本格的デジタルアーカイブを目指して：アジア歴史資料センターの実験

牟田　昌平

Towards a full-fledged digital archives
Experiments at the Japan Center for Asian Historical Records

Shohei MUTA

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 through the Internet the index and the digital image information of historical records about modern Japan-Asia relations held and made available to the public by Japanese national institutions such as the National Archives of Japan, the Diplomatic Records Office of the Ministry of Foreign Affairs, and the National Institute for Defense Studies of the National Defense Agency. The center is a full-fledged digitalized archives providing about 2 million images and 100,000 catalogues as of March 2002 solely relying on Internet access. The function of the center is to provide users with digitalized multi-page documents in a form that is faithful to the original, easily accessible, and searchable. In order to fulfill its objectives, 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 a digital dictionary for synonymous and related words specialized for Japanese modern history. This paper introduces these unique features of the center.

1．はじめに

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画像データとしてホームページ上で公開するとともに、目録データベースの整備などアジア歴史資料に関する統括的な情報サービスを目的として１９９９年１１月３０日設立が閣議決定された施設である。現在、約10万タイトル、２００万画像を提供する本格的なデジタルアーカイブである。本論では、センターの情報提供システムの概要と特長を中心に紹介する。

2．設立の背景と課題

アジア歴史資料センター設立は１９９４年「歴史図書・資料の収集、研究者に対する支援等」を行う事でアジア近隣諸国の人々との関係改善を目的とする村山総理談話に端を発する。政治・外交的な要請、公文書館制度、社会の情報化の現状を踏まえて、まず国立公文書館、外交史料館、防衛研究所図書館が所蔵する「アジア歴史資料」をインターネット

で提供することが閣議決定、２年間の準備の後、昨年１１月３０日に開設された。現在、２００万画像、１０万件を越える目録データを提供する本格的なデジタルアーカイブである。毎年ほぼ同程度データが追加されていく予定である。既にこれまで一部の専門家にしか存在が知られていなかった日米開戦経緯に関する米国側外公文書を暗号解読した日本側の資料が提供されるなど内外の歴史研究者から日本近代史研究のあり方そのものを根底から変える試みであるとして注目されている。(「特殊情報」で検索)

我が国の文書館を取り巻く環境は厳しいものがある。文書館そのものに対する社会的な認知も低く、博物館や図書館に比較して財政的な基盤も軟弱である。また、「アーキビスト」は博物館学芸員や図書館司書のように制度として確立したものではない。センターのデータベース構築に協力が不可欠な資料所蔵各館も慢性的な専門職員不足に悩まされている。所蔵する資料の整理分類作業だけでも手一杯で新しいセンターに資料を提供するための作業負担は困難であった。限られた人的資源を前提に、いかに短期間で大量の画像と目録データを処理するかが最大の課題であった。

３．情報提供システムの基本コンセプト

アジア歴史資料センターは、インターネットで情報提供を行う本格的な「バーチャルアーカイブ」である。成功の鍵は利用者が必要とする資料を簡便に検索し早く入手することが出来るかにかかっている。そこで、以下に紹介する情報提供システムの基本コンセプトを採用した。(注１)

統一整理分類体系

センターが提供するのは電子画像化された公文書資料である。図書と異なり原則として１点しかないユニークなものである。また、所蔵機関特有の方法で整理分類されており日本１０進分類法のような公文書館共通の整理分類方法はない。そこで既存の分類体系を横断的に整理分類するために提唱されたのが７階層からなる共通整理分類体系である。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)）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綴じられたもの）を基本の共通単位として７階層からなる「目録データ階層構造モデル」を設定した。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「出所原則」（フォンド尊重）を壊すことなく、異なる所蔵機関の目録データ横断検索が可能となった。

目録項目は、各所蔵機関で運用されている目録分類体系を整理しISAD(G)が提唱する２６要素から我が国の文書管理の実態に即して選ばれた１５項目とインターネット対応型書誌項目 Dublin Core が提唱する１５項目の特性を生かして提案された。さらにシステム搭載のための管理項目として５項目が設定された。ISAD（G）では全ての階層に渡って目録データが作成されるがセンターでは実用性を検討の上、第６階層（ファイル、簿冊）と第７階層（アイテム、件名）の２つのレベルに対して作成されている。その他の階層についてはインターネット上で階層別の解説を入れ、ツリー構造を階層に沿って検索する「階層検索」を導入することで対応した。(注２)

