Chaon 実装の歩み：XEmacs CHISE から libchise へ

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

Chaon 実装開発の歩み

1. UTF-2000 以前
2. XEmacs UTF-2000 の登場
3. CHISE: 環境への展開

背景 (1) 1990 年台前半頃の状況

計算機の処理能力の向上によって多字・多言語処理が現実化してきた

1990 年	JIS X 0208:1990, JIS X 0212 制定
1990 年頃?	CDP データベースの開発開始
1991 年	Mule の開発開始
1993 年	Mule 1.0
	ISO/IEC 10646-1:1993 制定
	Windows NT 3.1
1994 年	Windows NT 3.1 日本語版
1995 年頃	IRIZ 漢字ベース

→ 大規模コード、統一コード待望論

背景(2) Code Wars

1984 年	ISO 10646 の開発開始
1987 年?	ISO DP 10646
1990 年	JIS X 0208:1990, JIS X 0212 制定
	ISO 1stDIS 10646 vs. Unicode 1.0
1991 年	（芝野の野望!?）
	ISO 1stDIS 10646 否決
	CJK-JRG 発足、統合漢字 v1.0
1992 年	ISO 2ndDIS 10646
1993 年	ISO/IEC 10646-1:1993 制定
1996 年	ハングルの大移動、UTF-16, Unicode 2.0
1997 年	JIS X 0208:1997

→ 文字コードをめぐる宗教対立!?の激化

背景(3) 漢字符号化の困難さの認識

代替アーキテクチャの希求

- 部品による漢字合成
  - 中央研究院の CDP (1990 年頃)
  - fj.kanji での盛り上がり、漢字 ML 設立 (1994 年)

  → グリフ合成の問題、文字表現力の問題
- 文字符号に依らない文字表現
  - SGML の実体参照の利用

  → CES 依存性は解決できても、CCS 依存性はなかなか解決できない

→ もっと抜本的な解決はないもんかいな

ラディカル・アプローチ

汎用文字コードを基礎にコンピューター・システムを考えるのではなく…

- 汎用じゃない文字コードを考えてみる（全部外字）
- すごく長い bit 長で文字を表しても良いかも
- *My Symbolic System*
  - シンボルを文字列ではなく画像で表現した Lisp 処理系を考えてみる
    - シンボルはアイコンで画面上に表現されている
    - 白紙窓に絵を書いて OK ボタンを押すと、その絵を表現とするシンボルが生成される
    - もし文字が必要なら、この仕組の中で定義する

UTF-2000

“My Symbolic System” のアイディアの内、とりあえず「文字を定義する」という部分を現実化してみよう

- 文字をオブジェクトとして扱う
  - 文字の持つ性質（文字素性）の集合で文字を表現する
- 文字列はオブジェクトの列にする
  - 整数に対する算術的変換 (ex. UTF-8) を使ってオブジェクト ID を符号化してマルチ・バイト列を作ることができる

