Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2009 Proceedings

International Conference on Information Systems (ICIS)

2009

Towards Quality of Data Standards: Empirical Findings from XBRL

Hongwei Zhu
Old Dominion University, hzhu@odu.edu

Liuliu Fu
Old Dominion University, lxxfu001@odu.edu

TOWARDS QUALITY OF DATA STANDARDS: EMPIRICAL FINDINGS FROM XBRL

Research-in-Progress

Hongwei Zhu
Old Dominion University
Norfolk, VA 23529, USA
hzhu@odu.edu

Liuliu Fu
Old Dominion University
Norfolk, VA 23529, USA
lxxfu001@odu.edu

Abstract

Certain data standards can help improve the quality of the data created according to the standards. But data standards do not always improve data quality. We introduce the notion of "quality of data standards" and argue that quality of data is affected by the quality of the standards used. We develop metrics for assessing quality of data standards. The metrics are evaluated empirically using company financial reports created using the eXtensible Business Reporting Language (XBRL) data standards. Our findings show the use frequency of standard data elements roughly follows a power law distribution. Tradeoffs exist between relevancy and completeness dimensions and between a single user perspective and user community perspective.

Keywords: data quality, interoperability, data standards, XBRL, power law distribution, long tail

Introduction

In Desert Storm, an aerial observer located an enemy unit and sent a bombing request to the artillery headquarters. Using the enemy location's coordinate received from the artillery headquarters, the Navy ship off the coast fired two rounds, but both missed the target by 527 meters (Herrera 2003), a distance way greater than the expected precision!

What went wrong? It turned out that the artillery headquarters and the Navy used different geo-coordinate systems with which the same coordinate represents different locations on earth. This is a form of data quality problem caused by lack of interoperability between organizations and between systems.

Aside from battle grounds, data sharing among organizational units and between organizations is also required in ordinary day-to-day operations. Data from disparate sources must have high quality and be unambiguously interpretable for the users to make good decisions. Data standardization has the potential of ensuring quality and enhancing interoperability of data from disparate sources.

The Department of Defense (DoD) later found a simple solution to the data quality problem mentioned earlier. In a project dubbed "Cursor on Target", the DoD standardized target data exchanged among different branches of the armed forces. The standard is rather simple and consists of only three entities and 13 attributes (Rosenthal et al. 2004). Regardless of the coordinate system used internally, the target data exchanged using the standard can be correctly interpreted by any branch that receives it. Another successful DoD project developed a standard for exchanging meteorological (i.e., weather-related) data among different systems. The standard consists of approximately 1,000 elements and took five years to develop (Rosenthal et al. 2004). This standard worked very well partially because the concepts represented by the data elements are commonly understood with precise scientific definitions

Do data standards always improve data quality and interoperability of disparate sources, especially when the standards are large with thousands of data elements? This is an under-researched area and there is evidence showing that the impact of standards on data quality can go either way, depending on the "quality of standards".

In this paper, we introduce the concept of "quality of standards", develop a framework for assessing the quality of standards, and empirically evaluate the framework using data standards and corresponding data instances. The empirical evaluation uses a collection of eXtensible Business Reporting Language (XBRL) (XBRL International 2006) taxonomies (the data standards) and XBRL financial statements (the data sources) filed by approximately 140 public companies to the Securities and Exchange Commission (SEC). Our analysis show that many companies only use a small fraction of the data elements in the standard taxonomies and have introduced a large number of additional data elements not defined in the standard taxonomies. As a result, the financial statements from different companies cannot be easily compared. The lack of interoperability among financial statements of different companies is largely due to the low quality of the standard taxonomies measured using our metrics for quality of data standards.

Quality of Data Standards

Most data quality research focuses on data, not the standards used to create and organize the data. Data quality is a multi-dimensional concept that goes beyond accuracy. Prior research has identified 16 dimensions (e.g., consistency, interpretability, completeness, relevancy, etc.) of data quality (Wang and Strong 1996). Data quality perceived by users of different roles within an organization can be assessed using survey instruments (Lee et al. 2002). Quality of database schemas is discussed in (Redman 1996). Although a database schema is a type of data standard, it is mainly used within a single organization to organize and store data in a database. In contrast, the main objective of many data standards is to allow for meaningful exchange of data among multiple organizations so that the data from different organizations are interoperable.

