Interoperability of Data Created using Extensible Data Standards

Research-in-Progress

Hongwei Zhu and Harris Wu
Old Dominion University
{hzhu, hwu}@odu.edu

Abstract

As an important data quality dimension, semantic interoperability of data can potentially be improved using data standards. Certain data standards are extensible, allowing users to introduce custom data elements. How interoperable is data created using an extensible data standard? How does extensibility of a data standard affect interoperability of data created using the standard? How does a standard evolve and what is its impact on interoperability? This research address these questions by empirically measuring semantic interoperability of real world data created using a large extensible data standard. It also uses case studies to develop an understanding of how users extend data standards. Preliminary results of the impact of standard evolution on data's interoperability are also discussed.

Keywords. Data standards, semantic interoperability, data quality, standard evolution.

Interoperability of Data Created using Extensible Data Standards

Research-in-Progress

Introduction

Interoperability is an important data quality dimension (Wang and Strong 1996). Lack of interoperability creates significant challenges to information integration. Nearly 40% of IT budget in an organization is used to address these challenges (Bernstein and Haas 2008). Data standards are a form of metadata to reduce schematic and semantic heterogeneity of data from multiple sources. Data standards have the potential to solve interoperability problems as standards-based data are expected to be interoperable. Large-scale data standards, such as those within the US Department of Defense (Rosenthal et al. 2004) and across the real estate mortgage industry (Markus et al. 2006), include many data elements and are intended for use by a large number of organizations.

Since a data standard requires a certain level of consensus from a community of users, it is often a common denominator. Information not captured by the data standard is lost when data from multiple sources are integrated (Bernstein and Haas 2008). Thus, some data standards are extensible, allowing users to capture all information that they would like to retain and share. Apparently, extensibility is a double-edged sword as it can potentially reduce semantic interoperability of data created using an extensible data standard (Debreceny et al. 2005).

How interoperable is data created using an extensible data standard? How does extensibility of a data standard affect interoperability of data created using the standard? How does a standard evolve and what is its impact on interoperability? Driven by these questions, this research attempts to develop an understanding of the interplay of data standards, standards' extensibility, standard evolution, and interoperability of the resulting data. Specifically, it makes two contributions. First, using the metrics to measure the interoperability of standard-based data (Zhu and Wu 2011a; Zhu and Wu 2011b), we measure the interoperability of data instances created using a large-scale, real world, extensible data standard. Second, we conduct industry-based analysis and case studies of how users extend the standard to understand factors that affect interoperability and standard evolution.

The data standard is the US Generally Accepted Account Principles (GAAP) Taxonomy, encoded in eXtensible Business Reporting Language (XBRL) (XBRL International 2006). The data standard has been adopted by the Securities and Exchange Commission (SEC), which mandates all publically traded companies in the U.S. to use the Taxonomy to create their financial statements. Thus the data instances are financial statements submitted to the SEC. Besides the original GAAP Taxonomy released and adopted by SEC in 2009, a new 2011 version of the taxonomy was released this year and has been approved by SEC for use in financial statements.

Interoperability Metrics

A standard often adopts a uniform syntax for data representation. Thus it is relatively trivial to achieve syntactic interoperability. In addition, a standard also defines a set of data elements with their semantics agreed upon by all users, aiming to attain semantic interoperability. But in many cases when users are allowed to choose among different elements in the standard or to extend the standard with custom elements, semantic heterogeneity problems will arise.

In this paper, we focus on the *comparability* aspect of semantic interoperability. A set of data instances is interoperable if the instances use the same set of data elements defined in a data standard. Interoperability measures the extent to which the data instances have overlapping data elements defined in a standard. This definition allows us to measure interoperability directly without relying on unreliable semantic matching techniques (Rahm and Bernstein 2001; Rahm et al. 2004).

