IETFにおける
多言語ドメイン名規格化動向

2001年10月3日

日本ネットワークインフォメーションセンター
米谷嘉朗 <yone@nic.ad.jp>

背景

- インターネット利用者層の拡大
  - 誰もが簡単にインターネットを利用可能
  - 身近な存在化
- ドメイン名に対する社会的要求の変化
  - ブランド名、サービス名としての使われ方
  - 覚えやすい、覚えてもらいやすい名前
- なじみのある文字による表現の模索
  - いろいろなサービスの登場

標準化の状況

- インターネットではインターオペラビリティが重要
  - プロトコルの標準化が必要
  - IETFにおける標準化の活動へ
  - Internet AreaのIDN (Internationalized Domain Name) WGで作業中

IDN WGの歴史

- 第46回IETF(1999/11)にIDNS WG設立のBoF
  - その後すぐにML開始
- 2000年2月にInternet AreaでIDN WG設立
- 第47回IETF(2000/3)に最初のWG Meeting
  - Requirements(要求条件)の取りまとめが当初の目的
  - その後、プロトコルの規定も目的に追加
- これまでの活動成果
  - WG Internet-Draftは10数本
    - 個人I-Dを含めると40本以上
  - RFCはまだ(0本)

IDN WGの活動

- 国際化(多言語)ドメイン名の要求条件取りまとめ
  - REQUIREMENTS – 継続中
- ドメイン名国際化のためのさまざまな提案
  - 文字の表現方法、文字列の比較方法、DNSでの取扱方法など
- 各提案の評価と選定
  - 第51回IETF Meetingで議論 – 継続中

REQUIREMENTS

- draft-ietf-idn-requirements-08.txt
- ドメイン名の国際化（多言語化）のための要求条件定義がスコープ
  - 既存のDNSプロトコルと互換性があること
  - 文字セットはISO-10646/Unicodeであること
  - 正規化が行われること
  - 既存のドメイン名空間に容易に追加できること
  - など30項目
- 留意点
  - 既存のプロトコルとの互換性維持（RFC2825）
  - ドメイン名空間の一意性維持（RFC2826）

提案の評価

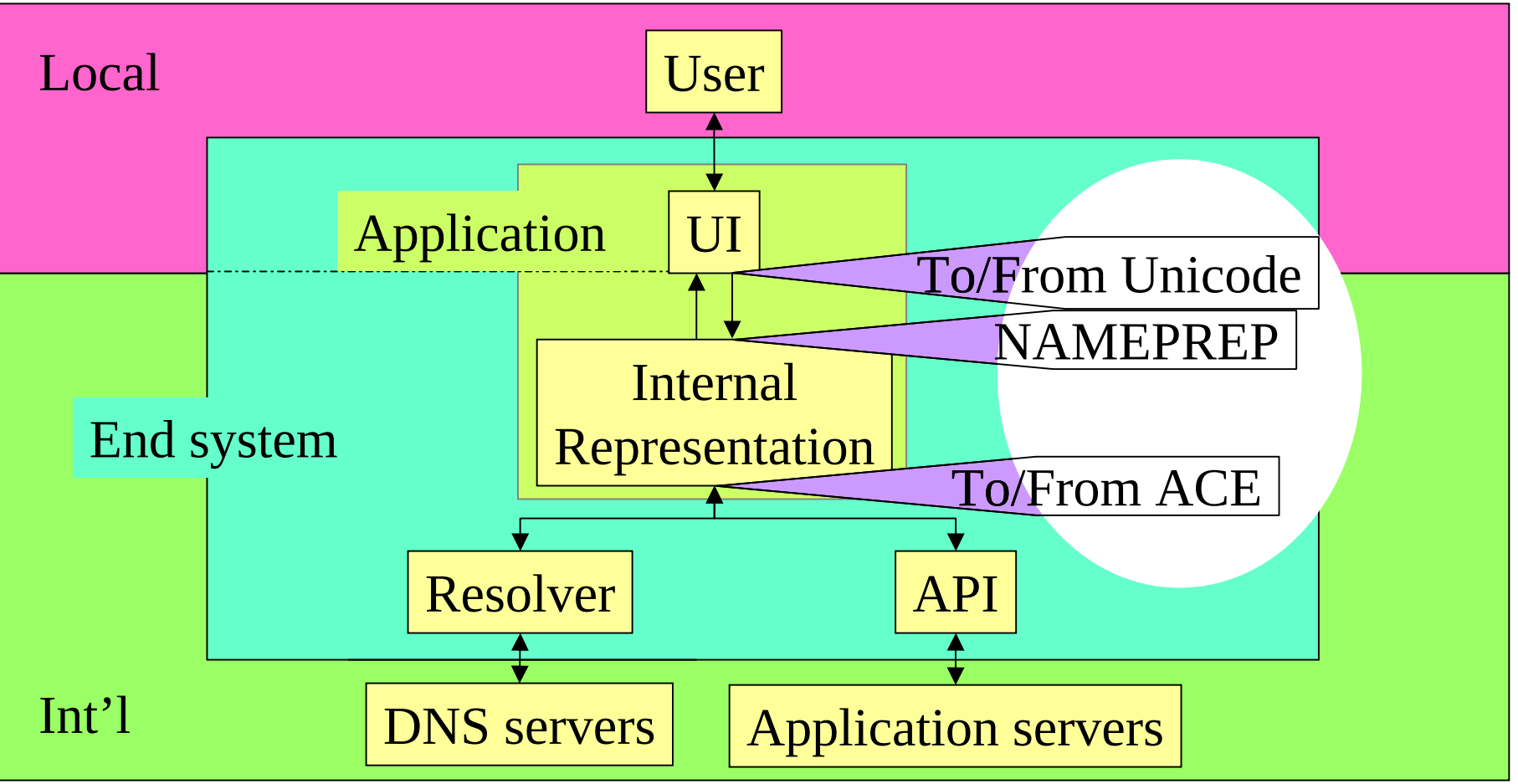
- 第51回IETF Meetingで実施
  - 数多くの提案の中から今後WGが注力していくものを選択
  1. DNSプロトコルでの取扱方法
     - IDNA vs UDNS vs UNAME
  2. 文字列の比較方法
     - NAMEPREP or not
  3. 文字の表現方法
     - AMC-ACE-Z vs DUDE

