2007－CVIM－157 (8)
2007／1／11

# ディジタル図書館のための情報アクセス基盤の構築

前田 亮

立命館大学 情報理工学部

本稿では，立命館大学21世紀COEプログラム「京都アート・エンタテインメント創成研究」のサブプロジェクトとして著者らが行っている「京都学ディジタル図書館の構築」プロジェクトの概要について述べる．本サブプロジェクトでは，平安時代の貴族の日記である『兵範記』のディジタル図書館を構築している．また，この成果を基に，伝統的モンゴル文字で書かれた古文書のディジタル図書館の構築を開始している．さらに，これらを含む人文科学に関するデータベースのメタサーチを実現する手法について研究を行っている．本稿では，これらのプロジェクトにおけるこれまでの研究の進捗状況および今後の課題について述べる．

# Information Access Infrastructure for Digital Libraries

Akira Maeda

College of Information Science and Engineering, Ritsumeikan University

In this paper, we describe about a research project called “Building a Digital Library of Kyoto Studies”, which is a sub-project of the Ritsumeikan University the 21st century COE program “Kyoto Art Entertainment Innovation Research”. In this sub-project, we are developing a digital library of “Hyohanki”, which is a diary written by an aristocrat during the late Heian era (1132-1184). In addition, we recently started developing a digital library of traditional Mongolian documents based on this research. Moreover, we are investigating techniques for realizing a meta-search of the databases related to the humanities, including “Hyohanki” and traditional Mongolian documents. This paper describes the current developments and future challenges of this research project.

## 1. はじめに

近年，ディジタル図書館やディジタルアーカイブが注目され，さまざまな文化的資料のディジタル化や保存に関する研究が盛んに行われている．しかしながら，それらのコンテンツに対して容易で効率的なアクセス手段を提供するという観点からの研究はまだ多くはない．コンテンツの量が膨大になればなるほど高度なアクセス手段が要求されることは，現在のWebの状況を見ても明らかである．本稿では，筆者が21世紀COEプログラムにおいて進めている，文化的資料のディジタルコンテンツに対して高度な情報アクセスを実現するための研究プロジェクトの概要について述べる．

本研究プロジェクトでは，主に「京都学ディジタル図書館システム」の開発を行っている[1]．本システムでは，平安時代の貴族の日記である

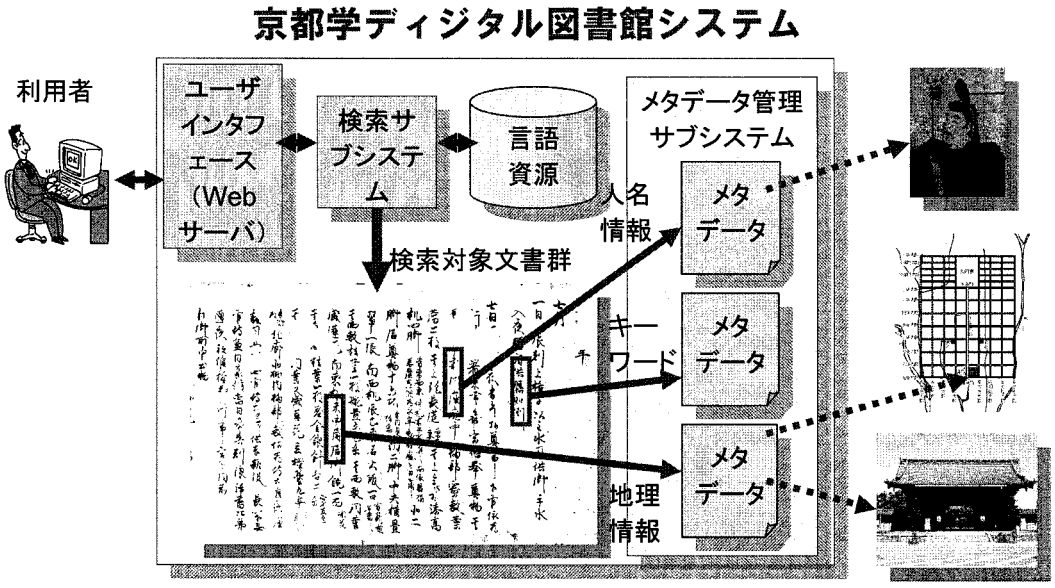


図 1:京都学ディジタル図書館システムの概要

『兵範記』を例として，単なる文字列マッチングではなく，文書全体あるいは単語単位，さらには文字単位で意味を解析することにより，現代語によって検索する機能や，現在の文字コードに含まれない文字を含む文書を検索する機能を実現する．また，本文中に様々な表現で現れる人名・地名・建造物名などの自動抽出，および本文中に現れる単語の語義の推定を行う．さらに，京都歴史地図のGIS（地理情報システム）と連携することで，歴史都市京都に関するさまざまな研究成果を統合する[2].

