文字知識に基づく文書処理環境の現状と未来

– Character Information Service Environment Project –

守岡 知彦

京都大学 人文科学研究所 附属漢字情報研究センター

はじめに

CHISE (CHaracter Information Service Environment) Project:

文字データベースに基づく文字処理技術の開発

- **文字符号から文字知識へ**

背景（その 1）

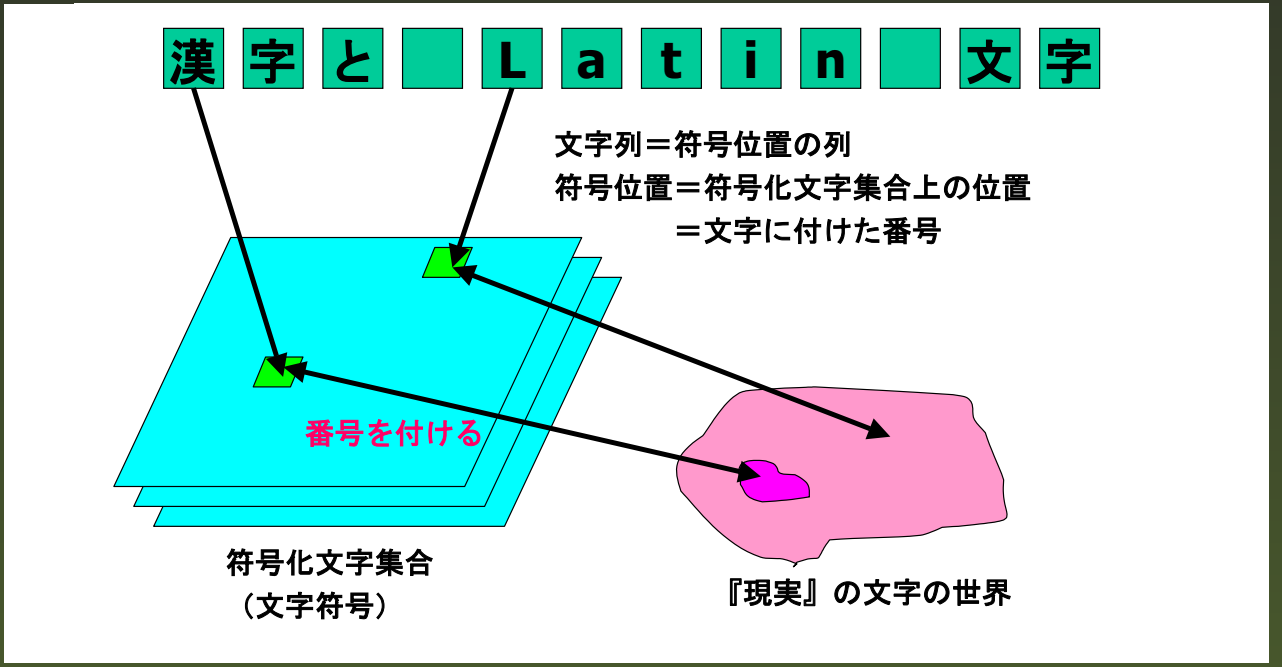
- 電子テキストは 素晴らしい !!!
  - 検索
  - 再利用
  - 配布

  が 容易
- でも 文字符号に 縛られる
  - 外字問題
  - Code Wars
- 文字問題の 平和的解決!?に 向けて
  - 文字符号に 頼るのを 止めよう !?

そもそも文字符号って何?

