TEI を用いた『渋沢栄一伝記資料』テキストデータの再構築

金甫榮（渋沢栄一記念財団） 中村覚（東京大学）
小風尚樹（千葉大学） 橋本雄太（国立歴史民俗博物館）
井上さやか・茂原暢（渋沢栄一記念財団） 永崎研宣（人文情報学研究所）

本研究は，近現代日本語資料として『渋沢栄一伝記資料』のテキストデータの，TEI ガイドラインに準拠したエンコーディングを通じその構造化を図り，データの活用を促進させることを目的とする．本稿はその中間報告として，エンコーディング方針およびテキストデータの構造分析と再構築方法を示し，その過程から得られた成果について述べる．

Reconstructing Text Data of the Shibusawa Eiichi Denki Shiryo Using TEI

Boyoung Kim (Shibusawa Eiichi Memorial Foundation) Satoru Nakamura (The University of Tokyo)
Naoki Kokaze (Chiba University) Yuta Hashimoto (National Museum of Japanese History)
Sayaka Inoue / Toru Shigehara (Shibusawa Eiichi Memorial Foundation)
Kiyonori Nagasaki (International Institute for Digital Humanities)

This research targets modern Japanese text, Shibusawa Eiichi Denki Shiryo, in compliance with TEI (Text Encoded Initiatives) guidelines to structure and utilize the text data. As an interim report, this paper describes the encoding policy and how the text data was structurally analyzed and reconstructed as well as the impacts on the text data.

１．まえがき

近年 TEI を活用した研究事例は多く見られるようになったが，近現代日本語資料を対象にした事例は多くない．そこで本発表では，幕末から昭和初期（19 世紀後半から 20 世紀前半）を中心とした記録である『渋沢栄一伝記資料』（以下『伝記資料』）の事例研究について報告する．

『伝記資料』は，全 68 巻（本編 58 巻，別巻 10 巻），約 48, 000 ページにわたる膨大な資料集であり，本編 1～57 巻（58 巻は索引のため除外）は 2016 年からインターネット[1]を通じて閲覧できるようになった．一方，別巻の 10 冊（日記や書簡，談話，講演，遺墨，写真など）は，まだ公開には至っていない．

本研究[2]はこれら未公開の資料の中で，別巻第 1～2 に掲載されている日記および集会日時通知表（スケジュール表のようなもの）を対象にする．ここには 1868 年（慶応 4）から 1931 年（昭和 6）に至る渋沢栄一の日々の出来事が記されており，関連人物や組織，地域に関する膨大な情報が含まれている．これが，本資料が渋沢研究のみならず，日本の近現代史および経済史研究において重要な資料として評価されている所以であるが，未だそのテキストデータを利用した可視化や分析は行われたことがない．そこで，本研究ではそのテキストデータを TEI のガイドラインを踏まえて構築しつつ，テキストデータのあり方や公開，活用方法などを考察することを目的としている．

本稿はその中間報告であるため，TEI を用いたエンコーディング方針とテキストデータの再構築方法について述べ，その過程からもたらされた成果を踏まえながらテキストデータのあり方につい考察することとする．

２．エンコーディング方針

『伝記資料』のデジタル化は 2004 年から進められ[3]，現在は全 68 冊のテキストデータが TXT 形式で存在している．そのデータに TEI を適用する理由の一つは，構造化にある．その構造を考える上では次の 2 点を最優先し，エンコーディングの方針とした．

(1) 1 冊の書籍としての物理的構造（表紙，凡例，解題，目次，本文，奥付など）だけではなく，内容的構造（編・部・章・節など）をも表現すること
(2) 『伝記資料』に引用された原資料の来歴情報や出所といったコンテクスト情報を明確に表現すること

