Integration of Structured and Unstructured Data in the Financial Analysis Domain - A State of the Art

Research-in-Progress

Maryam Heidari
TU Bergakademie Freiberg
maryam.heidari@bwl.tu-freiberg.de

Carsten Felden
TU Bergakademie Freiberg
carsten.felden@bwl.tu-freiberg.de

Claudia Koschtial
TU Bergakademie Freiberg
claudia.koschtial@bwl.tu-freiberg.de

ABSTRACT

There is a consensus (Hoffman and Strand, 2001; Hannon, 2002; Bovee et al., 2005; Willis, 2005; Cox, 2006) that XBRL (Extensible Business Reporting Language) as a technical standard for facilitating transfer and analysis of financial statements could improve the speed and quality of transmitting, analyzing, and even more accurate financial reports by providing machine readable documents. Footnotes, which include important explanations about financial values, have still an unstructured format and are an obstacle for analysts and other stakeholders who want to benefit from analyzing financial statements. It is of interest how data integration approaches can support and facilitate the process of data extraction from footnotes automatically to gain accurate and reasonable analysis to avoid manual tasks. To address this issue, a state of the art is needed to identify and cluster relevant existing methods in terms of structured and unstructured data integration. It is shown that most of the existing literature is focused on a storage level of data integration. Other researchers deal with methods and tools to integrate and analyze structured and unstructured data separately. But, no identified paper illustrates an unstructured data integration solution to support analytical tasks based on XBRL documents.

Keywords
unstructured data integration, state of the art, footnotes, XBRL, use case

INTRODUCTION

In the recent decade with the advent of XBRL, financial analysis has observed a great deal of development in analyzing phase of financial data (Janvrin and Mascha, 2010). Although, implementing XBRL applications can make the analyzing processes faster and accurately (Hoffman and Strand, 2001; Hannon, 2002; Bovee et al., 2005; Willis, 2005; Cox, 2006), but there still exist unstructured parts (notes), which are not separable from any financial statement. Each financial document consists of two different kinds of data: structured (values, percentages, etc.) and unstructured, e.g. (foot)notes. The understanding and usage of both data types is critical to users who benefit from using financial information like analysts, investors, auditors, or external decision makers. Without footnotes, financial reports are usually not fully understandable and might be somehow misleading (Putra, 2008). Footnotes are complementary disclosures in context of understanding and interpreting financial values.

According to Weglarz (2004), XBRL provides an infrastructure to standardize financial reports of companies, but the automated relation between values and footnotes is an obstacle except to physically read them. “When asked what additional information they would like to see reported in XBRL, many analysts ‘wish lists include footnotes.’” (NRI survey, 2011). Financial footnotes are divided into two main categories: first, the main evaluation policies used by the company (i.e. LIFO, FIFO) are explained and have to be made clear in footnotes. Second, some additional explanations, which are not possible to put into the main body of the financial statements (i.e. lease payments, interest rate, due dates, etc.). This makes obvious that the list of relevant items, which can be defined in footnotes, are relatively long. (Putra, 2008) Therefore, it is the paper’s goal to categorize relevant research in terms of an integration of structured and unstructured data in context of XBRL to gain an...
overview and identify a research gap of a meaningful IT support within the domain of financial analysis. Hereby, to address this issue, this paper contributes to find out and cluster approaches and methods which existing in data integration literature and consider that to what extent they are mentioned to analytical aspects of data integration to support analysts, decision makers and stakeholders who benefit from analyzing financial documents. Thus, these research questions will be answered through the paper:

RQ1: How existing data integration approaches will be helpful in context of XBRL financial analysis?
RQ2: To what extent these approaches support analytical aspect of data integration?

The course of the paper is as follows: after illustrating motivation and relevance of the research through use cases in the following section, the research method will be introduced and enhanced by a literature search and analysis. Afterwards, the identified articles will be explained and discussed to determine the specific field. Finally, further work possibilities will be discussed and summarized in conclusion section.