また，本研究プロジェクトにおけるこれまでの研究成果を基に，昨年より伝統的モンゴル語のディジタル図書館の構築を開始した[3][4][5].

また，本研究プロジェクトの成果を広く世界に向けて発信するために，コンテンツの翻訳版を用意することなく検索を可能とする言語横断情報検索技術[6]，さらにこの技術を応用した時代・文化横断型検索技術についても研究を行っている.

一方，前述の京都学ディジタル図書館システムと，本学および他の研究機関で公開されている人文系データベースを，標準的なメタデータ記述項目および情報検索プロトコルを用いることで，統一的に検索する手法について研究を行っている.

本稿では，これらの研究プロジェクトの進捗状況と今後の課題について述べ，ディジタル化された文化的資料への情報アクセスに関わる現状の問題点およびその解決の見通しについて考察する.

## 2. 京都学ディジタル図書館

前節で述べたように，近年さまざまな文化的資料のディジタル化が急速に普及し，古文書や古記録などの古典史料についてもディジタル保存やデータベース化が進められている．膨大かつさまざまな種類・分野におよぶコンテンツに対して，容易で効率的なアクセス手段を開発することが重要な課題となっている.

京都学ディジタル図書館は，京都に関するさまざまな文化的資料をディジタル化して，その情報を広く世界に向けて発信するシステム

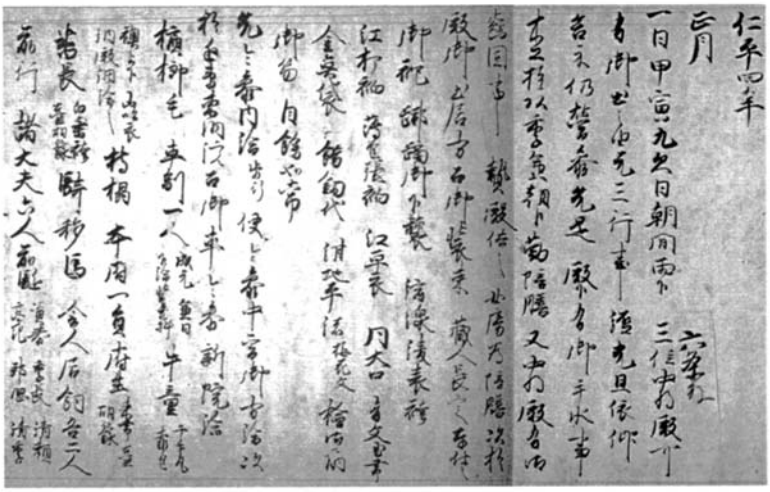

図 2:『兵範記』影印本の一部

の構築を目指している．その際，重要な課題の一つが，膨大なコンテンツに対して容易で効率的なアクセス手段を開発すること，なかでも古典史料を検索する技術開発である．現在開発している京都学ディジタル図書館の概要を図1に示す．

本研究では，古記録・古文書の文字情報を対象として，高度な情報アクセスを実現する手法の確立を検討している．即ち，古記録に対して単なる文字列マッチングではなく，文書全体あるいは単語単位，さらには文字単位で意味を解析することにより，たとえば現代語による検索や，人名や地名・建造物名などを自動抽出し，関連情報へリンクする機能などを提供する．また，将来的には，これらを基に『兵範記』に関するオントロジを構築することを目指している．

