Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2004 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-5-2004

Electronic Disclosure and Financial Knowledge Management

Jerome Yen

Percy Yuen

Belinna Bai

Electronic Disclosure and Financial Knowledge Management

Jerome Yen, Percy Yuen, Belinna Bai

Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China
{jyen, ltyeun, lbai}@se.cuhk.edu.hk, Tel: (852) 2609-8237, Fax: (852) 2603-5505

ABSTRACT

In this paper we report the benefits of using eXtended Markup Language (XML) to support financial knowledge management, which include indexing, organizing, association generation, cross-referencing, and retrieval of financial information to support the generation of knowledge. The current searching engines cannot provide sufficient performance, such as, recall, precision, extensibility, etc, to support users of financial information. XML is able to partially solve such problem by providing tags to create structures. XML provides a vendor-neutral approach to structure and organize contents. XML authors are allowed to create arbitrary tags to describe the format or structure of data, rather than restricted to a specific number of tags given in the specification of HTML. A prototype of XML-based ELectronic Financial Filing System (ELFFS-XML) has been developed to illustrate how to apply XML to model and add value to traditional HTML-based financial information by cross-linking related information from different data sources, which is an important step in moving from traditional information management to knowledge management. We compared the functionality of XML-based ELFFS with the original HTML-based ELFFS and SEDAR, an electronic filing system used in Canada, and recommended some directions for future development of similar electronic filing systems.

Keywords: Financial Knowledge Management, Electronic Financial Filling Systems, XML

1. INTRODUCTION

Corporate filings, news articles, trading data, information about the listed companies, etc, are some most frequently requested data by investors. Majority of financial data or information are disseminated through Internet hoping that they will be widely available. However, searching, retrieving, and utilizing financial data have become a major challenge to both users and researchers. Such challenge includes inssues, such as, efficiency, precision, recall, relevance, and value creation. According to the 10th WWW User Survey, only around 25 percent of people used the WWW to access financial information on daily basis. The same survey also pointed out that more than 45 percent of the respondents felt that they were not able to find the information they needed. To solve these problems, content creation, representation, and management can be few areas to attack. It is possible to automate or semi-automate the process of content creation, collection, indexing, organizing, storage, and dissemination.

In this paper, we dissect the problems mentioned above into the life cycle of financial information and proposed a framework to use eXtended Markup Language (XML) to rectify these shortcomings. It is also hoped that it will help users to better manage financial *data* by systematically convert, synthesize, and create financial *knowledge* from *data and information*. It is a very important element in knowledge management [9]. To support this research, a prototype of XML-based ELectronic Financial Filing System (ELFFS-XML) has been developed to illustrate how to apply XML to model traditional HTML-based financial information and providing value added services. We compared the functionality of XML-based ELFFS with the original HTML-based ELFFS and SEDAR and recommend some directions for future development of similar electronic filing systems.

For the rest of the paper, Sections 2 provides a short discussion about related research works of financial filing systems as well as the issues and problems of organizing financial information on the Internet. Section 3 discusses the Financial Knowledge Management Framework. The implementation and major technologies of ELFFS-XML are discussed in Section 4. Experiment and user evaluation are discussed in Section 5. This paper is concluded with a discussion about future research.

2. FINANCIAL INFORMATION MANAGEMENT AND ITS ISSUES

Transparency and efficiency are two extremely important issues to the users of financial information [8]. Transparency can be defined as the visibility of a financial market and efficiency refers to the time and efforts that needed by users to search and retrieve the needed information. In order to make high quality decisions and narrow the gaps between private investors and corporate investors, both issues need to be solved.

Financial information/data were formerly disseminated through traditional media, such as five nationwide Securities and Exchange Commission (SEC) reading rooms in the US. To obtain the needed information, investors had to travel to these locations, search, and then make photocopies, which was extremely inefficient. Private companies, such as Mead Data Central, also provided financial information via dedicated connections with extremely high charge that not every investor could afford.

Since the advent of the Internet and hypertext, most financial information were digitised and disseminated through Internet. In the following, we will discuss the existing Electronic Filing Systems and current financial information management issues.

