メタデータブラウザの多言語対応に向けた課題への取り組み

○阪口 哲男, 樋爪 育恵, 加藤 大博

Approach to Problems of Developing a Browsing System of Multilingual Metadata

○ Tetsuo Sakaguchi, Ikue Hizume, Hiromichi Kato

Abstract

The Research Center for Knowledge Communities at University of Tsukuba introduced the Knowledge Community Information System (KCIS) in February 2003. KCIS inherited the metadata of the digital library system of the University of Library and Information Science (ULIS-DL). In KCIS, metadata are able to be written in several languages because they are expressed in XML and Unicode. The browsing system of metadata of ULIS-DL were designed to handle metadata written in Japanese. This paper discusses problems of applying the browsing system to multilingual metadata. One of the problems is that the system extracts words from metadata with Japanese morphological analyzer. Another problem is synonyms between different languages. This paper describes a prototype system which are developed to make the issues clear. It also describes development of a experimental system to display multilingual XML documents. It is needed for using the browsing system of metadata with lightweight terminals.

1 はじめに

図書館情報大学ディジタル図書館 (以下 ULIS-DL)[1] では図書館情報学及びその関連分野の情報資源のメタデータを作成・蓄積してきた。このメタデータは、図書館情報大学と筑波大学の統合に伴い、2002 年 10 月に開設された筑波大学知的コミュニティ基盤研究センター[1]が運用する知的コミュニティ情報システム (以下 KCIS, 2003 年 2 月稼働開始) に引き継がれた。

KCIS ではメタデータを XML 形式で記述し、文字コードとして UTF-8 (Unicode) を使用している。そのため、これまで蓄積してきたメタデータの記述言語は主として日本語と英語であったが、今後はより多様な言語による記述を取り扱うことが可能となる。

これまで ULIS-DL の一部として運用してきたメタデータブラウザ [2] は日本語と英語で記述されたメタデータを対象としている。今後 KCIS で作成・蓄積されるメタデータを対象とするにはメタデータブラウザを多言語対応したものが必要となる。そこで、本稿ではこれまでのメタデータブラウザを多言語対応する際の問題について検討し、現在までに開発を進めてきた試作システムについて述べる。

2 多言語対応に関する問題

メタデータブラウザはメタデータ中に含まれる文や句から単語を抽出し、その単語を一覧形式で利用者に提示する。利用者はその中から自分の要求に最も適切と思われる語を選択し、メタデータを検索する。ここで多言語化の際に問題になるのは主に以下の 2 点と考えられる。

- 単語の抽出
- 同義で表記が異なる語

[1] http://www.kc.tsukuba.ac.jp/

単語の抽出過程は、言語毎に記述方法が異なるために一律の方式を用いるのは困難である。特に日本語のような分かち書きをしない言語の場合にはそれぞれの言語における語の変化形や連接規則、文法などの処理も単語の境界を定めるために必要となる。従来のメタデータブラウザでは単語抽出に日本語形態素解析器の茶筌[2]を用いている。茶筌では特に英文の処理は行わないが、英文を入力した場合に空白を語の区切りとみなして分割したものを未知語として出力する。そこで、これまでのブラウザではこの出力に若干の記号の処理を加えたものを単語として取り扱っている。そのため、英単語については変化形も別の語として扱ってきた。

以上のようにこれまでは日本語対応を基本としてきたので、英語についても十分な対応はできていない。また、茶筌は辞書に基づいて形態素解析を行うため、専門用語など辞書にない語が多用されている場合においては、意味をなさない単位に分割されることも生じている。

同義で表記が異なる語に関しては日本語の単語同士でも生じるが、多言語の場合にはより顕著になる。この問題については従来より cross-lingual information retrieval として数多くの研究がなされている。メタデータブラウザでは、語を利用者が投入するのではなく一覧から選択するため、抽出した語と同義の語の両者をどのように提示するかが問題となる。

単語抽出や同義語の問題の他に、多言語対応に関しては語あるいは文字を正しく表示することが大前提となる。現在では例えばマイクロソフト社の Windows 2000 や XP のように Unicode 標準に基づいたある程度の多言語対応機能を備えている OS の普及が進んでおり、この問題の比重は少なくなってきていると考えられる。その一方で処理性能や記憶容量が乏しい携帯型端末も普及しており、誰もが気軽に多様な言語で記述された情報を文字化けのような誤りがない状態で見ることができるまでにはまだ時間がかかるものと思われる。

