情 報 学 基 礎 44－1
（1996. 11. 21）

WWW文書のための多言語ブラウザとそのゲートウェイサービス

前田亮†，藤田岳久‡，阪口哲男†，杉本重雄†，田畑孝一†
{maeda, take, saka, sugimoto, tabata}@ulis.ac.jp

†図書館情報大学
〒305 茨城県つくば市春日1-2

‡共立女子大学
〒101 東京都千代田区神田神保町3-27

ディジタル図書館にとって多言語文書環境の実現は必須の課題であるが，現実には十分な環境が整備されていない．我々はフォントをインストールすることなしに多言語の表示を可能とするWWW文書ブラウザを開発した．この多言語ブラウザは，WWWサーバとユーザ側のWWWブラウザの中間に位置しHTML文書に文字フォントを付加するゲートウェイ，およびWWWブラウザ上で動作する専用のビューアから構成される．また，1996年8月からこの多言語ブラウザをインターネット上で公開し，試験的にサービスを行っている．本稿ではこの多言語ブラウザの実現方式とゲートウェイサービスの概要について述べる．

Multilingual Browser for WWW Documents and its Gateway Service

Akira Maeda†, Takehisa Fujita‡, Tetsuo Sakaguchi†, Shigeo Sugimoto†, Koichi Tabata†

† University of Library and Information Science
1-2, Kasuga, Tsukuba, Ibaraki, 305, Japan

‡ Kyoritsu Women's University
3-27 Kanda-Jimbocho, Chiyoda-ku, Tokyo, 101, Japan

A multilingual document environment is a crucial aspect for the digital library. However, conventional computers and networks do not provide us with a sufficient multilingual document environment. The authors have developed a multilingual browser for WWW documents for users who have no multilingual fonts on their terminals. The browser has two major components, a gateway and a document viewer. The gateway located between a WWW client and a WWW document server transforms a source document into a packaged form which includes the source text and a minimum set of font glyphs for the document. The viewer realised as a set of applets receives the form and displays the source text. This paper presents the technological aspects of the browser and the experimental gateway service since August 1996.

1 はじめに

図書館は従来から多言語の資料を扱ってきており，従ってディジタル図書館においても多言語の文書を扱える環境は必須である．しかしながら現在のコンピュータ上の多言語文書環境は貧弱であると言わざるを得ない．現状では多言語を扱える文字コード系が普及しておらず，また通常コンピュータ上には自国語の文字フォントしか用意されていない．将来多言語を扱える文字コード系が普及したとしても，すべてのユーザの端末に全世界の言語の文字フォントを用意するのは非現実的であると思われる．また，図書館としては異体字や古典文字など既存の文字セットに含まれない文字も扱えなければならない．

これらの問題に対処するための一つの方法として，我々はユーザの端末側にフォントのインストールの必要がない WWW 文書のための多言語ブラウザの開発を行っている [1][2][3].

本稿ではこのブラウザの実現方式と試験的に行っているゲートウェイサービスの概要について述べる．

2 WWW 上での多言語文書環境の現状と問題点

現在の HTML の仕様である HTML 2.0[4] では，文書中に使用できる文字セットを ISO-10646 に含まれる文字としているが，その符号化の方法については定めていない．実際には，例えば日本語の場合，現在のところ ISO-2022-JP（いわゆる JIS コード），Shift-JIS，EUC の 3 つの符号化方法が用いられている[1].

ただしこれらの符号化方法では，複数の言語を一文書中に混在させることはできない．これを可能にする符号化方法としては，ISO-2022 に基づいて複数の文字セットをエスケープシーケンスによって切り替える ISO-2022-JP-2[5], ISO-10646 の文字セットの一部を 1 文字 16 ビットの固定長で符号化する Unicode（ISO-10646-1）などがある．しかし，ISO-2022-JP-2 で符号化できるすべての文字セットのフォントを用意することは困難であるし，Unicode についても，多くの問題点が指摘されており，広く使われるまでには至っていない．

