歴史的文献画像のための電子スクラップブックシステム

石川正敏[†] 波多野賢治[††] 天笠俊之[††]
植村俊亮[††] 勝村哲也[†]

本論文では，電子図書館などで公開されている東アジア圏の歴史的文献の画像と関連する書誌や注釈などの情報を共有するためのデータモデルを提案し，プロトタイプシステムの実装について報告する．提案モデルは，文献画像と注釈などの関係を管理する文献データモデルと，文献データの分類を管理する電子スクラップブックデータモデルからなり，このモデルに基づいた操作には，注釈の編集，文献データからの切り抜き，文献データの分類，検索がある．本研究では，提案モデルに基づくデータの記述に XML を利用するので，提案モデルの XML への写像について述べる．検索処理に関係データベースを利用するので，提案モデルに対応した関係表の設計と検索条件からの SQL の生成について述べる．このモデルに従った電子スクラップブックシステムのプロトタイプの実装を用いて，提案モデルの実用性を示す．提案モデルによって，歴史的文献の画像を用いた関連する情報の集約が可能となり，人文社会科学や歴史学研究における効率的な情報の収集と共有が可能になると考えられる．

An Electronic Scrapbook System for Historical Document Images

Masatoshi Ishikawa,[†] Kenji Hatano,[††] Toshiyuki Amagasa,[††]
Shunsuke Uemura[††] and Tetsuya Katsumura[†]

In this paper, we propose a data model and implementation of a prototype system for sharing images and text data related to East Asian historical documents. Our proposed model consists of a document image data model and an electronic scrapbook data model. A document image data model manages relationships between document images and their annotations. An electronic scrapbook data model manages document image data classified by a user. Our proposed model has editing operation for annotation, document image data clipping operation, and classifying operation of document image data. Our proposed model also has retrieval operation. To describe data based on our proposed model by XML, we explain mapping method from our data model to XML. In order to use a database for retrieving document image data, we explain mapping method from our data model to relational table, and translation function of retrieval equations into SQL. As an example, we implement a prototype system for illustrating usefulness of our proposed model. Our proposed model enables researchers in humanities disciplines and social sciences to collect and share historical documents, because the users can aggregate historical documents widely available in digital libraries.

1. はじめに

近年，図書館や大学などの研究機関によって，日本，中国，韓国などの東アジア圏の歴史的文献の電子化とインターネット上での公開が活発に行われている[1)~3)]．また，人文社会科学，歴史学研究におけるインターネット利用として，歴史地理情報の交換[4)]，古典の多言語検索[5)]，古文書の閲覧[6)] などが提案されている．

東アジア圏の歴史的文献はくずし字のように文字の判別が困難な記述や古語や地名など現代と読みや意味が異なる記述があるため，その文献の内容理解に辞書情報などの注釈が必要である．しかし，歴史的文献や辞書情報は研究機関ごとに独自の形式で公開されるため，関連のある情報であっても公開元が異なれば，その関連が十分に示されるとは限らない．したがって，利用者が独自に歴史的文献の関連する情報を編集し共有するための手法が必要であると考えられる．

本研究では，東アジア圏の歴史的文献を対象に，文献と関連する情報の関係を管理する枠組みと，それらを共有するための操作を提案する[7)]．本論文では，歴

† 島根県立大学総合政策学部
Department of Policy Studies, The University of Shimane
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

史的文献の画像と関連するテキストデータの関係を管理するための文献データモデルと，文献データの分類を管理する電子スクラップブックデータモデルについて述べる．このモデルに従ったデータの記述には，XML[8] を利用する．XML によって，文献の画像とテキストデータのように異なるメディアの関係をシステムに依存しない形式で記述することができる．

本論文では，文献画像の内容理解の支援のための注釈編集操作について述べる．さらに，関連する文献を収集することが，文献の内容理解の役に立つと考えられる．そこで，関連する文献データの収集操作として，文献データの一部から新たな文献データを生成する切り抜き操作，文献データの関連を表すための分類操作について述べる．データの共有支援として，検索処理について述べる．本研究では，提案モデルの検索に関係データベースを利用する．そこでこのモデルに従ったデータを関係データベースで管理するための関係表の設計と，検索条件からの SQL 文の生成について述べる．最後に，提案モデルに基づいたプロトタイプシステムの実装について報告する．プロトタイプシステムは，クライアント/サーバ形式のシステムであり，各データの編集，切り抜き，分類操作，文献データの検索機能を実装している．プロトタイプシステムを用いた文献データと電子スクラップブックの閲覧，および文献データの検索例を示し，提案モデルの有効性を示した．このモデルによって，人文社会科学や歴史学の研究における文献と関連する情報の関係づけを研究者自身によって行うことが可能となる．また，研究者間で提案モデルに従ったデータを共有することにより，共同研究に役立つと考えられる．

本論文の構成は次のとおりである．2 章では，東アジア圏の歴史的文献のための電子スクラップブックシステムの基本的な要求について考察する．3 章では，文献データモデルと電子スクラップブックデータモデルについて述べる．4 章では，XML による提案モデルの記述，関係表への変換および，提案モデルに従ったデータ編集処理と検索処理について述べる．5 章でプロトタイプシステムの実装について述べ，データの閲覧と検索の実行例を示す．6 章では，関連研究として他の電子文書モデルについて述べ，最後にまとめと今後の課題を示す．

2. 東アジア圏の歴史的文献について

本章では，東アジア圏の歴史的文献の利用について考察し，本論文で提案する歴史的文献のための電子スクラップブックシステムの基本的な要求をまとめる．

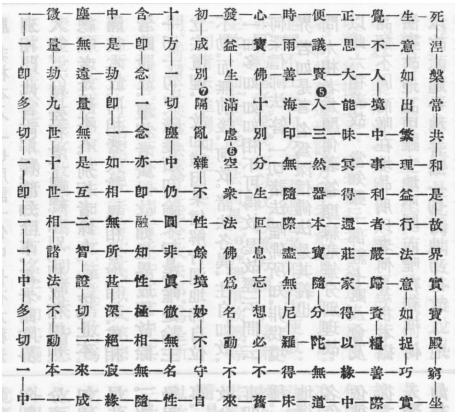

図 1 特殊なレイアウトの歴史的文献の例
Fig. 1 An example of historical document which has a unique layout.

