データベースシステム 117－7
（1999. 1. 22）

XML リンク機能による異種文書の統合方式

國島丈生　横田一正　白木善隆　劉渤江
岡山県立大学情報工学部　岡山理科大学総合情報学部

現在コンピュータ上で扱うことのできる文書には、さまざまなフォーマットが存在する。たとえば、ある研究発表に関する文書のように、意味的には関連しているがコンピュータ上では異なるフォーマットの文書として扱われるようなものや、同一の文書に関して変種が存在するようなものである。本稿では、このような意味的に関連する異種文書を統合的に管理するためのシステムの要件について論じ、XML のリンク機能を基礎とするシステム構成について提案する。

A Method for Integrating and Managing Heterogeneous Documents using XML Link Functions

Takeo Kunishima　Kazumasa Yokota　Yoshitaka Shiraki
Faculty of Computer Science and System Engineering
Okayama Prefectural University

Bojiang Liu
Faculty of Informatics
Okayama University of Science

We use many electronic documents with several document formats on computers. For example, documents which have different formats but relate each other from their meaning point of view. Another example, as we see in comparative literature, the same contents are represented by many versions of documents. We call these documents as **heterogeneous documents**. In this article, we discuss about requirements of the system for integrating and managing heterogeneous documents, and propose its architecture using XML link functions.

1 はじめに

現在コンピュータ上で扱うことのできる文書には、さまざまなフォーマットが存在する。

たとえば、ある研究発表に関する文書をコンピュータで作成する場合を例として考えると、一連の作業の中で、予稿の LaTeX ファイル、そこから自動生成される dvi ファイルや PostScript ファイル、講演に用いる PowerPoint ファイル、予稿を HTML 形式や PDF 形式に変換したファイルなど、互いに関連を持つ電子文書が多数作り出される。これらは意味的には関連しているが、コンピュータ上では、異なるフォーマットの文書として扱われ、それを閲覧・編集するアプリケーションも異なるのが普通である。

また、同一の文書に関して変種が存在する場合がある。たとえば、著者らが進めている文学データベースの研究 [8,10] では、同一の物語であっても書かれた時代や著者が異なると表現が異なり、その差異を調べることが比較文学研究の重要な研究要素である。

本稿では、このようにファイル形式の異なる電子文書、あるいは同一の文書の変種を異種文書と呼ぶこととし、意味的に関連する異種文書を統合的に管理する方法について議論する。

2 異種文書の統合管理

2.1 統合管理への要求

異種文書の統合管理について、われわれは次のような要求があると考えている。

2.1.1 リンクによる関連付け

通常、関連する異種文書は何らかの形でまとめて管理することが普通である。たとえば、前述の研究発表の例であれば、一連の作業によって生成されたすべての文書を同一のディレクトリに格納するなどして分類しておくのが普通であろう。しかし、これらの文書はフォーマットも閲覧・編集アプリケーションも異なることがほとんどであり、名前による関連付け、もしくは (WWW で行われているような) ファイル単位でのリンクによる関連付けが行える程度であった。関連付けの粒度をさらに細かくしたり、3 個以上のアンカー (文書、文書要素) 間でのリンクを張ることができれば、たとえば、次のような関連付けが (計算機に解釈可能な形で) 行えるようになる。

- 予稿文書、OHP 文書、PDF フォーマットによる予稿文書など、複数の関連文書の関連。
- 予稿 LaTeX 文書の第 2 章と OHP 文書の 3〜5 ページと予稿 PDF 文書の 4〜7 ページが対応する。

あるいは、紙の文書ではユーザがメモを書き込むということがあるが、電子文書ではメモ文書を作成し、元文書とリンクを張るということで、メモの書き込みに近い機能を提供することができる [10]。この場合、電子文書からメモへ、またメモから電子文書へ相互にリンクの巡航を行えるほうが望ましい。

2.1.2 文書群

関連する文書を文書群に構成し、まとめて扱いたいことがある。たとえば、変種であるような文書の集まりをグルーピングして検索などの操作対象にするなどの場合である。文書群の考え方を用いれば、たとえば、文書群同士でリンクを張るという関連付けを行う、文書群単位の検索を行うというようなことが可能となる。

