Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2004 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2004

Identification and Visualization of Corporate Structure: A Preliminary Investigation

Richard Holowczak
Baruch College, CUNY

Charlie Peng
City University of New York

Sheridan Yeates
Baruch College, CUNY

Identification and Visualization of Corporate Structure: A Preliminary Investigation

Richard Holowczak
Baruch College, CUNY
richard_holowczak@baruch.cuny.edu

Charlie (Qian) Peng
Baruch College, CUNY
qian_peng@baruch.cuny.edu

Sheridan Yeates
Baruch College, CUNY
sheridan_yeates@baruch.cuny.edu

ABSTRACT

Interpreting corporate reports has largely remained the domain of accountants, corporate lawyers and analysts. As we have unfortunately witnessed, even the best representatives of these groups can be mislead by opacity in reporting and corporate structures. As mergers and acquisitions, foreign investment and other activities complicate the corporate landscape, constituents of all skill levels should be entitled to clear explanations of events and decisions that potentially affect shareholder value. We are actively in the process of developing several approaches and tools to automate the extraction of certain types of corporate information and to present this data in a useful fashion. In this paper we will outline the major steps involved in automating the extraction of corporate subsidiary information and will demonstrate several approaches for visualizing subsidiaries and their relationships.

KEYWORDS

Information extraction, corporate reporting, annual reports, corporate subsidiaries

INTRODUCTION

Today corporations are facing what is likely the most intense scrutiny in history. The spectacular failures of Enron, WorldCom, Parmalat, and a growing list of others has brought, both in the US and abroad, legislation (e.g., Sarbanes-Oxley) and other forms of regulation with the goal of requiring strict adherence to good accounting practices, specifically as implemented in corporate reporting. Several anecdotal discussions of massive corporate failure have focused on the information contained (some contend "buried") in corporate reports such as 10-K and 10-Q reports. The principal difficulty in divining poor corporate behavior from such reports lies in the manner in which such reports are structured and written. It is our contention that scrutinizing ever more complex filings will require new tools.

The area of focus for this study includes the subsidiaries that make up a large corporation. As was the case in both Enron and, more recently, Parmalat, the dealings that inflicted the most damage and that obfuscated the true financial health of the corporation were found to be carried out by subsidiaries that were, in many cases based in countries known for their liberal tax structure.

This study began with the following broad goals in mind: First, we are interested in developing an automated means of extracting the subsidiary structure from 10-K and 10-Q filings submitted to the Securities and Exchange Commission (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. Given sufficiently robust parsing and information extraction techniques, a side benefit will be the ability to produce more formal structures of subsidiaries, perhaps formatting using eXtensible Business Reporting Language (XBRL) (XBRL 2003). Second, we are interested in finding ways to represent, graphically, the subsidiary structure. Many corporations have hundreds of subsidiaries (in 1997 Enron had 281) and presenting this large amount of information in a comprehensible fashion is a challenge. Finally, we are interested in characterizing corporations using the attributes of their subsidiary structure and how that structure has evolved over time. Our aim here is to determine if patterns of subsidiary evolution can be used as a reliable predictor of corporate malfeasance. To date we have achieved reasonable success on the first two items and are actively working on the third.

There are a number of commercial data sources and systems that offer information on corporate subsidiaries. Two examples are Bloomberg and the Directory of Corporate Affiliations that is published by LexisNexis Group. Such services are not

inexpensive and may not provide all of the subsidiary's details. For example, in the case of Enron, Corporate Affiliations shows only 21 subsidiaries (essentially 3 levels deep) whereas we were able to extract 281 subsidiaries nested 5 levels deep.

CORPORATE REPORTING, INFORMATION EXTRACTION AND KNOWLEDGE DISCOVERY

From a computer information systems (CIS) perspective, the world of data is generally stratified into one of three categories. Structured data appears at one end of the continuum and includes data stored in spreadsheet cells, database tables and columns, and formatted text and binary files. The major advantages of structured data are the ease of manipulating them through queries and other operations. Unfortunately, the vast majority of data in the world is not structured. Free text documents, books, letters, magazine articles, web pages and so on are considered to be unstructured data. Other types of data such as e-mail messages fall somewhere in the middle. They contain some structured elements such as the Subject line of an e-mail message but also contain unstructured data such as the text of the message itself.