(1) 歴史的文献の文書としての特徴と画像としての特徴の計算機上での利用

人文社会科学研究の対象となる文献の多くは作成された年代が古いため紙の劣化や汚れにより読解が困難であるが，人文社会学研究にとって，このような文献の特徴は，年代，地域，人物を特定するために重要な情報である．したがって，計算機を用いた歴史的文献の閲覧には，原文献の見た目の再現に適している画像を用いるべきである．本論文では，このような画像を文献画像と呼ぶ．たとえば,「華厳一乗法界圖」は文書を渦巻き状に記述した歴史的文献である（**図 1**）[9]．この文献では，2 次元空間に配置された渦巻き状の字の並びも仏教の精神を表現しているため，文字列としての特徴だけではなく，画像としての特徴も重要であることが分かる．一方，画像だけで歴史的文献を管理した場合，文献画像に対する効率的な検索は，期待できない．そこで，歴史的文献を計算機上で扱うためには，文献画像だけではなく，文献の内容を表すテキストデータもあわせて管理することが必要である．さらに，歴史的文献の内容は，古語や現代と異なる地名などがあるため，注釈を用いて内容を補足する必要がある．

(2) 外字情報の扱い

東アジア圏の歴史的文献は，外字と呼ばれる Unicode などの標準的な符号化文字集合にない文字を含むことが多い．一般に外字は文字列一致の対象として直接扱うことができないが，歴史的文献に対する検索効率を上げるには，外字も符号化文字と同様に文字列一致の対象とする必要がある．

(3) 歴史的文献収集の支援

歴史的文献の収集では，文献全体を収集するほかに，

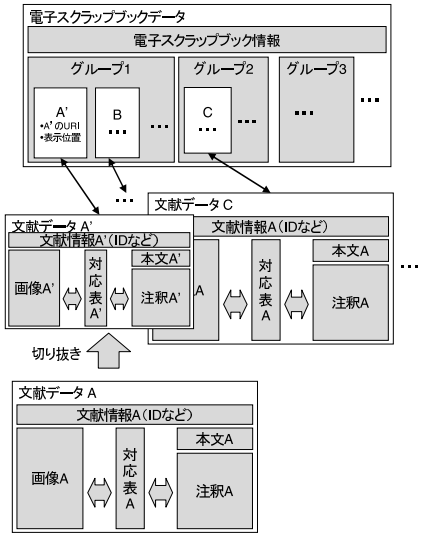

図 2 文献データモデルと電子スクラップブックデータモデル
Fig. 2 Document image data model and electronic scrapbook data model.

文献の一部だけを収集することがある．たとえば，百人一首から “春” に関する記述を抜き出し収集することがあげられる．このような文献収集の支援には，元の文献画像から任意の部分を切り抜く機能や，利用者が収集したデータの分類を管理する機能が必要である．さらに，文献画像を単に抜き出しただけでは元の文献画像と注釈の関係が失われるので，利用者は内容理解の手がかりを失ってしまう．そこで，文献画像の切り抜きにあわせて，その部分に関係付けられた注釈を抜き出す必要がある．

3. データモデル

本章では，文献データモデルおよび電子スクラップブックデータモデルについて述べる．文献データモデルは，歴史的文献の画像と関連するテキストデータを管理するためのデータモデルである．このモデルによって，歴史的文献の文書としての特徴と画像として特徴を計算機上で同時に扱うことを実現する．電子スクラップブックデータモデルは，文献データを分類し，それらの参照などの関連を表すためのモデルである．文献データモデルと電子スクラップブックデータモデルの関係を**図 2** に示す．文献データモデルは，文献画像，文献情報，本文，注釈を管理し，対応表で注釈と文献画像の位置関係を管理する（図 2 下）．電子スクラップブックデータは，このデータのメタデータである電子スクラップブック情報と文献データの分類を記述するグループからなる（図 2 上）．

図 2 の文献データ A’ は文献データ A からの切り抜きによって生成した文献データである．文献データ A’ も文献データモデルに従うので，他のデータと区別せずに管理できる．

3.1 文献データモデルの構成要素

(1) 文献情報

文献情報は，表題などの文献のメタデータを記述する．メタデータの項目に標準的な項目を用いれば，他の図書目録などのメタデータ DB と組み合わせた文献検索が可能になると考えられる．そこで本論文では，文献情報の属性に Dublin Core Metadata Element Set Ver. 1.1 [10] を利用する．

例 1. 図 1 の場合，文献に直接記述されていない著者や出典などの情報を記述する．

文献情報 = (Title: 華嚴一乘法界圖，Subject: 仏典，…)

(2) 文献画像

文献画像は，歴史的文献の画像を管理するための構成要素である．ここで管理する画像を用いて，テキストデータでは表現できないレイアウトや字形などの文献の見た目を計算機上で再現する．本論文では，文献画像の値として画像の URI を用いて間接的に管理する．

例 2. 図 1 が http://foo.ac.jp/hokai.jpg で公開されている場合，このモデルでもその URL を記述する．

文献画像 = http://foo.ac.jp/hokai.jpg

(3) 本文

この構成要素では，文献画像の内容をテキスト化したデータを記述する．

例 3. 図 1 の場合は，その文献画像の内容を記述する．

本文 = 佛爲名動… 性法

(4) 注釈

注釈は，文献画像の内容に関するテキストデータを管理するための構成要素である．注釈は，識別子と注釈文の組の集合である．識別子は，対応表で文献画像の領域と注釈の関係付けに利用される値である．注釈文は，文献の内容理解の支援や検索に利用されるテキストデータである．文献のある語は n 個の注釈が付く，逆に，注釈は関係付けられた語が文書中に出現する回数だけ関係がある．したがって，文献画像と注釈の関係は，一般に多対多の関係になる．

例 4. 図 1 の各文字に対する読みを注釈として記述する例を示す．

注釈={(1,（一文字目 読み “ふつ”）)，(2,（二文字目 読み “い”))，…}

(5) 対応表

対応表は，直接画像に記入できない注釈と文献画像

の領域の関係を管理する構成要素であり文献画像の領域と注釈の識別子の組の集合である．文献画像の領域は，文献画像の座標系を，(1) 画像の左上を原点，(2) 水平，左から右の方向を x 軸正の向き，(3) 垂直に上から下の方向を y 軸正の向きとした場合の長方形の原点に最も近い点と最も遠い点の座標の組で表す．注釈と文献画像の領域を明示的に関係付けることで，文献画像の一部を切り出すと同時に，その領域に関連付けられている注釈を抜き出すことができる．そのほかに文献画像の個々の文字ごとに領域を記述し，関係付ける注釈に文字の出現順を記述すれば，文献画像の読む順序を明示することができる．

