研究ノート

アジア歴史資料センター「辞書検索」構築の成果と課題

―ユーザーと歴史資料をつなぐインターフェースを目指して―

松浦　晶子
中野　良

はじめに

アジア歴史資料センター（以下、「センター」という。）は、日本における本格的な歴史資料デジタルアーカイブの草分け的存在として、2001年11月に設立された。主に、国立公文書館、外務省外交史料館（以下、「外交史料館」という。）、防衛省防衛研究所戦史研究センター（以下、「防衛研究所」という。）の３機関から提供されたアジア歴史資料（近現代における日本とアジア近隣諸国との関係にかかわる歴史資料として重要な公文書等）のデジタル画像を、アジ歴データベース（以下、「アジ歴DB」という。）を通じて公開している。2016年には、センターで公開中の３機関の資料数は、2008年以来目標としていた3000万画像を突破した[1]。センターではこうした膨大な資料のなかから、ユーザーが見たいと思った資料を探し出せるように様々な取り組みをしている。そのなかの一つに、「同義語」や「表記ゆれ」などのキーワードによるOR検索で網羅的に資料を探せる「辞書検索」がある。

「辞書検索」の開発の経緯は、アジ歴DBの「生みの親」と言える牟田昌平と小林昭夫らによる複数の論考[2]から知ることができる。キーワード検索において、ユーザーが入力した用語（検索語）と、資料から抽出されたテキストデータに現れる用語の乖離を埋めるために開発されたこの機能は、デジタルアーカイブに実装するシソーラス的機能として先駆的な試みであり、牟田と小林をはじめ、その整備にあたってきた諸先輩方にあらためて敬意を表したい。

他方、「辞書検索」はアジ歴DBに特化した機能として整備が進められてきたが、現在はセンター単独での構築から集合知的な構築へと進む段階に来ている。この20年で全国の公文書館や大学などが自前でデジタルアーカイブを公開し、辞書検索のような機能へのニーズが高まってきていると推察されるからである[3]。センターと同じタイプの「辞書」「関連語」といった機能を実装している機関は、センターへの資料提供機関である国立公文書館と防衛研究所のデジタルアーカイブに限られるが、広くシソーラス機能としてとらえれば、沖縄県公文書館など実装例は存在し、今後も拡大していくものと思われる。他方で、「辞書検索」や歴史資料に即したシソーラスの構築には多大な労力がかかるため、類縁機関のあいだでのノウハウ共有が効果的だと思われる。

こうした問題意識から、本論文においてセンターの取り組みについて情報発信することにした次第である。そこで本稿では、筆者らがセンターで勤務を開始した2017年から積み重ねてきた成果をもとに、2021年４月にアジ歴DBに新たに実装された「辞書検索（カテゴリ・五十音）」について紹介したい。なお、辞書構築の過程で見えてきた課題については「おわりに」で言及する。

１　「辞書検索」の特徴

まず、2023年７月現在の「辞書検索」をもとに、その基本的な特徴について確認しておきたい[4]。

センターでは、簿冊と簿冊内の文書の件名ごとに目録データ（メタデータ）を作成し、アジ歴DBに登録している[5]。そのうち件名の目録データには、資料画像の閲覧画面につながる「閲覧」ボタンが用意されているほか、資料原文に書かれた内容（資料の作成者名称、作成年月日、資料を管理していた組織などの名称、内容（資料の先頭300文字）、写真・図・表のキャプション、調書のタイトル）を採取したテキスト情報が搭載されている（【図１】）。

【図１】アジ歴DBに登録された外交史料館提供資料の目録データ詳細情報画面。

ユーザーのキーワード検索の対象はこの目録データであり、この豊富なテキスト情報によって検索結果が大幅に増加する効果が得られている。これがアジ歴DBの大きな強みになっている。

その一方、目録データが資料の文言そのままであるからこそのデメリットもある。ユーザーの検索語と当時使われた用語、すなわち資料に書かれた用語が異なることによる、検索漏れの発生である。

例えば、現代では一般に「太平洋戦争」と呼ばれる戦争を、当時の日本政府は「大東亜戦争」と呼んでいた。アジ歴DBで３機関の資料を検索してみると、「太平洋戦争」では451件しかヒットしないが、「大東亜戦争」にすると21,384件にまで膨れ上がる。つまり、資料に書かれた用語を使って検索しないと、膨大な量の検索漏れが生じてしまうのである。

また、用語には複数の呼称、すなわち同義語があり、「大東亜戦争」の他にも「対米英戦争」などがある。さらに、「大平洋戦争」などの表記ゆれもある。しかし、同義語や表記ゆれのすべてを自力

で調べ上げるのは、たとえその分野の専門家でもなかなか困難である。とくにセンターのユーザーは、内外の研究者のみならず卒論を書く学部生や一般の歴史愛好家、ファミリーヒストリーを調査する遺族など多岐にわたり、その歴史知識には幅があるため、同義語や表記ゆれの自力での発見は期待しがたい。

こうしたデメリットを補うため、「ユーザーと歴史資料をつなぐためのインターフェースとなる辞書」というコンセプトをもとに、「辞書検索」が構築された。ユーザーが当時の用語を理解したうえで、資料に書かれた用語を使って検索を行い、探している資料までたどり着けるようにするための検索支援ツールである。

