DECISION SCIENCES INSTITUTE

Financial Data Analytics and Extraction from Security Exchange Commission's (SEC) EDGAR Filings

Richard Green
Texas A&M University - San Antonio
Email: Richard.Green@tamusa.edu

Robert Burdwell
Texas A&M University - San Antonio
Email: Robert.Burdwell@tamusa.edu

Robert Vinaja
Texas A&M University - San Antonio
Email: Robert.Vinaja@tamusa.edu

ABSTRACT

This is a proposal to provide specific industry, business, and financial data to business students and faculty. The data are appropriate for empirical research into issues that involve publicly-owned corporations that are listed on public stock exchanges in the United States. This project involves the development of programs and scripts to make accessing and using the data of the Security Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) for the purpose of primary empirical academic research.

KEYWORDS: XBRL, EDGAR, Security Exchange Commission, Financial Statement Analysis

INTRODUCTION

This project involves the development of programs and scripts to make accessing and using the data of the Security Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) for the purpose of primary empirical academic research.
The EDGAR system holds all mandatory filing data for over 9,700 corporate entities. The data are complete from April of 2005 through March 2015. The EDGAR standard taxonomy includes nearly 14,000 terms and can be expanded. This project will start by extracting and analyzing the last five years.
Accessing this data is a daunting task for most would-be researchers due to the sharp learning curve of mastering the details of electronic filing, the extensible business reporting language (XBRL), downloading and parsing the filings.
In fact, the tasks of downloading and parsing of the data are quite straight-forward and can be easily performed by appropriate computer programs. However, at this time, there is NO freely-available, user-friendly set of programs for this purpose.

LITERATURE REVIEW

CRSP - The Center for Research in Security Prices provide a comprehensive database for historical security prices and returns information. Research Data Access CRSP provides data access.

Python- XBRL is a library for parsing XBRL documents, it is a set of scripts that use python to parse XBRL marked-up text. There is also open source python middleware to add XBRL handling to basic python.
Arelle is a project to provide an easy to use open source facility for XBRL. Arelle includes Edgar manual validation. We have downloaded Arelle got the RSS feed to work for all filings. The next step is to export the data to a database. There is also a CVS file of the column headings that can be downloaded. This feature provides a real time saver for the approach to extract the data to MySQL.
Pysec compiles a list of all SEC filings from EDGAR into SQL, downloads them, and parses 50+ key accounting terms from XBRL filings. Pysec is also an open source Python XBRL parser that can extract arbitrary XBRL terms. The "guts" of Pysec were coded by Charles Hoffman, who was one of the originators of XBRL. One of the objectives is to extend this code to extract more than 50 accounting terms. This approach should produce a demo program before the development of the final product.
This approach based on Python scripts does not require loading the data into MySQL or Oracle. One of the potential advantages in that this approach can save a significant amount of disk space and processing time. One of the disadvantages of this approach is that Pysec is a combination of Python scripting (through a version called Django) and Visual Basic. Another limitation is that Pysec does not read XBRL in a zipped folder. Rather, as one of its developers wrote, "...command sec_import_index compiles a list of all SEC filings from EDGAR into SQL, populating the Index model, and will lazily download them as you access them..." EDGAR filings are provided in a zipped file and the ability to search the filings in compressed format would be very efficient. After installing Python and Pysec, one of the required steps is to adjust the settings.py, and modify the DATA_DIR = '/you/directory/to/download/files/to' and set your database. Currently, we had a few issues with the Pysec Phython application. There are a few errors each time pages are run. Thus, even though the program assumes that a copy of the sec filings in XBRL zip files is resident on the local computer, it is parsing them into SQL.
Another approach is to put the data into a MySQL or Oracle Express database, even if only temporarily for processing. Oracle XBRL Extension uses the XML DB feature of Oracle to give solution for collecting, storing, querying, analyzing, and managing XBRL content.
XBRLAnalyst is a financial data platform for Excel that leverages XBRL. XBRLAnalyst allows you to download 20k filings per month for $10 per month. Another alternative is to publish these exports on a web server and then conduct data analysis with Tableau. The same vendor also features a VBA Analyst download and API integration for developers.
Another commercial product by Altova, Raptor XBRL validates the instances against the taxonomy and verifies the formulas and definitions