We borrow the *interoperability* metrics defined in (Zhu and Wu 2011a; Zhu and Wu 2011b). Interoperability between a pair of data instances is based on the common data elements used. The interoperability between users *i* and *j*, $\mathbf{I_{i,j}}$, can be defined as

$$I_{i,j} = \frac{|U_i \cap U_j|}{\sqrt{|U_i||U_j|}} \quad (1)$$

Clearly, $I_{i,j}=I_{j,i}$. The pair-wise interoperability for all users, $\mathbf{I_2}$, is defined as the arithmetic mean of pair-wise interoperability among all pairs. This definition can be extended to interoperability of any k-tuple (with k>=2) as

$$I_{i_1,\ldots,i_k} = \frac{|U_{i_1} \cap \ldots \cap U_{i_k}|}{\sqrt[k]{|U_{i_1}|\ldots|U_{i_k}|}} \quad (2)$$

The k-interoperability of all users, I_k, can be defined as the arithmetic mean of the k-interoperability among all *k*-tuples. We will limit our discussion to I_2 and I_3 because interoperability calculation is computationally expensive. For I_k, there are $O(n^k)$ k-tuples that need to be computed.

When a user is allowed to extend the standard, U_i can be partitioned into two sets: U_i^s (elements from the standard) and U_i^c (elements custom-made by the user). One may argue that custom elements tend to be specific to the user and people are largely interested in comparing data defined in the standard. Thus it is interesting to measure interoperability by just considering standard data elements. For this purpose, we define $I_{i,j}'$ and $I_{i,j,k}'$ by replacing all occurrences of U with U^s in equations (1) and (2).

XBRL, GAAP Taxonomy, and Data Collection

In this section, we provide background information about XBRL, GAAP Taxonomy, and methods for data collection and analysis.

XBRL and GAAP Taxonomy

XBRL is a *technology* based on XML Schema and XML Linking. It defines a business reporting language by specifying a set of data types, XML elements, and attributes for each element. For example, XBRL defines data types such as *monetaryItemType* and *sharesItemType* that are often used in business reporting. Using XBRL, any jurisdiction can develop its own reporting taxonomy as a *data standard* for companies to exchange business data.

An XBRL taxonomy consists of *taxonomy schemas* that define data elements and *linkbases* that specify various relationships among data elements and other resources. For example, below is the of the definition of the *Assets* data element in the schema of the GAAP Taxonomy:

```
<xs:element id='us-gaap_Assets' name='Assets' nillable='true' substitutionGroup='xbrli:item'
type='xbrli:monetaryItemType' xbrli:balance='debit' xbrli:periodType='instant' />
```

Below is how a company uses the Assets element to report its total assets in its financial statement:

```
<us-gaap:Assets id="Item-0039" contextRef="As_Of_12_31_2010" unitRef="Unit12" decimals= "-6">
35067000000</us-gaap:Assets>
```

When multiple companies use the *Assets* element in the GAAP Taxonomy to report their total assets, this piece of data is semantically interoperable among these companies. Conversely, if companies introduce their own custom elements, data reported using custom elements are not semantically interoperable because there is no effective and error-free method to identify semantic equivalence of custom data elements.

Data Collection and Analysis

The US GAAP Taxonomy is an open data standard that is publicly available. Financial statements submitted to the SEC are also publically available. To support our ongoing research, we have created a system that monitors the SEC website and automatically downloads financial statements. We analyzed all

official XBRL filings submitted to the SEC as of July 31, 2010, hereafter referred to as the July dataset. To identify any potential longitudinal trend, we also analyzed an additional snapshot as of February 26, 2010, referred to as the February dataset. We plan to capture and analyze another data snapshot for all official XBRL filings as of October 31, 2011. Most financial statements in the October 2011 data set will be based on the 2011 version of GAAP taxonomy, and thus will support our research in standard evolution.

