ARCHIVES

アジア歴史資料センターにおけるデジタル・アーカイブ

国立公文書館アジア歴史資料センター　　牟田　昌平

1　はじめに

アジア歴史資料センター（以下「センター」：www.jacar.go.jp）は、本年11月30日で開設2周年を迎えた。「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」（閣議決定：「アジア歴史資料整備事業の推進について」）をインターネットで「いつでも」「どこでも」「だれもが」「無料」で利用出来る本格的デジタルアーカイブとして、2003年11月末現在、画像データ数450万、目録データ約35万件を提供している。常に利用者の視点に立ちサービスを提供するために、定期的なアンケートや国内外での調査を行い、情報提供システムの改善に努めてきた。その結果、ホームページへのアクセス件数は、開設当初の一日平均220件から現在では800件を超えるまでになった。開設以来2年間のアクセス累計も37万件を超えている。提供するコンテンツ自体は一般に馴染みが薄い歴史的な公文書であるが、その重要性は着実に理解されてきている。本項では本格的なデジタルアーカイブとして設立に至った経緯と情報提供システムの根幹をなす画像提供システムや目録検索システムの概要について紹介するとともに、過去2年間に行った主な改善点と今後の課題について紹介する。

2　設立経緯

センターの開設は、1994年8月31日、アジア近隣諸国の人々との関係改善を目的とした談話で村山総理（当時）が過去の歴史を直視し、未来志向に立った対話が可能となる資料を提供するセンターの設立検討を表明したことに端を発する。具体的な検討は、15名の学識経験者からなる有識者会議に委ねられ、翌年、「日本とアジア近隣諸国等との間の近現代史に関する資料及び資料情報を、幅広く、片寄りなく収集し、これを内外の研究者をはじめ広く一般に提供すること」を目的とする「アジア歴史資料

センター」の設立が提言された。

有識者会議の提言を受けて、内閣外政審議室に担当部署が設置され、内外類似機関や歴史資料の保存公開に関する調査、センターの具体的な組織やシステムについて調査・検討が行われた。そして、1999年11月30日、「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「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が閣議決定された。センターは2年間の準備期間を終えて2001年11月30日に開設された。

3　近現代資料の保存・公開を巡る課題

閣議決定は、センターが「国立公文書館、外務省外交史料館、防衛研究所図書館等の国の機関が保管するアジア歴史資料を電子情報の形で蓄積するデータベースを構築し、インターネット等を通じて情報提供を行う」機関であるとした。しかし、大量の画像や目録データをインターネットで提供した先行事例もなく、準備作業は試行錯誤を重ねたものであった。

その中でも最大の課題は、本年4月に内閣府官房長が設置した「歴史資料として重要な公文書等の適切な保存・利用等のための研究会」の「中間とりまとめ」の指摘にもあるように、公文書を歴史資料として保存・利用するための基盤といえる「公文書館制度」に対する社会的認知の低さと未整備であった[1]。特に、資料検索に不可欠な目録情報を作成する専門家の不足や近現代資料の目録編纂記述に関する方法論の不備は、年間約17万件（初年度実績）という大量の目録データを作成する上で最大の障害であった。

外交史料館所蔵
レファレンス番号（B20020307503）

国立公文書館所蔵
レファレンス番号（A20030322714）

[1]内閣府の研究会及び「中間報告」については<http://www8.cao.go.jp/chosei/koubun/>を参照。

4　提供資料

センターでは、閣議決定にもとづき、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（旧陸海軍文書所蔵）の3館が所蔵する明治初期から太平洋戦争終了期までの「アジア歴史資料」を電子データで収集・提供している。

センターが提供する資料は、画像サンプルのように「国家機密」「極秘」「用済後焼却スベシ」等と注記されたものや伊14号潜水艦の図面が含まれているように、本来戦前の政府が非公開を前提として保存管理していた記録である。これら記録の多くは、戦後、米国に接収され後に返還され外交史料館や防衛庁図書館で利用に供されているものや、1971年の国立公文書館設置によって一般利用が可能になった公文書など一級の歴史資料群である。

