast, the WAIS indexing of company header files is increasingly cumbersome and slow. Similarly, as disk space becomes cheaper we are now making selected documents available for full text searching. This illustrates the challenge of adapting to new technologies while maintaining scalability.

Thus early technology choices can lock projects into inefficient solutions, while new technologies create new opportunities. The selection of specific technologies and timing the migration to newer platforms while maintaining older application services will present an ongoing challenge for all web information systems designers.

4.0 Web Information Systems: Value Migration and Implications for Competition.

While providing new capabilities for information dissemination, web information systems transform information vendors' requirements for value-addition. In the market for disclosure documents, the WWW and Internet EDGAR initiative has radically reduced the cost of disseminating EDGAR data to users. This in turn has diminished the relative value of timely and convenient access to EDGAR documents as a key source of product differentiation and competitive advantage. In response, market incumbents and new entrants are using web information systems to modify and differentiate their products on different dimensions to create competitive advantage. Specifically, the information vendors are modifying their products to:

- Increase human analysis and judgement
- Aggregate more proprietary content
- Support personalized and community interaction.

Figure 2 summarizes these three strategic directions for information vendors in web enabled

environments. These strategies are illustrated by examples of vendor responses to the EDGAR initiative.

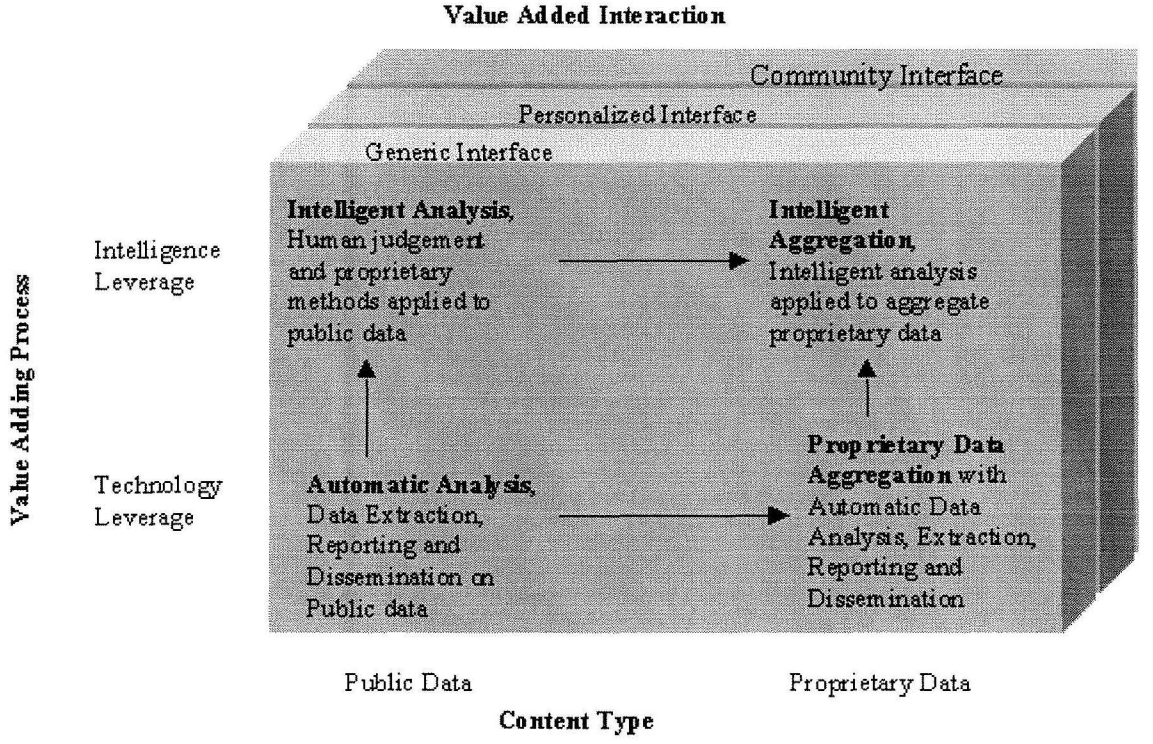

Figure 2: WIS Enabled Information Vendor Strategies

4.1 From Technology Leverage to Intelligence Leverage

As the Internet became widely adopted, the EDGAR data vendors undertook initiatives to leverage the improved performance of processing, storage, communications, interface and programming technologies for the WWW. In the EDGAR domain these ***technology leverage*** strategies include timely notification and document delivery, full text search, automatic data extraction and reporting, and data analyses.