The general trend over the past decades of research has been to find approaches for imposing structure on unstructured or semi-structured data in an attempt to gain the ease of manipulation advantages of structured data. Recent work in formatting documents using eXtensible Markup Language (XML), for example, is one such approach, however it is one that addresses the problem at the document creation stage. The information retrieval community has approached this problem through a sub-field called information extraction (IE - also termed "automated content extraction" and "text mining") (Grishman 2003). The goal of IE is to extract structured facts or information from unstructured data. IE techniques have been used in a number of domains including extracting facts from news articles (MUC 1995), classifying legal documents (Holowczak, Adam, 1997) and querying legal document collections. A wide range of techniques have been applied to this problem including statistical analysis of text, natural language processing or a combination of the two (Cardie 1997).

10-K and 10-Q reports, as submitted to EDGAR, are in the semi-structured category. Some of the header information such as the company name, submission date, document type and so on is formatted. However, the majority of text in the remainder of the document is not. Subsidiary information is most often contained in Exhibit 21 or Exhibit 22. An initial investigation of a sample of 10-K reports revealed that subsidiary information is presented using a wide range of representations. In some cases, subsidiaries are simply listed, one per line, as text. In others, subsidiaries are formatted with leading spaces indicating the parent-child company relationship among the subsidiaries. Others have more elaborate HTML markup with HTML tables containing COLSPAN and division tags used to indicate the parent-child relationships. These main methods are summarized in Table 1, however dozens of variations on these main methods have also been observed.

The specific tasks that are required to extract subsidiary structure include 1) locating and downloading the 10-K document from EDGAR, 2) searching within the document for the existence of an appropriate exhibit and/or a reference to subsidiary information, 3) isolating the subsidiaries and determining their formatting, 4) applying an appropriate information extractor to extract the subsidiary information, the relationships among the subsidiaries and any additional data (such as ownership and state/country of incorporation) that might also be encoded in the exhibit file.

To gain a better understanding of the magnitude of the problem, we built a large index of SEC filings from the EDGAR database from years 1994 to 2003 (3,230,176 records) and filtered out filings other than 10-K's. For example, we filtered out NT 10-K, 10-K/A (amendments) and all other form types. The results are summarized in Table 2.

To test the effectiveness of our software on the second step (extracting the subsidiary exhibit) we took a random sample of 227 10-K documents, processed them using our software extractors and produced recall and precision statistics to assess their performance. The software we developed uses custom Perl programs that look for a handful of specific word and phrase patterns in the 10-K documents Given a sample of *N* documents, we assign each case to one of four categories A, B, C, and D as shown in Table 3.

Recall indicates the percentage of all documents with subsidiary information that were identified	= A / (A + C) =	98.3%
Precision indicates the percentage of identified documents that indeed contained subsidiaries	= A / (A + B) =	99.1%
Error indicates the percentage of documents that were misidentified (category B and C)	= (B + C) / N =	1.3%
Accuracy indicates the percentage of correctly classified documents (1 – Error)	= (A + D) / N =	98.7%

Of the 227 10-K reports, our software automatically extracted 116 subsidiary exhibits. In all but one case, where an exhibit was extracted, it was found to contain subsidiary information. This leads to a very high precision (99.1%). In only two cases did subsidiary information appear in a 10-K document and our extraction routines were not able to locate them. This results in very high recall as well (98.3%). In the remainder of cases (109), no subsidiary information was present. Often subsidiaries have not changed since the previous filing so a reference is given in the current filing to this effect. In addition, several companies do not have subsidiaries at all.

The result of this initial study demonstrates that in our sample, we are able to automatically extract exhibits containing subsidiaries from a high percentage of the 10-K reports. A more rigorous and extensive study is presently being planned.

