e-Japanology における情報発信プラットフォームの試み

辻澤隆彦（東京農工大学　総合情報メディアセンター）

【キーワード】 e-Japanology、Web 検索技術、クローラエンジン、
インデックス化

1．はじめに

WWW（World Wide Web）の登場により種々の情報が加速度的に発信されるようになり、これらの情報を効率よく検索するための Web 検索技術が進展してきた [1]、[2]。Web 検索技術はクローラエンジン、ページリポジトリ、インデックス化モジュール、クエリーモジュール、ページランク付けモジュールなどから構成されることが一般的である。クローラエンジンは Web 上の文書を収集するソフトウェアであり、ロボットあるいはスパイダーなどと呼ばれている。このクローラが収集してきた文章を一旦ページリポジトリに蓄積し、インデックス化モジュールによりインデックス化して検索データベースとする。Web 検索では情報量が膨大であることから、一般にページランク付けモジュールを使って表示する順序などを決定している。一方、オフィス情報システムに目を移すと、企業内における情報の分散化が進展し、種々の部門において Web サーバやファイルサーバに種々の情報が蓄積されてきている。これらの分散化した情報への到達手法として、クローラエンジンの適用可能性が注目され、企業内システムへの取り組みも報告されている［3］。

e-Japanology の構想は多言語アクセスに対応した日本学のコミュニティ基盤の構築と多摩地区の日本学研究教育組織および外国人留学生コミュニティを活用した日本学知識の構築・蓄積、さらには多言語アクセス及び知識資源の継続的な累加のシステムの実現を目標とした構想である。現在、東京外国度大学、東京学芸大学、東京農工大学が共同でプロジェクトとして取り組んでおり、多摩地区の日本学研究教育組織の持てる知識資源を恒久的に集約・更新でき、且つグローバルに日本学知へのアクセスビリティを強化した仕組みを構築することで、日本学教育研究での価値創造を支援することを最終的な目標としている。

著者は、上述した Web 検索技術として活用されているクローラエンジンによるデータベース構築技術を活用し、拡充を図っていくことが e-Japanplogy 構想を実現する一つの方法であると考え、テストベッド構築の検討を進めてきた。ここでは、第一段階として進めた、Web 図書館コンテンツをインデックス化したテストベッドシステムについて報告する。検索対象とした Web および Web 図書館は近代デジタルライブラリ（http://kindai.ndl.go.jp/）・デジタルライブラリ（古典籍）（http://del.ndl.go.jp/）・国際日本研究センター Web（http://www.tufs.ac.jp/common/icjs/jp/index.html）・京都国際マンガミュージアム（http://mmsearch.kyotomm.jp/index.html）・えむえむブログ（http://d.hatena.ne.jp/kyotomm/）・Japanese American National Mu-

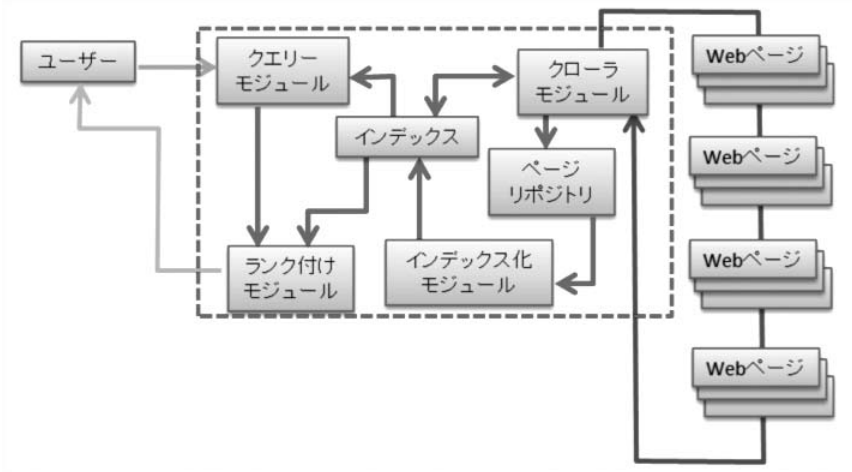
seum（http://www.janm.org/collections/）である。テストベッドでは個々のWeb図書館検索ページでの検索に比べ、一つの検索窓から複数のWeb図書館情報を検索できることが確認でき、e-Japanology構想を実現するプラットフォームとしての可能性を示すことができた。オフィス情報システムがそうであるように、クローラエンジン技術とインデックス化によるデータベース構築だけでは依然不十分ではある。今後、研究者の連携情報や、サブカルチャー情報から日本語研究情報までの幅広い日本語研究情報を扱うための階層化表示機能など、プラットフォームが持つべき機能についての検討とテストベッドへの実装を通した評価実験を進めていく予定である。

