Character Database 2.0

守岡 知彦

2005年10月13日

CHISE Project

Character Database 2.0 とは

Character Database 2.0 とは

ついカッとなって付けてしまった。

2.0 ならなんでも良かった。[1]

[1] http://d.hatena.ne.jp/yomoyomo/20050915/versiontwo

Character Database 2.0 とは

ついカッとなって付けてしまった。

2.0 ならなんでも良かった。

というのはさておき…

Character Database 2.0 とは

これからの CHISE 文字データベースが目指す地点

Character Database 2.0 を考える前に

Character Database 1.0

＝ 従来の文字データベース

について考えてみよう

従来の文字データベース

- マッピング・テーブル

従来の文字データベース

- マッピング・テーブル
- 異体字シソーラス

従来の文字データベース

- マッピング・テーブル
- 異体字シソーラス
- 文字属性データベース
  e.g. 例：Unicode Database

従来の文字データベース

- 文字コードをキーにしたもの
  - マッピング・テーブル
  - 異体字シソーラス
  - 文字属性データベース
    e.g. 例：Unicode Database

従来の文字データベース

- 文字コードをキーにしたもの
  - マッピング・テーブル
  - 異体字シソーラス
  - 文字属性データベース
    e.g. 例：Unicode Database
- Chaon モデルに基づくもの
  - CHISE 文字データベース

従来の文字データベース

- 電子化されたもの
  - 文字コードをキーにしたもの
    - マッピング・テーブル
    - 異体字シソーラス
    - 文字属性データベース
      e.g. 例：Unicode Database
  - Chaon モデルに基づくもの
    - CHISE 文字データベース

従来の文字データベース

- 電子化されたもの
  - 文字コードをキーにしたもの
  - Chaon モデルに基づくもの

従来の文字データベース

- 電子化されたもの
  - 文字コードをキーにしたもの
  - Chaon モデルに基づくもの
- 電子化以前のもの
  - 字書、韻書 等

従来の文字データベース

- 電子化されたもの
- 電子化以前のもの
  - 字書、韻書 等

なんらかの機能のためにまとめられている

従来の文字データベース

- 電子化されたもの
- 電子化以前のもの
  - 字書、韻書 等の工具書等
- なんらかの機能のためにまとめられている
- 特に漢字の場合、電子化された文字データベースは、今の所、紙ベースの辞書に負けている

従来の文字データベース

- なんらかの機能のためにまとめられている
- 電子化された文字データベースは、今の所、紙ベースの辞書に負けている

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

漢字辞典の良い所

- 形・音・義の３要素が揃っている
- 字源（歴史的変遷）に関する情報が載っている
- 用例が載っている

漢字辞典の良い所

- 形・音・義の３要素が揃っている
  - 形：字形だけでなくその異同情報も
- 字源（歴史的変遷）に関する情報が載っている
- 異説が載ってる場合もある

漢字辞典の良い所

- 形・音・義の３要素が揃っている
  - 音
    - 日本語音
      - 音（漢音、呉音、唐音、通用音等）
      - 訓
    - 中国語音（ピンイン、注音等）
    - 歴史的表記（反切等）
    - などなど

漢字辞典の良い所

- 形・音・義の３要素が揃っている
  - 義：
    - 意味記述
    - 異体字・類字関係の情報：種別（本字、古字、略字、俗字、誤字、別字などなど）が出典情報込みで載っている
    - 用例が載っている
    - 字源（歴史的変遷）に関する情報が載っている
    - 異説が載ってる場合もある

漢字辞典を使う

例：版本に載っている謎の語彙を調べる

- ちょっとかすれてたりする場合も
- 時代、地域毎の異体字の問題
- 音がヒントになる場合も
- 似た用例がヒントになる場合も

漢字辞典を使う

例：版本に載っている謎の語彙を調べる

- ちょっとかすれてたりする場合も
- 時代、地域毎の異体字の問題
- 音がヒントになる場合も
- 似た用例がヒントになる場合も

形・音・義（+ その他のメタデータ等）を総合して扱わないと…

従来の文字データベースの問題点

- なんらかの機能のためにまとめられている
- 電子化された文字データベースは、今の所、紙ベースの辞書に負けている

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

従来の文字データベースの問題点

- なんらかの機能のためにまとめられている
- 電子化された文字データベースは、今の所、紙ベースの辞書に負けている

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

- 従来の文字処理が実現しようとした目標が貧しい

従来の文字処理の志の問題?