We realize that companies are from different industries. Thus we group companies according to industry. The Standard Industrial Classification (SIC) was first introduced in the 1930s and may not accurately represents industries of the modern economy. The North American Industry Classification System (NAICS), introduced in 1997, provides an improved industry categorization. SEC filings contain SIC codes. We extract these codes and use the mappings provided at www.naics.com to obtain the corresponding NAICS codes. Companies are then categorized according the first two digits of the NAICS codes. Not all SIC codes have a corresponding NAICS code. We group companies without NAICS codes into one category.

In addition to analyzing data instances to obtain interoperability measures, we conduct case studies on two companies to understand how and why they extend the GAAP Taxonomy. These detailed analyses allow us to explain impacts of standards' extensibility on standard-based data's interoperability. For this purpose, we analyzed the financial statements of Lockheed Martin and Northrop Grumman.

Preliminary Findings

In this section, we report preliminary findings of the study. We will update the results and findings after we analyze the October 2011 dataset and report them at the workshop.

Characteristics of Datasets

The 2009 version of the GAAP taxonomy specifies a total of 13,452 data elements, among which 2,653 are abstract and 346 were deprecated on January 31, 2009. The number of concrete elements can be used in financial statements is 10,799, of which 10,537 are active (not deprecated).

Certain companies have filed multiple times during the time periods of both snapshots. When constructing the dataset for the 10K's, we keep only the first valid 10K if a company has filed more than once. Several companies submitted the financial statement for the same reporting period twice with the second overriding the first one. In this case, we only include the valid filing and exclude the one being overridden. The characteristics of the two snapshots of SEC filings are presented in Table 1. The July dataset has more than twice as many companies and filings as the February dataset partly because it also includes the 10Q's of many medium-sized companies, which according to the SEC mandate should start to use XBRL starting June 15, 2010.

Table 1. Characteristics to Two Datasets

Dataset	# Companies	# Filings	# 10K's	# GAAP Elements in 10K	# Custom Elements in 10K
02/26/2010	483	1231	261	2,083	4,403
07/31/2010	1,119	2,884	452	2,690	7,508

As more companies submit their filings, more GAAP elements are used and more custom elements are introduced. On average, an individual 10K statement uses less than 200 data elements, among which 129 are GAAP elements. While the number of companies and filings nearly doubled between two snapshots, the number of unique GAAP elements used by the community of filers did not increase as much. The number of GAAP elements used by all companies increased 29.14%.

Interoperability of Financial Statements

Interoperability of all 10K's can be computed for each possible pairs and triples. Out of the 261 10K's of the February dataset, there are 33,3390 pairs and 2,929,290 triples. Out of the 452 10K's of the July dataset, there are 101,926 pairs and 15,288,900 triples. The summary statistics of these interoperability scores of the two datasets are reported in Tables 2 and 3, respectively. Note the bolded "Mean" row has values corresponding to I_2, I_2', I_3, and I_3', as previously defined.

Table 2. Interoperability of 261 10K's, 02/2010

	I_{ij}	I_{ij}'	I_{ijk}	I_{ijk}'
Min	0.1033	0.1266	0.0300	0.0374
Max	0.7646	0.8654	0.5150	0.5809
Mean	**0.3724**	**0.4250**	**0.2435**	**0.2781**
Median	0.3798	0.4340	0.2456	0.2811
Standard deviation	0.0837	0.0859	0.0629	0.0668

Table 3. Interoperability of 452 10K's, 07/2010

	I_{ij}	I_{ij}'	I_{ijk}	I_{ijk}'
Min	0.0770	0.1151	0.0181	0.0225
Max	0.7646	0.8654	0.5354	0.5936
Mean	**0.3637**	**0.4151**	**0.2360**	**0.2695**
Median	0.3692	0.4222	0.2349	0.2694
Standard deviation	0.0842	0.0852	0.0618	0.0649

The ranges of pair-wise and triple-document interoperability increase as the dataset grows. For example, the *min* of pair-wise interoperability I_{ij} is 0.1033 for the February dataset, and reduces to 0.0770 for the July dataset. The *max* has stayed the same. The mean interoperability slightly decreased from the February dataset to the July dataset. For the July dataset, the mean is 0.3637, meaning that on average, 36.37% data elements in the 10K's of randomly picked two companies are interoperable. More importantly, I2' is higher than I', and I3' is higher than I3'. Both results are statistically significant. This result shows that if no custom elements were introduced, interoperability of data instances would be higher.

