3C2-OS-13b-7

大規模会計 Linked Data のためのシステムアーキテクチャ

System Architecture for a Large Scale of Accounting Linked Data

鈴木 健太 Kenta SUZUKI　玉川 奨 Susumu TAMAGAWA　山口 高平 Takahira YAMAGUCHI

慶應義塾大学
Keio University

This paper discusses how to develop a RDF model of accounting domain and create its data from XBRL provide by EDINET, which is electronic disclosure system of Financial Service Agency in Japan, and also discuss how to handle a huge Kinked Data and design its architecture in relation to database and its application server.

1. はじめに

XBRL(eXtensible Business Reporting Language)による財務報告及び情報開示が金融庁 EDINET で正式に開始されてから 4 年が経過した. 従来 HTML と PDF によって開示されていた財務情報が, XBRL で構造化されて開示されたことにより, 機械的な処理が容易になった.

また, 一方で近年, Linked Open Data(LOD)が注目されている. Linked Open Data の活動は会計データの利用促進という面からも, 大いに意義があるといえる. XBRL から抽出したデータを Linked Data として公開し, 財務情報のドメイン外のデータと結びつけることで, 財務情報のみからは得られなかった知見が機械自動的に得られるようになる可能性がある. また, XBRL のデータは企業データのコアとしての役割を担うことも期待され, 財務分析及び報告といった現在の XBRL の利用シーン以外にも, XBRL をベースとしたデータが役立つものになると考えられる. Linked Data として XBRL のデータを開示するためには, 外部データとの連携に適した RDF モデルを構築する必要がある. XBRL における財務データの表現は XLink を利用し, 各種勘定科目の定義の情報などを相互のリンクを利用して記述している. RDF においてもこのような表現は可能であるが, あまりに XBRL 的な記述方法になってしまうと, 外部の開発者が利用する際にモデルの複雑さによって, データを扱いづらくなってしまう可能性がある. 財務情報を正確に示しつつ, かつ, 外部の開発者にとっても利用しやすい形でデータを提供するためのモデルを構築する必要がある.

構築した RDF モデルにデータを当てはめるため, XBRL からデータを抽出していく. Linked Data で利用するための RDF を構築するためには, XBRL パーサを実装し, 勘定科目の定義階層を含めて的確にデータを抽出する必要がある. 対象となる XBRL ファイルの数も多いため, パーサも効率の良いものを実装しなければならない. 将来的に XBRL ファイルから RDF ファイルへのシームレスな変換をするには, XBRL パーサのパフォーマンスは解決しなければならない課題の 1 つとなるだろう.

また, LOD を安定して供給するには, 適切な RDF モデルを構築した上でスケーラブルなシステム環境を作成しなければならない. Linked Data を構築し, データを公開するためのシステムは幾つか開発されているが, データの規模や利用方法によって適切なシステムアーキテクチャは異なる. エンドユーザからの SPARQL による問い合わせの傾向を分析し, システムの可用性を担保しつつ, 快適なパフォーマンスを保たなければならない. 今回構築する XBRL を基にした RDF データは非常に巨大なものであり, これを提供するためのバックエンドの構成についても熟慮を重ねる必要がある.

連絡先：鈴木健太, 玉川奨, 山口高平　慶應義塾大学理工学部　〒223-8522　神奈川県横浜市港北区日吉 3-14-1
E-mail : {k_suzuki,s_tamagawa,yamaguti}@ae.keio.ac.jp

2. XBRL

XBRL の基礎的な仕組みについて説明する.

2.1 XBRL の概要

XBRL は財務諸表の電子的な報告に XML を応用したものである. XBRL の導入により, 各種財務報告用の情報の作成・流通・利用が効率的に行われることが期待されている. XBRL が導入される前は, PDF 及び HTML の形式で財務報告が行われていており, 構造化されたデータを取得するには高いコストが必要であった. XBRL 標準で定義されたデータは, 財務情報を利用する企業間で効率的に流通させることが可能となる.

XBRL は大きく分けて2つの要素から構成される. 各企業の財務情報が記述されたインスタンス文書, 項目のタグ名（語彙）を定義したタクソノミ文書である. インスタンス文書には, タクソノミ文書はタクソノミスキーマとリンクベースによって構成されている.