IDNA

(Internationalizing Host Names In Applications)

- draft-ietf-idn-idna-03.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)
- アプリケーション以外(リゾルバなど)で行うと…
  - アプリケーションプロトコルでの対応が困難
    - 文字コードセットの自動判別
    - メールアドレス、HTTPのHostヘッダなど

IDNA


Local
User
Application
UI
To/From Unicode
NAMEPREP
Internal
Representation
End system
To/From ACE
Resolver
API
Int'l
DNS servers
Application servers

NAMEPREP

(Stringprep Profile for Internationalized Host Names)

- draft-ietf-idn-nameprep-06.txt
  - 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

NAMEPREP
(Stringprep Profile for Internationalized Host Names)

- 日本語の場合
  - 全角英数字とASCII
    - 「ＪＰＮＩＣ」と「JPNIC」
  - 半角カナと全角カナ
    - 「ﾆｯｸ」と「ニック」
  - 濁点、半濁点の合成
    - 「ｼﾞｪｰﾋﾟｰ」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
4. unassigned
   未定義文字の扱い

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

- 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
    - draft-ietf-idn-aceid-01.txt
  - ドメイン名の構造に識別子を持たせる方式（Zero Level Domain方式）はドメイン名空間分断の危険性

ACE変換例
(RACEの場合)

- 日本語ドメイン名.JP
  ↓
  BQ--3BS6KZZMRKPDBSJQ4EYKIMHTKQGQ.JP
- 混在EXAMPLE.JP
  ↓
  BQ--3BW7OVZIABCQAWAAIEAE2ACQABGAARI.JP
- ようこそ.ABCカンパニー.JP
  ↓
  BQ--GCEEMU25.*BQ*--GD7UD72C75B2X46RZP6A.JP

ACE変換の流れ
(RACEの場合)

ユーザー		日本語.JP	ローカル
アプリケーション	コード変換	日本語.JP	Unicode
	正規化	日本語.JP (*65e5 672c 8a9e*.JP)	Unicode
	圧縮	1101100001…00000.jp	ビット列
	BASE32	3bs6kzzmrkpa.jp	ACE
ネットワーク		*bq--*3bs6kzzmrkpa.jp	ACE(+識別子)

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

ACEの提案

ACE名 (略称)	正式名	I-D
RACE	Row-based ACE	03
DUDE	Differential Unicode Domain Encoding	02
AMC-ACE-Z	Adam M Costello's 26th ACE	01

