アジア歴史資料センター

—本格的なデジタルアーカイブを目指して—

牟田昌平¹, 小林昭夫²

著者抄録 : アジア歴史資料センターは、2001年11月30日、独立行政法人国立公文書館の組織として開設された。アジア近隣諸国との相互理解促進のために、政府が所蔵する戦前の公文書から、アジア諸国との関係資料をインターネットで「いつでも」「どこでも」「だれもが」「無料」で検索し画像データとして利用できる本格的なデジタルアーカイブである。閣議決定から開設まで2年間、最新の技術動向を踏まえながらも、できるかぎり既存の確立した技術を応用し、手書き文書も含めた文字情報の内容検索と閲覧、印刷、画像データダウンロードに機能を集中した情報提供システムである。本論では、センターの情報提供システムの特長と言える最新の画像圧縮技術を導入した画像提供システム、歴史用語と英語に対応する専門辞書、検索情報を充実させるための原文情報や英文検索対応の目録システムを中心に紹介する。

キーワード : デジタルアーカイブ, 歴史資料, 画像圧縮技術, DjVu, Dublin Core, ISAD(G), 電子辞書, 同義語

The Japan Center for Asian Historical Records

—Toward a full-fledged digital archives—

MUTA Shohei¹; KOBAYASHI Akio²

Author Abstract: The Japan Center for Asian Historical Records of the National Archives of Japan (JACAR) was established in November 2001 for the purpose of providing people of all over the world access through the Internet to official records from collections kept by various ministries and agencies of the Japanese Government on modern history of relations between Japan and various foreign countries, primarily neighboring Asian countries. As of July 2002, the center provides about 2.3 million pieces of image data and a catalog database of 160,000 items, which will be augmented on an ongoing basis. In order to fulfill its mission, the center introduced very unique technology and system such as the DjVu document compression technology; a catalogue system relying on the General International Standard Archival Description (ISAD(G)) and Dublin Core; and a digital dictionary for synonymous, related words, and English translations specialized for Japanese modern history.

Key words: digital archives, historical records, image compression technology, DjVu, Dublin Core, ISAD (G), digital dictionary, synonymous

原稿受理 (2002-08-19)

(情報管理 45 (7): 477-483)

1. はじめに

アジア歴史資料センター(以下センター)は、「過去の歴史を直視するために、歴史図書・資料の収

集、研究者に対する支援等」を行うセンターの設立を内外に確約した1994年の村山総理「平和友好計画」に関する談話に端を発する。政府は5年間の検討の末、1999年11月30日、政治・外交的な要請、

¹独立行政法人国立公文書館 アジア歴史資料センター (〒102-0093 東京都千代田区平河町2-1-2 住友半蔵門ビル別館4階) Tel. 03(3556)8801 E-mail: smuta@archives.go.jp

¹Japan Center for Asian Historical Records, National Archives of Japan (1-1, Hirakawa-cho 2-chome, Chiyoda-ku, Tokyo, 102-0093)

²総合研究開発機構(NIRA) 政策研究情報センター (〒150-6034 東京都渋谷区恵比寿4-20-3 恵比寿ガーデンプレイスタワー 34階) Tel. 03(5448)1741 E-mail: kob@nira.go.jp

²National Institute for Research Advancement (NIRA), Center for Policy Research Information (Yebisu Garden Place Tower, 34th Fl., 20-3, Ebisu 4-chome, Shibuya-ku, Tokyo, 150-6034)

公文書館や国の記録管理制度、社会の情報化の現状を踏まえて「アジア歴史資料整備事業」の一環として、まず、国の諸機関が所蔵する「近現代における我が国とアジア近隣諸国等との関係に関わる歴史資料として重要な我が国の公文書その他の記録」（アジア歴史資料）をインターネットで提供する「アジア歴史資料センター」を国立公文書館の組織として開設することを閣議決定した。センターは2年間の準備の後、2001年11月30日に開設された。現在、国立公文書館、外務省外交史料館、防衛庁防衛研究所図書館所蔵のアジア歴史資料を提供している。3館所蔵の対象資料だけでも優に2,700万画像を超えると推定される本格的なデジタルアーカイブである。