– 符合化文字方式 –

- 文字を 番号で 表現する



文字に 関する 知識は 符号化文字集合の 定義 の 中に

- 符号を いろんな用途で 使い 回す （ 表現、 表示、 検索、 再利用など ）

汎用文字符号からの脱却

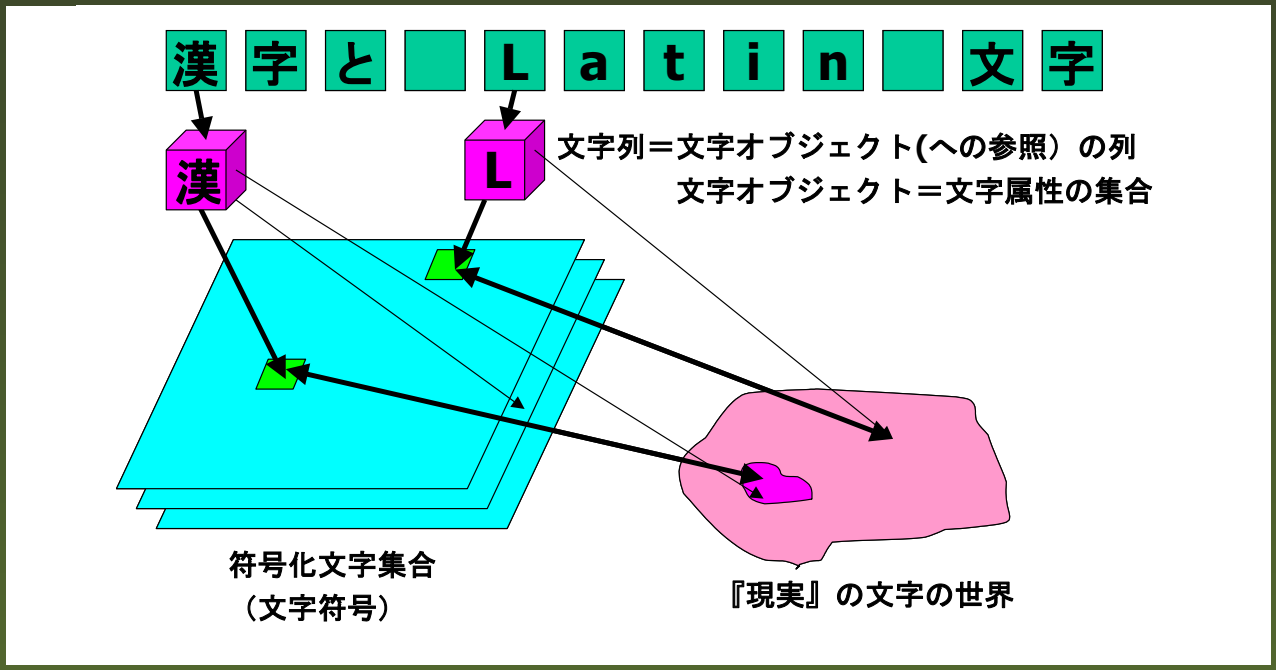
- 用途に応じて要求される性質が違う
  → 用途毎に違う表現を使えば？
  例：表現は RDF とか、表示は SVG とか、
  検索目的なら異体字を正規化とか
- 文字の性質や用字意図の明示化
  → 文字に関する知識を計算機の中へ