Data standards are meta-data that specifies the characteristics of data elements and their relationships. From the perspective of standard users, meta-data is also data. Therefore certain aspects of data quality are also applicable to data standards. Data standards also have their distinct characteristics that require different quality metrics or different measurement methods for the same metrics. Since a data standard is designed for use by multiple organizations, and quality is defined as fitness for use, it is necessary to observe and analyze how different organizations use the standard when assessing the quality of the standard.

In this study, we will focus on two quality dimensions that have significant impact on the interoperability of data created by multiple organizations that use a common standard: *completeness* and *relevancy*. When a standard has low completeness, users must introduce new elements, in which case the data instances are not interoperable. When a standard has low relevancy, users incur unnecessary cognitive cost and have increased propensity of misusing data elements, in which case the data instances contain Error.

In (Wang and Strong 1996), completeness is defined as "the extent to which data are of sufficient breadth, depth, and scope for the task at hand", and relevancy is defined as "the extent to which data are applicable and helpful for the task at hand". Schema completeness and pertinence (i.e., relevancy) are defined similarly in (Redman 1996). These definitions are useful at conceptual level, but they do not offer metrics for measuring quality along the defined dimensions. Below, we develop metrics that can be used to assess completeness and relevancy of a data standard. From standard user's perspective, the "task at hand" is to use the standard to create data instances that can interoperate with data created by other users. Thus we need to examine the data instances of various standard users to assess standard quality. Figure 1 illustrates this point.

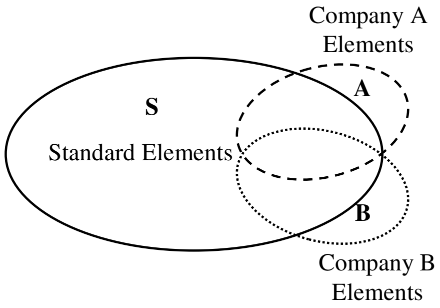


Figure 1. Quality of Data Standard

A data standard specifies a set of elements, **S**, represented by the solid oval in Figure 1. Suppose there are only two organizations that use the standard: company A and company B. Each company extends the standard by adding its own elements, which can be identified by examining the company specific standard. To identify the elements actually used by a company, we have to examine the data instances created by the company. The sets of elements used by companies A and B are represented in the figure as **A** (the dashed oval) and **B** (the dotted oval), respectively. The completeness and relevancy of the standard from the perspective of a given company may be defined straightforwardly. For example, from company A's perspective,

$$Completenss_A = \frac{|A \cap S|}{|A|}, \quad Relevancy_A = \frac{|A \cap S|}{|S|}.$$

From the standard developer's perspective, we need to evaluate different definitions of the metrics. For example, it may be tempting to define them as

$$Completenss_S = \frac{|(A \cup B) \cap S|}{|A \cup B|}, \quad Relevancy_S = \frac{|(A \cup B) \cap S|}{|S|}.$$

But the definition does not take into account the intersection of A and B. It may be more appropriate to take the harmonic mean or the geometric mean of the metrics measured for all companies.

In order to obtain the measurements of the metrics, we must identify the standard elements used by each company. Thus these metrics provide more information about how a standard is used than other studies that only report the average of the number of standard elements used by all companies. For example, an investigation on how U.S. companies used XBRL in their SEC filings (Boritz and No 2008b) finds that an average company used 162 standard elements and 190 extension elements. From the averages, we cannot infer the number of standard elements used by all companies because we do not know the degree of overlapping of the standard elements used by different companies. It will be incorrect to say that 8.1% (which is 162/2,000) of the 2,000 elements were used by the companies - this would be correct only if all companies used the same 162 elements.