一つの文書は複数の文書群に含まれ得る。たとえば、ある OHP 文書を複数の講演で使用することはよくある。このとき、それぞれの講演に関係する文書を文書群にすると、OHP 文書は双方の文書群に含まれる。また、文書の一部を共有するような場合も考えられる。今の例で言えば、OHP 文書の一部を複数の講演で使用するような場合に相当する。

2.1.3 ビュー

文書間の関連付けを表すリンクや文書群は、常に同じではない。たとえば、メモ付けの場合、

メモは基本的に私的なものであるから、ユーザごとにメモを管理し、必要に応じて共有を行うのが望ましい。これは、リンクや文書群に対して各ユーザごとなどにビューを与えることに相当する。

2.1.4 文書(群)に対する検索

リンクや文書群で関連づけられた異種文書に対してキーワードなどで検索を行うとき、個々の文書を単位とするだけでは不充分なことがある。たとえば、「本文とメモのどこかに2つのキーワード A, B が含まれている文書群」を検索したい場合、本文に A、メモに B が含まれていても該当するが、これは、従来 WWW で行われているようなページを単位とする全文検索だけでは不充分で、リンクや文書群といった意味的情報を考慮する必要がある。

2.1.5 文書の更新に伴う一貫性保持

文書が更新されたとき、リンク情報などにも影響が及ぶことがある。この場合の「影響」には、直接的にリンクのアンカーが変更されることによる影響の他に、意味的関連が変更されることによる影響もあり得る。このような、文書の更新に伴う一貫性保持支援(半自動の可能性も考えられる)が必要である。

2.2 異種文書統合管理基盤としてのXML

現状の文書フォーマットの大半は、先に述べた統合管理を行うのに必要なリンク機能を備えていない。そのため、統合管理を実現するためには、なんらかの技術的基盤が必要になる。

われわれは、この技術的基盤として、W3C (World Wide Web Consorsium) が提案する構造化文書のためのデータ記述言語 XML (eXtensible Markup Language) [2,7] を用いて研究を進めている。XML を用いている理由は、以下のようなものである。

文書スキーマの定義 文書に対する検索機能が必要になると考えられ、文書の意味的構造が明確になっていることが重要である。言い換えれば、文書スキーマが明確に定義できることが必要である。XML では、DTD (Document Type Definition) によって独自の文書型定義を行うことができる。

リンク機能 XML では、XLink [5], XPointer [6] という2つの規格によって、強力なリンク機能を提供している。具体的には、文書中の任意の粒度のオブジェクトをアンカーとできる、3つ以上のアンカーに対してリンクを張ることができる、文脈外リンク (out-of-line link)[1]によって元の文書に手を加えずにリンクを張ることができる、などの機能を有用と考えている。

インターネットでの技術的優位性 HTML と同様、WWW での利用を考慮して作られた規格であり、Netscape Navigator や Internet Explorer などの主要な WWW ブラウザでの対応が期待できる。また、オフィスアプリケーションでも将来の XML 対応を公表しているものもあり、文書フォーマットとしてデファクト・スタンダードとなりうることが期待できる。

レイアウト情報の定義 見栄えにこだわった WWW ページが氾濫している現状から分かるように、電子文書といえども、それを扱うのが人間である以上、レイアウト情報を保持する機能は必要になる。XML では、CSS (Cascading Style Sheet) [1,4], XSL [3] などのレイアウト定義言語を持ち、レイアウト情報を定義することができる。

2.3 XMLに基づく異種文書の統合管理システム

異種文書を統合的に管理するために、XML 形式の文書以外は XML 形式に(半自動で)変換して管理する。この変換は可逆的であるが、必要に応じて他の形式にも(半自動で)変換す

[1]XML に関する技術用語の日本語訳は [7] によった。

る。これは現状ではかなり強い方針であるが、先に述べたように近い将来 XML に対応した文書アプリケーションが現れると予想されること、構造を持つ文書 (SGML, HTML, 限定された形式の LaTeX など) と XML との相互変換を行うソフトウェアが出始めていること、などの理由から、いずれ現実的な仮定になると考えている。相互変換が容易でない異種文書の扱いについては、4 節で議論する。

XML のリンク機能を用いることにより、従来の WWW よりも強力な文書間の関連付けを行うことができる。しかし、異種文書の統合管理という面から言えば、これだけでは不充分なところがある。たとえば次のような点である。