背景（その２）

- 文字に関するさまざまな情報・知識の一元管理したい
  - mapping table
  - 字体・字形情報
  - 用例
  - その他各種属性など

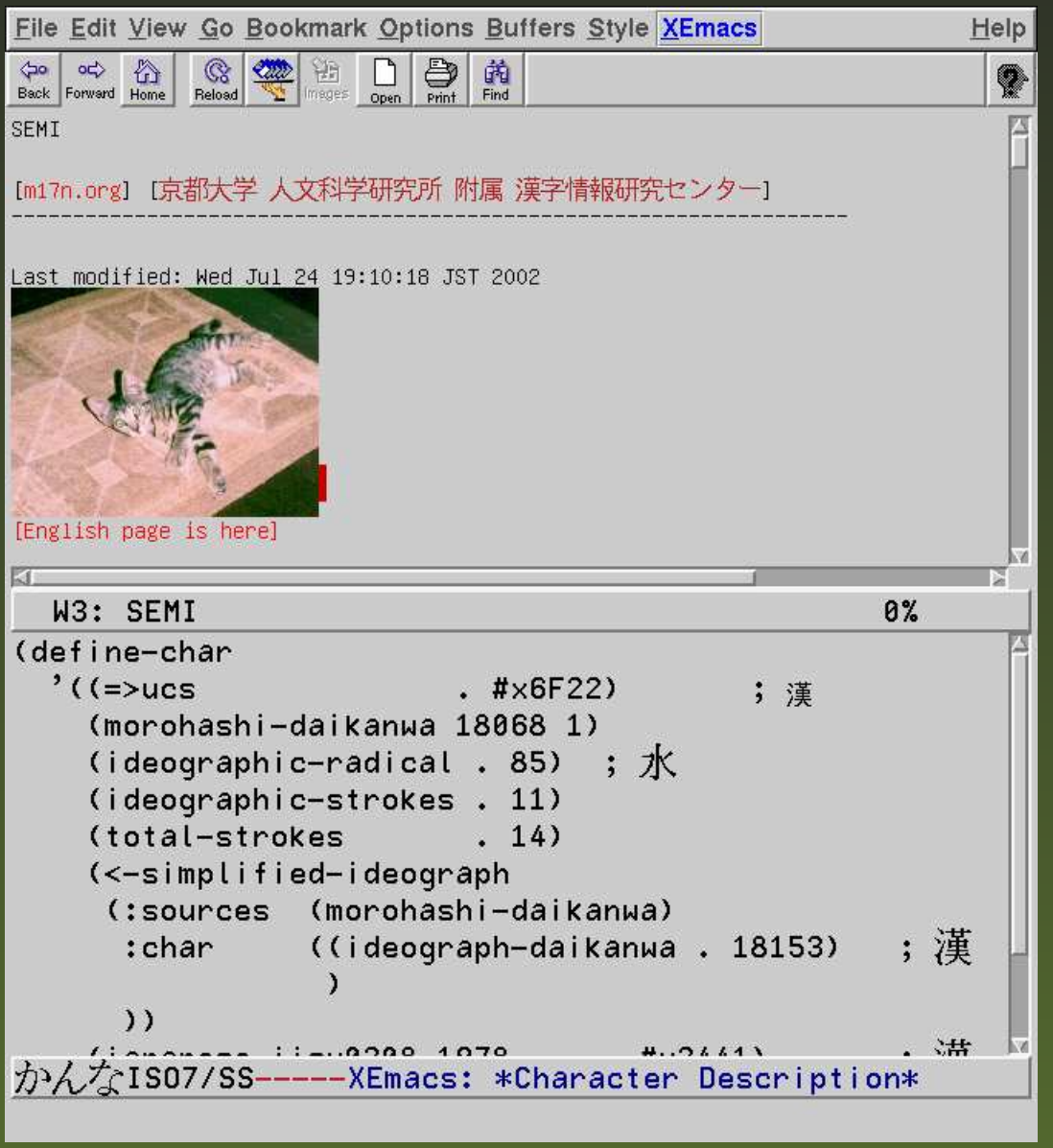
UTF-2000 方式

- 文字を 文字属性の 集合（ 大域 的に 情報交換 可能な もの ） で 表現
- 文字オブ ジェクト ID の 機 械的割 り 当て
- 文字の 内部表現へ の 直接ア クセスの 禁止



XEmacs UTF-2000

- XEmacs に基づく UTF-2000 方式の試験実装
- XEmacs-Mule に対する上位互換性
- 豊富な文字空間 (最大約 10 億文字定義可能)
- 4 byte 以内の任意の文字符号が利用可能



