A4-2

XML を用いた再構築可能な漢字文献データモデル

石川 正敏 ⋆　波多野 賢治 †　天笠 俊之 †　吉川 正俊 †,‡　植村 俊亮 †　勝村 哲也 ⋆

⋆ 島根県立大学 総合政策学部
† 奈良先端科学技術大学院大学 情報科学研究科
‡ 国立情報学研究所 ソフトウェア研究系

⋆ 〒697-0016 島根県浜田市野原町 2433-3
† 〒630-0101 奈良県生駒市高山町 8916-5
‡ 〒101-8430 東京都千代田区一ツ橋 2-1-2

⋆ {m-ishikawa, t-katsumura}@u-shimane.ac.jp, †‡ {hatano, amagasa, yosikawa, uemura}@is.aist-nara.ac.jp

あらまし　従来の図書館では公開が困難であった古典的な文献や資料をインターネット上で公開するために，文献の電子化が進んでいる．しかし，公開される文献の多くは閲覧だけが許されており，利用者による文献の引用などの二次利用を想定していないことが多い．そこで本研究では，電子図書館で公開される文献から必要な情報を自由に取り出すためのデータモデルと文献からの情報抽出操作を提案する．提案するデータモデルでは，様々な形式で公開される漢字文献に関する情報 (画像やテキストなど) を XML に変換することによって，文献に対する柔軟な検索と利用者による情報抽出の両立を実現する．利用者による情報の抽出操作は，提案モデルに従って電子化した文献から利用者が必要な部分を切り抜き，スクラップブックと同様な領域に貼り付ける操作である．また，本研究は漢字文献を対象に，利用者が情報を抽出するプロトタイプシステムを実装する．

キーワード　電子図書館，漢字文献，電子スクラップブック, 編集操作，文字列検索，XML

A Data Model for Reconstructable Kanji Documents using XML

Masatoshi Ishikawa⋆, Kenji Hatano†, Toshiyuki Amagasa†, Masatoshi Yoshikawa†,‡, Shunsuke Uemura† and Tetsuya Katsumura⋆

⋆ Department for Policy Studies, The University of Shimane
† Graduate School of Information Science, Nara Institute of Science and Technology
‡ Software Research Division, National Institute of Informatics

⋆2433-3, Nobara-cho, Hamada, Shimane 697-0016, Japan
†8916-5, Takayama, Ikoma, Nara 630-0101, Japan
‡ 2-1-2, Hitotsubashi, Chiyoda, Tokyo 104-8430, Japan

⋆ {m-ishikawa, t-katsumura}@u-shimane.ac.jp, †‡ {hatano, amagasa, yosikawa, uemura}@is.aist-nara.ac.jp

Abstract　Libraries have promoted classic literature as digital document that are difficult to be displayed in their original forms. These digital documents are now accessed through the Internet. However, such digital documents are available for reading only, not for citing nor editing. In this paper, we propose a data model for users to extract and edit digital documents from a digital library. The proposed model makes it possible to flexibly retrieve documents as well as to edit information by integrated managing of documents displayed as images or texts by XML. It also provides a method to cut a part of documents and paste it in an area of prototype systems, which is examined whether users can operate cut and paste using digital kanji documents.

Key words　Digital Library, Kanji Documents, E-scrapbook, Editing Method, String Retrieval, XML

1 はじめに

図書館の電子化が進み多くの古典的な文献が，テキストデータもしくは画像として公開さている．特に，日本，中国，韓国などの東アジア圏の古典的な文献は，Unicode [1] などの符号化文字集合にない文字や保存状態によって判別が困難な文字があるため，テキストデータと画像の両方の形式で電子化されることが多い．このように電子化された文献の多くは，電子図書館やデジタルアーカイブなどの形態でインターネット上で公開される．

しかし，インターネット上で公開される文献の多くは閲覧だけが可能であり，利用者による文献の引用などの二次利用を想定していない．しかし，研究活動などの情報収集は，文献に記述されている一部の文書だけを利用することが多いため，文献の二次利用の要求は高いと考えられる．