2. アジア歴史資料センターの意義

歴史研究者にとって、日本特有の不思議な「新発見」がある。年に2回、夏と冬に歴史的な文書が「新発見」されることである。例えば、2002年7月18日付け朝日新聞朝刊に『「機密」処分、日本軍が命令』という記事がある。「敗戦直後、日本軍が組織的に公文書を処分していたことが、米国で公開された資料から浮かび上がった」とトップニュース扱いである。「公開された資料」とは、米国国立公文書館で公開されている太平洋戦争開戦前から米陸軍省が解読していた日本の暗号文書である。2001年の12月には、日本側が米国の外交文書の解読に成功していた証拠書類が、これも米国の国立公文書館で発見され、解読文書が日本の外交史料館から「発見」されたとする報道があった（毎日新聞2001年12月5日）。

実は、この「新発見」メカニズムと、センター設立の理由でもある近隣諸国との間だけでなく国内でも繰り返される「歴史認識」論争とに同一の原因が潜んでいる。つまり、わが国における歴史記録の保存公開制度の不備である。「新発見」メカニズムの根底には、内外を問わず多くの歴史研究者が日本の戦前の記録には重要なものが残っていないと思込んでいることがある。確かに大量の公文書や重要な記録が戦火や意図的な焼却のために紛失している。しかし、今回の事業対象となった国立公文書館、外務省外交史料館、防衛研究所図書館の3館に所蔵される「アジア歴史資料」はマイクロフィルム換算で2,700万コマ、資料件数にして優に120万件^{注1}を超えると推定された。その中にはこれまで研究に使用されてこなかった多くの「国

家機密」、「軍機密」資料が含まれている。

これらの資料保存機関を取り巻く環境は、諸外国と比べると見劣りするもので、わが国では歴史資料の保存公開施設や「アーキビスト」（司書にあたる文書専門家）に対する社会的な認知も低いのが現状である。そのため所蔵機関の努力にもかかわらず、資料検索に不可欠な詳細目録や「検索手段」（finding aid）を整備することすら困難な状況である。よって、所蔵機関に足しげく通い資料に精通した研究者には周知の事実でも一般には知られておらず、「8月15日」や「12月8日」にスクープを探す記者にとっては宝の山であった。そこには、「従軍慰安婦」関係資料発見報道など「周知の事実」が社会問題化し、国際的な「歴史認識」論争の発火点となる「発見」となったメカニズムが働く。

センターは、これまで公開されていても一部しか利用されなかった公的歴史資料を、最先端の技術を活用することで「いつでも」「どこでも」「だれもが」「無料」で検索し、閲覧・印刷・画像データのダウンロードができるようにする試みである。これらの資料の多くは「国家機密」「軍事機密」「用済後焼却スベシ」等の注記があるように、公表を前提に作成されたものではない。第一級の一次資料と言える。外交史料館所蔵資料には、12月の報道にあった日本側による米国国務省の暗号解読文書も含まれている。2002年8月現在、約16万目録情報、230万画像を提供している。開設から7か月で6万件を超えるアクセスを記録している。英語での検索も可能となっており、「アジア諸地域の人々との歴史認識をめぐる対話を深め、21世紀における日本と世界との共生の基盤」（有識者会議報告書より）となるべくサービスを展開している。センターは「歴史認識」論争のあり方を根本的に変える可能性を秘めている。

3. システム基本設計

3.1. アジア歴史資料センターに求められる機能

センター情報システムの基本設計を行うにあたり、以下の案件を考慮した。

センターに求められる要件としては、

- ・ 画像データのための大容量記憶装置
- ・ 巨大なデータベース検索に堪えるサーバ
- ・ 巨大なデータを保全するバックアップ装置
- ・ 情報提供のための高速インターネット回線

