Internationalization in IETF contexts: Cary Grant and Audrey Hepburn explain it all to you.

Ted Hardie
March 22, 2007
Contexts for Internationalization

- Protocol descriptions
- Protocol elements
- Human elements
  - Characters
  - Languages
  - Identifiers
  - Searching
  - Comparison
Protocol descriptions

- The clip's protocol description was in at least four different languages, but only two were close to complete.
- To know that, though, you have to speak them all.

- IETF protocol documents are in English, as a definitive source language.

- There is a basic right to translate into any other language. From BCP 78, Section 7.1.c:
  - “to prepare or allow the preparation of translations of the Contribution into languages other than English”
Protocol Elements

- You don't internationalize a chin or an orange.
- Even if the protocol description is highly textual (like XML), treating the protocol bits as tokens, state machine transitions, or similar formal constructs simplifies protocol design.
- A UI can map the tokens to something appropriate if they must be displayed to users.
- The same approach works well for error processing; rather than have multiple natural language error messages, use a code and allow UI mapping to a specific context.
The Grey Area

- Things we think are protocol elements can sometimes become human elements.
  - URLs were tagged protocol elements hiding behind anchor text in HTML.
    - Some are now high-value brands.
    - Some, however, are AJAX goop.
  - This can change the interoperability strategy.
    - Strategy one has the presentation layer give a locally meaningful answer based on a token.
    - Strategy two is that one or both ends understand the local context enough to use it to give a meaningful answer, or fail gracefully.
Human Elements: Characters

- The IETF tends to talk about “charsets” (see RFC 2978): a set of characters in a specific encoding. We commonly use UTF-8 (STD 63), an encoding for Unicode.

- Unicode is a set of numbers (code points), each of which maps to a character.
  
  What “character” means is a little complicated.

- We do register other, more-limited charsets and use some special-purpose encodings, e.g. the ASCII-compatible encoding for IDNs.

- Subtitling our clip, a limited character set would work, but UTF-8 covers any likely contingency.
The IETF tends to talk about "charsets" (see RFC 2978): a set of characters in a specific encoding. We commonly use UTF-8 (STD 63), an encoding for Unicode. Unicode is a set of numbers, each of which maps to a character. What "character" means is a little complicated. We do register other, more-limited charsets and use some special-purpose encodings, e.g. the ASCII-compatible encoding for IDNs.
Human Elements: Characters

- The IETF tends to talk about “charsets” (see RFC 2978): a set of characters in a specific encoding. We commonly use UTF-8 (STD 63), an encoding for Unicode.

- Unicode is a set of numbers, each of which maps to a character.
  - What “character” means is a little complicated.

- We do register other, more-limited charsets and use some special-purpose encodings, e.g. the ASCII-compatible encoding for IDNs.

- Subtitling our clip, a limited character set would work, but UTF-8 covers any likely contingency.
Human Elements: Languages

- Language tags and registry defined in RFC 1766.
  - Human languages.
  - Content-language header defined.
  - May be for content that is not textual.
  - Not easy to represent multi-lingual content.
  - Usable in content negotiation.
- RFC 4646 obsoletes original registry.
  - Switched to subtag-based registry and construction of tags.
  - Matching algorithms are defined, but not required.
Human Elements: Identifiers

- Unstructured identifiers, like human names, have implicit context (e.g. Brian Cruikshank alias Peter Joshua alias Alexander Dyle alias Adam Canfield).

- Structured identifiers, like URIs, have some explicit context. Original URI spec said non-ASCII must be %-encoded, but did not limit the encoding. RFC 3986 assumes it is UTF-8 or ASCII superset.

- IRIs (RFC 3987) are closer to a presentation layer for identifiers.

- All URIs are IRIs; all IRIs must have a URI form; other rules defined on a scheme-specific basis.
Human Elements: Search

- Getting a protocol to see a match where a human would see a match can be tricky; see the IANA collation registry and RFC 4790.

- Substring matching is easy when it maps to octet-by-octet compare.
  - Unfolding may be necessary to get multi-octet sequences right.

- Normalization may be necessary to meet user expectations when both composed characters and combining characters are possible or when case insensitive matching is desired.

- Some issues are politically promoted but not viable (e.g. after one-way transform).
When do you need an internationalization section?

- Theoretically, when you deal with humans.
- Practically, it's like a plot summary: when you need to give a précis of the key points to date.
  - Historically, folks have documented how their decisions differed from the norm.
  - Since the norm has always been moving target, that's created a lot of lore.
- John will take us through the effort to create a new chapter in that lore now.