インターネット応用特論
５．文字通信：
文字コードと国際化（２）

太田昌孝

mohta@necom830.hpcl.titech.ac.jp

ftp://ftp.hpcl.titech.ac.jp/appli5.ppt

文字とは？

- 言語を図形で表現するための要素
  - 表音文字
  - 表意文字

スクリプト（用字系）とは？

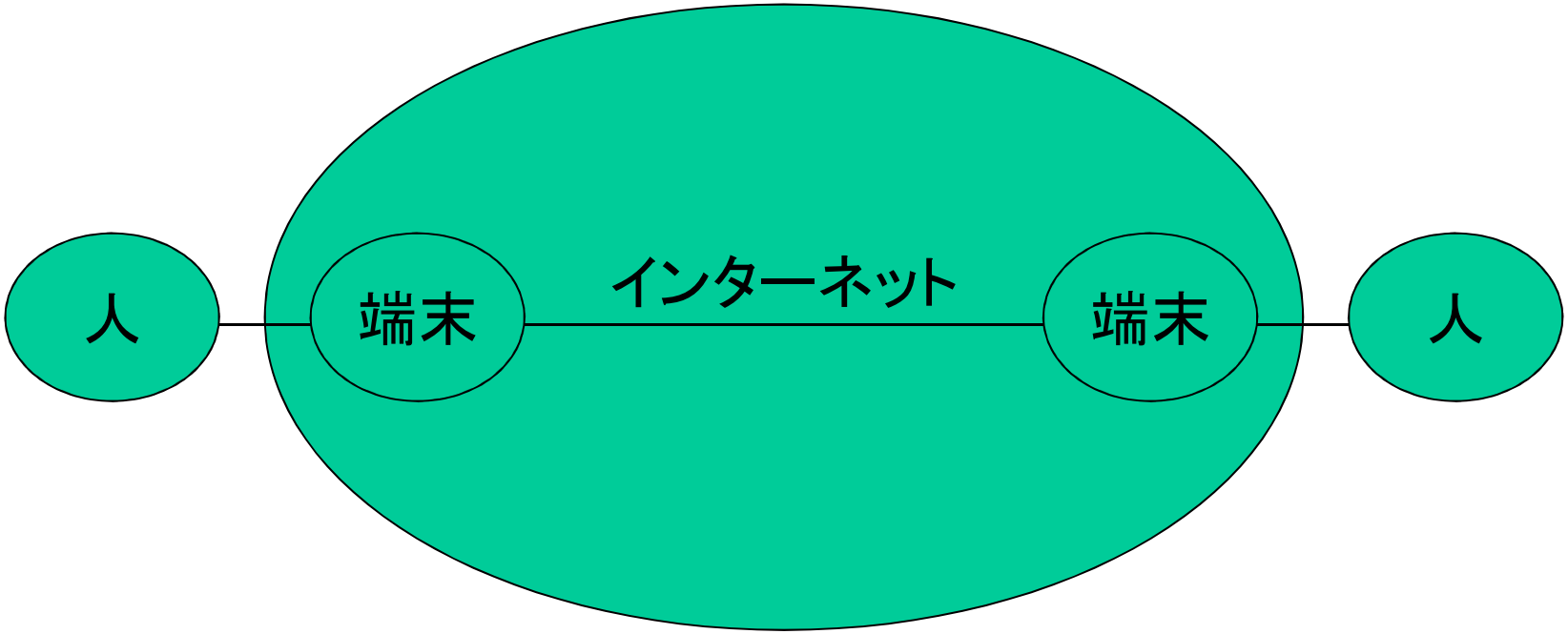
- 言語を文字で表記する体系
- 言語と文字を混同してはいけない
  - 「でぃすいずあぺん」はかなスクリプトによる英語
  - “Koreha pen desu.”はローマ字（ラテン）スクリプトによる日本語

日本語の表記に用いるスクリプト

- かな（ひらがな、カタカナ、万葉仮名）
- 漢字かな混じり
- ローマ字（ヘボン、訓令、他）
  - 「まさたか」をフランス語風に書くと“massata ka”
- その他、各地の文字による表音表記

インターネットと国際化

- インターネットにより
  - 世界中の端末が接続される
- 全ての端末は国際化されるべき？
  - たぶんそのとおり
- インターネットにより
  - 世界中の人が接続される
- 全ての人は国際化されるべき？
  - ある意味ではそうだが、、、



エンドツーエンド原理は端末では終わらない

国際化

- 国際化された人
  - 英語が使える人？
  - 主要六ヶ国語が使える人？
  - 全世界の全ての言語が使える人？
  - 個々の分野で国際的に使われている言語が使える人！
    - 日本文学を論じるのは国際的に日本語
      - 漫画やアニメも？
    - インターネットを論じるのは国際的に英語

