

リンク構造で関連性を表現する為のリンク要素間リンクの制御

大矢 一志
鶴見大学

oya-k@tsurumi-u.ac.jp

本稿では、人文資料の電子化でテキストデータと画像データの関連性をマークアップ言語で示す手法を TEI の提案を元に検討し、記述の線状性を生かしたハブデータの使用法を提案する。同時に、その際に必要となるリンク要素間リンクを再帰的に制御する手法について新たに属性 `a_handover` を含む提案を行う。これにより、従来表現しきれなかった情報をマークアップデータに記述することが可能となる。またリンク要素間リンクの制御で実現する複合データ単位は、複数のテキスト構造を容易に扱えるため、結果として単一テキストデータ構造中に含める内容の判定基準が生まれる可能性がある。

Management of links between link elements to represent correlation on link structures

OHYA Kazushi
Tsurumi University

oya-k@tsurumi-u.ac.jp

To represent correlation between marked-up text data and multiple image data, the marked-up data must expand the descriptive power beyond the restriction of tree data model. For such a method, management of links between link elements is a good candidate. To realize the management, new link features are required. Then, the objective of this paper is to propose the link features and show the usage. This would contribute to facilitation of digitizing humanities resources.

1 はじめに

人文資料の電子化において、電子ファクシミリ版とテキストデータ版の関連づけは重要である。本稿は、マークアップデータをハブデータとして使い、これらの関連づけを行うに必要な論点の整理と、その1つであるリンク要素間リンクの制御について提案を行う。はじめに、従来の手法のまとめと論点を示し、ハブデータとしてテキストと画像の関連性を示す手法を検討する。次に、ハブデータを使用する際の論点を整理し、課題の1つであるリンク要素間リンクの制御について、[大矢土屋 1999[7]]を拡張する。

2 人文資料の電子化

電子化された人文資料には、通称はないものの、いわゆる a) テキストデータ版、b) 電子ファクシミリ版、c) アプリケーション版がある。cは教育目的が多く、主に展示や学習、鑑賞を目的に作られてい

る。bは貴重資料、学術資料の電子版に多い。これは、いわゆるファクシミリ版で既に編集・整理法が培われ、また人文研究者もその利用に慣れていることがある。aは一番古い歴史ある手法ではあるが[12]、ファクシミリ版と比べれば、まだ実際の研究資料として上手く活用されていない。理由としては、編集・整理の段階で、利用者側に相当のコンピュータリテラシを求めること、テキストデータ版で主流のマークアップ言語を利用したデータの作成には、技術的な課題が残されていること[9]がある。更に、テキストデータ版と電子ファクシミリ版との関連づけに確立された手法がないことも理由にある。

2.1 従来の手法

電子ファクシミリ版の画像データは、a)(投射・撮)像領域、b) 解像度(粒度)を基準に作られている。a像領域は、多くの場合、資料全体を俯瞰1コマとするのが原則で、例えば書籍の場合、いわゆる

見開きの状態が1領域とされる¹。b 解像度は、必要に応じて複数種用意されることがある²。この何れの構成においても、画像データとテキストデータの従来の関連は、各画像データに対応する単位をテキストデータに用意する手法が採られてきた。例えば、テキストデータとして目録データのみを収録する場合、対応する資料の表紙や部分の画像が、それに関連付けられる。この派生として、目録データ1件に対して複数の解像度の画像をリスト化して対応付けるケースもある(e.g.[11])。また、テキストデータが資料のテキスト内容である場合、いわゆるページを像領域とした画像データが、当該ページにあるテキスト内容の冒頭単位と対応付けられる(e.g.[10])。

更に複雑な対応を求める場合³、現在は先述のいわゆるアプリケーション版として作られることが殆どで(e.g.[1])、テキストデータと画像データの関連づけを、アプリケーションから独立したテキスト内容と同列の(テキスト)データとして記述されることは少ない。人文資料の電子化の一つの志向として、このような複雑な対応関係もテキストデータとして記述するスタンスがある。例えば、TEIに代表される、マークアップ言語によるアノテーション記述の可能性を探るコミュニティーではこのスタンスが採用されている。

2.2 TEI P5の対応

TEI P5[3]では、テキストデータと画像データとの関連づけについて、4通りの記法が用意されている。

1. 任意の要素に属性 **fac**s の値として画像データを参照する⁴。

2. 要素 **facsimile** を主たるデータ内容として使用する。従って、当該 XML ファイルは画

¹見開きという単位は、資料の形態上の単位ではなく、編集上の単位である。電子版を作成する際、多くは、俯瞰1コマを像領域としてインタフェースとの対応を採りやすくし、資料の形態上の単位を無視することができた。これは、巻子体の電子化でも同様である。但し、後述するように、電子版の改裝を検討する際には、資料の形態上の単位も想定する必要がある。