その特徴は、日本語の「基本語」を立項し、基本語に対する「同義語」と「表記ゆれ」、さらに基本語と密接に関係する「関連語」を定義した辞書データを使って検索を行うというものである。例えば、ユーザーが「太平洋戦争」でキーワード検索をした場合（【図２】）は、

①目録データと辞書データから、「太平洋戦争」の文字列を含むデータが検索される。

②検索の結果、「太平洋戦争」を含む目録データがあれば、その結果一覧が表示される。

③検索の結果、「太平洋戦争」を含む辞書データがあれば、基本語「太平洋戦争」に定義された同義語、表記ゆれ、関連語が表示される。（【図３】）

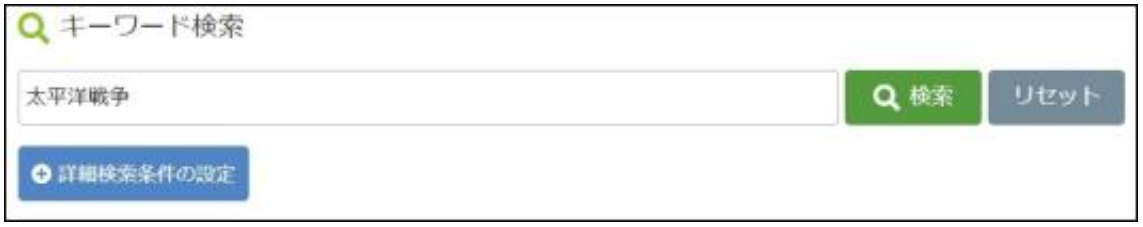
となる。また、仮に「太平洋戦争」が目録データに含まれていなかった場合でも、

①目録データと辞書データから、「太平洋戦争」の文字列を含むデータが検索される。

②検索の結果、「太平洋戦争」を含む目録データがない場合、結果一覧が表示されない。

③検索の結果、「太平洋戦争」を含む辞書データがあれば、基本語「太平洋戦争」に定義された同義語、表記ゆれ、関連語が展開される。（【図４】）

となる。



【図２】アジ歴DB「キーワード検索」の画面で、「太平洋戦争」と入力したところ。

【図３】「太平洋戦争」で検索を行ったときの検索結果一覧の画面。画面上部にある「辞書・表記ゆれを指定して検索」をクリックすると辞書データが展開され、「選択した辞書・表記ゆれで検索」をクリックすると辞書データによる再検索を行うことができる。

【図４】「太平洋戦争」で検索を行ってヒットしなかったときの検索結果一覧の画面。この場合、辞書データが自動で展開されて再検索が提案される。

このように、ユーザーが現代的な用語で検索を行ったとしても、検索結果一覧とともに表示された辞書データを見ることで、資料に書かれた用語の存在に気付くことができる。そして、基本語、同義語、表記ゆれのすべてにチェックを入れて、論理演算子 OR を選択して再検索を行えば、同義語や表記ゆれを自力で列挙する必要もなく、これらの用語を含むすべての目録データの検索結果を確認することができる。

また、資料にはユーザーが想定する用語が同義語・表記ゆれも含めて全く書かれていない場合も多い。そこで、関連語にもチェックを入れて再検索を行えば、周辺の資料にまで広がりをもたせるとともに、想定する用語が含まれていなかったとしても重要な関連資料にたどり着くこともできる。このように、「辞書検索」は簡便的確かつ網羅的な検索を可能にするのである。

ところで、「辞書検索」は、現在様々な分野の文献データベースに実装される、用語の同義関係や関連関係、階層関係などを定義したシソーラスと似た特徴をもつ。しかし、両者は主題索引の有無という部分では異なっている。

元来文献データベースにおいては、インデクサーが文献の主題を分析し、その概念を用語に変換して文献に索引語を付与し、ユーザーはそれを検索の手段に用いてきた[6]。そのなかでシソーラスは、インデクサーとユーザーが適切な索引語を選択するための索引語集としての役割を担っている。

他方、アジ歴 DB では資料に索引を付与していない。資料の主題を分析しようにも、手書き文書を含む近現代の資料を全文読解することが、現実的に困難なためである。また、たとえそれができたとしても、主題を分析し索引語を付与する際に、個人の解釈が入ることは避けられないため、センターの歴史資料に対する中立性を担保できない可能性も生じてくる。

こうした事情から、センターでは独自の「辞書検索」の構築を進めた。そして、その機能をブラッシュアップするための様々な改定を重ねてきた。次章では、2017 年から 2020 年までに行った取り組みについて述べる。

２　カテゴリ検索の構想

「辞書検索」は、センター設立期の 2001 年より構築が始められたが、それから 16 年が経過した 2017 年、ユーザーにとってより使いやすい機能にすることを目指して、「辞書検索」の機能を本格的に改善することにした。その際、主に、①検索手段の追加[7]と、②構築法の改定の２点に重点が置かれた。

まず①について、2017 年の時点では、辞書データを利用するための検索手段として、「キーワード検索」と「五十音検索」の２つが用意されていた。とくに前者は、ユーザーにとってもっとも基本的な検索手段である。

ところが、「キーワード検索」の画面からは、ユーザーが辞書データに登録されている用語を確認することはできない。第１章でその仕組みを述べたとおり、ユーザーが検索を行い、検索語が辞書データにある文字列と一致した場合に初めて、検索結果一覧とともに辞書データが表示される。これは、最初から辞書データを使って検索したいユーザーにとっては不便である。