そこで，本研究では東アジア圏の文献を漢字文献と呼び，このような文献から利用者が必要な情報の探し出し，引用などの二次利用するための手法を提案する．文献集合からの情報抽出は，漢字文献集合から利用者が注目するトピックもしくはフレーズが記述されている文献を検索し，その結果から利用者がさらに自分の意図に合った文献の選択および必要な部分の取り出すことによって行われる．このような情報収集を実現するために，本稿では，漢字文献の検索と閲覧ためのデータモデルを提案する．漢字文献の表示には，元の文献をスキャナ等で電子化した画像を利用する．しかし，漢字文献の検索では，画像の特徴を用いた検索だけではなく，漢字文献の内容に対する文字列検索も必要である．そこで，本稿で提案するデータモデルは，漢字文献の画像と，関連するテキスト情報を単一のXML 文書として記述する．提案モデルによって，元の漢字文献の見た目を計算機上での再現することと漢字文献の内容に対する文字列検索の両方を実現する．また，提案モデルに従った漢字文献データの交換と Unicode などで記述できない文字である外字に関する情報をテキスト中の記述するために XML [2] を利用する．

本稿で提案する漢字文献からの情報抽出は，利用者が漢字文献の画像に対して指定した領域に従って，漢字文献データから画像及び関連するテキストデータを取り出す操作である．提案方法に従って漢字文献から取り出した情報を電子スクラップと呼ぶ．さらに本稿では，複数作成される電子スクラップを管理するための電子スクラップブックを提案する．

本稿では，提案する漢字文献データモデルと電子スクラップブックを用いた文献検索と電子スクラップの操作を検証するためのプロトタイプシステムを作成した．プロトタイプシステムでは，提案する漢字文献データモデルを Relax NG [3] を用いて記述し，百人一首 [4] を対象に漢字文献データを作成した．また，文字列検索処理のために漢字文献をデータベースに格納した．最後にプロトタイプシステムの実行例を示す．

2 漢字文献データモデル

2.1 漢字文献データモデルの構造

図 1 に漢字文献データモデルの概略を示す．漢字文献データモデルは，書誌情報，表示情報，内容情報，表示情報と内容表の対応表から構成され，単一の XML 文書として表現される．書誌情報は，漢字文献のタイトルなどを記述し，表示情報では，利用者に漢字文献を示すための画像を示す．内容情報は，表示情報の内容を書き下した文書や，人名や地名などの重要な単語を抜き出した単語リストなど，複数の情報の記述を許す．対応表は，各内容情報に合わせて記述するので，内容情報と同数記述する．

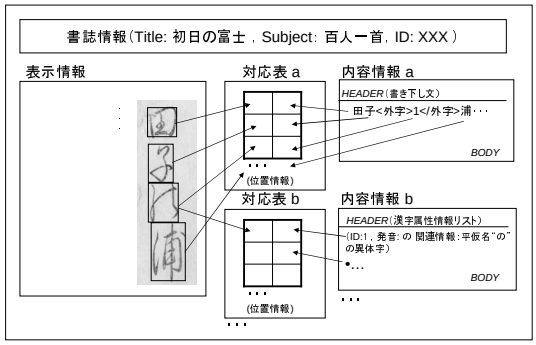

図 1: 漢字文献データモデル

図 1 は山部赤人の句を表す漢字文献データの例である．図 1 の表示情報の 3 文字目は，ひらがな ”の” の異体字であり，ここでは外字として扱う．

2.1.1 書誌情報

書誌情報では，識別子や製作時期などの文献に直接記述されない情報と，表題などの文献目録の作成に必要な情報を記述する．書誌情報で用いる要素名は，Dublin Core Element Set Version 1.1 [5] で定義された 15 項目を用いる (図 2)．このような既存の要素を用いることで他の目録管理システムなどとの情報交換が容易になると考えられる．

```
Title, Subject, Identifier, Creator,
Description, Publisher, Contributor, Date,
Type, Format, Source, Language, Relation,
Coverage, Rights
```

図 2: 書誌情報の要素

本稿では，文献を識別するための要素 Identifier，文献名を記述する要素 Title，文献の主題を記述する要素 Subject を必須とする．図 1 の書誌情報の記述例を以下に示す．

例 書誌情報の記述

```
<KDInfo>
<DC:Titel>初日の富士</DC:Title>
```

<DC:Subject>百人一首</DC:Subject>
<DC:Identifier>1</DC:Identifier>
</KDInfo>

各要素のプレフィックス ”DC:” は, Dublin Core Element Set Version 1.1 の名前空間を表す.

2.1.2 表示情報

表示情報は，利用者に漢字文献を示すための情報を記述する．本稿では，元の漢字文献をスキャナ等で電子化した画像を用いる．対象とする画像には以下のような種類がある．

(1) ラスタ画像
ラスタ画像には，bitmap，JPEG，GIF などの形式がある．この形式は，スキャナ等で作成でき，WWW 環境での画像交換でも広く利用されている．また，この形式は，境界強調などの画像処理に適しているので紙の変色や文字のかすれなどから判別が困難な文字を含む漢字文献の解読支援に利用できる．

(2) ベクトル画像
ベクトル画像は拡大縮小などの画像変換をしても形状が保存される形式なので，漢字文献の画像に記述されている文字の形状の詳細な調査などに利用できる．特に本稿では，ベクトル画像を XML 文書として管理できる SVG [6] の利用が適当であると考えられる．

表示情報における画像の記述は，MIME エンコードしたラスタ画像や SVG 画像を直接記述するか，画像の参照情報として URL を用いて間接的に記述する．図 1 の表情情報を URL を用いて記述する例を以下に示す.