XEmacs UTF-2000 の 問題点・課題

- でかい
- 文字をどう定義するか？（どんな文字属性を付けるか？）
- 文字属性情報をどう交換するか？
- 文字属性情報をどう利用するか？

lazy loading (1) – 目的

文字に関する各種属性および文字符合関連の情報を外部のデータベースに置き、必要な時に読み込むようにする

- 主記憶の節約
- 定義した文字や文字属性を読み書きしやすくする
- 他の利用者やアプリケーションとの情報の共有

lazy loading (2) – 枠組

- 情報が外部に存在することを示す Lisp オブジェクト Qunloaded を導入
- 文字属性表参照関数 get_char_id_table に load_char_attribute_maybe を追加
- 符号化文字集合復号関数 DECODE_DEFINED_CHAR に load_char_decoding_entry_maybe を追加
- map 問題への対策
- その他メンテナンス用、bootstrap 用等

lazy loading (3) – map 問題

map-char-attribute (map-char-table) が 非常に 遅く なる

- とりあえず 文字属性を 一括 load する 関数を 追加して、 on memory で 動かすように した
- でも map-char-attribute って あんまり 何回も 動かさないので、やっぱり 遅い (;_;)

  そもそも、世の 中に 存在する 全文字・全属性を 対象に 全数探索する ことは 無謀。
  こういうのを 多用するのは 間違っている !!!

lazy loading (4) – 結果

- バックエンド
  - XEmacs の database 機能を使用
    - 今のところ Berkeley DB でのみ動作を確認
- ダンプ後の実行形式

  XEmacs-Mule 10 MB (strip 後 6 MB)

  従来型 **XEmacs UTF-2000** 32 MB (strip 後 27 MB)

  lazy-loading 版 16 MB (strip 後 11 MB)

文字属性データベース

- XEmacs UTF-2000 附属の文字データベース
  - 自由に利用可能な各種データベースを統合・整理したもの
  - 約 10 万字（define-char 換算）収録
  - 漢字を細かく分離している（やりすぎ!?）
- 漢字構造情報データベース

漢字構造情報データベース

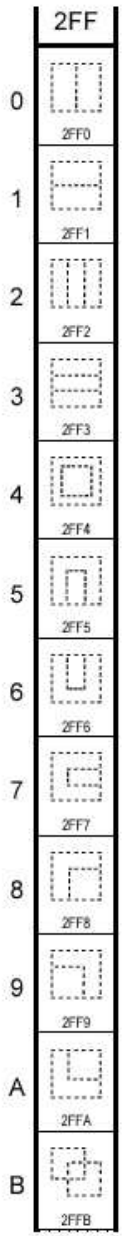
漢字の部品の組合せ構造の機械可読な表現

- 形の抽象的な表現
- 字義や音価にも関係している

表現方法としては IDS (Ideographic Description Sequence) を採用

- IDC (Ideographic Description Characters) をオペレーターとして用いた前置記法（入れ子状表現が可能）
- ISO/IEC 10646-1:2000 で定義されている

IDC



Ideographic description characters

These are visibly displayed graphic characters, not invisible composition controls.