TEI が採用している XML スキーマでは，タグによる階層構造としては一つしか記述できないため，内容的構造を主とする場合は空タ

グを用いることで物理的構造を共存させることになる[4]. また, 物理的構造を主とする場合には, XML の属性値を用いることで内容的構造を表現することも可能である. 『伝記資料』においては, 物理的構造は資料の書誌情報として必要である一方, 内容的構造は利用者の資料に対する理解を容易にするためには欠かせない情報である. 日記等の一連の内容が書籍をまたがって記載されていることがあるため, 本研究では, 書籍の単位では物理的構造でマークアップし, 一つの書籍の中では内容的構造に基づいて記述した上で, 内容的構造とオーバーラップする書籍内部の物理的構造については空タグでマークアップ (これについては 4 章で詳述する) する一方, 書籍をまたぐ内容的構造については, XML の属性値として一連の ID を与えることで, これらの異なる構造を同時に表現した.

以下は詳細タグを省略したものであるが, type="volume"は書籍単位の物理的構造でマークアップしたことを意味し, xml:id="DKB1"は別巻第 1 を, xml:id="DKB2"は別巻第 2 を指す ID である. また, n="41000000000"および n="42000000000"は, 『伝記資料』の全体構成の編, 部, 章の中で部を指しており, それぞれ日記と集会日時通知表という内容的構造を示すものである. 下記の TEI マークアップでは, 日記が巻をまたいでいること, また, 別巻 2 の途中でその内容が日記から集会日時通知表に変わることを表現している.

```
<! ファイル 1 >
<group>
 <text type="volume">
    <text n="41000000000" xml:id="DKB1">
       <text xml:id="DKB100001m">…</text>
       <text xml:id="DKB100002m">…</text>
    </text>
 </text>
<group>


<! ファイル 2 >
<group>
 <text type="volume">
    <text n="41000000000" xml:id="DKB2">
       <text xml:id="DKB200001m">…</text>
       <text xml:id="DKB200002m">…</text>
    </text>
    <text n="42000000000" xml:id="DKB2">
       <text xml:id="DKB200003m">…</text>
       <text xml:id="DKB200004m">…</text>
    </text>
 </text>
</group>
```

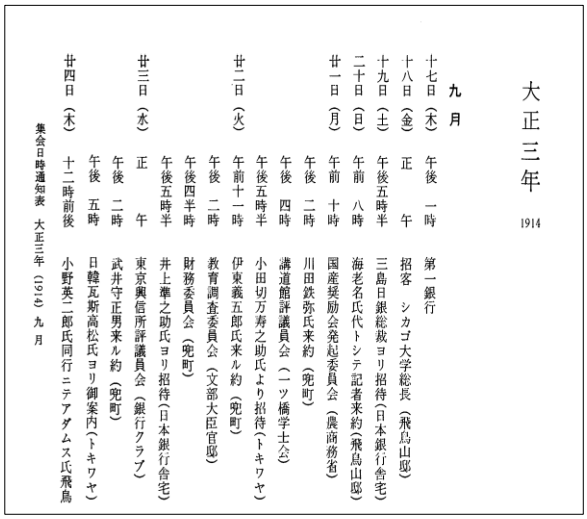
また, (2)を実現するためには, テキストのかたまりや階層, 原資料との関係性を示す必要がある. これは多様なタグを用いるより使うタグは最小限とし,「type」属性を活用することとした. 例えば, 日記は原資料の本文翻刻テキストだけではなく, その原資料のタイトルと資料の物理的特徴, 所蔵情報なども記載されている. そのため全体を(a)原資料のタイトル, (b)原資料の概要, (c)日記本文 (原資料の翻刻文) , (d)原資料を読解する為に付された『伝記資料』編纂時の補記の四つのかたまりで把握した. 以下が詳細タグを省略した大まかな構成である.

```
<text type="diary">
   <front>
      <head>原資料のタイトル</head>
   </front>
   <body>
      <div type="archival-description">原資
      料の概要</div>
      <div type="diary-entry">日記本文 (原資
      料の翻刻文) </div>
      <div type="note">編纂時の補記</div>
   </body>
</text>
```