²これを複数レイヤとして用意する、例えば Kodak PhotoCD の手法や、1枚の高品質(高圧縮)データを用意する、例えば VFZ(PixelLive)や JPEG2000 などの手法がある。電子版人文資料の公開方法も、データ形式に縛られる後者の手法から、Ajax を利用した前者の手法が今後は採用されることになるかもしれない。

³複数の画像、例えば、ページ全体と一文字部分をテキストデータの同一単位に関連付けるとき、a)異なる解像度が複数枚ある場合の他、b)同一像領域で部分領域を指定する場合があります、後者では部分領域のメタデータが必要となる。

⁴属性 **fac**s はグローバル属性。

像データの関連を示すもののみとなる。画像データの像領域、解像度をそれぞれ複数とることができる。

3. 要素 **facsimile** と資料のテキスト内容であるテキストデータが、ファイル内参照により関連付けられる⁵。

4. 資料のテキスト内容がメタデータ部分⁶を介して画像データと関連付けられる。

これらに、URL の機能を使い部分データ単位を参照する手法を併用することも可能である。

1の手法は、資料のテキスト内容データ中にある任意の要素に属性 **fac**s を使い画像データとの関連づけを行うものである。資料のテキスト内容を中心

```
<div type="chapter">
  <div type="section" facs="image1.eps">
    <p>...</p>
    <p facs="image2.eps">...</p>
  </div><!-- section -->
</div><!-- chapter -->
```

図 1: 属性 **fac**s による記述

として画像データとの関連づけを行う、直感的な手法である。但し、この手法では、資料のテキスト内容から複数の解像度への関連づけは困難である⁷。

2の手法は、TEI P5 で新たに導入されたもので⁸、画像データの関連のみをマークアップによる文書(木)構造で示すことができる。中心となる要素は **facsimile** である。

要素 **zone** により複数の像領域と解像度を持つことが可能である。これは画像データのメタデータを記述するスキームとして有用である例えば、アプリケーション版でこのメタデータを使用することで、資料の内容を示すテキストデータの独立性をより高めることができる。

画像データ G の各画像 $g_1, g_2, \dots, g_n$ が $g_1 \leq g_2 \leq \dots \leq g_n$ の様に並べられるとすると⁹、要素 **facsimile**

⁵メタデータを記述する要素 **teiHeader** とは独立して要素 **facsimile** を要素 **text** の外に使用することができる。これは、P5 から採用された大きなスキームの変更になっている。

⁶**teiHeader**。

⁷属性値 **IDREFS** と **ROLES** を用意することも可能ではある。但し、1) この意味を規定する手法が明らかでないこと、2) そもそもリスト型の属性値にはマークアップ言語の使い方として論議があること、がある。リンクによる関連性表示についての詳細は後述。

⁸P4 と P5 では TEI スキームが大きく変更されている。根要素も TEI.2 から TEI への変更された。

⁹正確には、画像データ G は像領域 A と解像度 S から構成され、 $a \in A, a = \{s_1, s_2, \dots, s_m \in S\}$ である。但し、 $a_n \subset a_{n+1}$ として、つまり便宜上各画像データの像領域は包含関係にあり、 $\neg \exists a(a = a_n \cap a_{n+1})$ とする。

```

<facsimile>
<surface>
<zone ulx="0" ulx="0" lrx="200" lry="300">
<desc>左ページ (低解像度)</desc>
<graphic url="image2Low.eps"/>
</zone>
<zone ulx="0" ulx="0" lrx="2000" lry="3000">
<desc>左ページ (高解像度)</desc>
<graphic url="image2High.eps"/>
</zone>
<zone ulx="200" ulx="0" lrx="400" lry="300">
<desc>右ページ (低解像度)</desc>
<graphic url="image1Low.eps"/>
</zone>
<zone ulx="2000" ulx="0" lrx="4000" lry="3000">
<desc>右ページ (高解像度)</desc>
<graphic url="image1High.eps"/>
</zone>
</surface>
</facsimile>
%</TEI>

```

図 2: 要素 facsimile による記述

の子要素 f は、記述の線状性より $f_1, f_2, \dots, f_i$ と並べられる。この時、 $g_x \leq g_{x+1}$ であれば対応するメタ記述も $f_n \leq f_{n+1}$ であり、この逆も真であることから、ここに単射の関連性が成立する。つまり、画像データを要素 facsimile の構造へ写し取することに成功している。