例 5. 図 1 の対応表の記述例は，次のとおりである．

対応表={ ((100, 100, 110, 110), 1), ((100, 110, 110, 120), 2), … }

例の領域情報の記述は，原点からの最近点の x 座標値，y 座標値，最遠点の x 座標値，y 座標値の順である．識別子は，先に述べた注釈の識別子である．

3.2 電子スクラップブックデータモデルの構成

電子スクラップブックデータモデルは，文献データの分類を管理するデータモデルである（図 2）．電子スクラップブックデータモデルによって，利用者は個々の文献を閲覧しただけでは発見の困難な文献間の参照関係などを表現できる．このモデルの構成要素は，電子スクラップブック情報とグループからなる（図 2）．以下にそれぞれの構成要素について述べる．

(1) 電子スクラップブック情報

この要素は，電子スクラップブックデータの所有者などのメタデータを記述する．属性には，文献情報と同様に Dublin Core Metadata Element Set Ver.1.1 [10] を用いる．文献データと属性を統一することで，メタデータに関して電子スクラップブックデータと文献データを区別せずに検索できる．

(2) グループ

グループは，0 個以上の文献データが属し，文献データの URI を用いて間接的に管理する．また，グループは，各文献データの URI にあわせて文献画像の配置情報を管理する．配置情報は，グループ単位で文献画像を閲覧するときに利用する．

3.3 文献データにおける外字情報の記述

東アジア圏の歴史的文献は外字を含むことが多いため，標準的な符号化文字集合だけで文献の内容を正確にテキスト化することは困難である．そこで，文献データモデルは，文献画像の内容に対応したテキストデータを記述するために外字の読みや画数などの文字属性を注釈として記述する．外字に関連する情報には，ekanji [11] や今昔文字鏡 [12] などの大規模漢字集合がある．また，このような大規模漢字集合をインターネットを介して交換する手法として “XML による画像参照交換方式” が提案されている [13]．そこで提案モデルでは，注釈を用いてこれら情報と文献画像中の外字を関連付ける．また，本文のテキストデータに含まれる外字は，外字情報を記述した注釈への参照として記述する．

例 6. 図 1 の本文データの 1 文字目が外字であった場合の本文データの記述について考える．まず外字情報は，注釈に,(100, { 読み：フツ，康熙字典コード：452}) と登録する．康熙字典コードとは，ekanji で定義した康熙字典の収録文字を管理するためのコードである．次に，本文に注釈で定義した識別子の値 100 を本文データ内で次のように記述する．

本文=(100) 爲名動 … 性法

このように注釈のデータと外字情報を結び付けることで，文字に関する文献固有の情報と一般的な文字の情報をあわせて記述することができる．

4. 設　　計

本章では，提案モデルに従ったデータを XML 文書として管理するための Relax スキーマの作成，検索のための関係表の設計，提案モデルの編集操作および検索処理について述べる．

4.1 提案モデルの XML による記述

提案モデルに従ったデータは，ネットワークを介して共有されること想定している．そこで本研究では，データの記述にインターネットでの情報交換に広く利用されている XML [8] を用いる．本論文では，提案モデルのスキーマ記述に，DTD に比べ柔軟なスキーマ記述が可能である Relax [14] を用いる．提案モデルから Relax スキーマを作成する手順を以下に示す．

(1) 文献データ，電子スクラップデータに対応したルート XML 要素をそれぞれ作成し，各構成要素をそのルート XML 要素の子とする．

(2) 各構成要素の持つ属性は，次のような方針で XML 要素を生成し，対応する構成要素の XML 要素の子とする．

(a) 本文などの属性を持たない構成要素は，その構成要素に対応した XML 要素自体が値を持つ．

(b) 文献情報のようにその記述できる属性が決まっているものは，その属性名を XML 要素として列挙する．

(c) 注釈などの構成要素の中で定義した属性の組が

複数回登場するものは，定義した属性の組を表す中間 XML 要素を定義し，中間 XML 要素を構成要素に対応した XML 要素の子として追加する．

文献データモデルの Relax スキーマの記述例を付録 A.1 に，電子スクラップブックデータモデルの Relax スキーマの記述例を付録 A.2 に示す．

文献データモデルの Relax スキーマは，文献情報，文献画像，本文，注釈，対応表の 5 つの部分からなる．文献情報の XML 要素は，Title などの 15 の XML 要素を子として持つ．文献画像や本文の XML 要素は，直接値を管理する．注釈と対応表の XML 要素は，それぞれの属性の組が，複数個出現するため，属性の組を保存する XML 要素を持つ．

電子スクラップブックデータモデルの Relax スキーマは，XML 文書のルート要素の子に電子スクラップブック情報とグループを持つ．グループは電子スクラップブックデータの中で 1 つ以上登場することを許す．電子スクラップブック情報は，Title などの 15 の属性に対応した XML 要素を持つ．グループは，文献データへの URI と表示情報の組が複数登場するので，それらの組を保存する XML 要素を子として持つ．

4.2 提案モデルの検索のための関係表の設計

提案モデルに従ったデータは XML 文書として記述するが，これらを検索する場合，XML 文書を個別に調べるより，検索に適した形式に変換した方がよい．そこで，本研究では検索のために XML 文書であるデータから必要な情報を抜き出し，関係データベースで管理する．本研究の検索は注釈や文献情報などの特定の要素に対する文字列一致であるため，データベースに格納する情報は，検索対象の要素名とその値である．検索対象となる要素の表現には，XPath[15] を利用する．提案データモデルを検索するための関係表には，以下のような表が考えられる．

(1) 属性に XML 文書の各要素までのパスとその要素の値を持つ関係表
(2) 表名に XML 文書の各要素までのパスを用い属性にその要素の値を格納する関係表

本研究の検索対象となる要素は，本文，注釈，対応表など値に文字列だけを持つ特定の要素なので，後者の方針に従って関係表を作成する．たとえば，表名 “/文献データモデル/文献情報/Title”，属性に “URI” と “Value” を持つ表が考えられる．表の属性 “URI” には，文献データの URI を記述し，属性 “Value” には文献データの要素 “Title” の値を記述する．

文献データの検索は，1 つの要素だけを対象にするのではなく，複数の要素を対象にすることの方が多いと考えられる．したがって，要素ごとに関係表を作成した場合，検索条件によっては要素ごとに JOIN 操作が必要となり，効率的な検索処理が期待できない．そこで，先に定義した表を統合して JOIN 操作を減らすことを考える．ここでは，各表名であるパスの最長一致を行い，その結果である共通部分を用いて新たな表を作成する．生成された表の属性は，パスの非共通部分の要素名と各データの URI を持つ．たとえば，文献データの文献情報の場合，文献情報の各要素は共通パスとして “/文献データモデル/文献情報” を持ち，非共通分として “Title” や “Subject” を持つ．したがって，表名 “/文献データモデル/文献情報” であり，属性に “Title” や “Subject” などの 15 の属性と文献データの URI を持つ表が作成される．