We hypothesized that the interoperability of filings for companies in the same industry might be higher due to their tendency of having similar assets and reporting structures. To test this hypothesis, we have classified the companies in the July dataset into 12 different industries according to their NAICS codes, and computed the interoperability within each industry. The results are summarized in Table 4. Since the three-document interoperability requires at least three documents, any industry with less than three 10K's are grouped into the "Other" category.

Table 4. Interoperability of Financial Statements within Industry

Industry	# 10K's	I_2	I_2'	I_3	I_3'
Mining	39	0.4058	0.4588	0.2776	0.3140
Utilities	33	0.3853	0.4590	0.2615	0.3117
Manufacturing	137	0.4247	0.4725	0.2931	0.3262
Wholesale Trade	5	0.4795	0.5400	0.3434	0.3863

Retail Trade	22	0.4426	0.4940	0.3080	0.3438
Transportation and Warehousing	18	0.3701	0.4368	0.2441	0.2883
Information	19	0.4263	0.4908	0.2971	0.3422
Finance and Insurance	72	0.2951	0.3619	0.1818	0.2232
Professional, Scientific, and Technical Services	7	0.4353	0.4869	0.3128	0.3502
Health Care and Social Assistance	3	0.4940	0.5432	0.3723	0.4095
Arts, Entertainment, and Recreation	5	0.4106	0.4838	0.2765	0.3260
Other	92	0.3968	0.4469	0.2629	0.2961
Total/Average	452	0.3937	0.4495	0.2657	0.3032
Total/Average, excluding Finance/Insurance	380	0.4224	0.4661	0.2816	0.3183

Except for Finance and Insurance industry, the within-industry interoperability scores are higher than those of all companies across all industries (which are the in the bolded row of Table 3 and represent the interoperability of randomly selected companies). The weighted averages of within-industry interoperability are also higher than the scores of randomly selected companies. For example, weighted average within-industry pair-wise interoperability is 0.3937, higher than 0.3637 of random comparison. The scores for the Finance and Insurance industry are lower mainly because we only considered the US GAAP taxonomy in the interoperability definition, but certain financial companies also used an investment taxonomy specific to their industry. When Finance and Insurance industry is excluded, the within-industry interoperability is higher than the interoperability of companies across all industries. For example, the weighted average of within-industry pair-wise interoperability is 0.4224, much higher than 0.3637, the overall interoperability among all companies. Thus our results provide evidence to support the hypothesis.

Case Studies: How Companies Extend GAAP Taxonomy

In the preceding subsection, we notice that the use of custom elements leads to lower interoperability. We analyze financial statements of Lockheed Martin (LMT) and Northrop Grumman (NOC) to understand why they extend the GAAP Taxonomy and whether the extensions are necessary. We also examine whether the 2011 version of the GAAP taxonomy has absorbed (or eliminated the need of) any certain custom elements and therefore will increase the interoperability in the future.

LMT used 49 unique custom elements in its SEC filings as of the second quarter of 2010. Most of the elements were used in the 10-Q filing dated July 29th, 2010, while a few were from prior filings and not used in later filings. For each custom element, we tried to manually identify potentially matching or closely related GAAP elements. Among the 49 data elements, we identified eight data elements that are not necessary as there are corresponding GAAP elements (see Table 4).