[1] これらの符号化方法で使用される文字はすべて ISO-10646 の文字セットに含まれている

3 多言語 WWW 文書ブラウザ

3.1 多言語 WWW 文書ブラウザの構成

我々が開発している多言語 WWW 文書ブラウザ（MHTML：Multilingual-HTML ブラウザ）は，ユーザの端末側で動作する MHTML ビューアと，WWW サーバと MHTML ビューアの中間に位置する MHTML ゲートウェイから構成される．MHTML ブラウザの構成を図 1に示す．

MHTML ゲートウェイはユーザから要求された HTML テキストをサーバから取得し，これに必要最小限のフォントを付加した MHTML テキストに変換し，ユーザ側に返す．現在の MHTML ゲートウェイは C 言語を用いて開発しており，UNIX ワークステーション上で動作する．

MHTML ビューアは WWW ブラウザ上で動作し，ユーザからの要求を MHTML ゲートウェイに送り，ゲートウェイから返ってきた MHTML テキストを表示する．MHTML ビューアには Java アプレットを用いており，Java に対応した WWW ブラウザ上で動作する．Java を用いることで，ビューアは最初にゲートウェイにアクセスした時に自動的にダウンロードされるため，ユーザ自身でインストールする必要はない．

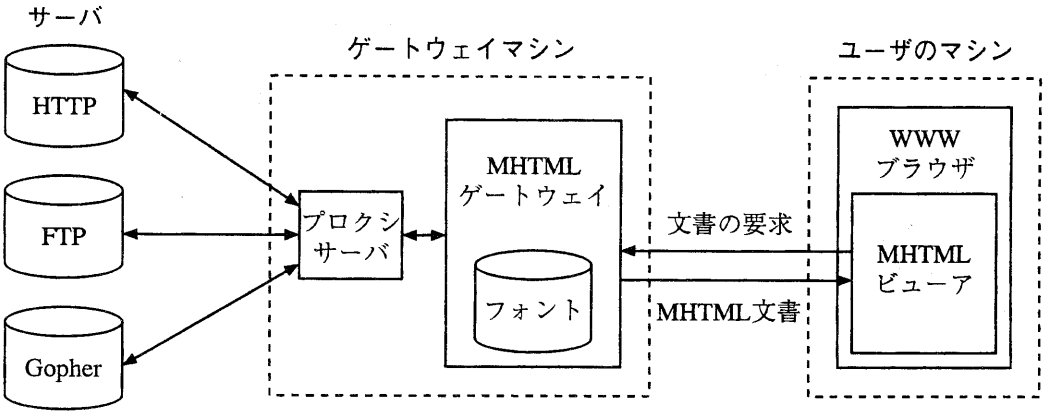

図 1: MHTML ブラウザの構成

3.2 他の方法との比較

フォントをインストールせずに多言語の表示を実現する他の方法として，文書の文字部分を一文字毎あるいは数文字分まとめてインラインイメージに変換する方法，ページ全体を一枚のインラインイメージに変換し，リンクをクリッカブルマップで実現する方法などが考えられる．前者の方法は実際に DeleGate [2] に付属の CII (Character by Inline Image) ライブラリ，および Shodouka [3] で実現されている．

これらの方法と我々が開発した MHTML ブラウザの方法について，同じ文書の転送に要するバイト数の比較を行った．比較は日本語 (ISO-2022-JP) で書かれた 6 つの論文について行った．CII (一時文字毎，数文字単位)，ページ全体の GIF イメージ，MHTML のそれぞれの方式について，元となる HTML 文書に対する転送バイト数の比率を比較した．CII の場合は，DeleGate で変換された各インラインイメージの大きさの合計とこれらを表示するためのHTML テキストの大きさを足したものを用い，ページ全体のイメージの場合はページを WWW ブラウザで表示させたもののスナップショットの GIF イメージの大きさを用いた．結果を図 2に示す．この図を見ると，DeleGate の CII ライブラリを用いた場合やページ全体の GIF イメージを用いた場合に比べ，MHTML による方法はより少ないバイト数で済んでいることがわかる．

また，一文字あるいは数文字単位でインラインイメージに変換する方法では，一文字あるいは数文字単位でネットワークのコネクションを張る必要があるため，一文書に必要なコネクション回数が多数となり転送に非常に時間がかかる．一方 MHTML の方法では一回のコネクションで済む．