ACEの比較
「日本語ドメイン名試験.JP」の変換例

RACE	*BQ--* 3BS6KZZMRKPDBSJQ4EYKIMHTKQGYUZU2CM .JP
DUDE	*DQ--* M5E5M72C0A9EJ0C9U1Q4V3L40D0A66PA13 .JP
AMC-ACE-Z	*????* ECKWD4C7C777U7MW04B0V4JI0AU09J.JP

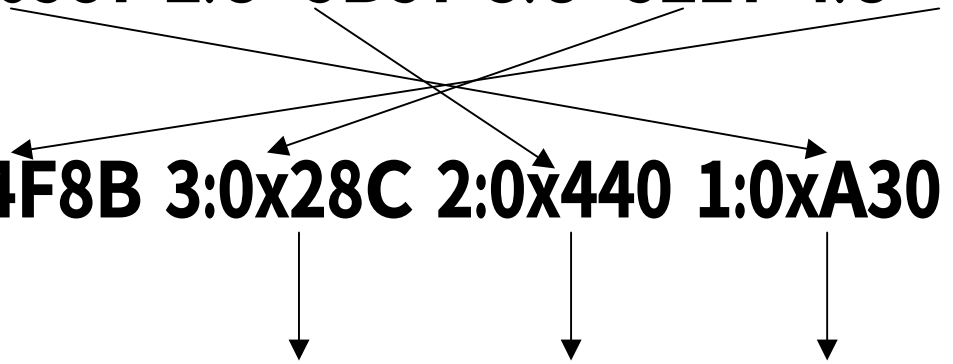
AMC-ACE-Zの特徴

- 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〈〈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)

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



並び替え

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

数値化 (差分*文字数 文字位置)

3. 0x13E30 0xA33 0x1102 0x28C1

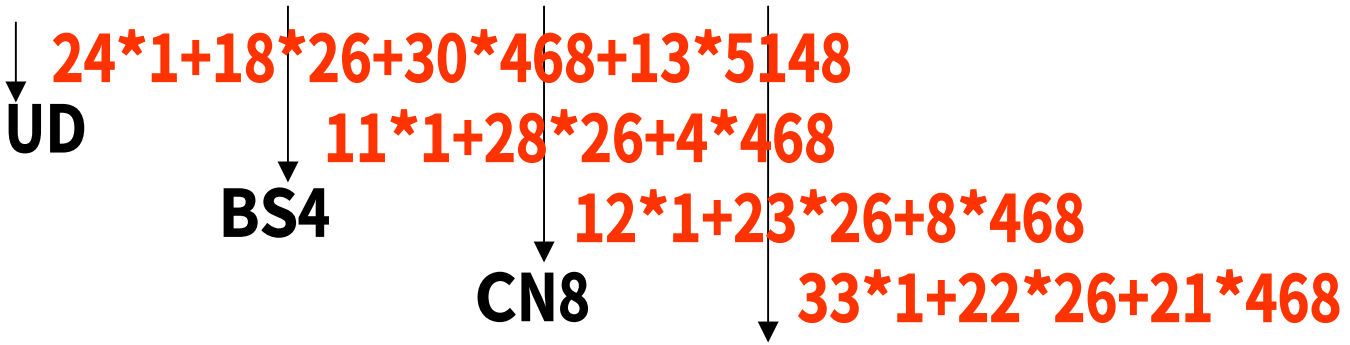
AMC-ACE-Zの考え方 (説明のために単純化)

