京都学ディジタル図書館プロジェクト

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
理工学研究科

概要 本稿では，筆者が現在進めている二つの研究プロジェクトの概要を紹介する．一つは21世紀COEプログラムのサブプロジェクト「京都学ディジタル図書館の構築」であり，もう一つはオープン・リサーチセンター整備事業のプロジェクトである「ディジタルコンテンツのための情報アクセス基盤に関する研究」プロジェクトである．COEのプロジェクトでは，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の構築を行っている．オープン・リサーチのプロジェクトでは，人文科学分野のデータベースのメタサーチを実現する技術の研究を行っている．本稿では，これら二つの研究プロジェクトの現在の状況と今後の展望について述べる．

Building a digital library of Kyoto studies:

Information Access Infrastructure for Digital Libraries

Akira Maeda
Graduate School of Science and Engineering

Abstract In this paper, we briefly describe two research projects that the author is involved in. One is a project of the 21st Century COE Program, which is called "Building a Digital Library of Kyoto Studies". The other is "Research Project on Information Access Infrastructure for Digital Contents" of the Open Research Center Program. In the COE project, we are building a digital library of "Hyohanki", which is a diary written by an aristocrat during the late Heian era (1132-1184). In the Open Research Center project, we are investigating techniques for realizing a meta-search of the databases related to the humanities. This paper describes the current developments and future challenges of these two research projects.

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プログラムおよびオープン・リサーチセンター整備事業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．

21世紀COEプログラムでは，「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『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

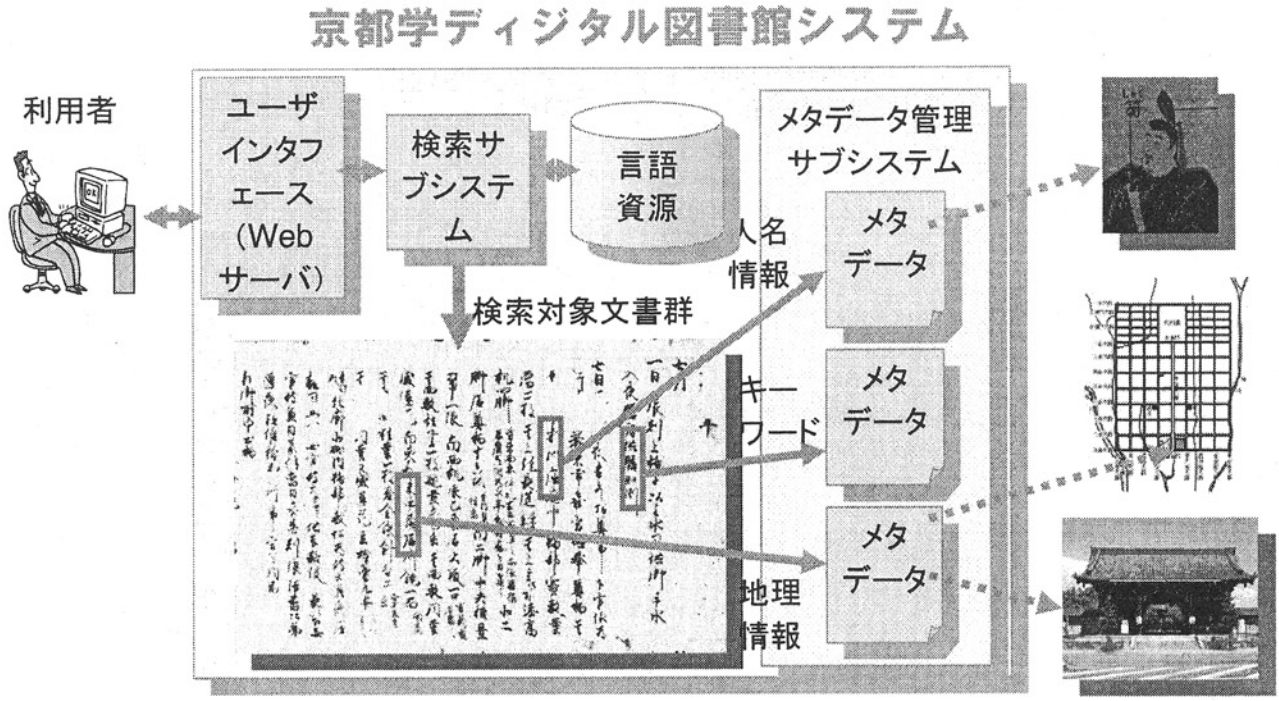

図 1:京都学ディジタル図書館システムの概要

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ことを目指す[2]．さらに，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について研究を行っている．

一方，オープン・リサーチセンター整備事業では，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2つ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