2FF0 IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT

2FF1 IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW

2FF2 IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT

2FF3 IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW

2FF4 IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND

2FF5 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE

2FF6 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW

2FF7 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT

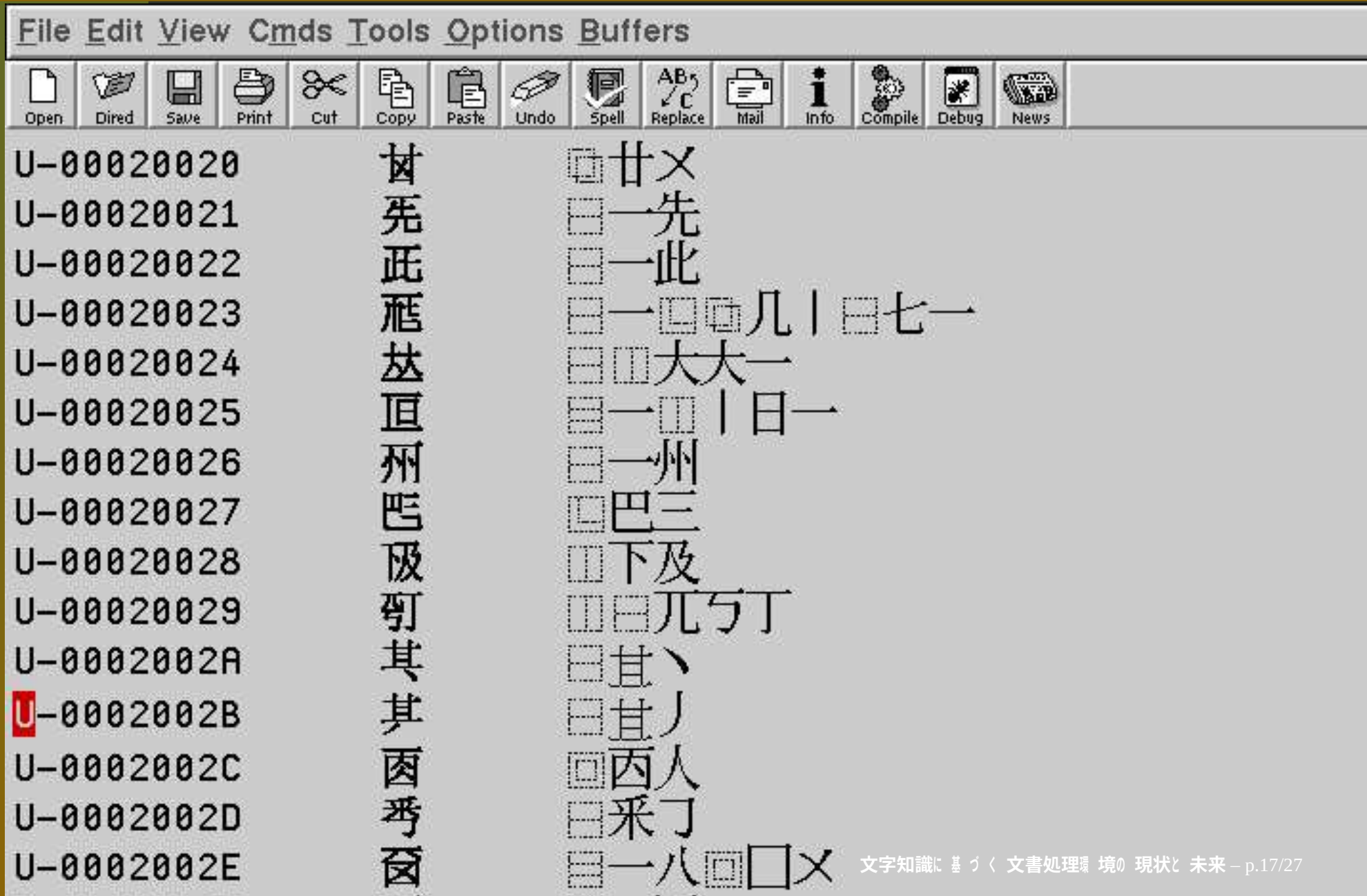
2FF8 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT

2FF9 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT

2FFA IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT

2FFB IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID

IDSの例


File Edit View Cmds Tools Options Buffers
Open
Dired
Save
Print
Cut
Copy
Paste
Undo
Spell
Replace
Mail
Info
Compile
Debug
News
U-00020020 廿 ⿲廿メ
U-00020021 𠂔 ⿱一先
U-00020022 𠂤 ⿱一此
U-00020023 𠂭 ⿱一⿰⿰几丨⿱七一
U-00020024 𠂇 ⿱⿰大大一
U-00020025 𠂋 ⿱一⿰ 丨日一
U-00020026 州 ⿱一州
U-00020027 𠂑 ⿰巴三
U-00020028 𠂔 ⿰下及
U-00020029 𠂢 ⿰⿰元ㄎ丁
U-0002002A 𠂣 ⿱甘丶
U-0002002B 𠂤 ⿱甘丿
U-0002002C 𠂥 ⿰西人
U-0002002D 𠂦 ⿱釆丁
U-0002002E 𠂧 ⿱一八⿰口メ