- 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



2. OIUD
3. BS4
4. CN8
5. XML

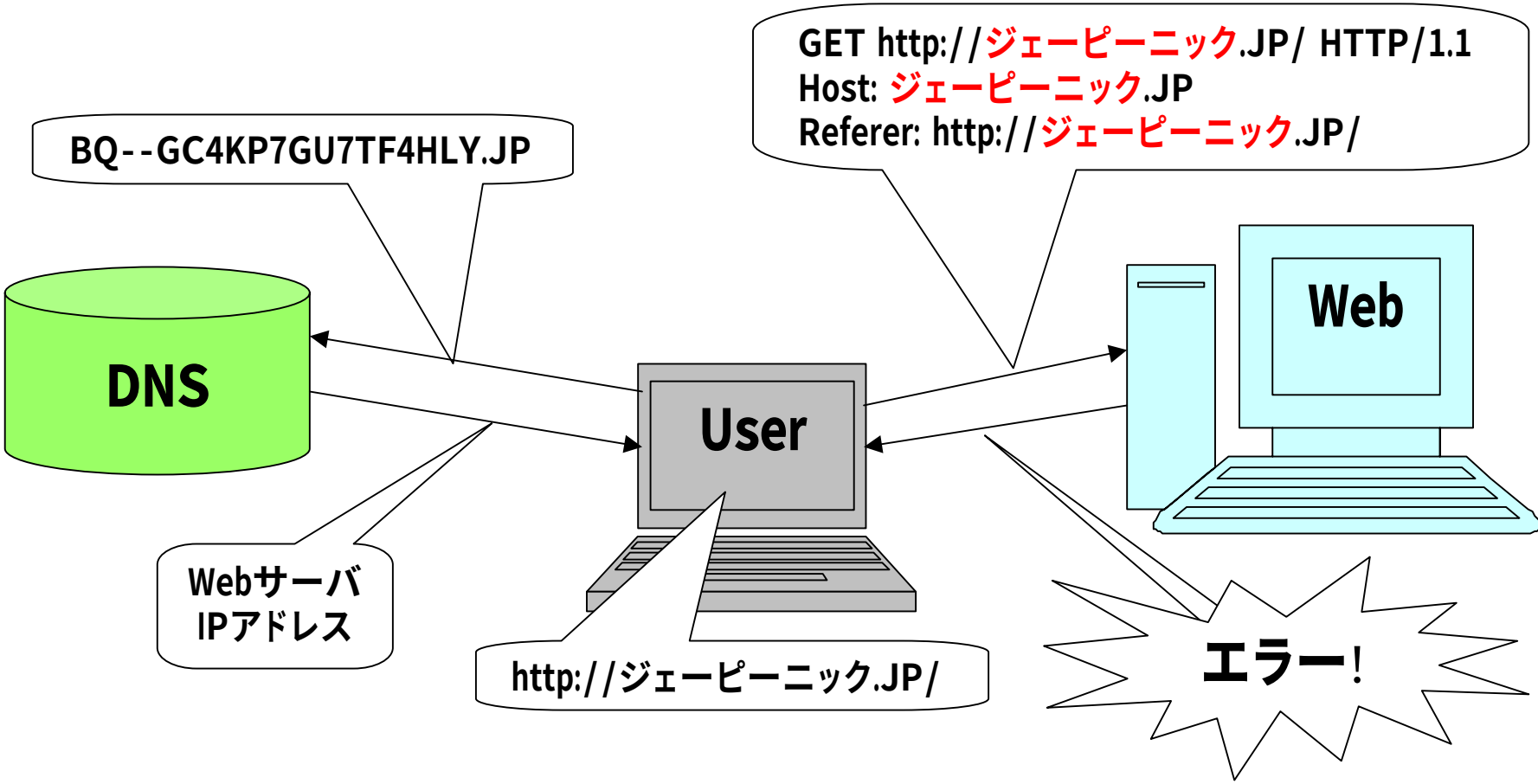
- 「文字列例」=>“OUIDBS4CN8XML”
  - 本当のAMC-ACE-Zでは“FSQW5D78MBSK”

IDNが標準化されたら終了か?

