ポスト文字コード時代の文書処理技術に関する展望

守岡 知彦

概 要

長らく、十分に漢字を表現するためには、符号化された漢字の数が足りないということが叫ばれてきたが、ISO/IEC 10646-2 統合漢字拡張 B の制定以降、標準として符号化された漢字数だけに関していえば、大漢和辞典の収録文字数をはるかに上回るものとなっている。そのかわりに、文字の同一性や異体字の問題などが以前よりも重大な問題となってきている。このため、符号化された文字数が足りなかった時代以上に文字知識のデータベース化および文書技術との統合が重要なものとなっている。本発表ではこのような背景をもとに研究・開発を進めてきた CHISE (CHaracter Information Service Environment) の概要と今後の展望に関して述べる。

1 はじめに

パーソナル・コンピューターとインターネットの普及は電子テキストの利便性を広く示すこととなり、多くの分野で研究情報の電子化の機運をこれまで以上に高めている。東方学関連においても幾つかの大規模データベースが作られ、こうした動きは今後ますます盛んになって行くと思われる。

東方学における各種テキストや論文およびこれらのメタデータを電子化する際、従来より漢字の符号化の問題がたびたび指摘されてきた。殊に、文字符号に収録された漢字の数が足りないということがしばしば問題となり、e 漢字、文字鏡、GT, 漢字庫などの中・大規模の文字集合が提案されてきた。一方、国際的な事実上の標準として登場した Unicode [9] が当初符号空間を 16 bit 固定長 (UCS-2 [5]) としたために収録可能な漢字数が約 2 万字に制限され、無理な漢字統合を行ったため、漢字文献の電子化や国際的な情報交換を行う上で幾つかの問題があり、批判を浴びていたこともあった。

しかしながらこうした数の問題に限っていえば、ISO/IEC 10646-2 [6] で制定された統合漢字拡張 B の登場以降解消されたといえる。現在、UCS [5] / Unicode に収録された漢字は約 7 万字となり、数だけでいえば既に大漢和辞典や e 漢字、GT, 漢字庫を越えている。また、統合漢字拡張 C1 の制定作業が行われており、やがて文字鏡に収録された漢字の数を越えることとなろう。全ての文字レパートリーを UCS/Unicode に統合する動きは今まで以上に進み、やがては 10 万字を越えるような規模になるかも知れない。UCS/Unicode が当初符号空間を 16 bit 固定長としていたため、古い UCS/Unicode 系の実装は統合漢字拡張 B 以降をすぐに利用できないこともあるが、こうした問題は徐々に解消されて来ており、Mac OS X や Windows XP などでは既にフォントさえ存在すれば 7 万字以上の漢字が特別な外字セットなしに利用可能な環境が整いつつある。

このような状況を考えれば、漢字符号化における問題の中心は符号化された漢字の数の問題から次の段階に移ったのではないかと考えられる。もちろん現状でも表現できない漢字は存在するし、それが既に符号化された漢字の異体字にもなっていないという例も存在するであろうが、こうした事態は段々少なくなってきているのではないだろうか。むしろ既に符号化されているにも関わらず漢字が見つけ出せず新たに外字を作ってしまい、重複符号化してしまうというようなことの方が起こりやすくなっているのではないかと思われる。また、真に『足りない漢字』を見つけた場合、その出典・用例を示してその漢字の追加提案をする方が社会に対する貢献となるであろう。

一方、符号化された漢字が大幅に増えてしまったために引き起こされた問題がある。そのひとつが今まである UCS の符号位置で包摂されていた漢字を事実上分離してしまったために引き起こされた問題であり、この結果、異なるマッピング・テーブルを使う場合、同じ形の漢字が違う符号位置で表現される事態が生じており、ある意味、従来の漢字統合よりも深刻な問題を引き起こすようになった。また、そもそも異体字が増えたことによって従来よりも真剣に異体字処理を行う必要が生じてきたといえる。漢字を含むテキストを電子化する場合、ある文字をどの符号位置で表現すべきかについて自由度が増えたため、異体字を正規化するにせよしないにせよ、漢字をどう表現するかに関して今まで以上に真摯に取り組む必要が生じてきたといえる。

2　用途に応じた表現とその統合

文字符号は通信用の符号をその起源とし、本来、少ない情報量で文字に関する必要最低限の情報を伝えるためのものであったといえる。文字符号は本来視覚的情報を包含していなかったが、パーソナル・コンピューターなどの普及により、視覚的情報を伝えるものとしても使われるようになった。このことから、漢字においては多数の異体字を符号として区別することが要請されることとなり、『漢字が足りない』という問題が起こったといえる。やがて、検索や再利用やインターネットでの情報交換の利便性が認識されると、文字符号は文字の視覚的属性を含むさまざまな属性を背負わされることとなったといえる。そして、文字符号に期待する役割は利用者や分野によって微妙に異なり、文字符号制定の場は相反する利害対立を調整するための政治的な場となり、その結果である文字符号は妥協の産物となることを余儀なくされている。こうした問題は文字符号で全てを表現しようという問題設定によって生じているといえる。つまり、少ない情報量で文字に関する必要最低限の情報を伝えるための仕組を機械的に拡張して、少ない情報量で文字に関する多数の情報を伝える仕組として使おうとしたことに無理があったということである。