**MOTIVATION**

We analyzed existing literature to identify approaches, which deal with analytics within the domain of financial analysis. The approaches use structured data of balance sheets and also unstructured data, which are located in footnotes and are relevant for calculating financial ratios, analytical tasks, forecasts, or support decision making processes of financial issues in companies.

Franceschetti, Koschtial, and Felden (2012) proposed a method to regain trust and to recognize insolvency in business transactions. This method is called “Break-up Analysis (BUA)” . This procedure is a method to find out, whether the company is able to pay all liabilities or debt at any time, which result in a better recognition of insolvency or bankrupt of those companies. To deal with this challenge, analysts should use data provided by balance sheets, because they contain information about the liquidation values of companies. By applying BUA, the asset side of the balance sheet is reviewed and the liability side will be integrated by memo accounts. The important part is that, some assets should be removed in the balance sheet. They are called cut-off elements. Some of these cut-off elements are identified in footnotes, for example, financial fixed assets and current assets receivable. Therefore, the analyst should be manually read footnotes and extract them from footnotes.

Gujarathi (2008) clarifies, whether the financial ratios are consistent with different company’s policies and business environments and therefore in risk analysis and profitability. It is a case study in international financial statement, which compares two companies’ financial statements: Sachiko, a Japanese company, and Radiance, a U.S. corporation. According this use case, income statements, balance sheets, and footnotes for both companies are reviewed through the latest accounting period. One of the extracted items from footnotes is Financial Leases, based on this comparison; it is to recognize that most of the leases in the Japanese company are treated as operating leases (despite the U.S. Company, which leases are mostly shown as capital leases). To find out the possible effects of this difference on financial statement analysis, the case study calculated some ratios like interest expenses. For calculating this ratio, we should know about the interest rate (e.g. the interest rate on the long term secured debt of companies). This information can be manually obtained from footnotes.

Another use case, done by (Duke et al., 2012), illustrates the effects of the new lease standard by the FASB (Financial Accounting Standard Board) on existing operating lease by applying this new standard on two firms (FedEx and UPS), which rely on operating leases. This introduced standard proposes that all firms report their existing operating leases as capital leases. According to this case study and to recognize the impact of this standard, analysts should calculate some key financial ratios such as retained earnings, debt-to-equity, return-on-assets, and interest coverage ratios. They need to be identified in the footnotes for calculating ratios.

Bennett and Bradbury (2003) present the financial statement impact of constructive capitalization for 38 firms listed on the New Zealand Stock Exchange and propose two methods for their capitalization. They demonstrate that for discovering the impact of this capitalization, data like estimating the lease liabilities and estimating the income effect are needed. They collected all the information about operating leases of all 38 firms, such as properties, plant and equipment. New Zealand firms are required to disclose future operating lease rental in four continues periods. They should check footnote disclosures to estimate the present value of lease obligations.

Based on reviewed ratios and the derived cases from literature, it is obvious that information contained in footnotes and information contained in structured elements need to be integrated among the analysis to gain a complete understanding of the financial state of a company. According to these use cases, a manual extraction of important information from unstructured footnotes is costly, time-consuming, and error prone. It is also demonstrated that understanding and reading the unstructured part of financial statements, which contain important information for analysts, is a non-separable part of any financial analytical tasks. There is no automatic or computer-based framework except reading footnotes and due to the lack of research in the literature, which is more focused on an automated integration of the unstructured part of financial documents.
with structured values, and also based on a technical integration by XBRL (providing both data types, structured and unstructured), there is a strong need to define and search for an appropriate solution in favor of a meaningful automated integration of footnotes and structured values to support the analytical financial processes.

**RESEARCH METHOD AND RELATED WORK**

Due to this fact, that XBRL documents are based on both structured and unstructured data types, which are both necessary to comprehensive analysis and considering to XBRL technology as a technical attempt to make financial statements in a unified and machine readable platform as well, the state of the art is needed to recognize previous attempts in the area of unstructured data integration and evaluating to what extent these methods are applicable in financial statements based on XBRL to extract information from textual parts and integrate them with structured values to facilitate analytical tasks.

