# 有価証券報告書に含まれるデータの企業間比較における課題について

佐藤栄作 [1]　木村泰知 [1]
[1] 小樽商科大学
g2020149@edu.otaru-uc.ac.jp　kimura@res.otaru-uc.ac.jp

## 概要

本研究は, 有価証券報告書に含まれるデータを企業間で比較する状況において生じる課題を明らかにした. 具体的には, 企業間の比較ができるデータを「要素 ID とコンテキスト ID の両者が完全一致しているデータ」と定義し, そのようなデータが TOPIX100 算出対象企業の有価証券報告書にどの程度含まれているのかを可視化した. 加えて, 要素 ID とコンテキスト ID が付与されていても, 企業間の比較ができないケースについても考察した. また, 改善策として XBRL と CSV を利用したデータセットの構想を示した.

## 1　はじめに

有価証券報告書は，金融商品取引所において株式を公開している企業が，事業年度ごとに外部に公開する重要な資料である．これは，企業間の比較分析において非常に重要な役割を果たしている. 有価証券報告書は, 原則として，EDINET [1)] [1] への電子提出が義務付けられており, PDF 及び XBRL[2)] の形式で EDINET 上に公開されている．EDINET のサイトからは，これらの報告書を直接ダウンロードすることが可能であり, また，EDINET API を利用することでも PDF や XBRL のダウンロードも可能である. XBRL ファイルにはタクソノミやインスタンスの情報が含まれており, 表内の必要なデータを特定し取得することが可能である. XBRL におけるタクソノミとは, いわば雛形のようなものであり, そこに数値や期間, 単位などのインスタンスが入れられるといった仕組みである. このように, 表内のデータに対してタクソノミが付与されることにより, 構造的な正確さを保持したデータ抽出が可能となる.

しかしながら, XBRL を用いても企業間の比較が困難な状況が散見される. それには主に 2 つの要因が考えられる. 一つは, 全ての表にタクソノミが付与されているわけではないという点である.TOPIX100[3)] の算出対象である企業の有価証券報告書に含まれる表において, タクソノミが付与されている表は全体のおよそ 18 %に過ぎないとの報告がある [2]．つまり, 必要なデータが取得できないために, 企業間の比較ができないという問題が生じている. もう一つの要因は, 企業が独自に定義することができるタクソノミや, 少数の企業のみが使用しているタクソノミが存在しているという点である. これは, タクソノミが付与されていたとしても, それが必ずしも企業間の比較を可能とするわけではないということである.

本研究の目的は, 有価証券報告書に含まれるデータについて企業間の比較を行う際の課題を明らかにし, その解決策を提案することである. 本研究の貢献は, 以下の 3 点である.

- 有価証券報告書における企業間の比較可能なデータを定義した (3.1 節）.
- 企業間でデータを比較するうえでの課題を明らかにした（4 章）.
- 解決策として, 企業間の比較に関するデータセットの構想を示した（4.4 節）.

## 2　関連研究

門脇らは, 有価証券報告書に含まれる表の各セルを Metadata, Header, Attribute, Data の 4 つに分類する TDE タスクを提案した [3]. しかしながら, このタス

1)　EDINET（Electronic Disclosure for Investors' NETwork）とは, 金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システムである.

2)　XBRL（eXtensible Business Reporting Language）とは, 各種財務報告用の情報を作成, 流通及び利用できるように標準化された XML ベースの言語である.

3)　TOPIX100 とは, TOPIX（東証株価指数）を構成する銘柄のうち, 流動性と時価総額の高い 100 銘柄を算出対象として選定した株価指数を指す.

クでは, 表の構造理解には取り組んでいるものの, データの抽出までは行っていない.

木村らは, 有価証券報告書内の本文と該当箇所のセルを関連付けるタスクとして, TTRE（Text-to-Table Relationship Extraction）タスクを提案した [2]. しかしながら, このタスクは抽出対象が単一の有価証券報告書に限定されており, 複数企業の有価証券報告書を跨いだ情報抽出は考慮されていない.