「アーキビスト」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。これは、図書目録

にキーワードを付与する代わりに目次データを入力するようなものである。専門家の手を煩わすことなく内容検索対象となるデータを大量に増やす事が可能となった。これは文書の先頭３００文字に歴史資料を特定するための基本的な情報である何時、何処で、誰が、何のために作成したかが記載されていることが多いとの調査結果から決定された方法である。専門家による作業とは質的に比較できないが大量の目録情報処理のアウトソーシングには不可欠の手段であった。

同義語・関連語辞書

さらにインターネットで一般的な自由後検索を可能とする為に原資料に含まれる歴史用語と現在使用されている歴史用語との乖離を埋めるための同義語･関連語辞書を作成した。例えば、一般に認知されている「太平洋戦争」は公文書では使用されていない。そこで閣議決定で正式名称として採用されている「大東亜戦争」を同義語として展開して検索できるようにした。そこで辞書の編纂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、出来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。さらに基本語の解説から同義語および関連語を収録した。最後に資料の件名データを形態素解析し基本語に対応する同義語・関連語を大学院生レベルの専門家を動員して収録した。辞書編纂作業は今後も継続されより内容の充実が図られる。

同義語・関連語の定義

	定義と事例
同義語	当該語を置き換えても意味（概念）が変わらないもので、普通名詞および固有名詞（実際の同義語には、読み、英語訳等も含む） ①漢字表記の不統一（誤字が歴史的に使われた場合も含む） ・柳条湖事件／柳条溝事件／柳條湖事件 ②歴史的に用いられた名称と現在使われている名称（旧称の使用） ・シンガポール／昭南／新嘉坡 ③同じ事象を指し示す異称 ・日露講和条約／ポーツマス条約 ④歴史用語と現代用語 ・大東亜戦争／太平洋戦争 ⑤正式名称と通称等（個人名も含む） ・下関講和条約／下関条約／日清講和条約
関連語	当該語から類似・関連・連想される語で普通名詞および固有名詞 ・石井・ランシング協定→門戸開放／石井菊次郎／ランシング／九ヶ国条約 ・西安事件→国共合作／抗日民族統一戦線／張学良／蒋介石

検索がヒットするかは最終的には目録データと検索や検索語から展開される同義語・関連語との整合性である。そのため双方のデータの抽出にあたっては共通の基準が不可欠と

考えられた。ISAD(G)では資料を特定するための用語や名称（アクセス・ポイント）として人名、団体名、地名の３つを「オーソリティ・データ」として定義付けている。さらにISAD(G)と対をなす「国際標準：団体、個人、家に関する記録史料オーソリティ・レコード」ISAAR(CPF)では記録作成者（組織、個人、家）に標準化された表記様式（典拠となる名称：オーソリティ・エントリー）を定めている。センターではこれらをもとに、歴史資料のオーソリティーエントリーといえるトピック、時間、場所、組織、人からなる「５つの軸」を目録データ抽出のための基準とした。（注３）

「５つの軸」と考え方と目的

軸	意味
① トピック（事象･事件）	歴史的な事件や事象等、一般的にキーワードとして検索に使われる。資料が何のために書かれたかを定義することが可能。
② 時間	資料固有の作成年月日や編纂された時期の情報で資料の時間的な位置を定義する。
③ 場所	資料が作成された場所や資料が関係した場所を特定する。
④ 組織	作成した組織を特定し組織の指示命令系統などを明らかにする。
⑤ 人	各資料に関係した人に関する情報で報告者や報告書の作成者、報告の対象となる事件等の関係者などを特定する。

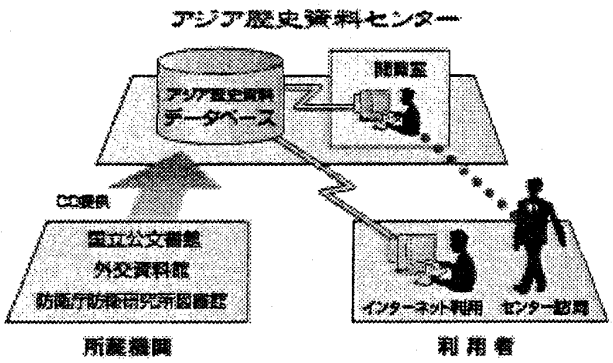
高度画像圧縮技術およびマイクロ・デジタル化の標準仕様

センターの主たる機能は文字コンテンツをいかに早く提供するかである。紙質や色の復元など高画質が必要とされる貴重書画等を扱うバーチャルミュージアムとは異なる。そこで文字情報の電子化にあたってはマイクロフィルムや画像電子化技術の成熟度を勘案して文字が研究に支障ないレベルで判読でき情報提供量を最大化できる妥協点としてモノクロ２値、４００dpi、TIFF ファイルを標準仕様として採用した。さらに現在最高レベルの画像圧縮性能（モノクロ２値ではJPEGの２０倍程）DjVuを採用した。DjVuは文字データを高度に圧縮するために開発された次世代型画像フォーマットで文字情報が判読可能な状態で高度圧縮が可能である。資料の拡縮（ズーム）やページ送り、サムネイルによる指定ページへの移動、判読が困難な部分だけの拡大機能など資料閲覧のための高度な機能を備えている。（注４）