こうした場合に便利なのが後者である。「五十音検索」は、辞書データの基本語を五十音順にリスト化したものであり、基本語をクリックすることで、基本語による検索が実行されて、検索結果一覧とともに辞書データが表示されるようになっている。

ただし、あらかじめ基本語が判明している場合は便利だが、漠然とした関心をもとに資料を探そうとする場合は、五十音の「あ」から順に基本語をたどっていくのは効率的ではない。この問題を解決するためには、辞書データを主題別に分類した検索手段を用意することが有効であると考えられた。

そこで、2017 年 12 月に「アジ歴地名・人名・出来事事典」（以下、「アジ歴事典」という。）[8]という WEB コンテンツを公開した（【図５】）。これは、辞書データに登録されている基本語の一部を、「地名」、「人名」、「出来事」のカテゴリに区分し、３階層構造のツリー形式で表示したものである。

「地名」、「人名」、「出来事」の下位カテゴリは、当時のセンター公開資料が主に明治維新期から太平洋戦争終結時までに作成されたものであったことをふまえ、その時代の呼称や概念をもとに区分されている。そのため、例えば「地名」の下位カテゴリを構成する「地域名」には、「樺太」、「満洲」、「南洋群島」などが含まれており、現在の呼称と一致していない点に注意する必要がある。また、これらの分類は政府の公式見解ではないことをお断りしておく。

【図５】「アジ歴地名・人名・出来事事典」のトップページ。

特筆すべきは、基本語から直接的に検索を行うのではなく、検索を行う前に基本語についての情報をまとめたページ（以下「基本語ページ」という。）を経由するようにしたことと、ユーザーが基本語の意味を理解したうえで検索ができるように、従来の辞書データにはなかった解説を基本語ページに搭載したことである。「ユーザーと歴史資料をつなぐためのインターフェースとなる辞書」という「辞書検索」のコンセプトをさらに具体化したといえるだろう。

例えば、「満洲」の地名を調べたいときは、【図６】にあるように「地名を検索」を選択したうえ

で、

地名

└地域名「満洲」

└地名「長春」「ハイラル」「ハルビン」「奉天」「満洲里」

と上位から下位へとカテゴリをたどれば、「満洲」の地名一覧を確認できる。

【図６】「地域名」（満洲）から、「地名」の一覧を確認する画面。

【図７】が、「満洲」の地名「ハルビン」を選択して表示された基本語ページである。基本語とその解説、さらに基本語に定義された同義語、関連語と表記ゆれが表示される。「このキーワードで検索」をクリックすると、基本語「ハルビン」での検索が実行されて、資料の検索結果一覧が表示される。もし同義語や表記ゆれを使いたい場合は、辞書データからそれらを選択して再検索を行うことができる。

【図７】「ハルビン」についての基本語ページ。

「アジ歴事典」を公開した翌月の 2018 年１月分の「アジ歴事典」トップページへのアクセス数は、937 件であった。センターで公開している全 WEB コンテンツのアクセス数の平均が 419 件だったのに比べて比較的多く、ユーザーの一定の需要があることが分かった。

ただし、「アジ歴事典」は独立した WEB コンテンツとして公開されたため、アジ歴 DB に実装された「五十音検索」との機能的連携が十分ではなかった。そのため、将来的には「アジ歴事典」がもつ機能をアジ歴 DB に実装することが期待された。これについては第５章で述べる。

３　構築法の改定

次に、②構築法の改定である。センター設立期より辞書データの整備は、牟田・小林らによる構築法（以下、「旧構築法」という。）を軸に進められた[9]。当時、センターの資料公開数は 200 万画像を超えていたとはいえ、目録データの情報はまだ十分に揃っていないなかで、歴史用語辞典などにある現代用語をもとに、5,600 件の基本語を辞書データに登録したものである[10]。

それ以降も辞書データの整備が進められたが[11]、この間にも資料の公開が進み、前述のとおり 2016 年には 3000 万画像を突破した。さらに、従来はアジ歴 DB の性能により目録データに搭載できる文字数に制限があったが、技術の発達とともに搭載できる文字数が漸次増大した。そのため、目録データに出現する用語の種類は増加の一途をたどり、2017 年３月時点で辞書データの基本語には 11,327 件が登録された。