このような問題を解決するためには、用途毎に適した表現を用いることが考えられる。例えば、中に書かれた内容に対する検索を行うことが目的ならば、異体字はなるべく正規化すべきだといえる。もちろん、異体字データベースを構築して検索時に対処することはできるが、検索のコストが増えることもあるし、どこまで異体字を分離するかの基準が入力する人によって異なったために異体字表現の品質を整えるためにコストがかかるということもある。逆に視覚的情報に意味がありそれを忠実に表現することが目的だとすれば、画像を利用する方が適切であるといえる。この場合、そのままでは検索ができないので目的に応じて適切なメタデータを付けるなどの工夫が必要である。

しかしながら、電子テキストの利便性のひとつが再利用性であり、本来とは違った目的のデータに転用することを考えれば、特定用途専用というのもまた問題である。このようなことを考えれば、それぞれの用途に適した表現が可能であることと、そうした各表現を統合する枠組の存在が重要であるといえる。このような枠組として有望なもののひとつが XML (eXtensible Markup Language) [10] である。XML では用途に応じたタグセットを定義することができ、実際に各種分野用にさまざまなタグセットが提案され利用されている。人文科学で扱うさまざまなテキストに対しても TEI (Text Encoding Initiative) コンソーシアム [8] がタグセットに関するガイドライン [7] を作成している。XML は本来文字で書かれた文書のマークアップを想定していたが、マークアップする対象は（文脈自由文法で記述可能な）なんらかの構造を持ったデータならば何でも良く、画像に対しても SVG (Scalable Vector Graphics) [11] というタグセットが制定されている。

SVG は Postscript や PDF の XML 版といったものであるが、他のタグセットと共同で運用す

ることが考慮されており、SVG 用プラグインを利用したり SVG 対応 WWW ブラウザーを利用することによって XHTML などに埋め込んで利用することが可能となっている。よって、テキストの論理構造を TEI などで表現し視覚的構造を SVG で表現することが可能で、内容の論理構造と視覚的構造の双方の側面をそれぞれに適した形で表現しつつ、両者を統合することが可能となる。そこで著者は SVG を用いた文字画像のマークアップ手法を試みている。

