多言語の中の日本語―Aleph を日本語環境で使う―

佐藤（さとう）　康之（やすゆき）
（メディアセンター本部課長）

1　はじめに

Aleph は多言語に対応した図書館システムである。前システムの KOSMOS II においても主にラテン語系言語（英独仏語など）で使用される音標符号付き文字を入力して Web OPAC で表示できたが，Aleph は Unicode[注1]が採用する言語であれば非常に多くの文字を入力，検索，表示できる。特に KOSMOS II で実現できなかった中国語の簡体字，繁体字，韓国語のハングルが扱えることは，大きな進歩と言える。Aleph で日本語を扱うことは，日本語の慣習を意識して Aleph の多言語機能に「日本語を追加する」ことと位置づけて Aleph の開発元である Ex Libris 社（以下，E 社）と共同で取り組んだ。

2　Aleph の多言語機能

Aleph ではデータの表現形式として Unicode の符号化方式の一つである UTF-8[注2]を使用する。UTF-8 では CJK（中国語，日本語，韓国語）文字は 3 バイトで扱われ，データ項目の入力可能文字数に制限がある場合，CJK 文字を多用する日本では不利となり，システムが表示する日本語メッセージを短い表現にするなどの工夫が必要となる。

Aleph では一般的なパッケージシステムと同様，各種のシステムテーブルファイルによって様々な機能の動作を設定する。多くのシステムテーブルファイルは使用する言語によって個別に設定でき，各業務クライアントソフトウェア[注3]や Web OPAC に表示する各種メッセージも同様に設定できる。Web OPAC の利用者向けメッセージは，言語毎に設定されたシステム内のディレクトリ[注4]に置かれる HTML テンプレートファイル[注5]により，利用者が選択した言語で表示される。導入にあたっては，これらの日本語用ファイルを作成することからスタートした。

Aleph の多言語検索機能は，ラテン語系言語用をベースに中国語，韓国語用に拡張されている。一般的な検索システムにおいては，利用者の入力する検索語とインデックスを形成する索引語を，各言語の文章記述の中からどのように抽出するかが重要となる。ラテン語系言語のように単語間の区切りが空白などによって明確な場合は容易だが，区切りが明確でない CJK 言語は特殊な手法が必要となる。Aleph は中国語，韓国語を検索する際の検索語，索引語抽出にバイグラム[注6]方式を採用し，一部中国語のために辞書方式（予め用意された辞書に合致した語を単語として抽出）も併用している。この多言語検索機能に日本語を追加することが最大の課題となる。

3　日本語環境で使う

Aleph を日本語環境で使うために考慮した主なポイントを紹介する。

図書館システムにおいて，目録をどのような書誌レコード形式で記述するかは最も重要な要件の一つである。Aleph は，MARC21 のほか UNIMARC など世界標準の書誌レコード形式を扱うことができるが，日本語の場合は基本となる漢字かな混じりの記述（以下，漢字記述）とカタカナ記述，ローマ字記述の少なくとも 3 種類の記述が必要であり，日本語対応を進めるには書誌レコード形式を最初に決める必要がある。KOSMOS II では漢字，カタカナ，ローマ字の記述のために MARC21 のタグを拡張していたが（例えばタイトルのタグ 245 のカタカナはタグ K245 とした），Aleph ではこれができないため，タグ中のサブフィールド$9 にカタカナは K を（$9K のように），ローマ字は R を指示することで対応することとした。また，各タグの繰り返し（例えば一人目の著者の漢字とカタカナ，二人目の著者の漢字とカタカナなど）は，サブフィールド$6 に繰り返し番号（$601，$602 のように）を付加することとした。

索引語抽出のためのバイグラム方式は，日本語の検索において無駄な検索結果（以下，ノイズ）を多く含むとの一般的な評価がある。この影響を最小限にする手法が必須と判断し，後述する形態素[注7]方式を追加することとした。