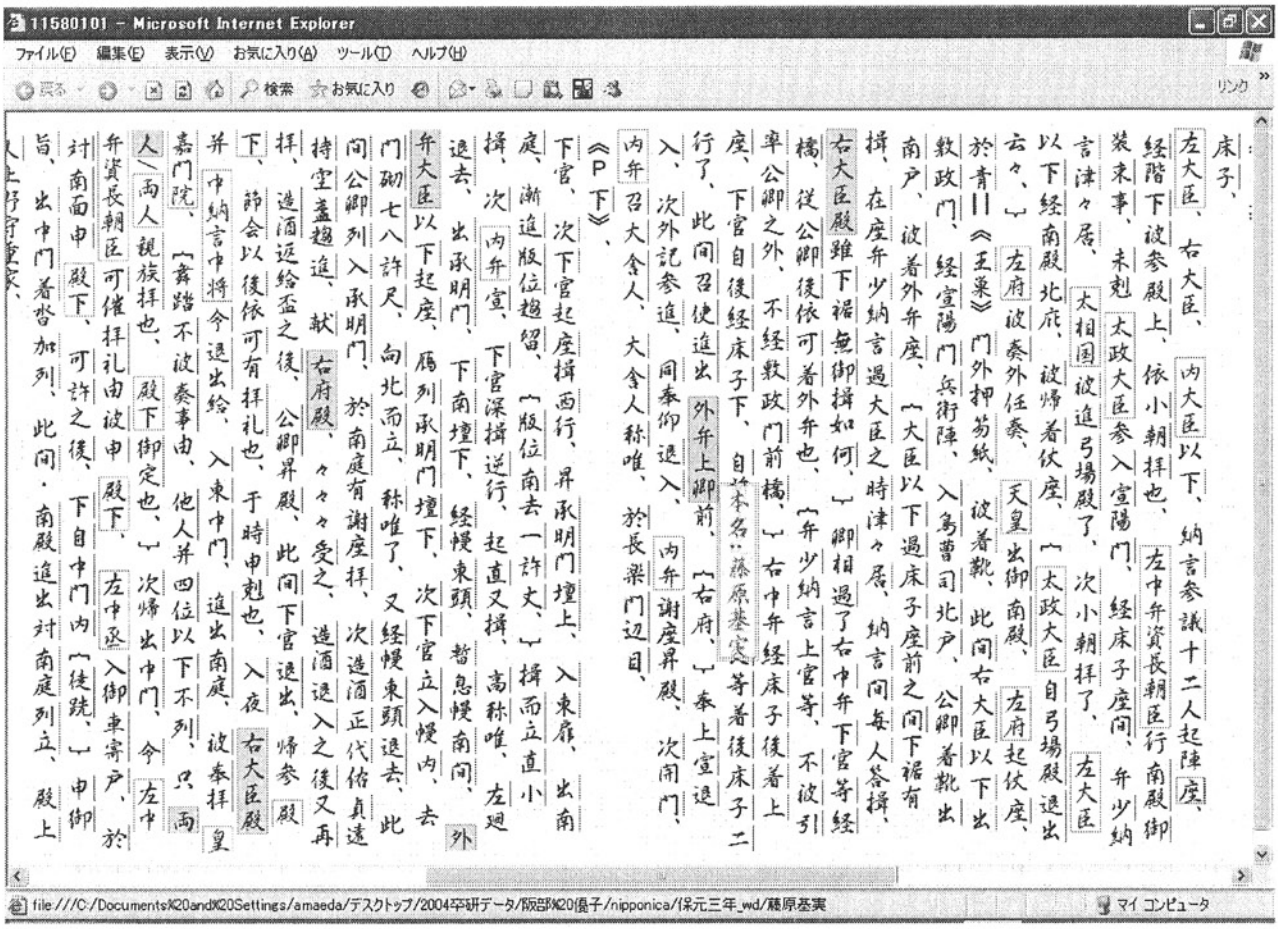

図 2:『兵範記』本文のハイパーテキスト

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ている. 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

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.

3.1. 人名・地名・建造物名の抽出

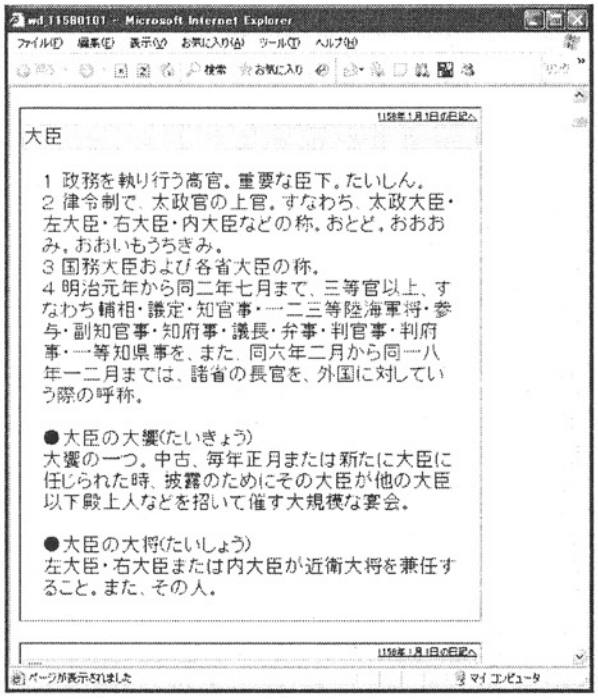

図 3:本文中に現れる単語の語義の表示

人名については,「立命館文学」において兵範記の人名索引が出版されており,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検討している.