注1) 外交史料館所蔵資料は1簿冊を1件として換算

- ・セキュリティの強化に配慮
- ・拡張性に配慮

が考えられた。

また、このほかに情報システム関連として、

- ・検索部分における使用手法とその実装
- ・検索の中立性と精度を上げるための目録内容
- ・使用文字コード体系
- ・XMLやZ39.50への対応
- ・同義語・関連語・英語訳辞書システム

などを考慮した。

3.2. 検索システムと専用辞書

利用者の側に立ってわかりやすいユーザインタフェースや海外への対応を考え、資料閲覧のキーとなる検索においては、歴史資料検索にターゲットを絞った同義語、関連語、英語からなる専用辞書システムを新規開発・装着し、辞書のアシストがある検索システムを構築することとした。これらを実現するために資料の整理分類体系を最初から見直し、Dublin CoreとISAD (G) (General International Standard Archival Description)を参考としてセンター独自の整理分類体系を構築した。これは今後のメタデータへの対応を念頭に置いたもので、歴史資料の真正性、信頼性を担保する「オーソリティデータ」と言えるトピック（事象や事件、何を）、時間（いつ）、場所（どこで）、組織（だれが）、人（だれが）の「5つの軸」を目録データ項目

に組み入れた。これにより、所蔵機関の整理分類体系の階層に沿った検索と、所蔵機関に関係なく、「盧溝橋事件」のようなテーマや「1941年12月8日」（太平洋戦争開戦日）や「満州鉄道」などの「時間」や「組織」等の軸を組み合わせた横断的検索が可能となった。さらに「慰安婦」、「七三一部隊」、「南京大虐殺」等「歴史認識」の問題になりかねない特定キーワードの目録項目「内容」への付与は行わず、原文に使用された歴史用語を各文書の先頭300文字から原則抽出し、「内容」に付している。

センターの情報システムは構成図のとおりである（図1）。次に、特にシステム設計上重要となった画像ファイル関連の基本的な考え方と辞書のアシストのある検索システムについて詳細に述べることとする。

4. 基本設計における画像ファイルの取り扱い

基本設計にあたっては、当初予定の640万件3,200万ページにおよぶ画像の取り扱いが大きな問題となった。画像の1点は簿冊と呼ばれる文書類を合本したものとなり、A4からA3程度のサイズとなる。内容は第二次大戦前の公文書（国の行政文書）となり、手書き文字（毛筆含む）あるいは活版で印刷した形態のものが多い。

図1 システム構成図

4.1. 画像デジタル化

センターは、各資料所蔵機関から画像データと画像番号やタイトル等を収録した基礎目録情報を電子情報として受け取ることとなっている。

画像データは、各機関でモノクロのマイクロフィルムからデジタル化される。すなわち原本を一度マイクロフィルムに撮影し、そのマイクロフィルムをデジタル化することになる。マイクロフィルムを用いることにあたっては、撮影・閲覧・保存・高速デジタル化技術等のマイクロフィルム関連技術が成熟していること、今回の膨大な量の処理に関して、機材や技術者の数を考慮しても十分に大量処理が可能な点、原本を直接電子化するよりもはるかに価格格的、人的、技術的にも有利なことからマイクロフィルムを使用することが決定された。

4.2. 画像ファイル標準仕様

画像ファイル作成について、どのような解像度や階調の画像ファイルを標準とすべきかという問題があった。本画像ファイルは研究目的での利用を前提としており、できるかぎり原本に近いことが望ましいことは言うまでもない。当然、高精細な画像ファイルはサイズが大きくなる。しかし、3,200万点を保存すること、インターネット上で公開することを考慮すると、大きなファイルサイズでは、すべてを保存するには膨大な量の記憶媒体が必要であり、設計当時の一般的な記憶媒体の容量やインターネットへの接続速度を考えると現実性が薄くなってしまう。また、資料1点あたりのページ数も、多いものでは数百ページにわたるものがあり、この点からもサイズを小さくすることが必須であった。