例 表示情報の記述

```
<KDimage>http://foo.ac.jp/foo.jpg</KDimage>
```

2.1.3 内容情報

内容情報は，漢字文献の内容に関連したテキスト情報である．例えば，漢字文献の内容を表したテキストや，英訳したテキスト，人名や地名などの重要な単語に対する辞書情報などが挙げられる．従って，本稿では一つの表示情報に対して複数の内容情報が存在する．そこで内容情報の構成は，漢字文献データ内で固有のキーワードを記述するヘッダと，本文の組として記述する．また，ヘッダは，他の内容情報と識別するために利用する．

内容情報の本文に記述する内容には以下のような種類がある．

書き下し文 書き下し文は，表示情報の内容を記述した文書である．文書の記述には，Unicode などの符号化文字集合を用いる．文書中の外字がある場合，次に述べる漢字属性情報リストの識別子を参照情報として記述する．参照の記述には，XML を用いる．さらに，文字によっては複数の解釈が存在することが考えられるので，参照情報は，一つ以上の漢字属性情報リストの識別子を記述することがある．従って，書き下し文は，符号化文字と漢字属性情報リストへの参照で構成された XML 文書であると言える．図 1 に関する書き下し文の記述例を以下に示す.

例 書き下し文の記述

```
<KDcontent>
<header>書き下し文</header>
<body>田子<gaiji>1</gaiji>浦</body>
</KDcontent>
```

要素 gaiji は，次に示す漢字属性情報リストへの参照の記述例である.

漢字属性情報リスト 漢字属性情報リストでは，漢字文献に記述されている文字の形状や発音，意味などの文字に関連する情報を記述する．漢字の属性情報は，漢字文献中の文字の識別子と，部首，総画数，四角号碼などの漢字の形状に関する情報，漢音，呉音，訓，pinyin などの読みに関する情報，漢字の意味や異体字，eKanji [7] や今昔文字鏡 [8] などの大規模漢字集合への参照などの関連情報の組として記述する．図 1 に関する漢字属性情報リストの記述例を以下に示す.

例 漢字属性の記述

```
<KDcontent>
<header>漢字属性情報リスト</header>
<body>
<KanjiInfo><ID>1</ID>
<pronunciation>の</pronunciation>
<relatedInfo>"の"の異体字</relatedInfo>
<KanjiInfo>
<KanjiInfo>...</KanjiInfo>...
</body>
</KDcontent>
```

要素 KanjiInfo は一つの文字情報を表し，要素 ID は他の内容情報が参照するために利用する．また，要素 pronunciation と要素 relatedInfo はそれぞれ文字に関する発音および関連情報を記述する．この例では，文字の形状情報を省略している.

単語リスト 漢字文献には，人名や地名，仏教用語のような専門用語などの重要な単語が含まれている．これらの単語に関する意味や注釈は，文献を検索する上で重要な手がかりとなる．そこで，本稿では，漢字文献中の単語と意味の組をリスト形式で記述する．また，単語に含まれる外字は，書き下し文と同様に漢字属性リストの識別子を記述する．図 1 に関する単語リストの記述例を以下に示す.

例 単語リストの記述

```
<KDcontent>
<header>単語リスト</header>
```

```
<body>
<word>
<string>田子</string><info>...</info>
</word>
<word>...</word>...
</body>
</KDcontent>
```

要素 word は，一つの単語情報を表し，要素 string，要素 info はそれぞれ表示情報に記述されている単語と，単語の意味を記述する．

2.1.4 対応表

対応表は，内容情報と表示情報の関連を表形式で記述したものであり，内容情報から表示情報への写像として扱える．対応表に記述する内容情報に関する情報は，文字の出現位置である．内容情報が書き下し文の場合は，文字列の先頭から順に与えた番号を用いる．漢字属性情報や単語リストなどのリスト形式の情報は，各項目の順に従って番号を与える．