3 の手法は、テキストデータと画像データの構造を、同一ファイルに記述する手法である。これは、メタデータ記述の部分に、画像データの構造に関する部分が付加された、と見ることができる。

```

<facsimile>
<surface xml:id="shot1">
<graphic url="image1.eps"/>
<zone xml:id="page1" ulx="0" ulx="400" lrx="300"/>
</surface>
<surface xml:id="page2" ulx="0" ulx="200" lrx="300"/>
</surface>
<surface xml:id="shot2">
<zone xml:id="page3" ulx="0" ulx="200" lrx="400" lry="300">
<graphic url="image3.eps"/>
</zone>
<zone xml:id="page4" ulx="0" ulx="0" lrx="200" lry="300">
<graphic url="image2.eps"/>
</zone>
</surface>
</facsimile>
<text>
<body>
<div type="chapter" facs="#shot1">
<div type="section">
<p>...pb facs="#page3"/>...</p>
<p>...pb facs="#page4"/>...</p>
</div><!-- section -->
</div><!-- chapter -->
</body>
</text>
%</TEI>

```

図 3: 要素 facsimile と text による記述

要素 text の構造は、資料のテキスト内容の構造を反映したものとといえるので¹⁰、これは資料のテキスト内容を要素 text の構造へ写し取することに成功している¹¹。この手法により、画像データとテキストデータの構造を記述することができる。但し、これらの関連性は属性タイプ ID、IDREF によるリンク構造で示される。従来、リンク構造の記述はこ

れで十分とされていたが¹²、図 3 で示されているリンク構造が、処理上どのように扱われるのか、記述上からは分からない。一般に、現行マークアップ言語が記述・構成するリンク構造については、1) リンク経路の構成要素が ID-IDREF(S) だけであること¹³、2) 記述されたリンク経路と処理時に発生するリンク経路の論議が未整理である¹⁴、という課題がある。例えば、TEI では従来から、資料のテキスト内容中には外部依存の情報¹⁵は記述せず、メタデータを記述する要素 `teiHeader` 中にまとめるというスタンスを採っているが(先の 4 番目の手法、cf. 図 4)、現行のリンク素性では有向リンクを作るしかない。この構造で無向リンクが記述できたとす

```

<TEI>
<teiHeader>
<fileDesc>
<title><title></title></title>
<publication><publication></publication></publication>
<sourceDesc>
<list>
<item>
<sourceDesc>
<fileDesc>
<teiHeader>
<text>
<body>
<list>
<item>
<app xml:id="app1">
<rdg wit="#a #b">あ</rdg>
<wit>資料 a 資料 b</wit>
<rdg wit="#c #d">い</rdg>
<wit>資料 c 資料 d</wit>
</app>
<app xml:id="app2">
<rdg wit="#a #b #c">え</rdg>
<wit>資料 a 資料 b 資料 c</wit>
<rdg wit="#b">お</rdg>
<wit>資料 b</wit>
</app>
</list>
</body>
</text>
</TEI>

```

図 4: `teiHeader` を使った記述

ると、例えば図 5 の様に記述することができる¹⁶。リンク構造の記述は入り組むものの、各マークアップ部分の役割毎に内容を分けて記述することができる¹⁷。しかし、依然として関連性を表現するリンク構造のリンク経路には先述の(記述する仕組みが

¹² 正確には、HyTime[4] 以降の現行マークアップ言語の関連規定において、リンク構造がよく論議されてこなかった。RDF で十分、またはグラフを必要とするデータをマークアップ言語で記述する不要である、リンク制御は処理上の問題でマークアップ言語で記述する必要はない、とされてきたか。勿論、その背景には、リンクそのものの哲学的考察への忌避、defacto standard の慣性、実装の難しさがある。

¹³ すなわち、有向リンクが構成される。同一 ID によるリンク経路は認められていない (ID 名前規則違反による) ため無向リンクは構成できない。

¹⁴ 後述。

¹⁵ ファイルロケーションなど。

¹⁶ この例は現行 TEI には準拠していない。

¹⁷ このように、リンクの解決処理を含意する有向性が、実際の記述に制約を課すことがある。これは現行 XML が、記述的マークアップ言語 (descriptive markup languages) として未

¹⁰ overlap 等、木構造に合わない内容の記述方法という問題とは別に、少なくとも必要なテキスト内容は記述されているという意味。

¹¹ この証明は、先の画像データのほぼ同じ手管でできる。

```

<TEI>
<teiHeader>
<fileDesc>
<titleStm><title></title></titleStm>
<publicationStm><publisher></publisher></publicationStm>
<sourceDesc>
<listWit>
<witness xml:id="baseA"><ref target="URLa"/></witness>
<witness xml:id="baseB"><ref target="URLb"/></witness>
<witness xml:id="baseC"><ref target="URLc"/></witness>
<witness xml:id="baseD"><ref target="URLd"/></witness>
</listWit>
</sourceDesc>
</fileDesc>
</teiHeader>
<facsimile>
<surface xml:id="a" xml:base="#baseA">
<graphic url="partA"/>
<zone xml:id="A_1" ulx="200" ulx="100" lrx="400" lry="300"/>
</surface>
<surface xml:id="b" xml:base="#baseB">
<graphic url="partB"/>
<zone xml:id="B_1" ulx="20" ulx="10" lrx="40" lry="30"/>
</surface>
<surface xml:id="c" xml:base="#baseC">
<graphic url="partC"/>
<zone xml:id="C_1" ulx="300" ulx="200" lrx="500" lry="400"/>
</surface>
<surface xml:id="d" xml:base="#baseD">
<graphic url="partD"/>
<zone xml:id="D_1" ulx="30" ulx="20" lrx="50" lry="40"/>
</surface>
</facsimile>
<text>
<body>
<list>
<item>
<app xml:id="app1">
<rdg wit="#a #b"></rdg>
<wit></wit>
<rdg wit="#c #d"></rdg>
<wit></wit>
</app>
</item>
</list>
</body>
</text>
</TEI>