3 問題解決への取り組みと試作システム

前節で述べた問題を解決する取り組みの手始めとして、2 種類のシステムの試作を行った。一つは日本語のみを対象としているが、前述の単語抽出と同義語の問題の低減あるいは解決のための機能をメタデータブラウザに備えさせたものである (以下、改良版メタデータブラウザ)。もう一つは、様々な環境における多言語 XML 文書を表示可能とするシステムである (以下、多言語 XML ブラウザ)。

3.1 改良版メタデータブラウザ

改良版メタデータブラウザではまず抽出する単語の精度を上げるために、形態素解析用辞書の語彙を増やすことの有効性の確認を行った。実際には、茶筌で用いている IPA 品詞体系日本語辞書 (IPADIC) に対して、EDR 電子化辞書[3]に含まれる語の追加を行った。メタデータブラウザでは名詞 (未知語を含む) を用いているので、茶筌の辞書に既に含まれている名詞約 21 万語に対して約 8 万語を追加した。

2003 年 2 月時点のメタデータ約 3 万 7 千件から抽出された利用者に提示される語は約 3 万 2 千語から約 4 万語に増加した。「トーキング」という語を例にとると、以前は一語としては抽出されず、「トー」(25 件) と「キング」(23 件) のみが抽出されていたが、「トーキング」(16 件)、「トー」(2 件)、「キング」(5 件) となり、改善されていることがわかる。しかしながら、まだ「トー」のように意味をなさないものも抽出されていることから、この手法では単語の抽出エラーを完全になくすことが難しいことがわかる。

同義語に関しては、抽出した語の一覧の際に同義語が存在する語にマークをつけ、そのマーク

[2] http://chasen.aist-nara.ac.jp/
[3] http://www2.crl.go.jp/kk/e416/EDR/

を選ぶことにより同義語一覧を表示し、その中から語を複数選択可能とした (図 1)。図では、最初の一覧から「母親」を選び、同義語を表示した上で「マザー」も選んでいる。

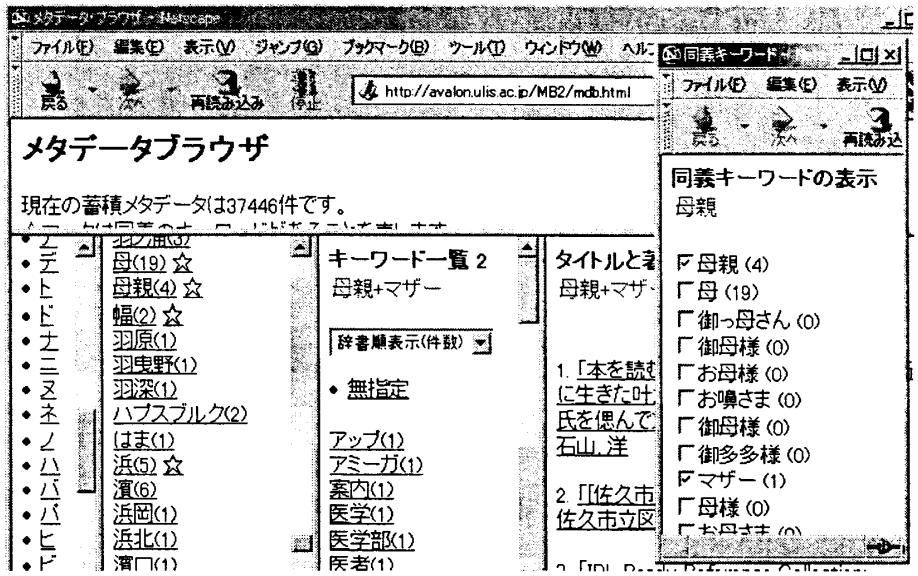

図 1: 改良版メタデータブラウザの画面

本システムでは EDR 電子化辞書の日本語単語辞書に付与されている概念識別子が同一の単語を同義語としている。単語には多義性があるので、利用者が同義語の中で適切なものを選択できるようにしている。本システムは Java Servlet として構築し、利用者インタフェースには WWW ブラウザを用いている。

3.2 多言語 XML ブラウザ

多様な利用者環境において様々な言語の HTML 文書の表示を可能としたものが MHTML ブラウザである [3]。MHTML ブラウザでは Java Applet を実行可能であればあらかじめ文字フォントを備えていない環境でも様々な言語の Web ページを見ることができる。この方式に基づいて多言語 XML 文書を対象としたブラウザ、多言語 XML ブラウザを試作した。MHTML ブラウザでは書字方向や禁則処理、ワードラップ処理が省かれていたが、それらへの対応機能も含めた。