2.2 タクソノミスキーマ

XBRL のスキーマ記述には XML Schema が利用されている. タクソノミスキーマでは, XBRL のもつデータの定義を記述している. XBRL においては名前空間及び各リンクベースファイルのインポートの機能を利用し, 財務情報定義を導入している. タクソノミスキーマは3つの階層をなしており, GAAP タクソノミ, 産業別タクソノミ, 企業別タクソノミによって構成される. 各タクソノミは, 5 種類のリンクベースをインポートし, 当該 XBRL インスタンスの各勘定科目や文書情報の定義に関する情報を提供する.

GAAP タクソノミ及び産業別タクソノミについては, 会計規則に基づいたものを金融庁が毎年 1 回発表している. 企業別タクソノミは各企業独自の勘定科目を定義するものであり, 開示書類ごとに個別に含まれている.

2.3 リンクベース

リンクベースには表示リンク, 計算リンク, 定義リンク, 名称リンク, 参照リンクの5つがある.

表示リンクでは各勘定科目の表示順を定義する. 計算リンクは各勘定科目間での重みつき加算式を定義している. 定義リンクは項目間の同義性や出現規則などが示されている. 名称リンクは様々な言語における各勘定科目の呼称が定義されている.

*1 http://info.edinet-fsa.go.jp/

参照リンクは各勘定科目の会計概念定義の根拠となっている文献へのリンクが示されている.

タクソノミスキーマに定義された項目に対して, 各項目間の関係や, 各項目に対する追加情報などを, XLink[DeRose, S ら 2001]の外部リンク機能を利用して定義したものがリンクベース

である. 具体的には, 各勘定科目の表示順序や, 計算方法, 勘定科目として表示される値のラベルの定義などを行う. これらの定義は, タクソノミスキーマとは別のファイル(=リンクベースファイル)として作成される. XBRL ではこれらのリンク定義を個別のリンクベースとして, ファイルを分けて作成することができる.

XBRL を使って財務情報を作成する場合, 各国の会計制度に対応したタクソノミ文書に加え, 業種などで共通化されたタクソノミ文書を利用する. さらに, 自社独自の情報については, 自社タクソノミ文書を作成する. こうして作成されたタクソノミ文書を元にインスタンス文書を作成する.

3. 関連研究

関連研究としては, Publishing XBRL as Linked Open Data[R. Garcia ら 2009]があげられ, EDGAR を対象として XBRL データを RDF に変換している. Publishing XBRL as Linked Open Data では XBRL から RDF への変換を XML の構造をそのまま置き換えることによって実現しているため, XBRL の記述形式に近い RDF を出力している. しかし, このデータは財務報告データの一部分のみをマッピングしたものであり, XLink によって関連付けられている XBRL の全データを変換するには至っていない.

Representing Financial Reports on the Semantic Web [J. Bao ら 2010] においては, XBRL を意味的に解釈し, XBRL のスキーマを OWL に変換を行っている.

また, 会計ドメインにおける RDF モデルの構築と Linked Data との連携[鈴木 2011]では, 会計ドメインにおける RDF モデルの構築について言及した. 本稿では Linked Data としてクエリングの効率性などを考慮し, RDF モデルを変更している.

4. 提案システム

本稿では、金融庁 EDINET の公開している XBRL を対象として、RDF への変換を行う. システムの実装には Java を使用し, データベースには TDB[*2] を利用している. また, SPARQL を利用するための HTTP ラッパーとして Joseki*3 を採用した.

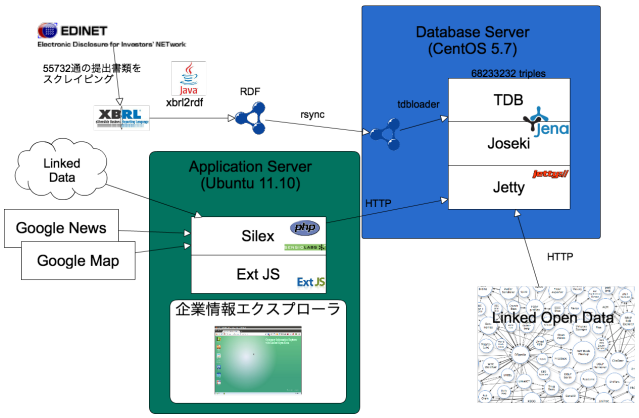

図 1 システムの全体像

4.1 システムの全体像

本システムの全体像を図 1 に示す. 本システムは大きく分けて Linked Data の生成部と利用部に分けられる. Linked Data は以下の手順で作成した.