```

図 5: 3ヶ所からリンク構造を構成する記述

ないため、結果として書けないことによる) 曖昧さがある。

3 関連性表示の課題

3.1 リンクで示される関連性の課題

関連性をリンク構造で表現する時、先述したマークアップ言語で記述されるリンク構造が持つ曖昧さ¹⁸に加えて、リンクそのものが持つ抽象性¹⁹が問題となる。マークアップ言語で表現されるリンク構造は、もともと記述の線状性という制約を超えるものとして作られたもので、一般には線的でない情報を示している²⁰。しかし、これを実際に扱う(処理する)際には線的な状態への変換が必要になるが、従来はこれをマークアップ言語の記述として明記す

熟であることを示す一例であろう。マークアップ言語一般の論議はまだもっと必要である。

¹⁸1)ID-IDREF(S) によるリンク経路、2) 記述されたリンク経路と処理時に規定されるリンク経路。

¹⁹後述するように、正確には、リンクが未規定のまま使われていることと云うべき。勿論、それを補う素性は XLink, TEI, HyTime などいくつか提案されている [7]。

²⁰例えば、参照元 a から参照先 b への有向リンクを、順序に似たリンク構造として $a < b$ と表せば、ここに $a < a$ と $a < b \wedge b < c \Rightarrow a < c$ が成り立つ。また、 $a < b \wedge b < a \Rightarrow a = b$ とすれば、可逆なリンク(または双方向リンク)を参照先と参照元を同じく扱えるものとして表現できる。ここに方向性はない。

ことはなく実装依存としてきた。しかし、処理時にリンク構造が決定されたとしても、その全てを実装依存とするのではなく、記述されている(できる)ことは重要である²¹。例えば、これと似た現象として「(記述上の) 記述内容の継承」が現行規格にある²²。今回の例でいえば、資料のテキスト内容から画像データへの関連は、必ずしも 1 対 1 ではなく、ある文字部分のテキストは当該文字画像の他、見開き画像や、さらには章全体の画像へとも関連付けられる²³。この程度の関係は記述できるべきであろう。

3.2 示される関連対象の課題

現行リンクそのものが持つ課題の他、関係を構成する対象の分析とその記述方法の検討が十分ではない。例えば、資料として冊子体を想定し、その画像データを見開き単位で用意したとする。この時、1) ある一節または段落にあるテキスト内容のデータと画像データとの関係と、2) ある大領域たとえば章や上部にあるテキスト内容のデータと画像データとの関係を考えたとき、現行 TEI P5 では、これを記述することができない。

1 は、実画像データよりも関係するテキスト単位が小さいケースである。例えば当該テキスト対象が文字である時には、図 6 のように要素 `zone` を使い記述することができる²⁴。これが段落にあるテキス

```

<char id="a">
<gloss>かな江</gloss>
<desc>Unicode にはない文字 (括字)。大阪朝日新聞。</desc>
<graphic ref="#b"/>
</char>
<!-- ----- -->
<surface xml:id="b" xml:base="baseA">
<graphic url="partB.eps"/>
<zone xml:id="B_1" ulx="200" ulx="100" lrx="210" lry="110"/>
</surface>
<!-- ----- -->
<g ref="#a">え</g>

```

図 6: 部分領域の指定

ト内容を想定した場合、複数の実画像データの部分

²¹すると、処理依存(処理時決定)と実装依存(実装時決定)は別であるから、記述にはない処理時に決定される、いわゆる一時的リンクが想定できる。本稿は、この一時的リンクを Ajax で実装する試みで検討された、ハブデータの持ち方の論議で得られた成果も元になっている。

²²実際の記述はなくとも、そこに特定の記述があると想定して、処理を行っている。

²³卷子体から冊子体、その逆の改裝は、電子版であるからこそ、自由に行うべきと考えている。例えば、線的情報(e.g. ストーリー性、時間性)を多く持つ資料、例えば楽譜やマンガといった資料は、本来卷子体の方が自然である。現在、マークアップを利用した電子版上での改裝を試みている。

