IDN - what's up?

Patrik Fältström
paf@cisco.com

Old stuff (what is IDNA)

- What is it?
- What implications do we get?

IDNA uses Unicode 3.2

Protocol issues

- Old protocols can only handle a subset of US-ASCII (A-Z etc)
- People want to use more characters when addressing resources (use Unicode)
- Two possible solutions:
  - Change protocols
  - "Encode" characters in US-ASCII

Encodings - 1

- With encoding one mean “translate something into something else, so the original data can be retrieved again by inversing the translation”, like creating a mapping function

Encodings - 2

- Example:
  - Say 1=A, 2=B etc
    - Instead of sending B C, we send 2 3
    - Receiver get 2 3, and converts to B C
- This mechanism is used in Email

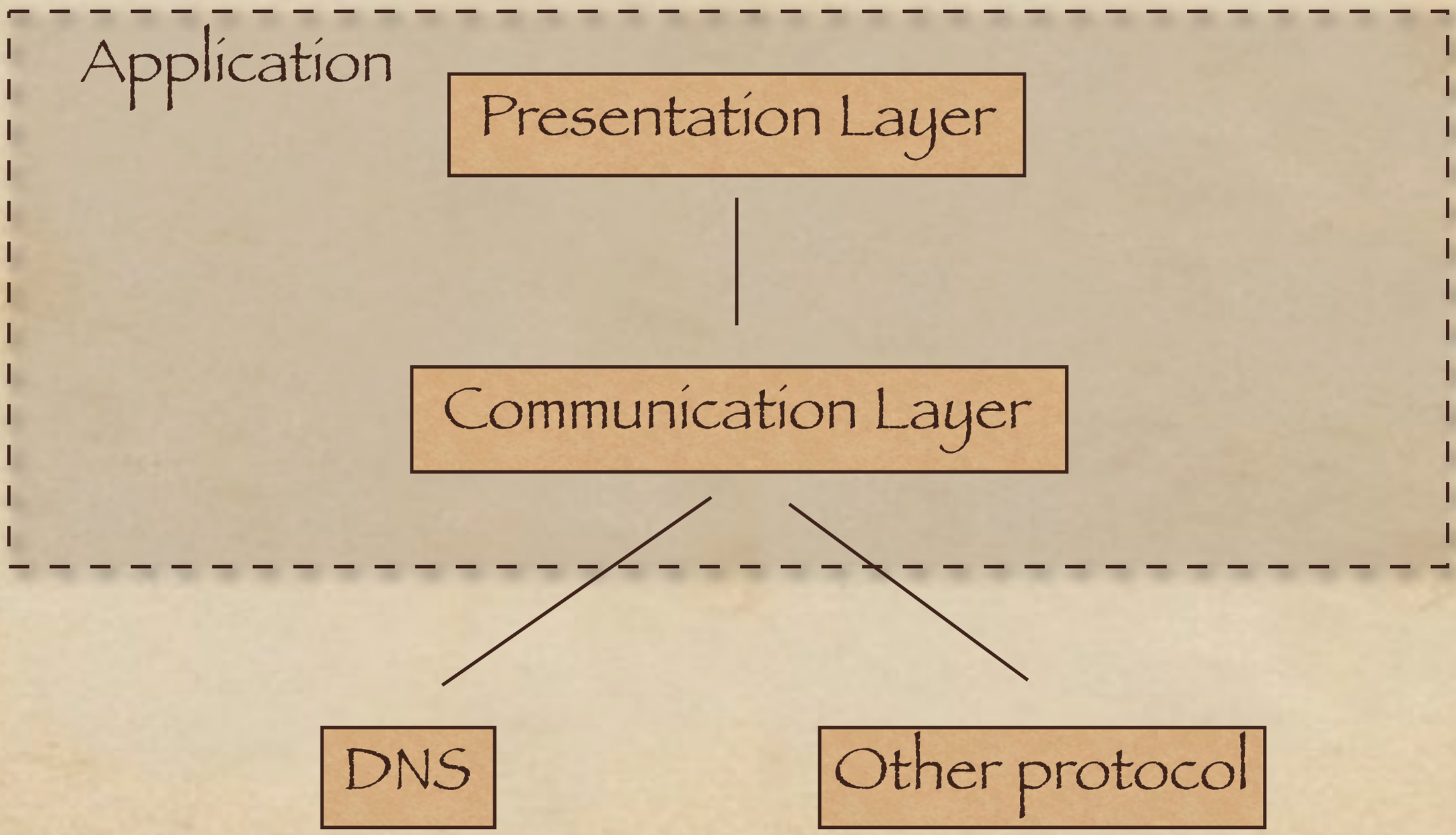
Before sending

- Sender types domain name in application
- Text is translated into Unicode
  - If it is not Unicode already
- The Unicode string is encoded in US-ASCII

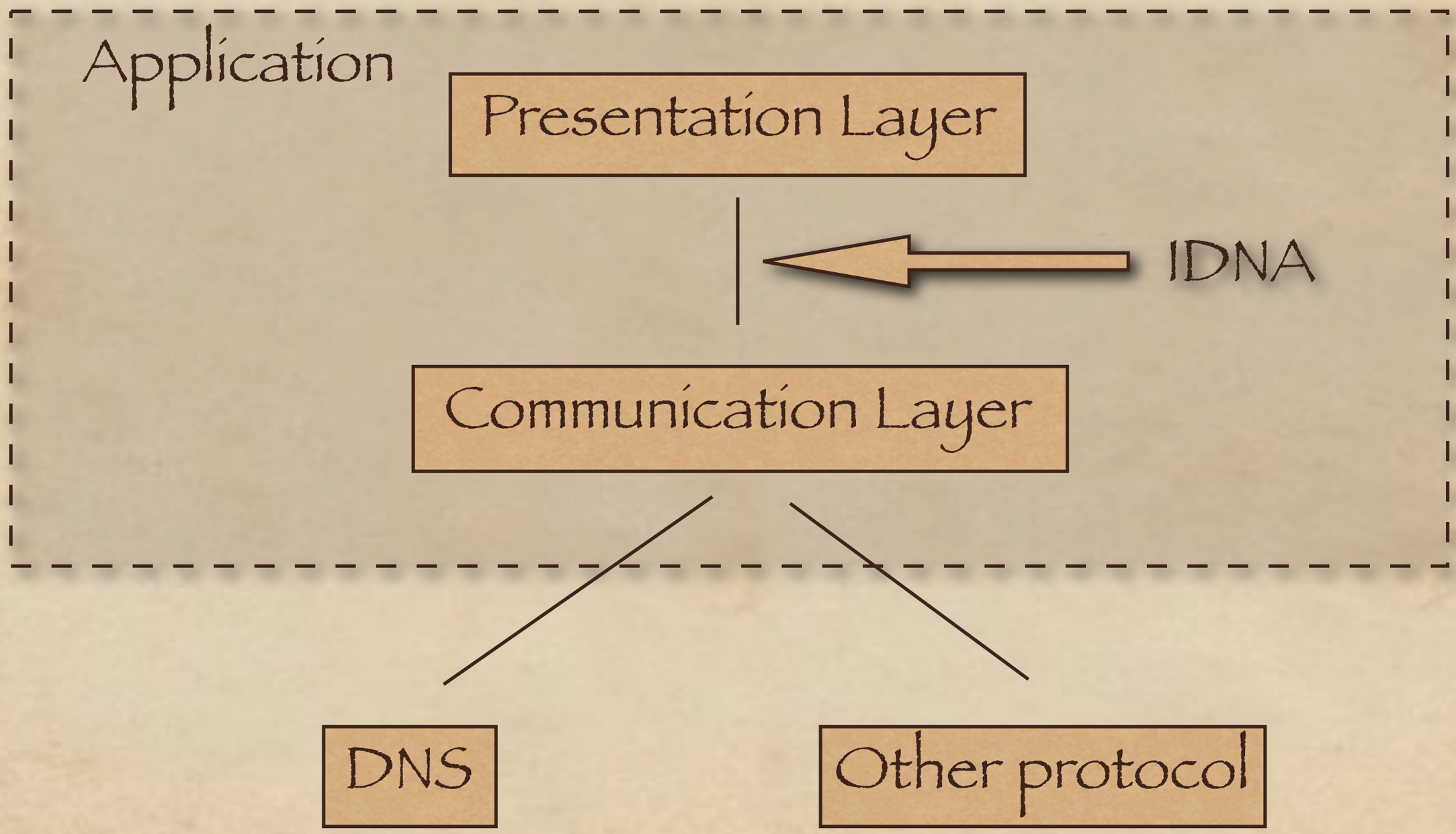
After receiving

- Receiver decodes the US-ASCII string
- Receiver translate text from Unicode to local charset
  - If not Unicode can be used directly
- The domain name is presented to the receiver

Where is this applied?