To this, a state of the art is performed to identify existing solutions, methods, and tools for integrating structured and unstructured data without considering a specific domain.

The state of the art is done in five phases (Cooper, 1998):

1. **Problem formulation**: This defines the academic goal and research relevance.
2. **Literature search and data collection**: Appropriate researches which are related to the problem formulation have been studied in the second phase.
3. **Literature evaluation**: In this phase, the acquired literature will be proofed of relevance and categorized.
4. **Interpretation**: The results are analyzed and interpreted.
5. **Presentation**: Finally in the last phase, the results are presented in a suitable fashion.

**Literature Search**

The used scientific data bases are EBSCO, AIS Digital Library, and Science direct (SD). The search domains are scientific articles, which mostly demonstrate methods and tools in terms of *unstructured data integration*. The used search terms are: structured and unstructured data integration and financial analysis. The analyzed period is 1994 until 2012. After removing doublets, 16 articles are remaining, which each one addresses an approach to data integration issues. According to the literature, it seems to be obvious that the field of data integration is not a new area in information science. Research and tools already exist in this context during the last decade. The selected publications are listed in Table 1.

<table>
<thead>
<tr>
<th>No.</th>
<th>Reference</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Seng and Kong, 2009)</td>
<td>The paper presents two main categories in data integration domain.</td>
</tr>
<tr>
<td>2</td>
<td>(Inmon, 2006)</td>
<td>Five possibilities for matching structured and unstructured data.</td>
</tr>
<tr>
<td>3</td>
<td>(Soibelman et al., 2008)</td>
<td>They propose a framework (TIIM) to analyze unstructured data in construction projects domain.</td>
</tr>
<tr>
<td>4</td>
<td>(Kloptchenko et al., 2002)</td>
<td>Two separate methods for analyzing structured and unstructured data in financial reports based on clustering methodology.</td>
</tr>
<tr>
<td>5</td>
<td>(Yang et al., 2008)</td>
<td>A comprehensive study about text mining tools with their strengths and limitations.</td>
</tr>
<tr>
<td>6</td>
<td>(Sukumaran and Sureka, 2007)</td>
<td>An infrastructure for data integration (CDW) based on text tagging and annotation techniques.</td>
</tr>
<tr>
<td>7</td>
<td>(Chang et al., 2009)</td>
<td>The article implies a content analysis framework within the customer relationship management domain.</td>
</tr>
<tr>
<td>8</td>
<td>(Curry et al., 2009)</td>
<td>The paper supposes an approach for different data types integration.</td>
</tr>
</tbody>
</table>
In order to have an organized evaluation of reviewed papers, it has to be considered that there are different categorizations to analyze these articles. Here, we explain two main different approaches in term of data integration in these papers, which will be the basis for further investigations. Then, we discuss these approaches through each paper. Afterwards, we argue to what extend this categorization are helpful and how they can be used in term of an XBRL financial analysis.

Considering the entire area of data integration, it is realized that there are two main approaches. On the one hand side, there are technical-driven approaches, which are concentrated on technical solutions and methods to integrate different data sources into integrated storage. On the other hand side, there are business-driven approaches, which are focused on an integrated data usage in order to support and improve business activities in context of analytical purposes. Figure 1 demonstrates the two dimensional framework. It should be implied that most of them integrate single data type sources (numbers within the brackets refer to the number of references in Table 1).
Technical-driven approach

All these articles which are placed under this category, discuss solutions and methods, which enable technically integrate different data sources. Their main focus is to integrate different data sources into a single storage option.

Seng and Kong (2009) imply that two main solutions within the data integration domain are available. The first one is Distributed Information Integration, which is based on so called object models. The second one is so called XML-based information integration. The term of object model refers to models, which are based on classes and their relationships. The purpose of an object model is to facilitate data management within the application. It is discussed that with the advent of XML, an increase of switching from object models to XML models for data integration is considerable. XML (Extensible Mark-up Language) is used as a standard to provide a common format for presenting data structures and contents. This supports the integration of different types of internet based data (Bertino and Ferrari, 2001). According to this categorization, there is a comparison between different information integration methods. Some of these methods are examples of Distribution Information Integration category and others are examples of an XML-based integration (Table 2).

