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
Internationalizing Web Addresses

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
  - 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

Transcribable URI

- Scheme-dependent mapping
- Excluded and Non-ASCII characters are escaped (%HH)

URI as a sequence of characters with syntax restrictions
Usable: [a-zA-Z0-9] | "" | "" | "" | "" | "" | "" | "" | "" | "" | ""
URI Delimiters: | | | | | | | | | |

URI mapped to octets

Potentially map octets to original characters
(requires encoding knowledge)
Internationalizing Web Addresses

**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


**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/

**Authority**

- `<scheme>`:<scheme-specific-part>
  <scheme>://<authority><path>;<query>!fragment>

  * 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>!fragment>

  * 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

  - segment = * (Letters, Digits and Hyphen (LDH))
Internationalizing Web Addresses

Fragment

\[ \langle \text{scheme} \rangle : // \langle \text{auth} \rangle \langle \text{path} \rangle? \langle \text{query} \rangle \# \langle \text{fragment} \rangle \]

\[ \text{fragment} = "( \text{pchar} / " / "?" )" \]

– 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, its media type and is independent of the scheme

Internationalizing Schemes

\[ \langle \text{scheme} \rangle : // \langle \text{authority} \rangle \langle \text{path} \rangle? \langle \text{query} \rangle \# \langle \text{fragment} \rangle \]

• 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
Internationalizing Web Addresses

**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**

- 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
**Internationalizing Web Addresses**

### IDNA Architecture

- **http://日本語.jp**
- Convert to Unicode
- **Nameprep**
- Case fold, Mapping, NFKC, Removal
- ACE (Punycode, profile of Bootstring)
- Convert to ASCII, Prepend ‘xn--’

### Nameprep Character Folding

1. Case folding to lower case (**UAX 21**)
2. Additional folding
   - Certain Greek characters
   - Symbols which include latin characters
   - $b = \text{NFKC}(a); c = \text{NFKC}(b)$; 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)
  - “Å+” U+0041,U+030A (A+combining ring above)
- Singletons
  - “Å” U+212B (Angstrom)

### Compatibility Equivalence

- Width (_CHARSET)
- Ligature (fi)
- Font variants (¥¥)
- Breaking differences (-)
- Cursive forms (ٖ)
- Circed (珺)
- Size, rotated (¥~)
- Super/subscripts (¥)
- Squared (珺)
- Fractions (%)
- Others (dż)

### 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
**Internationalizing Web Addresses**

### 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
- String 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

#### Japanese Domain Name

- Convert to Unicode: 65E5 672C 8A9E 002E 006A 0070
- Nameprep: xn--wgv71a119e.jp
- FF2A FF30

---

**Examples International Domain Names**

- afghanistan http://افغانیстан.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://تشاد.icom.museum
- china http://中国.icom.museum
- comoros http://коморы.icom.museum
- cyprus http://κύπρος.icom.museum
- czechrepublic http://českárepublika.icom.museum
- egypt http://مصر.icom.museum
- greece http://ελλάδα.icom.museum
- hungary http://magyarország.icom.museum

---

**ICU Demo**: www.ibm.com/software/globalization/icu/demo/domain
Examples
International Domain Names

- iceland http://ísland.icom.museum
- india http://भारत.icom.museum
- iran http://ايران.icom.museum
- ireland http://éire.icom.museum
- israel http://ישראל.icom.museum
- japan http://日本.icom.museum
- jordan http://اﻷردن.icom.museum
- kazakhstan http://қазақстан.icom.museum
- korea http://한국.icom.museum
- kyrgyzstan http://казакстан.icom.museum
- laos http://ລາວ.icom.museum
- lebanon http://لبنان.icom.museum
- macedonia http://македонија.icom.museum
- mauritania http://موريتانیا.icom.museum
- mexico http://мексикo.icom.museum
- mongolia http://монголулс.icom.museum
- morocco http://المغرب.icom.museum
- nepal http://नेपाल.icom.museum
- oman http://اُمِّان.icom.museum
- qatar http://قطر.icom.museum
- romania http://رومانيا.icom.museum
- russia http://россия.иком.museum
- serbia montenegro http://србијаицрнагора.иком.museum
- sri lanka http://sri.lanka.icom.museum
- spain http://españa.icom.museum
- thailand http://tland.icom.museum
- turkey http://turkey.icom.museum
- ukraine http://украина.icom.museum
- vietnam http://việtnam.icom.museum

The previous IDNA examples are courtesy of:
- Cary Karp, President, Museum Domain Management Association, Sweden
- http://musedoma.museum/
- From his presentation (session A1) at 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

URI Path

<tr xmlns="">
<url>
<scheme>
<authority><path?><?query>#<fragment></url>

- 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

International Resource Identifiers

Tex Texin
Internationalization Architect, Yahoo!
Internationalizing Web Addresses

**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

**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

**Example Scenarios**

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

- **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/](http://xn--wgv71a119e.jp/D%C3%BCrst/)
Internationalizing Web Addresses

**IRI Query, Fragment**

```html
<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/](http://www.w3.org/2003/06/mod_fileiri/)
    - [www.w3.org/2003/Talks/0904-IUC-IRI/slide19-0.html](http://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.
Internationalizing Web Addresses

session://iuc27#Questions?

Text Texin
Internationalization Architect, Yahoo!

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