By examining both the standards and the instances, we can also identify ways of improving standard design and areas where advanced technologies are needed to enable interoperability of data sources in the presence of "imperfect" standards. For example, a standard with low completeness will cause the standard users to introduce a large number of elements not defined by the standard. The interoperability of the data created by the users will be low in this case. We need to *match* up these elements amongst different users to enable interoperability. Technologies are desired to automate the matching process and reconcile other subtle semantic differences (Zhu and Madnick 2006).

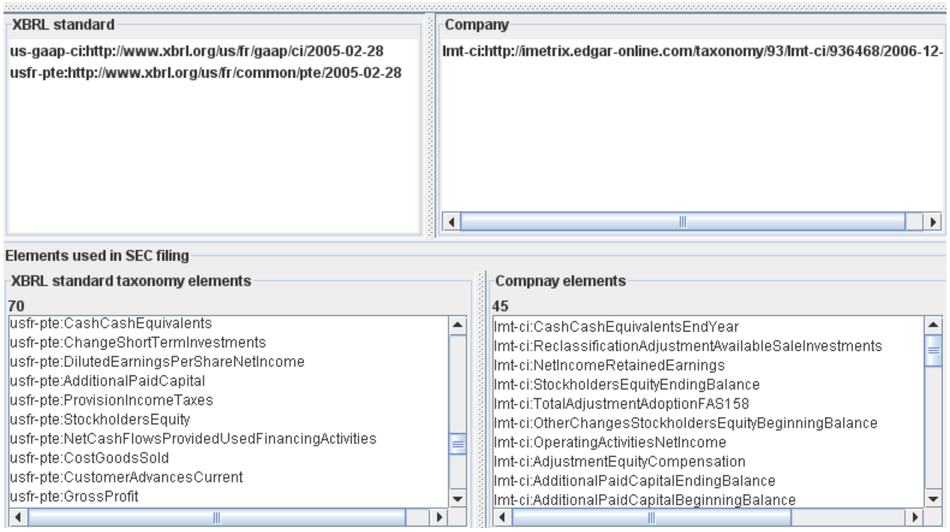
Evaluation Method

In the financial accounting community, there has been an effort since the late 1990s to develop a set of data standards so that financial statements from different companies can be easily compared (i.e., interoperable). The data standards, known as taxonomies, are specified using an eXtensible Markup Language (XML)-based language called the eXtensible Business Reporting Language (XBRL) (XBRL International 2006). There are country-specific taxonomies to accommodate differences in accounting methods and reporting requirements. In the U.S., there are a taxonomy for commonly used financial reporting terms (approximately 2,000 elements) and a set of industry-specific taxonomies. Collectively, the taxonomies form a standard for companies publicly traded in the U.S. to create electronic financial statements (which are called XBRL instances). Since the standard may not define all possible elements, companies are allowed to extend the standards with their own elements. Unlike meteorology, the financial accounting domain has concepts that lack precise and uniform definitions.

With the interest of adopting XBRL, the Securities and Exchange Commission (SEC) of the U.S. established a voluntary XBRL filing program in 2005 to allow companies to submit their financial statements in XBRL. As of April 30, 2009, 140 companies had submitted XBRL taxonomy extensions and XBRL instances, and many of the companies have submitted multiple filings for different accounting periods.

By analyzing the XBRL instances submitted to the SEC, we can identify the data elements used by the companies and evaluate the proposed metrics for measuring the quality of data standards. There are two additional benefits of using actual company data. First, it allows us to investigate if the XBRL standards have helped with the creation of high quality and interoperable financial statements. Preliminary studies show that there are still interoperability challenges (Debreceny et al. 2005; Zhu and Madnick 2007) and more than a half of the XBRL instances filed to the SEC have errors (Boritz and No 2008b; Chou 2006). Second, we can discover how data standards are used, and from the usage pattern we can further develop methods and guidelines for designing high quality data standards.