CHISE 漢字構造情報データベース

- 利用可能な既存のデータベースをなるべく変換
  - CDP (Chinese Document Processing) Database (台湾中央研究院の謝清俊らによる；約55000字収録)
  - CBETA 外字データベース (台湾の中華電子佛典協會 (CBETA) による；約13000字収録)
- 新規入力 (このために部品および IDC 入力用の四角號碼 quail を開発 [Wittern 氏による])
- 現在、UCS 収録の約7万字に対して一応入力完了
- cvs.m17n.org で公開中
- GPL で利用可能

複雑な文字知識の 表現

- 文字属性記 述に 関 す る ガイ ドラ イ ン
- TopicMaps の 利用

文字属性記述に関するガイドライン

- 属性の命名規則や値の型を定める
  - 写像: `=>foo`, `<=foo`, `=foo`
  - 関係: `->foo`, `<-foo`
  - 文字指定
  - 文字参照

  など
- 文字属性の整理・分類

TopicMaps の 利用 (Wittern 氏担当)

TopicMaps

- *topic* 定義 や *topic* 間 の 関 係に よ っ て 情報の 構造を 表現す る 方法
  - SGML 版: ISO/IEC 13250:2000
  - XML 版 (XTM (XML TopicMaps)) も 出て い る

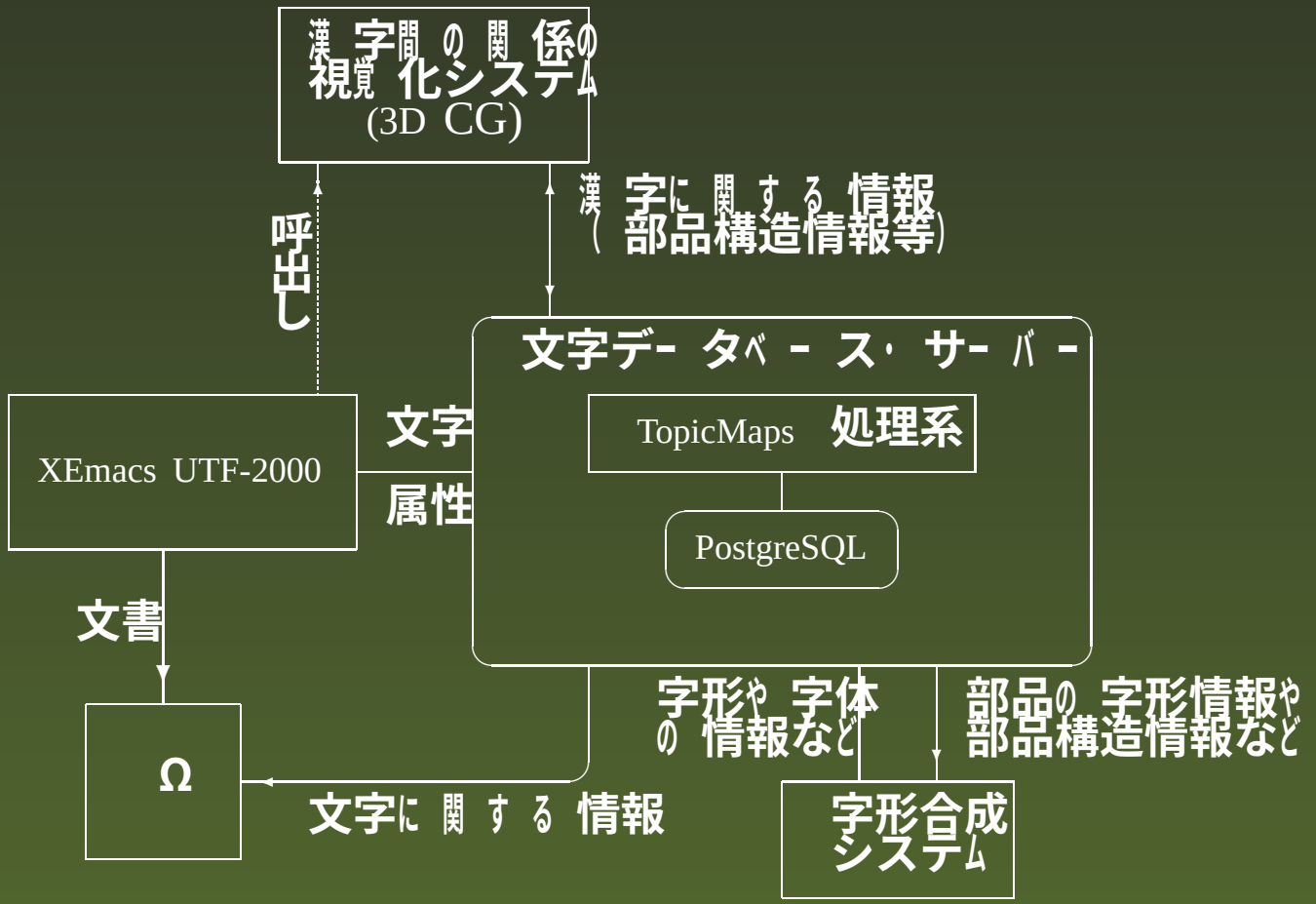
    → XTM を 用い る

文字や 属性間 の 関 係を TopicMaps で 表現した い

- そ の た め に TopicMaps 処理系を 開発す る
  - Zope を 用い た プ ロ トタイ プ
  - PostgreSQL を 用い た 実装 (現在開発中)
  - AxKit を 用い た 実装 (現在開発中)