5　インターネット配信を前提とした画像仕様

提供される資料は、画像サンプルからも明らかなように手書き文書や青焼き図面など多種多様な形態や材質の資料が含まれている。そこで問題となったのがインターネット上で画像を提供するための仕様であった。画像精度は、出来る限り原本に近いことが望まれる。一方、大量の画像データを収納し、インターネット上で公開する事を考慮すると、一点あたりの画像データは出来る限り小さい方が良い。そこで、記述内容（文字）が歴史研究用途として原本に忠実で判読可能であり、ファイルサイズを出来る限りコンパクトにしたモノクロ2値、400dpi、TIFF（Tagged Image File Format）形式を採用した。さらに実際の情報提供に当たっては、高度圧縮と操作性に優れ標準的なインターネットブラウザーに対応しているDjVu形式を採用した[2]。なお、JPEG形式での利用も可能である。

6　インターネット検索を視野に置いた目録項目

目録項目は、記録資料を検索する上で最も重要なツールである。歴史研究者の中に「目録作成作業は困難ではあっても、史料学の専門家にゆだねる必要があろう」とする意見がある[3]。しかし、近現代史料専門家の絶対的不足と方法論の未整

[2] DjVuの技術的な詳細については<http:www.lizardtech.com>または<http:www.lizardtech.co.jp>を参照。

備は既に指摘した通りである。そこで提案されたのがセンターの情報提供システムは、目録と画像データの内容とを合わせて目的の資料にたどり着くための「検索のためのツール」と見なすという考え方である。

そして採用したのが国際公文書館会議（ICA）が提唱する「国際標準記録史料記述：一般原則」（ISAD（G））（General International Standard Archival Description）[4]とわが国の公文書整理の基本単位である簿冊（主題別や時系列に整理された綴り）を基本の共通単位とした7階層からなる「目録データ階層構造モデル」（図1参照）である。これによって、文書資料整理の国際的な規則となっている「原秩序尊重の原則」を壊すことなく、異なる所蔵機関の目録データの横断検索が可能となった。

図１：各階層と各所蔵機関の目録構造比較対照表

レベル	階層名	各所蔵機関の目録レベル			
①	Super Fond	アジア歴史資料センター			
②	Fond 所蔵機関名	国立公文書館	外交資料館	防衛研究所図書館	
③	Sub-fond 出所	（例）内閣	外務省記録 調書	陸軍 海軍	
④	Series シリーズ	太政類典 公文録 公文類聚 公文雑纂 その他	旧記録 明治・大正期８門式分類 「１門　政治／１類帝国外交」 新記録 昭和戦前期16門分類法	陸軍省大日記類 陸軍資料 海軍省公文備考類 海軍資料	
⑤	Sub-Series サブシリーズ	各シリーズの細分類項目「外国交流」等	項 「１項　一般政策」等	大日記類79シリーズ 陸軍資料75シリーズ 海軍公文備考類18シリーズ 海軍資料42シリーズ	
⑥	File 簿冊	・簿冊名 「太政類典・第二編～」等	・ファイル（簿冊）件名 「幕末外交関係雑件」等	・簿冊名 「密大日記明治27年」等	
⑦	Item 件名	・簿冊内の個別件名 「清国ニ送ル国書」	・ファイル内の件名や細目 次等「遣欧使節一件」等	・簿冊内の個別の件名 「参謀本部より軍事停車場建設の件」等	

さらに、ISAD（G）の記述項目とインターネット対応型書誌項目として提唱されていたメタデータセットであるダブリンコア[5]の要素に、我が国の文書管理の実態を考慮して15目録項目（要素）を採用したのがセンターの目録項目である。管理項目を除き利用者が利用できるのは図2に表示された項目である[6]。階層構造は「機関」「出

[3] 歴史学者の立場からセンターの目録のあり方に関して問題の指摘がなされている。例としては、檜山幸夫「台湾史史料の共用化への模索」『台湾の近代と日本』（台湾史研究部会編、中京大学社会科学研究所、2003年）

[4] 『記録史料記述の国際標準』アーカイブズ・インフォメーション研究会編訳、北海道大学図書刊行会、2001年。

[5] 永田治樹「アーカイブズと図書館情報学：メタデータの相互運用」『アーカイブズの科学（上）』国文学研究資料館史料館編、柏書房、2003年pp.219-244.