<table>
<thead>
<tr>
<th>Distributed Information Integration (based on object model)</th>
<th>TSIMMIS</th>
<th>The Stanford-IBM Manager of Multiple Information Sources</th>
<th>The goal of this system is to facilitate the integration of heterogeneous information based on mediator/wrapper structure through an Object-Exchange Model (OEM).</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM</td>
<td>Information Manifold</td>
<td>It uses a knowledge representation technology to integrate information from different sources.</td>
<td></td>
</tr>
<tr>
<td>DISCO</td>
<td>Distributed Information Search Component</td>
<td>It provides scaling access to heterogeneous data sources and based on mediator/wrapper structure.</td>
<td></td>
</tr>
<tr>
<td>Garlic &amp; Clio</td>
<td>It is developed at IBM Almaden Research center</td>
<td>It integrates data from different relational and non-relational sources through mediator/wrapper structure with an object oriented language.</td>
<td></td>
</tr>
<tr>
<td>YAT</td>
<td>Yet Another Tree-based</td>
<td>It is based on mediator/wrapper architecture to facilitate data conversion. It is a semi-structured model and uses named trees with labeled nodes.</td>
<td></td>
</tr>
<tr>
<td>MIX</td>
<td>Mediation of Information using XML</td>
<td>It deploys XML as a data model for integration purposes. It uses XMAS (XML Matching and Structuring) language for user queries.</td>
<td></td>
</tr>
<tr>
<td>Agora</td>
<td>It is developed in the Caravel Project at INRIA Rocquencourt</td>
<td>It processes and integrates XML data sources based on the relational technology and translate users XQuery into SQL query on a generic relational schema.</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison between two different integration categories

Inmon’s approach (2006) demonstrates five possibilities for matching structured and unstructured data: no matching, coincidental textual matches, hard wired matches, probabilistic matching, and metadata matching. It is argued that using each method depends on different aspects and also depends on which of them is an appropriate solution for a particular organization or purpose.

A paper by Sukumaran and Sureka (2007) introduces an infrastructure that integrates both structured and unstructured data, which is called Complete Data Warehouse (CDW). They use text tagging and annotation techniques, which are called Named Entity (NE), as a fundamental step for unstructured data integration. The main idea behind this concept is bringing unstructured data into structured format and then using BI applications for analyzing this integrated data.

Curry et al. (2009) supposed an approach for data integration purposes, which is applicable to different data types. They argue that the architecture of data integration system should be consisting of two components: data preparation and
integration phase to convert different kinds of data sources into a unified format, and a query and user interface module to apply to integrated data using SPARQL.

Baars and Kemper (2008) discussed three approaches in the field of CRM to the integration of structured and unstructured data and put them into the three layer BI framework (data layer, logic layer, access layer) and compared them according defined criteria. These approaches are: integrated presentation, analysis of content collection, and distribution of analysis results and analysis templates. This approach in access layer, where analytical tasks have been done, can also be categorized as business-driven approach.

The Stanford-IBM Manager of Multiple Information Sources (TSIMMIS) is another project (Chawathe et al., 1994) for integrating heterogeneous information sources. It uses wrapper-mediator structure to facilitate integration process, which includes both structured and unstructured data. One limitation is a requirement for a more human participation.

Langegger (2008) proposed a system called Semantic Web Integrator and Query Engine (SemWIQ), which is supposed for meaningful data sharing in scientific collaboration. This system is based on the mediator-wrapper architecture and SPARQL query processor.

Another framework in this category is presented by Dominovic (2009), which demonstrates a data integration framework called ARCH by means of an ontology, which provides mapping process more efficient. The limitation of this framework is the only integration of heterogeneous structured data in different databases without considering unstructured data.