²⁴正確には図 6 も、属性 `ref` の部分で TEI を拡張している。例であるから如何様にも書けるが、できるだけ TEI に準拠した記述にしている。

領域を指定する必要があるが、現行 TEI ではこの記述ができない。例えば、図 7 のような記述が必要になる²⁵。

```
<div type="phrase">
  <list type="image">
    <item>
      <ptr target="image001.eps">
        <zone xml:id="B_1" ulx="10" uly="50" lrx="110" lry="300"/>
      <desc>1 ページ目</desc>
    </item>
    <item>
      <ptr target="image002.eps">
        <zone xml:id="B_1" ulx="150" uly="50" lrx="200" lry="300"/>
      <desc>2 ページ目</desc>
    </item>
  </list>
</div>
```

図 7: 複数像領域の指定例 1

2 は、実画像データよりも関係するテキスト単位が大きいケースである。例えば、当該テキスト対象が第 1 部全体である時には、図 8 のように複数の画像データを示す必要がある²⁶。

```
<div type="part">
  <list type="image">
    <item>
      <ptr target="image001.eps">
        <zone xml:id="B_1" ulx="10" uly="50" lrx="110" lry="300"/>
      <desc>1 ページ目</desc>
    </item>
    <item>
      <ptr target="image002.eps">
        <desc>2 ページ目</desc>
      </item>
    </item>
    <item>
      <ptr target="image003.eps">
        <desc>3 ページ目</desc>
      </item>
    </item>
    <item>
      <ptr target="image004.eps">
        <zone xml:id="B_1" ulx="10" uly="50" lrx="200" lry="300"/>
        <desc>4 ページ目</desc>
      </item>
    </item>
  </list>
</div>
```

図 8: 複数像領域の指定例 2

1 と 2 は、実画像データに対してテキスト構造の単位が小さい場合と大きい場合で、共に複数の像領域 (画像データ+部分指定) との関連づけが必要になっているが、現行 TEI ではこの記述ができない。本稿では、リンク構造でテキストデータと画像データの関連性を示す際に、1) ハブデータの併用と、2) リンク素性を拡張することにより、関連性の詳細な記述データを用意し、同時に処理時の解決を予測する記述が可能になる手法を提案したい。リンクについては、より広い考察が必要となるが、本稿ではリンク素性に関して [大矢土屋 1999[7]] の拡張を以下で提案する。関連対象の記述方法については、ハブデータとして仮想単位の導入を提案する。

²⁵ テキスト構造を資料の構造上に重ねたと解釈し、資料の構造上の当該 1 点に例えば画像データへのリンクを記述する手法が、TEI はかで採られている。TEI ではこれをマイルストーン型 (model.milestoneLike) とポインタ型 (model.ptrLike) の、いわゆる空タグで記述する。図 7 は TEI 準拠ではない。

²⁶ この例も TEI 準拠ではない。

4 ハブデータによる関連付け

これまでに、資料のテキスト内容と画像データは全単射の関係を作らないことから、テキストデータ中で属性 `facs` を使うことでは関連性が記述できないこと、現行 TEI で提案されている手法で資料のテキスト内容の構造と画像データの構造はそれぞれ独立にテキスト構造に写し取ることができること、但しリンク構造には曖昧性が伴い、複数像領域の記述法が検討されていないことを確認した。本稿では手法のひとつとして、記述の線状性をできるだけ利用したハブデータ skeleton を使い関連性の記述を試みる。

4.1 Skeleton

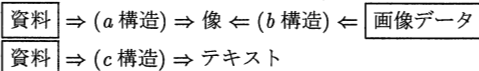
skeleton とは、資料のテキスト内容を削除したテキスト構造と、画像データ構造とを同じテキスト構造中に記述したものである。例えば、図 9 のようになる。skeleton 上には資料のテキスト内容の構

```
<text>
  <body>
    <pb corresp="#1_1">
      <list type="image" xml:id="#1_1">
        <item><ptr type="low" target="image001.eps"/></item>
        <item><ptr type="high" target="image001.eps"/></item>
      </list>
      <p shape="point" coords="200,60">←注意 →</p>
    </body>
    <seg corresp="#1_1.1">
      <list type="image" xml:id="#1_1.1">
        <item><area shape="rect" type="low" coords="200,100,190,110"/></item>
        <item><area shape="rect" type="high" coords="400,200,380,220"/></item>
      </list>
      <seg corresp="#1_1.2">
        <list type="image" xml:id="#1_1.2">
          <item><area shape="rect" type="low" coords="190,100,196,110"/></item>
          <item><area shape="rect" type="high" coords="380,200,392,220"/></item>
        </list>
      </seg>
      <note corresp="#1_1.3">
        <list type="image" xml:id="#1_1.3">
          <item><area shape="rect" type="low" coords="196,100,200,110"/></item>
          <item><area shape="rect" type="high" coords="392,200,400,220"/></item>
        </list>
      </note>
    </seg>
    ... <pb/> ... <lb/>
    <pb corresp="#1_2">
      <div type="page" xml:id="#11_2">
        </div>
      <lb/> ... </p>
    </body>
  </text>
```