4.3 処　　理

4.3.1 編 集 処 理

(1) 注釈の編集

文献データの注釈の編集として挿入と削除について述べる．注釈の挿入は，入力に文献画像の領域と注釈文を与え，(a) 対応表と注釈にそれぞれの値を登録するための属性を追加し値を挿入する，(b) 追加した注釈文に対し識別子を与える，(c) 識別子を対応表に追加することによって処理する．注釈の削除は，関係する対応表と注釈の各要素を削除である．この場合，入力として注釈の識別子を与える．ただし，領域情報に複数の注釈が関連付けられている場合は，対応表の領域情報は削除しない．

(2) 文献データの切り抜き

切り抜き操作は，文献収集を支援するための操作であり，この操作で得られるデータを切り抜きデータと呼ぶ．切り抜きデータは，元の文献データの URI を持つので，2 つのデータ間の参照関係は記録される．この処理は，文献データから指定された情報を抽出する処理と，その情報に基づいて新たな文献データを生成する処理からなる．

(a) 情報の抽出

入力として文献画像から切り抜く領域を指定する．この領域に従って情報を文献データから，(i) 文献画像から取り出した部分画像，(ii) 対応表から指定された領域に内包される領域情報と注釈の識別子，(iii) 先に取り出された識別子を持つ注釈文を取り出す．

(b) 文献データの生成

空の文献データを作成し，先の操作で得た情報を文献データに登録する．切り抜き元の文献データの

URI を切り抜きデータの文献情報の属性 Refrence に登録する．

この操作には，切り抜き後の元の文献データの編集結果を切り抜きデータに反映させないか反映させるかの違いから，静的な切り抜きと動的な切り抜きが考えられる．これらの処理は，それぞれ切り抜きデータを物理的に別のデータとして作成する処理と，データの閲覧の度に新たな切り抜きデータを生成する処理によって実現できる．

(3) 分類操作

ここでは，分類操作として文献データの分類を管理するグループの分割処理と結合処理について述べる．

(a) グループの分割

この処理では，入力として分割対象のグループと，移動させる文献データの URI の集合が与えられる．分割処理は，空のグループを電子スクラップブックデータに追加したうえで，指定された文献データの URI の集合を移動させる処理である．

(b) グループの結合

この処理は，入力として与えれた結合元のグループと結合先のグループの組に対して，(i) 結合元のグループに登録されている文献データのすべて結合先のグループに移動させ，(ii) 空になった結合元のグループを削除する処理である．

4.3.2 検索処理

提案モデルに従ったデータに対する検索処理について述べる．文献データの検索は，文献データの注釈などのテキストデータを対象にした文字列一致である．

(1) 検索条件として検索対象の要素と検索キーである文字列の組の集合を与える．
(2) 各組で指定された属性の値と検索キーが一致するかどうかを検査する．この検査は，データベースへの問合せという形式で処理する．問合せ処理は次項で述べる．一致するデータがあれば，そのデータの URI を一時的な検索結果して取り出す．
(3) すべての組に対して (2) を繰り返し，各検査結果で得られる URI を集計する．
(4) 集計結果で上位になった URI を文献データを検索結果として利用者に返す．

次に，外字を含むテキストデータの文字列一致について述べる．ここでは文献データの本文を対象に考える．この処理で与える文字列は，直接，読みなどの外字情報を記述したものを考える．

(1) 与えられた文字列から外字情報部分を抜き取る．
(2) 残った部分による部分一致を行い一致する文献データを取り出す．また，本文の値から一致した部分を抜き出す．
(3) 取り出した文献データの注釈要素に対して，(1) で抜き出した外字情報を用いた検査を行う．
(4) 外字情報と一致したものがあった場合，一致した注釈の識別子を取り出す．
(5) 最後に，(4) で取り出した識別子が，(2) で取り出した文字列に含まれるかを検査し，含まれているものを検索結果として取り出す．

最後に，電子スクラップブックデータの検索について述べる．電子スクラップブックデータは，文献データを URI によって間接的に管理しているため，直接，文献データを含む電子スクラップブックデータを検索することはできない．そこで，電子スクラップブックデータの検索は，次のように処理する．

(1) 文献データの検索を行い，該当する文献データの URI を取り出す．
(2) 取り出した文献データの URI を検索キーとして，電子スクラップブックデータのグループの検索を行う．
(3) 検索条件を満たした電子スクラップブックデータを結果として取り出す．

4.3.3 問合せ処理

文献データの検索で関係データベースを利用するために検索条件から SQL 文を生成する．検索条件から SQL を生成するために取り出す値は，検索対象の要素名と，その要素の値と文字列一致をするための文字列である．問合せの結果は，条件を満たした要素を含むデータの URI の集合である．与えられた要素名は，関係の表名と属性を連結したものである．本論文では，検索条件に要素名から得られる表名が同じであり属性名の異なる条件が複数あれば，AND 検索として，以下のような SQL 文を生成する．

```
SELECT Identifier
FROM 表名
WHERE 属性名 1 LIKE ' 値 1' AND
属性名 2 LIKE ' 値 2' AND ...
```

上記の SQL 文の表名には，与えられた要素のパスとデータベースにある表名を比較し一致した部分を用いる．一方，属性名は一致しなかった部分を用いる．SELECT 句の Identifier は，文献情報の属性 Identifier の外部キーであり，文献データの URI である．

検索条件に含まれる対象の要素の中に同じ要素を指定したものが複数ある場合は，OR 検索として処理する．本論文の検索では，問合せで得られる URI を集

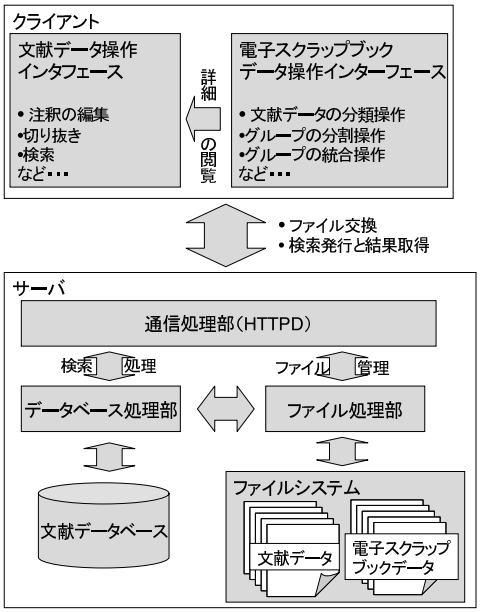

図 3 プロトタイプシステムの構成
Fig. 3 Prototype system architecture.