以下、2．ではWeb検索エンジンの概要について、3．ではテストベッドの詳細について述べる。

2. Web検索エンジンの概要

図1はWeb検索技術の概要を示した図である。以下簡単に各モジュールの機能について説明する。

図1　Web検索エンジンの概要



2.1　クローラモジュール

広く分散するWebページについてどのページをどのようにとってくるかが記載されているプログラムであり、同時に、どのような頻度でWebページをクロールするかも含まれる。一般にWebページの情報を取得する場合、http getというプロトコルが使われる。Webページによっては、特定のページやディレクトリをクロールさせたくない場合があるが、この場合はクローラをブロックすることや検索結果を表示させないように制御することがWebページ側で可能である。

2.2　ページリポジトリ

クローラにより収集された新しいWebページを一時的に蓄えるのがページリポジトリで

ある。その後、インデックス化モジュールに Web ページが送られるまで、ページリポジトリに蓄積された情報は残される。

2.3　インデックス化モジュール

ページリポジトリから送られた Web ページを基に、必須記述子を抽出する機能を持つ。ここで抽出された記述子などはインデックスとして出力され、インデックス内に格納される。

2.4　インデックス

インデックスには Web ページのキーワード、題名、鍵となる文章の内容、画像インデックス、PDF インデックスなどが格納される。

2.5　ランク付けモジュール

Web ページをある基準に従ってランク付けする機能をもつ。Google が採用しているリンクポピュラリティに基づく PageRank はその例である。リンクポピュラリティは Web ページ間のリンクを一種の人気投票とみなし、多くの質の高いリンクを集める Web ページは高い支持を受けているとして Web ページのランク付けを行う手法である。

2.6　クエリーモジュール

ユーザからの自然言語による質問に回答する機能を持つ。インデックスにアクセスし、対応する Web ページをランク付けモジュールの結果と関係づけてユーザに回答する。

3. e-Japanology 情報発信用テストベッド

2. では Web 検索エンジンの概要について述べた。e-Japanology 情報発信用テストベッドでは、図1に示した Web 検索エンジンの中のランク付けモジュールを除いた機能により、システムの構築を行った。Web ページ検索と異なる点は、Web クローラで収集できない対象（例えば対象 Web ページが検索ページである場合など）では対象に特化したプログラムを埋め込むなどのカスタマイズ化が必要となったことである。テストベッドの構築ではクアンタムテクノロジー社が開発をした N-gram 方式に基づく検索エンジン A-trek を使った。

3.1　A-Trek の特徴

A-Trek が持つクローラには、Web 検索エンジンで使われている Web クローラの他に、Web クローラでは収集できない Web ページに対応するためのカスタムクローラ、Windows などの共有フォルダ上のファイルを収集する共有ファイルクローラ、リレーショナルデータベース中の情報を取り出すためのデータベースクローラなど各種ローラーが用意されているが、テストベッドの構築では Web クローラと検索 Web ページを対象としたカスタムプログラムを使ったカスタムクローラを用いた。

3.2 クローリング対象サイトとカスタムクローラ

テストベッドにおいてクローリングの対象とした Web サイトは以下6 サイトとした。

- 近代デジタルライブラリ （http://kindai.ndl.go.jp/）
- デジタルライブラリ（古典籍） （http://del.ndl.go.jp/）
- 国際日本研究センター （http://www.tufs.ac.jp/common/icjs/jp/index.html）
- 京都国際マンガミュージアム （http://mmsearch.kyotomm.jp/index.html）
- えむえむブログ （http://d.hatena.ne.jp/kyotomm/）
- Japanese American National Museum （http://www.janm.org/collections/）

以下、各サイトのクローリングに対応したカスタマイズについて具体的に記述する。

3.2.1 近代デジタルライブラリとデジタルライブラリ（古典籍）

これらのサイトは、サイト自体が検索システムとして提供されている。検索システムとして提供されたサイトは、検索条件を与えなければならないため、汎用の Web クローラでは、情報を収集できない。ここでは、検索条件を工夫し、全件のデータを表示させるためのプログラムが必要であった。これらのサイトは、検索結果の最大数が1,000 件であるという縛りがあり、結果が1,000件以下になるよう検索条件を調整しなければならなかったが、NDC 番号と出版年を組み合わせて検索結果を表示させその内容をクローリングすることで実現した。ただし､「デジタルライブラリ（古典籍）」では､絞り込みに指定する条件がすべてのデータには存在していなかったため、全件を取得することはできなかった。図２は近代デジタルライブラリのトップページを、図３は近代デジタルライブラリ NDC 番号を示した図である。図2からわかるように、近代デジタルライブラリのトップページには検索システムが用意されている。図４には NDC 番号と年代から検索した近代デジタルライブラリの検索結果を示した。