# 3　データの企業間の比較について

## 3.1　企業間の比較が可能なデータとは

本節では, 企業間の比較を行う場合, どのような情報が必要となるのかを示す. EDINET の API は, 2023 年 8 月 21 日に EDINET API version2 が公開されており, XBRL から取得できるタクソノミやインスタンスの情報を CSV 形式でダウンロードできるようになっている. ここでは, EDINET で公開されている有価証券報告書の XBRL ファイルおよび CSV ファイルを用いて説明する.

例として, 大和ハウス工業第 81 期の有価証券報告書を取り上げる. 図 1 は連結経営指標等の 1 行目に記述されている内容である. 図 2 は CSV ファイルの一部を抜粋したものであり, 図 1 で示されているデータの部分を抜粋している.

ひとつの有価証券報告書の範囲内では, 当該データの意味要素を定義する要素 ID は一意になっておらず, データの文脈要素 (いつのデータか, 時点・期間どちらか, 個別・連結どちらかなど) を定義するコンテキスト ID と合わせることで一意な識別子となる. 例えば,5 つのデータは全て「jpcrp_cor:NetSalesSummaryOfBusinessResults」の要素 ID を持つが, データの期間がそれぞれ異なるため, コンテキスト ID と合わせることで一意となっている.

したがって, 異なる企業の有価証券報告書から, 要素 ID ＋コンテキスト ID が同一のものを取得してくることにより, 企業間の比較を行うことができる. つまり, 企業間の比較が可能なデータとは, タクソノミによって要素 ID とコンテキスト ID が設定されており, 他の企業の有価証券報告書にも同様の要素 ID とコンテキスト ID を持つデータが存在するデータであるということがいえる. 本研究では, 以降,「企業間の比較が可能である」ということと「要素 ID ＋コンテキスト ID が一致する」ということを同義で扱う.

## 3.2　企業間の比較が不可能なデータとは

前節では, タクソノミによって要素 ID とコンテキスト ID が付与されていれば企業間の比較が可能であると述べたが, 逆にタクソノミが付与されていないデータは要素 ID やコンテキスト ID を取得することができないため, 当然企業間の比較もできないということがいえる.

また, 仮にタクソノミが付与されているデータであったとしても, 企業間の比較を行うことができない場合が考えられる. なぜなら, 他の有価証券報告書に同じ要素 ID ＋コンテキスト ID を持つデータが必ずしも存在するとは限らないためである. タクソノミには「EDINET タクソノミ」と,「提出者別タクソノミ」の 2 種類が存在する. 前者は, EDINET から提供される標準的なタクソノミである. 後者は, 前者の中に適切なタクソノミが存在しなかった場合に, 企業ごとに作成されるタクソノミである. 提出者別タクソノミでは, 企業独自の要素 ID が適用されるため, データに提出者別タクソノミが付与された時点で, 他の企業との比較が難しくなってしまう状況にある.

# 4　有価証券報告書データの分析

本章では, 実際の有価証券報告書に含まれるデータを分析し, 企業間の比較を行う際に課題となりうる点を調査する. 分析対象とする有価証券報告書は, TOPIX100（2021 年度）の算出対象企業, 計 100 社の 2020 年度提出分の有価証券報告書である. したがって, 対象となる有価証券報告書の数は 100 である.

## 4.1　メンバー要素

コンテキスト ID には, 末尾にメンバー要素が付与されているものが存在する. メンバー要素とは, ディメンション（ある概念をカテゴリ化, セグメント化したもの）の構成要素となるデータを記述する際に付与される要素である. タクソノミが付与されているデータ 200,279 個のうち, メンバー要素が付与されていないものは 67,920 個, 付与されているものは 132,359 個であった. また, メンバー要素の種類は, EDINET タクソノミに限定した場合, 2,495 種類であった. そのうち, 100 社全てが共通して使用していたのは 17 種類, 1 社のみが使用していたのは 2,207 種類と大半を占めていた. 前者のメンバー要素であっても,「Row1Member（1 行目のメンバー）」,

| 回次 | 第77期 | 第78期 | 第79期 | 第80期 | 第81期 |
|---|---|---|---|---|---|
| 決算年月 | 2016年３月 | 2017年３月 | 2018年３月 | 2019年３月 | 2020年３月 |
| 売上高 （百万円） | 3,192,900 | 3,512,909 | 3,795,992 | 4,143,505 | 4,380,209 |