多言語 XML ブラウザはゲートウェイとビューワから構成され、ゲートウェイでは XSLT スタイルシートによって XML から XHTML へ変換し、書字方向や禁則処理のための指示用のタグを付与する。そして利用者端末上で実行されるビューワに変換後のデータとその文書内に含まれる文字フォントを送る。ビューワは Java Applet であり、ゲートウェイから送られたフォントを用いて表示を行う。言語毎の表示上の規則などはゲートウェイで解釈し、ビューワはどの言語かに関わらずゲートウェイが付与したタグの指示によって表示を行う。ゲートウェイとビューワの通信プロトコルには HTTP を用いる。

図 2は試作システムの画面例である。現時点ではプレーンテキスト相当の表示機能のみであるが、右から左へ記述するアラビア語やヘブライ語に対応し、日本語の禁則処理、欧文のワードラップにも対応している。禁則処理やワードラップに関しては改行の許可・禁止を示すタグを、書字方向についても方向を表すタグを用いることで、新たな言語に関してはゲートウェイの修正のみで対応可能となっている。試作システムではまだ縦書き表示やハイパーリンクなどが未実装であ

るが、ビューワのプログラム自体は約40KB程度であり、現在の携帯電話ではまだ難しいがPDA程度の携帯型端末での利用には問題が少ないと思われる。

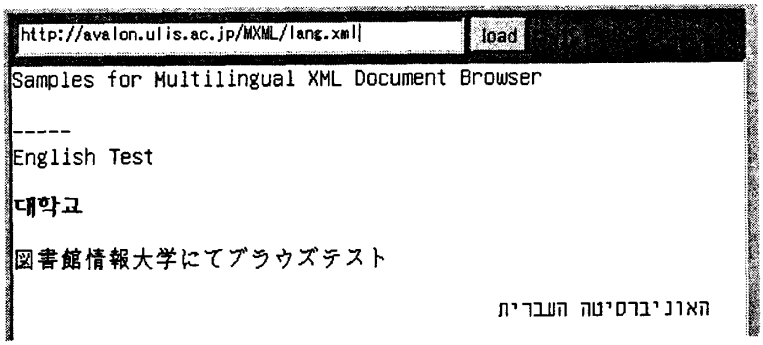

図 2: 多言語 XML ブラウザ画面

4 おわりに

これまで日本語で記述されたメタデータを対象としていたメタデータブラウザについて、多言語対応について考えられる問題点とその解決方法について検討し、システムの試作を行った。中でも単語の抽出に関しては、現在の方式では各言語毎の形態素解析器を必要とすること、そしてその辞書の整備が課題であることがわかった。

同義語に関しては試作システムではメタデータ中に含まれる語から同義語を提示する機能を準備した。しかしながら、利用者がメタデータに含まれていない語を最初に思い浮かべることも考えられるため、より多様な方法で同義語をアクセス可能にすることが必要と思われる。

参考文献

[1] 平岡博, 真中孝行, 横山敏秋, 阪口哲男, 杉本重雄, 田畑孝一. 図書館情報大学ディジタル図書館システム. 情報管理, Vol.42, No.6. p.471–479. 1999.

[2] 阪口哲男, 嶋田恭子, 沼尻花名, 田畑孝一. メタデータのブラウジングシステムの構築. 情報知識学会第 8 回 (2000 年度) 研究報告会講演論文集. p.65–68. 2000.

[3] 前田亮, Myriam Dartois, 太田純, 藤田岳久, 阪口哲男, 杉本重雄, 田畑孝一. クライアントにフォントを必要としない多言語 HTML 文書ブラウジングシステム. 情報処理学会論文誌. Vol.39, No.3. p.802–809. 1998.

阪口 哲男. 筑波大学図書館情報学系 (〒 305-8550 茨城県つくば市春日 1-2).
樋爪 育恵. 図書館情報大学 (2003 年 3 月卒業).
加藤 大博. 同上.
Tetsuo Sakaguchi. Institute of Library and Information Science, University of Tsukuba.
(saka@ulis.ac.jp / saka@slis.tsukuba.ac.jp)
Ikue Hizume. University of Library and Information Science.
Hiromichi Kato. University of Library and Information Science.