UTF-2000 の現実化

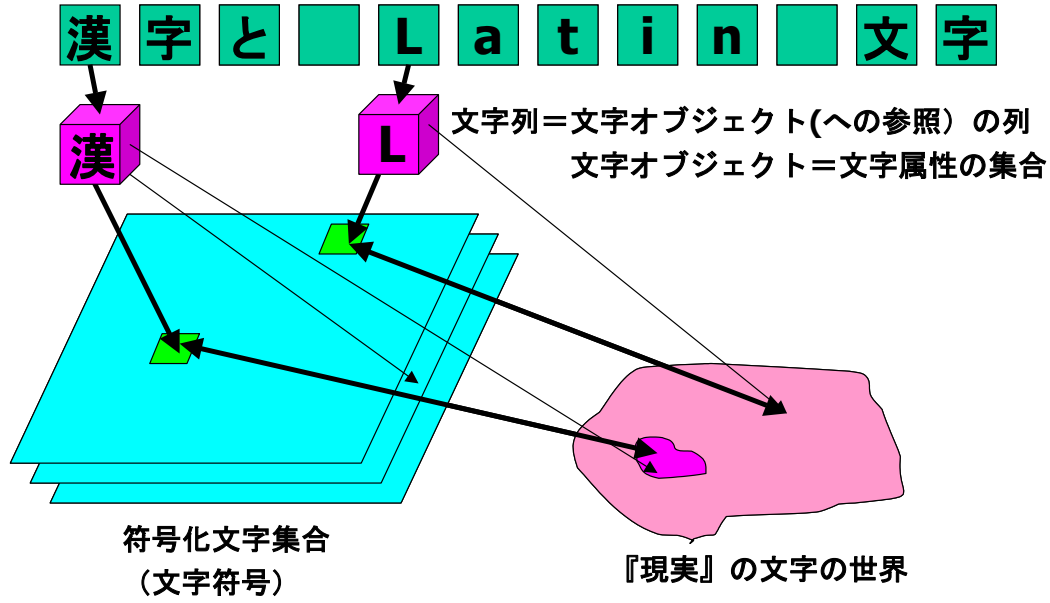
1998 年 4 月	utf-2000 mailing list 開設
1998 年 4 月	UTF-2000 based on GNU Emacs 20.2.90 (g 新部氏による)
1999 年 5 月	XEmacs UTF-2000 の実装作業開始
1999 年 6 月	XEmacs 21.2.16 (Sumida)
	UTF-2000 Version 0.2 (JR 難波)