画像ファイルの仕様において、歴史研究に堪える階調と解像度について歴史研究者や画像技術者により各種の組み合わせの比較検討を行い、文字が研究用途として原本に忠実で判読可能であり、サイズをできるかぎりコンパクトにする解決策として、モノクロ2値、400 dpiを標準とすることにした。なお、モノクロマイクロフィルムは、原則としてモノクロ2値として考えるが、微細なレベルでは4階調程度の表現力があるとの話もあった。

画像ファイル形式については、納入事業者より納品される形式と、センターにおいて保存・インターネット公開する形式を別個に定めることとし、納入形式は機種や特定ソフトウェアへの依存性の少ない（広く利用されている）TIFF（Tagged Image File Format）形式とすることとなった。公開形式は

DjVu形式を新しい試みとして採用することとした。

4.3. 高度画像圧縮技術

事業者から納入された画像ファイル形式はTIFF形式となり、このままインターネットで公開しても、画像閲覧に関して標準的なブラウザ（Internet Explorer, Netscape）がTIFF形式に対応していないことと、何よりもサイズが非常に大きくなるため、前述のとおりインターネット公開用の画像形式を別項定めることとした。

ブラウザでサポートされている画像形式のGIF, JPEG, PDFにおいても、調査の結果、サイズがTIFFより大きくなるケースも判明し、センターでは納入された画像を高度画像圧縮技術を利用した公開形式（DjVu形式）に変換して保存・公開することとなった。

4.4. DjVu形式

本形式は、モノクロ2値の文書情報圧縮ではJPEG形式の10～20倍の圧縮率となり、現状では最高レベルの圧縮性能を持つと言える。また、JPEG形式は非可逆であるがDjVu形式の文字部分は可逆画像となる。

DjVu形式は、画像ファイルを白黒の文字画像部分とその他カラー画像部分（前景と背景と呼ぶ）の3つに分解し、それぞれの部分に最も適合した圧縮方式を使用するため、非常に高い圧縮率が実現可能となる。画像を見る場合は、これらのファイルを見る側の専用ビューアーで合成して表示することになる。

約14 cm×21 cm（B5より一回り小さい程度）のサンプルを300 dpi、フルカラーでスキャンすると表1のようなファイルサイズとなる。

DjVu形式は、一般的に画像を100分の1～1000分の1にまで圧縮可能であるとされている。DjVu形式の画像を表示するには専用ビューアーが必要となる。専用ビューアーは単体でも動作するが、ブラウザにプラグインとしても組み込むことが可能であり、ホームページ閲覧時においても違和感なくDjVu形式画像を閲覧することが可能である。ビュー

表1 ファイルサイズ比較

画像形式	ファイルサイズ
非圧縮TIFFイメージ	約11.5M
JPGイメージ	約640KB
DjVuイメージ	約66.8KB

アー画面はサムネールの表示・非表示や拡大・縮小・印刷などに対応し、虫眼鏡のように一部分を拡大することも可能である（図2）。

4.5. 価格など

DjVu形式画像の作成・表示には専用のソフトウェアやビューアーが必要である。専用ビューアーは無料でダウンロードが可能であり、だれでも自由に取得して使用することができる。DjVu形式画像作成については専用の作成ソフトウェアが必要となる。DjVu形式画像作成については、作成ソフトウェアの購入で課金されずに、作成したDjVu形式画像の数での課金となる。すなわち1つ画像ファイルを作成するとそれに対して課金が発生し、権利購入数まで画像ファイルが圧縮できる仕組みとなっている。