こうした状況を受けて、2017 年から辞書データの整備を本格的に行うことにした。整備にあたっての基本的な方向性は、以下の２点である。

・既存の基本語の改定（不要と考えられる基本語の削除、基本語への同義語・表記ゆれ・関連語の追加、基本語と同義語の関係整理）を行うこと。

・ユーザーのニーズが高いと考えられる用語を、新規に基本語として立項すること。

これらの方針に沿って、目録データの情報が豊富に揃っている現状に合わせた新しい構築法（以下、「新構築法」という。）をもとに整備を行った（【表１】）。

【表１】旧構築法と新構築法における基本語、同義語・表記ゆれ、関連語の基本原則の対照表。

旧構築法[12]	新構築法	主な変更点
基本語	基本語	基本語
利用者が検索に用いる用語と、歴史専門用語辞典等に記載される普通名詞および固有名詞。特定の歴史事象に関する現代用語や正式名称に対する略称等も含む。また英語検索に頻繁に利用される用語に対応する日本語の普通名詞及び固有名詞も基本語とする。	目録データに出現する、地名・人名・出来事・組織などの固有名詞、ならびに目録データ上で現在と異なる表記が確認された普通名詞。同じ意味の用語が複数存在する場合は、正式名称または現代で一般に知名度が高い名称を基本語とする。ただし、正式名称または一般に知名度が高い名称であっても、ヒット数が０件あるいは極端に少ない場合は、同義語として登録する。	・旧構築法ではユーザーの検索語や歴史用語辞典などから抽出した現代用語であるのに対し、新構築法は目録データ中から抽出した資料用語とした。 ・新構築法では、基本語のヒット数が０件あるいは極端に少ない場合は、その用語を基本語にはせず同義語とする。
同義語・表記ゆれ	同義語・表記ゆれ	同義語・表記ゆれ
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。	目録データに出現する、基本語と置き換えても同じ意味・概念になるもので、普通名詞および固有名詞。	変更なし。
関連語	関連語	関連語
当該語に関わりのある語で普通名詞および固有名詞。関連語の抽出は原則として目録データに含まれる用語を対象とする。関わりのある語とは当該語を歴史的に特定する要素である「事象、人物、場所、組織、時間」を示す用語。	目録データに出現する、基本語と関係する地名・人名・出来事・組織などの普通名詞および固有名詞。	新構築法では、「時間」は対象外とした。

まず、既存の基本語については、アジ歴 DB で基本語とそれに定義された同義語、表記ゆれをすべて使って検索しても、何もヒットしなかった場合に、その基本語ごと削除した。１件以上ヒットした場合は、その基本語に対して目録データから抽出した新たな同義語や表記ゆれ、関連語を追加す

る作業を行った。また、１件以上ヒットしても、基本語のヒット数が０件あるいは極端に少なく、同義語のヒット数が多い場合は、両者の関係を整理して、基本語と同義語を置き換える作業を行った。

さらにこのとき、基本語に定義されていた関連語で不要と思われる用語、すなわち旧構築法に定められていた関連語の「時間」も削除することとした。旧構築法では、関連語は「……当該語を歴史的に特定する要素である「事象、人物、場所、組織、時間」を示す用語である」と定義されていたため、関連語には「時間」としての年月日（西暦と和暦の「〇年〇月〇日」）が登録されていた。

しかし、例えば基本語に条約名を登録する場合、条約を関連資料から考察するためには、国家間の交渉、条約の調印、批准、批准書の交換といった締結までの過程が重要となる。そのため、関連語に特定の年月日だけを登録することは適切ではないと考えられた。だからといって、関連する日付をすべて登録すると繁雑になってしまう。これは条約に限らず他の出来事についてもいえることであった。

次に、新規の基本語について述べる。センターはレファレンス業務の蓄積をふまえ[13]、ユーザーの関心が高いと考えられる陸海軍や植民地関係の資料などを対象に、地図や組織図などを通じて関係する用語にたどりやすくすることでキーワード検索を支援する「アジ歴グロッサリー」というWEBコンテンツを、2015年から隔年で公開していた（【図８】）。

【図８】2018 年１月公開のアジ歴グロッサリー「公文書に見る外地と内地－旧植民地・占領地をめぐる人的還流－」より、植民地官僚の経歴図。官僚の姓名をキーワードにしてアジ歴 DB で検索ができる。

そこで紹介される用語は、資料に関連の深い人物名（植民地官僚など）や組織名（陸海軍の部隊、植民地に設置された行政機関や民間企業など）である。キーワード検索で利用されうる基礎的な用語は、2001 年以降の整備で辞書データに登録済みであったが、アジ歴グロッサリーで紹介される用語はあまり登録されていなかった。そこで、ユーザーが辞書データを使ってこうした用語で検索ができるように、新規に基本語を立項した。

このように、既存の基本語の削除・追加・修正の作業と、新規の基本語の立項を行うことで、辞書データを充実させていったのである。この整備の結果、2020 年 10 月時点で登録された基本語は 11,589 件となった。2017 年３月時点の 11,327 件から大きな変動はないが、実に膨大な作業量であった。ただし、この整備は、2021 年４月のアジ歴 DB のシステム更新を受けて、仕切り直しとなった。これについても第５章で述べたい。

４　歴史資料における地名と人名の書かれ方

2017 年からの整備の過程で、公文書をはじめとする近現代の歴史資料において用語がどのように書かれるかにつき、様々な事例を蓄積することができた。そこで、そのなかから特徴的な事例をいくつか紹介してみたい。

辞書データの整備にあたっては、専門知識をもつスタッフが、目録データから基本語とその同義語、表記ゆれを抽出し、辞書データに登録を行う。基本語のなかには、同義語や表記ゆれを大量に含むものがある。とくにカタカナ表記はゆれの幅が大きく、人力で抽出することが困難なときもある。その場合は、n-gram[14]やレーベンシュタイン距離[15]といった統計的手法を用いて、目録データから表記ゆれ候補を抽出したリストを作成している。ただし、あくまで機械的に抽出したものであるため、最終的にはスタッフがリストにある用語を一語ずつ検証してから登録を行う必要がある。