さらに，古記録には現在の文字コードに含まれない文字が多く含まれるが，これら「外字」を含む電子化史料に対して効率的な検索を行う手法，人名・地名・建造物名などにおける同一物の複数表記（呼称）に対する効率的な抽出方法，史料文言に対する概念検索システムの開発が必要である．具体的には，『兵範記』を素材として，これらの技術開発に向けた基礎的な研究を行っている．現在は，全文検索システムOpenText7を用いて兵範記本文のデータベースを作成し，上記抽出方法や，利用しやすい検索表示方法としてKWIC（KeyWord In Context）表示の実装などを行った．今後，電子図書館システムInfoLibとの連携によって，メタデータ検索を可能にする．これにより，京都歴史地図のGISをはじめとする歴史的空間情報に関する様々なコンテンツとのリンクが可能になる．

## 3. 『兵範記』検索システム

### 3.1. 『兵範記』とは

兵範記（「へいはんき」もしくは「ひょうはんき」と読む）は，平安時代後期の長承二年（1132）から元暦元年（1184）までの間，平信範が記した日記であり，54巻が現存する．『人車記』『平洞記』『北隣記』などとも呼ばれる．平信範は，朝廷実務の要職である蔵人・弁官を長期間勤め，鳥羽・後白河院の院司，また摂関家累代の家司（家政機関職員）としても活動した人物である．彼の中級貴族・実務官僚という立場に基づき，院政期の行政，たとえば政策決定にいたる推移や行政文書の写し，要人の見解などの情報と，公家有職，たとえば朝廷・院・摂関家に関する儀式次第などに関する精確・詳細な記述が見られる．『兵範記』影印本[7]の一部を図 2に示す．

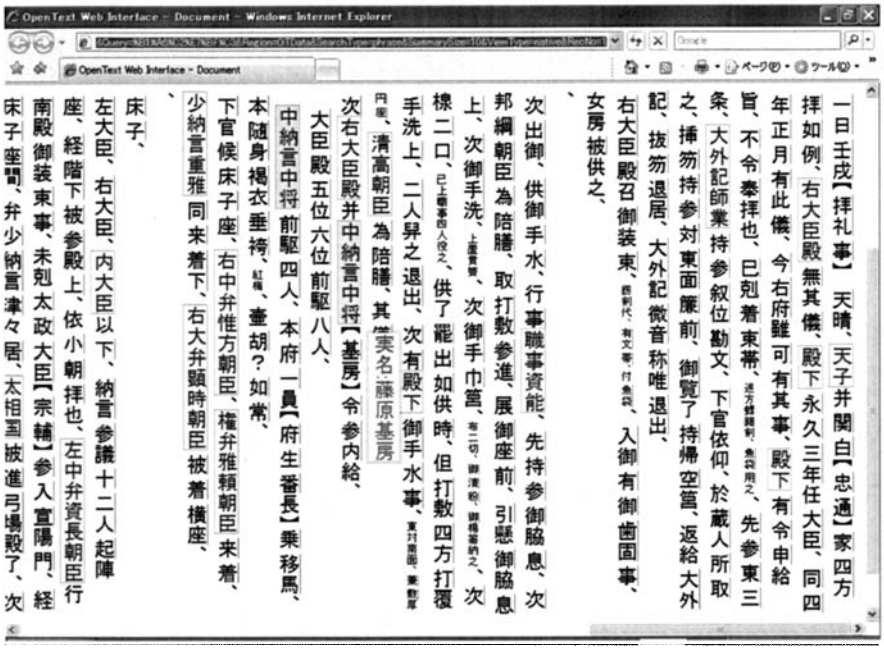


図 3:『兵範記』本文のハイパーテキスト

### 3.2. 『兵範記』検索システムの概要

本研究の主要な目的の一つは，古記録・古文書の文字情報を対象として，高度な情報アクセスを実現する手法の確立である．具体的には，古記録・古文書に対して，単なる文字列マッチングではなく，文書全体あるいは単語単位，さらには文字単位で意味を解析することにより，たとえば現代語による検索や，人名や地名，建造物名などを自動抽出し関連情報へリンクする機能などを提供することを目指している．この実現のために，XML，メタデータ，セマンティックWebなどの技術を用いることを検討している．また，古文書には現在の文字コードに含まれない文字が多く含まれるが，これら「外字」を含む文書に対して効率的な検索を行う手法についても検討している．