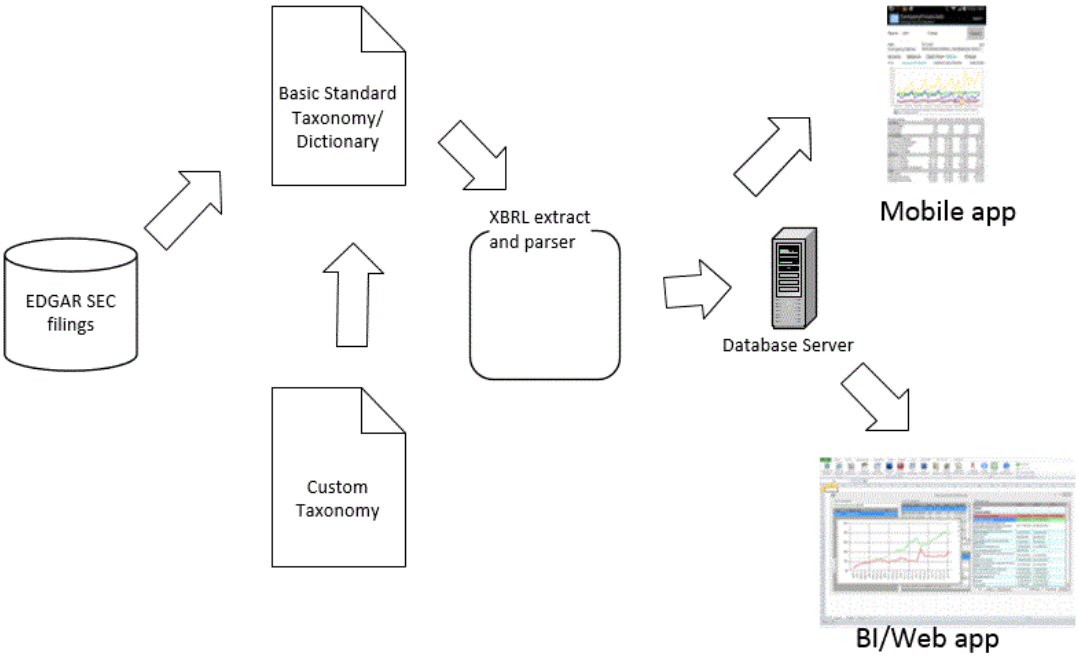
METHODS AND MODEL

The principal investigator has already developed Python scripts that complete the functions of downloading, saving, and searching the EDGAR data. The next objective is to parse the data once it is downloaded. An additional objective is to make the scripts “user-friendly”. A future step is to create an interface to extract these EDGAR filings.
The user interface should allow the user to search for specific companies. Also, the user should be able to search by keywords, then identify the taxonomy items, and use checkboxes to select a specific item from the taxonomy, for instance: depreciation related items. The interface should also provide a search capability to search and navigate the taxonomies.

If the project is installed on the university platform, there will be some small additional demand on the university network servers, as the approximate size of the unzipped database will be

approximately 52 GB. The compressed database with EDGAR files is about 14 GB (about three years' worth of data), and unzipped is 52 GB. The unzipped data will be hosted on MySQL. Since an XBRL element can have a variable number of items, it would be difficult to map directly it to a relational table, because the table with variable columns. The database will be regularly updated by using RSS. The database size will grow over time. At this time, a copy of the proposed database is maintained on a desktop computer which is a 4-core Intel i7 processor and a 600 GB hard disk.
More text about this . . . Figure 1 is a tabular summary of the frequency of work in this stream.

Figure 1: Implementation Model



RESULTS

This project plan includes the following steps:

1. Familiarize all participants in the project with XBRL and EDGAR.
2. Select a program for the project from those currently available, including C, C++, Java, and the Perl and Python scripting languages.
3. Develop a description of the desired Graphical User Interface (GUI)
4. Develop a "pseudo-code" (flow chart and scripting) description of the code to be written.
5 Write and test the code.
6. Enlist "Beta" testers from faculty.
7. Debug and address issues from the Beta test phase.
8. Finish and distribute the program to interested faculty and students.