端末の国際化（１）

- コンピュータ言語や文字ベースのプロトコル（FTP等）の国際化
  - 人工言語（文字数は限定的でかまわない）
  - 自然言語とは無関係
- 米国のＫ＆Ｒ　Ｃと国際化ISO　Ｃ
  - 一部の文字（ISO　646の国別部分（‘￥’など）を利用しなくてもよくなった
  - ASCIIが国際的に流通しているため無意味

端末の局所化（１）

- 局所化コンピュータ言語はあってもいいかもしれないが、誰も使わない
  - カナCOBOL
  - 世界的に流通するプログラムにならない
- 局所化プロトコルは無意味
  - インターネットは全世界の端末をつなげる

端末の局所化（2）

- 局所的に使われている文字を人間に出力
  - 局所文字コード
  - ビットマップディスプレイ＋局所文字フォント
    - 昔はキャラクタジェネレータROMを多用
- 局所的に使われている文字を人間が入力
  - 局所化キーボード？
  - 字書情報（漢字などの場合）

局所文字コード

- ISO　646の各国版
  - 12文字の自由度
    - 欧州でも少なすぎ
- 日本は世界に先駆けて
  - 2バイト漢字コードJIS　X　0208を制定

ASCIIの（英語のラテン文字表記の）単純さ（１）

- 文字数が少ない
- 横書きのみ
- 片方向（左から右）のみ
- リガチャー（文字の形の前後関係みよる変化）が不要
- 文字同定とその字形に共通認識がある

ASCIIの（英語のラテン文字表記の）単純さ（2）

- 大文字小文字の対応が明確で規則的
  - 飾り文字もない
- 文字幅が一定
- 広く普及していてどこでもつかえる

JIS X 0208の複雑さと単純さ（1）

- 文字数は多い
- 横書きと縦書き
  - あまり真面目に考えてない
- 片方向（左から右）のみ
  - 現代ではそう
- リガチャー（文字の形の前後関係による変化）が不要
  - 合成用丸はあるが、、、

JIS X 0208の複雑さと単純さ(2)

- 文字同定とその字形に共通認識がない
  - これは大問題
- カタカナひらがなの対応が明確で規則的
  - ヴには対応するひらがなはないが、、、
  - 飾り文字はない((半)濁点は合成済み)
- 文字幅が一定
- 広く普及していてどこでもつかえる
  - 現在の日本では

端末の国際化（２）

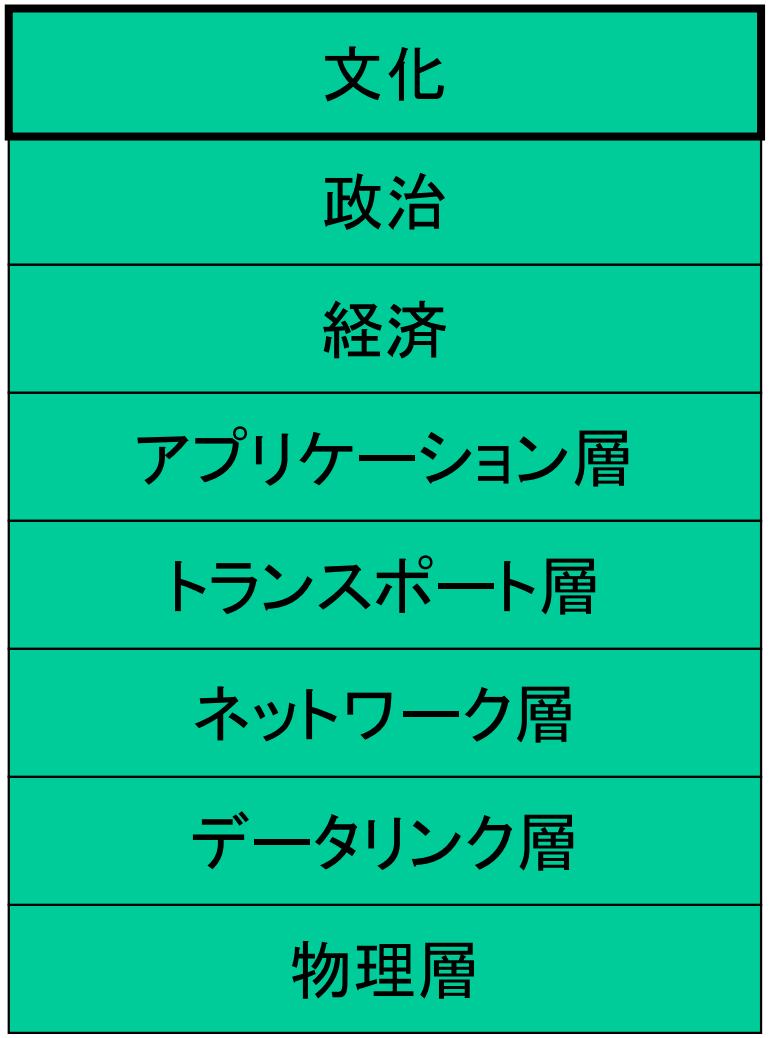
- 人間相手の文字出力をする端末では
  - 全世界の各地で局所的に使われている文字の出力の網羅的サポート
    - 全世界の各地で局所的に使われている文字を網羅した文字コード体系が必要
      - ISO2022