1. XBRL ファイルの取得
2. XBRL インスタンスから発見可能なタクソノミ集合(DTS: Discoverable Taxonomy Set)の識別
3. 各ファイルのパース
4. RDF のファイルシステムへの出力
5. データベースへの格納

DTS とは, XBRL インスタンスから参照されている全ての XML スキーマファイル及びリンクベースファイルのことを指す. XBRL をパースするためには, XBRL インスタンスだけではなく, すべての DTS をパースし, 各種定義情報を取得しなければならない. DTS を発見する方法は, 拡張可能な事業報告言語(XBRL) 2.1)[JISX7206]の第 3 章において, 定義されている.

抽出対象とした XBRL ファイルの規模は表1のとおりである.

表 1 対象とした XBRL ファイル

対象期間	2008 年 4 月～2011 年 4 月
1ファイルの生成にかかる時間	約 3500msec
1年分の XBRL の変換のためにかかる時間	5532279msec
総ファイル数	57678 ファイル
総行数	87982410 行

XBRLから生成したLinked Dataを利用したアプリケーションについても構築を行った. アプリケーションについては 4.3 で紹介する.

4.2 RDF モデル

XBRL インスタンスから得られた情報を RDF として出力する. 会計ドメインオントロジーとの整合性を担保するため, 以下の情報をアウトプットする.

図 2 に出力する RDF のモデルを提示する. 独自に定義したプロパティについての説明は表 2 のとおりである. なお, xbrl-owl の名前空間 URI は http://www.yamaguti.comp.ae.keio.ac.jp /xbrl_ontology/owl# としている.

モデルについての具体的な説明を行う. 各企業が提出した書類について xbrlowl:hasReport プロパティでその関係を示し, 値に提出書のインスタンスを配置した. 提出書のインスタンスは xbrlowl:report を rdf:type と関連付けられており, 提出書類であるということを明示している. 提出書類のインスタンスは複数のコンテキストをもつ. コンテキストには連結と非連結, 損益計算書と貸借対照表, そして会計期間の3つの情報が含まれている. 各コンテキストは財務諸表上の1つ1つの項目に参照される.例えば資本金は貸借対照表の資本の部に含まれるが, 非連結と連結の場合に共に存在し, かつ, 複数の会計期間で現れるものである. 提出書類1つの中にも複数の会計期間の財務諸表が掲載されているため, それらを区別するためにこのようなコンテキストという概念を導入している. このコンテキストの概念自体は XBRL の仕様に含まれているものであるが, このようにグラフモデルで表現することでさらに柔軟かつ容易に利用することができるようになっている.