XML による異種文書ラッパー さまざまな異種文書を XML の枠組みで統合する。詳細は 4 章で述べる。

文書群のビュー 3.1 節で示したリンク定義文書を複数用いると、ある文書に関連する文書群を複数定義したり、一つの文書群に対して複数のリンク定義を与えることが可能となる。これを用いて、文書群に関する「ビュー機能」を提供することが可能になる。これは、たとえば、メモのようにユーザごとに異なる文書群を定義するという場合に有効である。

リンク情報の一貫性管理 3.2 節で示した拡張ポインタの変更を伝播する機構を実現するには、ECA 機構などを用いて、イベント (元文書にどのような変更が加えられたか) とそれに対する処理 (どのファイルにどのような変更を加えるか) を定義し実行するシステムが必要になる。

文書群に対する検索機能 通常の文書検索は個々の文書を検索単位として行われるが、2.1.4 節で述べたように、関連する文書群に含まれる文書に対して検索を行ったり、文書群を検索単位として検索を行う、というような要求がある。

われわれは、これらの問題を解決するため、図 1 のような統合管理システムを構想し、研究を進めている。

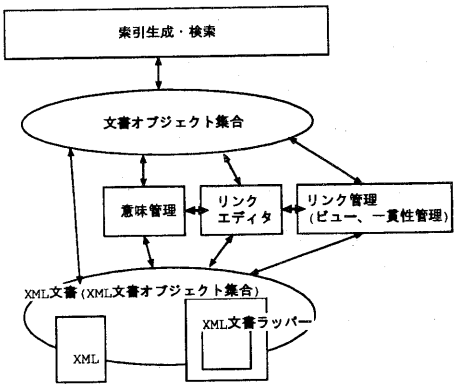

図 1: 統合管理システムの概要

3 章で述べるように、XML における文書要素はかなりの部分で文書管理に関するわれわれの要求を満たすが、完全ではない。したがって、XML 文書要素を基にして意味管理やリンク管理、それを支援するリンクエディタなどのシステムを構築し、結果として構成された文書オブジェクトに対して索引生成や検索などを行う。

また、すべての異種文書が XML と可逆変換が行えるわけではないので、その支援のために、XML を基にした文書ラッパーを実装し、それによって XML と可逆でない文書フォーマットも、この文書管理の枠組みの中に取り入れる。

3 拡張リンクによる文書間の関連付け

関連する異種文書群を定義するために、XML の拡張リンク (extended link) を用いる。また、各文書内で互いに関連する文書要素も、拡張リンクを用いて関連付けを行う。拡張リンクを用いるのは、3 つ以上の異種文書群や文書要素が関連することがあるためである。

図 2 に、拡張リンクを用いた関連付けの例を示す。図中、`<docgroup>`タグが関連異種文書群の定義を、`<related>`が関連する文書要素の定義をそれぞれ表す。

```
<docgroup xml:link="extended">
<locator xml:link="locator"
  href="main.xml"/>
<locator xml:link="locator"
  href="ohp.xml"/>
<locator xml:link="locator"
  href="www.xml"/>
</docgroup>
<related xml:link="extended">
<locator xml:link="locator"
  href="main.xml#ID(section_2)"/>
<locator xml:link="locator"
  href="ohp.xml#span(ID(p3),ID(p5))"/>
<locator xml:link="locator"
  href="www.xml#ID(section_2)"/>
</related>
```

図 2: 拡張リンクによる関連付けの例

```
main.xml:
...
<dbs9901 xml:link="group">
<docgroupdef xml:link="document"
    href="dbs9901-link.xml"/>
</dbs9901>
...
```

図 3: 拡張リンクグループの指定

このとき、次の 2 点が問題となる。

1. どの文書中に拡張リンクの定義を含めるか。
2. アンカーの指定をどのように行うか。

それぞれについて、次節以降で検討を行う。

3.1 拡張リンクを定義する文書

リンクの定義がどの異種文書からも参照できる必要があるので、拡張リンクの定義のみを集めた XML 文書 (リンク定義文書) を文書群ごとに一つ設け、文書群中の各文書から参照するという方法がよいと考えられる。たとえば、図 2 の XML 記述がリンク定義文書 `dbs9901-link.html` に書かれているとき、予稿本体の文書 `main.xml` では図 3 のように拡張リンクグループを指定する。