図2 近代デジタルライブラリトップページ

図3　近代デジタルライブラリにおける NDC 番号

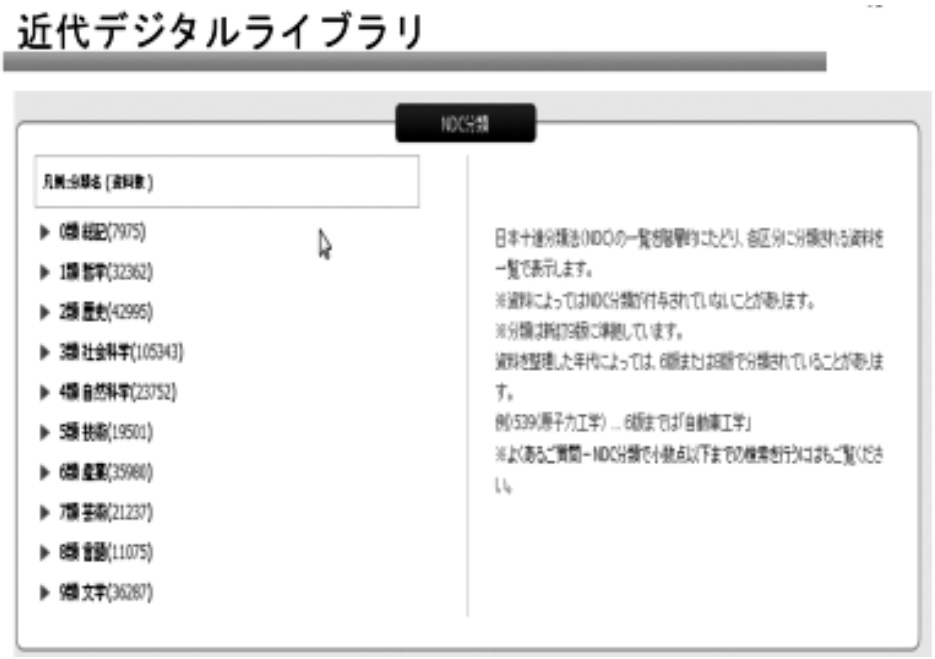
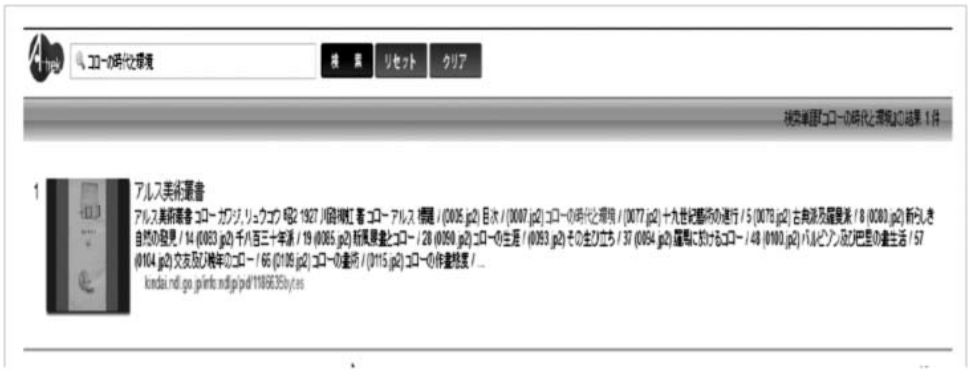


図4　NDC 番号と年代から検索した近代デジタルライブラリの検索結果

図５は構築したテストベッドを使った検索結果である。図４の最後段にあるアルプス美術選書の中に記載されている『コローの時代と環境』を検索キーワードとして検索した結果を示している。

図5　テストベッドを使った検索結果



3.2.2 国際日本研究センター

汎用の Web クローラでも可能なサイトであったが、PDF を検索対象として HTML 上の文字列をそのタイトルにしたり、HTML からリンクされている画像をサムネイルとしたりするために、カスタマイズを行った。図6は『東京外国語大学府中キャンパス留学生日本教育センター・さくらホール』を検索語として検索した結果を示した図である。図7は『日本の子育て』を検索語として検索した結果を示した図である。PDF ファイルやビデオ教材ファイルのトップページがサムネイルとして表示されていることが確認できる。