Where is this applied?



Encoding...

1. Input from user

   Fa¨ltström

2. Apply Nameprep

   fältström

3. Apply Punycode

   xn--fltstrm-5wa1o

Implications

- Two different strings in Unicode might be "equal" according to the rules
- Two strings "looking" the same might be different Unicode strings and different strings according to the rules

Implications

- Example:
  - Fältström and fa¨ltström
    - xn--fltstrm-5wa1o
    - Today Faltstrom and faltstrom are equal
- IDNA does not change DNS rules

Implications

- Example:
  - CYRILLIC SMALL LETTER IE (U+0435)
  - LATIN SMALL LETTER E (U+0065)
- Also of course a font issue...

More implications

- What is "domain name" and what is in zone file are two different things
  - fältström.se

    xn--fltstrm-5wa1o.se
  - 费思哲.se

    xn--xwrt3x2r0b.se

Example

What registrant
wanted to register
Fältström.se

What someone
might type in
Fa¨ltström.se

What's in the zonefile
xn--fltstrm-5wa1o.se

What one get which
decode the domain name
fältström.se

Spot The Difference…

Spot The Difference…

Spot The Difference…

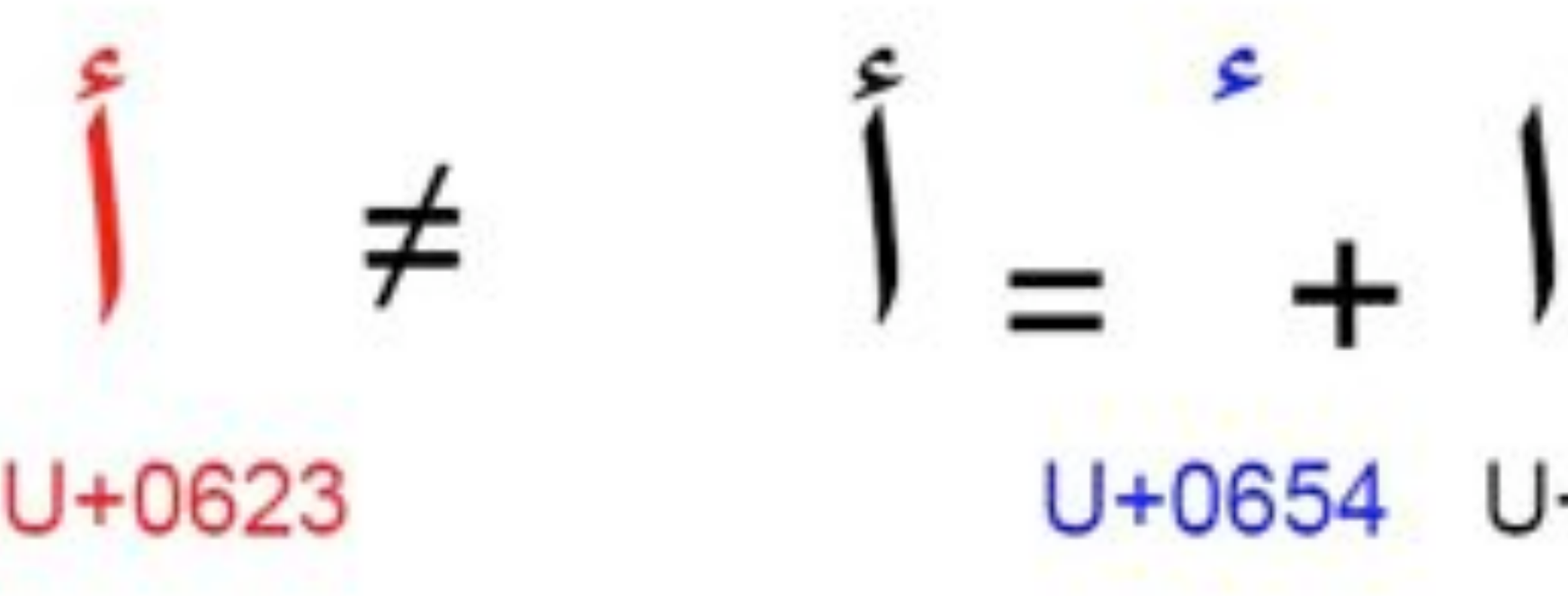
Spot The Difference…

≠

input[0] = U+0627
input[1] = U+0654
input[2] = U+066e
input[3] = U+06ec
input[4] = U+0627

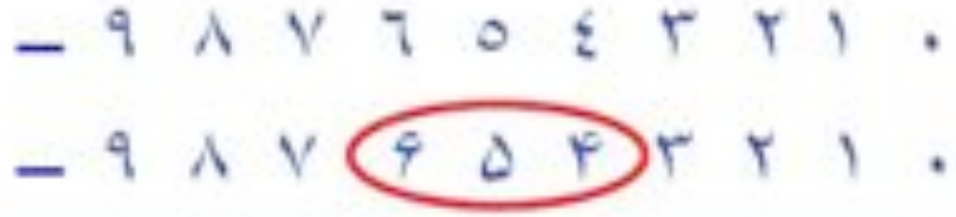
input[0] = U+0623
input[1] = U+0646
input[2] = U+0627

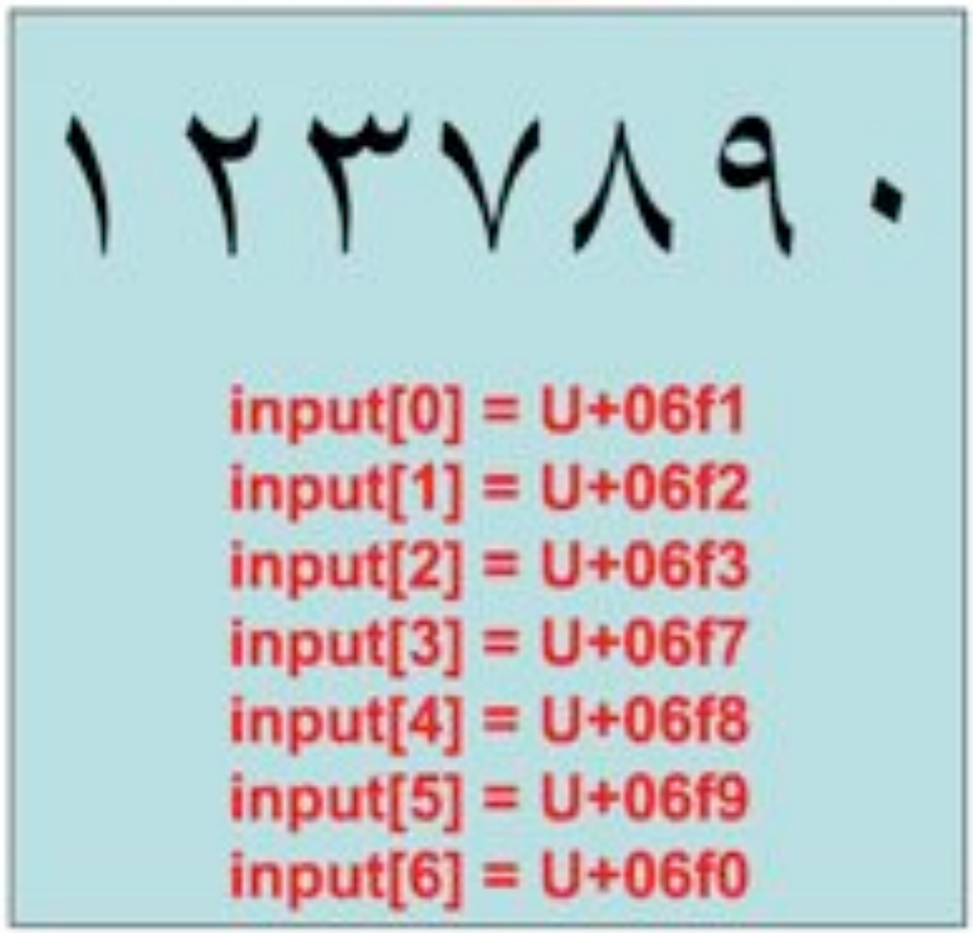
They Look The Same To Us … But Not To A Computer