対応表に記述する表示情報の領域は，表示情報に描かれている部分画像の範囲を表現する．そのために本稿で用いる座標系は，画像の左上を原点とし，水平方向の左向きを X 軸の正の向き，垂直方向下向きを Y 軸の正の向きとする．領域の形状は長方形であり，原点から最も近い点 $(minX, minY)$ と最も遠い点 $(maxX, maxY)$ の 2 点の座標値を対応表に記述する．例えば，画像中の文字の範囲を記述する場合は，文字を囲む Minimum Bounding Rectangle の原点に最も近い点と最も遠い点の座標を用いる．また，対応表は複数存在するので，どの内容情報と表示情報の対応表であるかをヘッダに記述する．

例である図 1 を対象に書き下し文と表示情報に関する対応表の記述例を以下に示す．下記の記述例では，書き下し文の本文の 1 文字目である ”田” に対応する表示情報の領域を記述している．

例 対応表の記述

```
<KDcorrespondence>
<KDheader>書き下し文</KDheader>
<row>
<cposition>1</cposition>
<ipositon>
<minX>100</minX><minY>100</minY>
<maxX>120</maxX><maxY>120</maxY>
</iposition>
</row>
<row>...</row>...
</KDcorrespondence>
```

要素 KDheader は，どの内容情報に関する対応表であるかを表している．要素 row が対応表の一つ対応情報であり，要素 cposition が内容情報の位置，要素 iposition が，対応する表示情報の領域を表す．

2.2 漢字文献の検索

漢字文献の検索は，各内容情報や表示情報などの一つ情報に対する検索と，書き下し文と漢字属性情報リスト，漢字属性情報リストと単語リスト，表示情報の領域などのような二つ以上の情報に対する検索を組み合わせる方法が挙げられる．特に提案モデルでは，外字を含む文字列検索のように漢字属性情報リストと組み合わせた検索が多く行われると考えられる．また，ここでは検索で得られる結果は，すべて表示情報の検索に一致した文字が描かれている領域とする．

以下に主な検索処理について述べる．

- 漢字属性情報リストの参照を含む検索キーによる検索
  書き下し文や単語リストなどの各内容情報は，テキスト情報であるため，書き下し文に対する検索，漢字属性情報リストの関連情報に対する検索などどれも同じ文字列検索として処理できる．また，検索キーが，符号化文字だけで構成されているのであれば，一般的な文字列検索の手法が利用できると考えられる．そこで，ここでは検索キーに漢字属性情報リストへの参照が含まれる検索を考える．この場合の検索処理は，次の通りである．(1) 検索キーに含まれる漢字属性情報リストへの参照部分を任意の文字と一致するもとみなして文字列検索を行う．(2) (1) の結果として得られる内容情報から検索キーと一致した部分を取り出す．(3) 取り出した文字列の中に，参照のための要素がありかつ参照値が含まれているものを選択する．
- 外字の属性情報を含む検索キーによる検索
  検索キーの外字の記述に，読みや画数などの外字の属性情報を用いる場合を考える．この検索では，書き下し文と漢字属性情報リストを組み合わせた検索になる．その処理は，検索キーから外字の属性情報の部分を取り出し，それに基づいて漢字属性情報リストの文字列検索をする．その結果として漢字属性情報リストの ID が得られるので，外字の属性情報を含む検索キーは，漢字属性情報リストの参照情報を含む検索キーに変換できる．変換された検索キーを用いて，先に述べた漢字属性情報リストの参照を含む検索キーによる検索をすることで処理できると考えられる．
- 表示情報の領域に着目した検索
  検索キーとして表示情報の領域を用いる検索を考える．このような検索は，写本の比較などのように同じ場所に同じ語彙が出現すると期待される文献の検索に利用できると考えられる．ただし，表示情報で用いる画像のサイズがすべて同じあるとは限らないので，領域を検索キーとして利用するには領域の位置を正規化する必要がある．領域の正規化は，指定した領域の座標値を表示情報の縦横それぞれの大きさで割る

ことである．また，領域だけでは，そこに含まれる文書の内容が不明なので，領域の座標値を検索キーに対応表を検索し，書き下し文に記述されている文字列を取り出す．検索処理は，次の通りである．(1) 取り出した文字列を用いて書き下し文の検索を行い，文字列が出現する漢字文献と各文献の表示情報の領域を得る．(2) (1) の各検索結果に対し領域の位置を正規化する．(3) 正規化された各漢字文献の領域と，検索キーとして与えた領域を比較し二つの領域が重なれば，その漢字文献を検索結果として選択する．

- 複合検索
  複合検索は内容情報に対する検索と表示情報に対する検索を組み合わせた検索である．内容情報の検索結果は，対応表を用いることで表情報の領域に写像できる．また，表示情報への検索は，表情報の領域を検索キーに用い，検索結果も表示情報の領域である．従って，複合検索では，それぞれの検索結果から得られる領域の内包関係を調べることで実現できると考えられる．