Table 4. Unnecessary Custom Elements

LMT Element	Corresponding GAAP Element
LossContingencyDamagesSoughtCompensatory	LossContingencyDamagesSought
QualifiedDefinedBenefitPlanContributionsByEmployer	PensionContributions
FairValueAssetsAndLiabilitiesMeasuredOnRecurringBasisTextBlock	FairValueAssetsMeasuredOnRecurringBasisTextBlock
EarningsPerShareTextBlock	EarningsPerShareTextBlock
LongTermDebtBeforeUnamortizedDiscount	LongTermDebt
OtherComprehensiveIncomePostretirementBenefitPlansReclassification AdjustmentForRecognitionOfPriorPeriodAmountsNetOfTax	OtherComprehensiveIncomeDefinedBenefitPlans AdjustmentNetOfTaxPeriodIncreaseDecrease
OtherComprehensiveIncomePostretirementBenefitPlans UnrecognizedAmountsNetOfTax	OtherComprehensiveIncomeDefinedBenefitPlansNet UnamortizedGainLossArisingDuringPeriodNetOfTax
OtherComprehensiveIncomePostretirementBenefitPlans UnrecognizedAmountsTax	OtherComprehensiveIncomeDefinedBenefitPlansNet UnamortizedGainLossArisingDuringPeriodTax

The remaining 41 custom data elements were deemed necessary extensions. We identified four reasons whey these extensions were necessary, and the number of custom elements for that reason:

- "Missing parent" issue (14 elements): e.g., FairValueAssetsMeasuredOnRecurringBasis. GAAP has an element for each type of asset and liability and its "fair value" and "measured on a recurring basis", such as FairValueAssetsMeasuredOnRecurringBasisDerivativeFinancialInstrumentsAssets, but it does not have one for all assets, all liabilities, or net assets (which is calculated as assets minus liabilities). LMT develops a custom element to sum up all fair value assets measured on a recurring basis.
- High level of detail (14 elements): e.g., SharesAuthorizedForRepurchase. GAAP taxonomy has an element for total number of shares, but does not have any elements for the number of shares authorized for repurchase.
- Wording limitations (4 elements): e.g., ComprehensiveIncomeTableTextBlock. This custom element is to include both table and text. GAAP taxonomy has ComprehensiveIncomeTextBlock, which is for text only.
- Unique situation (9 elements): e.g., PercentageOfEnvironmentalRemediationCostsReimbursed. This custom element refers to a certain environment law that applies to LMT.

We made several observations based on the case study of LMT financial statements. Most custom elements were indeed necessary. A majority of custom elements were due to high level of detail absent in GAAP taxonomy, or the "missing parent" issue in GAAP. Generalizing these observations as hypotheses, we studied financial statements of Northrop Grumman (NOC), a company in the same industry as Lockheed Martin (LMT). The NOC case study confirmed the same observations (Table 5). Among its 55 custom elements, only five were unnecessary. We further hypothesized that the 2011 version of the GAAP taxonomy could potentially address these issues by adding standard elements that describe detailed concepts common to the filing companies, completing existing hierarchies of accounting concepts with high-level parent elements, and clarifying or refining certain elements to avoid wording limitations. Since custom elements are best suited to describe unique situations of each filing company, we did not expect the 2011 version of the taxonomy to add any new standard elements to address situations unique to certain companies.

We tested our hypotheses when the 2011 version of GAAP taxonomy was released in March 2011. Indeed, the new version of GAAP taxonomy addressed several "missing parent", "high level detail" and "wording limitation" issues. For example, GAAP 2011 added AssetsFairValueDisclosureRecurring, which should eliminate the need of LMT's custom element FairValueAssetsMeasuredOnRecurringBasis. For another example, LMT's custom element SharesAuthorizedForRepurchase can now be replaced by a new element in GAAP 2011, StockRepurchaseProgramNumberOfSharesAuthorizedToBeRepurchased. A summary of the results on how the 2011 version of GAAP taxonomy may reduce the need of custom elements (and therefore increase interoperability) is presented in Table 5.