また, 集会日時通知表 (図 1) は (e)タイトル (年) (f)月と日, (g)時間と予定, (h)備考の四つのかたまりに加え, 日記同様に(i)編纂時の補記で構成されている. 中で月と日は一つのかたまりとし<div1>, <div2>のようなタグを用いて階層構造で表現することも可能である. しかし, 機械可読性を高めるため タグだけを使い[5], そこに「type」属性を与えることで各 の内容と関連性に対する人間可

読性の向上を図ることとした．以下が詳細タグを省略した大まかな構成である．

```
<text type="schedule">
   <front>
      <head>タイトル（年）</head>
   </front>
   <body>
      <div type="month">月</div>
      <div type="day">
           <head>日</head>
           <listEvent>時間と予定</listEvent>
      </div>
      <div type="memo">備考欄及び欄外の記載文</div>
   </body>
</text>
```


大正三年　1914

九月

十七日（木）午後　一時　第一銀行
十八日（金）正　午　招客　シカゴ大学総長（飛鳥山邸）
十九日（土）午後五時半　三島日銀総裁ヨリ招待（日本銀行舎宅）
二十日（日）午前　八時　海老名氏代トシテ記者来約（飛鳥山邸）
廿一日（月）午前　十時　国産奨励会発起委員会（農商務省）
午後　二時　川田鉄弥氏来約（兜町）
午後　四時　講道館評議員会（一ツ橋学士会）
午後五時半　小田切万寿之助氏より招待（トキワヤ）
廿二日（火）午前十一時　伊東義五郎氏来ル約（兜町）
午後　二時　教育調査委員会（文部大臣官邸）
午後四半時　財務委員会（兜町）
午後五時半　井上準之助氏ヨリ招待（日本銀行舎宅）
廿三日（水）正　午　東京興信所評議員会（銀行クラブ）
午後　二時　武井守正男来ル約（兜町）
午後　五時　日韓瓦斯高松氏ヨリ御案内（トキワヤ）
廿四日（木）十二時前後　小野英二郎氏同行ニテアダムス氏飛鳥

集会日時通知表　大正三年（1914）九月

図１　集会日時通知表の例

さらに詳細なタグ付けを行う際には，テキストデータの活用方法を視野に入れる必要があろう．日記および集会日時通知表において，可視化と分析の手がかりになるのは，日時および人名，地名であると判断し，これら三つの要素のタグ付けに注力した．

３．自動処理

『伝記資料』のテキストデータは膨大であるため，自動処理が欠かせない．既存の TXT ファイルには，すでに統制記号（○，【】，▲など）や，ID，改行などを用いた区分けがされており，自動処理に必要なパターン化がある程度可能な状態にあったため，その特徴を利用し Python で専用のプログラムを組み，自動でタグ付けをした後人力で確認を行った．作業は次の三つに分けて行った．(1)テキストの構造を決めるマークアップ，(2)人名と地名のマークアップ，(3)日付のマークアップである．

(1)の作業では，特に典拠資料が変わる部分で見られる階層の不揃いを整える作業，また，各階層に含まれる本文範囲を定める作業は，自動では処理することが難しく手動で行った．この事例については 4 章で述べる．

(2)の人名と地名の処理は，オープンソース形態素解析エンジンである MeCab[6]と近代文語 UniDic[7]を組み合わせて使用しているが，判定ミスも含まれるため，最終的には目視で確認修正を行なう予定である．特に，人名にも地名にも判定し得る用語や，略称，通称については人による判断が必要となる．

(3)の日付・時間の処理は，次の 2 つのプロセスに分けて実施した．(a)『伝記資料』の構成に依存しない汎用的な方法でのマークアップ，(b)『伝記資料』の構成に依存したマークアップ情報の補完である．

(a)のプロセスでは，正規表現を用いて日付，時間に関する文字列を抽出し，<date>，<time>タグを用い，さらに「when」属性に「yyyy-mm-dd」「hh:mm:ss」の形式で正規化した値を持たせた．この時，「月日」または「日」しか表記されていない日付については，「XXXX-mm-dd」「XXXX-XX-dd」のように文字「X」を仮に与えた．