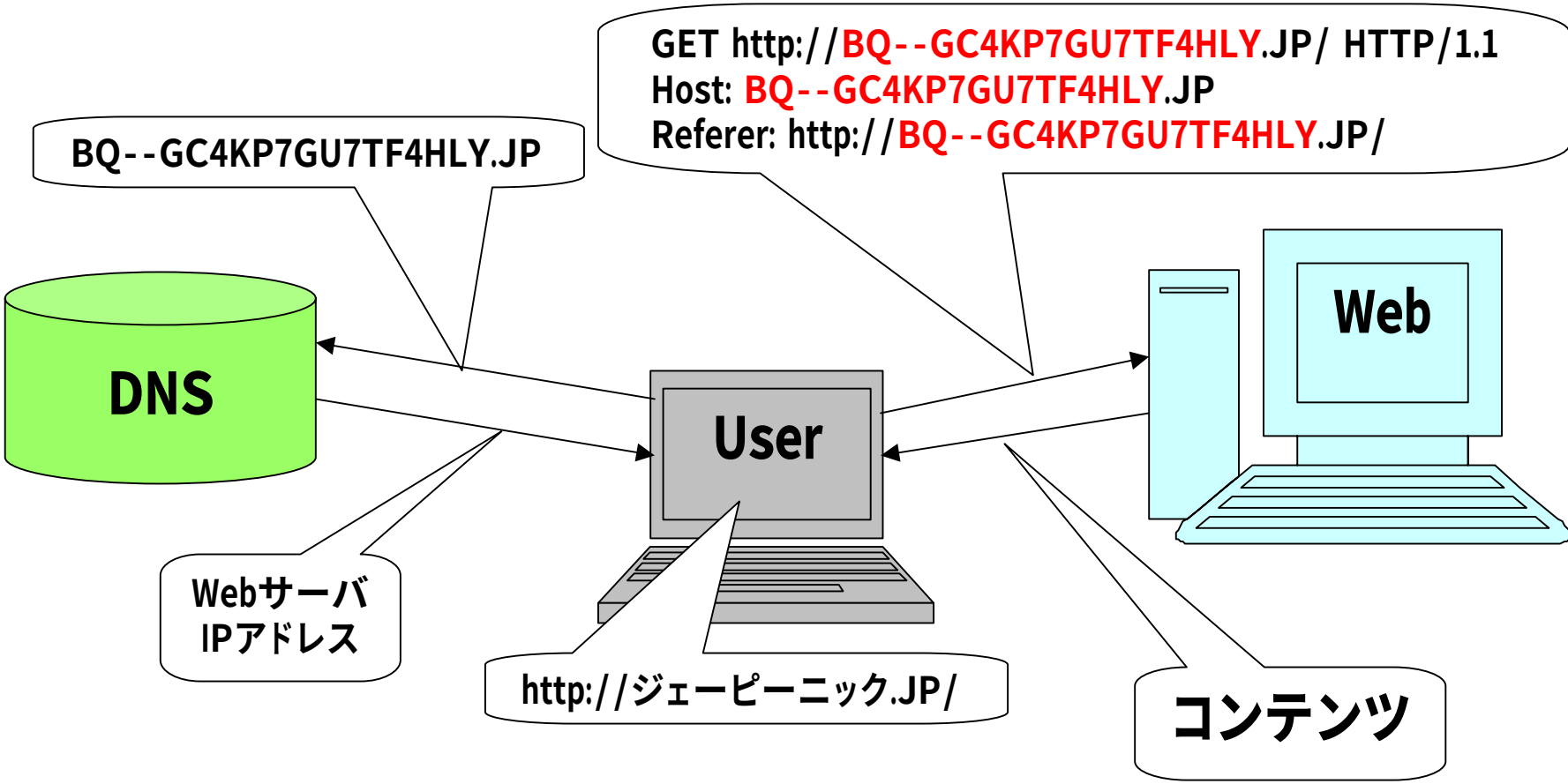
- 利用者がIDNを使うのはアプリケーション
- IDNAはアプリケーションの変更を要求
- アプリケーションプロトコルにおけるIDNの取扱方式を決定していく必要がある

IDNの標準化終了は普及(アプリケーションの対応)スタートである!

HTTPリクエスト


GET http://ジェーピーニック.JP/ HTTP/1.1
Host: ジェーピーニック.JP
Referer: http://ジェーピーニック.JP/
BQ--GC4KP7GU7TF4HLY.JP
DNS
User
Web
Webサーバ
IPアドレス
http://ジェーピーニック.JP/
エラー!

HTTPリクエスト


GET http://BQ--GC4KP7GU7TF4HLY.JP/ HTTP/1.1
Host: BQ--GC4KP7GU7TF4HLY.JP
Referer: http://BQ--GC4KP7GU7TF4HLY.JP/
BQ--GC4KP7GU7TF4HLY.JP
DNS
User
Web
Webサーバ
IPアドレス
http://ジェーピーニック.JP/
コンテンツ

日本語ドメイン名協会 (JDNA)

