

RootLGR における JGP の具体的課題定義(案)

1 背景

1.1 歴史

国際化ドメイン名(Internationalized Domain Name、以降 IDN と記載)は、英語以外の言語を母語とする人々にとって利用しやすいアドレスを提供するものとして、1998 年にアジアを起点として検討が開始された。2000 年にインターネットプロトコルの標準化を行っている IETF(Internet Engineering Task Force)で WG が設立され、2003 年に標準が RFC 3490、3491、3492 として制定された¹。

IDN 標準化の際に、中国語圏から IDN の利用には異体字²問題の解決が必要だとの主張が行われた。中国語における異体字問題とは、中国本土で使われている簡体字と台湾や香港などで使われている繁体字は基本的に 1 対 1 に対応付くものであり、簡体字の地域と繁体字の地域間でのコミュニケーションのためには簡体字のアドレスと繁体字のアドレスは同一のものとしてアクセスできる必要がある、というものである。しかし、この異体字のあるべき姿は、言語ごとに異なり、たとえば、共通的に漢字を使用する中国語と日本語においても、異体字のとらえ方は異なっている。このような背景から、IETF における標準は世界中どこでも共通して使えるプロトコルである必要があるため、中国語の異体字問題は地域化(Localization)であるとして、IDN プロトコルの検討対象外とされた。

中国語の異体字問題は、漢字を使う言語文化圏の中国・日本・韓国・台湾の NIC(Network Information Center)が中心となって結成した JET(Joint Engineering Team)で検討が行われ、ドメイン名登録時に異体字を含む文字列のバリエーションを同一登録者に紐付けて、そのまとまり(パッケージ)を不可分とするという運用方式が合意された。この方式は JET Guidelines としてまとめられ、2004 年に RFC 3743 として発行された。また、この考え方は ICANN の IDN Implementation Guidelines にも取り入れられ、IDN 登録を行うレジストリは登録可能な文字のコードポイントとその異体字リストのテーブルを IANA に登録することが推奨されている。

1.2 JET Guidelines の目的

JET Guidelines は、異体字を持つ漢字が含まれるドメイン名の登録時に、登録申請された

¹ 2010 年に、RFC 5890、5891、5892、5893、5894 で更新された。これらの標準を区別するため、2003 年版を IDNA2003、2010 年版を(方式が決定した年から)IDNA2008 と呼んでいる。

² 一般には、同音同義の漢字であって、字体だけが異なり、どの文脈でも交換可能である漢字と定義される。本書では、ある言語において同音同義の漢字であって、Unicode のコードポイントは異なるが同一の文字として取り扱われるべき漢字と定義する。

ドメイン名(文字列)と異体字で置き換えた文字列(群)を一つのまとまり(パッケージ)として、同一登録申請者に紐付ける運用方式を提案するものである。JET Guidelines では、登録可能な文字とその文字の異体字を結び付ける IDN テーブルのフォーマットと、パッケージを生成するアルゴリズムを規定している。JET Guidelines で規定する IDN テーブルの構成を以下に示す。

➤ 構成要素

- ✓ その IDN テーブルが適用される言語
- ✓ その IDN テーブルに含まれる文字の出典のリスト(参照番号で区別)
- ✓ その IDN テーブルのバージョン番号
- ✓ 3 カラムからなる IDN テーブル本体

➤ IDN テーブル本体の各カラムの意味

第 1 カラム：登録可能文字(Unicode のコードポイント)と、出典の参照番号

第 2 カラム：ドメイン名での使用が推奨される異体字(群)と、その出典の参照番号

第 3 カラム：ドメイン名では使用されず予約される異体字(群)と、その出典の参照番号

各カラムはセミコロン (;) で区切られる。異体字が複数ある場合は、各異体字はカンマ (,) で区切られる。以下に、例を示す(JET Guidelines から引用)。

Language Variant Table for zh-tw # 言語は台湾の中国語

Reference 1 CP950 (commonly known as BIG5) # 出典の参照番号 1

Reference 2 zVariant, zTradVariant, zSimpVariant in Unihan.txt

Reference 3 List of Simplified Character Table (Traditional column)

Reference 4 zTradVariant in Unihan.txt

Version 1 20020701 # July 2002 # バージョン番号

以下、IDN テーブル本体

#登録可能文字;推奨される異体字;予約される異体字

5718(1);5718(4);56E2(2),56E3(2) # 團;團;团,团

60F3(1);60F3(1); # 想;想;

6559(1);6559(1);654E(2) # 教;教;教;

6E05(1);6E05(1);6DF8(2) # 清;清;清

771F(1);771F(1);771E(2) # 真;真;真

806F(1);806F(3);8054(2),8068(2) # 聯;聯;联,聯

96C6(1);96C6(1); # 集;集;