2.1 Existing Electronic Filing Systems

As Internet appeared, a systematic solution – the electronic financial filing system was emerged to aid parties, such as, filers, regulators, and investors to collect, process, store, and organize financial information. Filers usually are the companies that need to file financial documents to the regulatory bodies, such as Securities Exchange Commission (SEC). They are required by law to submit information for disclosure to the public. The regulators are responsible for collecting the submissions, processing them, guaranteeing the information received are correct, valid, and sufficient, and making them available to the public.

Different countries have different electronic financial filing systems. SEDAR [3] is a Canadian project that focuses on building financial information repository to support investors. It collects financial reports or documents from Canadian listed companies, which can be in different formats, and creates a web accessible archive for users. EDGAR [2] (Electronic Data Gathering, Analysis and Retrieval) is the US efforts in supporting collecting, indexing, organizing, and disseminating financial filings through Internet. It provides facilities for filing parties to format their reports in SGML. ELectronic Financial Filing System (ELFFS), was an initiative at the University of Hong Kong endorsed by the Securities and Futures Commission (SFC) of Hong Kong Government to support digitising, collecting, archiving, and disseminating financial information using traditional HTML [8].

2.2 Current Management of Financial Information on the Internet

We have carefully analyzed and studied the strengths and weaknesses of each filing system, and found there are still lot of rooms for all the filing systems to improve. An ideal electronic filing system should allow collecting documents from submitters via Internet to increase the speed of transmission as well as lower the transmission costs. It should support processing and validating the submissions, store and create archives, organize, index, and extract the useful clues, trends, and relationships buried in the data. This is a crucial step, which can be defined as knowledge generation.

From the day when the World Wide Web was born, a HyperText Markup Language (HTML) was needed to "mark up" contents in order to create the format that can be presented on the web. HTML, deals with two different views of the same piece of information: *what* the content is and *how* the content is presented. This is what *metadata* [4] is about. HTML documents are actually a mixture of both content and presentation instructions, which makes it very difficult to process by both human beings and machines. As some of the syntaxes used in HTML are not too strict, data extraction from a HTML document may not be even possible.

Over the past few years, Internet has become a major information source for investors and it provides information that covers almost all the major financial markets in the world. People started to create value from the huge amount of raw information, but processing it to create *knowledge* is difficult. Raw data exist in many forms, and when it comes to unstructured documents like HTML and all those generated by presentation-based software, content extraction becomes even more difficult. Quick adoption of Internet also accelerated the explosion of digital contents and information overload has become a major challenge. Data redundancy is another problem, e.g. one copy in the raw information and another copy in the presentable page. Keeping two sets of data not only cost more storage, but also cost more manage efforts.

2.3. A Survey about the User Issues

We conducted a survey to study the situation on the usage of financial information. 42 people are interviewed for the major problems they encountered when accessing financial information over Internet. 30 of them were investors, 10 were financial software developers and the remaining were information system administrators. Based on a scale of 1 to 10 (from least serious to the most serious), we found that the financial information on the Internet faced the following issues, the number in the bracket indicates the average score of severity:
The

1. **Too time consuming to gather** (8.13)
2. **Volatile to catch** (7.64)
3. **Distributed and fragmented in nature** (7.69)
4. **Bulky to store** (7.68)
5. **Too fast to collect** (7.54)
6. **Easily perishable** (7.49)
7. **Not unique** (7.34)
8. **Varying in content representation/format** (7.29)
9. **Difficult to be processed by human** (7.16)

3. FINANCIAL KNOWLEDGE MANAGEMENT

According to Thomas Davenport [7], knowledge management is a systematic attempt to create, gather, distribute and use knowledge. Matt Richtel [15] defines knowledge as information that is accessible and can be directly used to solve a problem while it is just raw and impenetrable. In simpler words, knowledge management is the methodology to add value to raw data and convert them into useful knowledge.

