

IETF IDN WGでの標準化状況

2001年12月4日

JPNIC 米谷嘉朗 <yone@nic.ad.jp>

日本語ドメイン名とIDN

- 日本語ドメイン名
 - IDNのサブセット
 - 通常の日本語で使われている文字を使ったドメイン名
 - 漢字、平仮名、片仮名、それらに準ずる文字(記号)
- IDN(Internationalized Domain Name)
 - ドメイン名を国際化(ドメイン名に使える文字を従来の英数字から拡張する)
 - IETFで標準化作業中

IDN WGの方向性

- 要求条件の取りまとめ (*REQUIREMENTS*)
- 「ホスト名」で使える文字の拡張
- 拡張に使用する文字セットはUnicode
- アプリケーションで対応 (*IDNA*)
 - DNSプロトコルでの対応は避ける
- 文字列は正規化 (*NAMEPREP*)
- ネットワーク上では基本的にASCII互換表現 (*ACE*)
- 第53回IETFまでには結論付けよう

2001年12月4日

日本語ドメイン名解説

3

IDNA

(Internationalizing Host Names In Applications)

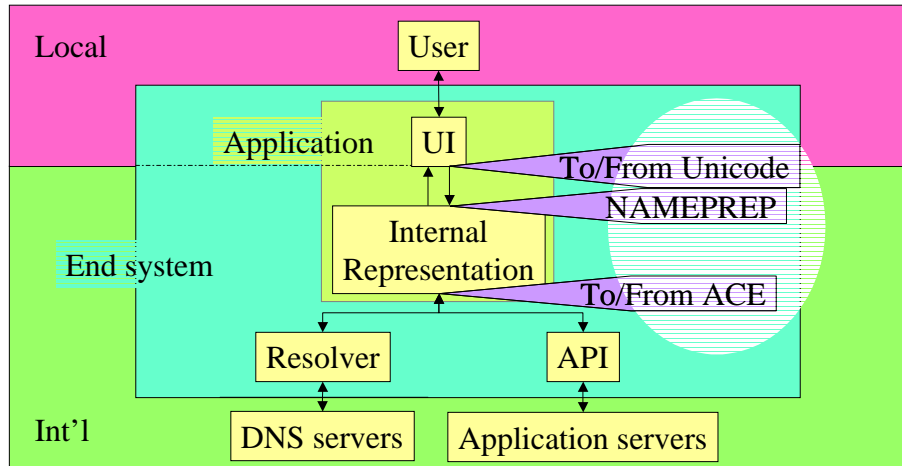
- draft-ietf-idn-idna-05.txt
- IDNの処理をアプリケーションで行う
 - ローカルコードセット(JISなど)とUnicodeの変換
 - 文字列の正規化(*NAMEPREP*)
 - IDNエンコーディング変換(*ACE*)

2001年12月4日

日本語ドメイン名解説

4

IDNAイメージ



2001年12月4日

日本語ドメイン名解説

5

NAMEPREP

(Stringprep Profile for Internationalized Host Names)

- draft-ietf-idn-nameprep-06.txt
 - 05まではPreparation of Internationalized Host Names
 - 06からSTRINGPREP (Preparation of Internationalized Strings) と対に
 - draft-hoffman-stringprep-00.txt
- 文字列の比較を容易にするため、意味的、表示的に同じ文字列の表現形式を統一
 - 文字種(大文字、小文字)
 - 互換文字(全角、半角)
 - 合成文字

2001年12月4日

日本語ドメイン名解説

6

NAMEPREP

(Stringprep Profile for Internationalized Host Names)

- 日本語の場合
 - 全角英数字とASCII
 - 「JPNIC」と「jpn ic」
 - 半角カナと全角カナ
 - 「ニック」と「ニッ ク」
 - 濁点、半濁点の合成
 - 「シ ューピー」と「ジェーピー」