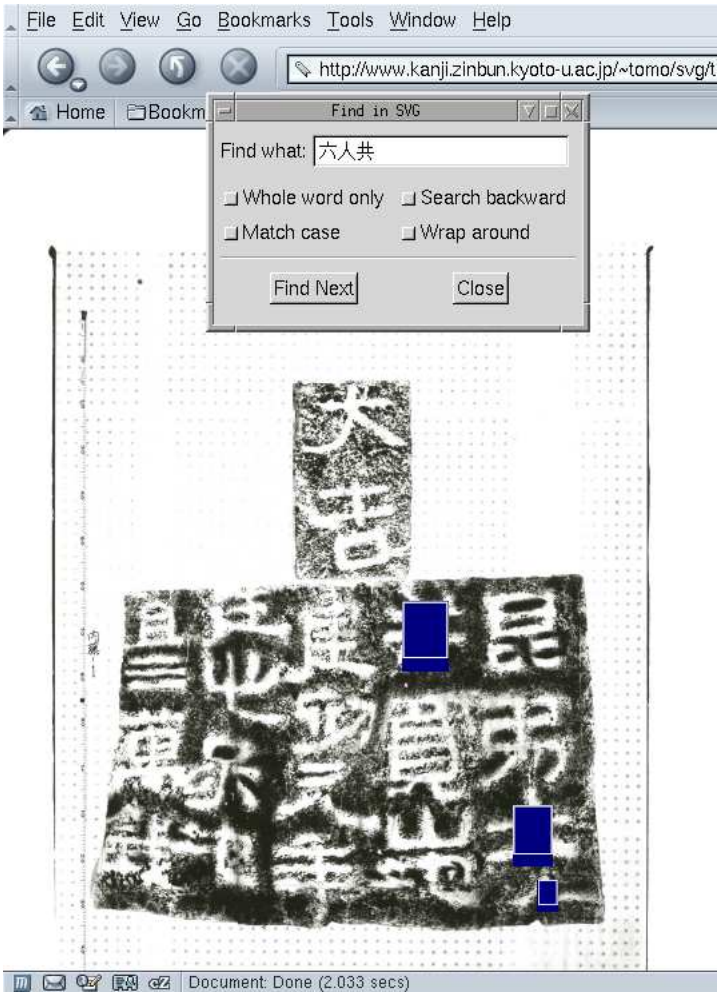

図 1: SVG を用いた文字画像の例

3　文字データベースに基づく文字処理アーキテクチャ

XML のようなマークアップ手法はテキストの持つ多様な性質を用途毎に適切な形で表現するのに有用であり、こうして記述された構造化テキストは異なる形式に変換することも容易であるし、異なる構造化テキストを統合したハイパーテキストを構成することも可能である。一方、テキストと同様に文字にもさまざまな側面があり、内部に構造を持ったり、他の文字との関係が存在し、こうした性質は用字系 (script) や書記系 (writing system) の構造を形作ったり、テキストの性質に影響を与えたり与えられたりする。特に、漢字は形状（形）でもって発音（音）と意味（義）を表現する『表語文字』であるため、アルファベット系表音文字と異なり、形態素や単語に似た性質を

持っている。その端的な例が異体字である。異体字を『同じ意味で使われる文字が異なる形状で表されるもの』と考えるならば、意味の文脈依存性によって必然的に異体字は文脈依存性を持つことになる。このため、漢字を精密に記述しようとするならば、形音義をはじめとするさまざまな属性を明記しその文字をどういうものとして捉えているかを明示することが重要であると考えられる。理想的にはテキストにおける文字の出現毎に文脈依存の属性も含めて表現するのが望ましく、そうした文字の性質の表現がテキスト層におけるマークアップと連携することも重要であるといえる。

文字をこのようなものとしてとらえるならば、それを処理するシステムは文字に関するさまざまな属性を動的に参照・処理できるようなものである必要があるといえる。このような観点に基づき、著者は 1999 年から **UTF-2000** [1] と呼ぶ、文字を属性の集合として表現・処理する新しい文字処理アーキテクチャの開発を行ってきた。また、2001 年からはこれを発展させた形で **CHISE (CHaracter Information Service Environment)** プロジェクト [13] を開始した。

CHISE プロジェクトは、UTF-2000 プロジェクトで開発してきた文字データベースに基づく文字オブジェクト技術に基づき、文字に関するさまざまな情報・知識の編集や、文書の編集・組版・印刷などの各種処理を一貫して行なうことが可能な総合的な文字操作環境の開発を目指している。

3.1 XEmacs UTF-2000

XEmacs UTF-2000（図 2, 図 5）は著者らが開発している文字データベースに基づく文字オブジェクト技術を利用した多言語文書編集環境である。これは文書編集系 XEmacs を元に、文字・文字列・バッファの内部表現を変更し文字属性を管理するデータベース機構を付けるなどして、文字属性の集合で文字を表現する手法である『UTF-2000 方式』に基づく文字処理を実現したものである。

XEmacs UTF-2000 では文字は文字属性の集合によって定義される文字オブジェクトとして扱われ、各文字オブジェクトには固有の「文字 id」が割り当てられる。この文字オブジェクトは最大 2^{30} = 約 10 億個定義可能であり、非常に大量の種類の文字を同時に扱うことが可能となっている。なお、文字 id は内部的なものであり、利用者は文字 id ではなく文字の属性を使って文字を参照・処理するようになっている。

文字を定義するために XEmacs UTF-2000 では `define-char` という組込み関数を用意している。この他、文字属性の参照・設定関数、文字属性に対する map 関数、探索関数など文字符号を隠蔽して文字を処理するための関数群を追加しており、文字符号に依存しないプログラムが書けるようになっている。また、従来の XEmacs に対する上位互換性を持っており、XEmacs 用に書かれた多くのプログラムがそのまま動作する。

XEmacs UTF-2000 は文字を属性の集合として扱っているため、基本的には全文字の全属性を記憶しなければならなず、多くの記憶資源を要することになる。また、こうした文字データベースを XEmacs UTF-2000 内部の記憶空間に抱えている場合、XEmacs UTF-2000 の外部と文字属性データベースを共有することができない。このような問題点を解決するために、XEmacs UTF-2000 において外部の文字データベースを利用するための機構を開発している。これは外部の文字データベースから文字属性を必要な時に情報を獲得する (lazy-loading) ための枠組と、外部文字データベースの種類毎の実装（バックエンド）からなる。現在、XEmacs の `database` 機能（Berkeley DB のような属性値を保持するための単純なデータベースに対する抽象）を利用したバックエンドが実装されている。また、今後、PostgreSQL バックエンドの開発を予定している。