### 3.3. 人名・地名・建造物名の抽出

人名については，『立命館文学』別巻において兵範記の人名索引[8]が出版されており，すでにテキスト化が行われている．これは表形式のデータベースとして格納されており，本文中に現れた人名に対して，それが表す人物の本名，出現した日付，その他付加情報などが記載されている．兵範記に現れる人物の数は膨大であり，また人名は実名で書かれることはほとんどなく，様々な表記で記述される．現段階では，本文中に現れた約3,600名について，約2万件の出現のデータを抽出している．

また，地名・建造物名については，現段階で約270の建造物について，読み・分類・現在地名などの付加情報が記載されている．

本研究では，これらを基に，本文中に様々な表記で現れる人名・地名・建造物名に対して，そのメタデータや地図上の位置へのリンクを自動的に付与する手法を開発している．

### 3.4. 古記録・古文書の概念検索

古記録・古文書を現代語で検索するためには，文書中に現れる単語の意味を知る必要があるが，これを現在の自然言語処理技術で自動的に行うことは難しい．しかしながら，通常の情報検索においても，文書あるいは単語の意

味をシステムが理解した上で検索しているわけではなく，語義の曖昧性を残したままで検索を行っているのが現状である．情報検索では質問に対する完全な答えを求める必要はなく，関連すると思われる文書あるいは文書中の部分を返すものであるため，曖昧性を必ずしも完全に解消する必要はない．

本研究では，古文書の概念検索への第一歩として，国語辞典などの既存の辞書を用いて，すべての文字あるいは単語について可能性のある語義をすべて索引に登録し，これと質問との文字列マッチングを行うことで，関連する可能性のある文書中の部分を検索結果とする．また，単語共起傾向を用いることで，古文書における語義の曖昧性を解消することを検討している．

## 4. 『兵範記』本文の表示

本システムでは，前節までで述べた手法により抽出されたメタデータ，あるいは人名・建造物名などの固有表現を，利用者が容易にかつ効率的に発見・利用できるような本文の表示形式を実装した．具体的には，HTMLのスタイルシートを利用し，縦書き，フォントの指定，マウス移動時の付加情報のポップアップ表示，同一人物の色付けによる識別などを実現している．

### 4.1. 本文の表示機能

実際の本文の表示例を図 3に示す．この例では，図中央付近の「中納言中将」という文字列にマウスカーソルが置かれている状態を示している．ここでは「中納言中将」は藤原基房のことを指すが，マウスカーソル位置の右下に実名である「藤原基房」がポップアップ表示されており，さらに本文中で「藤原基房」を指す人名の表記が色つきで表示されている．これらの表示は，マウスカーソルの移動に応じて動的に表示が切り替わるようになっている．

また，本文中の右線（横書きの場合のアンダーラインに相当）の部分には，小学館の『国語大辞典』に記載されている語義へのリンクが張られている．これは，本文と辞書の見出しとの最長一致文字列を抽出することで実現している．

### 4.2. 今昔文字鏡による外字の表示

古文書や古記録には現在の文字コードに含まれない文字が多く含まれるが，これらの文字をコンピュータ上で表示することは容易ではない．本システムでは，約14万字からなる「今昔文字鏡」フォントを用いることで，『兵範記』におけるほぼすべての外字を一般的なWebブラウザ上で表示することを可能にした．

## 5. 断簡の復元

兵範記は，50余年の期間のうち28～29年分しか現存せず，その中でも「断簡」（何らかの事情で本来つながっていた日記の一部が切断され，ばらばらになったもの）が存在する（図 4）．テキスト化を進めるにあたって，この断簡を復元することは重要であるが，手がかりの少ない多数の文書の断片から復元するのは非常に困難な作業である．また，史学研究を進める上で，日記間の前後関係や年代を特定することは非常に重要である．そこで，本研究では，個々の断簡における文字の大きさや癖，紙の質感などをデータ化し，これを断簡復元の際のヒントとして用いることを検討している．つまり，これらのパラメータを特徴量とし，その類似度の高いものが，おそらく同じ文書中の断簡であろうという推定を行う．

文字の大きさや癖のパラメータ化のためには，古文書文字認識の技術を応用することを検討している．また，紙質の特徴量の抽出のために，武田ら[9]によるフラクタル次元を用いた特徴量抽出技術などを用いることを検討している．

これらの推定は，最終的には人手で確認する必要があり，コンピュータによる処理はあくまでもそれに対する補助的なものであるが，膨大な数の断簡をある程度対応付ける現実的な手段として有効であると考えている．