４． センター概要（検索の実際）

「アジア歴史資料データベース」には2002年３月現在、国立公文書館、外務省外交史料館、防衛庁防衛研究所所蔵のアジア歴史資料、約200万画像、10万件を超える目録がデータベース化されている。このデータベースはインターネットを通じて、いつでも、だれでも無料で利用できる。データの軽量化のために無償提供されるビューアー（プラグイン）以外に特定のOSに依存しないオープンデータベースである。

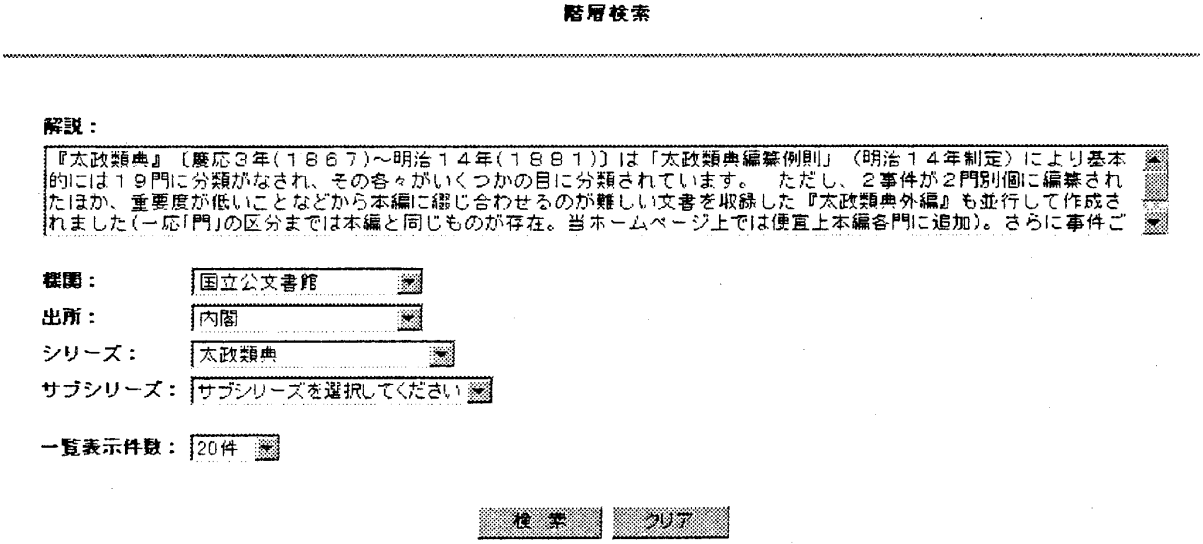
アジア歴史資料センター画像提供システムのイメージ図



さらにセンターの特長は、利用者の利便性に配慮した３つの検索システムの提供にある。

「階層検索」　各所蔵機関毎に分類をたどりながら目的とする資料群を検索するシステムである。所蔵機関→出所→シリーズ→サブシリーズと階層を下りながら絞り込んでいく方法である。選択した階層毎に簡単な解説が表示される。第６階層のファイル（簿冊）では対象となる簿冊名がリストとして表示され、第７階層のアイテム（件名）では選択された簿冊に含まれる資料の一覧が表示される。この検索は各所蔵機関の資料体系を熟知している研究者や所蔵機関別に漏れなく資料を閲覧することを想定して採用された。