**図 1**　有価証券報告書に含まれる表の例（大和ハウス工業第 81 期有価証券報告書 連結経営指標等より売上高の行を抜粋）

両者を組み合わせることで、一意なIDとなり比較・抽出に利用できる

| 要素ID | 項目名 | コンテキストID | 相対年度 | 連結・個別 | 期間・時点 | ユニットID | 単位 | 値 |
|---|---|---|---|---|---|---|---|---|
| jpcrp_cor:NetSalesSummaryOfBusinessResults | 売上高、経営指標等 | Prior4YearDuration | 四期前 | その他 | 期間 | JPY | 円 | 3192900000000 |
| jpcrp_cor:NetSalesSummaryOfBusinessResults | 売上高、経営指標等 | Prior3YearDuration | 三期前 | その他 | 期間 | JPY | 円 | 3512909000000 |
| jpcrp_cor:NetSalesSummaryOfBusinessResults | 売上高、経営指標等 | Prior2YearDuration | 前々期 | その他 | 期間 | JPY | 円 | 3795992000000 |
| jpcrp_cor:NetSalesSummaryOfBusinessResults | 売上高、経営指標等 | Prior1YearDuration | 前期 | その他 | 期間 | JPY | 円 | 4143505000000 |
| jpcrp_cor:NetSalesSummaryOfBusinessResults | 売上高、経営指標等 | CurrentYearDuration | 当期 | その他 | 期間 | JPY | 円 | 4380209000000 |

**図 2**　EDINET からダウンロードできる CSV の例（大和ハウス工業第 81 期有価証券報告書 CSV より一部を抜粋）

「Row2Member（2 行目のメンバー）」といった, 具体的なメンバーを指定しての情報抽出が難しいと思われるメンバー要素も見られた. したがって, メンバー要素の付与はコンテキスト ID の一致率が低下する大きな要因となっているということが言える.

更にメンバー要素が企業間の比較を複雑化させる例を挙げると, 図 3 では, 要素 ID は同じであるが, コンテキスト ID に付与されているメンバー要素（橙枠）がそれぞれ異なる 6 つのデータが示されている. この部分の PDF を見ると 4 通常は要素 ID に依存していた各データの項目名が, 要素 ID ではなくメンバー要素に依存していることがわかる.

本研究では, 企業間比較の際には同一の要素 ID ＋コンテキスト ID（メンバー要素を除く）を有していれば, 同一ディメンション内のデータであるため複数のデータを取得してもかまわないと仮定し, 異なるメンバー要素を持つコンテキスト ID は, それらを省略してメンバー要素のないひとつのコンテキスト ID とみなすこととする.

## 4.2　EDINET タクソノミ付与データ

EDINET タクソノミが付与されたデータの総数は, 192,221 個であった. 要素 ID とコンテキスト ID の組み合わせの種類は, 4,907 種類であった. 図 5 は, 要素 ID ＋コンテキスト ID の種類数を, 使用している企業の数ごとに示したグラフである. 全ての企業（100 社）に共通して使用されている要素 ID ＋コンテキスト ID の種類数は, 231 種類であった. 逆に, 1 社にのみしか使用されていない要素 ID ＋コンテキスト ID の種類数は, 1,034 種類であった. EDINET タクソノミは, 提出者別タクソノミと異なり, 全ての企業に共通していることを鑑みると, 要素 ID ＋コンテキスト ID の一致率に向上の余地があると考えられる.