図 4:『兵範記』の断簡の例

## 6. 伝統的モンゴル文字文書のディジタル図書館

本研究プロジェクトにおけるこれまでの研究成果を基に，昨年より伝統的モンゴル語のディジタル図書館の構築を開始している．本システムの大きな特徴は，現代モンゴル語とは異なるスクリプトである伝統的モンゴル語で書かれた文書に対して，キリル文字で書かれた現代モンゴル語による問合せでの検索を実現する点にある．この実現のために，京都学ディジタル図書館において開発した概念検索および言語横断情報検索技術を応用する．

### 6.1. 伝統的モンゴル文字とモンゴル語

モンゴル国，中国及びロシア連邦の複数の地域に住んでいるモンゴル人の間ではモンゴル語が使用されている．モンゴル語はアルタイ言語の一つである．モンゴル語には独特の伝統的モンゴル文字とキリル文字を使用したモンゴル語の2種類があり，いずれも表音文字である．モンゴル文字の文字集合は子音文字が27文字，母音文字が8文字で合計35文字からなる．

モンゴル国では1946年からキリル文字が公式に使用される文字になった．ロシア語のキリル文字集合にモンゴル語固有の母音を表す「ө」と「ү」の2文字が追加されている．文字集合は母音文字13文字，子音文字20文字，記号2文字の35文字からなる．語順は日本語と同じく，主語－補語－述語のSOV型である．

モンゴル文字の書記体系はCJKの書記体系と共通点があり，西洋システムと大きく異なる．モンゴル文字は縦に書かれるが，日本語と違い左から右に行が進行する．また，モンゴル文字は単語中の位置によって異なる形をとる．こ

図 5:伝統的モンゴル文字で書かれた文書の例

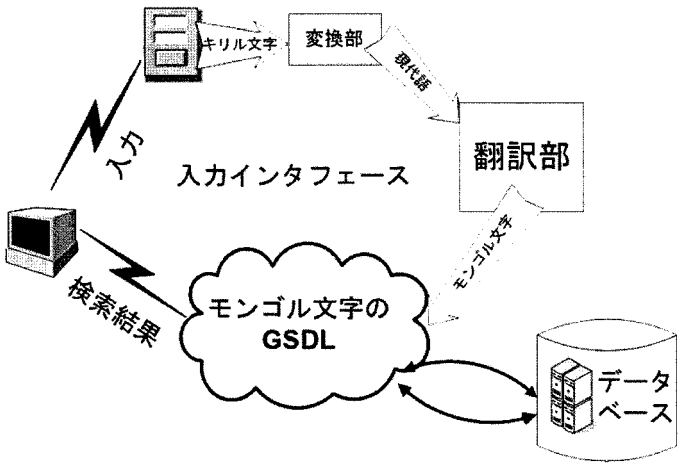


図 6:伝統的モンゴル文字文書のディジタル図書館の構成

れは独立形，語頭形，語中形，語尾形である．また，ある子音が母音と結合し，弓形の形態をとる場合もある．Unicode規格には基本文字セット，句読記号と数字のみが登録されている．

### 6.2. システムの構成

本システムではWaikato 大学で開発されたGreenstoneディジタル図書館（GSDL）システム[1]を使用する．GreenstoneシステムはUnicodeに対応した，ディジタル図書館コレクションの構築や配布のためのオープンソースソフトウェアである．

利用者がキリル文字もしくはラテン文字の検索キーワードを入力すると，変換部により現代モンゴル語の検索質問に変換され，翻訳部に送られる．ここで，検索質問を単語に分割し，辞書により伝統的モンゴル文字による単語に変換する．これを用いてGreenstoneシステムで検索が行われ，最終的にモンゴル文字文書の検索結果が表示される（図 6）．

## 7. ディジタルコンテンツのための情報アクセス基盤

本研究プロジェクトでは，前述の『兵範記』や伝統的モンゴル語の古文書を含む各種メディアから構成されるディジタルコンテンツに対する効率的な情報アクセスを実現するための基盤技術についても研究を行っている．従来のテキストにとどまらず，イメージ，映像などのあらゆる種類のディジタルコンテンツに対する効率的なアクセス手段の確立を目指している．具体的には，メディアの種類に依存しないメタデータ表現のための枠組みを基盤として利用し，さらにメタデータの語彙の関係を記述するオントロジを構築する．これにより各種メタデータの相互運用性が確保されることで，ネットワーク上に分散して蓄積されているディジタルコンテンツに対して統一的な情報アクセス手段を提供する．