NAMEPREPでの処理

1. map
大文字・小文字の文字種を統一化 (Case Folding, UTR#21)
2. normalize
複数の表現形態をもつ文字列を統一化 (NFKC, UAX#15)
3. prohibit
ドメイン名として不適切な文字の除外
 - unassigned
未定義文字はprohibitと同様に扱う

ACE

(ASCII Compatible Encoding)

- 非ASCII文字をASCII文字(英数字・ハイフン)のみで表現する方式
 - 既存のDNSで実現可能
 - DNSプロトコル、既存のアプリケーション等への影響が最小
- ベースとする文字セットはUnicode
- ラベル中で使える文字数は減少
 - 8bitデータを英数字のみ(5bit)で表現するための損失

ACEの問題

- ASCII文字列なので既存のドメイン名と区別可能な識別子が必要
 - 区別できなければ逆変換(デコード)できない
 - 識別子のつけ方、識別子の決め方が課題
 - 識別子を含めてASCII文字列のため、提案された瞬間に登録される可能性
 - 実際にgTLDで発生
 - IDNA-05では“**--”形式のプレフィックスを推奨

IDN WGにおけるACEの選択

- ACE選択基準
 - 簡易なアルゴリズムであること
 - 実装が容易なこと
 - 現実的なドメイン名に対して効果的な圧縮が効くこと
 - ラベル中で使える文字数を増やすため
 - エンコードとデコードが1対1に対応づくこと
 - 一つのドメイン名が複数のエンコード結果を生じないこと
- 第51回IETFでAMC-ACE-ZがWGの選択となった
 - 候補はDUDE、MACE、AMC-ACE-Z

2001年12月4日

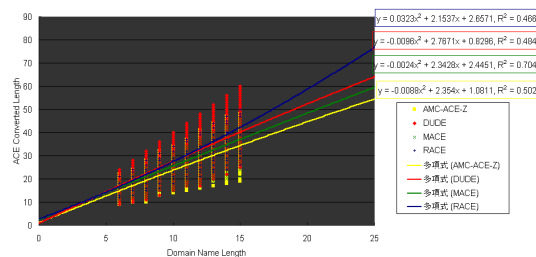
日本語ドメイン名解説

11

ACEの比較

「日本語ドメイン名試験.JP」の変換例

RACE	BQ--3BS6KZZMRKPDBSJQ4EYKIMHTKQGYUZU2CM.JP
ACE-Z	ZQ--ECKWD4C7C777U7MWO4BOV4JIOAU09J.JP



日本語JPドメイン名での評価結果

2001年12月4日

日本語ドメイン名解説

12

AMC-ACE-Zの特徴

- draft-ietf-idn-amc-ace-z-01.txt
- 圧縮アルゴリズム
 - 文字をコードポイントの小さい順に取り出し、直前に処理した文字との差分と文字位置を数値化
 - 英字、数字、ハイフンは先にくりだしておく
- ASCII文字化アルゴリズム
 - 整数を一意に決まる可変長で表現するGeneralized variable-length integersという考え方を採用
 - BASE36(A-Z, 0-9の36文字)

2001年12月4日

日本語ドメイン名解説

13

AMC-ACE-Zの考え方 (説明のために単純化)

- 「文字列例」を変換してみる
- 圧縮

1. 1:U+6587 2:U+5B57 3:U+5217 4:U+4F8B

2. 4:0x4F8B 3:0x28C 2:0x440 1:0xA30

3. 0x13E30 0xA33 0x1102 0x28C1

並び替え

数値化
(差分*文字数+
文字位置)

2001年12月4日

日本語ドメイン名解説

14

AMC-ACE-Zの考え方 (説明のために単純化)

- Generalized variable-length integers化
 - 10進数の12345は
 $1*10^4+2*10^3+3*10^2+4*10^1+5*10^0$ という表現
 - すべての桁の数字が0-9なので、12345は123と45なのか、1234と5なのか区別ができない
 - かつ、012345と12345は表現は違うが同じ値
 - この問題を解決する方法
 - 各桁ごとに閾値があり、それ以下の数字が現れると区切りとみなす方法
 - 閾値は基数以下の適当な値

