Internationalizing Web Addresses

Tex Texin
Internationalization Architect, Yahoo!

Objectives

- Brief overview of web addresses,
- Architecture for domain names and URI
- Recent recommendations for use of non-English characters,
- How Unicode fits in
- Current state of the art.

What is a Web Address?

- **Web Address** – lay term for **URI**
- **URI** – Uniform Resource Identifier
- Two types of URI
  - **URL** – Uniform Resource Locator
  - **URN** – Uniform Resource Name

What are they? What is a resource?

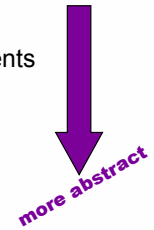
And what are IDNA and IRI?

Resource

- Anything that has identity
- Conceptual mapping to an entity (or set of entities)
  - Not necessarily the entity
  - Not necessarily network retrievable
  - Providing conceptual mapping is unchanged:
    - resource does not need to physically exist
    - content can change

Resource Examples

- Files of all types (e.g. *.txt, .jpg, .htm, …)
- Devices (e.g. printers, …)
- Databases, Database contents
- Applications, Services
- People, Corporations
- Books, DVDs
- etc.



Resource Identifier

- An object that can act as a reference to something that has identity
  - A name
  - A locator
  - Both

URL and URN

- Uniform Resource Locator (URL)
  - Subset of URI that identify resources via a representation of their primary access mechanism (e.g., their network "location")
- Uniform Resource Name (URN)
  - Subset of URI that remain **globally unique** and **persistent** even when the resource ceases to exist or becomes unavailable
  - URNs are not necessarily retrievable

Uniform (or Universal)

- Naming scheme that supports different types of identifiers
  - in the **same** context, and in **many** contexts
    - common syntactic conventions
    - consistent semantic interpretation
    - independent of access mechanism
  - extensible
    - new types do not break existing uses

Examples Of Uniformity

- Different identifier types
  - http://www.yahoo.com/
  - https://calendar.yahoo.com/textexin
  - mailto:textexin@yahoo.com
  - ftp://ftp.yahoo.com/public/
  - http://search.yahoo.com/search?p=texin
  - file:///D:/tex/index.html#toc
  - urn:example:animal:ferret:nose

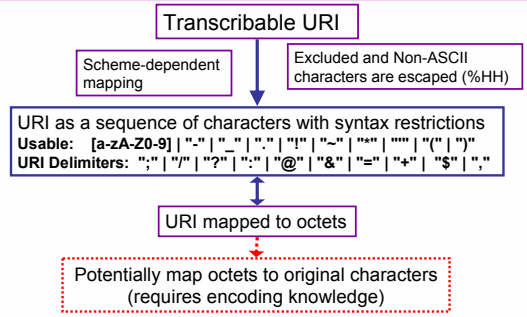
Uniform Resource Identifier

- Achieving Uniformity
  - Characters required for **Transcribability**
    - Napkin-compatible
    - Memorable
  - Common syntax across both schemes and contexts
    - Implies syntax restrictions, and
    - Character escape mechanisms

Uniform Resource Identifier

Transcribable characters conforming to a restricted syntax used for **uniformly** identifying an abstract or physical **resource**

URI Character Representation



URI Syntax

- <scheme>:<scheme-specific-part>
- <scheme>://<authority><path>?<query>
- New in RFC 3986 (replaces RFC 2396)
  <scheme>://<authority><path>?<query>#<fragment>
  - Note: with key=value pairs, **value can be URI**
    http://search.yahoo.com/search?p=**http://www.brad-pitt.com**

URI Components

- Scheme: method to access the resource
- Authority (Domain Name or IP Address)
  - Name of the machine hosting the resource
- Path: resource name, given as a path
- Query: Info. interpreted by the resource
- Fragment
  - indirect identification of a secondary resource by reference to a primary resource and additional identifying information

Each part has its own syntax!