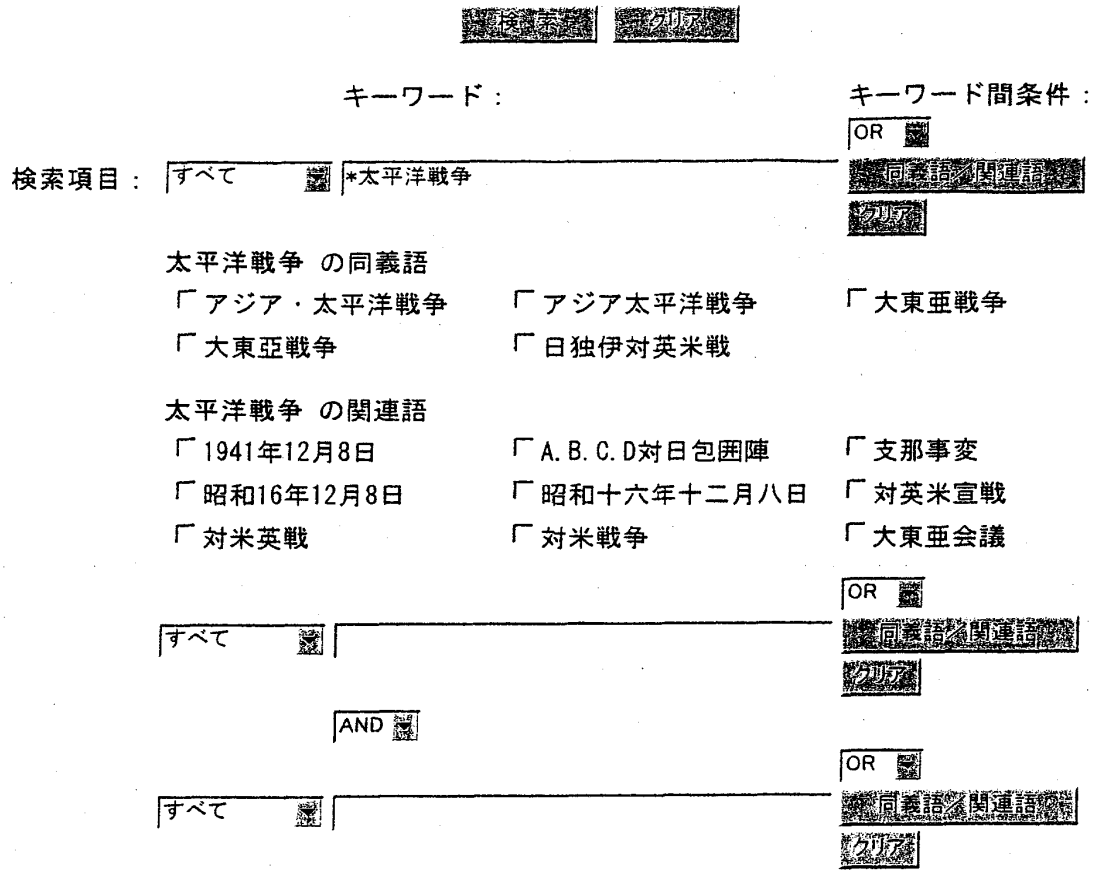
階層検索画面



「キーワード検索」　自由語検索に同義語・関連語辞書の展開機能、年代域による絞込み機能を付与している一般の利用者を想定した検索システムである。検索は第６、７階層の目録データに収録された情報について行われる。

「キーワード詳細検索」　検索対象機関の指定、検索項目（表題、作成者、内容、組織歴・履歴）毎のキーワード設定、入力したキーワードの同義語や関連語の展開や削除、年月日まで特定可能な年代域の指定など検索条件をより絞り込んだ検索が可能である。異なる所蔵機関の資料を統一された項目で検索できるアジア歴史資料センターの情報システムの特性を生かした検索方法である。

「キーワード詳細検索」での同義語・関連語辞書の展開



5． 今後の展開と可能性

時間的な制約や文書館を取り巻く環境、情報基盤の現状を考えると、センター情報提供システムに多くの改善すべき点があることは避けられない。そのためモニター制度などを活用して利用者の生の声を聞きサービスの改善を進めていく予定である。また、広く海外からの利用を促進するため多言語化への対応を検討する。特に英語での情報提供はインターネットで情報提供を行う以上不可欠と考えられる。そこで件名の英語化を行い、英語での検索を可能とするためのシステム開発を行っている。センターの情報提供システムは，大量の文字コンテンツのインターネットでの提供を主目的とするものである。文字がコンテンツの中心となる大量の行政情報公開等に新しいモデルを提供するものである。

注１．詳細に関しては「アジア歴史資料センターについて」『アーカイブズ』第８号、国立公文書館、平成１４年３月を参照

注２．目録の国際記録標準の詳細に関しては『記録史料記述の国際標準』（北海道大学図書刊行会、アーカイブズ・インフォメーション研究会編訳、２０００年）を参照

注３．ISAAR(CPF)に関しては『記録史料記述の国際標準』p.132-136 を参照

注４．Djvu の技術的な詳細は www.lizardtech.com　または www.keiyou.co.jp