Description: Indented text using number of spaces to indicate relationships

Example:

```
                        ENRON CORP. SUBSIDIARIES

 Atlantic Commercial Finance B.V. (The Netherlands)
    Enron Colombia Transportation B.V. (The Netherlands)
    Enron Power Holdings C.V. (The Netherlands) (99.9%)
    Enron Power Honduras C.V. (The Netherlands) (99%)
       Enron Power Honduras S. de R.L. de C.V. (Honduras) (99%)
    Offshore Power Production C.V. (The Netherlands) (99.9%)
       Enron Mauritius Company (Mauritius)
          Dabhol Power Company (India)
```

Description: Fixed width text columns

Example:

```
            OLYMPIC CASCADE FINANCIAL CORPORATION
Subsidiaries of the Registrant                    Percentage
                                                  of Voting
                                     State of     Securities
SUBSIDIARY NAME                    INCORPORATION    OWNED
National Securities Corporation     Washington       100%
WestAmerica Investment Group        California       100%
```

Description: Column position markers with relationship indicated by the column OWNER

Example:

ENTITY NAME	OWNER	%
<S>	<C>	<C>
Acajutla Company (Cayman Islands)	El Paso Energy Acajutla Company	100
AES/Sonat Power L.L.C. (VA)	El Paso Merchant Energy Holding Co. (DE)	50
	Unaffiliated Party	50
Agua del Cajon Company (Cayman Islands)	El Paso Neuquen Holding Company	50
	Unaffiliated Parties	50
Aguaytia Energy del Peru S.R.Ltda. (Peru)	Aguaytia Energy L.L.C.	22
	Unaffiliated Parties	77
Aguaytia Energy L.L.C. (DE)	EPED Aguaytia Company	24.3
	The Maple Gas Development Corporation	16.96
	Unaffiliated Parties	58.74
Ajax Corporation S.A. Sucural (Argentina)	Ajax Corporation S.A. (Uruguay)	100

Description: HTML table using COLSPAN and DIV tags to indicate relationships

Example:

Name	State or Jurisdiction of Entity
The Goldman Sachs Group, Inc.	Delaware
Goldman, Sachs & Co.	New York
Goldman Sachs (Asia) Finance Holdings L.L.C.	Delaware
Goldman Sachs (Asia) Finance	Mauritius
Goldman Sachs (UK) L.L.C.	Delaware
Goldman Sachs Holdings (U.K.)	United Kingdom
Goldman Sachs International	United Kingdom
GS Financial Services L.P. (Del)	Delaware
Goldman Sachs Capital Markets, L.P.	Delaware
Goldman Sachs (Japan) Ltd.	British Virgin Islands
J. Aron Holdings, L.P.	Delaware

Description: Paragraphs with relationship indicated in the text

Example:

SUBSIDIARIES OF THE REGISTRANT
All of the Company's subsidiaries are wholly owned by the registrant or a subsidiary of the registrant:
Vermont Pure Springs, Inc. ("Springs") - incorporated in the State of Delaware.
Crystal Rock Spring Water Company ("Crystal Rock") - incorporated in the State of Connecticut.
Excelsior Springs Water Company, Inc. - incorporated in the State of New York and wholly owned by the Company's subsidiary, Springs.
Adirondack Coffee Services, Inc. - incorporated in the State of New York and wholly owned by the Company's subsidiary, Springs.

Table 1 Examples of Subsidiary information included in 10-K reports

Document Type	Description	Documents	Companies
NT 10-K	Notification of Late Filing	15,585	7,542
10-K/A	Amendments to prior 10-K filing	41,248	8,249
10-K	Annual report	57,023	16,610
Total		113,856	20,339

Table 2 10-K and related documents and companies identified in SEC EDGAR database (1994-2003)

	10-K Contains Subsidiary Exhibit	10-K Does Not Contain Subsidiary Exhibit
Extracted Subsidiary Exhibit	A = 115	B = 1
Did NOT Extracted Subsidiary Exhibit	C = 2	D = 109

A is the number of cases where subsidiary exhibits were correctly identified and extracted (correct)
B is the number cases where the extractor misidentified subsidiary exhibits when none were present (error)
C is the number of cases where subsidiary exhibits were not identified (missed) by the extractor (error)
D is the number of cases where subsidiary exhibits were not identified and no subsidiary information exists (correct)