Jayaprabha and Saradha (2010) propose an approach to convert unstructured tourism data into semantic web format by means of the Resource Description Framework (RDF), which is based on ontology languages. This framework has three levels for tourism unstructured data integration as follows:

1. downloading unstructured text using on line search (html)
2. converting html into XML using different convertors like SAX parser
3. mapping process with the help of ontology (travel2.owl). In this level XML tags are mapped and converted into RDF to be available for semantic search.

**Business-driven approach**

The term of data integration does not affect technical issues, but deals also with business aspects. Today’s business processes depend on the use of technology. Business-driven approaches are supposed to deal with such solutions to facilitate a usage of integrated data. The papers of this category deal with single data types. There is no identified paper, which covers both data types within a single technology like XBRL, to improve business activities in analytical tasks within financial analysis domain.

According to Soibelman et al. (2008), data grow in terms of volume and complexity within the construction projects domain, results in efforts to manage and analyze unstructured data precisely. Due to this aim, they proposed the framework Text Information Integration Methodology (TIIM) to analyze unstructured data. It has five main components: project document preparation, project model preparation, classification, retrieval and ranking, and association. The methodology is verified through developing a prototype called Unstructured Data Integration System (UDIS) based on 25 construction data bases and 30,000 electronic documents.

There is also a research in the financial domain as well, which is supposed to analyze both kinds of structured and unstructured data in financial reports. They are based on clustering methods (Kloptchenko et al., 2002). It applies a Self Organizing Map (SOM) to analyze structured data and uses a prototype matching text clustering methodology as a text mining tool for unstructured information within a financial report. To interpret future financial performance, the results of both analyses are combined.

A study by Yang et al. (2008) deals with a comprehensive overview of some text mining and visualization tools. They try to compare different text mining tools based on six criteria which are: type of tools, capabilities (method), data sources, result output, and potential users. Accordingly, it has discussed perceived strengths and potential limitations of each tool to increase the understanding of existing text mining tools. According to these tools, it is supposed that unstructured data should be separately analyzed through different text mining tools without integrating with structured data.

Chang et al. (2009), proposed the Content Analysis Framework to analyze text customer data in the domain of customer relationship management (CRM) by applying four steps:

1. data collection from different types of customer’s data.
2. content analysis by category building, coding and reliability and validity tests.
3. data movement and data transformation.
4. building business value with the help of decision tree.

To verify the effectiveness of this approach, a case study proposed in this research.
It is recognized that a great range of data integration approaches in recent years are based on ontology-aided structure. Maier (2003) argues that “Ontologies are the best representation models to support and fulfill the integration process within organizations”.

In the following, there are research papers which all of them use ontology for data integration purposes. Li and Brook (2006) applied a framework to integrate both text mining (document management systems or content management systems) and data mining techniques in a unified platform for decision makers to gain business intelligence. The core part of this framework is using a taxonomy, which semantically links these two kinds of data. According to Felden (2006), “unstructured data have to be identified, classified, and evaluate to clarify their grade of interest and usefulness”. He introduces Systematic Analysis and Research Tool (SMART) in the German energy market area. It uses ontology-based user profiling to be understandable by both users and information systems. There is also a triggering system for unstructured data, which works for each profile. Cheng et al. (2002), develop a system to analyze sentences in unstructured documents through ontology for semantic classification as well. To verify this system, a prototype is implemented into three distinct layers (application, physical storage, core engine) so called Ontology-based Semantic Classification (OSC), which there is an interface via XML between these components.

Addressing the first research question, technical-driven data integration approaches are not appropriate for our research goal, because integrating structured and unstructured data already have been done technically with XBRL. Furthermore, business-driven approaches, which are also reviewed, focus on single data type and are not applicable for XBRL financial documents which tackle with both kinds of data. Considering the field of financial analysis, there is a study by Piechocki (2007) which visualizes the financial reporting process (Figure 2). There exist different steps before using information for financial reports: information need, information preparation and information transmission, and unifying the formats. All of these steps have been done using XBRL. The final step is based on using financial information for specific purposes, e.g. the interests of analysts.

The financial reporting process can be technically supported in term of data integration. But as it is discussed in use cases, the problem resides in analysis level (final step), where analysts should extract relevant information from footnotes by reading textual part manually to use them for calculating financial ratios.