To assist with data analysis, we have developed an XBRL processing tool to extract data elements from XBRL instances and separate standard elements from those defined by the company. Figure 2 shows the partial user interface after the elements of an XBRL instances have been extracted and separated. In this particular example, the company used 70 elements defined in XBRL data standard and 45 elements defined by the company.



us-gaap-ci:http://www.xbrl.org/us/fr/gaap/ci/2005-02-28
usfr-pte:http://www.xbrl.org/us/fr/common/pte/2005-02-28

lmt-ci:http://imetrix.edgar-online.com/taxonomy/93/lmt-ci/936468/2006-12-

70

usfr-pte:CashCashEquivalents
usfr-pte:ChangeShortTermInvestments
usfr-pte:DilutedEarningsPerShareNetIncome
usfr-pte:AdditionalPaidCapital
usfr-pte:ProvisionIncomeTaxes
usfr-pte:StockholdersEquity
usfr-pte:NetCashFlowsProvidedUsedFinancingActivities
usfr-pte:CostGoodsSold
usfr-pte:CustomerAdvancesCurrent
usfr-pte:GrossProfit

45

lmt-ci:CashCashEquivalentsEndYear
lmt-ci:ReclassificationAdjustmentAvailableSaleInvestments
lmt-ci:NetIncomeRetainedEarnings
lmt-ci:StockholdersEquityEndingBalance
lmt-ci:TotalAdjustmentAdoptionFAS158
lmt-ci:OtherChangesStockholdersEquityBeginningBalance
lmt-ci:OperatingActivitiesNetIncome
lmt-ci:AdjustmentEquityCompensation
lmt-ci:AdditionalPaidCapitalEndingBalance
lmt-ci:AdditionalPaidCapitalBeginningBalance

Figure 2. Screenshot of XBRL Process Tool

Evaluation Results

The results are based on the analyses of the XBRL instances of 140 companies collected from the SEC's voluntary filing program. Most companies have more than one filing. We treat all filings of each company as opposed each filing as a data point. For a company that has more than one filing, the set of data elements used by the company is the union of the elements used in all filings.

Before we present the details, we first highlight the main findings: there were large variations across accompanies and the use frequency of standard elements roughly followed a power law distribution.

Figure 3 shows the distribution of the number of filings per company (i.e., number of XBRL instances per company). Most companies only filed once or twice, but there are two companies that filed more than 16 times.

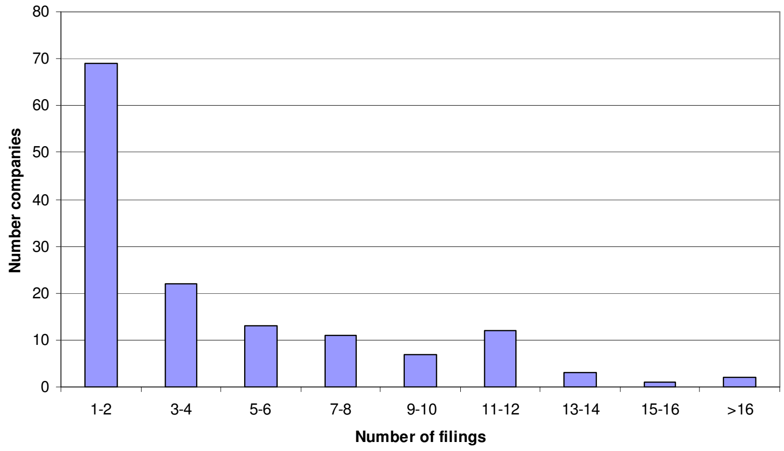


Figure 3. Histogram of Number of SEC Filings per Company

The use of standard elements and company-specific elements also varied substantially across companies. Table 1 provides the summary statistics. Both extreme cases exist: company that only used elements from the standard and company that did not use standard elements at all! An average company used 128 elements from the standard (which defines ~2,000 elements) and introduced 64 elements of its own. To the average company, the completeness of the standard was 66.67% (which is 128/(128+64)), and the relevancy of the standard was 6.40% (which is 128/2000).