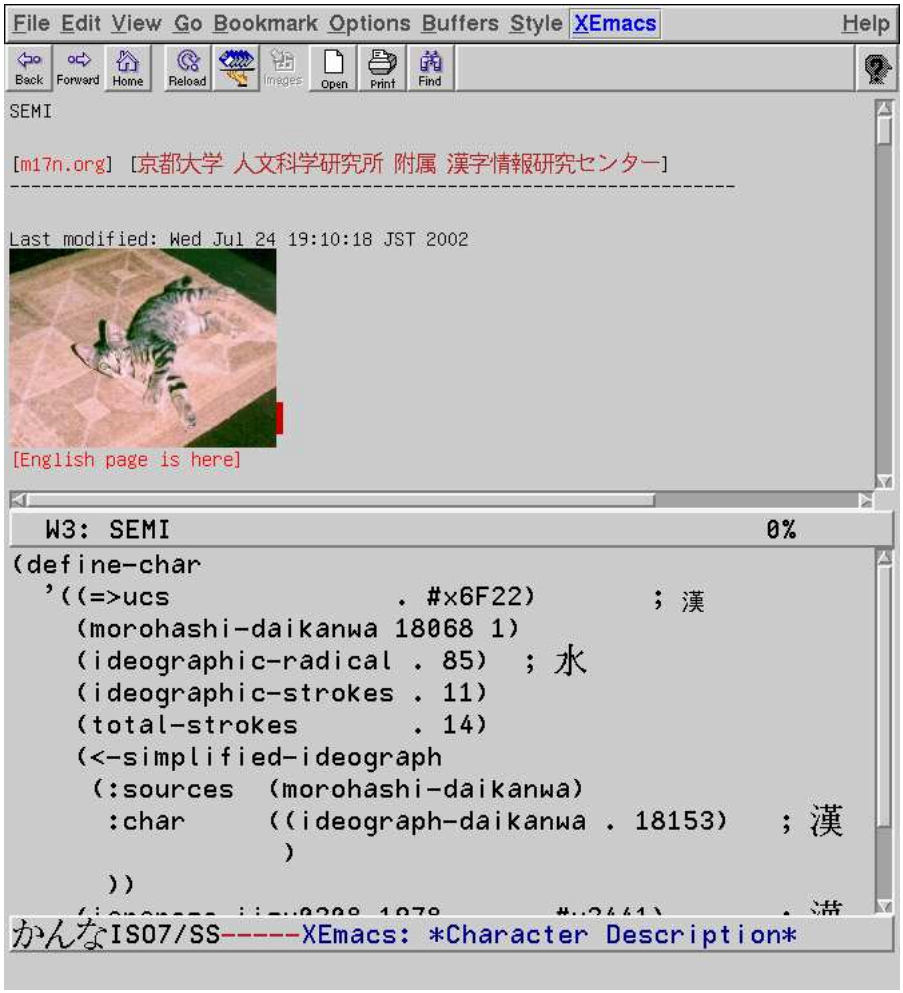

図 2: XEmacs UTF-2000

3.2 TopicMaps

Topic Maps [4] は対象となる情報の構造を topic（『話題』）の集まりとして捉え、topic の定義や topic 間の関係などを記述することで、さまざまな構造を持った各種情報リソースを表現しようとするものである。

TopicMaps に関する標準としては、SGML [2] および HyTime [3] Architectural Forms に基づく形式の DTD (Document Type Definition) 表現による仕様が ISO/IEC 13250:2000 [4] として制定されている。また、独立したヴェンダーのグループによって XML 版も開発され 2000 年 12 月に XTM (XML Topic Maps) 1.0 として公開されている。これは 2001 年 12 月に ISO 標準の修正 (amendment) として承認されている。そこで、CHISE プロジェクトではこの XML 版を採用することにした。

XEmacs UTF-2000 という実装が既に存在している文字属性表現および define-char 形式の場合と異なり、TopicMaps に基づく文字知識表現形式は今のところ処理系が存在していない。そこで、2001 年度から Christian Wittern 氏は Topic Maps エンジンの開発をはじめ、 Zope (Zope Object Publishing Environment) [12] を用いたプロトタイプを開発した。しかしながら、数万〜数十万の文字を対象に数十万〜数百万の topic を扱うには現状の Zope は不十分であり、また、XML の処

理を行う上で必須といえる UTF-8 の処理にもバグがあり、このプロトタイプはあまり実用的なものとはいえない。このため、今後は PostgreSQL を用いた新たな Topic Maps エンジンや XEmacs UTF-2000 上で動作する Emacs Lisp で記述した Topic Maps 編集システムの開発を予定している。

このような事情から、現在のところ TopicMaps に基づく文字知識表現手法の開発はまだあまり進んでいない。とりあえず現在のところ、抽象文字、文字インスタンス、異体字形、文字構造、言語、読み、意味、時代、空間、良く使われる用例、符号化文字集合への写像、辞書への参照 などの文字属性軸が定義されている。

今後は文字属性記述に関するガイドラインと TopicMaps に基づく文字知識表現形式の双方の開発を進めるとともに、両者間の相互変換を実現し、文字属性と TopicMaps の両方の視点で文字知識を操作できるようにしたいと考えている。