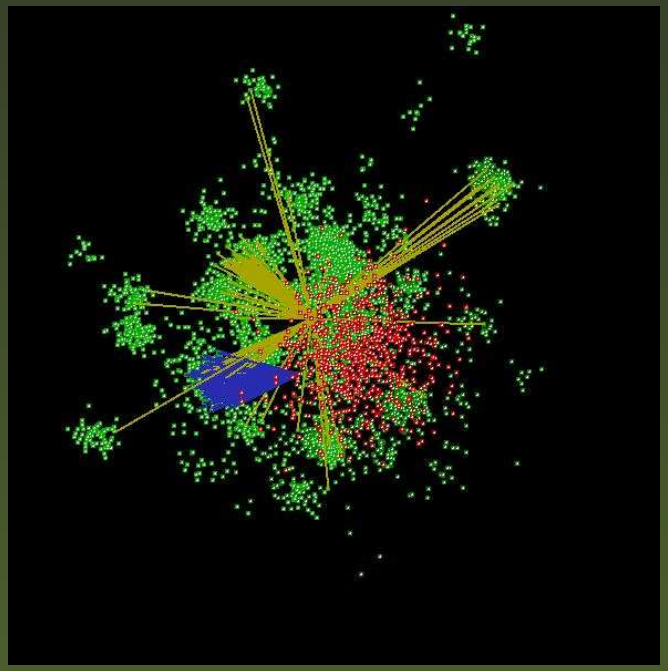
今後の展開

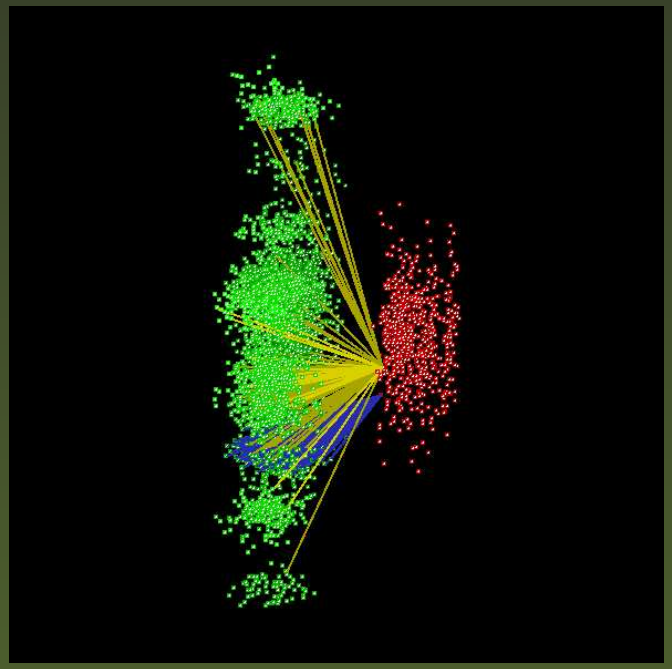
文字データベースに基づく総合的な文書処理環境の実現



文字知識情報の 解析と 可視化

文字間の関係を解析し、視覚化することで、文字知識情報を多角的に把握・編集するためのツールを実現する（藤原氏、鈴木氏、江渡氏らによる）





- 学習、検索、包摂規準の制御等

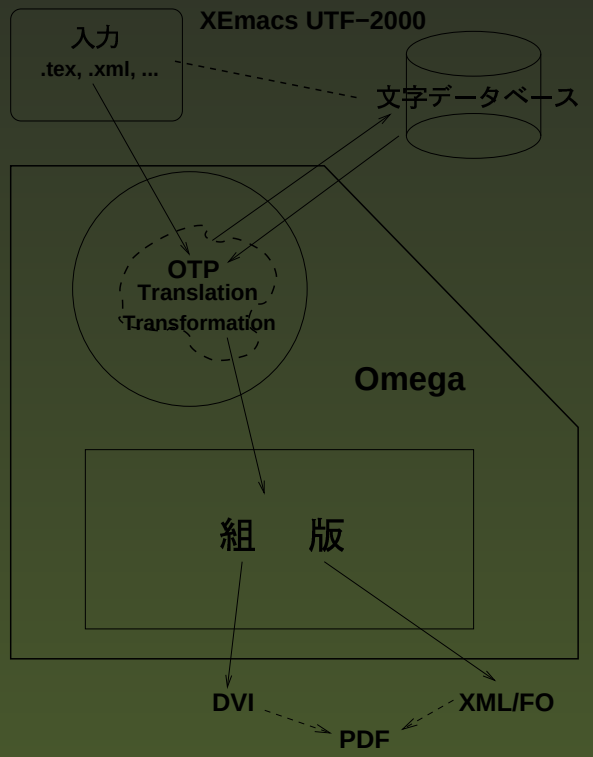
字形合成・管理システム

- 漢字構造情報を用いた字形合成（上地氏、師氏らによる）
  - 漢字構造情報と部品字形から複合漢字の字形を合成
  - 入力に使う情報や合成結果を文字データベース・サーバーで管理
    → グリフ・字形情報を統合した文字データベースへ
  - （バッチ処理を目的とした）単独の実装
  - （Ωとの連動を目的とした）ΩTPを利用した実装

を開発する

Ω（苫米地氏、宮崎氏らによる）

文字定義に基づいた組版の実現



文字層とテキスト層の連携の実現

- TEI P5 の WSD（書記系定義）実装の実現を目指す

期待される効果

- 定義した文字がちゃんと印刷・交換できる→符号化文字技術からの本質的な脱却
- 文字データベース・サーバーへの文字知識集約化の実証→計算機環境全体のCHISE化のための一里塚
- 漢字構造情報の応用
  - 漢字構造情報データベースのリファイン
  - FQDNのような分散型漢字記述の実現へ

CHISE Project に 関 す る 情報

- WWW
  - http://cvs.m17n.org/chise/
  - http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/
  - http://mousai.as.wakwak.ne.jp/projects/chise/
- 和英 2 言語に よ る mailing lists (参加方法は 上述 の 頁を 参照して 下さ い)
- 各種成果物の 最新版が anonymous CVS で 入手 可能
- 参加者募集中!!!