Table 3 Categories of documents used for performance evaluation

Identifying which reports contain subsidiary information and extracting the exhibit is only a part of the problem, however. As mentioned above, subsidiary information is presented in a wide range of formats and styles. Five main styles were revealed in our sample data set as shown in Table 4. Presently, we have in place extractors capable of handling indented text such as that shown in the Enron example, as well as subsidiaries aligned in a column format as in the third example in Table 1. The output of these extractors is a database-ready table of data containing all of the subsidiary information, relationships and ancillary data (state or country of incorporation, etc.).

Subsidiary format	**Number of subsidiary exhibits (total=115)**
Indented list	35
Fixed width columns	34
Column indicators	29
HTML	10
Paragraphs	7

Table 4 Subsidiary exhibit formats observed in sample

Once subsidiary information has been extracted, information extraction techniques can then be employed elsewhere in the document (and other related documents) to automatically extract and collect additional facts about each subsidiary. Ideally, the subsidiaries and all of their related information will be stored in a database to facilitate further manipulation and retrieval. At this time, we are exploring additional extraction techniques, including bootstrapping methods (Yangarber, et. al. 2000) to carry out this step.

Once this information has been transformed into a structured format, a wide range of opportunities exist to manipulate and present the data. Expressing the subsidiary data in XBRL would be relatively straightforward and facilitate interchange of such data with third parties. Visualizing subsidiaries is another such task that is described in the following section.

VISUALIZATION OF CORPORATE STRUCTURE GRAPHS

Many companies have a large number of subsidiaries. Representing such a large number of subsidiaries, their relevant information and their relationships presents several challenges. Initially, we treated the parent-child subsidiary relationships

as precedent relationships that are easily represented using the DOT language (Koutsofios and North 2002). DOT is a part of the GraphViz graph drawing software developed at AT&T Research. The DOT algorithms automatically produce a layout of the graph nodes (subsidiaries) and arcs (ownership relationships) and can write to most standard image file formats such as bitmap, graphic interchange format (GIF) and Adobe Acrobat files.. The result is a clear depiction of all of the relevant subsidiary information. A partial example for Goldman Sachs Group 2001 is shown in Figure 1. Rectangles represent domestic US subsidiaries, ovals represent non-US subsidiaries and arcs are labeled with the percentage ownership. Each node is further labeled with the name of the subsidiary and the country or US State where it is incorporated.

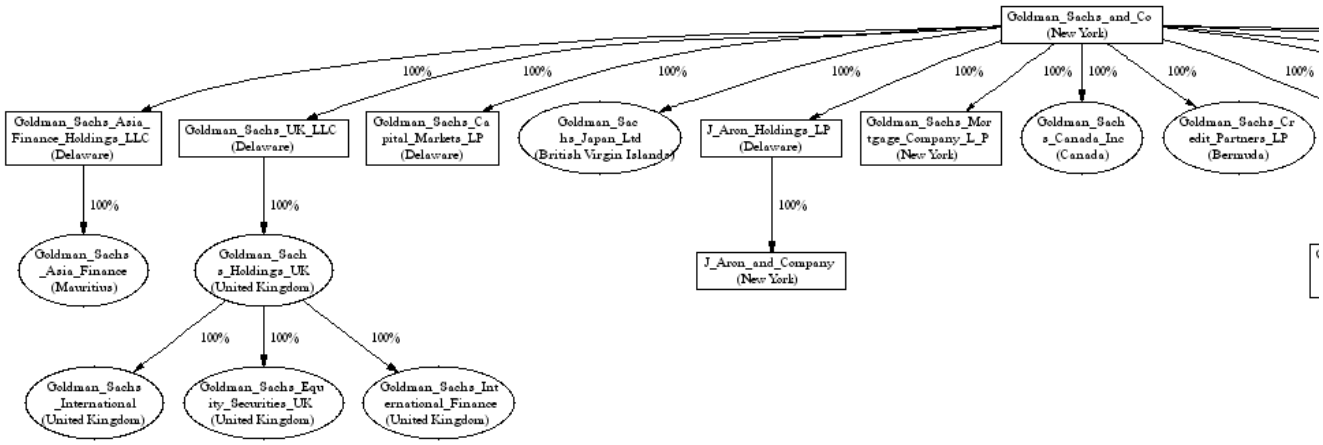

Figure 1 The 2001 Goldman Sachs Group subsidiaries produced by DOT (partial)