計し，その集計に基づいて利用者に示す検索結果を作成するので，1 つの SQL 文にまとめるのではなく，要素の個数だけ SQL 文を生成する．生成する SQL 文の構造は先に示したものを利用する．

5. プロトタイプシステム

5.1 プロトタイプシステム構成

プロトタイプシステムは，文献データの編集や閲覧を支援するクライアントと，文献データと電子スクラップブックデータを利用者間で共有するためのサーバからなる（**図 3**）．

(1) クライアント

クライアントは，文献データ操作インタフェースと電子スクラップブックデータ操作インタフェースからなる．前者は，利用者による文献データの閲覧，切り抜き，注釈の編集を支援するインタフェースであり，後者は，利用者による電子スクラップブックデータの閲覧や分類を支援するインタフェースである．

(2) サーバ

サーバは，(a) クライアントとの通信を処理する通信処理部，(b) 文献データなどのデータベースへの登録と検索処理をするデータベース処理部，(c) 文献データなどのファイルの入出力を管理するファイル処理部，(d) 効率的に検索を処理するための文献データベースからなる．文献データベースは，一般に広く利用されている関係データベースを用いる．

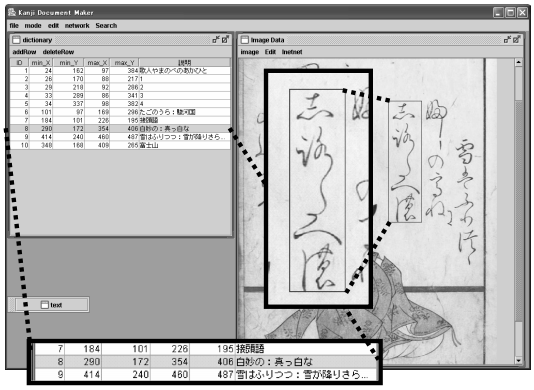

図 4 文献データの閲覧例
Fig. 4 Example of document image data.

5.2 プロトタイプシステムの実装環境

プロトタイプシステムの実装には，Java1.4 を使用した．クライアントとサーバ間の通信処理には，Apache2.0 と Tomcat4.0 を利用し，関係データベースには MySQL4.0.4 を使用した．データベース処理部やファイル処理部は Servlet として実装し必要に応じて通信処理部から実行される．

プロトタイプシステムのクライアントとサーバは，同一の PC（OS：WindowsXP，CPU：Intel Pentium III 844 MHz，メモリ：512 Mbyte，ハードディスク：30 Gbyte）上で実装している．ただし，クライアントとサーバは別のプロセスで実行し，PC 内部で HTTP による通信をしているので，クライアントとサーバを異なる計算機上で実行することも可能である．プロトタイプシステムは，文献画像として玉川大学図書館で公開されている百人一首を利用した[16]．

5.3 実 行 例

5.3.1 データの閲覧

図 4 は文献データ操作インタフェースを用いた百人一首の文献データの表示例である．このインタフェースは，文献画像，書誌情報などを表示するウィンドウ群からなる．対応表と注釈は，注釈の識別子に基づいて 1 つの表にまとめている．このインタフェースには，文献閲覧モードと注釈編集モードがある．前者は，文献データの注釈編集を禁止し，切り抜き操作を許可するモードである．後者は，前者の逆のモードである．図 4 は，文献データの対応表の利用例として選択した注釈に関連付けられている文献画像の領域を視覚化している．

図 5 は，電子スクラップブック操作インタフェースの実行例として，百人一首から切り抜いた著者名を並べて表示している．図に示す電子スクラップブック

図 5 電子スクラップブックデータの閲覧例
Fig. 5 Example of electronic scrapbook data.

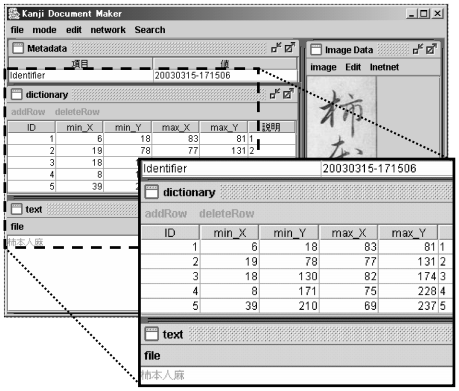

図 6 電子スクラップブックデータ内の文献データの詳細の表示例
Fig. 6 Details of document image data from electronic scrapbook data.

データは，1 つのグループに 3 つの文献データが登録されており，インタフェースの下部にそのデータの URI を示し，上部に対応する各文献画像を表示している．このインタフェースでは，グループの追加，削除，文献データの移動，削除の機能を実装している．電子スクラップブックデータに登録された文献データの詳細の閲覧には，文献データ操作インタフェースを利用する（**図 6**）．図 6 で示している文献データは，ある百人一首からの切り抜きデータである．図 6 の表部分は，文字の出現を記した注釈とそれに対応する文献画像の領域を示している．

利用者はこれらのインタフェースを用いて作成した文献データや電子スクラップブックデータをクライアントもしくはサーバのどちらかで保存する．サーバ側で保存した場合は，ネットワーク上にデータが公開されたものとして，文献データベースに登録する．

5.3.2 検　　索

本項では，プロトタイプシステムに実装した文献データの検索例を示す．**図 7** は検索インタフェースであり，文献データ操作インタフェースの一部として実装した．このインタフェースを用いて，利用者は，検索対象の要素と検索キーとなる文字列を指定する．図 7 は，“歌人” もしくは “鳥” に関する注釈を含む文献データの検索例である．文字列に含まれる “%” は 0 文字以上の任意の文字列と一致することを表す．作成した検索条件は，文献データモデルに従った XML 文書としてサーバに送信される．このインタフェースによって，利用者は，文献画像の注釈編集と同様の操作で検索条件を記述できる．

図 7 検索インタフェースの実行例
Fig. 7 Example of search for document image data.