![Figure 2. General reporting process](image)

Regarding the reviewed literature, no paper could be identified, addressing an appropriate method to deal with extracting footnote automatically and integrate it with structured values to do financial analytical tasks in XBRL financial reports. Considering the financial analysis domain and considering XBRL, both structured and unstructured parts of financial reports are delivered already in a unified platform. As a result, there is a strong need to focus on developing such models or conceptual frameworks, which are business-driven and facilitate the usage of integrated delivery of structured and unstructured data so that financial analysis can be done in an automatic fashion. Summarizing to second research question it can be stated that the reviewed approaches support analytical aspects of data integration in a limited manner.

After accomplishing the literature review, further investigations are necessary to analyze and identify requirements of data integration for financial analysis based on XBRL documents. Possible solution should consider an enhancement of an analytical usage of an automatic data integration to reduce human interfere and therefore improve the financial analysis as a conceptual framework. A next step is to assess research results and consolidate and visualize outputs and arguments of further research potentials within this area.

**CONCLUSION**

The introduction of XBRL enables a unified platform for processing financial statements and to increase the speed of financial analysis and reduce cost and time consuming manual processing of financial data (www.xbrl.org).
This paper clarifies obstacles; financial analysts are faced in term of using unstructured data of XBRL financial documents, because they have to integrate them with the structured values to be able to calculate financial ratios.

Considering the fact that data integration has different aspects and because it happens in different steps of the information value chain, two approaches (technical- and business-driven approach) are introduced to cluster relevant papers. The results show that most current data integration methods concentrate on a technical-oriented integrated data storage level. The business-driven approaches focus on an integrated usage of data, but does not integrate structured and unstructured data in a unified platform. In this context, there is no identified paper placed in business-driven approach, which deals with both data types in a single platform.

Furthermore, Piechocki’s financial reporting process implies that the financial reporting processes can be improved by using XBRL. But this leads to an approach to facilitate an automatic content extraction from footnotes and integration with structured values to support data analysts.

In this context, it seems to be necessary to concentrate on business-driven approaches and on data usage instead of data storage.

This paper is a research in progress and does not deliverer a particular solution. Instead, it clarifies and identifies the necessity of integrated usage of data in financial documents. Further research is necessary to develop a conceptual framework for integrating structured and unstructured data to support the financial analysis domain.

REFERENCES

5. Chang, Che-Wei; Lin, Chin-Tsai, Wang, Lian-Qing, Mining the text information to optimizing the customer relationship management, Journal of Expert systems with applications, No.36, pages 1433-1443
10. Curry, Edward; Harth, Andreas; O’Riain, Sean (2009) “Challenges ahead for converging financial data”, W3C Workshop on improving access to financial data on the web
11. Dominovic, Christiane, ARCH: incorporating usability into a data integration framework, iiWAS 2009 Proceedings
22. Kloptchenko, Antonina; Eklund, Thomas; Bach, Barbro; Karlsson, Jonas; Vanharanta, Hannu; and Visa, Ari, “Combining data and text mining techniques for analyzing financial reports” (2002), AMCIS 2002 Proceedings paper 4
23. Langegger, Andreas, Virtual data integration on the Web: Novel methods for accessing Heterogeneous and distributed data with rich semantics, iiWAS 2008 Proceedings
24. Li, Quanzhi and Brook, Yi-fang, Information mining: Integration data mining and text mining for business intelligence, AMCIS 2006 Proceedings
25. Maier, Andreas; Aguado, J.; Bernaras, A.; Laresgoiti, I.; Pedrinaci, C.; Pe. and Smithers, T., Integration with Ontologies, 2nd Conference Professional knowledge management (WM 2003), April 2003
32. Sukumaran, Sreekumar; Sureka, Ashish (2007) “Integration structured and unstructured data using text tagging and annotation”, The data administration newsletter by TDAN.com
33. Weglarz, Geoffrey, Two worlds of data: unstructured and structured, DM Review, September 2004