3.3 文字データベースに基づく文書処理環境

文字データベースを XEmacs UTF-2000 の外部に保持する目的のひとつは、必要な文字の必要な情報だけを保持するようにして記憶資源の効率化を計ることであるが、文字に関する知識をさまざまなアプリケーションから利用できるようにする上でも非常に大きな意味がある。

この観点からいえば、現行の Berkeley DB に基づく文字属性毎の単純なデータベース機構は管理のしやすさやスケーラビリティーの点で問題があると考えられ、現在、CHISE Project では PostgreSQL の採用を検討している。

また、PostgreSQL に基づく文字データベース・サーバーのクライアントとして XEmacs UTF-2000 の他に、Ω に基づく多言語組版系、グリフ・字形情報の管理・合成系、漢字間の関係の視覚化システムの実現を目指している。これらの連携により図 3 に示すような文書処理環境を構成すれば、既存の符号化文字集合に依存せずに文字やテキストを一貫して編集・処理可能な環境が実現できると考えられる。これが CHISE プロジェクトの当面の目標であり、近い将来（2003 年中）における実現を目指している。

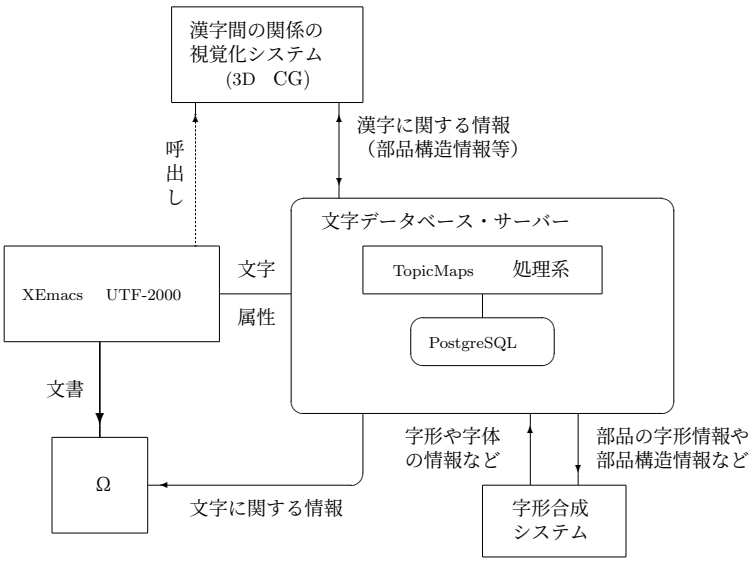

図 3: CHISE System 2003

4　文字知識のデータベース化

文字データベースに基づく文字処理システムを活用するためには文字知識のデータベース化が不可欠であり、我々はこうした文字データベース・コンテンツの開発にも力を入れている。2001年度から漢字の部品の組合せ構造を表現した漢字構造情報データベースの開発をはじめ、現在、ISO/IEC 10646-1:2000 [5] 基本統合漢字（Unicode の例示字形）, 同 統合漢字拡張 A, および ISO/IEC 10646-2 [6] 統合漢字拡張 B に対する入力を一応完了している。

現在のデータベースは、Unicode Database, CNS 11643 と諸橋大漢和辞典の対照表、CDP データベース 、CBETA 外字データベース、CHINA3 外字データベース、著者らがこれまで作成してきたその他の雑多なデータベース等を統合し、互いの矛盾点を修正するものである。まだ誤りも多く、品質は高くはないが、現時点で約 10 万字分の定義が存在する。

4.1　漢字構造情報データベース

多くの漢字は偏と旁などの部品の組み合わせによって構成されている。こうした漢字の部品の組合せ構造は形の抽象的表現となるだけでなく、字義や音価にも関係しており、字源に基づく文字構造の分析は「解字」と呼ばれ、そうしたデータは重要な辞書記述の１つである。そこで我々は、UTF-2000 基本データベースに収録されている全ての複合（会意・形声）漢字に対し漢字構造情報を付けることを目標に、漢字構造情報データベースの開発をはじめた。

漢字構造情報に基づく符号化手法の試みは 1970 年代に遡るが、漢字符号化の主流とはならず、標準的な記法も確立されて来なかった。その後、ISO/IEC 10646-1:2000 [5] においてはじめて漢字構造情報の標準記法である IDS (Ideographic Description Sequence) とそのためのオペレーター群である IDC (Ideographic Description Characters) （図 4）が定義された。そこで、我々はこの IDS に基づく方法と、これを S 式（Lisp 表現）化し付加情報を許した *ideographic-structure* 形式、および、これを Topic Maps 化したものを用いることにした。