5. アジア歴史資料センター公開資料への応用

5.1. 既存画像形式との差異

資料所蔵館よりのサンプルデータをモノクロ2値・400 dpiでデジタルイメージ化を行ったところ、JPEG形式で1ページあたり1 M～2 Mバイト程度となった。GIF、PDF形式でもサイズは同様に大きく、TIFF形式（LZW圧縮）ではその半分程度のサイズであった。ブラウザがサポートしているJPEG形式で1ページあたりの容量を平均1.5 Mバイトと見積もると、3,200万ページでは48 Tバイト（テラバイト）となり、当時の記憶装置とバックアップ装置の選定に難があった。このため解像度を下げる方

向での検討もあったが、同一サンプル群のDjVu形式のテストでは1ページあたり300 Kバイトを想定すれば問題ないという結果となった。これから記憶容量を算出すると9.6 Tバイトとなり、機器選定とインターネットでの公開の見通しがたった。

5.2. 公開画像データの現状

開設より現在まで約230万ページのデジタル化を行い、DjVu圧縮をして約400 Gバイト（ギガバイト）の容量の使用となっている。平均画像サイズは約175 Kバイトとなる。モノクロマイクロフィルムよりモノクロ2値への画像変換時に背景がほとんど残らない。その結果、DjVu形式圧縮の前景と背景のデータがほとんど0に近いため高い圧縮率が出ている。現在、センターのハードディスクは容量に余裕があるためミラーリングで運用を行っている。なお、計画では毎年200万画像を処理していく予定となっている。

6. 専用辞書による検索のアシスト

従来、大量で特定のデータを検索する場合、「キーワード」（検索用語）による検索が行われてきた。しかし、場所、時間、利用者が限定されないWebの検索環境においては、シソーラス配布による検索の運用が困難であり、利用者が自由に用語を入力できる「自由語」による検索が一般的である。このような「自由語」の検索語も、検索するための鍵となる用語という意味で一般的に「キーワード」として呼ばれている。問題は、検索に利用されると想定される言葉（キーワード）と実際に資料の

図2 画像表示画面例

中に利用されている言葉(検索対象語としてのキーワード)が異なる点であった。特に、広く一般の利用者を想定したために、現在使われている歴史用語と実際に資料に記されている歴史的用語との差を埋めるための検索システムを備えることが不可欠と考えられた。

検討の結果、原資料に使用されている歴史的な用語と現在使用されている歴史用語からなる同義語・関連語辞書を編さんし、これに読みや英語訳等を加えた、アジア歴史資料の特長を反映した辞書を作成して検索システムに搭載することが決定された。例えば「満鉄」は、「南満州鉄道株式会社」(正式名称)の同義語や「満鉄設立委員会」のような関連語に展開される。

センターの検索システムの成否は、検索対象となる目録項目データと同義語・関連語辞書の相互の整合性に大きく左右される。検索のために利用者が使うキーワードと共通用語辞書とがうまく関連づけられ、目録データに該当する言葉が抽出されていなければ検索でヒットしない。そこで、辞書の編さんにあたっては同義語と関連語について詳細な検討が行われ、できるだけ解釈が入る作業を排除する方法論が検討された。まず、入力されるキーワードと同様の基本語の抽出が必要である。

基本語は歴史用語辞典等典拠が明確な出典を利用して抽出された。利用者の対象を高校生レベルまでとしたことから、高校歴史教科書に対応する歴史用語集に収録された用語も基本語に含まれている。さらに、6か月間実際に利用された検索語約1万8,000語を形態素解析等を利用した分析を行い、歴史研究に必要と考えられる一般名詞も含めた600語を追加して、現在、基本5,600語が編さんされている。

辞書を用いたキーワード詳細検索を「タイ」を例にとって説明すると「タイ」というキーワードの辞書展開が行われ、同義語に「シャム」、「暹羅」、「泰国」、「泰國」、「siam」が表示される。英語訳「Thailand」とひらがな読み、ローマ字読みは表示されない。関連語に「ピブン元帥」、「泰緬鉄道」等の関連語が表示される。利用者はボタン操作で任意の用語を選んで展開し検索することができる。検索対象となるのは目録に収録された原文に含まれる言葉である(図3, 4)。