52

53 この例で、「清(U+6E05³) 真(U+771F) 教(U+6559)」という文字列が登録申請されたとする
54 と、登録者(ここでは登録者 A とする)に紐付けられる IDN のパッケージは以下となる。

55

登録文字列：清(U+6E05) 真(U+771F) 教(U+6559)

予約文字列：清(U+6E05) 眞(U+771E) 教(U+6559)

清(U+6E05) 眞(U+771E) 教(U+654E)

清(U+6E05) 真(U+771F) 教(U+654E)

清(U+6DF8) 真(U+771F) 教(U+6559)

清(U+6DF8) 眞(U+771E) 教(U+6559)

清(U+6DF8) 眞(U+771E) 教(U+654E)

清(U+6DF8) 真(U+771F) 教(U+654E)

56

57 登録者 A とは異なる別の登録者(ここでは登録者 B とする)が後から「清(U+6DF8) 眞

³ U+16 進数は、Unicode のコードポイントを示す。

(U+771E) 教(U+654E)」を登録申請しても、当該文字列中の文字はいずれも登録可能文字に含まれていないため、登録できない。もし、いずれかの文字が登録可能文字であったとしても、その文字列は既に登録されている登録者 A のパッケージに含まれているため、登録はできない。

1.3 各 TLD における異体字の定義

いくつかのレジストリは、そのレジストリが運用する TLD 配下(主には SLD:第二レベルドメイン名)で IDN を使用可能としている。.CN、.TW を中心とする中国語ドメイン名を登録可能としているレジストリでは、それらレジストリが協力して共通の IDN テーブルを定義し、そこから各レジストリが自レジストリで使う部分を切り出して JET Guidelines にしたがって定義し、それをルールとしている。

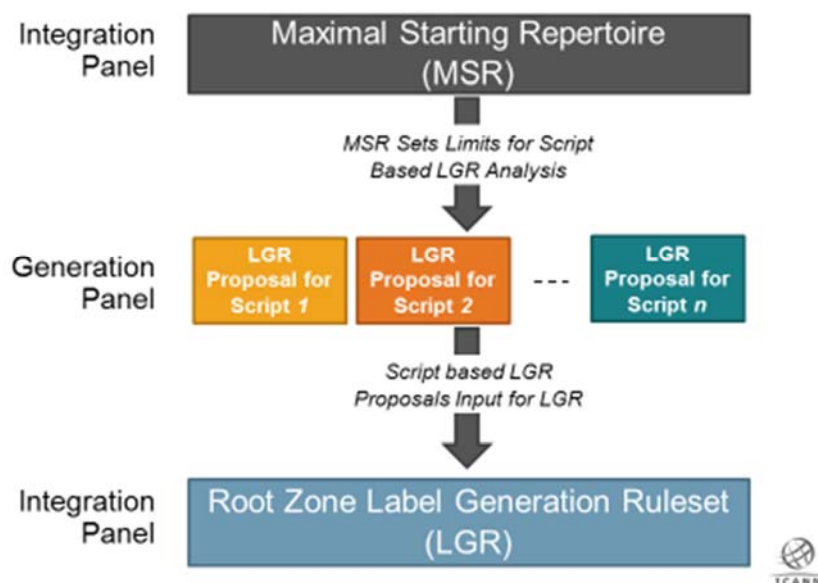
.JP では、2001 年に導入された汎用 JP ドメイン名において日本語 JP ドメイン名を使用可能とすることを JPNIC が決定し、専門家チームが日本語 JP ドメイン名のあり方を検討し、外部の方々意見も取り入れつつ、ルールを決定し、JPRS がそのルールに基づき運用している。そのルールの特徴は、使用可能文字を JIS 第 1 水準と第 2 水準とし、異体字は設けないというものである。異体字を設けない理由は、日本では異なる文字コードを持つ文字は異なる文字として扱うことが適切であるという判断をしたためである。

2 RootLGR と言語ルール

2.1 RootLGR とは

[RootLGR プロジェクトの背景、RootLGR の目的、RootLGR プロジェクトを構成する統合パネルと各言語生成パネルの関係、現在の状況などをここで説明する予定。]

LGR Development Process



2.2 言語ルール

RootLGR は Root ゾーン(TLD 文字列)用の統合ルールのことであり、各言語パネルが作成した言語ルールを統合パネルが統合して作成する。各言語パネルが作成する言語ルールは JET Guidelines の IDN テーブルとほぼ同等であり、以下の情報から構成される。