[6] 小川千代子「ISAD（G）の実装：アジア歴史資料センターの階層検索システム」『レコード・マネジメント』記録管理学会、No.45, Nov.2002

所」等と「記述レベル」(⑥⑦階層)として記述されている。(図2参照)

史料専門家による要約が必要とされる「内容」のデータ作成にあたっては、各資料の本文先頭から300文字程度を原文のまま抽出することを原則とした。(図2「内容」部分を参照)勿論、300文字の妥当性については、利用者の間で議論がある。しかし、これまでのアンケート等の回答から専門家による精緻な目録作成に手間をかけ公開が遅れるより公開を優先すべしとの意見が多くこの方法は支持されているといえる。

図2：目録データサンプル

7　画像変換および目録データ作業作成の流れ

センターで提供する目録・画像データベースは、各所蔵機関で作成された目録と画像のデジタル情報を元に作成される。(図3参照)

[画像データ作成作業の流れ]

各所蔵機関は、資料をまずマイクロフィルムで撮影する。提供する資料がデジタル画像であることから当初資料収集に当たっては、原本から直接デジタル撮影を行うことも検討した。しかし、数年前の状況では技術的に確立しているフィルム撮影からスキャナーによって機械的にデジタル化したほうがコスト的にも品質的にも優位であった。デジタル技術の急速な発展の結果、コスト的な優位性は無くなってきているが公的記録に不可欠な保存媒体として信頼性や、簡単に改竄できる電子データとは異なり、マイクロフィルムの持つ真正性などを考慮すると当分の間は作業の流れの基本仕様であろう。

マイクロフィルムは、各所蔵機関によってTIFFファイルにデジタル化されセンター

に提供され、フィルム自体は、各所蔵機関で保存管理される。センターでは、TIFFファイルをDjVu形式のファイルに変換しデータベースに投入する。

［目録データ作成作業の流れ］

各所蔵機関ではデジタル画像をセンターに提供するにあたって、簿冊や件名単位で整理したリンクデータ（基本データ）を同時に提供する。リンクデータは、簿冊名、件名、CD番号、フォルダ名、ファイル番号、ファイル数など画像データを特定し管理する基礎データである。センターでは目録作成仕様にそって、作成者、作成年月日、内容、階層情報などの必要事項を抽出し目録データを作成したのちデータベースに投入する。これらの作業は基本的にアウトソーシングで行われる。

図３：アジア歴史資料情報提供の流れ

８　情報提供システムの基本機能要件

センターは、「いつでも」「どこでも」「だれもが」「無料」でデジタル化された画像データを利用出来ることを目的としている。これらの基本要件を満たすために次のようなシステムを採用した。

（1）インターネット24時間接続による情報提供システム：

原則24時間365日停止させないために、画像、目録データの追加・修正等のオンライン作業が可能なシステム、ハッカー等の攻撃から守るための24時間の監視体制、事故等によるサービス停止を最小限にするためのシステム上および組織的な対応、回線の二重化等によるインターネットとの接続の冗長性の確保等を行った。

(2)同義語・関連語・英語を含む辞書機能:

専門家だけでなく一般の利用者に対して、自由かつ簡易に目的とする当時の歴史用語で記述されている歴史資料が検索できるようにするための機能である。検索対象となる目録データに含まれる件名や先頭300文字には現代では用いない用語が含まれている。そのため目録データ検索を補助するシソーラス(類義語集)的な機能を持つ。同義語とは、当該語句を置き換えても文書(概念)の意味が変わらないものである。例えば「太平洋戦争」と「大東亜戦争」である。また、アジア/亜細亜/亞細亞のような表記の違いも含まれる。

関連語とは、ある言葉から類似・関連・連想される語で普通名詞および固有名詞を指す。例えばある事象(真珠湾攻撃)から連想される場所(真珠湾)、組織(アメリカ太平洋艦隊、連合艦隊)、日時(昭和16年12月8日または米国時間の同7日)が関連語となる。現在、基本語5600語が辞書としてシステムに搭載されている。基本語には英語検索のための英訳またはローマ字読みが付与されている。

(3)一般、学生、歴史研究専門家、外国人研究者など、多様な利用者を想定した検索システム:

(ア)階層検索

ISAD(G)の階層構造を適用して、センターを頂点(スーパーフォンド)とする7階層(レベル)からなる体系にそって各所蔵機関の資料を探索できるようにしたものである。図1にそって説明する。国立公文書館(レベル②所蔵機関名)を例に取ると「内閣」(レベル③出所)→「太政類典」(レベル④シリーズ)→「外国交際」(レベル⑤サブシリーズ)→「レベル⑤のサブシリーズに含まれる簿冊のリストの表示。例えば太政類典第二編等」(レベル⑥簿冊)→「選択した簿冊に含まれる件名リストの表示」(レベル⑦件名)となる。これによってサブシリーズ単位や簿冊単位で資料の整理された秩序にそって閲覧が可能になった。資料群の内容の全体を把握するのに便利な方法である。

(イ)キーワード検索

インターネット検索サイトで一般的に利用される自由語検索に辞書機能と年代域での絞り込み機能を付けた一般利用者を念頭に置いた検索システムで、最も利用

される検索システムである。3館が所蔵する資料を⑦レベル（件名）で横断的に検索することが出来る。歴史用語に不慣れな一般の利用者も、辞書機能を利用することで容易に当時の資料を検索することが可能となっている。例えば「太平洋戦争」（現在の歴史用語）で検索すると45件、同義語辞書（大東亜戦争、大東亞戦争等）を展開して検索すると8198件（いずれも11月末現在の検索数）となる。

（ウ）キーワード詳細検索

検索する所蔵機関の自由な組み合わせ（1館または複数）、年代域での絞り込み、さらに検索項目（すべて、表題、作成者、内容、組織歴・履歴）からの選択、辞書機能の展開、および同義語・関連語の選択など多くの検索絞り込み機能を備えており研究者を対象とした検索システムである。

（エ）レファレンスコード検索

レファレンスコードは⑦レベル（件名）の一件ごとに付与された半角英数文字12文字（例：B20020307503）で該当資料を直接検索する方法である。論文への引用や既知の資料を利用する場合の利便性を考えて追加された機能で、検索結果から各資料の「前資料」や「次資料」への移動が可能となっており前後に関連資料がないか探すことが可能である。

（オ）英語検索

日本語による検索環境がない利用者（海外の日本研究者等）に対して検索を可能にするための検索システムである。階層検索、キーワード検索、レファレンスコード検索を提供している。英語目録には、簿冊名、件名などの基本情報が英訳されている。ただし、日本語目録にある「内容」はコストなどの制約があり英訳していない。そのため、英語件名のみでは検索対象となるデータが不足し、日本語検索結果とのずれが生じる。例えば、「Pacific War」で検索すると20件しか検索されない。これは「Pacific War」に対応する用語である「太平洋戦争」が英語訳の対象となる件名（表題）に20件しかないためである。そこで辞書を展開すると「Pacific War」の訳である「太平洋戦争」（基本語）の同義語辞書に含まれる「大東亜戦争」等で日本語目録データを検索するようになっている。その結果、8136件（11月末現代）が検索される。日英目録構造の違いから日本語検索で同義語辞書を展開した場合と英語の場合に若干の差が生じるが99%以上の検索結果を得ることが可能となった。今