3.2. 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ことを考えている.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検討している.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.

実際の本文の表示例を図 2に示す.この例では,図中央付近の「外弁上卿」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「外弁上卿」は藤原基実のことを指すが,マウスカーソル位置の右下に実名である「藤原基実」がポップアップ表示されており,さらに本文中で「藤原基実」を指す人名の表記(「右大臣殿」「外弁大臣」「右府殿」)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

表 1:利用者評価の結果

評価項目	システム適用前					システム適用後				
	1	2	3	4	5	1	2	3	4	5
本文の古文らしさ	6	11	3	0	0	0	0	4	14	2
本文に登場する人物がわかる	7	12	1	0	0	0	0	0	0	20
本文に登場する呼び名の違う同一人物が発見できる	18	2	0	0	0	0	0	0	0	20
大まかな本文の意味がわかる	15	5	0	0	0	0	1	15	3	1
一行(一部分) でも意味がわかる	13	5	1	0	0	0	0	8	10	2
得られる情報量	16	3	1	0	0	0	0	6	12	2

字列を抽出することで実現している.「大臣」という単語の語義を表示した例を図 3に示す.

5. 評価実験

前節で述べたメタデータ抽出手法を『兵範記』の本文へ適用し, システム導入前と比較してどの程度情報量が増えたかなど, システムの有用性について検証した.

5.1. 実験の概要

この実験は, メタデータ抽出手法を『兵範記』の本文へ適用し, 適用した本文と適用する前の本文とを比較し, どのように変更されたか, どのような利点があるかについて客観的な立場から利用者に評価をしてもらい, システムの有用性を検証するという目的で行った.

評価方法としては, 表 1に示す6項目について, 5 段階評価で比較・評価してもらう. 比較対象となる本文は, 1158 年11 月16 日の日記を利用し, システム適用前の本文はプレーンテキスト形式のものとし, システム適用後の本文は図 2のようなHTML 形式のものとする.

5.2. 評価結果

利用者評価は, 日本語の読める人を対象に20人に評価を行ってもらった. 評価結果は表 1の通りである. この結果から, どの項目においても評価が上がっていた. また, この項目だけでなくシステム適用後の本文を読んだ感想を聞くと,「古文の知識や読めない漢字があっても単語帳を見ることで少しは意味が分かった」「自分で辞典を引かなくて良いので楽だった」「同一人物がすぐ分かるのがよい」などの肯定的な意見が多く見られた.

6. ディジタルコンテンツのための情報アクセス基盤

オープン・リサーチセンター整備事業における研究プロジェクトでは, 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研究を行っている. 冒頭で述べたように, すでに膨大な量になりつつあるディジタルコンテンツを有効に利用するためには, それらに対する効率的なアクセス手段が必要であることは, 現在のWebの状況を見ても明らかである. 本研究プロジェクトでは, 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的とする. 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ことを目指す.

本研究は, メディアの種類に依存しないメタデ

表 2:Dublin Coreの基本エレメント

エレメント	説明
Title	情報資源に与えられた名前. タイトル.
Creator	情報資源の内容の作成にあたった主たる責任を持つもの. 著者, 作者.
Subject	情報資源の内容のトピックス. 主題, キーワード.
Description	情報資源の内容の記述. 内容記述.
Publisher	情報資源を利用可能にするにあたって責任を持つもの. 出版者, 公開者.
Contributor	情報資源の内容への寄与に対して責任をもつもの. 他の関与者, 寄与者.
Date	情報資源の何らかの事象に関連付けられた日付.
Type	情報資源の内容の性質もしくはジャンル. 資源タイプ.
Format	物理的表現の形式, もしくはディジタルにおける表現形式.
Identifier	与えられた環境において一意に定まる情報資源に対する参照. 資源識別子.
Source	現在の情報資源が作り出される源となった情報資源への参照. 情報源, 出所.
Language	該当情報の内容の言語.
Relation	関連情報への参照. 関係.
Coverage	情報資源の内容が表す範囲あるいは領域. 対象範囲. 空間・時間的範囲.
Right	情報資源に含まれる, もしくはかわる権利に関する情報. 権利管理.

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.

6.1. 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. Dublin Core の基本的な記述項目である15 項目の基本エレメントは表 2の通りである.

6.2. Z39.50情報検索プロトコル

Z39.50とは, クライアント・サーバ環境における情報検索システムの標準的なプロトコルとしてANSIが制定したもので, ISOの規格にもなっている. Z39.50の特徴として, この規格に対応した複数のデータベースに対する統合検索(横断検索)が可能であることが挙げられる. 米国では多くの図書館でZ39.50対応のOPACが公開されているが, 国内ではあまり普及していない.

6.3. 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.

7. おわりに

本稿では, 筆者が21世紀COEプログラムおよびオープン・リサーチセンター整備事業において

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横断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, Dec. 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, Dec. 2004.