3 電子スクラップブック

3.1 電子スクラップブックと漢字文献の関係

図 3 に漢字文献データと電子スクラップブックの関係を示す．図 3 に示す通り，電子スクラップブックは，複数の電子スクラップからなる．電子スクラップは，切抜き元の漢字文献データから，漢字文献の識別子，内容情報，対応表の要素から関連する要素を抜き出したものである．さらに，電子スクラップ独自の情報として，他の電子スクラップと区別するための識別子とこのスクラップを視覚化するときの位置情報であるオフセットを持つ．電子スクラップの内容情報に外字が含まれる場合は，切抜き元の漢字文献の漢字属性情報を参照する．また，このデータモデルでは切抜きの表示に必要な表示情報を直接記述せず，電子スクラップに記述している切抜き元の漢字文献の識別子と，対応表の領域情報を用いて間接的に参照する．

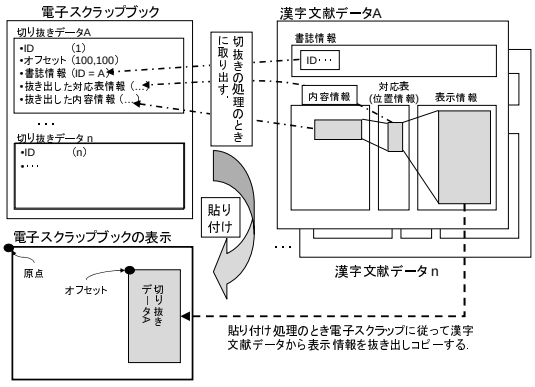

図 3: 漢字文献データと電子スクラップブック

3.2 切抜きと貼り付け操作

漢字文献データから，利用者が必要な情報を取り出すための切抜き操作について述べる．切抜きの対象である漢字文献は，XML 文書であるため，この操作は，XML 文書から必要な要素を取り出す操作である．

1. 切抜き要求.
   利用者は，漢字文献の識別子と表示情報に対して切抜き領域を指定する．
2. 漢字文献データから要素の抽出.
   指定された漢字文献識別子に一致しかつ，指定された領域に内包される座標が記述されている対応表の一部を抜き出す．次に抜き出した対応表の内容情報の出現情報の最大値と最小値を調べる．書き下し文である内容情報と先に求めた文字列の領域に基づいて切抜きに対応した内容情報を取り出す.
3. 電子スクラップブックへの登録
   抜き出したデータに対し，電子スクラップの ID とオフセットの初期値を追加した上で切抜きを電子スクラップとして電子スクラップブックに追加する．

次に貼り付け操作について述べる．この操作は，電子スクラップを視覚化する操作である．電子スクラップ自身は，表示情報を持たないため，この操作をする場合，切抜き元の漢字文献データから間接的に表示情報を取得しなければならない.

a. 表示位置の指定
   貼り付ける電子スクラップの ID を指定し，スクラップの表示位置を指定する．
b. 漢字文献データからの切抜き画像の生成
   電子スクラップが参照している漢字文献識別子と，対応表に記述している座標値をすべて取り出す．取り出した座標値の集合から X 軸，Y 軸それぞれの座標値の最大と最小を求める．求めた値の組によって得られる長方形の領域に従って参照元の漢字文献データの表示情報から画像を切取る．最後に切取った画像を電子スクラップブック上に表示する．

複数の電子スクラップの関連の分析や分類をするために，電子スクラップの貼り付ける位置は自由に変更できる．

4 プロトタイプシステム

4.1 プロトタイプシステムの構成

プロトタイプシステムは，漢字文献データを管理するためのサーバと，漢字文献を閲覧するためのインタフェースを提供するクライアントの構成にした (図 4)．サーバは，漢字文献データをデータベースに格納するための DB 登録エンジン，漢字文献を管理する漢字文献 DB，クライアントとの通信と問合せの変換等の処理をする漢字文献管理システムからなる．クライアントは，漢字文献データ

の閲覧と検索をするための漢字文献閲覧インタフェースと，電子スクラップを操作する電子スクラップブックインタフェースからなる．クライアントとサーバ間の通信プロトコルは，HTTP を用いる．これにより利用者はネットワークのセキュリティーポリシーを意識することなく漢字文献の閲覧等の操作ができると考えられる．また，電子スクラップブックは，ファイルとしてサーバ側で管理することで，複数の利用者による共有が可能になると考えられる．