Figure 1 demonstrates the primary limitation of this representation. To remain legible, the graph must be drawn at a reasonable size. Goldman Sachs had at this time 27 subsidiaries and a readable printout of such a graph requires two 8.5x11 pages be laid side by side. Enron's 281 subsidiaries require a dozen such pages to display legibly.

A second approach taken was to feed the subsidiary information into the Microsoft Visio drawing program. This technique can be automated so as to produce a reasonably good layout with virtually no human intervention. An example of this is shown in Figure 2.

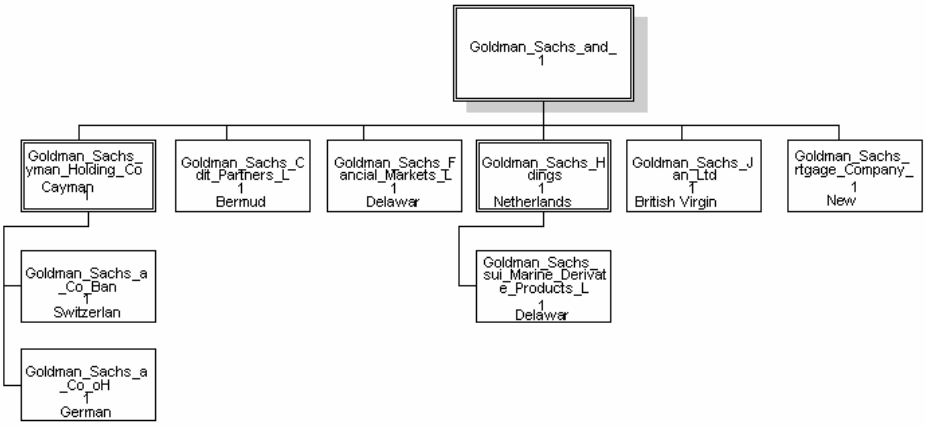

Figure 2 A portion of the 2001 Goldman Sachs subsidiaries produced in MS Visio

A third approach taken was to build a web application on top of the subsidiaries database and make use of an expandable outline Java applet to reveal or hide the subsidiaries. As shown in Figure 3, clicking on a subsidiary causes related company information to appear in the right hand window. Additional information such as the subsidiary company website or facts on the subsidiary pulled form the 10-K reports could also be displayed.