本研究は，メディアの種類に依存しないメタデータ表現として次節で述べるDublin Coreを利用し，さらにそれらの関係を記述するオントロジを構築することで，さまざまなメディアに対する統一的なアクセス手段を提供する点に特徴がある．さらに国際標準の情報検索プロトコルであるZ39.50を用いることにより，さまざまな機関で蓄積が進んでいるディジタルコンテンツの相互運用性が確保され，複数のディジタルアーカイブやディジタル図書館のコンテンツを同時に検索するなど，これまで不可能であった機能が実現できる．

[1] http://www.greenstone.org/

### 7.1. Dublin Coreメタデータ

Dublin Core とは，1995 年にオハイオ州ダブリンにあるOCLC（Online Computer Library Center）で開催されたワークショップで提案されたメタデータの規格である．Dublin Core の特徴は，インターネットのような巨大情報空間において，分野を超えて情報資源を探し出す要求のために開発されたため，様々な分野のメタデータの記述項目が統一されていることである．そのため，分野によらない共通したメタデータを作成することができる．

### 7.2. 人文系データベースの統合検索

上述のDublin Coreに従ってメタデータを設計し，Z39.50に対応した情報検索システムを構築することにより，複数のディジタルアーカイブやディジタル図書館のコンテンツを統一的に検索する手段を提供することができる．これにより，ディジタルアーカイブやディジタル図書館の利用の促進に寄与することが期待できる．本研究プロジェクトでは，前述の京都学ディジタル図書館システムと，本学および他の研究機関で公開されている人文系データベースとの統一的な検索の実現を目指している．

## 8. おわりに

本稿では，筆者が21世紀COEプログラムにおいて進めている，文化的資料のディジタルコンテンツに対して高度な情報アクセスを実現するための研究プロジェクトについて述べた．

本研究の今後の課題として，京都学ディジタル図書館システムにおける『兵範記』オントロジの構築と，それによる概念検索や言語・時代・文化横断型情報検索などのより高度な情報アクセスの実現，Dublin CoreおよびZ39.50を用いた本学および他の研究機関で公開されている人文系データベースとの統一的な検索の実現などが挙げられる．

## 参考文献


[1] 前田 亮，佐古 愛己，杉橋 隆夫：京都学ディジタル図書館の構築と多言語情報アクセス．人文科学とコンピュータシンポジウム論文集，Vol. 2003, No. 21, pp. 195-202（2003）

[2] 佐古 愛己，河角 龍典，前田 亮，杉橋 隆夫：古記録データベースと歴史的空間情報のGIS化．人文科学とコンピュータシンポジウム論文集，Vol. 2004, No. 17, pp. 9-16（2004）

[3] Garmaabazar Khaltarkhuu，前田 亮：伝統的モンゴル文字文書の現代モンゴル語による検索手法の提案．第5回情報科学技術フォーラム講演論文集（2006）

[4] Garmaabazar Khaltarkhuu and Akira Maeda: Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text. Proc. 9th International Conference on Asian Digital Libraries (ICADL2006), pp. 478-481（2006）

[5] ハルタルフー・ガルマーバザル，前田 亮：伝統的モンゴル文字文書のディジタル図書館の構築．人文科学とコンピュータシンポジウム論文集，(2006)

[6] 木村 文則，前田 亮，宮崎 純，吉川 正俊，植村 俊亮：Webディレクトリを言語資源として利用した言語横断情報検索．情報処理学会論文誌：データベース，Vol. 45, No. SIG 7 (TOD 22), pp. 208-217（2004）

[7] 京都大学文学部国史研究室編：「兵範記二」，京都大学史料叢書2，思文閣出版（1988）

[8] 兵範記輪読会編：「兵範記人名索引」Ⅰ～Ⅲ，『立命館文学』別巻（1987, 1991, 1999）

[9] 武田 哲也，村山 健二，岡田 至弘，坂井 利之：大規模古文書画像データベース構築のためのパターン解析手法の検討，京都大学大型計算機センター 第 57 回研究セミナー報告，pp.11-15（1997）