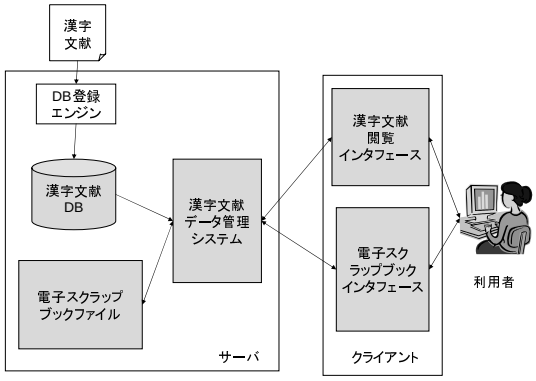

図 4: プロトタイプシステムの概要

4.2 実装環境

プロトタイプシステムは，Java1.3 を用いて実装した．クライアントとサーバ間の通信には，Apache と Tomcat を利用した．漢字文献データを管理するデータベースには，Oracle 8.1.6 を使用した．

本稿で使用したデータは，表示情報に玉川大学図書館 Web サイトの漢籍・和装本コレクションとして公開されている百人一首の画像 [4] を利用し，内容情報には，百人一首の内容を書き下したテキストを用いた．漢字文献データのスキーマ記述には，Relax NG [3] を利用した．Relax NG は，XML 文書のスキーマ定義言語であり，DTD に比べデータの記述能力が高く，XML 文書として記述できる．また， Relaxer [9] などの関連技術を用いることで XML 文書に関する操作を容易に実装できる．図 5 に本稿のプロトタイプで用いた Relax NG で記述したスキーマの主要な部分を示す．図 5 の 要素 KDinfo では書誌情報として，Title，Subject，Identifier を記述する．要素 KDimage には，百人一首の画像が格納されている場所の URL を記述する．要素 KDcontent は，百人一首の画像に対応したテキストデータを記述する．要素 KDcorrespondence は，百人一首の画像の内容と内容情報との対応を記述する．

4.3 漢字文献データのデータベースへの格納

大量の漢字文献への検索処理のためには，データベース技術の利用は不可欠であると考えられる．データベースシステムにはオブジェクト指向データベースや XML 文書処理に特化したデータベースシステムなどがあるが，今回は，一般的に広く普及している関係データベースに漢字文献データを格納した．漢字文献データの書誌情報，表示情報，内容情報，対応表の各要素は，それぞれ独立した要素である．そこで本稿では，漢字文献の要素ごとに分割し，対応した関係表に格納した．各関係表に格納される値が，どの文献に属するかを明示するために書誌情報の ID を外部キーとして参照する．図 6 は， 図 5 の各要素に対応した関係スキーマを示す．

```
<element name="KanjiDocument">
<element name="KDinfo">
<element name="Title"><text/></element>
<element name="Subject"><text/></element>
<element name="Identifier"><text/></element>
</element>
<element name="KDimage"><text/></element>
<element name="KDcontent">
<element name="header"><text/></element>
<element name="body"><text/></element>

</element>
<element name="KDcorrespondence">
<element name="KDheader"><text/></element>
<oneOrMore>
<element name="row">
<element name="cposition"><text/></element>
<element name="iposition">
<element name="minX"><text/></element>
<element name="minY"><text/></element>
<element name="maxX"><text/></element>
<element name="maxY"><text/></element>
</element>
</element>
</oneOrMore>
</element>
</element>
```

図 5: プロトタイプシステムで利用した Relax NG スキーマ

書誌情報 (Identifier, Title, Subject, filename)
表示情報 (Identifier, imageInfo)
内容情報 (Identifier, Header, Body)
対応表 (Correspondence_Identifier, Identifier, KCNumber, maxX, maxY, minX, minY)

図 6: 漢字文献データの格納に用いた関係スキーマ

図 6 で下線が引かれた属性は，その関係表の主キーである．関係**書誌情報**では，図 5 で定義した書誌情報の要素と漢字文献データのファイル名を記述する．関係**表示情報**には，漢字文献の識別子と画像の格納されている URL を記述する．関係**内容情報**には，漢字文献の内容情報の各要素と対応する漢字文献の識別子を記述する．関係**対応表**の内容情報の各要素とこの関係表独自の識別子を記述

する．

4.4 漢字文献検索の実装

今回のプロトタイプシステムで実装した漢字文献の検索について述べる．プロトタイプシステムで実装した検索機能は，書き下し文だけを対象としている．検索結果は，検索キーが含まれる文献と文献中に検索キーが現れる領域を表す座標の組の集合である．