まず、地名の書かれ方からみていきたい。地名のうち外国地名には、同義語と表記ゆれが多い。とくに戦前期には、外国地名の日本語表記に統一基準がなく、様々な書かれ方をされたためである。

表記ゆれが多い外国地名としては、ロシアの地名が代表的であろう。例えば、現代は一般的に「ウラジオストク」と書かれる地名の表記ゆれは、現在センターが把握しているだけでも、以下の 34 語がある。

①ウラジウォストク②ウラジウオストツク③ヴラジヴオストツク④ヴラジヴオトツク⑤ウラシオストーク⑥ウラシオストツク⑦ウラジオストツク⑧ウラジオストック⑨ウラシホストク⑩ウラシホストツク⑪ウラジホストツク⑫ウラジボストツク⑬ウラジボストック⑭ウラジワストク⑮ウラシワストーク⑯ウラジワストーク⑰ウラシワストック⑱ウラジワストック⑲ウラジヲストツク⑳ウラジヲストック㉑ウラヂウォストーク㉒ヴラヂヴオストーク㉓ウラヂウォストック㉔ウラチーウヲストツク㉕ウラチーヴヲスーツク㉖ウラヂオ㉗ウラチオストク㉘ウラヂオストク㉙ウラヂオストーク㉚ウラヂオストツク㉛ウラヂワストク㉜ウラヂヲストツク㉝ウラデイワストツク㉞ウラデオストック

ロシア語に関する研究を参照しつつ大まかに分類すると、ロシア語表記の「Владивосток」を、文字どおりカタカナに翻字したとみられる表記（ヴラヂヴオストークなど）や、実際の発音を転写したとみられる表記（ウラヂワストクなど）の少なくとも２パターンがあると分かる[16]。

また、カタカナ表記の他に、漢字表記や漢字＋カタカナ表記も存在している。現在把握しているのは以下の 14 語である。これらは「ウラジオストク」の同義語として辞書データに登録されている[17]。

①海参崴②海参威③烏拉日阿斯徳④烏拉地阿斯徳⑤浦塩⑥浦塩斯徳⑦浦塩須徳⑧浦潮⑨浦潮斯徳⑩浦潮須徳⑪浦汐⑫浦汐斯徳⑬浦塩ストツク⑭浦塩ストック

アジ歴 DB でこれらの用語を使って３機関の資料を検索してみると、「ウラジオストク」は 109 件がヒットするが、表記ゆれの①～㉞だと 234 件、同義語の①～⑭だと実に 15,227 件がヒットする。「ウラジオストク」だけで検索を終えてしまうと、膨大な検索漏れが発生することが分かる。少なくとも外国地名には漢字表記、漢字＋カタカナ表記、カタカナ表記が存在する可能性があり、それらが無数のバリエーションで書かれていることをおさえて、検索を行う必要がある。

次に、人名の書かれ方である。外国人名のカタカナ表記は、地名と同じく表記ゆれの幅が大きい。他方、日本人の書かれ方も日本人ならではの特徴がある。ここでは日本人の事例について取り上げてみたい。

日本人は、姓名（フルネーム）で書かれるほか、組織に所属している人物であれば、組織での役職もあわせて書かれることが多い。もっとも正式な表記法は、役職＋姓名（例：文書課長田中太郎）である。フルネームで書かれているため、人物の特定をしやすい。注意すべきは、姓＋役職（例：田中課長）という表記法で、一見してどこの誰なのか分からないこともあるため、前後の内容などから判断し、人物を特定する必要がある。

姓＋役職の書かれ方は、中央官庁の職員か、在外の外交官か、軍人かによっても異なってくる。中央官庁の職員は、姓＋役職（局長、部長、課長など）で書かれることが多い。他方、外務省には本省に勤務する職員のほか、在外公館に勤務する外交官がおり、階級（大使や公使など）があるときは姓＋階級で書かれる。

例えば、外交官の「白鳥敏夫」は以下のように書かれている。①～⑧は在外公館時代の階級や役職、⑨～⑯は外務省本省時代の役職にあたる。そして、階級や役職のそれぞれの書かれ方にもバリエーションがある。アジ歴DBで「白鳥敏夫」で検索すると229件、①～⑯で検索すると226件となり、ほぼ同等の件数がヒットする。

①白鳥大使②在伊白鳥大使③白鳥公使④白鳥両公使⑤白鳥特命全権公使⑥在瑞典白鳥公使⑦瑞典国駐箚白鳥公使⑧白鳥外交官補⑨白鳥情報部長⑩白鳥外務省情報部長⑪白鳥部長⑫白鳥情報局長⑬白鳥文書課長⑭白鳥翻訳課長⑮白鳥課長⑯白鳥事務官

軍人の場合は、中央官庁である陸軍省や海軍省、参謀本部などに在籍している者の場合は、姓＋役職で書かれることも多いが、基本的には姓＋階級（大将、大佐など）で書かれる。姓＋陸軍（海軍）＋階級、姓＋兵科＋階級などで書かれることもある。さらに、部隊や艦隊での役職（師団長、連隊長、司令長官、艦長など）があるときは、姓＋役職で書かれる。

例えば、「東郷平八郎」は以下のとおりである。①～⑦が軍人の階級、⑧～㊱が官衙や艦隊での役職にあたる。また、東郷平八郎のように著名な軍人の場合は、㊶のような尊称もある。アジ歴DBで「東郷平八郎」で検索すると2,480件、①～㊶で検索すると1,027件となる。