(b)のプロセスでは，『伝記資料』の構成に応じてマークアップ情報の補完を行った．例えば先の「年」が表記されていない日付について，その日付の表記が含まれる上位階層の構成から，多くの場合において「年」を推定することができた．また時間について，「午前」「午後」が表記されていないことが多いが，それらの文字列が登場する順序から，「午前」「午後」の情報を推定した．もちろん，これらの推定結果には誤りが含まれる可能性があるため，「evidence」属性に「inferenced」値を与えることで，推定結果であることを明示した．これらに対して人手によるチェックを加えることで，日付に関するマークアップの半自動化を実現した．

４．本研究の成果と意義

TEI エンコーディングを通じてプレーンテキストデータが抱えていた問題点を解決できたことは，本研究の大きい成果と言える．

既存の独自テータ形式が持つ一つ目の問題点は，書誌情報およびテキストデータに関する背景情報を保持できなかったことである．しかし，<teiHeader>に含まれる書誌情報やテキストが作成された状況，エンコーディング情報などを記述することで，そのデータがどういうものであるかを明確にし，データの信頼性を向上させることができた．以下は既存のテキスト形式では含まれていなかった情報である．

```
<teiHeader>
   <fileDesc>
      <titleStmt>テキストデータのタイトル，
      責任情報</titleStmt>
      <publicationStmt>発行者情報
      </publicationStmt>
      <sourceDesc>元となった資料情報
      </sourceDesc>
   </fileDesc>
   <encodingDesc>
      <projectDesc>エンコーディングの目的や
      経緯に関連する情報</projectDesc>
      <editorialDecl>エンコーディングの方針
      や方法</editorialDecl>
   </encodingDesc>
</teiHeader>
```

二つ目の問題点は，縦書きのテキスト表示スタイルが持つ意味が，横書きに変換する過程で失われ，テキストを構成する要素間の関係性を把握することが困難となり，段落や各要素を読む順番に変化が生じたことである．

図 2 は 1868 年（明治元）9 月 16 日と 17 日の日記で，図 3 はその TXT データである．TXT データ作成においては，本文を横にした際に見える形のまま再現する方法を取っている．これは内容の解釈が入る場合，人によって判断にばらつきが生じる可能性があるため，誰もが同じ結果を見出すためには合理的な方法であろう．しかし，割注を表現する過程では，TXT 形式が 1 行を複数に分けることが不可能なことから，行を分けるのではなくその逆の発想で全体の行数を増やし，その中にテキストを配置している．つまり，2〜3 行割注を表現するため 3 行を使用している．図 3 の 7 行目から 9 行目にあるのは 2 行割注の例であるが，全体行数および記載位置の変化により，7 行目がどの段落に含まれるものかが不明確な状態となってしまった．これは上の段落後に改行を入れることである程度は解消される可能性はあるが，根本的な解決方法とは言えない．