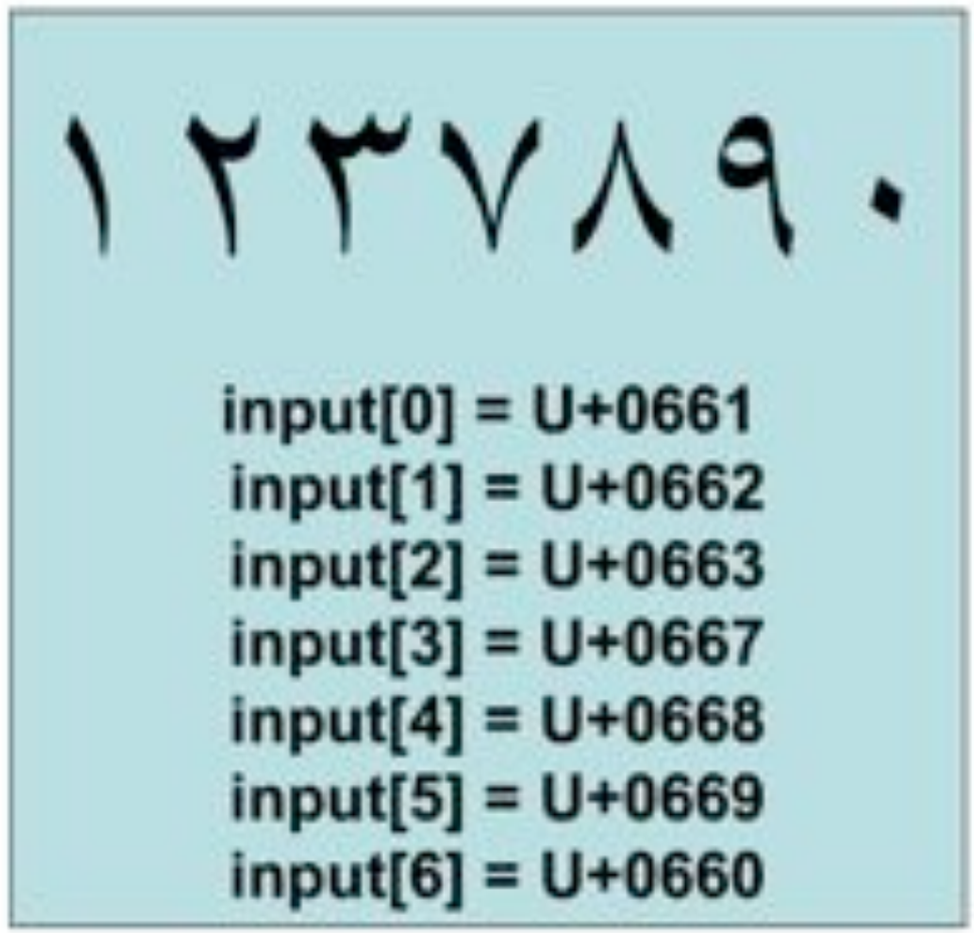
When 1 is not 1…

Arabic-Indic VS. Eastern Arabic-Indic digits





≠



The Arabic Language is only a part of the Arabic Script table

Accepted characters for Arabic, Persian, Urdu, and Pashto

In the beginning

3454 Preparation of Internationalized Strings ("stringprep"). P. Hoffman, M. Blanchet. December 2002. (Format: TXT=138684 bytes) (Status: PROPOSED STANDARD)

3490 Internationalizing Domain Names in Applications (IDNA). P. Faltstrom, P. Hoffman, A. Costello. March 2003. (Format: TXT=51943 bytes) (Status: PROPOSED STANDARD)

3491 Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN). P. Hoffman, M. Blanchet. March 2003. (Format: TXT=10316 bytes) (Status: PROPOSED STANDARD)

3492 Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA). A. Costello. March 2003. (Format: TXT=67439 bytes) (Status: PROPOSED STANDARD)

What is this?