3.1 Financial Knowledge Management Framework

From the point of view of an information producer,

we break up the Financial Knowledge Management into seven stages: *information creation, information collection, information storage, knowledge generation, knowledge dissemination, knowledge presentation and feedback*. It was derived from the knowledge management processes that discussed in previous section and was adapted specifically to suit the needs for Financial Disclosure and the development of the prototype (Figure 1). We have briefly categorized the information consumers into three groups, namely, individuals, financial regulators and value adders. Due to the scope of the paper, we will only discuss the needs of individual investors.

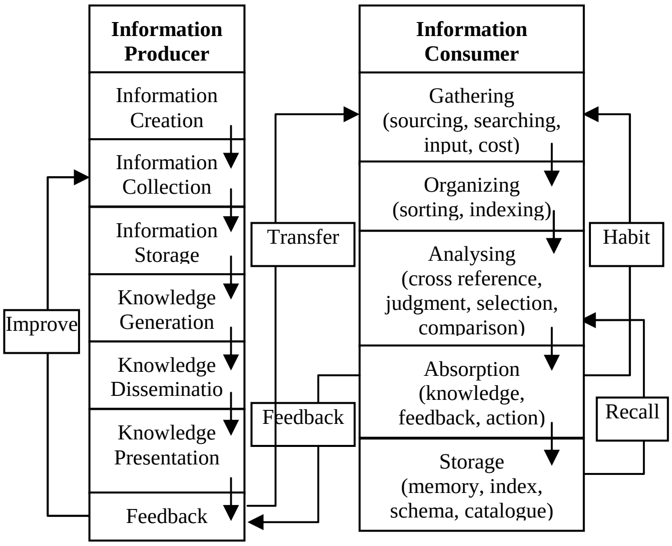
Information Producer
Information Creation
Information Collection
Information Storage
Knowledge Generation
Knowledge Disseminatio
Knowledge Presentation
Feedback
Improve
Transfer
Feedback
Information Consumer
Gathering (sourcing, searching, input, cost)
Organizing (sorting, indexing)
Analysing (cross reference, judgment, selection, comparison)
Absorption (knowledge, feedback, action)
Storage (memory, index, schema, catalogue)
Habit
Recall

Figure 1: Financial Knowledge Management Framework

3.2. Design of the Knowledge-based ELFFS-XML

Based on the Financial Knowledge Framework, we designed the general architecture of ELFFS-XML (Figure 2). In the following, we will discuss in detailed tasks in the each layer of the architecture.

3.2.1. Information Creation

In ELFFS-XML, we aimed to provide facilities for content creators to create structured contents using XML. The first step was the conversion of the Java-based cross platform Formatting Wizard for financial report filers. Users can input the data using a browser without knowing the details of XML.

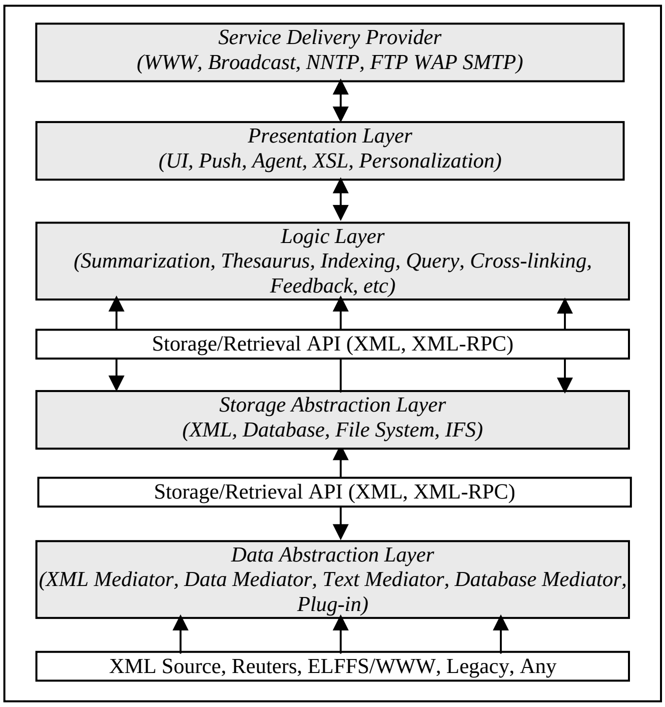
Service Delivery Provider
(WWW, Broadcast, NNTP, FTP WAP SMTP)
Presentation Layer
(UI, Push, Agent, XSL, Personalization)
Logic Layer
(Summarization, Thesaurus, Indexing, Query, Cross-linking, Feedback, etc)
Storage/Retrieval API (XML, XML-RPC)
Storage Abstraction Layer
(XML, Database, File System, IFS)
Storage/Retrieval API (XML, XML-RPC)
Data Abstraction Layer
(XML Mediator, Data Mediator, Text Mediator, Database Mediator, Plug-in)
XML Source, Reuters, ELFFS/WWW, Legacy, Any