①東郷元帥②東郷老元帥③東郷大将④東郷海軍大将⑤東郷海軍中将⑥東郷将軍⑦東郷中佐⑧東郷提督⑨東郷海軍軍令部長⑩東郷軍令部長⑪東郷連合艦隊司令長官⑫東郷連合艦隊長官⑬東郷連合艦隊司令官⑭東郷連合長官⑮東郷連長官⑯東郷連長長官⑰東郷同司令長官⑱東郷第一艦隊司令長官⑲東郷第一艦隊司令官⑳東郷常備艦隊司令長官㉑東郷常備艦隊司令官㉒東郷第三艦隊司令官㉓東郷横須賀鎮守府司令長官㉔東郷舞鶴鎮守府司令長官㉕東郷佐世保鎮守府司令長官㉖東郷佐鎮司令長官㉗東郷佐世保鎮守府シレイテウカン㉘東郷司令長官㉙東郷長官㉚東郷丁長官㉛東郷呉鎮守府参謀長㉜東郷浪速艦長㉝東郷鳥海艦長㉞東郷艦長㉟東郷海軍技術会議議長㊱東郷技術会議々長㊲東郷議定官㊳東郷御学問所総裁㊴東郷総裁㊵東郷侯爵㊶大東郷

このように、日本人の人名は、フルネームだけでなく、姓＋役職で書かれることが多く、フルネ

ームだけでキーワード検索を行うと、不十分な結果となる恐れがある。人名を検索するときは、その人物の履歴をふまえ、姓＋役職のキーワードも使って検索を行うのが適切であるといえる。

新構築法のもとではこうした表記ゆれや同義語の存在にも注意を払い、より網羅的な検索を可能にする辞書データの整備につとめている。旧構築法に比べ、「辞書検索」の機能を使うメリットは格段に高まっているといえるだろう。

５　「辞書検索（カテゴリ・五十音）」の実装

2021 年４月にアジ歴 DB システムが更新されることを受け、2020 年 10 月から各種データの移行作業が開始された。このとき、辞書データをそのまま移行するのではなく、2017 年から検討を重ねてきた成果を反映したうえで移行することにした。

まず、新しいアジ歴 DB に、従来の「キーワード検索」と「辞書検索（五十音）」に加えて、新規に「辞書検索（カテゴリ）」を実装した（【図９～11】）。これは、第２章で述べた「アジ歴事典」のカテゴリ検索の構想をアジ歴 DB 上で実現したものである。

【図９】新しいアジ歴 DB に実装された「辞書検索（カテゴリ・五十音）」の画面。

【図 10】「辞書検索（カテゴリ）」の画面。

【図 11】「辞書検索（五十音）」の画面。

「辞書検索（カテゴリ）」のカテゴリ区分は、「アジ歴事典」を土台にしているが、変更した部分が２点ある。１点目に、「辞書検索（カテゴリ）」の上位カテゴリに、従来の「地名」、「人名」、「出来事」の他にも、新規に「組織名・機関名」と「条約・協定」を追加し、全部で５つのカテゴリを表示するようにした。２点目に、上位カテゴリはそれぞれ下位カテゴリを有しているが、「アジ歴事典」が３階層構造であるのに対し、「辞書検索（カテゴリ）」は３階層以上が可能になっている。例えば、「アジ歴事典」にある「地名」カテゴリは、

地名
　└地域名「ベトナム・ラオス・カンボジア」
　　└地名「仏領インドシナ」「サイゴン」「ハイフォン」「ハノイ」
　└地域名「タイ」
　　└地名「タイ」
　　⋮

とあるように、「地名」カテゴリの直下の「地域名」カテゴリに国名が並んでいたが、「辞書検索（カテゴリ）」では、

カテゴリ１「地名」
　└カテゴリ２「東南アジア」
　　└カテゴリ３「ベトナム・ラオス・カンボジア」
　　　└カテゴリ４「仏領インドシナ」「サイゴン」「ハイフォン」「ハノイ」「ベトナム」etc.
　　└カテゴリ３「タイ」
　　　└カテゴリ４「タイ」「チェンマイ」「バンコク」「プーケット」
　　　⋮

とあるように、カテゴリ２「東南アジア」という地域区分を用意することで、分かりやすく整理した。

次に、「辞書検索（カテゴリ・五十音）」からカテゴリあるいは五十音をたどっていくと、最終的に用語詳細ページにたどり着く（【図 12】）。用語詳細ページもまた「アジ歴事典」の基本語ページにならったものであるが、「アジ歴事典」から２点ほど変更を加えた。

【図 12】用語詳細ページ。「解説」、「参考資料」、「基本語（日本語）」、「基本語（英語）」、「同義語」、「関連語」、「表記ゆれ」、「上位カテゴリ」などの項目がある。

１点目が、「見出し語」の追加である。これは、基本語を検索するためのいわば索引であり、とくにユーザーが想定しにくい用語が基本語として登録されていた場合に、効果を発揮する。例えば「日中戦争」の場合、現代用語の「日中戦争」と資料用語の「支那事変」でアジ歴 DB をそれぞれ検索すると、後者は 40,046 件がヒットするが、前者は２件しかヒットしない。そして第３章で述べたとおり、新構築法ではヒット数の極端に少ない用語は基本語にはしないため、

基本語：「支那事変」
　└同義語：「日中戦争」etc.