図6　テストベッドを使った国際日本研究センターページにおける

図7　テストベッドを使った国際日本研究センターページにおける

3.2.3 京都国際マンガミュージアム

この Web ページも検索機能を用いた Web ページで、汎用的な Web クローラでは収集できない Web ページになっている。近代デジタルライブラリにおいて行ったと同様に年度を検索条件に指定して検索を行い、その結果をクローリングするカスタマイズを行った。この Web ページでは、最終的な情報（検索結果からリンクされているページ）が一意の URL になっていないため、情報を収集しテストベッド側で検索するシステムを作成した。しかしながら、検索結果からオリジナルソースにリンクできないという問題が発生した。テストベッドでキャッシュしているデータへはリンクできるが、著作権などの問題を考慮して NOT FOUND を表示するようにした。

3.2.4 えむえむブログ

ブログでは一般に、投稿をいろいろな方法で表示できる。トップページには最新の何件かを表示したり、カレンダーを選択することで特定の日の投稿を表示したり、カテゴリづけにより特定のカテゴリの投稿だけを表示したりできる。また、これらのページは同じ URL であっても動的に変化するものが一般的である。 ブログが持つこのような性質のため、汎用の Web クローラを使ってブログを収集すると全てのページを取得してしまい同じ投稿が複数のページに含まれることとなる。さらに、動的に変化する URL の場合ではクローリングした時刻との関係で目的の HTML が表示されないことや存在しない可能性もある。このため、一投稿ごとに一意の URL を持ったページだけを収集する必要があり、ブログ専用のカスタムマイズが必要であった。ここでは、ブログに含まれる最初の画像をサムネイルとして収集する機能も取り込んでいる。

3.2.5 Japanese American National Museum

この Web ページでは、特定の URL からリンクされる全情報を取得することで必要な情報を得られるが、各ページのサムネイルを取得するためにカスタマイズを行った。図８は Japanese American National Museum のサイトの Clara Breed によるコレクション（第２次大戦

中米国における日本の子供たちからの手紙など）の一部を示したものである。図9は JPEG イメージをサムネイルとして蓄積したテストベッドにおいて、検索結果を示したものである。