AMC-ACE-Zの考え方 (説明のために単純化)

- Generalized variable-length integers化
 - 基数を36、閾値を10、18、25、25とすると
- | | | | | |
|----|------------------|--------------------|--------------------|-------------------------|
| 1. | 0x13E30 | 0xA33 | 0x1102 | 0x28C1 |
| | ↓ | ↓ | ↓ | ↓ |
| | $24*1+18*26$ | $(=1*(36-10))$ | $+30*468$ | $(=26*(36-18))+13*5148$ |
| | $(=468*(36-25))$ | | | |
| 2. | OIUD | BS4 | CN8 | XML |
| | | ↓ | ↓ | ↓ |
| | | $11*1+28*26+4*468$ | $12*1+23*26+8*468$ | $33*1+22*26+21*468$ |
| 3. | | | | |
| 4. | | | | |
| 5. | | | | |
- 「文字列例」=>“OUIDBS4CN8XML”
 - 本当のAMC-ACE-Zでは“FSQW5D78MBSK”

AMC-ACE-Z変換例

- 日本語ドメイン名.JP
↓
ZQ--ECKWD4C7C777U7MWO4BC84J.JP
- 混在EXAMPLE.JP
↓
ZQ--EXAMPLE-W48LF90O.JP
- ようこそ.ABCカンパニー.JP
↓
ZQ--P8J0A4AZM.ZQ--ABC-MJ4B9F9A7RSC.JP

2001年12月4日

日本語ドメイン名解説

17

最近の話題

次回Meetingの焦点？

最近の話題 (次回WG Meetingの焦点?)

- WGチャーター更新
 - 今後の活動項目と期限の明確化
- Reordering
 - 統計情報を使ってACEの効率を向上させる提案を採用するか
- 言語要求
 - 中国語圏の要求 (簡体字と繁体字の等価性)を標準として受け入れるか
 - 「国」と「國」、「东」と「東」など
- 3層モデル

2001年12月4日

日本語ドメイン名解説

19

チャーター、 マイルストーンの見直し

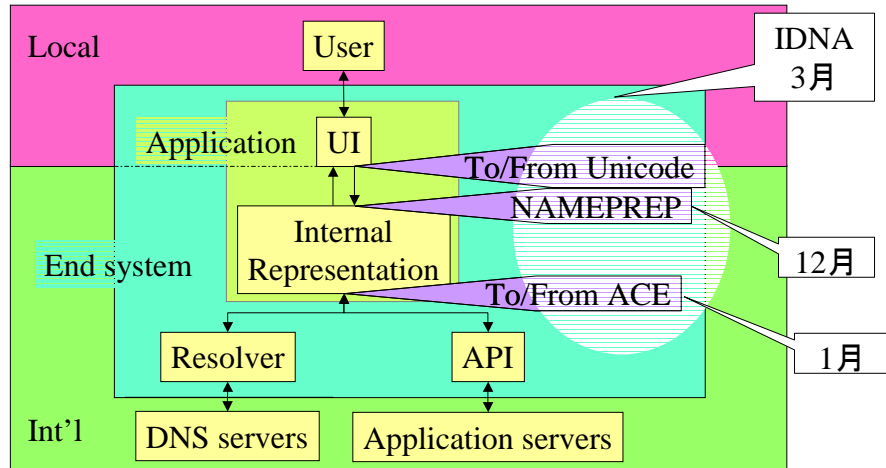
- 今後の活動項目と期限の明確化を目的
- WGドキュメントのIESG提出目標時期(IESG承認を経てRFC化される、承認まで最低2～3ヶ月)
 - NAMEPREP – 12月
 - ACE – 1月
 - IDNA – 3月