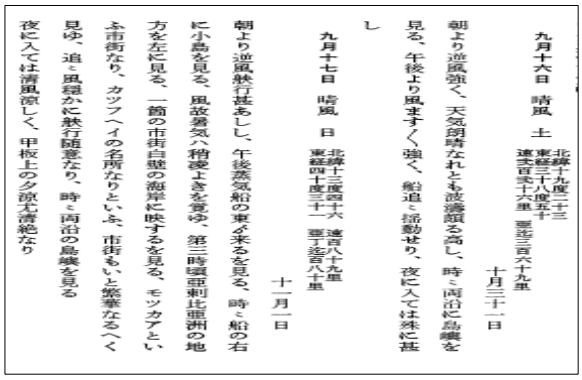

九月十六日　晴風　土　北緯十九度二十三 東経三十八度五十 速弐百弐十六里　亜迄三百六十九里
十月三十一日
朝より逆風強く、天気朗晴なれとも波濤頻る高し、時々両沿に島嶼を見る、午後より風ますます強く、船追々揺動せり、夜に入ては殊に甚し
九月十七日　晴風　日　北緯十三度四十六 東経四十度三十一　速百八十九里 亜丁迄百八十里
十一月一日
朝より逆風航行甚あしし、午後蒸気船の東より来るを見る、時々船の右に小島を見る、風故暑気ハ稍凌よきを覚ゆ、第三時頃亜剌比亜洲の地方を左に見る、一箇の市街白壁の海岸に映するを見る、モツカアといふ市街なり、カッフヘイの名所なりといふ、市街もいと繁華なるへく見ゆ、追々風穏かに航行随意なり、時々両沿の島嶼を見る
夜に入ては清風涼しく、甲板上の夕涼尤清絶なり

図 2　本文

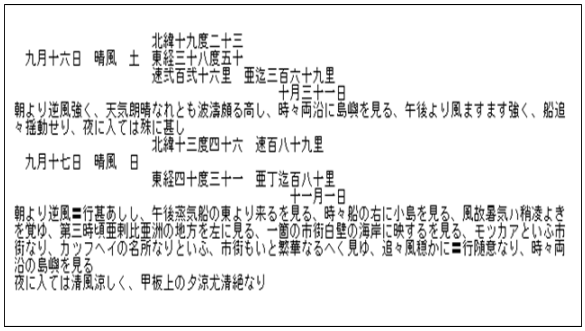

九月十六日　晴風　土　北緯十九度二十三 東経三十八度五十 速弐百弐十六里　亜迄三百六十九里
十月三十一日
朝より逆風強く、天気朗晴なれとも波濤頻る高し、時々両沿に島嶼を見る、午後より風ますます強く、船追々揺動せり、夜に入ては殊に甚し
九月十七日　晴風　日　北緯十三度四十六 東経四十度三十一　速百八十九里 亜丁迄百八十里
十一月一日
朝より逆風ニ行甚あしし、午後蒸気船の東より来るを見る、時々船の右に小島を見る、風故暑気ハ稍凌よきを覚ゆ、第三時頃亜剌比亜洲の地方を左に見る、一箇の市街白壁の海岸に映するを見る、モツカアといふ市街なり、カッフヘイの名所なりといふ、市街もいと繁華なるへく見ゆ、追々風穏かにニ行随意なり、時々両沿の島嶼を見る
夜に入ては清風涼しく、甲板上の夕涼尤清絶なり

図 3　TXT データ

また，読み順にも変化が生じた．図 4 の①は和暦（旧暦）の日時・天気・曜日，②〜③は緯度情報，④は移動速度，⑤は目的地までの距離，⑥は西暦の日付である．これらの実際の読み順は①→⑥の順であるが，TXT 形式では横書きの左上から右下へと読む順に従うため，②→④→①→③→⑤→⑥の順になってしまう．図 4 の例ではこれにより意味が変化するような重大な問題にはなっていないが，元のテキストが伝える意味が正確に表現されたとは言えない．

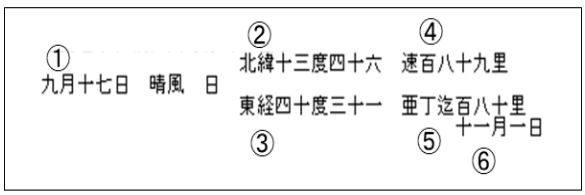


図 4 読み順

TEI エンコーディングでは，この二つの問題点に対して， タグで曖昧な段落の境界とその構成要素を明確にすることができたが，読む順の改善にまでは至っていない．この事例では，緯度と速度を示す北，東，速の言葉の出現順番を決めることで自動処理できる可能性があるが，これも根本的な解決策とは言えないだろう．この点については TEI のガイドラインにおける縦書きおよび割注に対する更なる検討が必要であると思われる[8].