Schemes

- Declares the type of resource and the access method.
- Defines the syntax and semantics of the rest of the URI
  <**scheme**>:<scheme-specific-part>
- Definitions are in IETF RFCs
- Scheme registry is at:
  - **www.iana.org/assignments/uri-schemes/**

Schemes

- acap, cid, crid, data, dav, dict, fax, file, ftp, go, gopher, h323, http, https, im, imap, ipp, iris.beep, ldap, mailto, mid, modem, mtqp, mupdate, news, nfs, nntp, pres, prospero, rtsp, service, snmp, soap.beep, soap.beeps, tag, tel, telnet, tftp, tip, pop, opaquelocktoken, sip, sips, urn, vemmi, wais, xmlrpc.beep, xmlrpc.beeps, z39.50r, z39.50s

Authority

<scheme>:<scheme-specific-part>
<scheme>://<**authority**><path>?<query>#<frag.>

- **authority** = server | reg_name
- server = [[userinfo "@"] **host** [":" port]]
- **host** = **hostname** | IP address
- hostname=*(domainlabel ".") toplabel ["."]
  - **Labels consist of Letters, Digits and Hyphen (LDH)**

Path and Query

<scheme>://<authority><**path**>?<**query**>#<frag.>

- **path** is specific to the authority (or scheme, if no authority), and identifies the resource within the scope of that scheme and authority
  path = segment *("/" segment)
- **query** is a string of information to be interpreted by the resource

Fragment

<scheme>://<auth.><path>?<query>#**<fragment>**

fragment = *(pchar / "/" / "?")

- indirect identification of a secondary resource by reference to a primary resource and additional identifying information
- can be a **portion** or **view** of the resource or a **reference** to another resource
- semantics depends on the primary resource, it's media type and is independent of the scheme

Internationalizing Web Addresses

Tex Texin

Internationalization Architect, Yahoo!

Internationalizing Web Addresses

- Native names are needed for usability
- Many people don't know Latin characters
  - This makes it difficult to:
    - Input, write, spell, pronounce,
    - recognize, remember, and comprehend
  - Branding, image association desired
- But backward compatibility is key!

Internationalizing Schemes

<scheme>://<authority><path>?<query>#<fragment>

- International **scheme names** not **strongly** needed

International Domain Names in Applications

Tex Texin

Internationalization Architect, Yahoo!

IDNA Goals

- Provide international standard
- Backward compatible
  - Existing DNS and application protocols continue
- One architecture worldwide
  - Independent of region, country and language

IDNA Design

- No changes to existing DNS architecture
- Applications and/or protocols compensate
  - The name character repertoire is expanded
  - A mapping to the old syntax is defined
  - New names can only be used where the application or protocol has been upgraded

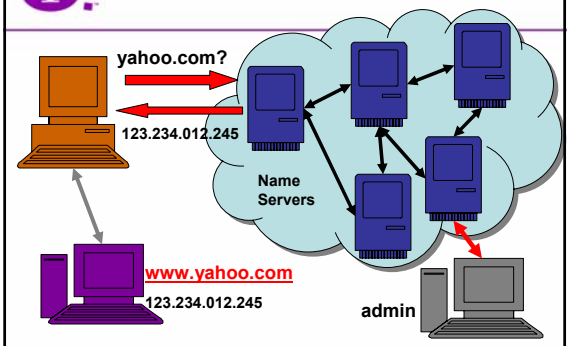
Domain Name Systems

- DNS name is hierarchical, friendly identifier for computer IP address
  - e.g. search.yahoo.com, www.kelkoo.co.uk www.kelkoo.de, 123.145.167.189
- DNS name is different from Hostname
  - DNS allows any octet, case-sensitive
  - Application (http, srv, etc.) can restrict further
    - Hostname restricts to ASCII, case-insensitive

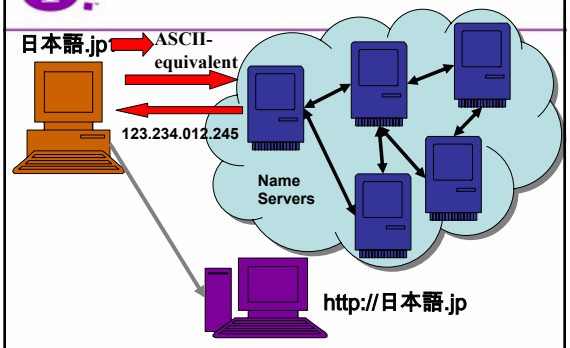
DNS Name Resolution

- DNS information maintained as a vast distributed database
- Name resolved to IP address by lookup
  - Client accesses name server(s)
  - Name servers access other name servers
  - Name servers retrieve, share and update Domain Name and IP Address information
- IDNA introduces a layer over DNS