図 9: skeleton の例

造が全て記述され、同時に画像構造 (像領域、解像度) が要素 `list` により写し取られている。TEI のアプローチと同様に skeleton でもテキスト構造から実存物への関連は空タグで記述される。但し、要素 `list` を資料のテキスト構造中に埋め込むことで、未定義のまま記述されるリンク経路に依らず、テキスト構造により明示化している。例えば、図 5 や図 6 では、リンクを複数わたる (traverse) ことを前提に記述の関係全体が表現されているが、これを図 9

では、記述に於ける属性の継承 (@target)²⁷として記述している。但し、この skeleton においても、いちテキスト単位が複数の像領域に対応する関連性は記述することは難しい。例えば、図 9 では段落を要素 p で記述しているが、ここに複数の画像データとその領域指定記述を併記することは、記述の継承を利用して難しい²⁸。skeleton 上に写し取られるテキスト構造 (c) と画像構造 (a,b) が、以下のような位置づけにあるとすれば、



複数像とテキスト単位との関連を記述する手法が新たに求められるといえる。これを一般には実装依存とするケースが多く、マークアップ言語で記述する例を聞かない。本稿では、もう少しマークアップ言語による記述に固執する。リンク要素をハブデータを介することで仮想単位を形成 (記述) し、関連性を示すアプローチを採用。

4.2 複合データ単位

複合データ単位とは、いわゆるハブ (テキスト) データの一種で、「構成要素としてリンク要素を持ち、全体でリンクとして振る舞うテキスト単位」のことである [7]。例えば、図 10 の要素 `div@xml:id=img1` がこれにあたる。当該要素 `div` は、内部構造を持

```
<a href="img1">見開き</a>
<!-- ----->
<div xml:id="img1">
  <item target="page1.eps"/>
  <item target="page2"/>
</div>
<!-- ----->
<div xml:id="page2">
  <item target="page2.eps"/>
  <item target="page2_supp.eps" loc="top">
    <desc>張り出し部分、伊豆半島</desc>
  </item>
</div>
```

図 10: 複合データ単位例

つテキスト単位 (要素) で、構成要素としてあるリンク要素では、当該単位を構成する実データが参照されている。このような単位は、独立した単位として扱われる他、時には所在情報を示す単なる仲介データとして機能することが期待される。例えば要素 `a` が選択されると、結果として、3 つの `eps` ファイルが得られることになる。これを実現するには、

²⁷ 正確には、図 9 において要素 `item` は属性 `target` の値をテキスト構造上は継承していない。記述の線状性を前提とした継承になっている。

²⁸ 1) 記述の方向に逆行する記述の省略はあり得ず、2) 同じ定義の要素が同じ構造レベルに複数有り、あるものは属性の記述を継承し、あるものは属性の記述が明記されているという記述は、未定義属性の判断が付かないため、処理を困難にする。

リンク要素間にあるリンク要素を制御するリンク素性が必要となる。

5 リンク要素間リンクの制御

[大矢土屋 1999][7] では、リンク要素がリンク経路を構成し、そのリンク経路上にあるリンク要素、すなわちリンク要素間リンクに対して、再帰的にリンク素性を規定し、当該リンク要素の振る舞いを制御することを提案した。これにより、HyTime でのようにわたり数²⁹を事前に知ることなく、リンク経路上にある全リンク要素の動作が制御できるようになる。[7] では、リンクを以下のように捉えた³⁰。

```
Link(Reference, Referents);
Referents = Referent++;
Reference = {ReferTo+, LinkDef};
ReferTo(Locator, Referent);
LinkDef = {Actant, Process};
Locator = LocatorName, LocatorValue;
Activate(Actant, Resolution(ReferTo));
Resolution(ReferTo) = Bind(Process, ReferTo);
```

リンクには動的経路と静的経路があり、それを `Link` は静的経路³¹、`Activate` は動的経路として示している。これにより、3.1 節にあるリンク構造が持つ曖昧さが解消される。再帰的にリンク要素の挙動を制御するため、1) リンク素性 `a_cue` と `a_timing`³²を導入し、2) これらリンク素性はリンク経路上と記述の構造上で継承されることを前提にすることを提案した。素性 `a_cue` は、ロケーションを解決する `Activate` に関するもので、素性 `a_timing` が有効であるかどうかを、値 `local` と `stop` で示すものとした。素性 `a_timing` は、ロケーション解決のトリガー (`cue`) となる主体 `Actant` に関するもので、解決を依頼するものがアプリケーションがユーザかを値 `auto` と `user` で示すものとした。詳細は [7] を参照して頂きたい。これらにより、3.1 節にある抽象性による課題に対処できる。また、複合データ単位が図 11 の様を実現可能となる。