2001年12月4日

日本語ドメイン名解説

20

論点の対応箇所 (マイルストーン/IESG提出時期)



2001年12月4日

日本語ドメイン名解説

21

正規化アルゴリズム

- NAMEPREPにおける正規化アルゴリズムは副作用の少ないNFCがよいか互換文字も対象とするNFKCがよいか
- NFCは正規等価のみ正規化
 - 「シ`」と「ジ」など(合成文字)
- NFKCはNFC + 互換等価を正規化
 - 「J」と「J」など(全角・半角)

2001年12月4日

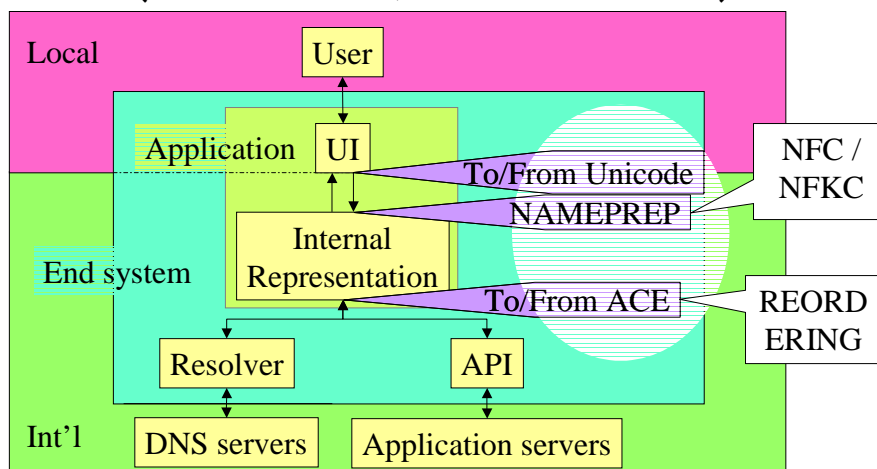
日本語ドメイン名解説

22

Reordering

- ACE化時の文字収容効率向上提案
- 文字出現頻度に基づいた文字並び替え
 - ACE変換時のプリプロセスとして実現

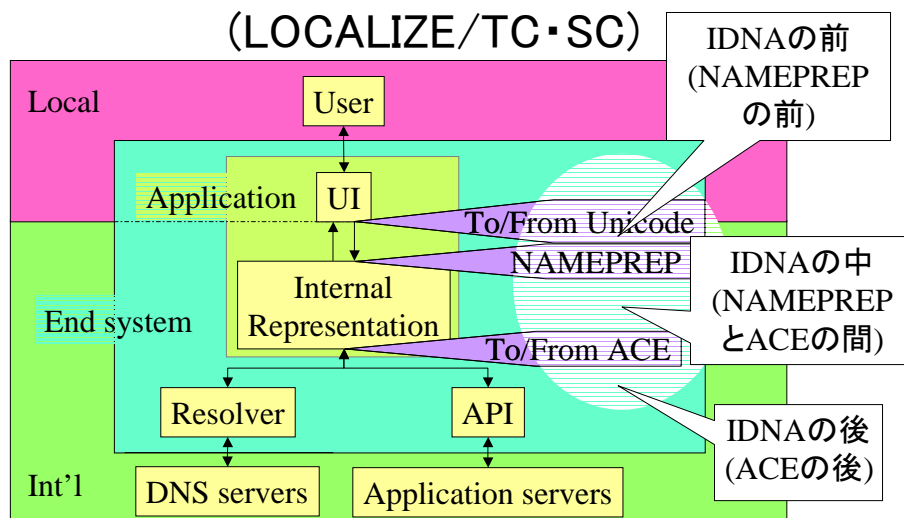
論点の対応箇所 (NFC・NFKC / REORDERING)



IDN WGでの論点

- Localization/言語的要求
 - IDNAの前、中、後のどこで実施するか
 - IDNAの前後なら標準化作業への影響は小
 - TC(Traditional Chinese)/SC(Simplified Chinese)の等価性と一意性の両立を実現したいという要求

