Alephで日本語を扱う―日本語分かち書き機能：導入について―

田中　真紀（たなか　まさき）
（メディアセンター本部）

1　はじめに

Alephはイスラエルの Ex Libris 社（以下 Exl 社）が開発した統合図書館システムである。Alephは Unicode に対応しているため，日本語を含め多言語を取り扱うことができる。日本語データの登録・表示から，スタッフが利用する各種画面の細部まで，図書館側で自由に設定が可能である。しかしながら，日本語特有の処理に関する機能が標準搭載されているわけではない。この為，Exl 社と慶應で日本語特有の処理について，一つ一つ打合せ，仕様を決め，開発，調整を行う必要があった。本稿では，その中でも最も苦労した，「分かち書き」機能の導入について記述する。

2　導入にあたって

慶應では目録データ上で分かち書きを作成するにあたって，形態素解析ソフトウェア Happiness[注1] を利用していた。Happiness に日本語テキストを入力すると，分かち（以下，W），カナ（以下，K），ローマ字（以下，R）が自動生成される。このデータを元に目録の WKR 部分を作成する。このソフトウェアを Aleph でも継続して利用するためには，連携部分の開発が必要となった。

(1) 連携について

当初 Exl 社からは，Aleph と Happiness を直接通信させる仕様を提案された。この方法は，処理速度が速く，至極まっとうな方法ではある。しかし，直接連携した場合には，以下のような問題が発生することが予想された。

・Aleph や Happiness のバージョンを変更することが困難となる（連携に障害が発生する可能性がある）

・分かち書きの調整が全て Aleph に依存してしまう（Exl 社への追加機能要求が増える）

・将来，分かち書きのソフトウェアを変更することが難しい（選択の自由が無くなる）

上記のような事を考慮し，より自由度の高い方法で連携することを選択した。

(2) 仕様について

「分かち」機能を恒常的に確保するため，新規アプリケーション Happiness ゲートウェイ（以下，G/W）を独自に開発し，Aleph と Happiness の間を取り持つ事にした。仕様を決めるにあたっては，メンテナンスの容易さを考慮してスタンダードな技術を採用した。通信は Web で使われている HTTP プロトコルを利用し，入出力フォーマットは MarcXML[注2] 形式とした。これにより，Exl 社は Aleph に MarcXML の入出力部分を開発するだけで，日本語を一切意識せずに分かち書き機能をサポートすることが可能となった。分かち書きの主たる機能を担う G/W の仕様については慶應側で全て決定した。

(3) 開発について

仕様の取り決め後，すぐに Exl 社から連携テストの要求があったため，慶應側でアプリケーションを早急に開発しなければならなかった。突貫工事で G/W を作成するしかなく，Happiness を利用しない暫定的な分かち書き環境を作成し連携テストに対応した。並行して，本格的な G/W のコア部分の開発を平和情報センター[注3]（以下，FIPS）に依頼した。慶應と FIPS 間で仕様変更・機能追加を重ねながら開発が完了した。

最初に連携部分の仕様を決めていたため，慶應・Exl 社・FIPS がそれぞれ別々に並行開発を行うこと事ができた。出来上がった G/W（図 1）は単純に Happiness を呼び出すだけではなく，分かち書きのタグ指定や参照する辞書の種類の指定，その他詳細な設定が可能となっている。

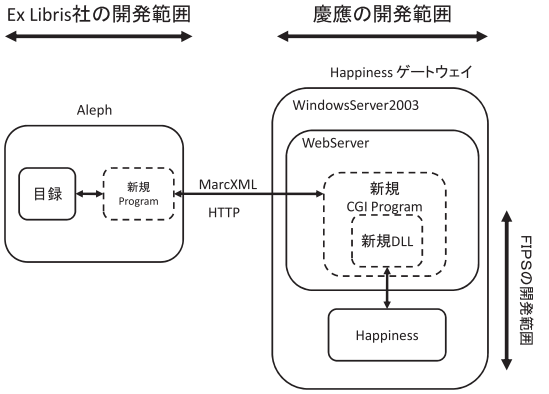

図 1.　Aleph-Happiness ゲートウェイの連携図

3　導入後

新しい分かち書き機能の導入は比較的スムーズに完了したが，前システムでの分かち書きとは生成条件や手順が大きく変わった。中には Aleph 側での制限により変更となった部分もある。以下に今回の改善点，および以前との手順の違いを上げる。

(1) 改善点

・対象タグの一括自動生成
・カナ生成の改善
・ローマ字生成に機能を追加
・辞書設定の変更（タグ・サブフィールド別）

(2) 分かち書き生成の手順

・旧手順
1. WKR を生成するタグを選択
2. レギュラータグ（原文）から W を自動生成
3. 確定した W から K を自動生成
4. 確定した K から R を自動生成
5. 対象タグ毎に上記手順を繰り返す

・新手順
1. 分かち書き対象の全てのタグを一括 WKR 生成

このように新しい手順では，一回の処理で自動的に WKR 生成されるため，手数がかなり減らすことが可能となった。

4　まとめ

新しいシステムを導入する際には，今まで当たり前のように使われてきた手順や機能は，必ずしも維持・搭載されるわけではない。しかし，分かち書きの導入は慶應・FIPS 側で開発したからこそ，仕様変更も無難にこなしリリースすることができた。もし Exl 社当初の提案を受けざるを得なかったとしたらここまでスムーズな導入は出来なかっただろうし，改良も容易ではなかっただろう。日本語処理に特化した重要な部分は，日本語を扱う難しさを分かっている慶應側がどのように維持・進化させていくかを考えていく事が重要である。

（度重なる仕様変更にも柔軟に対応していただいた FIPS さんには感謝したい。）

5　今後の展望

まだ新しい分かち書き機能はリリースしたばかりで，まだ改良の余地がある。今後は効率よく分かち書き生成ができるよう，辞書のメンテナンスはもちろんのこと，G/W の仕組みも調整しながら Aleph との共存を図りたい。

参考文献

1) 沓沢尚明，飯田一幸．ソフトハウスが始めた手作りデータベース―HiNET & HAPPINESS 開発秘話．情報の科学と技術．vol. 58, no. 9, 2000, p. 460-464.
2) 入江伸．ライブラリーシステム研究会の経過とシステムの課題―図書館システムの標準化に向けて．MediaNet. vol. 9, 2000, p. 8-11.
3) 保田明夫.“形態素解析と分かち書き処理”．http://wordminer.comquest.co.jp/wmtips/pdf/H15_01-4.pdf,（参照 2010-08-18）.

注

1) Happiness
株式会社平和情報センターが開発した形態素解析ソフトウェア
2) MarcXML
http://www.loc.gov/standards/marcxml/,（参　照 2010-10-26）.
3) FIPS
開発を依頼した時点には,「株式会社平和情報センター」であったが，最終納品時点では「富士通エフ・アイピー・システムズ株式会社」（省略名 FIPS）に社名が変更された.