2003 年 2 月に “XEmacs UTF-2000” は
“XEmacs CHISE” と改称した

Chaon モデル

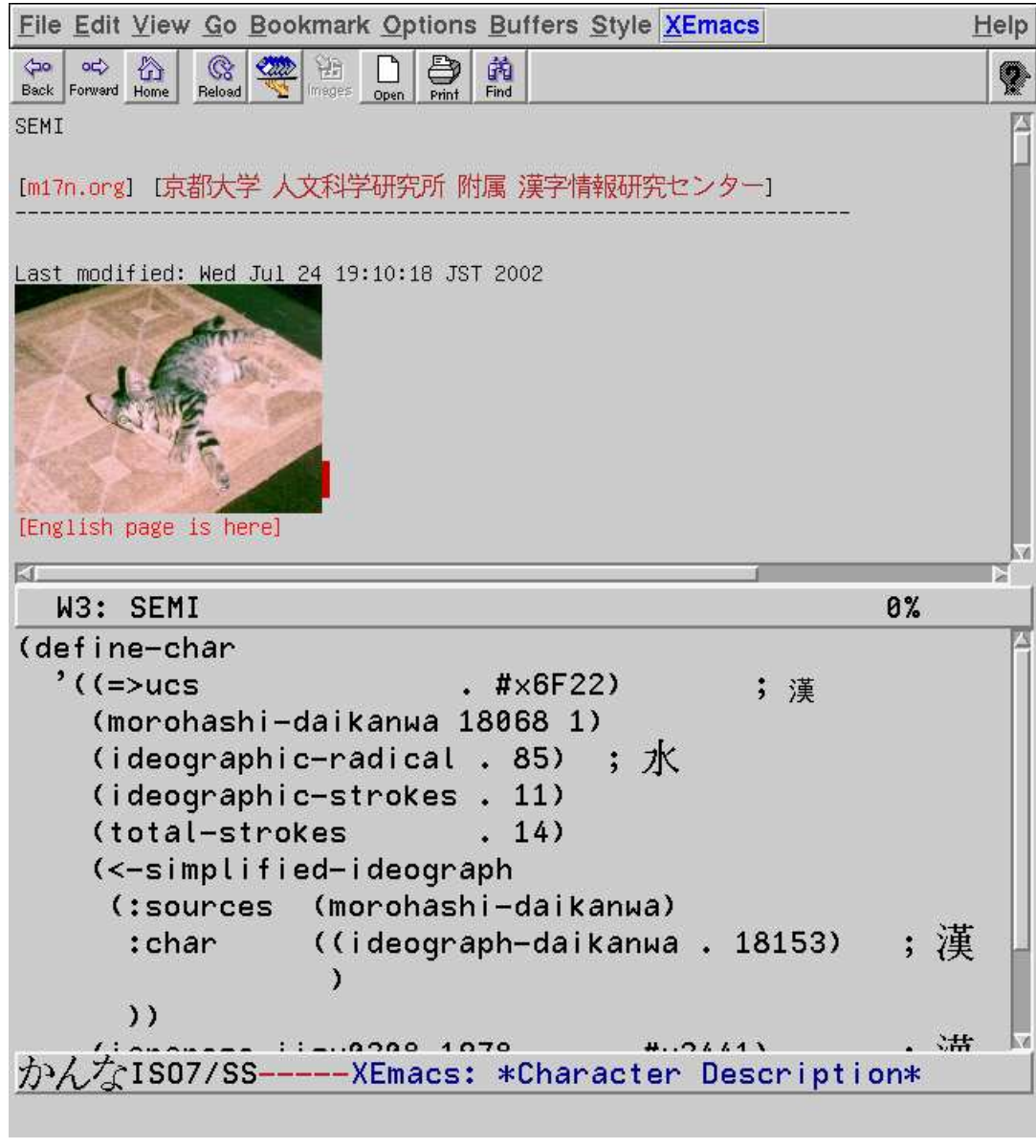
UTF-2000 のエッセンスを取り出したもの

- 文字を文字素性の集合（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 CHISE

- XEmacs-Mule に基づく多言語文書編集系（実用的側面）
- 文字符号に依存しない文字処理技術の試験実装およびテストベッド（実験的側面）



XEmacs CHISE の特徴

- XEmacs に基づく Chaon モデルの試験実装
  - 文字素性データベース機能
  - 豊富な文字空間（最大約 10 億文字定義可能）
- XEmacs-Mule に対する上位互換性
  - 既存の Emacs Lisp package が利用可能
  - Mule-charset の拡張 (coded-charset)
  - coding-system の拡張

文字素性データベース機能

- 文字定義
- 文字素性の設定
- 文字素性の参照
- 文字素性による文字の検索

Mule-charset の拡張 (coded-charset)

- 文字オブジェクトと code point の写像・逆写像
- 内部表現に対する依存性の解消
- 4 byte 以内の任意の文字符号が利用可能
- 文字素性の一種として扱われる
- 別名
- 継承