とあるように、基本語に「支那事変」を登録し、同義語に「日中戦争」を登録している。しかし、「辞書検索（カテゴリ・五十音）」を利用しようとするユーザーは、恐らく「日中戦争」の方を想定

して探す可能性が高いと考えられる。つまり、「辞書検索（五十音）」を利用したときに、「さ行」の「し」ではなく、「な行」の「に」から探すはずである。そうすると、「支那事変」を基本語にするのは都合が悪い。そこで、ユーザーが用語詳細ページにたどり着きやすくなるように、索引としての見出し語を追加することにした（【図 13】）。

【図 13】画面上部の太枠で囲った部分が見出し語。「このキーワードで検索」をクリックすると、見出し語ではなく基本語で検索される。

とはいえ、「日中戦争」のような例は少なく、多くの辞書データでは見出し語と基本語は一致している。また、見出し語はあくまで索引用であり、見出し語を使って検索することはできず、基本語で検索されるようになっている。

２点目が、英語版の作成と「基本語」（英語）の追加である。アジ歴 DB は海外ユーザーの需要も大きいことから、英語版も作成している。また、目録データに出現する用語については、センターが組織した「データ検証委員会」で検証を重ね、標準的な英訳を作成している[18]。そこで、用語詳細ページについても英語版を作成するとともに、日本語版でも「データ検証委員会」で検証済みの英訳を基本語についてのみ提示することにした。これにより、海外ユーザーに対する利便性向上を図った。

以上のように、2020 年 10 月から、カテゴリ・五十音の区分や、見出し語、英訳を追加しながら、辞書データの移行作業を行った。2023 年 7 月現在、移行が完了した基本語は 8,814 件である。現在もなお移行作業は続いている。

おわりに

本稿では、2017年から取り組んできた「辞書検索」構築の過程とともに、2021年に新しいアジ歴ＤＢに実装した「辞書検索（カテゴリ・五十音）」について紹介した。ユーザーは「辞書検索（カテゴリ・五十音）」という２つの検索手段から用語詳細ページにたどり着き、そのなかから適切な用語を選択し、資料の検索を行うことが可能になった。これにより、「ユーザーと歴史資料をつなぐためのインターフェースとなる辞書」というコンセプトをほぼ実現できたといえよう。最後に課題について述べたい。大きな課題は２点ある。

１点目が、新たな辞書データ作成原則の確立である。現在、辞書データは新構築法に則して順次登録と修正を進めているが、「どのような用語について辞書を作成するか」という観点からすると、必ずしも系統的もしくは論理的な原則を確立できていない。実態としては、アジ歴グロッサリーの新規作成にあたって検索語として選定された人名や組織名などを機械的に辞書データとしても登録するか、日常業務のなかで同義語・表記ゆれの存在に気付いた用語をその都度登録するほかない状況である。この点において、歴史用語辞典などから一定の規範性にもとづいて基本語を選定していた旧構築法を完全には代替できていないことになる。このような進め方では担当者の関心や知識によって整備される領域に偏りが出てしまい、ユーザーにとって有益な辞書データの整備が進まないケースも出てくることが危惧される。新構築法のコンセプトをふまえつつ、より系統的な辞書データ整備の原則や枠組みを構想する必要がある。

こうした辞書データの体系性に関連して、辞書のカテゴリ分類も課題である。現在の辞書データでは、戦前の呼称や概念をもとにカテゴリを区分しているが、戦後に一般化した現在の地域区分などとの整合性はほとんど考慮されていない。他方、2017年８月から外交史料館『戦後外交記録』の公開がアジ歴DB上ではじまり、今後さらに戦後資料の割合が高まるため、地域区分を再考する必要が生じると思われる。

２点目が、効率的な同義語・関連語の抽出である。本論でも述べたように、外国語の表記ゆれに関しては、膨大な目録データのなかから効率的に抽出する方法として、n-gramやレーベンシュタイン距離の有効性が確認されており、それにもとづいて多数の辞書データが作成されてきた。それに対し、同義語や関連語については同様な形で活用できる手法がいまだなく、本論に示したようなコンセプトにもとづいて担当者が手作業で同義語・関連語の抽出や選定を行っている。抽出された情報の信頼性・確実性という点で、性急な自動化には慎重であるべきなのは勿論だが、限られた人員と時間のなかで効率的に辞書データを作成していくためには、より合理的な手法を検討・開発していくことも求められるだろう。

これに関連して、「辞書検索」の構築にかかる多大な労力を削減するために、各機関がもつ辞書データを蓄積して共有する集合知的なデータベースを共同構築したり、辞書データを公開したりすること[19]も有効ではないかと考えられる。こうした辞書データの共有と公開は、検索支援のいっそうの進展や構築にかかる労力削減だけではなく、自然言語処理や言語学といった研究分野の言語資源として活用されるなど、様々な利点がある。「辞書検索」ないしはシソーラスがデジタルアーカイブの基本的機能として実装されることが今後も見込まれるのであれば、このような仕組みを作ることも一案であろう。