英文検索では、「Thailand」で検索するとタイトルおよび基本データのみが英訳されている英文目録を検索する。さらに、同義語や関連語辞書を展開すると辞書にある日本語データで展開され、日本語目録の内容を検索しヒットした目録の英文タイトルを表示する(図5)。

図3 検索画面例(1)

図4 検索画面例(2)

Organization: ☒ National Archives
☒ Diplomatic Record Office of the Ministry of Foreign Affairs
☒ National Institute for Defense Studies

Keyword: ☒ Synonyms ☒ Related Words

Search Method: ☒ Synonyms ☒ Related Words

☐ Synonyms ☐ Related Words

Date: - (Year.Month.Day)

図5 英文検索画面例

解説:
 「満洲問題」に関する文書関係資料や「済南事件」出兵に関する中国各地の状況報告や「排日及排貨」運動についての中国だけでなくアジア諸国、欧米各国、国内での反響についての報告書、「支那国慶及記念日ニ於ケル排日状況」に関する報告書、昭和2年3月の「南京事件」や同4月の「漢口事件」に関する報告、「帝国ノ対支外交政策」に関する省内の意見書や報告書類、柳井澤事件に始まる「満洲事変」から「支那事変」に関する文書、調査報告書類が含まれてい

機関:
 出所:
 シリーズ:
 サブシリーズ:
 一覧表示件数:

検索 クリア

図6 分類体系別表示画面例

7. おわりに

センターの情報提供システムの特長は、利用者の利便性を考えた多様な検索手段と大量の画像データ配信機能にある。整理分類体系が異なる組織の文書資料を、それぞれの文書整理の秩序を壊すことなく検索できるだけでなく、組織を横断しての検索が可能となっている（図6）。

目録についてはWeb上での情報検索の記述規則であるDublin Coreを参考に「表題」、「作成者」、「組織歴」、「年代域」等の基本データだけでなく、文書内容を先頭300文字程度目録に抽出することで検索対象となるデータ量を圧倒的に増やすことができた。その結果、これまで公開されていたにもか

かわらず「発見」されなかった資料をだれもが簡単に検索し利用できるようになった。歴史研究のあり方そのものを変えるものとして注目を集めている。さらに最先端の画像圧縮技術であるDjVuの採用によって、現在の通信環境でもストレスなしに大量の画像データを提供することが可能となった。センター情報提供システムは、対象物の色、材質感、立体感等を再現しなければならない「デジタルミュージアム」には不向きである。しかし、コンテンツの中心がモノクロ2値の文字情報である図書館や文書館に応用できるモデルと言える。特に、異なる組織で個々に整理された手書きも含めたいろいろな形態の文書を保存公開していかなければならない文書館にとってモデルとなるものである。

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. (オンライン), 入手先 <<http://www.jacar.go.jp/>>, (参照 2002-09-11).
- 2) 秦 郁彦. 慰安婦と戦場の性. 初版. 東京, 新潮社, 1999, 444p.
- 3) 松尾由美子. “アジア歴史資料センター構想の揺らぎとその要因”. 慶應義塾大学小島朋之研究プロジェクト卒業論文集. 2001. (オンライン), 入手先 <<http://jun-gw.sfc.keio.ac.jp/DB/ug/pdf/0907.pdf>>, (参照 2002-09-11).
- 4) 株式会社経業社. (オンライン), 入手先 <<http://www.keiyou.co.jp/>>, (参照 2002-09-11).
- 5) ZELLER, E. A History of Greek Philosophy Vol.2.1. CWRU Preservation Department Digital Library, (online), available from <<http://www.cwru.edu/UL/preserve/stack/HistGreekP2.html>>, (accessed 2002-09-11).
- 6) KAKASI project. “KAKASI—漢字→かな(ローマ字)変換プログラム”. (オンライン), 入手先 <<http://kakasi.namazu.org/>>, (参照 2002-09-11).