For example, a number of vendors responded to the Internet EDGAR initiative to develop a more

timely EDGAR service by purchasing Mead Data Central's primary EDGAR feed and releasing the data instantaneously on the Internet combined with notification services. Users were charged a premium price for this service, until FreeEDGAR (a service initiated in June 1997) provided same day feeds with advertiser sponsorship. All the major commercial vendors such as Disclosure, Moody's, and new Internet entrants such as Edgar Online, and Internet Financial Network provide both alert mechanisms and timely document delivery. FreeEDGAR, like the original EDGAR project, leverages the low cost telecommunications and publishing capabilities of the Internet and WWW to disseminate EDGAR data. This example highlights the limits of a technology leverage strategy where technology costs decline. Under such conditions barriers to entry fall, allowing new entrants to enter the market at lower costs than the original incumbent. This in turn creates new competitive pressures, eroding profit margins. In the case of EDGAR, real-time access to filings becomes a low value commodity, where the customer's price is virtually zero. This undercuts the traditional subscription or pay-per-use revenue for timely notification and delivery of documents.

A second technology leverage strategy is to add structure, format and extract valuable data from filings. The simplest level of formatting pertaining to the Internet is to add HTML tags to improve the look and feel of a document and enable easier navigation to different document sections as Disclosure's online offerings. This is especially useful for long filings. Another extension is to enable filings to be downloaded with reasonable formats into Microsoft Word or Excel. For example Financials Online, a small entrepreneurial venture, provides HOTEDGAR which consists of software tools for parsing financial data from tables in the filings and reformatting it for use in Microsoft Excel spreadsheets. Similarly, Price Waterhouse's Technology Center has developed Edgarscan, building off filings provided by the original Internet EDGAR project to create a sophisticated software tool employing multiple parsing strategies to automatically extract major textual and numerical sections of filings. EdgarScan was developed in Prolog and C++ and is the most comprehensive automatic extraction tool operating on EDGAR filings available over the Internet. The tool is used extensively within Price Waterhouse to track specific patterns in company filings and identify opportunities for Price Waterhouse's services to clients. As discussed earlier in this paper, there are a number of difficulties inherent in the structure of the filings that limit the effectiveness of automatic extraction of data from electronic filings. Thus none of these methods consistently extract the data correctly without human intervention.

Third, vendors can enable better search and value added reporting of the data. Vendors of EDGAR CD-ROMs have typically included full text search, analysis and visualization software on their CD-ROMs. These features are beginning to migrate to WWW services. Full text search is increasingly implemented on EDGAR web sites. As data warehouse software and online analytic processing tools for the WWW improve, well structured EDGAR data can be presented in multiple ways to users for analysis enabling comparisons, trend and other forms of data analysis.

Technology leverage strategies which primarily rely on exploiting innovations in technology capabilities are easily observed in the market place and generally easily imitated by well financed competitors. As WWW development and information dissemination tools improve, the entry costs for competitive vendors offering similar products continues to fall. However, with

the lowering of the barriers to entry, the potential return is also lowered. The market will probably not attract many new potential new entrants. Thus, these strategies have limited potential to create sustainable competitive advantage and the lowest differentiation and lowest barrier to entry strategies for data vendors.

In contrast ***intelligence leverage*** strategies create new products and services based on applying intelligence to data. These can be proprietary reports based on human interpretations or non-observable methods applied to the data. One example of this strategy is Disclosure's construction of Insider Trading Indexes. This product aggregates information on a weekly basis from 5,000 insider SEC filings, then combines it with a proprietary weighting system and expert analysis to gauge executive sentiment across industry sectors and regions. Such a ratings system depends on the design of the rating algorithm, and human judgment to establish ratings with adequate predictive power to correctly signal market directions. If the rating establishes a reputation as a good predictor or useful market signal, Disclosure's proprietary product will be harder to replicate, unless a rating system with better predictive power and reputation can be established. This would be a costly undertaking for competitors. Since the competitive advantage is based on intangibles such as reputation and human interpretation. This is an example of a resource-based strategy (Barney 1991) that is not easily replicated by competitors. Barney (1991) notes that a firm's resources that are valuable, rare, inimitable, and non-substitutable provide the basis for sustained competitive advantage. Intelligence leverage strategies create such assets.