DNS Name Resolution



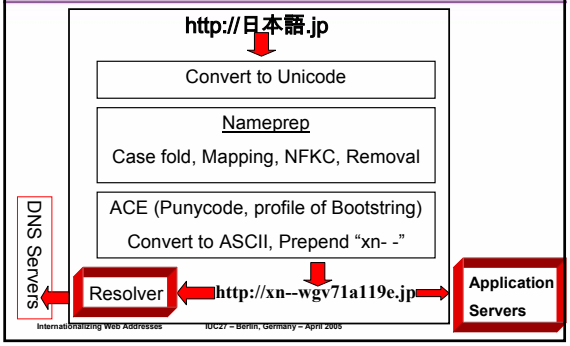
DNS Name Resolution



IDNA Architecture

- International Domain Name entered
- Conversion to Unicode, if needed
- Nameprep (DNA profile of Stringprep)
  1. Characters are folded or removed
  2. NFKC Normalization is applied (UTR15)
  3. Prohibited characters removed
- Unicode to ACE (ASCII-Compatible Encoding) Conversion

IDNA Architecture



Nameprep Character Folding

1. Case folding to lower case (UAX 21)
2. Additional folding
   - Certain Greek characters
   - Symbols which include latin characters
   - b = NFKC(a); c = NFKC(b);
     If c <> b then add a map a => c
3. Reduce typographic variations
   - line spacing, variant selectors… e.g. zwsp

Normalization (UAX 15)

- Equivalent strings are put into a single standardized form NFKC
  - Allows fast binary comparison
  - Reduces visual ambiguity
- Unicode defines two equivalences
  - Canonical and Compatibility
  - NFKC normalization standardizes both

Canonical Equivalences

- Composed vs. Combining characters
  - "Å" U+00C5 (A-ring pre-composed)
  - "A+˚ " U+0041,U+030A (A+combining ring above)
- Singletons
  - "Å" U+212B (Angstrom)

Compatibility Equivalence

- Width (ｶ カ)
- Ligature (fi)
- Font variants (𝔇)
- Breaking differences (-)
- Cursive forms (ن ﻧ ﻨ ن)
- Circled (⑪)
- Size, rotated (﹠ ⁀)
- Super/subscripts (₉ ⁹)
- Squared (㌐)
- Fractions (⅝)
- Others (dž)

Prohibited Characters

- Characters prohibited before IDNA
- Space, replacement & control characters
- Private use characters
- Non-character and surrogate code points
- Inappropriate characters (not for plain text, display variants)

  interlinear annotation, ideographic description, left-to-right mark, activate arabic form shaping, ideographic full stop

Classes of Characters

- **Based on Unicode 3.2**
- AO – Allowed characters
- MN – Characters Mapped to Nothing or normalized away
- D – Disallowed Characters (prohibited)
- U - Unassigned code points

Nameprep Versioning

- Unassigned code points become AO, MN or D when assigned **&** Nameprep is updated.
  - Applications treat unassigned code points as allowed
- Only allowed code points in name servers
  - Names are not registered until IDNA and servers are updated
  - Assumes additional case folding is minimal

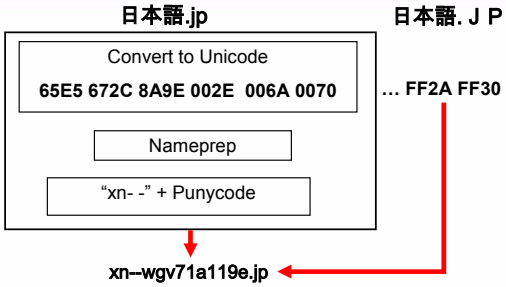
ASCII Compatible Encoding (ACE)

- ACE maps large character set to ASCII
- Bootstring algorithm minimizes length
  - DNS labels are 63 bytes, max. 255 bytes
    - Approximately 16 ideographs/63 bytes.
- Punycode parameterization for DNS
  - ASCII unchanged
  - Non- ASCII mapped to: a-z, 0-9, hyphen
- Prefix chosen to identify IDN: "**xn - -**"

Punycode

- Compression algorithm.
  - Extract characters in ascending codepoint order
  - Encode difference of codepoint from previous characters and position in an integer.
  - Extract Letters, Digits and Hyphen as bootstring.
- ASCII conversion algorithm.
  - Introduces 'Generalized variable-length integers'.
  - BASE36 (A-Z, 0-9).