図8　Japanese American National Museum のサイトの Clara Breed によるコレクション例

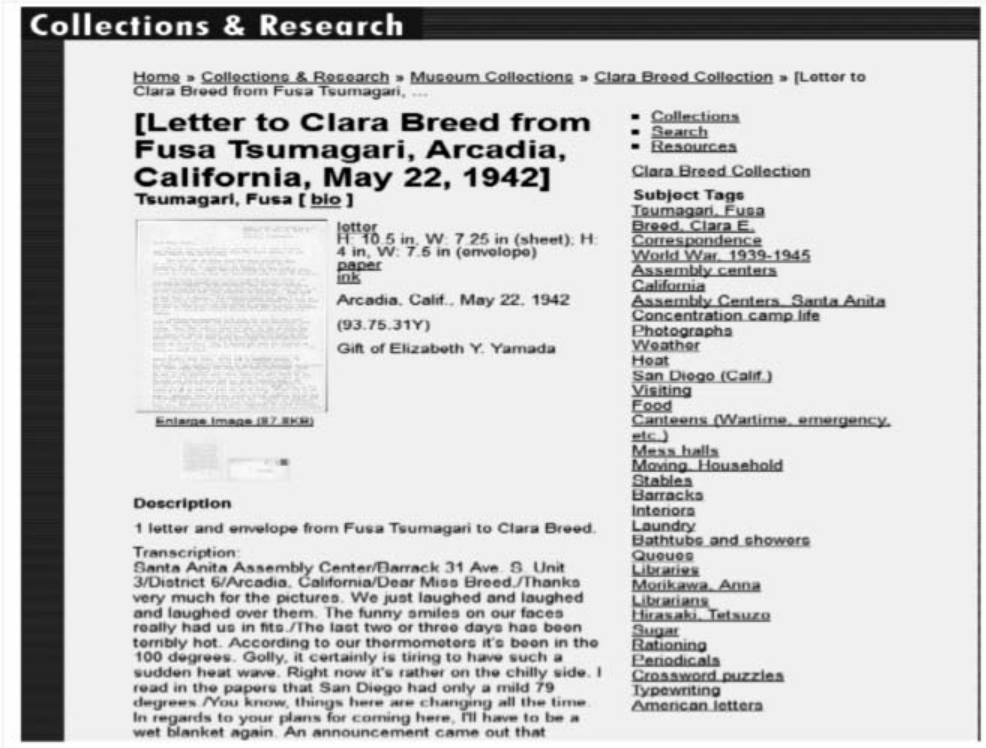


図9　テストベッドを使った検索結果

4. e-Japanology における情報発信基盤として　- 今後の進め方 -

3。では情報発信テストベッドとして構築したシステムの概要について説明した。そこでは、既存の Web ページあるいは Web サイトを対象に情報のハーベスティングを行い、PDF ファイルを含む情報のクローリングと検索のためのデータベース（インデックス化）を作成した。既に述べたように、Web サイトあるいは Web ページのクローリングでは種々のカスタマイズが必要であった。今後、情報発信基盤として e-Japanology に関する種々の情

報をクローリングしていく上では既存の Web サイトへの対応と同時に、これから Web ページなどを構築していく場合への対応について検討していく必要がある。

4.1　既存の Web ページあるいは Web サイトへの対応

検索 Web サイトなどへの対応は今後もクローラのカスタマイズが必要になる。このため、計画的な対応を検討していく必要がある。カスタマイズにおいては、クローリングの対象や範囲についての検討をしながら進めていくことが必要となる。

4.2　新規に Web サイトや Web ページを構築する場合の対応

ここでは、以下の2つの方針について述べたい。

4.2.1　ブログ形式による Web ページ構築

ブログは一度装飾を決めることで、その後、情報を記述するだけでページが作成されるという特徴を持つ。このため、どの投稿も統一感のあるデザインになるだけでなく、自由なカテゴリによる分類や時系列による管理、また、画像添付ファイルの付加や他者によるコメント追加などが可能であり、表現の自由度が高い。

このことから、同一のブログシステムにより情報蓄積を進めることも情報発信基盤を整備していく上では必要であると考える。このことで、クローラの準備が容易になるという利点もある。商用のブログソフトによっても、e-Japanology に興味を持つメンバーによる情報発信も可能となる。

4.2.2　LMS の活用による共通的な情報発信基盤構築

e-Japanology プロジェクトでは日本語教材活用の可能性を検証する目的で LMS「Sakai」を利用した e-Japanology Gate Way を開設してきた。そこでは、ビデオ教材を含め種々の教材をアップロードすることができる。この e-Japanology Gate way を使った情報蓄積が新たな Web サイト構築とそのためのクローラのカスタマイズよりも効率的情報発信プラットフォームを構築できるものと考える。この Gate Way をクローリング対象とすることで、情報の蓄積とインデックス化を効果的に進めることができる。LMS の活用は利用者の認証が可能となることも利点の一つとなる。

5. むすび

e-Japanology の構想は多言語アクセスに対応した日本学のコミュニティ基盤の構築と多摩地区の日本学研究教育組織および外国人留学生コミュニティを活用した日本学知識の構築・蓄積、さらには多言語アクセス及び知識資源の継続的な累加のシステムの実現を目標とした構想で、東京外国度大学、東京学芸大学、東京農工大学が共同でプロジェクトとして取り組んできている。本稿では Web 検索技術として活用されているクローラエンジンによるデータベース構築技術に着目し、この技術が e-Japanplogy 構想を実現するための情報プラッ

トフォーム構築の一つの方法であると考え、進めてきた評価用テストベッド、具体的には、Web 図書館コンテンツをインデックス化した検索システムについて述べたものである。検索対象とした Web および Web 図書館は近代デジタルライブラリ（http://kindai.ndl.go.jp/）・デジタルライブラリ（古典籍）（http://del.ndl.go.jp/）・国際日本研究センター Web（http://www.tufs.ac.jp/common/icjs/jp/index.html）・京都国際マンガミュージアム（http://mmsearch.kyotomm.jp/index.html）・えむえむブログ（http://d.hatena.ne.jp/kyotomm/）・Japanese American National Museum（http://www.janm.org/collections/）である。テストベッドでは個々の Web ページ図書館検索ページでの検索に比べ、一つの検索窓から複数の Web 図書館情報を検索できることが確認でき、e-Japanology 構想を実現するプラットフォームとしての可能性を示すことができた。しかしながら、既存の Web ページコンテンツを対象とすると、Web ページによってはクローラをカスタマイズする必要があり、すべての既存ページコンテンツを対象とすることの困難さも明らかになった。このため、既存のページと今後新規に構築する場合についての対応を検討する必要性についても述べた。その中で、新規に Web ページ構築を行う場合の基本的な考え方について示した。