文字コードとは？

- ある文字集合に含まれる文字からなる文字列のデジタル化の規則
  - 個々の文字にコード（数値）を振ることだけではない
- 文字集合の含む文字の数が問題
  - 多いと多数のビットが必用
  - 少ないと多様な文字が表現できない
    - 同種の文字の細かな違いも表現できない

国際化文字コードとは？

- 極めて多数の文字を含んだ文字コード
- 文字コードへの要求は文化ごとに違う
  - 多用な要求をどう統一するか？
    - 現実的には各国ばらばらの文字コードをもつしかない
    - 最低限の枠組みは欲しい
      - ISO 2022
      - 有限状態



インターネットの上のレイヤリング構造

ISO 2022

- 各国の文字集合規格を切り替える規格
- G0、G1、G2、G3の4つの7ビットバイト文字集合（含多バイト文字）を同時に使える
  - G0はそのまま、G1は8ビット目をたてて
  - G2、G3はSS2文字、SS3文字に続けて
- Gn集合への文字集合の割り当てはエスケープシーケンスで切り替え
  - ASCIIのG0への割り当てはESC＋＄＋B

シフトコード

- ISO　2022のエスケープシーケンスによる切り替えは国際化には必須
  - 局所化ならもっと簡単（2～3文字集合）でいい
- 局所化された端末内では
  - 1～2バイトの可変長コードが好ましい
- シフトJISが発端、BIG5、GBKも
- 通信にも使われた

ISO 2022とEUC

- EUC（Extended Unix Code）
  - ISO2022でGnへの割り当てを固定したもの
    - 実際に使われるのはG0とG1のみ
- シフト文字コードと異なり国際標準準拠
- 日本での並存は混乱をもたらしただけ
- EUC－GBは、シフト化されたGBKへ拡張

文書の構造

- 文書にはさまざまな構造がある
  - 改行
  - フォント切り替え
  - アンダーライン
  - 縦横の切り替え
  - 左右の方向性
  - 章、節、小節、段落、、、
  - 字サゲ

プレーンテキストと構造化テキスト

- 構造がないのがプレーンテキスト？
  - 実際には改行程度はある
  - UNIXではアンダーラインくらいまである
- 構造化テキストはなんでもやれる
- どこに境界線を引くべきか？
  - 抽象的議論にはきりがない

文字コードを何につかう？

- 文字の入力
- 文字の出力
  - ここまでなら、コード化しない画像でも可
- 文字列の検索その他の処理
  - これこそ、文字コードを使う意義

文字コードと有限状態性

- 文字列が有限状態だと
  - すべてが決定的
  - 検索も楽
- 文字列が有限状態でないと
  - 処理に時間がかかる
    - 例えば、LBAの検索はO(L^3)
  - ある種の処理は、そもそも計算不能
- 文字コードとは有限状態であるべき！

プレーンテキストのもてる構造

- PDAが必要なネストは無理
  - 改行
  - フォント切り替え
  - アンダーライン（１重）
  - 縦横の切り替え
  - 左右の方向性（ネストなし）
  - 段落
  - 字サゲ（タブ程度）

文字コードの国際化の統一的枠組みでは

- 世界中の文字の入出力が可能なのは当然として、、、
- 有限状態性は死守しないといけない
  - ISO2022の拡張ではネストした双方向性もサポート
- プレーンテキストと構造化テキストの切り分けが重要

端末の国際化（３）

- 人間相手の文字入力をする端末では
  - 国際化文字コードはもちろん必要だが
  - 全世界の各地で局所的に使われている文字の入力の網羅的サポート？
    - 標準的な入力デバイス？
      - GUIは子供のおもちゃ
      - ASCIIキーボード
    - 文字入力方法は文字や人によって多種多様
      - ソフトウェアダウンロード？
        - » 機種によってバイナリは違う

インターネットと国際化

- インターネットにより
  - 世界中の端末が接続される
- 全ての端末は国際化されるべき？
  - たぶんそのとおり
- インターネットにより
  - 世界中の人が接続される
- 全ての人は国際化されるべき？
  - ある意味ではそうだが、、、

文字環境の国際化のために

- 文字コードのインターネット化が必要
  - 文字フォントをインターネットでアクセス可能に
  - 字書情報をインターネットでアクセス可能に
  - 入力プログラムをインターネットでアクセス可能に

まとめ

- インターネットの普及により国際化された文字コードが必要
- 国際化された文字コードはインターネット化が必要
- 国際化と局所化を取り違えてはいけない