- なんらかの機能のためにまとめられている
- 電子化された文字データベースは、今の所、紙ベースの辞書に負けている

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

- 従来の文字処理が実現しようとした目標が貧しい

従来の文字処理の問題点

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

- 従来の文字処理が実現しようとした目標が貧しい
  - レイヤ化の問題

従来の文字処理の問題点

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

- 従来の文字処理が実現しようとした目標が貧しい
  - レイヤ化の問題
    - 例えば、自然言語処理と文字処理の縄張争い

従来の文字処理の問題点

→ コンピューターでできる文字処理は、紙ベースの辞書をひきながら人間が行う文字処理に機能的にだいぶ負けてる

- 従来の文字処理が実現しようとした目標が貧しい
  - レイヤ化の問題
    - 例えば、自然言語処理と文字処理の縄張争いというか、やっかいなものの押しつけ合い

従来の文字処理の問題点

- レイヤ化の問題
  - 「文字」という概念を適切に定義すると、

従来の文字処理の問題点

- レイヤ化の問題
  - 「文字」という概念を適切に定義すると、
    文字層で扱いたくない問題をグリフ層だとかマークアップ層だとか、自然言語処理層に押しつけることができる

従来の文字処理の問題点を解決するには

- レイヤ化の問題
  - 「文字」という概念を適切に定義すると、
    文字層で扱いたくない問題をグリフ層だとかマークアップ層だとか、
    自然言語処理層に押しつけることができる

→レイヤとレイヤの狭間を埋める

- 文字の周縁のサポート
- 他のレイヤ・モジュールとの連係

従来の文字処理の問題点を解決するには

→レイヤとレイヤの狭間を埋める

- 文字の周縁のサポート
- 他のレイヤ・モジュールとの連係

従来の文字処理の問題点を解決するには

→レイヤとレイヤの狭間を埋める

- 文字の周縁のサポート
- 他のレイヤ・モジュールとの連係

→ CHISE の目指す場所

CHISE の目標

→レイヤとレイヤの狭間を埋める

- 文字の周縁のサポート
- 他のレイヤ・モジュールとの連係

→ CHISE の目指す場所

- 多面的な文字データベースの実現
- 文字データベースの情報が文字処理やその他の処理に直接的に反映される環境の実現

CHISE の概要

- 多面的な文字データベースの実現
- 文字データベースの情報が文字処理やその他の処理に直接的に反映される環境の実現

CHISE の概要

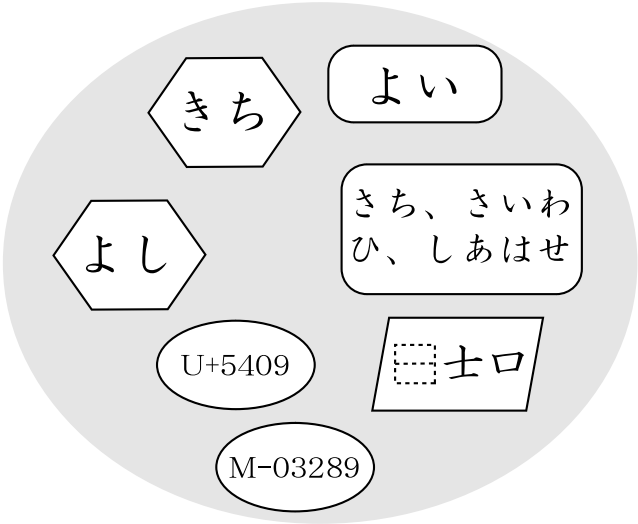
- 多面的な文字データベースの実現
  - CHISE 文字データベース
  - CHISE 漢字構造情報データベース (CHISE IDS)
- 文字データベースの情報が文字処理やその他の処理に直接的に反映される環境の実現

CHISE の概要

- 多面的な文字データベースの実現
- 文字データベースの情報が文字処理やその他の処理に直接的に反映される環境の実現
  - CHISE 文字データベースの環境内での共有
  - libchise を介した利用
  - これらを利用する各種アプリケーション

CHISE 文字データベース

Chaon モデルに基づく



文字を素性の集合で表す

CHISE 文字データベース

Chaon モデルに基づく

文字を素性の集合で表す

→ 当初は文字の表現モデルとして出発

CHISE 文字データベース

Chaon モデルに基づく

文字を素性の集合で表す

→ 当初は文字の表現モデルとして出発

→ 文字間のネットワークへ

CHISE 文字データベース

Chaon モデルに基づく

文字を素性の集合で表す

→ 当初は文字の表現モデルとして出発

→ 文字間のネットワークへ

文字間のネットワーク

紙の辞書は持っていた

但し、規範的

でも、一応、異説は書いてあった

出典情報を書けば、一応、複数の見解は共存可能

文字間のネットワークの複合体へ

出典情報を書けば、一応、複数の見解は共存可能

→ どれか中心を決めるのは嫌

- 学説や視点、用途によってどれを採用するかは違う
- 処理システムを複雑にしたくない

文字間のネットワークの複合体へ

出典情報を書けば、一応、複数の見解は共存可能

→ どれか中心を決めるのは嫌

→ ある断面がネットワーク（有向グラフ）となるネットワークの複合体にする

→ file system の mount みたいな処理を導入し、application 毎に固有のネットワークを実現できるようにしたい