既存の漢字構造情報を含んだデータベースとしては、台湾中央研究院の CDP データベースと台湾の中華電子佛典協會 (CBETA) の外字データベースがある。これらはそれぞれ独自の形式を採っている。なお、これらは GPL で配布されるプログラムで利用可能であるので、可能な限り変換して利用することにした。また、この他、日本でも「今昔文字鏡」があり、検字を目的としたは８万字以上の解字情報を持つが、今の所、自由ソフトウェアで利用することはできない。またこの他、和田研フォントや GT 書体など、フォント合成を目的にしたいくつかの試みが存在するようである。

4.1.1　CDP データベース

CDP (Chinese Document Processing) データベースは、1990 年から台湾中央研究院の謝清俊らが開発している文字データベースで、現在、漢語大辞典に収録されている文字を中心に 55500 字以上を含んでいる。

CDP データベースは Big5 と外字を利用して符号化されており、外字を利用して 14 種類のオペレーターを定義している。そのうち３種類は部品の結合を表現したもので、(1) 縦に並べる、(2) 横に並べる、(3) その他 を表現する。また 8 種類の反復記号がある。これは同じ部品を複数個配置することを表現するものである。また、この他に特殊記号が用意されている。なお、CDP データベースの漢字構造表記では入れ子を認めていない。また、IDS に比べ、結合オペレーターの種類

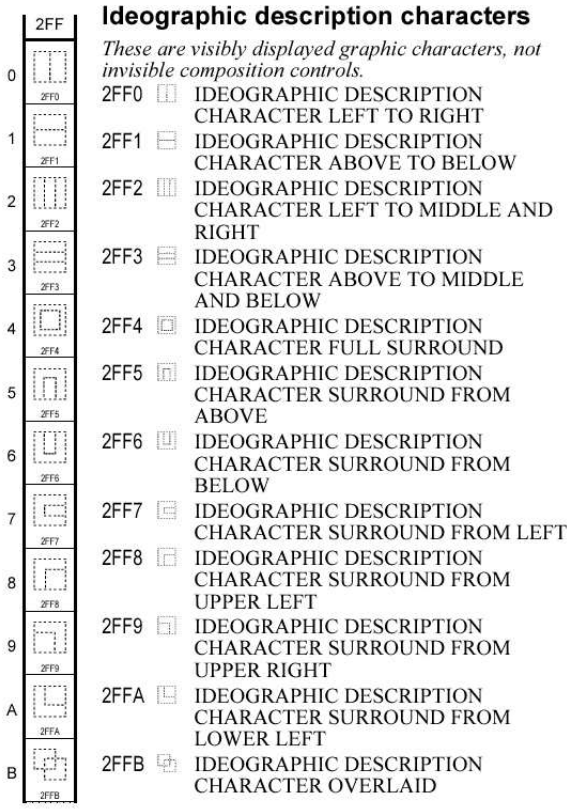

図 4: Ideographic Description Characters

が少なく、結合オペレータ (3) から IDC への変換に曖昧性があるため、機械的に IDS へ変換することはできない。

4.1.2 CBETA 外字データベース

台湾の中華電子佛典協會 (CBETA) は仏典データベースの作成上必要となった外字を対象に文字データベースを作成しており、この外字データベースには現在 13000 字程が収録されている。

CBETA 外字データベースも Big5 と外字を利用して符号化されているが、オペレーターは ASCII 文字を利用する。中置記法を使っており、‘(’ と ‘)’ を用いることで入れ子表現も可能である。結合オペレーターとしては CDP データベースのものと同様に、縦に並べることを表現する ‘/’ と、横に並べることを表現する ‘*’ と、その他を表現する ‘@’ の 3 種類である。この他に、部品の削除と置換を表現するためのオペレーターが存在する。部品の削除は $A - B$ という式で表現され、これは文字 A から部品 B を削除したものを表現している。例えば、草冠は 草 – 早 で表現できる。部品の置換は $A - B + C$ という式で表現され、これは文字 A 中の部品 B を C に置き換えることを表している。例えば、「花」は 草 – 早 + 化 で表現できる。この削除と置換を用いることで、3 種の結合オペレーターだけでは十分に表現できないような漢字でも、多くの場合において曖昧無く表現可能である。次により複雑な例として ((((瞭 – 目) – 小) – 日 + (工/十)) ∗ 支)/皿 を示す。CBETA 外字表現は少数の規則で驚く程多くの漢字をカバーできる半面、複数の式表現が生じやすいといえる。なお、なるべく ‘@’ を使わずに表現された式表現は、結合漢字の IDS 形式のデータ

ベースが存在すれば、機械的に IDS へ変換することが可能である。

4.1.3 CHISE 漢字構造情報データベース

我々は漢字構造情報表現として、IDS に基づく入れ子上の木構造形式を採用し、プレーン・テキストでは IDS 形式（図 5）、XEmacs UTF-2000 の文字データベースでは *ideographic-structure* 形式に基づく `ideographic-structure` 属性、XML では Topic Maps に基づく形式で扱うことにした。