4 MHTML ブラウザの実現方式

4.1 MHTML テキスト

MHTML テキストは，通常の HTML テキストにそのテキストで使用されている文字のみのフォントのビットマップ情報を付加したもの

[2] URL: http://www.etl.go.jp/etl/People/ysato@etl.go.jp/DeleGate/

[3] URL: http://www.lfw.org/shodouka/

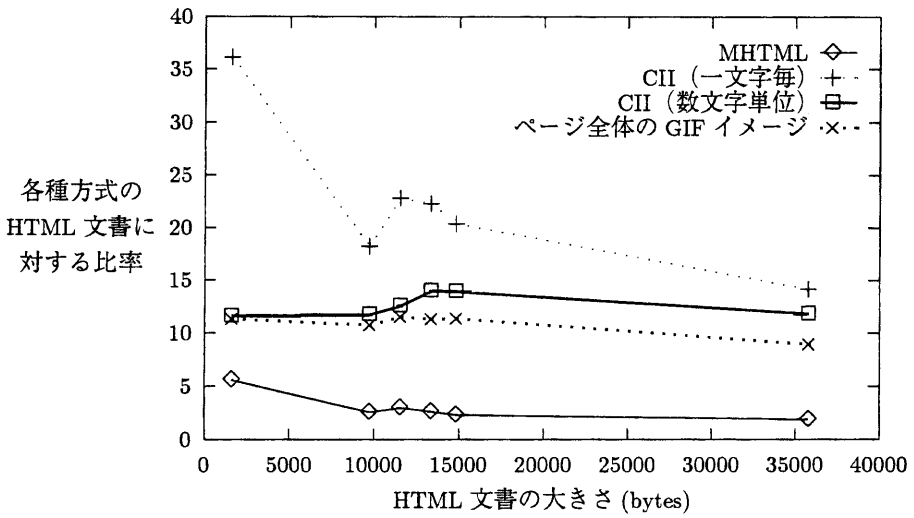

図 2: 転送に要するバイト数の他の方法との比較

である．これは MHTML ゲートウェイによって HTML テキストから生成され，MHTML ビューアに送られる．MHTML テキストの構造を図 3に示す．

MHTML テキストはヘッダ部，フォント部，テキスト部の 3 つの部分から構成される．ヘッダ部には，MHTML のバージョン番号，大きさ別のフォントの数，テキスト部へのバイトオフセット，各フォント情報が格納される．各フォント情報の部分には，フォントグリフの幅および高さとその大きさのフォントグリフの文字コードの数が格納される．フォント部には，文書を表示するための必要最小限のフォントグリフが格納される．テキスト部には，各テキスト毎に符号化される内部コード列が格納される．

テキスト部に格納されるコードは制御文字，ASCII コード，内部コードのいずれかである．HTML のタグの部分にのみ ASCII コードが用いられ，その他の文字の部分はすべて内部コードが用いられる．内部コードはフォント部に格納された順に一文字 16 ビットで割り当てられ