三つ目の問題点は，複雑な階層とその関連性を表現できなかったことである．特に日記では，前述のように複数のかたまりで本文を分けることができるが，その中で (a)原資料のタイトル，(b)原資料の概要，(c)日記本文（原資料の翻刻文）の包含関係を明示することは，テキストデータに原資料のコンテクスト情報を与えることにつながる．しかし，既存テキストデータにこのような情報は含まれていなかった．

TEI エンコーディングでは 2 章で述べたように と type 属性を用いてかたまりおよびその関係性を明確にすることが可能となっただけではなく，『伝記資料』の編集過程で起きた配置の誤りを改善することも可能となった．その例を図 5 と図 6 に示す．

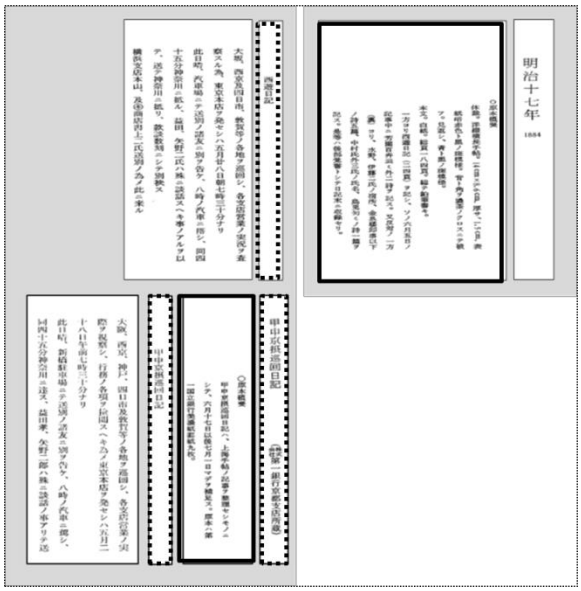


図 5 『伝記資料』の配置

図 5 は，1884 年（明治 17）の日記部分の構造を略図にしたものである．この年の日記は西遊日記と甲申京摂巡回日記の二つの日記帳が元となっており，本文構成も二つに分かれている．図の点線で囲まれている部分がそれぞれの日記のタイトルで，太い実線で囲まれている部分は原資料の概要である．しかし，最初（上段）に表出する概要は，1884 年（明治 17）全体（もしくは二つの日記帳）に該当するように読める．しかし，実際は西遊日記にだけ該当する内容であり，本来は西遊日記のタイトルの後に記載されるべき内容である．また，西遊日記のタイトルは一つだが，甲申京摂巡回日記のタイトルは二つ存在している．