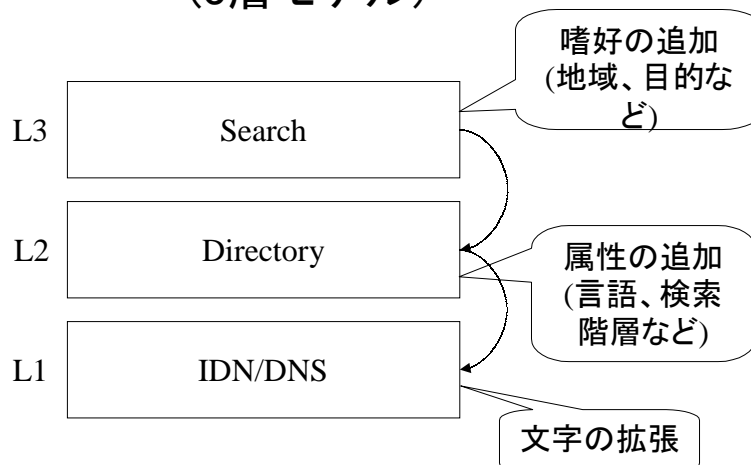
論点の対応箇所 (LOCALIZE/TC・SC)



3層モデル

- John Klensin氏の提案
 - draft-klensin-dns-search-02.txt
 - IRNSS(Internet Resource Name Search Service BOF)での議論
 - DNS自身に複雑な機構を持ち込むのではなく、その上の層(レイヤ)でのサポートを前提としてDNSで名前解決を行う方式の提案

論点の対応箇所 (3層モデル)



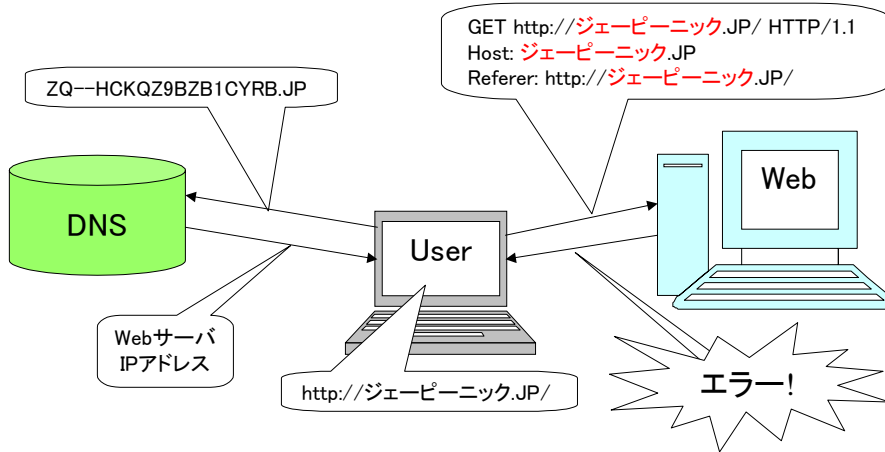
IDN WGの標準化が終了したら？

IDN標準化の終了は 日本語ドメイン名の始まり

- 利用者がIDNを使うのはアプリケーション
 - Web、メール、...
- IDNAはアプリケーションの対応が必要
- アプリケーションプロトコルにおけるIDNの取扱方式を決定していく必要がある

IDN標準化の終了は日本語ドメイン名普及(アプリケーションの対応)の開始である!