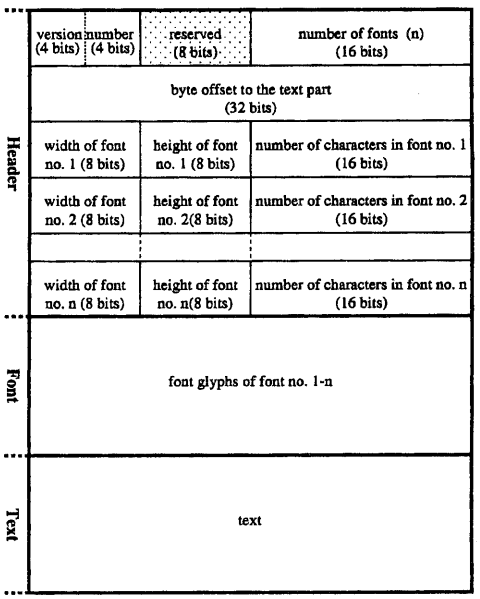

図 3: MHTML テキストの構造

る．この符号化はMHTMLゲートウェイによって各テキスト毎に行われるため，内部コードはそのテキスト内でのみ有効なものとなる．

テキスト部に含まれるすべての文字にフォントが付加されているため，MHTMLビューアでの表示にはユーザ側にローカルなフォントは全く必要としない．

4.2 MHTMLゲートウェイ

MHTMLゲートウェイはWWW，Gopher，FTPなどのサーバとMHTMLビューアの中間に位置する．ゲートウェイはユーザから要求されたURLに従ってサーバからHTML文書を取り寄せ，MHTMLに変換し，ユーザ側に返す．このとき，URLだけでは文書の符号化方法を判別できないため，ユーザがURLと同時にその文書の言語を指定する必要がある．ゲートウェイではこの情報に従って，サーバから取り寄せた文書を一旦ISO-2022-JP-2に変換してからMHTMLに変換する．サーバとの間にプロクシサーバを介することで，ゲートウェイからはHTTPプロトコルのみでGopherおよびFTPサーバへのアクセスを可能としている．

4.3 MHTMLビューア

MHTMLビューアはユーザのWWWブラウザ上のJavaアプレットとして動作し，通常のWWWブラウザと同様にURLの入力，ナビゲーションなどの機能を持つ．ビューアのプログラムコードは，ユーザがMHTMLゲートウェイに最初にアクセスした時にゲートウェイから送られる．MHTMLビューアの外観を図4に示す．

現在のMHTMLビューアはHTMLの主なタグに対応しているが，<TABLE>タグ，入力フォーム，クリッカブルマップなどは未対応である．また，インラインイメージはレイアウト処理の単純化のため，イメージそのものへのリンクとして実現している．

5 ゲートウェイサービスについて

我々は1996年8月8日からこの多言語WWW文書ブラウザをインターネット上で公開し，試験的にサービスを行っている[4]．現在（1996年10月20日）までに計1,439ページがMHTMLビューアによってアクセスされている（学内からのアクセスを除く）．

現在対応している言語（文字セット）は，日本語（JIS，Shift-JIS，EUC-JP），中国語繁体字（Big5），中国語簡体字（GB），韓国語（EUC-KR），タイ語（TIS）であるが，今後他の言語にも対応していく予定である．また，ゲートウェイのプログラムを移植可能なように整備し，他のサイトでのゲートウェイの立ち上げや対応する言語の追加を簡単に行えるようにする予定である．

6 おわりに

MHTMLブラウザの今後の課題として，入力フォームでの入力機能の実現，文書の印刷機能の実現などが挙げられる．このブラウザを用いて様々な応用が可能であると思われるが，現在我々は，日本語環境を持たないWWWブラウザから日本語資料の検索を行うためのOPACの開発を行っている[6]．

多言語対応ブラウザは将来のディジタル図書館にとって必須のツールであると言えるが，インターネット上でまだ多言語文書環境が一般的でない現状にあって，本稿で示した方法は多言語文書の表示環境を実現するための現実的な方法として有効であると思われる．