図 7 で与えられた検索条件は，同じ要素に対して 2 つの条件が示されているので，サーバでは次のような 2 つの SQL 文を生成する．問合せの結果は，与えられた文字列を注釈に含む文献データの URI の集合である．

(1) SELECT Identifier FROM ipan WHERE body LIKE ‘歌人%’;
(2) SELECT Identifier FROM ipan WHERE body LIKE ‘%鳥%’;

FROM 句の表名は，ルート要素からのパスを記述するが，記述の簡略化ために，付録 A.1 で定義した要素名だけを記述している．

サーバからクライアントに送信する検索結果の形式は，電子スクラップブックデータモデルに従う．**図 8** は，図 7 の検索結果を電子スクラップブック操作インタフェースで表示した例であり，検索結果である文献データの URI が 1 つのグループに登録されている．検索結果は，新たな電子スクラップブックデータであるので，利用者による分類や文献データの詳細の閲覧が可能である．文献データの詳細の閲覧は，文献データ操作インタフェースを用いるため，さらに注釈の追加や切り抜きなどの編集が可能である．

6. 関 連 研 究

インターネットを介した電子文書の交換に広く利

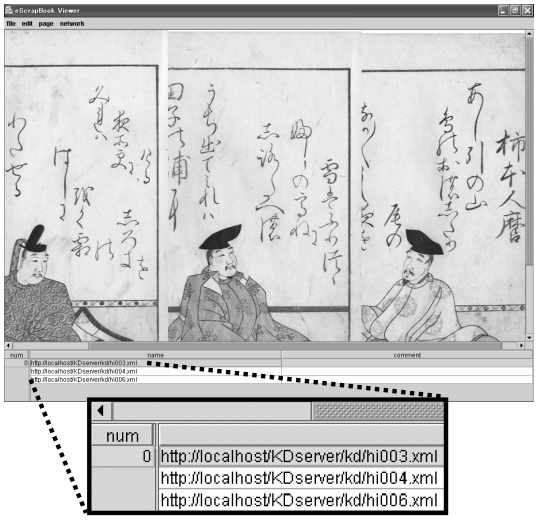

図 8 検索結果の表示例
Fig. 8 Example of search result.

用されているフォーマットとして Adobe 社の PDF（Portable Document Format）がある[17]．PDF は印刷に適した電子文書交換することを目的にしたフォーマットであり，文書交換のほかに，注釈の添付，校正などのためのマークの記入などが可能である．しかし，PDF 文書の切り抜き操作は，ページ単位でしかできず，ページ内の一部分だけを取り出す場合は画像もしくは文字列に変換しなければならない．一方，提案モデルは，PDF とは異なり文献データの任意の部分の切り抜きをすることができる．また，PDF が独自形式のデータモデルに対して，提案モデルに従ったデータは XML を用いて記述するので，HTML などの他の形式への変換が容易であると考えられる．

XML を利用した電子書籍モデルとして Open eBook がある[18]．このモデルは，インターネットを介した書籍の配布と閲覧を支援するためのモデルであり，一般的なビューワなどのアプリケーションは，著作権の保護のために文書の切り抜き操作ができない．一方，本論文は，利用者による情報の収集と整理の支援を目的としている．

7. ま と め

本論文では，インターネット上の東アジア圏の歴史的文献画像と関連する書誌情報や注釈などを，利用者間で集約し共有するためのデータモデルを提案し，プロトタイプシステムを実装した．提案したデータモデルは，文献データモデルと電子スクラップデータモデルからなる．前者は，文献画像と注釈などのテキストを関連付けて管理する．後者は，互いに関連する文献データの分類を管理する．本論文では，提案モデルに従ったデータの記述に XML を利用するための Realx スキーマの設計について述べた．XML を利用することでシステムに依存しない形式でデータが記述されるため利用者間のデータ共有に役立つと考えられる．関係データベースを利用した検索処理を実現するための関係表の設計について述べた．提案モデルの操作として，編集，切り抜き，分類，検索について述べた．編集操作は，歴史的文献に関連する情報の収集を支援するための操作であり，文献画像に対する注釈の追加，削除がある．切り抜き操作は，文献データの一部を利用して新たな文献データを作成する操作であり，分類操作は，利用者が収集した文献データを分類する操作である．検索は，文献データや電子スクラップブックデータを複数の利用者間で共有するための機能である．文献データの検索は，関係データベースを利用するので検索条件から SQL の生成について述べた．本論文では，こうしたモデルに基づいたプロトタイプシステムを実装し，文献データの閲覧などの実行例と検索処理の実行例から提案モデルの有用性を示した．提案モデルを用いれば，分散して公開されている歴史的文献に関する情報を，利用者が独自に集約することが可能となり，またこれを利用者間で共有することで，歴史的文献の相互関連の発見などが容易になると考えられる．

今後の課題として，外字を含む文字列に対する検索処理，複数利用者による書き込み制御などの共有する情報の更新機能，さらに，文献画像の特徴を考慮した柔軟な検索などがある．

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) 独立行政法人国立公文書館アジア歴史資料センター：アジア歴史資料センター (2001). http://www.jacar.go.jp/
2) 奈良文化財研究所：木簡データベース. http://acd.nabunken.jp/Open/mokkan/mokkan1.html
3) Academia Shinica Computing Centre: 漢籍電子文献. http://www.sinica.edu.tw/ftms-bin/ftmsw3
4) Johnson, I. and Osmakov, A.: Time Map. http://ecai.org/tech/timemap.html
5) 桶谷猪久夫，才藤千津子，Brown, D.：簡易型タグを利用した歴史史料の英日全文連携検索システムの設計と開発—古事記，日本書紀における事例，

じんもんこん 2001，人文科学とコンピュータシンポジウム，Vol.2001, No.18, pp.65–72 (2001).
6) 柴山 守，吉井良邦，ベンガッシュ・ラガワンほか：近世史料アーカイブズのためのバーチャル図書館，じんもんこん 2001，人文科学とコンピュータシンポジウム，Vol.2001, No.18, pp.109–116 (2001).
7) Ishikawa, M., Hatano, K., Amagasa, T., et al.: A Data Model for Reconstructable Kanji Documents Using XML, IASTED International Conference on Information Systems and Databases (ISDB 2002), Tokyo, Japan, pp.258–263 (Sep.25–27, 2002).
8) Bray, T., Paoli, J., Sperberg-McQueen, C.M., et al.: Extensible Markup Language (XML) 1.0, 2nd Edition (Oct. 2002). http://www.w3.org/TR/2000/REC-xml-20001006
9) 師 茂樹：電子テキスト概論，電脳中国学，漢字文献情報処理研究会，pp.196–204, 好文出版 (1998).
10) Dublin Core Metadata Initiative: Dublin Core Metadata Element Set, Version 1.1: Reference Description (1999). http://dublincore.org/documents/1999/07/02/dces/
11) 勝村哲也，丹羽正之：eKanji (2000). http://nohara.u-shimane.ac.jp/ekanji/
12) 文字鏡研究会：今昔文字鏡. http://www.mojikyo.org/html/index.html
13) 川俣 晶：XML による画像参照交換方式，日本工業規格協会，JIS-TRX0045 (May 2001).
14) Murata, M.: RELAX (Regular Language description for XML) (May 2001). http://www.xml.gr.jp/relax/
15) Berglund, A., Boag, S., Chamberlin, D., et al.: XML Path Language (XPath) 2.0 (2001). http://www.w3.org/TR/xpath20/
16) 玉川大学図書館：百人一首，漢籍・和装丁本コレクション (2000). http://www.tamagawa.ac.jp/sisetu/tosyo/w_index.htm
17) Adobe Systems Incorporated: PDF Reference, third edition, Adobe Portable Document Format Version 1.4 (1999). http://partners.adobe.com/asn/developer/acrosdk/docs/filefmtspecs/PDFReference.pdf
18) The Open eBook Forum: The OeBF Publication Structure 1.0.1 Recommended Specification. http://www.openebook.org/oebps/oebps1.0.1/download/