➤ 申請可能文字

MSR⁴の中から申請可能な文字の範囲を言語パネルが選択する。選択できる文字は、言語パネルがその言語で使用すると宣言した Script⁵の範囲に限られる。

➤ 異体字リスト

個々の申請可能文字について、その文字が持つ異体字を一覧(リスト)にしたもの。異体字リストに含まれる文字は、申請可能文字でなければならない。

➤ 各異体字の属性

割り当て可能(その文字を含む文字列を Root ゾーンに登録することが可能)もしくはブロック(その文字を含む文字列は Root ゾーンに登録することが不可能)の 2 種類であり、異体字リストに含まれる個々の文字について定義される。

⁴ 統合パネルが作成した、ルートゾーン(TLD 文字列)として使用可能な文字の最大集合を定義したもの。Maximal Starting Repertoire の略称。

⁵ Script は平仮名、片仮名、漢字などある言語を書き表すために使用される同種の文字(用字)のこと。

101 ➤ 文字列評価ルール
102 申請された文字列全体に対して適用されるルール。簡体字と繁体字の混在は許容しな
103 い(ブロックする)などの定義を行う。
104
105 「1.2 JET Guidelines」で例として示した IDN テーブルを言語ルールで記述⁶すると以下の
106 ようになる。

```
<language>und-Hani</language> <!-- 言語は中国語 -->
<data>
<char cp="6E05" tag="sc:Hani"> <!-- 申請可能文字は「清」 -->
  <var cp="6DF8" type="block" /> <var cp="6E05" type="both" />
  <!-- 異体字「清」はブロック、異体字「清」は簡体字と繁体字が同じ -->
</char>
<char cp="6DF8" tag="sc:Hani"> <!-- 清 -->
  <var cp="6DF8" type="block" /> <var cp="6E05" type="both" />
</char>
<char cp="771F" tag="sc:Hani"> <!-- 真 -->
  <var cp="771E" type="block" /> <var cp="771F" type="both" />
</char>
<char cp="771E" tag="sc:Hani"> <!-- 眞 -->
  <var cp="771E" type="block" /> <var cp="771F" type="both" />
</char>
<char cp="6559" tag="sc:Hani"> <!-- 教 -->
  <var cp="654E" type="block" /> <var cp="6559" type="both" />
</char>
<char cp="654E" tag="sc:Hani"> <!-- 敎 -->
  <var cp="654E" type="block" /> <var cp="6559" type="both" />
</char>
<rules>
  <action disp="block" any-variant="block" />
  <action disp="allocate" only-variants="simp both" />
  <action disp="allocate" only-variants="trad both" />
  <action disp="block" any-variant="simp trad" />
</rules>
</data>
```

⁶ IETF に提案されている Internet Draft "Representing Label Generation Rulesets using XML"(draft-davies-idntables)で形式が定義されている。

JET Guidelines の IDN テーブルと言語ルールの違いは以下の通りである。

- JET Guidelines では、予約される異体字は登録可能文字の一覧に現れなくともよいが、言語ルールでは異体字リストに含まれるすべての文字は申請可能文字でなければならない。
- JET Guidelines ではドメイン名登録を行うレジストリが独自の IDN テーブルを定義することとしているが、言語ルールは言語コミュニティ(当該言語を使用している国・地域などの言語文化圏)の合意に基づいて言語パネルが定義することとしている。
- JET Guidelines の IDN テーブルはそれ自身が単独で適用されるルールであるが、言語ルールは RootLGR を構成する要素であり単独で適用されることはない。

2.2 RootLGR

IDN TLD として申請された文字列に対して適用されるルールは、言語パネルが作成した各言語個別の言語ルールではなく、すべての言語ルールを統合パネルが統合した統合ルール(RootLGR)である。これは、TLD 文字列はさまざまな言語の文字列が混在するため、特定の言語ルールを適用できないためである。

RootLGR は、原則として各言語の言語ルールの和集合である。すなわち、異なる言語であっても同じ Script を共有するもの(中国語と日本語の漢字など)であれば、同じ文字(Unicode のコードポイント)に対して定義される異体字リストは、それぞれの言語で定義された異体字リストの和集合となる。

例えば、中国語ルールが「愛(U+611B)」と「爰(U+7231)」を異体字として

130

```
<language>und-Hani</language>
<data>
<char cp="611B" tag="sc:Hani"> <!-- 愛 -->
  <var cp="611B " type="trad" /> <var cp="7231" type="simp" />
</char>
<char cp="7231" tag="sc:Hani"> <!-- 愛 -->
  <var cp="611B " type="trad" /> <var cp="7231" type="simp" />
</char>
<rules>
  <action disp="block" any-variant="block" />
  <action disp="allocate" only-variants="simp both" />
  <action disp="allocate" only-variants="trad both" />
  <action disp="block" any-variant="simp trad" />
</rules>
</data>
```