Table 1. Statistics of Number of Standard Elements and Company Elements

	Standard	Company
Min	0	0
Max	528	593
Mean	128.28	64.26
Median	110	38
Stand deviation	78.32	85.78

Although on average a company only used a small fraction of the standard elements, the 140 companies used 1385 standard elements. That is, using the metric suggested earlier, the relevancy of the standard from the standard developer's perspective was 69.25% (i.e., 1385/2000). On the other hand, the companies introduced 8996 elements of their own. Assuming these elements were unique (a big assumption), the completeness of the standard from the standard developer's perspective was only 13.34% (i.e., (1385/(1385+8996)).

The use frequency of the 1385 standard elements is shown in Figure 4. There are a few elements that were used by more than a third of the companies. But most the elements were only used by a couple companies. This is a typical power law distribution, also known as the long tail distribution.

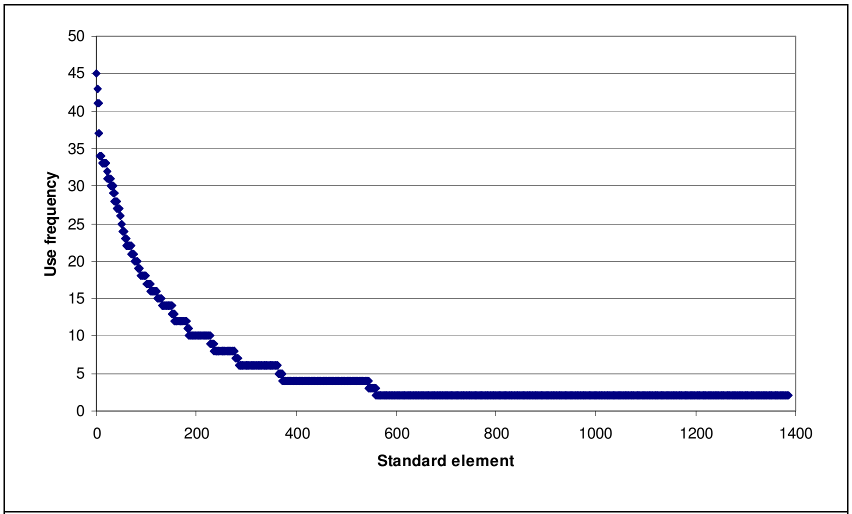


Figure 4. Use Frequency of Standard Elements

The top 50 most frequently used standard elements and their use frequencies are provided in Table 2 (e.g., *StockholdersEquity* is used by 45 companies). Elements are ordered according to descending order of use frequency.

Table 2. Top 50 Standard Data Elements and Use Frequency

StockholdersEquity	45	LiabilitiesCurrent	31
Assets	43	NetCashProvidedByUsedInFinancingActivities	31
CommonStockValue	41	RetainedEarningsAccumulatedDeficit	31
OtherAssetsNoncurrent	41	LiabilitiesStockholdersEquity	30
AccountsPayable	37	NetCashFlowsProvidedUsedInvestingActivities	30
Liabilities	37	NetIncome	30
CashCashEquivalents	34	ProvisionIncomeTaxes	30
NetCashFlowsProvidedUsedFinancingActivities	34	RetainedEarnings	30
NetCashFlowsProvidedUsedOperatingActivities	34	ComprehensiveIncomeNetOfTax	29
NetIncreaseDecreaseCashCashEquivalents	34	DepreciationAndAmortization	29
EarningsPerShareBasic	33	PaymentsToAcquirePropertyPlantAndEquipment	29
EarningsPerShareDiluted	33	AdditionalPaidCapital	28
Goodwill	33	CurrentLiabilities	28
IncomeTaxExpenseBenefit	33	DeferredIncomeTaxExpenseBenefit	28
InterestExpense	33	PropertyPlantEquipmentNet	28
LiabilitiesAndStockholdersEquity	33	TotalCurrentAssets	28
NetCashProvidedByUsedInInvestingActivities	33	AccountsReceivableNetCurrent	27
NetCashProvidedByUsedInOperatingActivities	33	CommonStockParOrStatedValuePerShare	27
NetIncomeLoss	33	CommonStockSharesAuthorized	27
PropertyPlantAndEquipmentNet	33	CommonStockSharesIssued	27
TreasuryStockValue	32	OtherLiabilitiesNoncurrent	27
AccumulatedOtherComprehensiveIncomeLossNetOfTax	31	IncomeLossContinuingOperationsBeforeIncomeTaxes	26
CashAndCashEquivalentsAtCarryingValue	31	IntangibleAssetsGoodwill	26
CashAndCashEquivalentsPeriodIncreaseDecrease	31	OtherNoncurrentLiabilities	26
IncomeLossFromContinuingOperationsBeforeIncomeTax	31	PreferredStockValue	25