coding-system の拡張

- UTF-8 など Unicode 系 CES
- CCS 変換に用いる coded-charset を指定可能
- 実体参照変換
- 文字結合（途中）

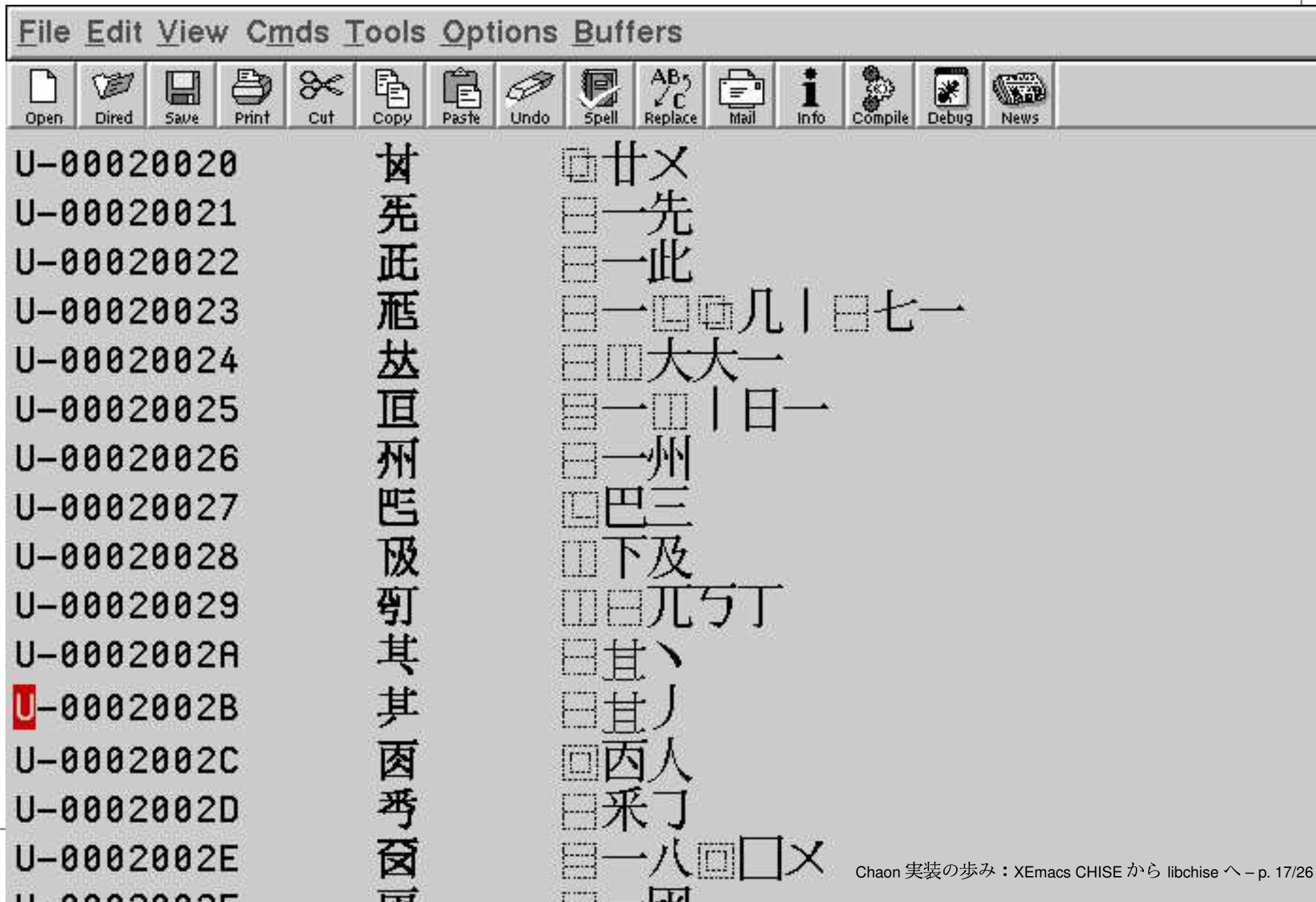
XEmacs CHISE の問題点・課題

- データベースの整備が必要
- 主記憶を浪費する
- XEmacs に閉じてしまっている

データベースの整備

- XEmacs CHISE 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

漢字構造情報のデータベース化



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:/cvs/chise で公開中 (ids module)
- GPL で利用可能