5.1 リンク定義の拡張

本稿で確認された要件で [7] にはないリンクに関する定義を以下に提案する。

1. リンク素性を持つものをリンク要素とする。

²⁹ `traverse` 数、リンクパスの数。

³⁰ 表記方法に問題があることは当時も今も承知しているが、当時のものを載せる。リンクそのものの規定方法はまだよく分からない。

³¹ 従って、`Link` は `Reference` と `Referents` の関係のみの構造になる。

³² 以前は `activate.cue` と `activate.timing` としたが、本稿では短く表記する。

```

<a href="#ing1">見開き</a>
<!-- ----->
<div xml:id="ing1" a_cue="local" a_timing="auto">
  <item target="page1.eps"/>
  <item target="page2"/>
</div>
<!-- ----->
<div xml:id="page2" a_cue="local" a_timing="auto">
  <item target="page2.eps"/>
  <item target="page2_supp.eps" loc="top">
    <desc>張り出し部分、伊豆半島</desc>
  </item>
</div>

```

```

<link id="here" cue="local" timing="auto">
  <link id="a" href="X"/>
  <link id="b" href="Y" cue="stop"/>
  <link id="c" href="Z"/>
</link>

```

図 11: リンク制御の例

現行では、参照解決処理の意味は、要素に付与されている。しかし、リンク素性を受け取る被参照単位は、必ずしも既定リンク要素とは限らない。とりわけ複合データ単位の利用を前提とする際には、属性に解決処理の意味を付与させる方が望ましい。従って、リンク素性をもつ属性はグローバル属性となる。

2. リンク素性はリンク経路上と記述の構造上で継承される。

本来、記述の構造上と記述の順序上の継承は異なるものである。例えば、SGML/XML における宣言の先取りルールは記述の順序上での制約で構造上の制約ではなく、属性の記述の省略はその逆である。この 2 つは規格上分けて規定されるべきで、とりわけ記述の再利用を容易にする際には重要となる [8]。本稿では、記述の構造上でリンク素性の継承を想定し、これとリンク経路上での継承を同じく扱いたい³³。

3. リンク素性の継承方向は、経路上は Reference から Referent、構造上は当該要素からその下位要素の 1 方向とする。

有向・無向リンクの定義と記述法について考察が十分ではないが、本稿の時点ではこの制約を採りたい。これにより、実装時に複合データ単位の性格付けが容易になる。

4. 属性 `a_handover` を導入する。

記述構造上、全対象にリンク素性を継承するのでなく特定要素に限定したい場合に、当該要素を指定する属性としたい。この素性は無くとも複合データ単位を使用すれば同等の記述は可能である。例えば、

```

<link id="here" handover="a c" cue="local" timing="auto">
  <link id="a" href="X"/>
  <link id="b" href="Y"/>
  <link id="c" href="Z"/>
</link>

```

の代わりに、

と記述することができる。但し、後述する処理時に使用されるリンク情報の記述を想定すると、リンク素性の継承のタイミングを指定する時にこの属性が有効に働くと考えられる³⁴。

5. 属性 `a_timing` の値には任意の `actant` が指定できる。

ユーザとブラウザ以外の `actant` を想定したい。

5.2 効果

複合データ単位と上記リンク素性により、単一テキスト構造下では表現が困難であった情報を、ハブデータである仮想単位 (複合データ単位) として記述することが可能になる。これにより、従来はマークアップデータとして記録されることがなかった (関連性に関する) 情報を記述することが可能となり、資料を電子化する入力・編集の段階で記録できる情報が増えることになる。また、複数のテキスト構造を容易に扱えるため、単一テキスト構造下に多くの情報を詰め込む必要はなくなり、結果として、単一テキストデータ構造中に含める内容の判定基準が生まれる可能性がある³⁵。

6 課題

マークアップ言語を使用した資料の電子化では、規格の論議とは別に必要となるマークアップ言語そのものの考察と同時に、対象の捉え方を検討すること³⁶が求められる。従って検討すべき項目は多く、本稿も論議の基盤部分に課題を多く残している。

1. 資料構造の参照モデル。

テキストデータと画像データとの関連性が自在に採れるようになると、電子版上で装丁を元資料のそれとは異なるものに、いわば電子版改装が行えるようになる³⁷。この際、当該資料を構成する

³⁴例えば、XInclude との併用時。

³⁵例えば、多言語内容を同一テキスト構造中に含めない、など。

³⁶アポストオリなデータ構造、すなわちデータ構造が自由に変わること前提とした研究手法を計算機科学上で確立しない限り人文情報学は成立しないのではない。

³⁷電子版資料であるからできる、紙媒体資料とは異なる新しい表現方法や実存形を検討する以前に、従来行われてきた紙媒体での改装を電子上で選択的に行うことの方が、資料が持つ内容の価値の判断に貢献するのではないだろうか。例えば、本書収蔵の江戸時代の版本『御馬印』(全 6 巻、うち零葉 37 葉) は、紙面には巻数と丁数が印刷されているが、本書以外の伝本は全て巻子体である。電子版にて当初の形態である冊子体へ改装する試みや、それと現存する巻子体とを鑑賞することで、技術の発展史は必ずしも利用者にとっての最善解を生む方向にはない

³³現時点では、脚注 20 にあるアイディアに留まり、まだ定義付けができていない。

単位 (e.g. 葉, 背や綴じ穴など) の参照モデルがあると便利である³⁸. 例えば, その試みとして [2] がある.