HTTPリクエスト (DNSリゾルのみ)

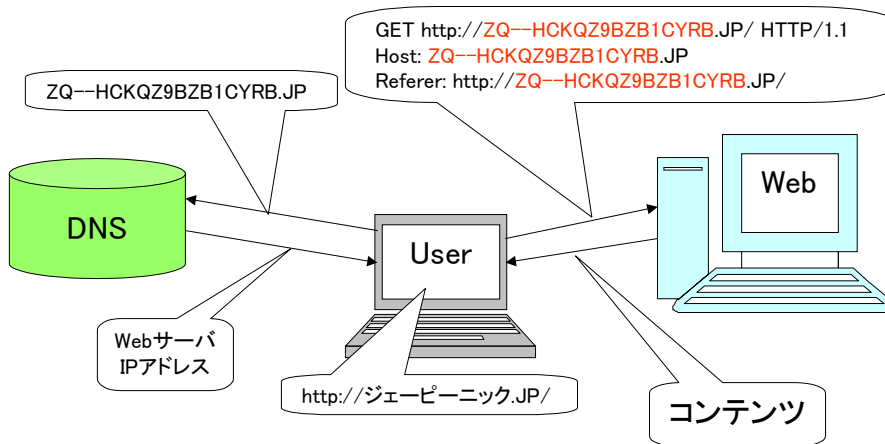


2001年12月4日

日本語ドメイン名解説

31

HTTPリクエスト (HTTPも対応)



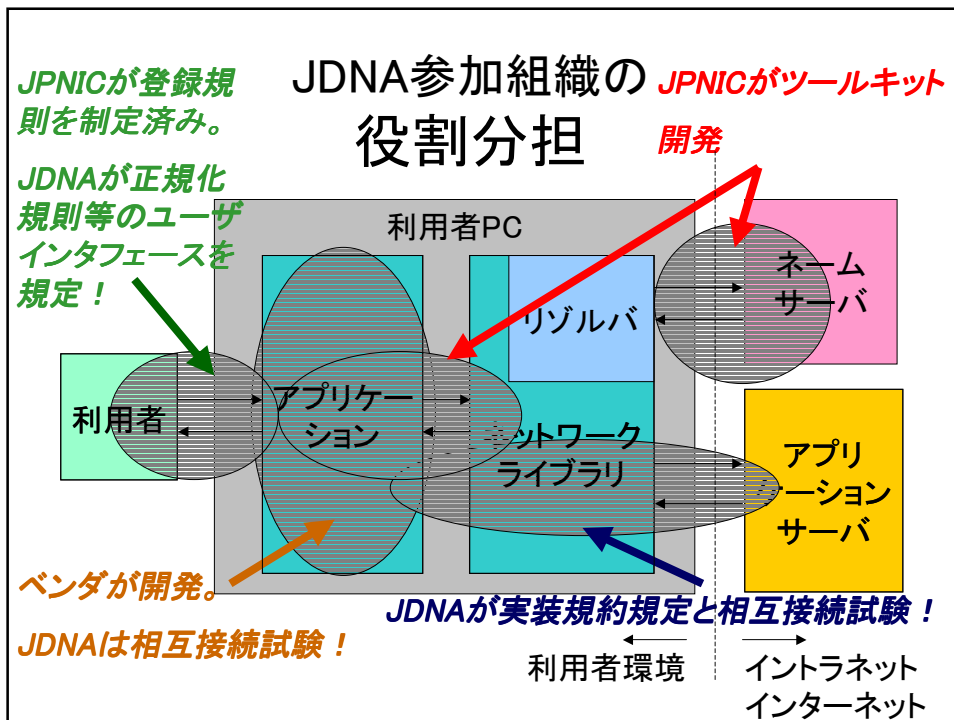
2001年12月4日

日本語ドメイン名解説

32

日本語ドメイン名協会 (JDNA)

- 日本語ドメイン名の普及を目的とした任意団体
- ドメイン名を扱うアプリケーションベンダ、サービスプロバイダ、レジストリが参加
- 2001年7月13日に設立



URL

- IETF IDN WG Web page
 - <http://www.i-d-n.net/>
- JPNIC IDN Web page
 - <http://www.nic.ad.jp/jp/research/idn/>
- JDNA Web page
 - <http://www.jdna.jp/>