付 録

A.1 Relax スキーマによる文献データモデルの記述

文献データモデルに対応した Relax スキーマの記述例を以下に示す．行 6〜8 は，文献データのルート XML 要素である．行 10〜19 は，提案モデルの 5 つの構成要素のルートの定義であり，bib は文献情報，im は文献画像，te は本文，an は注釈，ct は対応表を表す．行 21〜70 は，文献情報の属性であり，XML 要素 bib の子として定義する．各属性の属性名は，Dublin Core Metadeta Element Set Ver. 1.1 に定義された名前を用いた．行 72〜73 は，文献画像の定義であり，本論文ではこの XML 要素の値として文献画像の URI を記述する．行 75〜76 は，本文の定義であり，この XML 要素の値は文献画像の内容をテキストデータとして記述する．行 78〜92 は，文献データモデルの注釈を XML 要素で定義した例である．注釈は，注釈の識別子（行 89）と注釈の本体（行 91）の組の集合であるので，Relax スキーマでは，識別子と本文の組を保存するため XML 要素 row（行 83）を定義した．行 94〜108 は対応表の定義である．対応表は文献画像の領域 imp（行 101）と注釈の識別 tp（行 102）の組の集合であり，これらの組を保存するために XML 要素 ci を定義した．文献画像の領域の記述は，各頂点の座標値を文字列として記述する．

```
1:  <?xml version="1.0" encoding="UTF-8"?>
2:  <module
3:  moduleVersion="1.0"
4:  relaxCoreVersion="1.0"
5:  xmlns="http://www.xml.gr.jp/xmlns/relaxCore">
6:  <interface>
7:  <export label="ip"/>
8:  </interface>
9:  <!-- 文献データモデルのルート -->
10:  <tag name="ip"/>
11:  <elementRule role="ip">
12:  <sequence>
13:  <ref label="bib"/>
14:  <ref label="im"/>
15:  <ref label="te"/>
16:  <ref label="an"/>
17:  <ref label="ct"/>
18:  </sequence>
19:  </elementRule>
20:  <!-- 文献情報 -->
```

```
21:  <tag name="bib"/>
22:  <elementRule role="bib">
23:  <sequence>
24:  <ref label="Title"/>
25:  <ref label="Creator"/>
26:  <ref label="Subject"/>
27:  <ref label="Description"/>
28:  <ref label="Publisher"/>
29:  <ref label="Contributor"/>
30:  <ref label="Date"/>
31:  <ref label="Type"/>
32:  <ref label="Format"/>
33:  <ref label="Identifier"/>
34:  <ref label="Source"/>
35:  <ref label="Language"/>
36:  <ref label="Relation"/>
37:  <ref label="Coverage"/>
38:  <ref label="Rights"/>
39:  </sequence>
40:  </elementRule>
41:  <tag name="Title"/>
42:  <elementRule role="Title" type="string"/>
43:  <tag name="Creator"/>
44:  <elementRule role="Creator" type="string"/>
45:  <tag name="Subject"/>
46:  <elementRule role="Subject" type="string"/>
47:  <tag name="Description"/>
48:  <elementRule role="Description" type="string"/>
49:  <tag name="Publisher"/>
50:  <elementRule role="Publisher" type="string"/>
51:  <tag name="Contributor"/>
52:  <elementRule role="Contributor" type="string"/>
53:  <tag name="Date"/>
54:  <elementRule role="Date" type="string"/>
55:  <tag name="Type"/>
56:  <elementRule role="Type" type="string"/>
57:  <tag name="Format"/>
58:  <elementRule role="Format" type="string"/>
59:  <tag name="Identifier"/>
60:  <elementRule role="Identifier" type="string"/>
61:  <tag name="Source"/>
62:  <elementRule role="Source" type="string"/>
63:  <tag name="Language"/>
64:  <elementRule role="Language" type="string"/>
65:  <tag name="Relation"/>
66:  <elementRule role="Relation" type="string"/>
67:  <tag name="Coverage"/>
68:  <elementRule role="Coverage" type="string"/>
69:  <tag name="Rights"/>
70:  <elementRule role="Rights" type="string"/>
71:  <!-- 文献画像 -->
72:  <tag name="im"/>
73:  <elementRule role="im" type="string"/>
74:  <!-- 本文 -->
75:  <tag name="te"/>
76:  <elementRule role="te" type="string"/>
77:  <!-- 注釈 -->
78:  <tag name="an"/>
79:  <elementRule role="an">
80:  <ref label="row" occurs="+"/>
81:  </elementRule>
82:  <tag name="row"/>
83:  <elementRule role="row">
84:  <sequence>
85:  <ref label="id"/>
86:  <ref label="body"/>
87:  </sequence>
88:  </elementRule>
89:  <tag name="id"/>
90:  <elementRule role="id" type="string"/>
91:  <tag name="body"/>
92:  <elementRule role="body" type="string"/>
93:  <!-- 対応表 -->
94:  <tag name="ct"/>
95:  <elementRule role="ct">
96:  <ref label="ci" occurs="+"/>
97:  </elementRule>
98:  <tag name="ci"/>
99:  <elementRule role="ci">
100:  <sequence>
101:  <ref label="imp"/>
102:  <ref label="tp"/>
103:  </sequence>
104:  </elementRule>
105:  <tag name="imp"/>
106:  <elementRule role="imp" type="string"/>
107:  <tag name="tp"/>
108:  <elementRule role="tp" type="string"/>
109:  </module>
```

