Root Zone LGRおよび日本語生成パネル（JGP）について

- ご報告 & ご意見・コメント募集 -

2020年12月3日

日本語生成パネル チェア
堀田 博文 〈hotta@jprs.co.jp〉

背景

- ルートゾーンには様々な言語・scriptのラベルが存在
- いくつかの言語・scriptには同一視すべき(/した方がいい)文字同士が存在(「異体字」と呼ぶこととする)
  - 例1:字形は異なるが読み・意味が同じ文字(「国」と「國」と「圀」)
    例2:視覚的に同一(もしくは酷似)な文字(漢字「二」とカナ「ニ」?)

ラベルの適切さや同一視すべき文字をもつラベルをできるだけ自動的に判断できるように

- ルートゾーン用に、様々な言語・scriptのTLDラベルで使える文字の範囲および異体字を共通ルール(Root Zone Label Generation Rules; RZ-LGR)として決めておく
- scriptを共有する言語同士は統一ルールの作成において調整が必要
  - 例:CJK(中国語/日本語/韓国語)は漢字を共有

RZ-LGRとは

- 創設を申請されたIDN TLDラベルをルートゾーンの中でどのように取り扱うかを規定するルール
- 各言語やscriptに対する次の4つのルールからなる
  - 使用可能な文字の範囲(レパートリー)
    - たとえば、JIS第一水準と第二水準の文字?
  - 異体字の定義
    - たとえば、「国」と「國」と「圀」は異体字?
  - 異体字を含むラベルのうちどれをTLDとして使えるかの定義
    - たとえば、「国」と「國」が異体字であり「.中国」がTLDとして使用されている場合、「.中國」というTLDも使える?
  - TLD文字列全体に関するルールの定義
    - たとえば、中国語の簡体字と繁体字は1ラベル内では同時使用不可?

RZ-LGR作成の枠組み

- ICANN全体としてRZ-LGRの作成を推進
  - ICANN会合でRZ-LGRの解説と実装の呼びかけ
    - 2013年11月以降、各ICANN会合で情報共有会合開催
  - 英字以外を使用する国（地域）で作成活動開始
    - 2014年5月8日にJPNICオフィスでLGRのワークショップを開催
- IP : Integration Panel : 統合パネル
  - 各言語・スクリプトのLGR作成を支援し、全LGRを統合して1つのRZ-LGRを作成するチーム
  - 2013年10月設立
- GP : Generation Panel : 生成パネル
  - 各言語のコミュニティが設立する、LGRを作成するチーム
  - 日本語生成パネル(JGP)は、2014年より活動開始

JGPのドラフト提案
- 日本語LGRの概要 -

- 使用可能な文字の範囲(レパートリー)
  - JIS X 0208:2012の第一水準・第二水準の範囲（漢字、平仮名、片仮名、漢字および仮名に準ずる一部記号文字からなる約6,300文字）
- 異体字の定義
  - 日本語独自の異体字
    - 字形は異なるが読み・意味が同じ文字（「国」と「國」と「圀」））は異体字としない
    - 視覚的に同一（もしくは酷似）な文字（「二」と「ニ」））は最小限のものだけ異体字とする
  - 中国語/韓国語LGRで定義された異体字
    - それらを日本語LGRに取り入れる（日本語LGRでも異体字として扱う）
- ラベル文字列全体に関するルールの定義
  - 特にルールは定義しない

JGPのドラフト提案
- 視覚的同一/酷似文字の異体字定義 -

- フィールド実験を行い、視覚的同一/酷似文字を判断
  - 被験者数はランダムに選ばれた20人
  - 利用環境2種、フォント9種、文字サイズ3種を使用
  - Unicodeコンソーシアム公開の「錯視が起こりやすい文字対リスト」に記載されている文字対を対象
    - へ へ　・ ハ 八
    - ベ ベ　・ト ト
    - ペ ペ　・ ロ 口
    - ニ 二　・ タ タ
    - カ 力　・ エ エ
  - 上記10対は錯視が起こりやすいとの実験結果→異体字として定義
- 酷似する1ストロークの記号も異体字として定義
  - U+30FC(ー) U+4E00(一)　・ U+30FD(ヽ) U+4E36(丶)

JGPのドラフト提案
- 使用可能な異体字ラベル数の抑制 -

- TLDを申請し承認を得たTLD申請者は、
  - ルートゾーンの肥大化を防ぐため、申請ラベルの各文字を異体字に置き換えてできるラベル(「異体字ラベル」と呼ぶ)のうち、使用可能なのは、
    - 申請ラベル　及び
    - 申請ラベル内の文字(複数可)を常用漢字である異体字に置き換えたもの
  - 次が異体字同士とした場合、
    - 国　常用漢字
    - 國　常用漢字でない
    - 圀　常用漢字でない
    - 雲　常用漢字
    - 云　常用漢字でない
  - 「.圀云」を申請し承認を得たTLD申請者は、異体字の全組合せ(6つ)に対する権利を持つが、実際にTLDとして使用可能なのは2つとなる
    - .国雲　〇常用＋常用
    - .国云　×常用＋非常用
    - .國雲　×非常用＋常用
    - .國云　×常用＋非常用
    - .圀雲　×非常用＋常用
    - .圀云　〇申請文字列 (非常用＋非常用)

JGPのドラフト提案へのIPからのコメント

- 前述した内容のドラフト提案に対し、IPからコメントがあり、さらに検討中
- 主なコメント
  - 第一水準、第二水準に入っていない常用漢字(頬など4文字)をレパートリーに含めることの是非を検討すべき
  - 拗音(ャなど)、促音(ッなど)などの小字(捨て仮名)の特別ルールを検討すべき
    - 例1：捨て仮名のみからなる文字列はブロックする、
    - 例2：捨て仮名の前には通常の文字がなければならない、
  - 異体字として提案された10対以外にも、似た文字対 (才と⺣など) がセカンドレベルLGRでは定義されているが、それらを異体字に含めることの是非を検討すべき
  - 登録可能な異体文字列を常用漢字のみの組とした場合でもその組み合わせの数が膨大になる可能性があるため、さらなる削減策を検討すべき

ご質問?
ご意見?

提案内容) https://j-gp.jp/topics/20201015-01
意見受付) info@j-gp.jp