When companies introduce additional elements, these elements may overlap (i.e., multiple companies introduce the same data elements not included in the standard). Identifying the same elements introduced by different companies is very difficult because different names may be used by different companies for the same concept. Similarly, the same name may be used to refer to different concepts by different companies. We are in the process of developing matching algorithms to identify overlapping elements introduced by different companies. For now, let us use simple string matching and assume elements with an identical name are the same data element. The top 20 overlapping elements under this assumption are provided in Table 3 in descending order of the number of companies introducing the element.

Table 3. Top 20 Company-Introduced Elements and Frequency	
CashCashEquivalentsBeginningYear	20
CashCashEquivalentsEndYear	20
OperatingActivitiesNetIncome	16
PurchasePropertyPlantEquipment	16
AccumulatedComprehensiveIncomeBeginningBalance	12
AccumulatedComprehensiveIncomeEndingBalance	12
AdjustmentDepreciationAmortization	12
AdjustmentEquityCompensation	12
CashDividendAmountPerShare	12
CommonStockValueTotalBeginningBalance	12
CommonStockValueTotalEndingBalance	12
StockholdersEquityEndingBalance	12
AdjustmentAssetImpairmentCharge	10
AdjustmentDeferredIncomeTaxes	10
ChangeAccountsReceivable	10
ChangePrepaidExpensesOtherCurrentAssets	10
NetIncomeRetainedEarnings	10
RetainedEarningsBeginningBalance	10
RetainedEarningsEndingBalance	10
StockholdersEquityBeginningBalance	10

We observe that certain company-introduced elements are more specific than similar elements in the standard. For example, the first two elements in Table 3, *CashCashEquivalentsBeginningYear* and *CashCashEquivalentsEndYear*, are more specific than *CashCashEquivalents*, the 7th element in Table 2, used by 34 companies. However, these two elements are introduced unnecessarily because XBRL has a *context* mechanism to include the as-of date of *CashCashEquivalents* to indicate whether the value is for the beginning of the year or the end of the year. The matching algorithms we are developing will help us identify such elements that should not have been introduced by standard users. Introduction of such elements reduces interoperability of financial statements of different companies.

A more specific element is usually constructed by adding additional qualifiers to a more general element. If company-introduced elements tend to be more specific than those in the standard, company-introduced elements are expected to be longer and to have more sub-terms than those in the standard. A sub-term is a word from a vocabulary (English in this case) that is a part of an element name. For example, *CashCashEquivalents* has three sub-terms: cash, cash, and equivalents. We have analyzed the set of standard elements and the set of elements introduced by all companies using measures discussed in (Good and Tennis 2009). Our results show that on average, a standard element has 43.05 characters and 5.75 sub-terms, whereas a company-introduced element has 53.95 characters and 7.25 sub-terms. Thus it is likely that companies introduced certain elements because they felt those in the standard were too general to use. Further investigation using the element matching algorithms will reveal more information about the company-introduced elements.

Discussion

Completeness and relevancy of a data standard affect the interoperability of data created by multiple organizations that use the standard. Our empirical analysis shows that from an individual company’s perspective, the XBRL data standard has low completeness and relevancy. As a result, the financial statements from different companies have low interoperability because companies have introduced a large number of data elements not defined in the standard. The empirical analysis also reveals an interesting pattern of the usage of large data standards: a typical user only uses a small fraction of the standard, but when all users as a whole are considered, most of the standard is utilized. This raises an important question: given the high cost of standard development, how should we design a high quality standard with minimal cost? How should we make the trade-offs between completeness and relevancy from the perspectives of an individual user and all users as a whole?