後は、検索に頻繁に利用される英文用語に対応する日本語を基本語として増やし、辞書展開による英語検索の精度を高めるよう努めていく。

(4)「無料」での提供：

センターが提供する資料は公文書のため原則として著作権の問題がない。インターネットでの提供を原則とし資料が広く利用されることを目的としたセンターであることから「無料」での提供が原則となった。また、「無料」にすることで利用者の利用承認も不要とし、自由に利用出来る環境を確保するだけでなくプライバシー保護の点からも利用者の匿名性を確保している。

さらに、改竄や不当な複製利用などに対しては、システム上のセキュリティー強化だけでなく、利用者自身が画像データの真贋を「いつでも」「どこでも」「自由」かつ「無料」で確認できることで対応している。開設2年間、提供されている歴史的にセンシティブな資料が不当に改竄されたり利用されたりして問題になった事例はない。

9　改善点

センターの目的である広く一般に利用されるためには、まだまだ、多くの改善が必要である。そのため、センターでは定期的にアンケート調査を行うほか、内外での聞き取り調査を行い利用者の要望を反映したシステム改善に努めている。開設2年間に行った改善や改良の主なものをあげると次のようなものがある。

・レファレンスコードによる検索および検索資料の「前資料」「次資料」への移動を可能にした。コードを論文等に引用すればコードを検索するだけで求める資料を見ることが出来る。さらにその資料の前後の資料にどのようなものがあるか資料の原秩序を確認することが可能となった。
・実験的な英語検索システムを導入したことで日本語環境以外でも検索が可能となった。
・「初心者のための利用方法」を和英双方で作成。初心者でもセンター検索システムを簡単に使いこなせるように初めての利用者を念頭に説明をしている。
・中国語とハングルによるセンター解説ページの作成。既に検索までは無理として

もより詳細な解説ページが欲しいとの利用者からの要望があり現在対応を検討中である。

・「NOT」検索の導入。例えば「インド」を検索する場合ノイズとなる「インドネシア」を除くことが可能となった。

・検索に使用した用語を表示できるようにすることで絞込検索の経緯がわかるようにした。

・「不具合情報（誤字脱字を含む）フォーム」を設置し目録データの不備や画像データの不具合（画像の欠落等）に対して迅速に対応が出来る体制を整えた。

10　センターの課題と展望

センターは、急速な技術革新との競争にさらされている。予想以上の早さでブロードバンドが普及したことによって画像数が600ページを超える大きなファイルさえ利用者のコンピューターにストレス無く提供可能となった。さらに画像データのストーレッジ容量の巨大化と低コスト化が進んでいる。そのため、現在画像スペックの基本仕様となっているTIFF400dpi、2値画像をDjVuフォーマットに変換している現状のままでよいのか検討する時期に来ている。その際、重要となるのが、OSのバージョンアップ等の技術発展に伴い発生するデータのマイグレーション（変換）コストである。将来的には、知的所有権保護の下に囲い込みが進んでいる最先端技術へ過剰に依存することなく、システム全体をオープンな規格を基本とする方向性も検討対象となろう。

さらに、デジタル化によって発生する新しい問題もある。例えば、祖父の昔の自慢話を確認しようとして祖父の名前で検索したら過去の犯罪記録が出てくるような事態も想定される。電子化された社会での人権やプライバシー保護と歴史資料の情報公開原則との比較考量という新しい課題である。

このような新しい課題があるもののセンターは着実に歴史研究者だけでなく一般にも認知されてきている。アクセス件数も開設当初の一日平均220件から800件を超えるほどとなった。今後は、単に歴史資料のデータベースとしての機能だけでなく歴史資料に基づいた研究や学習を支援する機能の充実を図り、海外での利用を進めるための英語検索機能の強化を進めていくことも肝要である。