Figure 2: Architecture of ELFFS-XML

3.2.2. Information Collection/Storage

There are many financial information sources available. We have developed a set of retrieval agents that can connect to a variety of data sources including ELFFS financial report archive, Reuters, WWW, financial knowledge bases and other legacy sources. Each of them is working independently and the data they collect is converted into XML formatted files through a standard Application Programming Interfaces (API).

The XML files are placed in a traditional file system and an archive is formed gradually. They are recombined in real time using the”SYSTEM” entity feature of XML. Although it may increase access time, it actually increases the flexibility of handling these data. Until a general XML mass relational storage mechanism has been developed, we are staying with the primitive file system approach.

3.2.3. Knowledge Generation

We have a multiplexing engine that chooses from different kinds of data sources and invoke the predefined logic for each source, such as data summarization and cross-linking. An XSL style sheet is applied to an XML document that contains the annual report repository. Because XSL can be used to perform many operations on XML tags, it is equivalent to a logic layer. We used a Java-Servlet Cocoon to perform this XSL transformation on the server side.

3.2.4. Knowledge Dissemination/ Presentation

The Extensible Style Language is used again here. Using various tools, the XSL processor can output different well-presented formats such as the Portable Document Format (PDF), HTML, WAP pages or even XML.

4. IMPLEMENTATION AND MAJOR TECHNOLOGIES OF ELFFS-XML

In this section, we will discuss the implementation and major technologies of ELFFS-XML.

4.1 Document Format

The financial document is XML formatted. It starts with the <filing> tags. There are 2 major "sections" inside, <data> and <body>. Inside <data>, there are a number of vital data about the document, such as document identifier, document title, date, and a couple of keywords. Inside <body>, there are <sidebar> and <maintext> tags and inside <maintext>, there are <paragraph> tags, where content of the report is mainly placed. Some key performance data about the report have been tagged, such as profit last year (<perf-profit-lastyr>), profit this year (<perf-profit-thisyr>). After this XML page is fed into an XML processor, a XML document tree is generated. Document contents and data can be easily extracted.

4.2 Generation of Tags and Schemas

In reality, almost all the filers do not markup their file using XML. Even if they do, they are using different DTDs. This imposes a new XML standard called eXtensible Business Reporting Language (XBRL), which attempt to provide a framework defined by a standards group for XML business report. In this paper, we attempted a different approach. We tried to use a combination of three methods to generate the tags, namely, statistical and information retrieval approach, accounting and auditing practice approach and investor feedback approach.

4.2.1 Statistical Information Retrieval

A total of 5000 news article related to South East Asia available from Reuters, 200 governmental financial articles and 20 annual reports from different listed companies were pooled together to create the archive to support the identification of tags. We used several techniques to support tag generation. Two approaches are involved: the first one is calculating term frequency and document frequency of terms and phrases in all the documents; the second is using the concept space approach [11] as well as a couple of online financial and business dictionary to create our ELFFS thesaurus. Then we have performed object filtering and automatic indexing for each of our collected articles mentioned above. A list of concepts was generated after we performed co-occurrence analysis. Term frequency and document frequency calculation were used to find out the combined weights of terms and phrases in the concept space. All the subjects in the automatically generated thesaurus were then used to help identify rotationally duplicated terms.