Thus we expect information vendors operating in environments characterized by WIS generated technological discontinuities to migrate to intelligence leverage strategies and increase the content based on specialized interpretations and opinions of data. However, as long as the underlying data assets that form the basis of these opinions are not proprietary, and profit potential is sufficiently high, there is the potential for other competitors to establish an interpretation of better predictive power and accuracy on the same data.

4.2 Proprietary Data Aggregation Strategies

As FreeEdgar and similar projects makes access to timely EDGAR data a free resource, EDGAR data vendors are including their product as a part of more complex offerings. Specifically, information vendors are aggregating EDGAR data with proprietary information. Intellectual property protections such as copyright, patents and trademarks make certain types of information "proprietary" to owners, preventing others from consuming the information without purchase or owner's permission. Aggregating public information with proprietary information allows the aggregate product to be proprietary. This enables vendors to create unique products not easily replicable by competitors.

This data aggregation strategy is exemplified by Disclosure's combination of SEC data with the Articles Online collection of two million articles on companies and products from various newspapers, journals and other sources, such as equity research reports from IBES and Multex. Aggregation of different proprietary data sources and integration with the non-proprietary EDGAR data enhances convenience and value to the user by reducing their efforts in searching and acquiring related data. It enables economies of scale in searching from a user perspective.

Data aggregation also allows vendors to more easily sell subscriptions and reduce the buyer's adverse selection risks and transaction costs. Adverse selection risks and transaction costs arise from the information asymmetry between a buyer and a seller about the quality of the product. The buyers confront adverse selection risks as they often cannot know the relevance or quality of the information product without consuming it. Buyers incur transaction costs in acquiring information to verify product quality and writing contracts to mitigate against adverse selection risks. By bundling multiple branded sources of information, the vendor increases the buyer's likelihood of finding information of value within the bundle. Thus bundling reduces the buyer's perception of adverse selection risks and makes the buyer more amenable to purchase the bundle by subscription. Thus data aggregation strategies generally create more attractive products for customers and provide vendors with new dimensions for product differentiation. Data aggregation also creates new barriers to entry in the industry. These barriers to entry arise from high fixed costs of setting up a dissemination system for aggregate data including the costs of licensing content from multiple sources and effectively integrating the data into products. At the same time the low cost of communications and WWW development tools facilitate ongoing data aggregation from multiple sources and dissemination of data, allowing the vendor to take advantage of economies of scale.

Intelligent Aggregation is an extension of intelligent leverage and data aggregation strategies. This strategy combines information from proprietary and non-proprietary sources with proprietary analysis and human interpretations to create new signals on data and value-added information. For example, the Disclosure's insider trading indexes strategy could be extended to broader types of ratings similar to products offered by Moody's or S&P equity ratings. Alternatively, products can be generated that provide interpretations of and rate the accounting practices by the firm. The key asset in addition to proprietary data analysis methods and thoughtful interpretation of the data is the ability to generate a reputation of consistent practices for generating these new signals and good predictive power. Over time we expect information vendors on the WWW to migrate their value toward more "intelligent aggregation" strategies that combine analysis, thoughtful interpretations and useful new signals on data from multiple sources. These strategies, while costly to develop and implement, will also remain hard for competitors to effectively replicate.

4.3 Generic to Community Interaction

Web information systems enable varied forms of user interfaces and interactions between consumers and information vendors. Information vendors can leverage these capabilities to differentiate their products and services. We identify three levels of interaction: generic, personalized and community, which create value in different ways and require different levels of resources to implement.

Generic interactions apply common user interfaces to all or large subsets of users. The Internet EDGAR project generally provided generic interfaces to the EDGAR database with limited customization. In contrast, the Java Applet enabled a "personalized interaction", allowing users to define their preferences and presenting documents that matched their preferences first for review. This creates value by reducing the user efforts to search for documents on an ongoing basis. It also introduces a switching cost for users, who incur a cost to "personalize" their

interface, and thus are unlikely to incur the same cost with another vendor offering a similar level of service. Most for-pay EDGAR sites are adopting "personalized" interfaces for their clients.