IDNA Architecture Example



ICU Demo: www.ibm.com/software/globalization/icu/demo/domain

Examples International Domain Names

- afghanistan http://افغانستا.icom.museum
- algeria http://الجزائر.icom.museum
- austria http://österreich.icom.museum
- bangladesh http://বাংলাদেশ.icom.museum
- belarus http://беларусь.icom.museum
- belgium http://belgië.icom.museum
- bulgaria http://България.icom.museum
- chad http://تشادر.icom.museum
- china http://中国.icom.museum
- comoros http://القمر.icom.museum
- cyprus http://κυπρος.icom.museum
- czechrepublic http://českárepublika.icom.museum
- egypt http://مصر.icom.museum
- greece http://ελλάδα.icom.museum
- hungary http://magyarország.icom.museum

Examples
International Domain Names

- iceland http://ísland.icom.museum
- india http://भारत.icom.museum
- iran http://ایران.icom.museum
- ireland http://éire.icom.museum
- israel http://ישראל.איקום.museum
- japan http://日本.icom.museum
- jordan http://الأردن.icom.museum
- kazakhstan http://қазақстан.icom.museum
- korea http://한국.icom.museum
- kyrgyzstan http://кыргызстан.icom.museum
- laos http://[illegible].icom.museum
- lebanon http://لبنان.icom.museum
- macedonia http://македонија.icom.museum
- mauritania http://موريتانيا.icom.museum
- mexico http://méxico.icom.museum

Examples
International Domain Names

- mongolia http://монголулс.icom.museum
- morocco http://المغربl.icom.museum
- nepal http://नेपाल.icom.museum
- oman http://عمان.icom.museum
- qatar http://قطر.icom.museum
- romania http://românia.icom.museum
- russia http://россия.иком.museum
- serbia montenegro http://србијаицрнагора.иком.museum
- sri lanka http://இலங்கை.icom.museum
- spain http://españa.icom.museum
- thailand http://ไทย.icom.museum
- tunisia http://تونس.icom.museum
- turkey http://türkiye.icom.museum
- ukraine http://украина.icom.museum
- vietnam http://việtnam.icom.museum

Examples
International Domain Names

The previous IDNA examples are courtesy of:

- Cary Karp, President,
- Museum Domain Management Association, Sweden
- http://musedoma.museum/
- From his presentation (session A1) at IUC27 (http://www.global-conference.com/iuc27/program.html)

IDNA Issues

- Mapping Traditional-Simplified Chinese Characters
- Multiscript spoofs
  - www.PAYPAL.com using U+0391 Greek “A”
  - Recommendation for registry restrictions

International Resource Identifiers

Tex Texin
Internationalization Architect, Yahoo!

URI Path

<scheme>://<authority><path>?<query>#<fragment>

- URI path is ASCII-based
- %HH encoding for non-ASCII characters
  - Character encoding information is lost
  - so restoring original characters is risky
- No restrictions on equivalences
  - normalization, case folding
- Bidirectional scripts are problematic

Internationalized Resource Identifiers

- Solution: RFC 3987
  - Similar to IDNA, create a new construct Internationalized Resource Identifiers (IRI)
  - Distinct from URI, with a mapping to URI
  - Leave URI untouched and define where IRI can be used and when conversion to URI occurs.
  - Maintains backward compatibility

Internationalized Resource Identifiers

- **IRI requirements**
  - Distinguish IRI objects from URI
    - e.g. XML "anyURI"
      - www.w3.org/TR/xmlschema-2/#anyURI
  - **Don't use IRI where URI hasn't upgraded**
  - Support large character set
  - base encoding on UTF-8
    - for all URI,
    - scheme-specific URI,
    - URI component (e.g. fragment), or
    - specific URI within a scheme

IRI Usage

- Not in existing schemes, except by design
- Newly designed elements
- Presentation equivalents of existing protocols
- When used for retrieval, URI is generated
  - Scheme may have additional syntax restrictions
  - Validating URI eliminates defining equivalent IRI validation
  - Verify URI retrieval location
- Identification usage does not need URI