この方法を用いると、すでに存在する文書群内のリンクの定義の変更はリンク定義文書に対して行えばよく、元文書に手を加えなくてもよい。また、文書群の構成要素の変更も、変更対象となる文書とリンク定義文書にのみ手を加えればよい。

本稿で考えている文書管理では、文書群やリンクは動的に起こり得る。たとえば、ある文書とそれに関するメモを文書群として考えると、メモは動的に付加されたり削除されたりするものであるから、文書群を構成する要素が変更される。また、メモと元文書との関連を拡張リンクで実現するならば、リンクの定義もまた動的に変更される。

3.2 アンカーの指定方法

XML では、任意の文書要素をアンカーとすることができるため、すでに文書中に存在する文書要素をアンカーとして指定する場合は (文書要素に識別子が定義されていれば) 問題はない。

一方、メモを既存の文書に付加するような場合を考えると、アンカーは必ずしもすでに文書中に存在する文書要素であるとは限らない。この場合は、

1. 拡張ポインタ (extended pointer) によって、元文書に手を加えずにアンカーを指定する。
2. アンカーとしたい文書要素を (タグをつけることによって) 新たに定義する。

の 2 通りが考えられる。

1 の方法は、元の文書の DTD を変更することなく実現できることが利点である。しかし、現在の XPointer の規格では、元文書中の文書要素以外の部分をアンカーとして指定する場合は、要素名・文字列などを基にアンカーの候補を定め、その中で何番目の候補、あるいは要素や文字数をカウントするなどの方法を用いる。したがって、元文書に変更が加えられれば、拡張ポインタの定義にも変更を加えなければならない。

2 の方法は、アンカーとして指定できる部分が、あらかじめ DTD で定義された文書要素に

```
× <para>...<kw>...</para>
      <para>...</kw>...</para>
○ <para>...<kw>...</kw></para>
      <para><kw>...</kw>...</para>
```

図 4: アンカーの分割

限られる。そのため、たとえば2つの文書要素にまたがったアンカーを指定しようとすると、図4のようにアンカーを2つに分割して指定せざるを得ない。外部のプログラム (ブラウザなど) の実装で、これをあたかも一つのアンカーのように見せかける必要がある。

以上のように、1, 2いずれの方法でも、XML 処理系とは別に支援システムを用意しなければならない。

4 XML による異種文書ラッパー

現状では、2.3節で述べた方針はかなり強く、XML 形式と情報を落とすことなく相互に変換できる文書フォーマットは SGML などひじょうに限定される。

それ以外の文書を本稿の文書管理の枠組に取り込む方法として、XML による文書ラッパー (wrapper) というアプローチが考えられる。たとえば、`main.pdf` という文書に対して `main.xml` というファイルを用意し、`main.xml` の中に `main.pdf` から必要な情報を取り出して XML 形式に整形するプログラムの呼び出しに対応する文書要素を定義する。他の XML 文書からは `main.xml` 中に定義された文書要素に対してリンクを張る。

また、このアプローチでは、複数の文書をラッパーで仮想的に統合することができる。たとえば、PDF 形式や PostScript 形式の文書はレイアウト情報しかもっておらず、文書構造に関するデータを持っていない。そこで、文書構造に関するデータを記述した補助文書を用意し、必要に応じて補助文書中のデータを参照するような文書要素を `main.xml` 中に定義することで、外部からは、`main.pdf` が持っていない文書構造データを参照することが可能になる。

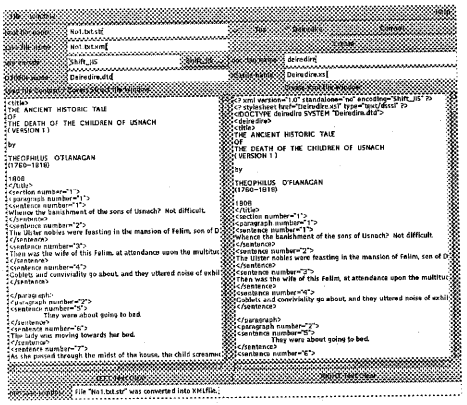

図 5: XML への変換システム

現在、文書ラッパーの実装の一貫として、XML への変換プログラムと XML リンクエディタを開発中である [9]。

図5に、我々が現在開発中である、XML への変換を行うプログラムの実行画面を示す。このプログラムでは、限定されたスタイル (典型的な article.sty, jarticle.sty を使って書かれたもの) の LaTeX 文書や、別プログラムを用いて構造化されたケルト民族の伝承文学であるデァドラ伝説のテキストを読み込み、与えられた DTD などの情報を基に XML 形式に変換する。向かって左のテキスト領域が変換前、右が変換後のテキストをそれぞれ表示している。