書誌レコード以外に利用者や取引業者も日本語で検索できるようにするために，同様にバイグラム方

式による検索機能の追加開発を行った。

国際的な図書館システムで日本語対応を検討する際には，日付と通貨の表現形式の検討も必要となる。日付はAlephが対応しているISO8601[注8]形式（YYYYMMDD）を採用した。通貨は，米ドルなどで使用される小数点以下の表示（例えばUS$10.50）を抑制できるようE社へ要請した。

４　検索機能の追加開発

バイグラム方式による検索のノイズとしては，検索語に「京都」を指定した場合に「東京都」が検索結果に含まれるような事例がある。バイグラム方式は，検索漏れが少ない反面，カタカナによる検索時などはノイズが多すぎて問題となる検索事例が多い。そのため，よりノイズの少ない形態素方式を追加するようにE社へ要請した。形態素方式は書誌レコードに漢字，カタカナ，ローマ字記述のほかに分かち書き記述を追加し，分かち書きにしたがって索引語の単語抽出を行うとともに，検索語を入力した際の単語分解を不要とするため，分かち書きされた単語を記述の先頭から順次連結して索引語とする方式である。また書誌レコードのタグの定義にあたっても分かち書きとしてサブフィールド$9Wを含むタグを用いることとした。E社にとって新しい索引語抽出の開発となるが，Alephの索引語ファイルの構造を変更することはできないため，E社との細かな仕様の調整が必要となった。

なお書誌レコードに対する分かち書き記述の追加を容易にするため，別途Alephの目録クライアントソフトウェアと日本語形態素解析ソフトウェアを連携させる開発もE社に要請した。この仕組みでは前述した書誌レコードの各タグのサブフィールド$6および$9も自動制御するようになっている。

Alephの検索機能には索引語を一覧しながら検索を進めるブラウズ検索もある。日本語の環境ではカタカナヨミ順に索引語を配列することが利用者にとって望ましいが，Alephでは実現が困難なため，UTF-8のコード順に配列することとした。

５　文字正規化テーブルの準備

Alephは利用者が入力する検索語文字の「ゆれ」を吸収するために，各種の文字正規化テーブルを持っている。日本語を追加するにあたって，これらのテーブルの全面的な見直しを行い，全角記号の「取る・詰め」やひらがなのカタカナへの変換，繰り返し記号（ゝやヽなど）の正字への置換に対応するため，文字正規化テーブルを拡張した。独仏語などの音標符号付文字変換は既に装備されているテーブルをそのまま利用した。

さらに，CJK漢字を統合して検索するためにCJK漢字異体字変換テーブルを準備する必要があった。Alephは中国語の繁体字を簡体字に置換するテーブルを持っており，日本語で使用されている文字が簡体字に置換されてしまう問題がある。また，日本語では新旧字体の統合も必要なため，新たに異体字変換テーブルを準備することとした。これにはKOSMOS IIで使用していた約590組の異体字変換テーブルをベースに，京都大学人文科学研究所東アジア人文情報学研究センターが公開している「Unicode異体字統合テーブル」，国立情報学研究所の「統合漢字インデックステーブル」，さらにUnicode Consortium[注9]が提供する「Unihanデータベース」をプログラムで統合，再編集して約4,700組の異体字変換テーブルを作成した。但し人の目によるチェックを経ていないため，一部の文字変換で齟齬が生じており，順次テーブルの修正を行っている。

なお，ブラウズ検索の一覧表示においても記号などによる配列の「ゆれ」を防止するため，専用の文字正規化テーブルを準備した。また，AlephではMARC21のNonfiling characters Indicator[注10]を利用することができるため，ラテン語系言語の書誌レコードでは冠詞（TheやAnなど）を除いた配列が実現できている。