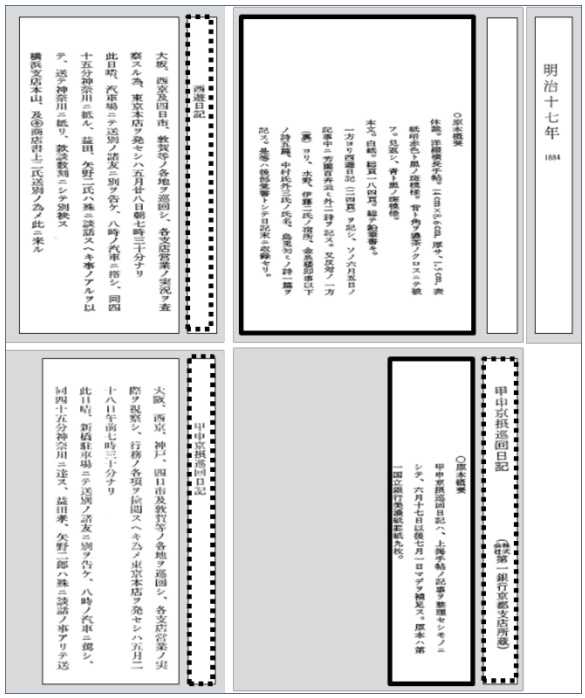


図 6 テキスト要素の再構成

TEI エンコーディングでは，この矛盾する構造を解決するため各要素を再構成した．図 6 のように空要素を補った上，階層を合わせると原資料の概要に含まれる範囲がより明確になる．以下はその結果を，細部タグを省略し大まかな構造だけで示したものである．

```
<group>
   <text>
      <front>
         <head/>
         <div type="archival-description">原
         資料の概要</div>
      </front>
```

```
      <body>
         <div>
            <head>西遊日記</head>
            <div type="diary-entry">日記本
            文</div>
         </div>
      </body>
   </text>
   <text>
      <front>
         <head>甲申京摂巡回日記…</head>
         <div type="archival-description">
         原資料の概要</div>
      </front>
      <body>
         <div>
            <head>甲申京摂巡回日記</head>
            <div type="diary-entry">日記本
            文</div>
         </div>
      </body>
   </text>
</group>
```

『伝記資料』は膨大な資料集であるにも関わらず，そのテキストの分析事例は未だ皆無である．それは利用者がテキストデータを一括で入手できないこと，そして，その量が膨大であるがゆえに内容を把握することが困難であることなどが理由として考えられる．しかし，本研究の成果により利用者はテキストデータを１冊丸ごと入手でき，その内容を構造的に理解できることが期待される．

このように既存テキストデータの問題点を抽出し，TEI を用いてその改善を図ったことは，テキストデータの信頼性および利用可能性の向上につながっている．これは，テキストデータの在り方を考える上で重要なプロセスであると考える．

５．あとがき

様々な資料のデジタル化が進んでいる中，その使いやすさのゆえデータの利活用の側面が強調される傾向にある．しかし，そのデジタルデータの信頼性が研究成果の信頼性にもつながることを忘れてはならないだろう．この点は，『伝記資料』のテキストデータのあり方を考える上で最も重要な軸となった考え方であり，エンコーディング方針にも大きく影響している．そして，テキストデータを分割しながらも，それらを再度つなぎ合わせることで，テキストデータが部分でありながら，全体としての情報をも維持できるように努めた理由でもある．

本研究の次の段階では，テキストデータの可視化や分析が要となる．例えば，タグ付けした人名と地名の分析や，日付によるタイムラインの表示など，多角的な研究アプローチを提示することが考えられるだろう．これについては今後の課題としたい．

※本稿は，2020 年度国立歴史民俗博物館総合資料学奨励研究「TEI を用いた『渋沢栄一伝記資料』テキストデータの再構築と活用」の成果によるものである．

参考文献

[1] デジタル版『渋沢栄一伝記資料』. https://eiichi.shibusawa.or.jp/denkishiryo/digital/main/, (参照 2020-11-09).

[2] 2020 年度国立歴史民俗博物館総合資料学奨励研究「TEI を用いた『渋沢栄一伝記資料』テキストデータの再構築と活用」

[3] 公益財団法人渋沢栄一記念財団編, “5 デジタル伝記資料（二〇〇四年～)”. 渋沢栄一記念財団の挑戦. 不二出版, 2015, pp. 91-100.

[4] “20 Non-hierarchical Structures”. P5: Guidelines for Electronic Text Encoding and Interchange. Version4.1.0. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html#NHVE, (参照 2020-11-09).

[5] Nancy Ide, C. Michael Sperberg-McQueen, Lou Burnard, 招待論文 TEI:それはどこからきたのか. そして，なぜ，今もなおここにあるのか？, デジタル・ヒューマニティーズ, 2018, Vol. 1, p. 3-28, https://doi.org/10.24576/jadh.1.0_3.

[6] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. https://taku910.github.io/mecab/, (参照 2020-11-09).

[7] 小木曽智信, 須永哲矢.『近代文語 UniDic』『中古和文 UniDic』を利用した総索引作成システムの開発. じんもんこん 2010 論文集, 2010, Vol. 2010, No. 15, p. 119-124.

[8] 王一凡, 永﨑研宣.東アジア文献への TEI の適用をめぐって. 研究報告人文科学とコンピュータ(CH). 2018, Vol. 2018-CH-118, No. 4, p. 1-4.