131

132 と定義し、日本語ルールが「愛(U+611B)」を異体字なしの

133

```
<language>und-Jpan</language>
<data>
<char cp="611B" tag="sc:Hani"> <!-- 愛 -->
  <var cp="611B " type="allocate" />
</char>
<rules>
  <action disp="block" any-variant="block" />
  <action disp="allocate" only-variants="allocate" />
</rules>
</data>
```

134

135 と定義した場合、RootLGR は以下ようになる。

```

<data>
<char cp="611B" tag="sc:Hani"> <!-- 愛 -->
  <var cp="611B" type="simp" when="und·Hani" />
  <var cp="611B" type="allocate" when="und·Jpan" />
  <var cp="7231" type="trad" when="und·Hani" />
  <var cp="7231" type="block" when="und·Jpan" />
</char>
<char cp="7231" tag="sc:Hani"> <!-- 爱 -->
  <var cp="611B" type="simp" when="und·Hani" />
  <var cp="7231" type="trad" when="und·Hani" />
</char>
<rules>
  <action disp="block" any-variant="block" />
  <action disp="allocate" only-variants="simp both" />
  <action disp="allocate" only-variants="trad both" />
  <action disp="block" any-variant="simp trad" />
  <action disp="allocate" only-variants="allocate" />
</rules>
</data>

```

138 この例の RootLGR で、「愛(U+611B)」という文字列が日本語で登録申請されたとすると、
 139 登録者(ここでは登録者 C とする)に割り当て可能な文字列は「愛(U+611B)」となり、「爱
 140 (U+7231)」は登録者に紐付けられるがブロックされる文字列となる。登録者 C とは異なる
 141 別の登録者(ここでは登録者 D とする)が後から「爱(U+7231)」を登録申請しても、登録者
 142 C に紐付けられているため、登録できない。

144 3 JGP の具体的課題定義(案)

145 3.1 登録可能文字の範囲の決定

147 JGP で登録可能とする文字の範囲を MSR の中から選ぶ。その範囲を選ぶ合理的な理由付
 148 けが課題である。例えば、以下のような範囲が考えられる。

- 150 ➤ 常用漢字の範囲
- 151 ➤ 常用漢字+人名用漢字の範囲
- 152 ➤ JIS X 0208:2012 の第一水準・第二水準漢字の範囲

➤ など

3.2 中国語ルール of 異体字が RootLGR に入った際の日本語への影響の理解

RootLGR に含む漢字の異体字として、JGP が CGP(中国語生成パネル)の異体字を容認するか個別の文字について協議するかは課題である。

中国語ルールで定義された異体字を容認とした場合は、TLD 文字列の登録申請を日本語だと宣言して行っても、中国語の異体字も区別なく異体字として扱われる。そのため、次のような状況が発生し得る。

➤ 登録者が意図していなかった文字を含む文字列が登録文字列に紐付けられブロックされる。

例：「航(U+822A) 空(U+7A7A)」と「桁(U+6841) 空(U+7A7A)」

日本語ルールが異体字を定義する・しないに関わらず、中国語ルールの異体字定義は日本語に影響を与えることを一般利用者が理解する必要があり、その情報を周知することも課題である。

3.3 日本語ルールで異体字を定義するか

日本語ルールで異体字を定義する場合でもしない場合でも、その合理的な理由付けは課題である。

また、どのような観点からその理由付けを行うかも課題である。例えば、以下のような観点がある。

➤ 日本語 TLD 登録申請者の利益を最大にすること

➤ 日本の一般利用者の混乱を最小にすること

➤ など

3.4 日本語ルールが異体字を定義する場合、個々の異体字の根拠

日本語ルールで異体字を定義する場合、個々の異体字が何を根拠とするかは課題である。これは、上記「3.3 日本語ルールで異体字を定義するか」で述べたとおり、どのような観点により異体字を定義するかによって異なる。また、JGP が独自に異体字リストを作成する場合、それをどのように権威付けるかも課題である。

3.5 日本語ルールで異体字を定義した場合の中国語への影響の把握

日本語ルールで異体字を定義した場合で、かつ、その定義が中国語ルールが定義した異体字と異なる場合、上記「3.2 中国語ルールの異体字が RootLGR に入った際の日本語への影響の理解」で述べたことと同様の状況が発生し得る。これを中国語パネルとどのように調整するかは課題である。

3.6 日本語ルール作成の迅速性

新 gTLD の次のラウンドは早ければ 2016 年に開始されることが見込まれている。RootLGR はそれまでに完成していることが必要であり、事前に CGP や KGP(韓国語生成パネル)との調整を行うことを考慮すると、遅くとも 2015 年上半期中の日本語ルール作成が必要である。迅速に日本語ルールを作成することは課題である。

[EOF]