*2 http://openjena.org/TDB
*3 http://www.joseki.org/

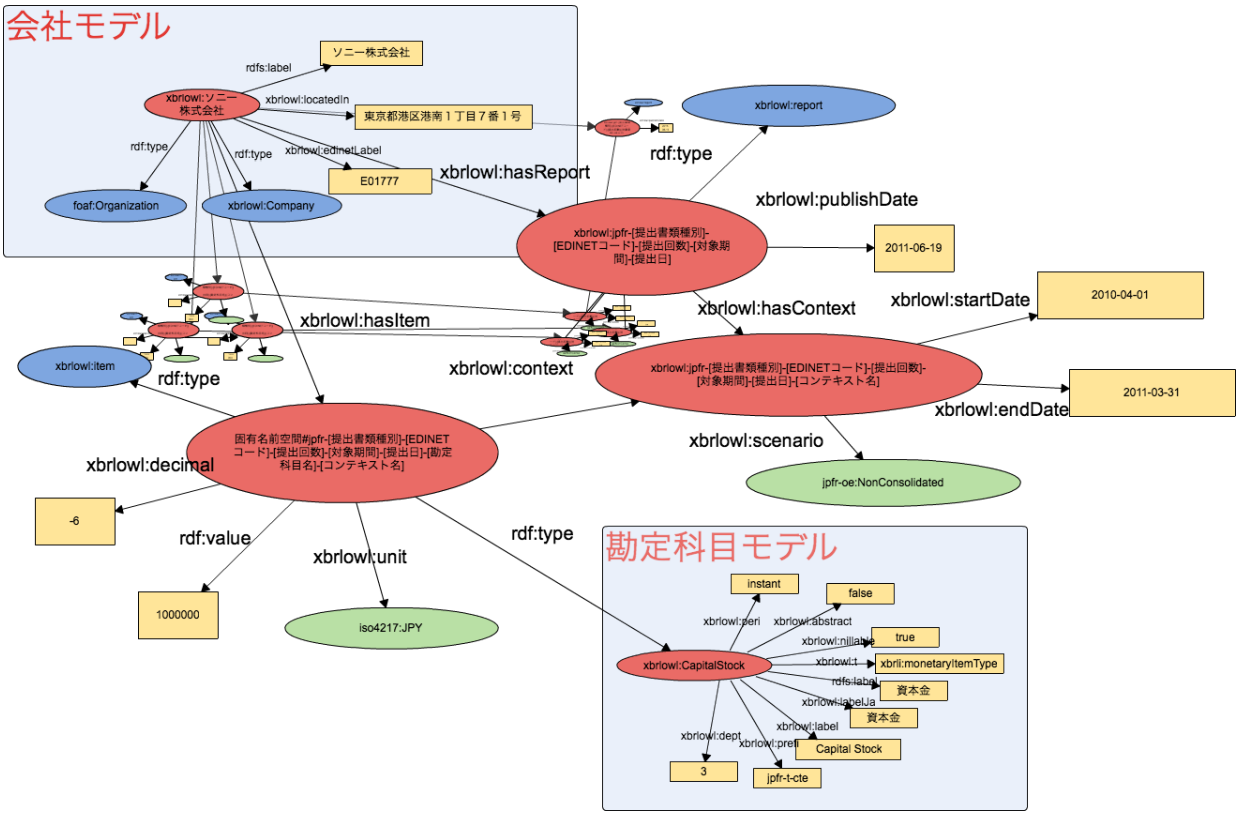

図 2 RDF モデル

表 2 独自定義プロパティ一覧

プロパティ名	説明
hasReport	会社リソースが提出した書類リソースとの関係
hasItem	会社リソースが提出した書類内で記述している財務諸表の各項目に関するインスタンスとの対応関係
context	財務諸表の各項目のインスタンスが属するコンテキストとの対応関係
hasContext	各提出書類インスタンスに含まれるコンテキストのインスタンスとの関係
unit	財務諸表の各項目のインスタンスの通貨単位情報を示す
publishDate	各提出書類インスタンスの提出日を示す
startDate	各コンテキストのインスタンスの会計期間の始まりを示す．このプロパティは貸借対照表のコンテキストの場合のみ存在する．
endDate	各コンテキストのインスタンスの会計期間の終わりを示す．このプロパティは貸借対照表のコンテキストの場合のみ存在する．
hasInstantDate	各コンテキストのインスタンスの会計期間を示す．このプロパティは損益計算書のコンテキストの場合のみ存在する．
decimal	財務諸表の各項目の表示単位情報を示す．通常は百万円単位となっている．
scenario	各コンテキストのインスタンスのシナリオ情報を示す．シナリオ情報とはそのコンテキストの連携・非連結の状態を示すものである．

財務諸表上の勘定科目１つ１つは会社インスタンスと xbrlowl:hasItem プロパティで関連付けられている．この具体的な勘定科目のインスタンスは xbrlowl:item クラスのインスタンスとして定義されており，rdf:type によって対応付けされている．xbrlowl:hasItem プロパティは会社インスタンスがそれまで提出した有価証券報告書などで記述した勘定科目インスタンス1つ1つについての対応関係を示すプロパティである．アプリケーションなどで各勘定科目の値を取得する場合には，このインスタンスから取得することになる．

RDF から値を取得する際には RDF モデル全体の構造を知っている必要がある．また，RDF モデルの構造を踏まえた上でクエリを行わなければならない．xbrlowl:item によって勘定科目のインスタンスを定義し，かつ，会社インスタンスと xbrlowl:hasItem を連結させた理由に利用者の利便性の向上という点がある．

4.3 アプリケーション実装

構築した Linked Data を利用し，企業情報を検索するためのサンプルアプリケーションを実装した．アプリケーション構築環境は以下の通りである．

OS: Ubuntu 11.10
Web サーバ: Apache 2.2
サーバサイド実装: Silex (PHP 製マイクロフレームワーク)
フロントエンド: ExtJS

プロジェクトページは http://suzuken.github.com/Corporate InformationExplorer/ にて公開している．Silex から LOD サーバに curl を利用して SPAQL を実行し，結果を取得するようにしている．企業に関する XBRL から取得される情報の一覧を閲覧す

ることができる．アプリケーションは企業検索，財務情報閲覧，ニュースリーダー等の機能を有している．

なお，クライアントサイドとサーバサイドとのデータのやり取りは JSON で行われているため，RDF を JSON にシリアライズし，クライアントサイド JavaScript での利用を容易にしている．

企業情報エクスプローラでは，企業の情報を横断的に検索することができる．現在は企業の財務情報はもちろんのこと，各企業のニュース情報を抽出することができる．ニュース情報の取得には Google News API を利用している．