The next step is to develop a mobile app and/or web application to access the data. Use business intelligence analytics software (like Tableau) to create dashboards for comparing financial ratios. Some of these important ratios include asset turnover ratio, inventory turnover

ratio, and liquidity ratio. Ratio analysis is used for selecting companies based on industry ratios, evaluating financial strength based on financial statement analysis based on ratios. A sample spreadsheet providing an example of financial statement ratio analysis is provided.
The availability of this application would facilitate research in additional areas such as: analyzing the frequency of errors in XBRL financial documents and the potential impact..

DISCUSSION AND CONCLUSIONS

Benefits of this project include enabling original primary empirical research to be conducted without having to invest tens of thousands of dollars in proprietary database products such as S&P Capital IQ Fundamentals ® from Standard & Poor's Financial Services. Additionally, this project requires no resources that are not already in use in the host university.
This project involves three faculty members but will also involve numerous students, primarily students in computer science. The students, under the direction of faculty, will develop additional programs to make this data easily available to interested users.
Costs of completing this project are minimal. One of the development options requires no software to purchase nor any database to license.

REFERENCES

Bharosa, N., van Wijk, R., Janssen, M., de Winne, N., & Hulstijn, J. (2011). Managing the transformation to standard business reporting: principles and lessons learned from the Netherlands. 12th Annual International Conference on Digital Government Research (dg.o 2011), 151–156. doi:10.1145/2037556.2037578

Borges, F. C. R., & Silva, P. C. da. (2012). A Framework for Processing Business Financial Rules (pp. 47–50). New York, NY, USA: ACM. doi:10.1145/2382636.2382648

Carrillo, E., Chaparro, F., & Santoyo, J. (2008). XBRL and financial information standards: a case success: university-enterprise-government in Universidad Autonoma de Bucaramanga. EATIS '08 Proceedings of the 2008 Euro American Conference on Telematics and Information Systems, 1–3. doi:10.1145/1621087.1621134

Data Access Tools. (n.d.). Retrieved April 8, 2016, from http://www.crsp.com/products/software-access-tools

EDGAR Dashboard. (n.d.). Retrieved April 8, 2016, from https://edgardashboard.xbrlcloud.com/edgar-dashboard/

Fan, W., Geerts, F., Li, J., & Xiong, M. (2011). Discovering Conditional Functional Dependencies. IEEE Trans. Knowl. Data Eng., 23(5), 683–698. Retrieved from http://dl.acm.org/citation.cfm?id=2460396.2460414

Felden, C. (2011). Characteristics of XBRL adoption in Germany. Journal of Management Control, 22(2), 161–186. doi:10.1007/s00187-011-0134-7

Fischer, H., & Mueller, D. (2011). Open Source & XBRL: the Arelle® Project (pp. 29–30). Retrieved from http://eycarat.faculty.ku.edu//myssi/_pdf/4-Fischer -Mueller-Open-source-ArelleProject.pdf

García, R., & Gil, R. (2010). Linking XBRL financial data. In Linking Enterprise Data (pp. 103–125). New York, NY, USA: ACM. doi:10.1007/978-1-4419-7665-9_6

GitHub - Arelle/Arelle: Arelle open source XBRL platform. (n.d.). Retrieved April 8, 2016, from https://github.com/Arelle/Arelle

GitHub - lukerosiak/pysec: Parse XBRL filings from the SEC's EDGAR in Python. (n.d.). Retrieved April 8, 2016, from https://github.com/lukerosiak/pysec

Hoffman, C., & Watson, L. (2009). XBRL For Dummies. For Dummies. Retrieved from https://sourceforge.net/projects/rivetdragonview/

Huang, M., Wang, D., & Wang, K. (2011). Ontology-based Semantic Retrieval of XBRL Data. 2012 Second International Conference on Business Computing and Global Informatization, 0, 363–366. doi:http://doi.ieeecomputersociety.org/10.1109/BCGIn.2011.97

index.html. (n.d.). Retrieved April 8, 2016, from http://www.oracle.com/technetwork/database/database-technologies/xbrl-extension/overview/index.html