[4]URL: http://mhtml.ulis.ac.jp/

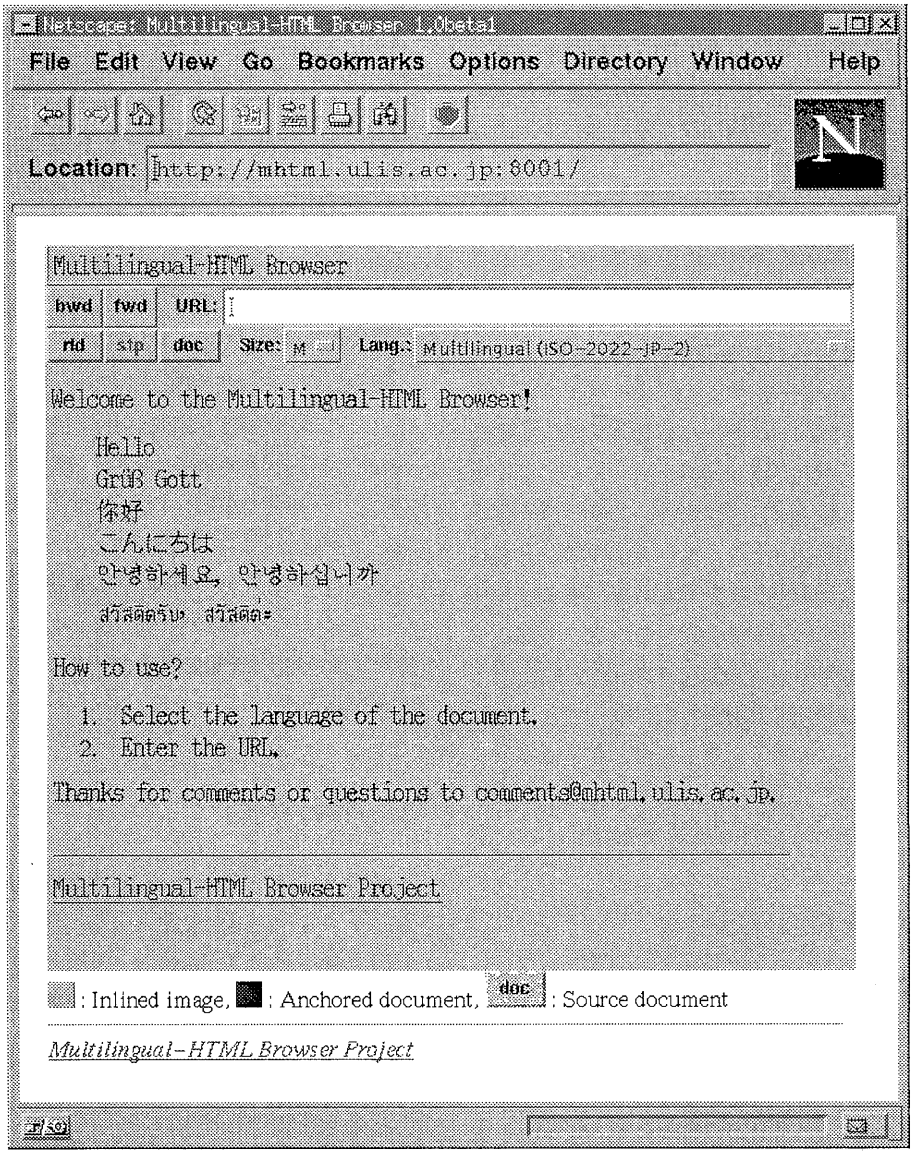

図 4: MHTML ビューアの外観

参考文献

[1] 前田亮，藤田岳久，リースエイチュー，阪口哲男，杉本重雄，田畑孝一：組み込みフォントを必要としないWWWのための多言語ブラウザ，ディジタル図書館，No. 4, p.21-25 (1995).
<URL:http://www.DL.ulis.ac.jp/DLjournal/No_4/maeda/maeda.html>

[2] Maeda, A., Fujita, T., Choo, L. S., Sakaguchi, T., Sugimoto, S. and Tabata, K.: A Multilingual Browser for WWW without Preloaded Fonts, in Proceedings of International Symposium on Digital Libraries 1995, p.269-270 (1995).
<URL:http://mhtml.ulis.ac.jp/papers/isdl95/isdl95.html>

[3] Sakaguchi, T., Maeda, A., Fujita, T., Sugimoto, S. and Tabata, K.: A Browsing Tool for Multi-lingual Documents for Users without Multi-lingual Fonts, in Proceedings of First ACM International Conference on Digital Libraries, p.63-71 (1996).

[4] Berners-Lee, T., Connolly, D.: Hypertext Markup Language - 2.0, RFC 1866 (1995).
<URL:http://ds.internic.net/rfc/rfc1866.txt>

[5] Ohta, M., Handa, K.: ISO-2022-JP-2: Multilingual Extension of ISO-2022-JP, RFC 1554 (1993).
<URL:http://ds.internic.net/rfc/rfc1554.txt>

[6] 藤田岳久，前田亮，阪口哲男，杉本重雄，田畑孝一：海外利用者のための日本語OPAC，ディジタル図書館，No. 6, p.32-39 (1996).
<URL:http://www.DL.ulis.ac.jp/DLjournal/No_6/take/take.html>