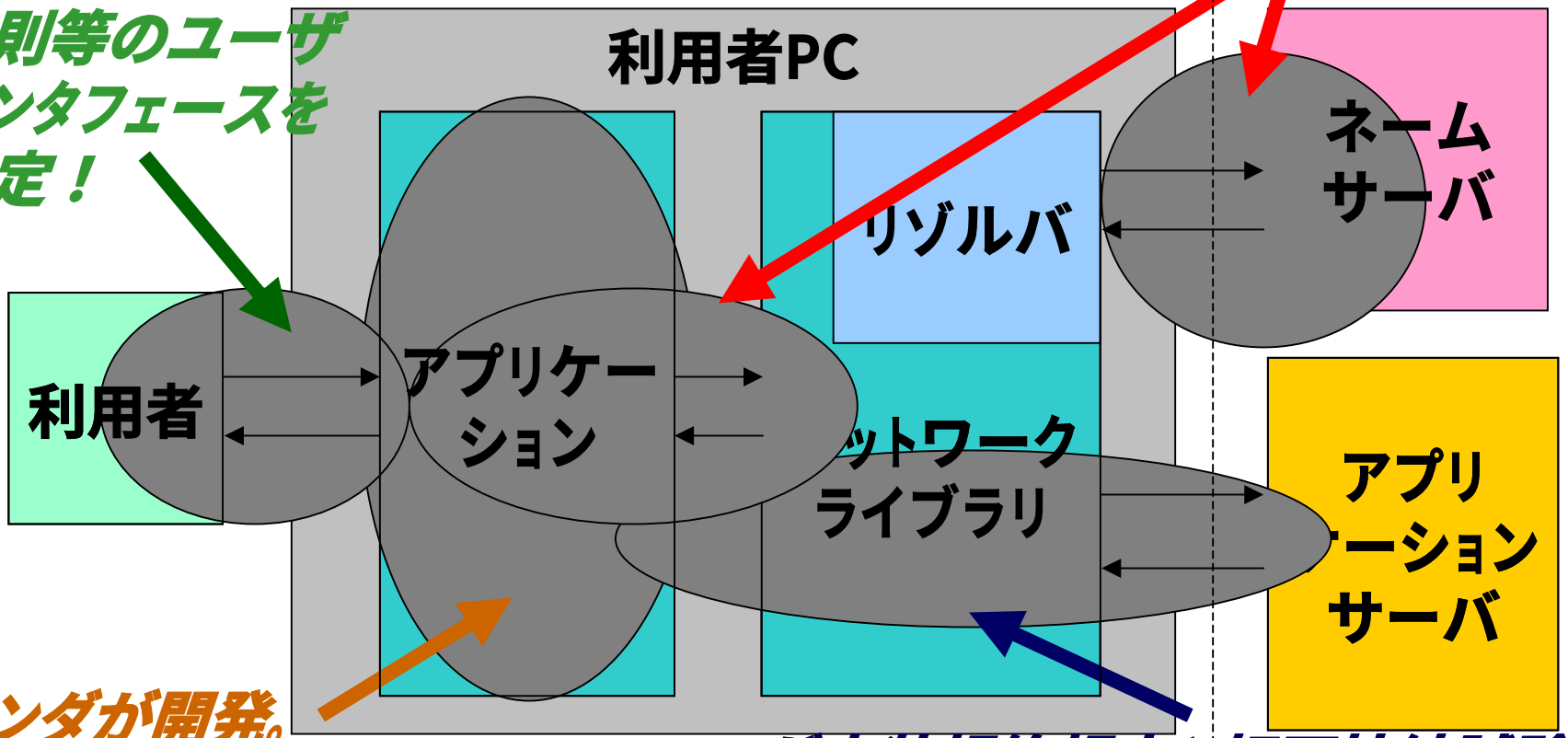
- 日本語ドメイン名の普及を目的とした任意団体
- ドメイン名を扱うアプリケーションベンダ、サービスプロバイダ、レジストリが参加
- 2001年7月13日に設立

JDNA参加組織の役割分担

PNICが登録規則を制定済み。

DNAが正規化規則等のユーザインタフェースを規定！

JPNICがツールキット開発



ベンダが開発。

DNAは相互接続試験！

JDNAが実装規約規定と相互接続試験！

利用者環境

イントラネット
インターネット

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/
- MINC
  - http://www.minc.org/

PR

- Internet Week 2001
  - 12/4 AMに日本語ドメイン名解説を行います
  - http://internetweek.jp/
- TAO(通信・放送機構)
  - 次世代DNSに関する研究開発
    - この一環でIDNのセキュリティ研究を行っています
  - http://www.shiba.tao.go.jp/prs13201.htm