## 4.3　提出者別タクソノミ付与データ

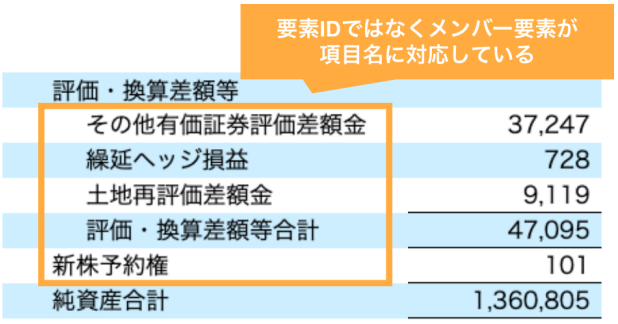
提出者別タクソノミが付与されたデータの総数は, 8,058 個であった. 提出者別タクソノミの要素 ID には, 企業独自の文字列が含まれるため, 当然ながら他者との比較が不可能である. 例えば, 大和ハウス工業の場合は,「jpcrp030000-asr_E00048-000:LandForDevelopmentCA」という要素 ID の「E00048」という部分が企業独自の文字列となる. 図 6 は, 要素 ID に含まれる企業独自の文字列をマスクしたうえで, 要素 ID ＋コンテキスト ID の種類数を, 使用している企業の数ごとに示したグラフである. 視認性を向上させるため, 1 社のみが使用している要素 ID ＋コンテキスト ID の種類数は省略している. 最大 11 の企業で同一の要素 ID ＋コンテキスト ID が設定されていたが, それらの種類数を合計しても, 1 社のみが使用している要素 ID ＋コンテキスト ID の種類数である 4,056 の 1 割にも満たないことが判明した.

## 4.4　課題解決に向けて

これらの課題に共通する要素としては, 企業間で完全一致した要素 ID ＋コンテキスト ID を付与, 並びに特定することの難しさが挙げられる. 理想的な解決方法としては, 全てのデータに企業間で統一したタクソノミが付与することであるが, その実現可能性は低い. タクソノミが付与されていないデータに関しては, データの特徴や文脈を考慮した暫定的な要素 ID とコンテキスト ID を自動で付与することにより, 企業を跨いでも所望のデータを得られる可能性がある. 既にタクソノミが付与されているデー


| 要素ID | 項目名 | コンテキストID | | 値 |
|---|---|---|---|---|
| jppfs_cor:NetAssets | 純資産 | CurrentYearInstant_NonConsolidatedMember | ValuationDifferenceOnAvailableForSaleSecuritiesMember | 37247000000 |
| jppfs_cor:NetAssets | 純資産 | CurrentYearInstant_NonConsolidatedMember | DeferredGainsOrLossesOnHedgesMember | 728000000 |
| jppfs_cor:NetAssets | 純資産 | CurrentYearInstant_NonConsolidatedMember | RevaluationReserveForLandMember | 9119000000 |
| jppfs_cor:NetAssets | 純資産 | CurrentYearInstant_NonConsolidatedMember | ValuationAndTranslationAdjustmentsMember | 47095000000 |
| jppfs_cor:NetAssets | 純資産 | CurrentYearInstant_NonConsolidatedMember | SubscriptionRightsToSharesMember | 101000000 |
| jppfs_cor:NetAssets | 純資産 | CurrentYearInstant_NonConsolidatedMember | | 1360805000000 |

**図 3**　要素 ID が同じであるが, コンテキスト ID にメンバー要素が付与されているデータの例（大和ハウス工業第 81 期有価証券報告書 CSV より一部を抜粋）




| | |
|---|---|
| 評価・換算差額等 | |
| その他有価証券評価差額金 | 37,247 |
| 繰延ヘッジ損益 | 728 |
| 土地再評価差額金 | 9,119 |
| 評価・換算差額等合計 | 47,095 |
| 新株予約権 | 101 |
| 純資産合計 | 1,360,805 |

