# 漢籍データベースにおける技術的課題

安岡孝一

## ★目標

漢籍データベースを WWW から高速に検索できるようにする

- 検索エンジンは OpenText を使用
- WWW ブラウザは Internet Explorer を仮定
- 回線速度は 64kbps 程度でも OK とする

## ★漢字コードをどうするか

- 検索エンジンで使える漢字コード

  Shift_JIS . . JIS X 0208 [漢字数 6355 字]
  EUC-JP . . . JIS X 0208 + JIS X 0212 [漢字数 6355+5801=12156 字]
  UTF-8 . . . . ISO 10646 [漢字数 20902+6582+42711=70195 字]

  ex)「鷗堂賸稾一巻」は、Shift_JIS では表現不可 (EUC-JP や UTF-8 では表現可)

- WWW ブラウザで使える漢字コード

  Shift_JIS . . JIS X 0208 [漢字数 6355 字]
  EUC-JP . . . JIS X 0208 のみ表示可能 (※ JIS X 0212 は使用不可)
  UTF-8 . . . . JIS X 0208 のみ表示可能 (※ Windows'98 以降は JIS X 0212 が追加)

  ex)「鷗堂賸稾一巻」は、いずれにおいても表現不可

∴検索エンジンでは UTF-8 が、WWW ブラウザとの入出力は Shift_JIS が、現実的

## ★ UTF-8 にあって Shift_JIS にない漢字をどうやって検索するか

- 「〓」によるワイルドカード処理
- 異体字による検索. . . UTF-8 と Shift_JIS の異体字対応表が必要

ex)「鷗堂賸稾一巻」は「鴎堂〓稿一巻」で検索可能に

## ★ UTF-8 にあって Shift_JIS にない漢字をどうやって表示するか

ex)「鷗堂賸稾一巻」の「鷗」と「賸」

- 画像で送る
  画像数が増えると回線速度が追いつかない
- 専用のフォントをインストールしておいてもらう
  必要な漢字全てが入ったフォントを事前に作らねばならず、非現実的
- フォント埋め込みを用いる
  表示画面ごとに最小のフォントを動的に生成できれば、速度・手間ともに最良