A.2 Relax スキーマによる電子スクラップブックデータモデルの記述

電子スクラップブックデータモデルを Relax スキー

マで定義した記述例を以下に示す．行 6〜8 は，電子スクラップブックデータモデルのルート XML 要素を定義である．行 10〜16 は，電子スクラップブックデータの構成要素である電子スクラップブック情報（行 13）と，グループ（行 14）の定義である．グループは，電子スクラップブックデータ内に 1 つ以上ある．行 18〜67 は，電子スクラップブック情報の属性に対応する XML 要素を定義している．属性名は，Dublin Core Metadata Element Set Ver. 1.1 に従う．行 69〜86 は，グループの属性に対応した XML 要素の定義である．グループの属性は，文献データの識別子を記述する XML 要素 ref（行 78）と複数の文献画像を一度に閲覧するための配置情報 x，y（行 76〜77）の組の集合である．この組を保存するために，XML 要素 grow（行 71）を定義した．

```
1:  <?xml version="1.0" encoding="UTF-8"?>
2:  <module
3:  moduleVersion="1.0"
4:  relaxCoreVersion="1.0"
5:  xmlns="http://www.xml.gr.jp/xmlns/relaxCore">
6:  <interface>
7:  <export label="gp"/>
8:  </interface>
9:  <!-- 電子スクラップブックデータモデルのルート -->
10:  <tag name="gp"/>
11:  <elementRule role="gp">
12:  <sequence>
13:  <ref label="cp"/>
14:  <ref label="sg" occurs="+"/>
15:  </sequence>
16:  </elementRule>
17:  <!-- 電子スクラップブックデータ情報 -->
18:  <tag name="cp"/>
19:  <elementRule role="cp">
20:  <sequence>
21:  <ref label="Title"/>
22:  <ref label="Creator"/>
23:  <ref label="Subject"/>
24:  <ref label="Description"/>
25:  <ref label="Publisher"/>
26:  <ref label="Contributor"/>
27:  <ref label="Date"/>
28:  <ref label="Type"/>
29:  <ref label="Format"/>
30:  <ref label="Identifier"/>
31:  <ref label="Source"/>
32:  <ref label="Language"/>
33:  <ref label="Relation"/>
34:  <ref label="Coverage"/>
35:  <ref label="Rights"/>
36:  </sequence>
37:  </elementRule>
38:  <tag name="Title"/>
39:  <elementRule role="Title" type="string"/>
40:  <tag name="Creator"/>
41:  <elementRule role="Creator" type="string"/>
42:  <tag name="Subject"/>
43:  <elementRule role="Subject" type="string"/>
44:  <tag name="Description"/>
45:  <elementRule role="Description" type="string"/>
46:  <tag name="Publisher"/>
47:  <elementRule role="Publisher" type="string"/>
48:  <tag name="Contributor"/>
49:  <elementRule role="Contributor" type="string"/>
50:  <tag name="Date"/>
51:  <elementRule role="Date" type="string"/>
52:  <tag name="Type"/>
53:  <elementRule role="Type" type="string"/>
54:  <tag name="Format"/>
55:  <elementRule role="Format" type="string"/>
56:  <tag name="Identifier"/>
57:  <elementRule role="Identifier" type="string"/>
58:  <tag name="Source"/>
59:  <elementRule role="Source" type="string"/>
60:  <tag name="Language"/>
61:  <elementRule role="Language" type="string"/>
62:  <tag name="Relation"/>
63:  <elementRule role="Relation" type="string"/>
64:  <tag name="Coverage"/>
65:  <elementRule role="Coverage" type="string"/>
66:  <tag name="Rights"/>
67:  <elementRule role="Rights" type="string"/>
68:  <!-- グループ -->
69:  <tag name="sg"/>
70:  <elementRule role="sg">
71:  <ref label="grow" occurs="+"/>
72:  </elementRule>
73:  <tag name="grow"/>
74:  <elementRule role="grow">
75:  <sequence>
76:  <ref label="x"/>
77:  <ref label="y"/>
```

```
78:  <ref label="ref"/>
79:  </sequence>
80:  </elementRule>
81:  <tag name="ref"/>
82:  <elementRule role="ref" type="string"/>
83:  <tag name="x"/>
84:  <elementRule role="x" type="string"/>
85:  <tag name="y"/>
86:  <elementRule role="y" type="string"/>
87:  </module>
```

(平成 15 年 3 月 25 日受付)
(平成 15 年 7 月 17 日採録)

（担当編集委員　加藤 俊一）

石川　正敏（正会員）

2000 年奈良先端科学技術大学院大学情報科学研究科博士後期課程単位取得退学．同年島根県立大学総合政策学部助手を経て，2003 年から同大学総合政策学部非常勤講師，北東アジア地域研究センター客員研究員．データベースシステムの研究に従事．電子情報通信学会，ACM，IEEE，日本データベース学会各会員．

波多野賢治（正会員）

1999 年神戸大学大学院自然科学研究科博士後期課程修了．同年から奈良先端科学技術大学院大学情報科学研究科助手．情報検索システム，データベースシステムの研究に従事．博士（工学）．電子情報通信学会，ACM，IEEE Computer Society，日本データベース学会各会員．

天笠　俊之（正会員）

1999 年群馬大学大学院工学研究科修了．同年から奈良先端科学技術大学院大学情報科学研究科助手．XML データベース，装着型コンピュータにおけるデータベース応用等の研究に従事．博士（工学）．電子情報通信学会，ACM，IEEE Computer Society，日本データベース学会各会員．

植村　俊亮（正会員）

1966 年京都大学大学院工学研究科修士課程修了．同年電気試験所（産業技術総合研究所）．1970 年マサチューセッツ工科大学電子システム研究所客員研究員，1981 年ソフトウェア部プログラム研究室長，1988 年東京農工大学教授を経て，1993 年から奈良先端科学技術大学院大学情報科学研究科教授．データ工学，データベースシステムの研究に従事．博士（工学）．情報処理学会フェロー，電子情報通信学会フェロー，IEEE Fellow．現在，情報処理学会理事，日本情報考古学会理事，データベース振興センター評議員等．

勝村　哲也

京都大学大学院文学研究科博士課程単位取得退学．京都大学人文科学研究所教授を経て，2000 年より島根県立大学総合政策学部教授．同大学メディアセンター長，2002 年北東アジア研究科長，北東アジア地域研究センター長を併任．東洋史学，中国文献学，漢字情報処理等の研究に従事．修士（文学）．京都大学名誉教授，京都大学人文科学研究所名誉所員，日本歴史学協会常任委員，仏教史学会評議員等．