4.2.2 Accounting and Auditing Practise

To improve the reality and practicability of the meta data, we examined the code of practices of several accounting firms and auditing firms. An additional list of tags was generated after this process.

4.2.3 Investors' Feedback

Investors are usually the final recipient and consumer of the financial information. We have interviewed 20 investors about what information they want to see from different types of financial information.

5. EVALUATION OF ELFFS-XML

ELFFS-XML is evaluated by different criteria: system efficiency, tag generation performance and user satisfaction. Before the evaluation, we first take a look at how ELFFS-XML compares to other electronic financial filing systems.

5.1 Comparison with Other Financial Information Disclosure Systems

ELFFS-XML is compared with different such systems using the Financial Document Disclosure Model. The ELFFS and the SEDAR system were used as references (Table 1). It is not difficult to see that ELFFS-XML, in many ways, has trivial advantages.

5.2 Systems Efficiency

We have conducted a simple experiment revealing the size of an XML document can cost significantly more resources than a binary encoded document (store, network transfer, read, process.) We definitely need some ways to reduce the impact of this problem, but these are outside the scope of this paper.

5.3 XML Tag Generation Performance Evaluation

We conducted the experiment with the thesaurus agent (computer), a group of three experts in the field of accounting, finance and investment, and a group of five individual investors with different backgrounds. The thesaurus agent is running on a SUN-Enterprise 6500 server with 8 ULTRAsparc-II 400MHz CPUs and 3 GB of memory. Each of the experts and investors were given an accounting and a financial investment dictionary, twenty annual reports of Hong Kong, a list of twenty finance related web sites with a machine to access those sites, a terminal that can access and view real time financial data from Reuters and 5000 pieces of archived Reuters news articles related to South East Asia market. After reading and examining these materials, the three groups of participants were asked to produce a list of tags they think useful. They are given a total of ten hours to come up with the list with no limit on items. Both human and computer participants are only allowed to view web pages that are directed linked from the list of specified web sites to increase the fairness of the experiment. After the list was produced, the naming of the tags was adjusted so that they are consistent with each other in terms of semantic meanings.

	SEDAR	ELFFS	ELFFS-XML
Information Creation	extremely varied output format	formatting wizard , Java based , cross platform	formatting wizard, Java based, cross platform, keywords tagged, other XML aware editors supported
Information Collection	modem terminal upload	WWW secured via SSL	WWW secured via SSL
Information Storage	file systems based (usually 1 file), database	file systems based (usually multiple files), database	file systems based (usually multiple files), XML archive, optional legacy database support
Knowledge Generation	document indexes	document indexes, full text indexes	document indexes, full text indexes, cross referencing with online news articles, keyword search, multiple documents comparison and querying
Knowledge Dissemination	WWW	WWW, SMTP (email)	WWW, WAP, SMTP (email)
Knowledge Presentation	Word documents, proprietary formats, need special proprietary viewer	HTML	XML/XSLT, server side transformation to HTML, browser independent presentation, document map, PDF for printing

Table 1: Comparison of Different Financial Information Disclosure Systems

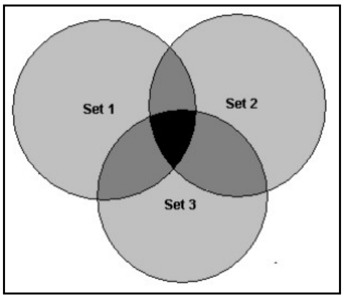

Figure 6: Different Sets of ELFFS-XML Tags

To quantify the performance of the different approaches, we view the selected tags from each of the three groups as separate sets as in Figure 6. Visually, the more dark area a set has with respect to its own size, the better is its performance. Mathematically, we use precision and recall as our evaluation metric. We slightly modified this concept and revised equations are:

$$P = \frac{number\ of\ good\ tags\ produced}{total\ number\ of\ tags\ produced}$$

$$R = \frac{number\ of\ good\ tags\ produced}{total\ number\ of\ good\ tags}$$

Visually, good tags are those tags inside the dark grey and black area. We used the F-measure to combine both ratings. The F-measure is defined as:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

β = 1 was chosen because we believe that precision and recall are equally important.