オフィス情報システムがそうであるように、クローラエンジン技術とインデックス化によるデータベース構築だけでは依然不十分ではある。今後、研究者の連携情報やサブカルチャー情報から日本語研究情報までの幅広い日本語研究情報を扱うための階層化表示機能など、プラットフォームが持つべき機能についての検討とテストベッドへの実装を通した評価実験が必要になる。

参考文献

[1] A. N. Langville, C. D. Meyer（著）、岩野、黒川利明、黒川洋（訳）、「Google Page Rank の数理」、共立出版、2009

[2] C. D. Manning, P. Raghavan, H.Schutze（著）、岩野、黒川、濱田、村上（訳）、「情報検索の基礎」、共立出版、2012

[3] 安藤、志賀、岩倉、岡本、「企業内情報検索の高度化手法の提案と評価」、情報処理学会研究報告、DD [デジタル・ドキュメント] 2010-DD-76（3）、pp. 1-6、2010

Setting a Platform for Transmitting Information in e-Japanology

TUJISAWA, Takahiko

Tokyo University of Agriculture and Technology

【keywords】 e-Japanology, web retrieval technology, crawler engine, indexing

A concept of e-Japanology is an initiative with the goal to achieve the community for Japanology studies accessible by multi-language. Tokyo University of Foreign Studies, Tokyo Gakugei University, and Tokyo University of Agriculture and Technology are working as a project jointly. The World Wide Web has become the largest information source in recent years and search engines are indispensable tools for finding needed information from the Web. We considered that the search engine technology, especially the crawler engine, will become one of the platforms for the community of Japanology studies, and tried to develop the test bed system using the crawler engine that has indexed the web content library.

In this paper, the test bed system which harvests the Japanology content accumulated in the web content library and makes indexed database is introduced. Several digital library, such as the digital library by National Diet library, the web site of International center for Japanese studies of Tokyo University of Foreign Studies etc., are object for data harvesting. Using the test bet system, it is understood that the needed information can be retrieved from several web sites by one search window.

東京外国語大学国際日本研究センター『日本語・日本学研究』第 4 号執筆者一覧

孫斐	北京大学大学院博士後期課程
ツォイ・エカテリーナ	東京外国語大学大学院博士後期課程
Hanan Rafik Mohamed	カイロ大学
葛茜	福州大学
篠原将成	国際基督教大学大学院博士後期課程
鈴木智美	東京外国語大学
花園悟	東京外国語大学
臼井直也	東京外国語大学大学院博士後期課程
谷口龍子	東京外国語大学
望月圭子	東京外国語大学
尹鎬淑	サイバー韓国外国語大学校
田中和美	国際基督教大学
ASADCHIH Oksana	タラス・シェフチェンコ記念キエフ国立大学
辻澤隆彦	東京農工大学

『日本語・日本学研究』国際編集顧問一覧（順不同）

趙華敏	北京大学
徐一平	北京外国語大学
蕭幸君	東海大学（台湾）
尹鎬淑	サイバー韓国外国語大学校
任榮哲	中央大学校（韓国）
于乃明	国立政治大学
金鐘德	韓国外国語大学校
陳明姿	国立台湾大学

編集後記　東京外国語大学国際日本研究センター『日本語・日本学研究』第4号をお届けします。／今号への公募論文の応募総数は14本（言語6、日本語教育3、文学3、歴史研究1、文化1）。うち8本が採用となりました。／また今号では、2013年7月31日から8月2日にかけて開催された夏季セミナー2013「言語・文学・歴史――国際日本学の試み」でおこなわれた院生発表会の要旨を掲載いたしました。国内外の院生の活気ある報告に私たちも大きな刺激を受けました。セミナー開催にあたってご協力いただいたみなさまに心から感謝申し上げます。（友常勉）

東京外国語大学国際日本研究センター
日本語・日本学研究 vol.4
Journal for Japanese Studies

発行：2014 年 3 月 31 日

編集者・発行者　東京外国語大学国際日本研究センター
代表者　野本京子
〒183-8534　東京都府中市朝日町 3-11-1　アゴラ・グローバル 2F
Tel/Fax: 042-330-5794

印刷・製本　㈲山猫印刷所
〒116-0014　東京都荒川区東日暮里 5-39-1
Tel: 03-5810-6945