2. リンクの規定.

リンクは処理と強く結びついている仕組みであるが, 実装を主眼とした研究においては, リンクの定義を論議する以前に必要な処理が実現するため, リンクを記述する必要性は感じられない. 一方, 本稿のように, データ作成を主眼とする研究においては, できるだけ人可読で記しておきたい³⁹. 複合データ単位はハブデータの可用性を高め, 従来の資料のテキスト内容を超えた情報がマークアップデータに記録される可能性がある. これは, これは知識処理のコーパスを充実させることに寄与するだろう.

3. リンク制御の手法.

再帰的にリンクを活性化させるリンク素性は, 現在では Ajax により現実化されたリンクによる部分単位の交換を, より容易に実現させる.

4. 処理時決定の仕組み.

本稿で挙げたリンク素性は, 必ずしも資料にある内容や構造を写し取るデータだけに記述されることを前提としない. 処理時に決定されるための記述として, またはリンクの振る舞いを決める処理を定義する記法上で使われることも想定している. 例えば, リンクを辿ることで各種処理が実現する様子を, リンク構造を示すマークアップ記述で表現できるのであれば, 複合データ単位の利用範囲は広がるだろう.

5. マークアップ言語のモデル論.

実存する資料や資料のテキスト内容以外の対象を, メタ記述で示される対象 (モデルの単位) とすると, マークアップ言語のモデル論が十分に検討されていないため, その記述方法に関する論議が頓挫する⁴⁰. 同時に, 記述レベルでも検討が十分とはいえない⁴¹.

ことを再認識できるかもしれない. 電子版上の改装で最良解が容易に得られるのであれば, 資料の電子化に新たな価値が加わることになる.

³⁸これは書誌学で用意されている用語 [5][6] を使うことになるだろうが, 電子編集上の視点から必要になる単位と, 書誌学上重要となる単位は異なる. 参照モデルとメタデータと分けて設定することに難しさがある.

³⁹近年は, スクリプト言語のコードを添付したマークアップデータも増えてきた.

⁴⁰XDM の整備により Infoset からは幾分かモデル論の論議がしやすくなったが, それでも十分ではないと感じている.

⁴¹例えば, 実記述, 記述の補完, スキーム, 実記述 (データ) の構造, 記述単位のモデルの全てに見通しの良い定義がない [9].

7 さいごに

本稿では, 人文資料の電子化で有効な手段となる複合データ単位を実現するために, リンク要素間リンクの制御について [大矢土屋 1999[7]] を拡張した. これにより, データ記述が容易になるだけでなく, 複数テキスト構造の扱いや, 部分データ単位の交換を容易にさせる可能性がある. 今後, 上記課題に取り組むと共に, 本稿にあるリンク素性と一時的リンクが実装上で共存可能であるか確認する予定でいる.

参考文献

- [1] エービーピーカンパニー, 2004, 『江戸明治東京重ね地図』, エービーピーカンパニー
- [2] S.Bauman and T.Catapano, 1999, “TEI and Encoding of the Physical Structure of Books”, *Computers and the Humanities*, vol.33, Kulver Academic Publishers
- [3] L.Burnard and S.Bauman eds., 2007, *The TEI Guidelines P5*, TEI
- [4] ISO, 1992, *ISO 10744 Hypermedia/Time-based Structuring Language (HyTime)*, ISO
- [5] 川瀬一馬, 1982, 『日本書誌学用語辞典』, 雄松堂出版
- [6] 長澤規矩也, 1979, 『図書学辞典』, 三省堂
- [7] 大矢一志 土屋俊, 1999, 「複合データ単位に於けるリンク要素間リンクについて」『情報処理学会研究会報告』 Vol.99, No.48, 情報処理学会
- [8] K.Ohya, 2001, “Necessities on a Descriptive Level for Reusing Metadata Descriptions”, *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*
- [9] 大矢一志, 2006, 「マークアップの課題を syntax から見た分類と解決のステップ」, 『TEI Day in Kyoto 2006 報告書』, 京都大学
- [10] E.Solopova ed., 2000, *The General Prologue on CD-ROM*, Cambridge University Press
- [11] 高崎直道ほか監修, 2002, 『聖語蔵経巻』, 丸善
- [12] T.N.Winter, 1999, “Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance”, *Classics and Religious Studies*, University of Nebraska-Lincoln