1. 検索キーの入力.
   検索キーの入力は，文字列を直接入力する方法と，漢字文献データを利用して間接的に入力する方法がある．特に後者の方法は外字を含む検索キーの入力に有効であると考えられる．しかし，今回のプロトタイプシステムでは外字処理は考慮していない．漢字文献データを用いた検索キーの指定は，次の通りである．(1) 漢字文献の識別子と検索キーが描かれた表示情報の範囲を指定する．(2) 指定された漢字文献識別子と一致しかつ，指定された範囲に内包される座標値を含む文字の出現位置 $p_1, ..., p_n$ を漢字文献の対応表から取り出す．(3) 指定された漢字文献の識別子に一致する内容情報の要素 Body の値である文字列 s を取り出す．(4) 文字列 s から文字の出現位置 $p_1, ..., p_n$ に含まれる文字を取り出し検索キーを生成する．
2. 内容情報への問い合わせ.
   検索キーを用いて各漢字文献の内容情報に対し問合わせをする．結果は，漢字文献の識別子と表題及び内容情報の要素 Body の値の組の集合である．ただし，単語リストや漢字属性情報への問合せの場合は，項目の出現位置を返す.
3. 検索キーの出現位置の検索.
   書き下し文への問合せの結果はその文書全体であるので，検索キーの出現位置を別途求めなければならない．そこで本稿では，ある文字列に対し先頭から検索キーと比較し，一致しなければ文字列の先頭の 1 文字を削除し残りの文字列と検索キーを再度，比較する．もし文字列比較が一致すれば，削除した文字数に 1 加えた値を文字列中の検索キーの出現位置として記録し，さらに検索を続ける．比較対照の文字列が検索キーの長さより小さくなったときに検索を終了する.
4. 表示位置の問合せ.
   3. の結果と対応表に対し問合せを行い対応する表示情報の領域を得る.

4.5 プロトタイプシステムの実行例

4.5.1 漢字文献の閲覧例

図 7 に漢字文献の閲覧例を示す．図 7 に示すインタフェースでは，右側のテキストエリアに検索結果を表示する．検索結果は，検索に一致した漢字文献の ID，タイトル，および検索キーがマッチした領域の座標を表示する．また左側には，漢字文献自身を表示する．検索キーと一致した語は，その部分を文字ごとに赤い線の長方形で囲む．図 7 は，百人一首のデータに対して 検索キー「天皇」で検索した結果を表示した例である．

図 7 の下部にあるボタンとテキストフィールドを用いて，漢字文献を操作する．検索キーの入力は，検索キーを直接入力する方法と，漢字文献データを用いて間接的に入力する方法がある．前者は，テキストフィールドに検索キーを入力し ボタン ”QBS” を押すことで実行する．後者の方法は，表示されている漢字文献に対し検索キーとなる範囲をマウスで指定し，ボタン ”QBI” を押すことで実行する．漢字文献の ID をテキストフィールドに入力した上で，ボタン ”Display” を実行することで，対応した漢字文献が表示される．漢字文献からの切抜きは，表示されている漢字文献に対し範囲をマウスで指定し，ボタン ”Cut” を選択することで実行する．切抜きは，電子スクラップブックに登録される．また，検索で得られる漢字文献への領域情報を用いて切抜きを行う場合は，ボタン ”CBR” を押す.

図 7: 漢字文献閲覧システム実行例

4.5.2 電子スクラップブックの操作例

図 8 は，百人一首からの切抜きを電子スクラップブック上に表示した例である．図 8 は，三つの百人一首データから和歌の冒頭と著者名を抜き出したものを整理した図である．電子スクラップブックインターフェースは，現在，管理している電子スクラップの一覧を図 8 の右側のテキストエリアに示す．また，左側が電子スクラップを表示する領域である．テキストエリアに表示される一覧は，項目番号と，内容情報の組を順に表示したものである．項目番号は，現在操作可能な電子スクラップを示すために電子スクラップインタフェースが一時的に割り振った値である.

図 8 の下部には，電子スクラップ操作のためのボタンとテキストフィールドが配置している．電子スクラップの表示は，電子スクラップの項目番号および表示位置を指定した上で ボタン ”Past” を選択することで行う．また，電子スクラップの削除は，項目番号を指定し，ボタン

”Delete” を押すことで実行する．ボタン ”Import” はサーバで保存している電子スクラップブックをクライアント側への取り込みを行い，ボタン ”Export” はクライアントでの操作結果をサーバに送信する機能を実行する．