そして、CDP 形式や CBETA 形式からの変換プログラムを作成し、機械的に変換可能な部分に関しては利用可能な既存のデータを用いることにした。しかしながら、前述のように CDP データベースは機械的に変換することができず、また、形式に則っていないと思われる部分もあり、我々の目的にとっては十分なものではなかった。一方、CBETA 外字データベースは仏典で用いられる特殊な文字が主であり、データソースの点で難点がある。また、IDS に変換するためには結合漢字を IDS で表現したデータベースが必要となる。そこで、CDP データベースから変換したデータを修正・補完することで、Unicode の基本統合漢字、ISO/IEC 10646-1 の統合漢字拡張 A, ISO/IEC 10646-2 [6] 統合漢字拡張 B, その他のレパートリの順に漢字構造情報データベースを開発することにした。

現在の所、基本統合漢字、拡張 A および拡張 B に関する入力作業はほぼ終了しており、現在、校正作業を行っている。また、この入力作業のために Christian Wittern 氏は CDP 外字や ISO/IEC 10646-1,2 の漢字を含む 7 万字以上の漢字を対象とした quail（GNU Emacs/XEmacs 用の標準的な入力システムの 1 つ）に基づく四角號碼方式の入力システムを開発した。

5 おわりに

XML に基づいて論理構造と視覚的構造の双方を表現可能な電子テキストを実現する試みを紹介すると共に、文字データベースに基づく文字処理技術を開発している CHISE プロジェクトの現状と今後の展開に関して説明した。

SVG を用いることにより、XML の枠組に基づいて画像をマークアップすることが可能であり、特別な文字コードやフォントを使うことなく原テキストにおける視覚的情報を再現可能である。また、テキストの論理構造を記述する他のタグセットと共同運用することにより、論理構造と視覚的構造を有機的に統合することも可能である。

一方、CHISE プロジェクトでは、全ての文字を文字データベースに基づいて処理する文字処理技術の研究・開発を進めているが、この手法は XEmacs UTF-2000 が実証しているように現在の計算機環境において十分な実用性を有していると考えられる。また、文字に関する知識を積極的に利用することで、さらなる可能性があると考えられる。

こうした可能性を現実のものとするためには、プログラム・データ双方のさらなる整備が必要であり、多言語化技術、国際化技術、データベース、分散化技術など各方面のハッカーや、言語学、文字学、文献学など各種言語や字書やテキストなどを扱う人文科学者など、多方面に渡る研究者の幅広い協力が不可欠である。

こうしたことを鑑み、CHISE プロジェクトはオープン・ソースで開発されており、その情報は

- http://cvs.m17n.org/chise/
- http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/