	# Correct	# Produced	Total Correct	P	R	F(1)
Investors	128	385	257	0.33	0.33	0.17
Experts	157	350	257	0.44	0.61	0.26
Agent	185	480	257	0.38	0.72	0.25

Table 2: Experiment Results for Different Tag Generation Approaches

6. CONCLUSION AND FUTURE RESEARCH

In this paper, we proposed a new approach to support indexing, organizing, and retrieval of financial information on the Internet using the eXtensible Markup Language (XML). The new approach includes models such as the document disclosure model, which was proposed to understand the underlying problems of information producers and consumers.

The strength of this new approach is that important data can be extracted relatively easily to facilitate easier creation of knowledge than the traditional HTML approach. ELFFS-XML can help to archive higher information transparency through the provision of filing standards and facilities to organize these information. End-users can gain benefits because they can obtain knowledge more easily. At the same time, it also provides multiple efficient delivery channels for end-users to receive the needed information at the lowest cost of time and money. It significantly bridges the gap between large corporate investors and small individuals, who are less resourceful.

We also found out that with the help of financial experts, computer agent, which automatically generate metadata or tags from a multiple documents collected from different sources, can provide very good performance in terms of precision and recall. Despite how ideal it sounds, there are still works need to be done. Currently, the DTD defined for ELFFS-XML is in an experimental stage, which lacks sophistication and details. Many financial performance tags are still missing. We are looking forward to expand it into a full-fledged XML vocabulary and let government agencies evaluate its usefulness.

Scalability is another issue. Searching through a large text-base XML repository on a file-system-based storage can be very I/O intensive, which may become the bottleneck to large systems. Moreover, there are still no standard ways to query XML archive. All we have now is a XML parser and a set of procedures about how to use the parser. We are currently implement dbXML and XML-QL, which are SQL-like languages for XML, to improve performance and provide more features.

REFERECES

[1]. Steve Lawrence and Lee Giles, "Accessibility and Distribution of Information on the Web", *Nature*, Vol 400, pp107-109, 1999.

[2]. Ajit Kambil , "The Direct EDGAR Access System: Final NSF Progress Report", http://www.stern.nyu.edu/~akambil/publications/nsfrep.pdf

[3]. "System for Electronic Document Analysis and Retrieval (SEDAR)", http://www.sedar.com/

[4]. Tim Berners-Lee, "Web architecture: Metadata", http://www.w3.org/DesignIssues/Metadata

[5]. Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen, "Extensible Markup Language (XML) 1.0 W3C Recommendation", http://www.w3.org/TR/REC-xml

[6]. Matt Richtel , "New Breed of Worker Transforms Raw Information Into Knowledge", New York Times, October 15, 1997.
http://chss2.montclair.edu/prdept/_pomo/00000008.htm

[7]. Thomas H. Davenport, et al. "Working Knowledge: How Organizations Manage What They Know", *Harvard Business School Press*, 1997

[8]. J. Yen, C. Yang, A. S. M. Lau, A. K. L Chung, K. Hwang, and D. Yeung, "Transparency and Efficiency in Information Transmission and Financial Digital Library", *Proceedings of the First Asia Digital Library Workshop*, pp195-206, 1998

[9]. Ikujiro Nonaka, Patrick Reinmoeller, and Dai Senoo , "The 'ART' of Knowledge: Systems to Capitalize on Market Knowledge," *European Mangement Journal*, Vol 16, No. 6, pp673-684, 1998

[10]. Narinder Singh, "Unifying heterogeneous information models: Semantic tags support knowledge webs", *Communications of the ACM*, Vol 41, No. 5, pp37-44, 1998

[11] Hsinchun Chen, Joanne Martinez, Tobun D. Ng, and Bruce R. Schatz , "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: Experiment on the Worm Community System", *Journal of the American Society for Information Science,* Vol 48, No. 1, pp17-31