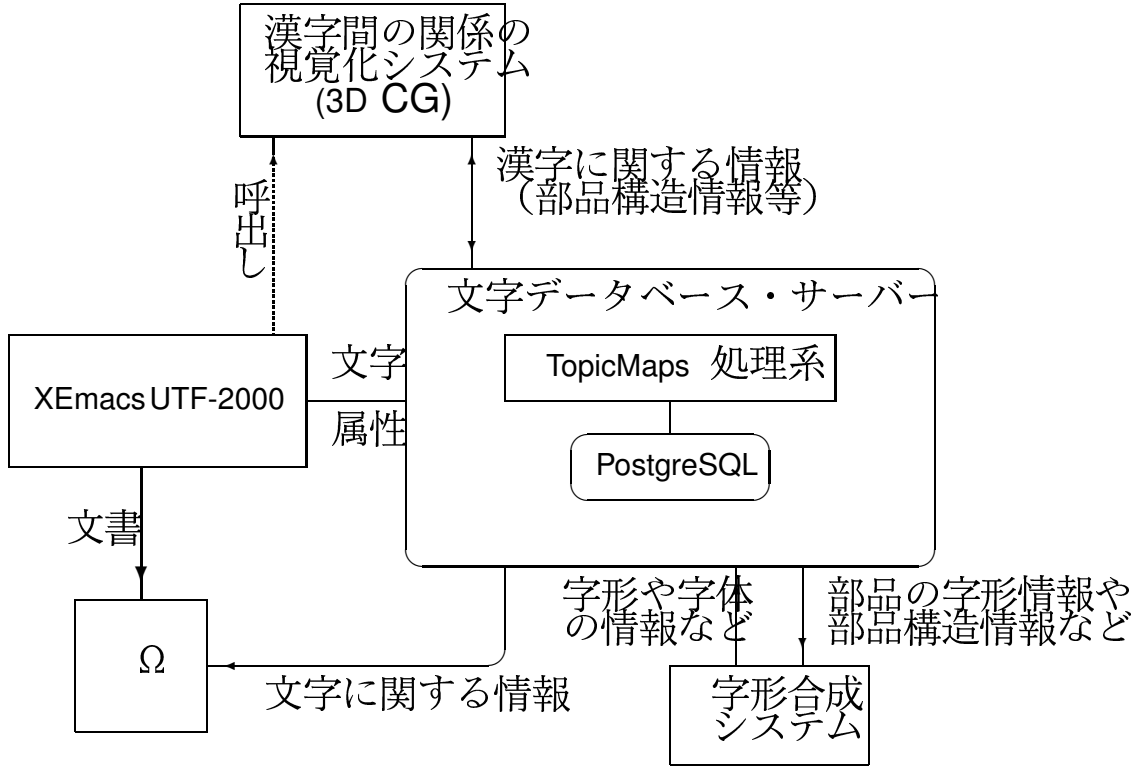
主記憶の節約

- 内部表現の効率化
- 外部文字データベース機能の実現
  - XEmacs CHISE で定義した文字素性データを XEmacs の外部のデータベースに保存
  - 必要な時に必要な分だけデータを取って来る（lazy-loading 機能）

  Berkeley DB を使用

XEmacs の外部への展開

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



CHISE プロジェクト構想

文字データベースの共有化

- PostgreSQL ベースの文字データベース・サーバー実現の遅れ
  - TopicMaps の実装が難しかった
- 当面、現状の Berkeley DB ベースの文字データベースを共有することになった
  - Ruby/CHISE, Perl/CHISE の登場

↓

- Berkeley DB の方が良いかも
- PostgreSQL 化よりも共通部分をライブラリ化を優先すべし→ libchise 構想

libchise の目標

CHISE 環境で共有される機能のライブラリ化

- XEmacs CHISE の機能のライブラリ化
  - 脱 Lisp 化
  - スケーラビリティ

↓

- 階層モデルの定義
- データ形式の定義

CHISE 階層モデル

第 4 層　オブジェクト層
　文字に対する抽象的なサービスの提供
　（関数、message 送信式、TopicMaps 等）
第 3 層　構造データ層
　複雑なデータ構造の実現（S 式、XML 等）
　（記憶管理や型システムの実現）
第 2 層　文字層
　素性名による文字データへのアクセスの実現
　（C の文字列レベルに限定、記憶管理無し）
第 1 層　データソース層
　（Berkeley DB や PostgreSQL などのデータベース）

データ形式

現状 文字素性値を構造化している

- 複数候補：集合を表現するリスト
- メタデータ：属性リスト

今後 なるべく文字素性値に
複雑なデータ構造を使わないようにする
→ 文字素性名の構造化

- 複数候補：素性名に選択指示子を付ける
- メタデータ：素性名にメタデータ指示子を付ける

libchise の現状

- 第 2 層の実装
  - 基本的なデータ参照機能の実装は完了
  - 書き込み機能、map 機能は近日中に実現予定
- libchise の利用
  - XEmacs CHISE の libchise 化（作業中）
  - libchise 版 Ruby/CHISE

今後の予定

- XEmacs CHISE 機能の libchise 化、第 3,4 層実装
- Kage 機能の libchise 化
- 文字素性データベースの再構築
- 文字素性データベースの libchise への移転
- 漢字の発音や意味に関する情報のデータベース化
- テキスト・データベースとの連携
- などなど