図 8: 電子スクラップブックシステムの実行例

5 関連研究

PDF [10] や eBook [11] は，インターネットを介して文献や書籍を交換するためのフォーマットである．これらは，電子的に作成された文書の画面での表示と印刷した結果を同じにすることを目的としたフォーマットであり，文献の見た目の情報とテキスト情報を同時に取り出すことができない．対して，本研究では文献の画像としての側面と文書としての側面を残したまま切抜きや貼り付け操作を実現すること目指す．

XLibris [12] は，active reading の支援のために電子文献への下線や注釈，切抜きなどの実際の紙と同様な操作環境を提供している．データ構造などで本研究との類似点も多いが，XLibris [12] は，ユーザインタフェースに着目した研究であるのに対して本研究では，データベースおよび情報検索の手法から文献の有効利用の実現を目指している．

6 まとめ

本稿では，実物の公開が困難である古典的な漢字文献を対象に，電子化した漢字文献を管理するためのデータモデルを提案し，そのモデルに対する検索及び切抜き操作と，その切抜きを管理する電子スクラップブックを提案した．漢字文献は，Unicode などの符号化文字集合だけで表現できないことがあるので，文献の表示には画像を用いる．しかし，漢字文献の画像だけを管理していたのでは，テキストによる文献検索ほど効率的な検索ができない．そこで本稿で提案した漢字文献データモデルでは，画像と漢字文献に関する情報とそれらの関連を一つの XML 文書として記述する．提案モデルに従った漢字文献は，元の漢字文献の見た目を計算機上で再現することと，文字列による文献検索の両方を実現できる．本稿では，漢字文献データに対する検索と電子スクラップブックに対する操作について述べ，プロトタイプシステムを実装し，実行例を示した．

今後の課題として，まず，本稿で述べた外字情報を含む検索キーを用いた検索や複数の検索キーを用いた検索などの今回のプロトタイプシステムでは実装していない検索機能の実現が挙げられる．他に，電子スクラップブックを有効に活用するために電子スクラップを貼り付ける機能だけではなく，電子スクラップブックに対して利用者独自のメモを貼り付ける機能の実現を目指す．メモを貼り付ける機能があれば，電子スクラップブックを複数の利用者で共有する場合の意思疎通の支援や，電子スクラップの分類について理解支援に利用できると考えられる．メモ機能は，電子スクラップの表示情報を漢字文献への参照だけではなく，利用者が独自作成した画像への参照を許せば実現できると考えられる．

謝辞 本研究の一部は，文部科学省科学研究費 (課題番号 11480088, 13780339) によるものである．ここに記して謝意を表す．

参考文献

[1] The Unicode Consortium: ”The Unicode Standard, Version 3.0”, Reading, MA, Addison-Wesley, 2000.

[2] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen et al.: ”Extensible Markup Language (XML) 1.0 (Second Edition)”, http://www.w3.org/TR/2000/REC-xml-20001006, 6 October 2000.

[3] James Clark, Makoto Murata: ”RELAX NG Specification”, the Organization for the Advancement of Structured Information Standards, http://www.oasis-open.org/committees/relax-ng/spec-20011203.html, 12/3, 2001.

[4] 玉川大学図書館: ”百人一首”, 漢籍和装丁本コレクション, http://www.tamagawa.ac.jp/sisetu/tosyo/w_index.htm, 2000.

[5] Dublin Core Metadata Initiative: ”Dublin Core Metadata Element Set, Version 1.1: Reference Description”, http://dublincore.org/documents/1999/07/02/dces/, 7/2, 1999.

[6] Jon Ferraiolo: ”Scalable Vector Graphics (SVG) 1.0 Specification”, http://www.w3.org/TR/SVG/,4 September, 2001.

[7] 勝村哲也，丹羽正之: ”eKanji”, 電子漢字研究会, http://nohara.u-shimane.ac.jp/ekanji/, 2000.

[8] 文字鏡研究会: ”今昔文字鏡”, http://www.mojikyo.org/html/index.html

[9] 浅海智晴: ”Relaxer”, http://www.asahi-net.or.jp/dp8t-asm/java/tools/Relaxer/index_ja.html

[10] Adobe Systems Incorporated: ”PDF Reference third edition Adobe Portable Document Format Version 1.4”, http://partners.adobe.com/asn/developer/acrosdk/docs/filefmtspecs/PDFReference.pdf, 1999.

[11] The Open eBook Forum: ”The OeBF Publication Structure 1.0.1 Recommended Specification”, http://www.openebook.org/oebps/oebps1.0.1/download/

[12] Morgan N. Price, Bill N. Schilit, and Gene Golovchinsky: ”XLibris: The Active Reading Machine”, CHI 98, ACM Press, pp. 22-23, Los Angeles, CA, April 18-23, 1998.