**図 4**　メンバー要素が項目名に相当している例（大和ハウス工業第 81 期有価証券報告書 貸借対照表より一部の当期項目のみを抜粋）

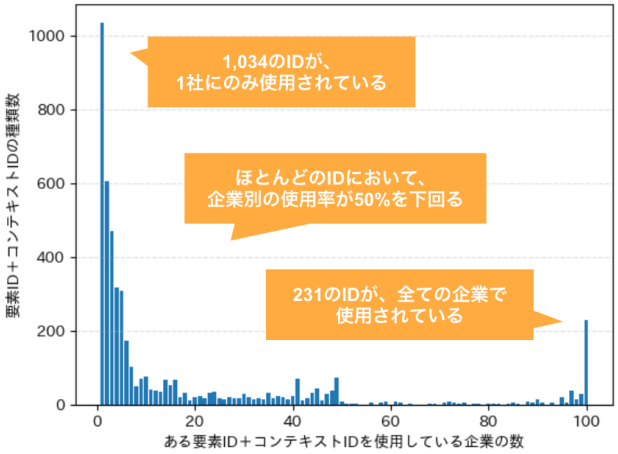



**図 5**　EDINET タクソノミにおける, 使用企業数ごとの要素 ID ＋コンテキスト ID の種類数

タに関しては, 要素 ID とコンテキスト ID の意味的側面に着目して, 企業を跨いで同類のデータを取得できる可能性がある.

したがって, 解決策として EDINET からダウンロード可能な XBRL と CSV を用いて, それらに活用できるようなデータセットを構築することが考えられる. 例えば, インライン XBRL（XBRL の情報を HTML で記述したもの）で記述された有価証券報告書と, それに該当する CSV を利用することで, 要素 ID やコンテキスト ID の命名パターンを文脈を踏まえて学習することや, 意味的側面の一致を構造的側面から裏付けることも可能になると思われる.

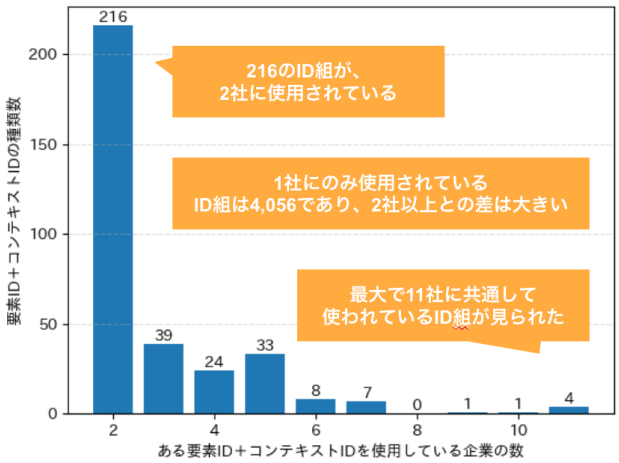



**図 6**　提出者別タクソノミにおける, 要素 ID の独自部分をマスクした際の, 使用企業数ごとの要素 ID ＋コンテキスト ID の種類数

## 5　おわりに

本研究では, 有価証券報告書に含まれるデータを企業間で比較する際の課題を明らかにし, その解決策を提案することを目的として調査を行った. 企業間の比較が可能であるデータは, 要素 ID ＋コンテキスト ID が一致しているデータであると定義した. 企業間比較の現状の課題としては, タクソノミが付与されていないデータが多くあること, メンバー要素によりコンテキスト ID の一致率が低下してしまうこと, EDINET タクソノミの要素 ID ＋コンテキスト ID の一致率が低いこと, 提出者別タクソノミが存在することを示した. これらの課題改善のため, データセットを活用した自然言語処理的アプローチを提案した. 今後の展望としては, 本研究の結果を勘案して実際にデータセットを構築し, タスクを策定することが挙げられる.

## 謝辞

本研究は JSPS 科研費 21H03769, および, 電気通信普及財団の助成を受けたものである.

## 参考文献

[1] 金融庁. Edinet 閲覧（提出）サイト. https://disclosure2.edinet-fsa.go.jp/WEEK0010.aspx. 最終アクセス日：2024 年 1 月 7 日.

[2] 木村泰知, 近藤隆史, 門脇一真, 加藤誠. Ufo: 有価証券報告書の表を対象とした情報抽出タスクの提案. 人工知能学会第二種研究会資料, Vol. 2022, No. FIN-029, pp. 32–38, 2022.

[3] 門脇一真, 木村泰知, 加藤誠, 近藤隆史, 乙武北斗. 有価証券報告書を対象とした表構造解析のためのデータセットの構築に向けて. 人工知能学会第二種研究会資料, Vol. 2023, No. FIN-030, pp. 100–105, 2023.