Janssen, M., & Tan, Y.-H. (2014). Dynamic Capabilities for Information Sharing: XBRL Enabling Business-to-Government Information Exchange. 2014 47th Hawaii International Conference on System Sciences, 0, 2104–2113. doi:http://doi.ieeecomputersociety.org/10.1109/HICSS.2014.266

Jimei, L., Yuzhou, H., & Meijie, D. (2013). XBRL in the Chinese Financial Ecosystem. IT Professional, 15(6), 36–42. doi:http://doi.ieeecomputersociety.org/10.1109/MITP.2013.59

Kotsiantis, S. B., Kanellopoulos, D., Karioti, V., & Tampakas, V. (2009). An ontology-based portal for credit risk analysis. Computer Science and Information Technology, International Conference on, 0, 165–169. doi:http://doi.ieeecomputersociety.org/10.1109/ICCSIT.2009.5234452

Li, J., Zhao, H., & Du, M. (2012). Analyzing Semantic Heterogeneity in XBRL Taxonomies: An Ontology Perspective. Management of E-Commerce and E-Government, International Conference on, 0, 264–268. doi:http://doi.ieeecomputersociety.org/10.1109/ICMeCG.2012.48

Liu, C., Yao, L. J., Sia, C. L., & Wei, K. K. (2013). The impact of early XBRL adoption on analysts' forecast accuracy - empirical evidence from China. Electronic Markets, 24(1), 47–55. doi:10.1007/s12525-013-0132-8

Madlberger, L., Thöni, A., Wetz, P., Schatten, A., & Tjoa, A. M. (2013). Ontology-based Data Integration for Corporate Sustainability Information Systems. Proceedings of International Conference on Information Integration and Web-Based Applications & Services - IIWAS '13, 353–357. doi:10.1145/2539150.2539208

Pinsker, R., & Li, S. (2008). Costs and benefits of XBRL adoption. Communications of the ACM, 51(3), 47–50. doi:10.1145/1325555.1325565

python-xbrl 1.0.2 : Python Package Index. (n.d.). Retrieved April 8, 2016, from https://pypi.python.org/pypi/python-xbrl/1.0.2

python-xbrl-middleware on PyPI - Libraries. (n.d.). Retrieved April 8, 2016, from https://libraries.io/pypi/python-xbrl-middleware

Rivet Software Dragon View XBRL Viewer download | SourceForge.net. (n.d.). Retrieved April 8, 2016, from https://sourceforge.net/projects/rivetdragonview/

Rosiak, L. (n.d.). OpenGov Voices: PySEC, bringing corporate financial data to the masses. Retrieved April 8, 2016, from https://sunlightfoundation.com/blog/2013/08/02/opengov-voices-pysec-bringing-corporate-financial-data-to-the-masses/

Wang, W., Huang, M., & Wang, Z. (2011). A Storage and Query Mechanism of XBRL Data Based on Native XML Database. 2012 Second International Conference on Business Computing and Global Informatization, 0, 497–500. doi:http://doi.ieeecomputersociety.org/10.1109/BCGIn.2011.131

XBRLAnalyst Delivers Financial Data To Excel | FinDynamics. (n.d.). Retrieved April 8, 2016, from https://findynamics.com/try-xbrlanalyst/

Yingchun, S., & Baohua, T. (2010). Research on the Disclosure Quality of Financial Reporting on the Internet Based on XBRL Technology. 2012 Fourth International Conference on Computational and Information Sciences, 0, 605–608. doi:http://doi.ieeecomputersociety.org/10.1109/ICCIS.2010.153

Zambrano, E., & Samper, J. (2007). XBRL: from common financial vocabularies to intelligent decision making. EATIS '07 Proceedings of the 2007 Euro American Conference on Telematics and Information Systems, 2–5. doi:10.1145/1352694.1352770

Zhu, H., & Wu, H. (2011). Quality of data standards: framework and illustration using XBRL taxonomy and instances. Electronic Markets, 21(2), 129–139. doi:10.1007/s12525-011-0060-4