Table 5: The reasons for custom elements and how GAAP 2011 may render custom elements unnecessary

	LMT		NOC	
	Custom elements	Resolved by GAAP 2011	Custom elements	Resolved by GAAP 2011
Missing Parent	14	4	6	0
Details Needed	14	3	26	5
Wording limitation	4	2	10	3
Unique Situation	0	0	8	0
Total	41	9	50	8

Among the small number of custom data elements of the two companies, we found a common 2011 version Taxonomy element: StockRepurchaseProgramNumberOfSharesAuthorizedToBeRepurchased. Although we cannot extrapolate from the case studies to estimate the number of custom data elements the 2011 version of the Taxonomy will help to eliminate, it is generally true that interoperability increases when companies use more Taxonomy elements and less custom elements. We plan to analyze the interoperability among financial statements in the October 2011 data set. We will expand our case study to include the financial statements in the October 2011 data set for LMT, NOC and other companies.

Conclusion

Semantic interoperability of data is needed to support effective use of data from multiple sources. Data standards have the potential of improving semantic interoperability of data. In this ongoing research, we try to understand to what extent a data standard can help improve data's semantic interoperability and how standard's extensibility affects data's interoperability. With the use of the GAAP Taxonomy as a data standard, more than 42% of the information between two companies within the same industry can be directly compared. Although this level of interoperability may seem low, it is in fact quite remarkable because without the data standard, no data can be reliably matched for comparison. A standard's extensibility allows for more flexibility to report and share additional information. Its side effect is that it reduces interoperability of data. In the case of financial reporting using the GAAP Taxonomy, the within industry interoperability would be more than 46% if extensions were not allowed.

There are other factors that affect interoperability of data crated using an extensible data standard. In our case studies of two companies, we observe unnecessary extensions, where a custom element is introduced when a standard data element should have been used. As of July 31, 2010, we observe that more than 30,000 custom elements had been introduced, many of which might be unnecessary. Certain technologies should be developed to aid users identify elements in a given data standard, which will minimize unnecessary introduction of data standards. We have also observed that certain custom elements are necessary and adding commonly introduced custom elements to an existing data standard can potentially increase data's interoperability. In practice, the Financial Accounting Standards Board maintains the GAAP Taxonomy and has an ongoing effort to determine which custom elements need to be admitted into the next version of the GAAP taxonomy. With the planned analysis for October 2011 dataset, we will be able to investigate the impact of new GAAP 2011 Taxonomy on companies' filing practice and interoperability. Hopefully, the new elements introduced in GAAP 2001 Taxonomy will reduce custom extensions and increase the commonality among different companies' statements. At the same time, we have to caution that when the size of data standard increases, its complexity increases, which can increase chances of misunderstanding and misuse. A standard's relevancy to an individual user also decreases with the size of the standard (Zhu and Wu 2011b). Our future research will examine these factors and develop a deeper understanding on how to create data standards to maximally achieve semantic interoperability of data.

References

Bernstein, P.A., and Haas, L.M. "Information integration in the enterprise," *Commun. ACM* (51:9) 2008, pp 72-79.

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.

Markus, M.L., Steinfield, C.W., Wigand, R.T., and Minton, G. "Industry-Wide Information Systems Standardization as Collective Action: The Case of the U.S. Residential Mortgage Industry," *MIS Quarterly* (30:Special Issue) 2006, pp 439-465.

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

Rahm, E., Do, H.-H., and Maßmann, S. "Matching Large XML Schemas," *ACM SIGMOD Record* (33:4) 2004, pp 26-31.

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.Y., and Strong, D.M. "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, 2006.

Zhu, H., and Wu, H. "Interoperability of XBRL Financial Statements in the U.S.," *International Journal of E-Business Research* (7:2) 2011a, pp 18-33.

Zhu, H., and Wu, H. "Quality of Data Standards: Framework and Illustration using XBRL Taxonomy and Instances," *Electronic Markets* (21:2) 2011b, pp 129-139.