We are currently working on the “big assumption” mentioned earlier: we assumed that the elements introduced by companies are unique. This may not be true. As seen in the preceding discussion, many company-introduced elements have identical names, indicating they are very likely identical concepts. A company may introduce a redundant element because the company may be unaware of its existence in the standard. This tends to happen when the data standard contains a large number of elements, as is the case of XBRL taxonomies. For example, the XBRL standard has the element *AccountantName* for “Accounts Information – Name”, yet a company introduced its own element *AccountsInformationName* (Boritz and No 2008a), which unnecessarily introduced redundancy. We plan to combine and enhance several syntactic and semantic similarity algorithms (Rahm and Bernstein 2001) to semi-automatically identify potential duplicate data elements. These algorithms will exploit certain characteristics of XBRL (e.g., the equational relationships among elements) (Zhu and Madnick 2007).

We have only considered two standard quality dimensions in this study. Future research should examine other pertinent dimensions and develop metrics to measure standard quality along these dimensions. Future research should also investigate the effects of standard quality along the additional dimensions on the quality of data instances created using the standard.

References

Boritz, E.J., and No, W.G. "Auditing an XBRL Instance Document: The Case of United Technologies Corporation," University of Waterloo.

Boritz, E.J., and No, W.G. "SEC's XBRL Voluntary Program on Edgar: The Case for Quality Assurance " in: *SSRN: http://ssrn.com/abstract=1163254*, 2008b.

Chou, K.H. "How Valid Are They? An Examination of XBRL Voluntary Filing Dcoments with the SEC EDGAR System," in: *The 14th International XBRL Conference*, Philadelphia, USA, 2006.

Debreceny, R.S., Chandra, A., Cheh, J.J., Guithues-Amrhein, D., Hannon, N.J., Hutchison, P.D., Janvrin, D., Jones, R.A., Lamberton, B., Lymer, A., Mascha, M., Nehmer, R., Roohani, S., Srivastava, R.P., Trabelsi, S., Tribunella, T., Trites, G., and Vasarhelyi, M.A. "Financial Reporting in XBRL on the SEC's EDGAR System: A Critique and Evaluation," *Journal of Information Systems* (29:2) 2005, pp 191-210.

Good, B.M., and Tennis, J.T. "Term based comparison metrics for controlled and uncontrolled indexing languages," *Information Research,* (14:1) 2009, p #359.

Herrera, X. "The Bottom Line for Accurate Massed Fires: Common Grid," *Field Artillery*:January-February) 2003.

Lee, Y.W., Strong., D.M., Kahn, B.K., and Wang, R.Y. "AIMQ: a methodology for information quality assessment," *Information and Management* (30:2) 2002, pp 133-146.

Rahm, E., and Bernstein, P.A. "A Survey of Approaches to Automatic Schema Matching," *VLDB Journal* (10:4) 2001, pp 334-350.

Redman, T.C. *Data Quality for the Information Age* Artech House, Boston, 1996.

Rosenthal, A., Seligman, L., and Renner, S. "From Semantic Integration to Semantics Management: Case Studies and a Way Forward," *ACM SIGMOD Record* (33:4) 2004, pp 44-50.

Wang, R., and Strong, D. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4) 1996, pp 5-33.

XBRL International "Extensible Business Reporting Language (XBRL) 2.1," XBRL International.

Zhu, H., and Madnick, S.E. "Improving Data Quality with Effective Use of Data Semantics," *Data and Knowledge Engineering* (59:2) 2006, pp 460-475.

Zhu, H., and Madnick, S.E. "Semantic Integration Approach to Efficient Business Data Supply Chain”," The 6th Workshop on eBusiness (WEB’07), Montreal, Canada, 2007.