将来的には他の Linked Data と連動して，企業に関する様々な情報を集約することを予定している．例えば，日本の新聞社が Linked Data を公開した場合，各企業に関する報道記事がよりシームレスに表示することが可能になるだろう．

図 3 企業情報エクスプローラトップ画面

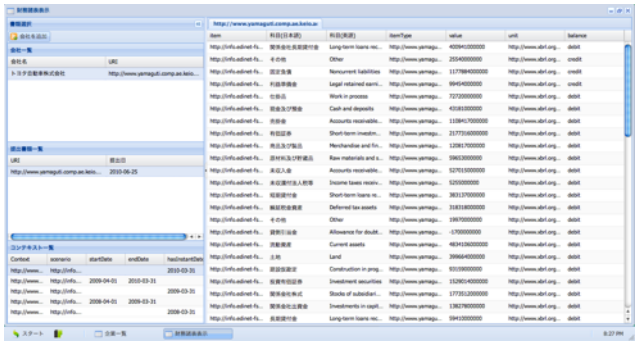

図 4 財務情報表示画面

5. 評価

5.1 作成したデータの規模

作成したデータの規模について述べる．名前空間別にまとめたものを表 3，概念別に区分けしたものを表 4 にまとめた．

表 3 作成したデータの一部（名前空間別）

プロパティ名	トリプル数
rdf:type	26169967
rdfs:label	21609
xbrlowl:hasCompany	55732
xbrlowl:edinetLabel	10486
xbrlowl:hasContext	1003154
xbrlowl:hasInstant	150778
xbrlowl:scenario	108915
xbrlowl:hasItem	8322798
rdf:value	7715434
xbrlowl:context	8322744
総トリプル数	68233232

表 4 作成したデータ一覧（概念別）

概念	数
勘定科目インスタンス	4268
会社インスタンス	10486
提出書類インスタンス	55732
コンテキストインスタンス	108915
item インスタンス	8322798

5.2 企業間比較への応用

企業間での財務諸表を比較する際には，いくつか代表的な指標を用いて比較するのが一般的である．これは経年比較においてもその時点での比較においても同様である．XBRL によって入手できる勘定科目の各値は，各企業の財務報告の形式に依存するため，必ずしもすべての企業において必要な値を取得できるとは限らない．今回構築した Linked Data では ROA や ROE などの基本的な経営指標の値は算出できるが、固定費や変動費についての区別を得ることはできない。

また，海外での LOD の広がり方を見ると，BBC や News York Times といったメディア各社が Linked Data を構築しており，このような領域のデータと今回構築した Linked Data を連携させることで，新しいサービスの実現が用意になるだろう．将来の展望として，海外メディアと同様に日本のテレビ局や新聞社が LOD を構築することになった場合，XBRL をベースとした LOD を Web サイトの企業活動関連の記事上に対応付けることに寄って，今までは財務諸表を参照しなければ確認することの出来なかった財務データを其の場ですぐに確認できるようにすることができる．テレビ番組でも，その番組内で報じられる企業のインスタンスが対応付けられていれば，その企業に関するニュースを閲覧したいユーザーにシームレスに情報を伝えることも可能となるだろう．

6. おわりに

XBRL のデータを利用し，Linked Data を構築した手順及び方法について述べた．また大規模な Linked Data の構築を行った．今後この Linked Data の運用を続けていく中で，さらにアーキテクチャについて考察しなければならないことも増えるだろう．

参考文献

[玉川 10]玉川 奨，桜井 慎弥，手島 拓也，森田 武史，和泉 憲明，山口 高平: 日本語 Wikipedia からの大規模オントロジー学習，人工知能学会論文誌，Vol. 25, No. 5, pp.623-636 (2010) .

[R. Garcia, R. Gil 09] Publishing XBRL as Linked Open Data, In *CEUR Workshop Proceedings*, volume 538 (2009)

[J. Bao, G. Rong, X. Li, L. Ding 10] Representing Financial Reports on the Semantic Web - A Faith ful Translation from XBRL to OWL RuleML2010, LNCS 6403, pp. 144-152 (2010)

[DeRose, S., Maler, E., Orchard, D. 01] XML Linking Language (XLink) Version 1.0., Technical report, W3C (2001)

[JISX7206] 拡張可能な事業報告言語（XBRL）2．1 (2005)

[鈴木 11] 鈴木 健太，山口 高平: 会計ドメインにおける RDF モデルの構築と Linked Data との連携，2011 年人工知能学会全国大会，3E3-OS20-5