A third capability enabled by the WWW is interaction among a community of users. While no traditional EDGAR vendor has implemented a community interaction strategy by becoming the organizer of a "virtual community" leveraging the EDGAR data set, some vendors provide EDGAR data to various virtual communities on investment and finance. For example EDGAROnline provides data to The Motley Fool (www.motleyfool.com), a popular virtual community for investors. The Motley Fool hosts a number of bulletin boards that enable users to interact and discuss specific companies, investments and disclosures such as earnings reports by companies. In addition to actively and directly interacting through bulletin boards, chat and other communications, a community of users may also interact more passively. Intelligent agent software such as Firefly can examine the behaviors of different users to collaboratively filter information and make recommendations to users on selecting information or other products based on the preferences of others with similar overlapping interests. Other web sites with similar "communal" features focus on a special area of interest. For example, www.techstocks.com focuses on high-technology and bio-technology stocks.

There are two main advantages to information vendors implementing community interaction strategies. First, it generates stocks of new, unique and proprietary content that is hard for competitors to replicate. Second, as noted by Hagel and Armstrong (1997) virtual communities can potentially create significant returns to scale in content production, reputation, and revenues for "community organizers". As a site becomes more useful it can generate more subscribers who in turn can contribute new useful content to the site generating positive externality benefits. This creates additional opportunities for advertising or transactional revenue. While community interaction strategies are in their infancy and there is much to be learned about organizing successful "virtual communities", we expect information vendors to increasingly migrate to implementing strategies to realize new sources of competitive advantage and revenue.

While there are many different ways to add value to data (see Taylor (1986) for a comprehensive analysis), the three directions identified in this section succinctly capture the key value-addition and differentiation choices confronted by information vendors in markets where competitors apply web information systems to deliver products to customers. While the low cost communications infrastructure of the Internet has made "access" to SEC data a free commodity to customers this has led data vendors to rethink the "value added" in their products as discussed in this section. Contrary to early expectations that the Internet would undercut existing SEC data vendors, it has provided them with opportunities to create new product that better serves existing customers and reach new customers to expand markets. Below we consider implications of the WWW for future public access to documents.

5.0 Web Information Systems: Implications for Public Access

Governments collect, process and generate large amounts of information. This information is a valuable resource to the public for decision-making, and a profitable resource for information vendors who add value to the data or disseminate it for a fee to the public. Beginning in the

1980s, the U.S. Federal Government increasingly privatized the dissemination of various types of electronic government information. In some cases private industry played a critical role in adding value to data and leveraging economies of scale in dissemination and specialized technical knowledge to efficiently disseminate data to the public for a fee. In other cases, private industry limited access by charging substantial fees for dissemination (See Love (1992) and Starr and Corson (1987) for reviews of private industry roles in disseminating data). Indeed the information industry has often sought to minimize direct government dissemination efforts based on the government's own guidelines for information management. Office of Management and Budget Circular A-130 provides government agencies with guidelines for electronic data dissemination requires agencies to try and recover the "cost of dissemination". Second it states that "As agencies are constrained by finite budgets, when there are several alternatives from which to choose, they should not expend public resources filling needs which have already been met by others in the public or private sector. Agencies have a responsibility not to undermine the existing diversity of resources." When dissemination was costly, these requirements were often used to limit government efforts at directly disseminating data. However, the 1993 version of Circular A-130 provides government agencies with a means of directly disseminating data "if legal consideration requires an official government dissemination product".

The Internet and WWW affects public access primarily by dramatically reducing the cost of dissemination to government and industry. Based on current industry practices, we estimate the fully loaded cost (staff, equipment and telecommunications) of the SEC Internet EDGAR dissemination system to be less than $1 million annually. If 400,000 files are transmitted daily this will approximate to less than a penny per filing for dissemination. As billing costs at present are higher, such a low cost makes it transaction cost inefficient to bill for and recover the cost associated with each filing disseminated to users. While the SEC EDGAR and other private systems provide free EDGAR systems that are near substitutes, there are a number of reasons why the SEC and other government should directly provide information over the Internet to the public. First, despite numerous alternatives and day-delayed documents, access to the SEC site continues to grow. This illustrates the public's confidence in data from the SEC. . Second, only the government site can be relied upon to maintain a free and exhaustive repository without access charges to the public in consistent and well documented formats. This enables specialized entrepreneurs to build diverse products and services that further expand the information industry. Third, similar public and private sector sites are not mutually exclusive on the Internet. The public site adds greater choice to the consumer of information, and provides a back-up to free resources elsewhere in the event of technical or network failure. Fourth, the interactive capabilities of the Internet allow the agencies to get valuable feedback on the types of information collected and disseminated, and ways of maximizing the value of filings to customers.