６　開発協議の過程

日本語対応のための機能追加は，システムテーブルの調整によって実現できたものもあるが，多くは新たな追加開発をE社へ要請する必要があった。E社としては，これらの開発を慶應向けということではなく，Alephの標準的な日本語対応機能の開発という位置付けで作業が進められた。但しE社の開発担当者は日本語を理解しないために，要望を的確に伝達することには多くの手間と時間がかかった。開発のための協議は2008年12月の導入プロジェクトキックオフ会議から始まり，最終的な機能仕様書が固まったのは2009年6月であった。この間，日本に

在勤するE社担当者と協議して要求仕様書を作成し，これを何度かレビュー，引き続き開発担当者が機能仕様書を作成し，やはり何度かレビューを実施，その後に実際の開発が始まり，完成した機能をE社内の開発用サーバで検証するといった一連の手続きを行った。英文による文書の交換は20数回，emailについては数え切れない。さらに，数回の会議での検討があり，2009年9月にはイスラエルにあるE社を訪問して開発した機能の検証を集中的に実施した。

7　プロジェクトを終えて

追加開発した日本語対応機能については今後も不具合対応の必要な部分が残っているが，概ね予定した機能は実現できた。一部の機能については今後のService Pack（改善プログラムの提供）で実現される予定である。Alephは索引語生成規則の設定など検索の根幹に関わる部分でもシステムテーブルによる調整ができるなどの柔軟さを持っているが，反面，設定のノウハウを要求される難しさもある。E社の日本語対応機能の開発に対する姿勢は，単に文字として日本語を扱うということではなく，日本語の慣習に配慮されたものだったと評価している。様々な言語が使われるヨーロッパを中心にユーザを持っていることが背景にあるのかもしれない。開発協議における英語でのコミュニケーションは想像以上に骨が折れたが，プロジェクトメンバーが，それぞれの専門性を発揮して要求を整理して伝えることによって乗り切ることができたと思っている。曖昧な要求は，言葉を尽くしても通じないことを痛感した。要求仕様をまとめるために，十分な業務経験に裏付けられた専門性が今後もこの種のプロジェクトで重要なことに変わりはないだろう。

最後に，漢字異体字テーブルを準備する過程で感じたCJK漢字の日本語で使われる漢字の存在感に対する懸念を付記しておく。圧倒的に中国語で使われる漢字が多い中で，日本語の記述で使用する字体の持つ意味が，中国語の漢字に埋もれることのないようにしなければならない。このような事例は既にOCLCなどの書誌レコードにも散見されるが，日本語の文化を守るためにもUnicodeのCJK漢字の字体統合については漢字文化圏全体の連携による整理が早急に必要な段階にきている。

注

1）コンピュータ上で文字を取り扱うためのコード体系のひとつ．世界中の文字を単一のコード体系で表現することを目標に開発されている．
2）Unicodeの文字を実際にコンピュータ上で扱うためにデジタルデータへ変換する方式のひとつ．
3）Alephの閲覧，目録，受入/雑誌業務のプログラムは，サーバ上で稼動するサーバソフトウェアとPC上で稼動するクライアントソフトウェアで構成されている．
4）コンピュータ上のファイルの格納場所．PC上のフォルダと同義．
5）IEなどのブラウザが表示するHTMLファイルのテンプレート．Alephの Web OPACはテンプレートファイルに必要な情報を動的に付加してブラウザへ送信し，PC上でページが表示される．
6）単語の意味と無関係に2文字単位に分割する方式で，検索対象とする文章を文字単位に分解するN-Gram（N文字インデックス法）のひとつ．
7）言語において意味をなす最小の言葉の単位．
8）日付と時刻に関する国際規格．
9）Unicodeを開発した米国の非営利団体．
http://www.unicode.org/,（参照 2010-08-18）
10）MARC21書誌レコード形式のインディケータで，処理対象から除外する文字数を表す指標．
http://www.loc.gov/marc/bibliographic/concise/bd245.html,（参照 2010-08-18）