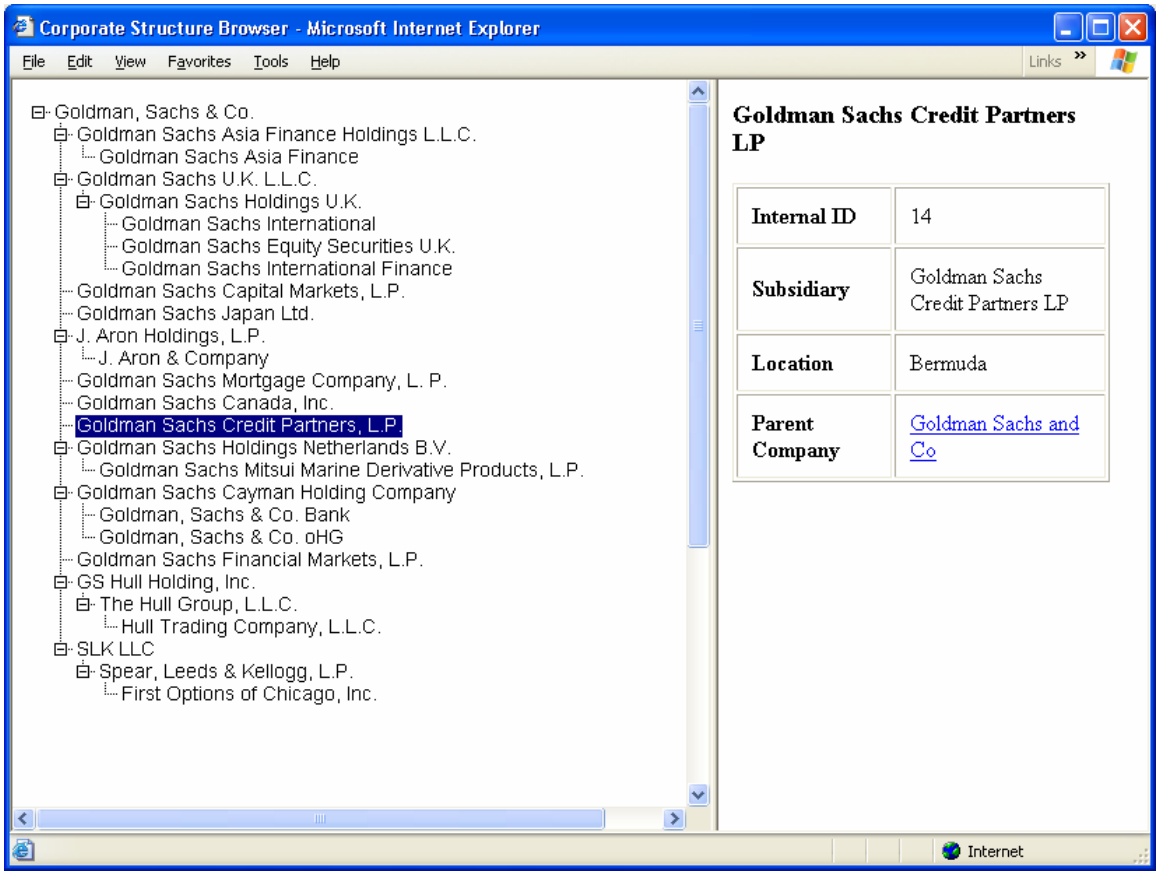

Figure 3 The 2001 Goldman Sachs subsidiaries displayed in a web application

CONCLUSION

In this paper we have outlined the motivation for, and challenges of, automating the extraction of corporate subsidiary information from annual reports. Such work presents a novel intersection of computer information systems and accounting research and clearly there is significant work that remains to be accomplished before completely automated extraction can be achieved. Even if subsidiaries in future corporate reports are encoded using XBRL, analysis of subsidiaries over time will still necessitate approaches such as what we have outlined in this paper.

REFERENCES

1. Advanced Research Projects Agency (1995) *Proceedings of the Sixth Message Understanding Conference* (MUC 6). Morgan Kaufman. November 1995. Columbia, Maryland, USA.

2. Cardie, C. (1997) Empirical method in Information Extraction. AI Magazine. 18(4). Page 65-80.

3. Extensible Business Reporting Language (XBRL) 2.1 Recommendation 2003-12-31. (2003) Editors: Engel, P., Hamscher, W., Shuetrim, G., vun Kannon, D., and Wallis, H.. Available on http://www.xbrl.org/resourcecenter/specifications.asp?sid=22

4. Grishman, R. (2003) Discovery Methods for Information Extraction. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo Institute of Technology, Tokyo, Japan, April 13-16, 2003.

5. Holowczak, R. D. and Adam, N. R. (1997) Information Extraction based Multiple-Category Document Classification for the Global Legal Information Network. *Proceedings of the Ninth Annual Conference on Innovative Applications of Artificial Intelligence* (IAAI-97). July 1997. Providence, Rhode Island. Page 1013-1018.

6. Koutsofios, E. and North, S. C. (2002) Drawing Graphs with dot, Available on research.att.com in dist/drawdag/dotguide.ps.Z

7. Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000) Automatic acquisition of domain knowledge for information extraction. *Proceedings for the 18th international conference on computational linguistics*. Germany, July-August 2000. Page 940-946.