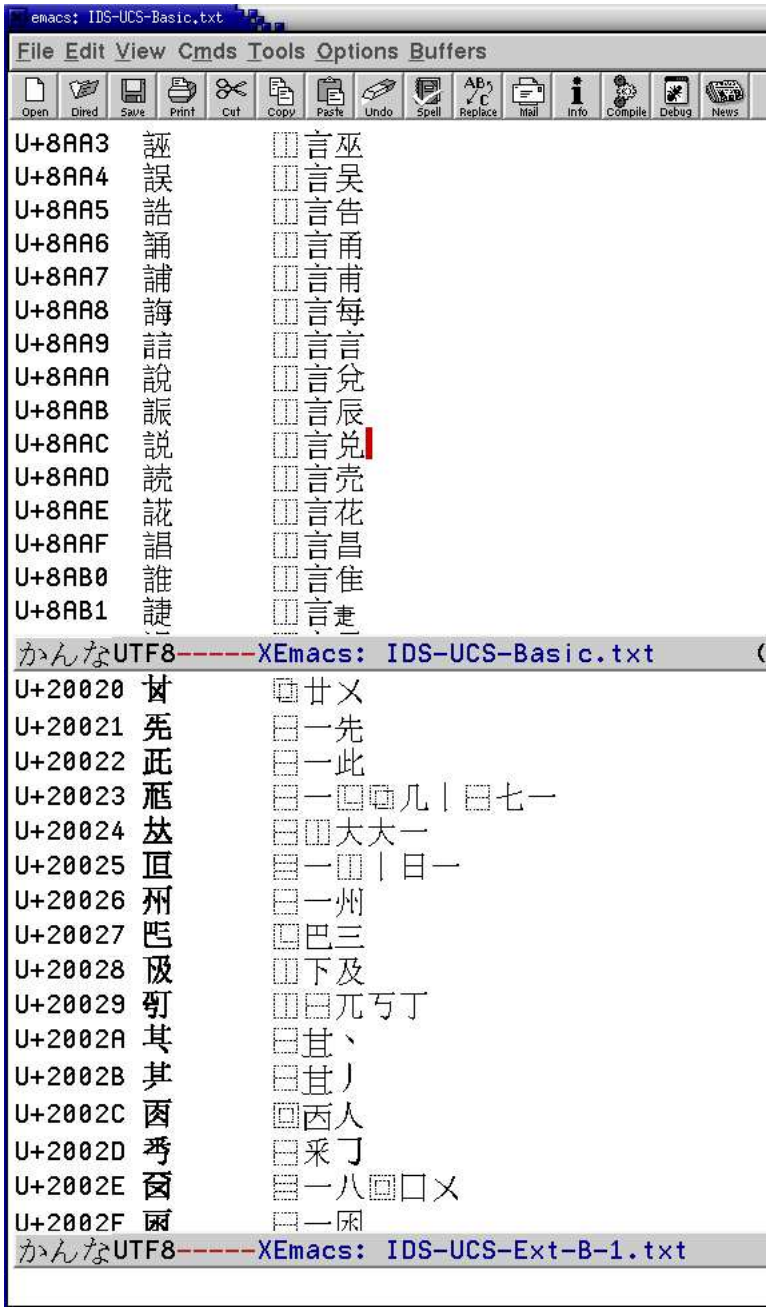

図 5: 漢字構造情報データベース

- http://mousai.as.wakwak.ne.jp/projects/chise/

で公開されている。これらの WWW 頁群を含め、各種成果物は CVS で管理されており、最新の開発状況を知ることができる。日本語用と英語用の２つのメーリングリストも用意されており、参加方法は上述の WWW 頁で説明されている。CHISE プロジェクトに興味を持たれた方は、是非、お気軽に御参加願いたい。

謝辞

本論文で述べた CHISE プロジェクトの研究の一部は 2000 年度に旧通商産業省工業技術院電子技術総合研究所（現 独立行政法人 産業技術総合研究所）からの受託研究として行われ、2001 年度には情報処理振興事業協会の「未踏ソフトウェア創造事業」の助成を受けた。

また、忙しい中 CHISE プロジェクトに主要メンバーとして御参加頂いた Christian Wittern 氏、江渡浩一郎氏、上地宏一氏、鈴木泰博氏、苫米地等流氏、藤原義久氏、宮崎泉氏、師茂樹氏に感謝する。また、2001 年度に未踏ソフトウェア創造事業のプロジェクトマネージャーとして、その後も、プロジェクト遂行において貴重なご助言と多大な御助力を頂いた g 新部裕氏に感謝する。

また、横田裕思氏やしおざきかずひこ氏をはじめとする UTF-2000 mailing list の参加者に感謝する。また、XEmacs UTF-2000 の開発にあたって貴重なご助言と御助力を頂いた産業技術総合研究所の戸村哲氏、半田剣一氏、錦見美貴子氏、高橋直人氏に感謝する。

参考文献

[1] bit 別冊「インターネット時代の文字コード」, 第 9 章「文書編集系における文字コード」. 共立出版, 2001.

[2] International Organization for Standardization (ISO). *Information processing — Text and office systems — Standard Generalized Markup Language (SGML)*, 1986. ISO 8879:1986.

[3] International Organization for Standardization (ISO). *Information processing — Text and office systems — Hypermedia/Time-based Structuring Language (HyTime)*, 1997. ISO 10744:1997.

[4] International Organization for Standardization (ISO). *Information technology — SGML Applications — Topic Maps*, January 2000. ISO/IEC 13250:2000.

[5] International Organization for Standardization (ISO). *Information technology — Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane (BMP)*, March 2000. ISO/IEC 10646-1:2000.

[6] International Organization for Standardization (ISO). *Information technology — Universal Multiple-Octet Coded Character Set (UCS) – Part 2: Supplementary Planes*, November 2001. ISO/IEC 10646-2:2001.

[7] C. M. Sperberg-McQueen and Lou Burnard, editors. *TEI P4: Guidelines for Electronic Text Encoding and Interchange — XML-compatible edition.* University of Oxford, 2002.

[8] The TEI consortium. http://www.tei-c.org/, May.

[9] The Unicode Consortium. *The Unicode Standard, Version 3.0*, February 2000.

[10] The World Wide Web Consortium (W3C). *Extensible Markup Language (XML) 1.0 (Second Edition)*, October 2000. http://www.w3c.org/TR/2000/REC-xml-20001006.

[11] The World Wide Web Consortium (W3C). *Scalable Vector Graphics (SVG) 1.0 Specification*, September 2001. http://www.w3.org/TR/SVG/.

[12] Zope. http://www.zope.org/.

[13] 守岡知彦, クリスティアン・ウィッテルン. 文字データベースに基づく文字オブジェクト技術の構築. 情報処理振興事業協会 平成 13 年度 成果報告集. 情報処理振興事業協会, 2002. http://www.ipa.go.jp/NBP/13nendo/reports/explorat/charadb/charadb.pdf.