一方、**XML** リンクエディタは変換後の XML 文書に拡張リンクを張る作業を支援するプログラムである。実行画面を図6に示す。

現在のバージョンでは、XML 文書中にすでに存在する文書要素を root 要素からのパス表現で列挙し、これらの中からアンカーとするべき要素を選択して必要な情報を補うことで、拡張リンクを表す XML 文書を自動生成する。

5 おわりに

本稿では、互いに関連のある複数の異種文書を統合的に管理する利点、管理の枠組みについて述べ、構造化文書言語として現在注目されて

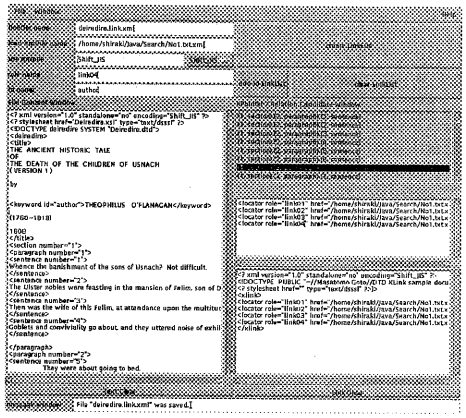

図 6: リンクエディタ

いる XML を用いた枠組みの実現方法について議論を行った。

XML は SGML や HTML に比べて規格が制定されてからまもないため、ブラウザその他の処理系が実装されて利用できるようになるにはまだ時間がかかると予想される。また、HTML のように、規格自体が今後拡張されていくことも考えられる。したがって、本稿の枠組みは、現状ではまだ現実的でない面がある。しかし、著者が聞く限りにおいては、アメリカなどでは XML への対応が本格的になってきている印象を受けており、近い将来現実的な枠組みになるであろうという期待を持っている。

謝辞

さまざまな議論を頂く岡山県立大学横田研究室の皆様に感謝します。なお、本研究の一部は文部省科学研究費特定領域研究 (A)「高度データベース」および基盤研究 (C) による。

参考文献

[1] B. Bos, H. W. Lie, C. Lilley, and I. Jacobs. Cascading Style Sheets, level 2. W3C Recommendation REC-CSS2-19980512, May 1998. available from http://www.w3.org/.

[2] T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0. W3C Recommendation REC-xml-19980210, Feb. 1998. available from http://www.w3.org/.

[3] J. Clark and S. Deach. Extensible Stylesheet Language (XSL). W3C Working Draft WD-xsl-19981216, Dec. 1998. available from http://www.w3.org/.

[4] H. W. Lie and B. Bos. Cascading Style Sheets, level 1. W3C Recommendation REC-CSS1-961217, Dec. 1996. available from http://www.w3.org/.

[5] E. Maler and S. DeRose. XML Linking Language (XLink). W3C Working Draft WD-xlink-19980303, Mar. 1998. available from http://www.w3.org/.

[6] E. Maler and S. DeRose. XML Pointer Language (XPointer). W3C Working Draft WD-xptr-19980303, Mar. 1998. available from http://www.w3.org/.

[7] XML/SGML サロン. 標準 XML 完全解説. 技術評論社, May 1998.

[8] 本行, 白木, 田槙, 国島, 横田. 文学データベースのためのプロトタイプシステムの実装. 第 57 回情報処理学会全国大会予稿集, Oct. 1998.

[9] 白木, 本行, 田槙, 國島, 横田. 異種文書データの統合管理システムの構想. 平成 10 年度電気・情報関連学会中国支部連合大会予稿集, pp. 111–112, Oct. 1998.

[10] 田槙, 本行, 白木, 國島, 横田. 文学データベースのための索引・検索機能の拡張. 平成 10 年度電気・情報関連学会中国支部連合大会予稿集, pp. 113–114, Oct. 1998.