このように、残された課題はまだまだ多いが、辞書機能をさらに向上・発展させていくためには

取り組むことが不可欠な課題でもある。それらを少しずつでも実現させていくべく、挑戦を続けていきたい。

---

[1] 波多野澄雄「Ⅰ．序文―世界に開かれた「デジタル文書館」として―」（『アジア歴史資料センター20年の歩み』国立公文書館アジア歴史資料センター、2021年）、1頁を参照。3000万画像という目標が示された高度情報通信ネットワーク社会推進戦略本部の「重点計画2008」は同書36頁。また同書は、https://www.jacar.go.jp/about/documentstable/JACAR_20years_of_history.pdf（最終アクセス日：2023年11月30日）からも閲覧できる。

[2] 「辞書検索」に関する論考として、牟田昌平「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―」（『情報知識学会第10回（2002年度）研究報告会講演論文集』2002年）、牟田昌平「インターネット上での国際的な歴史記録の共有を目指して―アジア歴史資料センターの事例―」（『じんもんこん2002論文集』2002、2012年）、牟田昌平・小林昭夫「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―」（『情報管理』45-7、2002年）、牟田昌平・小林昭夫「専門語を含むデータの検索における辞書の有用性―アジア歴史資料辞書その編纂作業から運用まで―」（『情報知識学会誌』14-2、2004年）などがある。

[3] センターと横断的検索を行うためリンク提携している機関に限っても、北海道立文書館、北海道立図書館、琉球大学、神戸大学、滋賀大学、大分大学、東洋文庫、アジア経済研究所などが独自のデジタルアーカイブを公開している。

[4] アジ歴DBのヒット件数は、2023年７月１日現在の件数である。

[5] アジ歴DBの目録項目のうち、３機関から提供されるのは「簿冊・件名標題」「資料群階層」「所蔵館における請求番号」のみであり、それ以外の「内容（先頭300字）」や「作成者」「組織歴」、「写真・図・表のキャプション」「調書」などの項目はセンターが独自に設定し、外部業者に発注して作成しているものである。

[6] F.W.ランカスター著、松村多美子・鈴木祐滋訳『情報システムのためのシソーラスの構築と利用』（情報科学技術協会、1989年）、１～３頁を参照。

[7] 検索手段の追加については、大野太幹「アジ歴データベース担当としての回顧」（前掲注１『アジア歴史資料センター20年の歩み』）でも簡易報告されているので参照されたい。

[8] 「アジ歴地名・人名・出来事事典」https://www.jacar.go.jp/dictionary/index.html（最終アクセス日：2023年７月１日）。

[9] 旧構築法の概要は、前掲注２にある各種の論考を参照されたい。

[10] 前掲注２「インターネット上での国際的な歴史記録の共有を目指して―アジア歴史資料センターの事例―」、206頁を参照。

[11] 2008年までの辞書データの整備については、平野宗明「アジア歴史資料センターから見たデジタル・アーカイブズの現在と展望」（『アーカイブズ学研究』10、2009年）を参照されたい。

[12] 前掲注２「専門語を含むデータの検索における辞書の有用性―アジア歴史資料辞書その編纂作業から運用まで―」、38頁を参照。

[13] ユーザーからのセンターに対する問い合わせのうち、研究者からの学術的な資料レファレンス以外のものとしては、ファミリーヒストリーに関する質問が多い。例えば、戦地や外地にいた家族がどこの地域のどの部隊に所属していたか、どのように復員・引揚したのかを知りたいといったも

のである。

[14] n-gramは、テキストを連続するｎ個の文字列で分割する手法である。例えば、「ウラジオストク」を2gramで分割して「ウラ」「ラジ」「ジオ」「オス」「スト」「トク」とする。これらの文字列を、目録データのテキスト情報と比較して、似ている文字列を抽出する。

[15] レーベンシュタイン距離は、ある文字列を別の文字列にするために、挿入・削除・置換の作業を何回実行すればよいかという文字列間の距離を求める手法である。例えば、「ウラジオストツク」を「ウラジオストク」にするには「ツ」を削除するため１回、「ウラシオストツク」は「シ」を「ジ」に置換して「ツ」を削除するため２回となる。回数が少ないほど２つの文字列が似ていることになるため、目録データのテキスト情報を解析して回数が少ない文字列を抽出する。

[16]「Владивосток」の日本語の漢字・カタカナ表記の変遷については、シャルコ・アンナ「第五章　外国地名の漢字表記と和風化―「樺太」（カラフト）、「浦潮」（ウラジオストク）を中心に―」（『日本語の表記体系における漢字の機能―外国地名・人名の表記を中心として―』早稲田大学博士学位論文、2022年）に詳しいので参照されたい。

[17] なお、本章では日本語表記のみ対象として論じているが、センターでは外国語で書かれた資料も公開しているため、「Владивосток」や「Vladivostok」といった外国語表記の存在も確認している。これら外国語表記についても、辞書データに同義語として登録している。

[18] データ検証委員会で検証された英訳を公開するWEBコンテンツとして、2021年６月に「歴史用語に関する日英対訳集」https://www.jacar.go.jp/Japanese-English_dictionary/index.html（最終アクセス日：2023年７月１日）を公開した。あわせて参照されたい。

[19] センターでは、「アジ歴地名・人名・出来事事典」で登録された辞書データを、前掲注８のサイト上でCSVファイル形式で配布する試みをすでに行っている。

松浦晶子（アジア歴史資料センター研究員）
中野　良（アジア歴史資料センター研究員）