The falling costs of dissemination indicate that the critical debate should not center on whether the government should implement Internet dissemination systems. Instead, the critical question in the web-enabled world is how much value added should government owned or operated Internet sites provide? Given the advance of technology, we believe at a minimum government sites should provide timely data with full text indexing and retrieval of text documents. In the case of EDGAR, more timely access to documents--at a minimum should be 20 minutes delayed,

as is common for stock price information--should be combined with search methods for documents indexed by ticker symbols, SIC code, etc., and full text search. As discussed in the previous section, in a competitive market, technology leverage strategies provide only short term advantages to vendors. In response we expect vendors to migrate their strategies toward proprietary data aggregation, intelligence leverage and community strategies. Thus we do not see adverse impacts on innovations or markets in the information industry if governments limit themselves to providing generic public interfaces to aggregate or single sets of government data through currently inexpensive and widely available software technologies. These services would effectively be limited to the bottom left quadrant on the cube in Figure 2.

As illustrated by the challenges to create additional value to the EDGAR data, a more important role for government is to enhance public access over the WWW by focusing on the methods of collecting and generating public data. Much more effort must be spent on the design and adoption of electronic filing processes to take advantage of low cost digitization and ubiquitous networks. Specifically we encourage the SEC to investigate WIS for creating electronic disclosure filing mechanisms integrated with the workflow of disclosing organizations. The SEC and other government organizations should also enforce stronger standards for data quality and specify tags to classify objects of interest in the data disclosed to the public. The Extensible Markup Language (XML) (Khare and Rifkin, 1997) initiative to specify data formats for structured document interchange on the WWW shows promise as a means of implementing such tags in documents at the point of document creation. Such efforts are of value to the public and data vendors as it reduces the costs of extracting and using relevant data from documents, helps ensure data quality and enables easier migration of the information to higher value added products.

Finally, as government information becomes more accessible to the public, more effort is required to simplify the information where possible so that it is comprehensible to interested citizens. In the case of EDGAR, such efforts should be closely integrated with efforts to specify standards to mark-up key data within documents submitted to the SEC.

Conclusions

EDGAR on the Internet clearly demonstrates the capacity of the Internet to effectively disseminate government information in a timely and inexpensive way. As tools for web information systems development, and the infrastructure for data dissemination improve, government-owned web systems will satisfy the basic needs of the public for access to key government information. As WIS capabilities increase, data vendors will have to migrate value from simple value addition mechanisms which leverage the features of technology to competitive models that leverage "intelligence" applied to select "aggregate bundles" of data, customized to individual users and communities. These transformations will enhance customer value and create new markets for information vendors.

Acknowledgements

The authors would like to thank Harvey Neville, James Risser, and the anonymous reviewers for their comments and suggestions.

Center for Digital Economy Research
Stern School of Business
Working Paper IS-98-13

REFERENCES

Akerlof, G. The Market for Lemons: Quality, Uncertainty and the Market Mechanism, *Quarterly Journal of Economics*, *84,* 1970, 488-500.

Barney, J. Firm Resources and Sustained Competitive Advantage, *Journal of Management*, 17, 1991 p99-120.

Hagel, J. and Armstrong, A. *Net Gain: Expanding Markets through Virtual Communities.* Harvard Business School Press, 1997

Khare, R and Rifkin, A. XML: A Door to Automated Web Applications. *IEEE Internet Computing,* May-June 1997

Love, J. The Marketplace and Electronic Government Information, *Government Publications Review,* 19, 1992 p397-412.

Love, J. Pricing Government Information. *Journal of Government Information*, 22, No. 5, 1995 p363-387.

Neville, Harvey. Davis Polk & Wardwell. Private Communication, 1998.

Risser, James. Benjamin N. Cardozo School of Law, J.D. Candidate. Private Communication, 1998.

Silver, M. *Systems that Support Decision-Makers: Descriptions and Analysis*. John Wiley and Sons, Chichester, 1991

Starr, P., and Corson, R., "Who Will Have the Numbers? The Rise of the Statistical Services Industry and the Politics of Public Data," in *The Politics of Numbers*, ed. William Alonso and Paul Starr, Russell Sage Foundation, New York, 1987.

Taylor, R. Value Added Processes in Information Systems, ABLEX Publishing Corporation, Norwood, New Jersey, 1986

Venkatraman, N. and Kambil, A. The Check is Not in the Mail: Strategies for Electronic Integration, *Sloan Management Review,* Winter 1991