Internationalized Resource Identifiers

<scheme>://<authority><path>?<query>#<fragment>

- **IRI Path Component**
  - Unicode NFC (not NFKC) Normalization
    - Canonical equivalence only
    - Applied to legacy encodings only
  - Transcode to UTF-8 before %hh encoding
- Steps can also apply to other components

Example Scenarios

- http://www.w3.org/People/Dürst/
  - Web Server using UTF-8
    - IRI: http://www.w3.org/People/Dürst/
    - URI: http://www.w3.org/People/D%C3%BCrst/
  - Web Server using ISO 8859-1
    - URI:http://www.w3.org/People/D%FCrst/

Example Scenarios

- http://日本語.jp/Dürst/
  - The entire (UTF-8) string can be %hh encoded
  - Then the domain name mapping applied
  - Or the order can be reversed.
    - Normalization sequence of domain (NFKC) vs IRI (NFC) is independent…

http://xn--wgv71a119e.jp/D%C3%BCrst/

IRI Query, Fragment

<scheme>://<authority><path>?<**query**>#<fragment>

- Backward compatibility limits the restrictions that can be imposed
- The resource itself may impose restrictions:
  - E.g., query may be processed by cgi and a database based on ISO-8859-1
  - Fragment may reference a Japanese label in an euc-jis resource
- If all components can be represented as utf-8, then IRI. If any component is not, the URI.

Bidirectional IRI

- Apply, except query, fragment
- Use logical order
  - Rendering
    - Unicode bidirectional algorithm
    - Present as if embedded Left-To-Right
  - Host names
    - Labels should not mix LTR and RTL chars
    - Labels with RTL characters should start and end with RTL characters

IRI and Web Servers

- The IRI is UTF-8 based, but the file system may not be.
  - e.g. Unix/Linux file system is just bytes.
  - File names are in the user's locale/encoding.
  - Therefore each web resource name may use a different (user's) encoding on disk.
- A different mapping may be required from IRI path to each filename's encoding.

Example Solution Apache mod_fileiri

- Martin Dürst created a patch for Apache.
  - Encodings of files are named in **.htaccess** file. Web server can then map IRI to filename of each file.

  www.w3.org/2003/06/mod_fileiri/

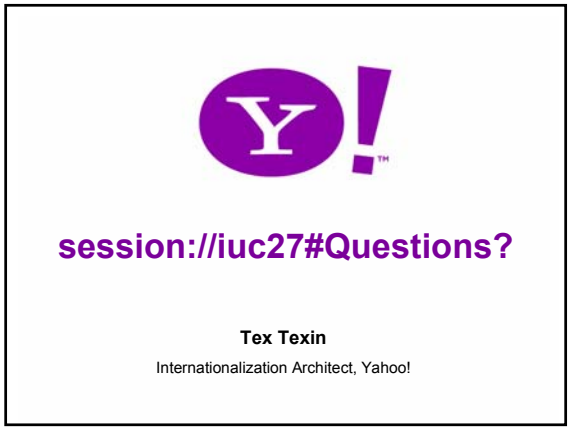
  www.w3.org/2003/Talks/0904-IUC-IRI/slide19-0.html
- If Unix (or other) file system is UTF-8, conversion is not needed.
- IIS and Apache 2 work as-is on Win 2000/XP

Support

- IDNA
  - Mozilla, Opera, Safari;
  - IE only with plugin
- IRI
  - IE, Mozilla can configure to use UTF-8
  - Opera and Safari
- IBM ICU, open source: uidna_
- Verisign list of supporting products

Detection

- How to detect IRI vs. URI?
- Two possibilities:
  - Generally assume if byte pattern fits UTF-8, it is likely UTF-8.
    - Not reliable for short strings, esp. Chinese, or when listing large numbers of URI as Yahoo! does.
  - Convert address to escaped form both ways. (UTF-8, native encoding). Do server requests.



References

- RFC 3490 IDNA
- RFC 3454 Stringprep
- RFC 3491 Nameprep
- RFC 3492 Punycode
- Intro. to Multilingual Web Addresses
- www.dns.net/dnsrd/rfc/
- RFC 3986 URI
- RFC 3987 IRI
- IDN and URI [PDF], Michel Suignard
- W3C Character Model, Resource Identifiers
- Numerous papers at Unicode Conferences

Internationalizing Web Addresses IUC27 – Berlin, Germany – April 2005 64