- 3454 Specifies overall algorithm - stringprep
- 3490 Specifies IDN algorithm - IDNA
- 3491 Specifies Nameprep
- 3492 Specifies Punycode

stringprep

- With profiles, any Unicode based string can be converted to another Unicode string so that they can be compared
  - Include illegal codepoints
  - Include mapping table
  - Give ability to create profiles
- Used for IDN, LDAP and other protocols

idna

- Algorithm for how to convert a domain name with Unicode codepoints to ascii
- How to use the stringprep profile and unicode
- Includes specification on how to handle unallocated codepoints
- “core” to IDN standard

nameprep

- Specific stringprep profile for unicode based domain names
- Convert a domain name with unicode codepoints to one of
  - Illegal domain name
  - Domain name with Unicode codepoints

punycode

- Converts a label with unicode codepoints to a domain name in ascii
- Example:
  - fältström
  - xn--fltstrm-5wa1o

What happened?

4690 Review and Recommendations for Internationalized Domain Names (IDNs). J. Klensin, P. Faltstrom, C. Karp, IAB. September 2006. (Format: TXT=100929 bytes) (Status: INFORMATIONAL)

In short...

- Explains the problems in the earlier standards
  - Bidirectional scripts
  - Non-spacing codepoints
- Explains the problems with scripts not yet created when IDNA was written
- Explains problem with versioning of Unicode
  - Old standard based on Unicode 3.2

Example

- If a label include a character that has right to left directionality, both first and last character of the string has to have right to left directionality
- Creates problem if for example the string ends with a codepoint with no directionality

יְיוָאָ

U+05D9 HEBREW LETTER YOD (R)

U+05D9 HEBREW LETTER YOD (R)

U+05B4 HEBREW POINT HIRIQ (NSM)

U+05D5 HEBREW LETTER VAV (R)

U+05D5 HEBREW LETTER VAV (R)

U+05D0 HEBREW LETTER ALEF (R)

U+05B8 HEBREW POINT QAMATS (NSM)

- Note that last codepoint has no directionality (Non Spacing Mark)

U+05D9 HEBREW LETTER YOD (R)

U+05D9 HEBREW LETTER YOD (R)

U+05B4 HEBREW POINT HIRIQ (NSM)

U+05D5 HEBREW LETTER VAV (R)

U+05D5 HEBREW LETTER VAV (R)

U+05D0 HEBREW LETTER ALEF (R)

U+05B8 HEBREW POINT QAMATS (NSM)

- Note that last codepoint has no directionality (Non Spacing Mark)

New IDN standard

- Will consist of a few documents
- Will not change punycode
- Backward compatible

New documents

Current versions

- draft-ietf-idnabis-rationale-08
- draft-ietf-idnabis-protocol-10
- draft-ietf-idna-bidi-02
- draft-ietf-idnabis-tables-04

draft-ietf-idnabis-rationale

- In fact named “Rationale and issues...”
- Addresses the concerns in the IAB document RFC 4690
- Explain how the issues are resolved

draft-ietf-idnabis-protocol

- Replaces the IDNA specification
- Core specification of new IDN standard

draft-ietf-idna-bidi

- Gives specifics for bidirectional scripts

draft-ietf-idnabis-tables

- Defines algorithm to use to calculate whether a codepoint in Unicode is in one of the categories
  - PVALID (Protocol Valid)
  - CONTEXTO / CONTEXTJ
  - DISALLOWED
  - UNASSIGNED

But IDNA2003 had mappings

- Mappings are not part of IDNA200x
- Labels MUST be stable under NFC
- Codepoints in label MUST pass bidi requirements
- Codepoints MUST be ok according to algorithm specified in tables document (which might include contextual rules)
- We MIGHT see a separate document on mapping, recommended behaviour for different applications etc

Why is this needed?

- IDNA standard must be independent of Unicode version
- IDNA standard must handle bidirectional scripts
- ...plus other things mentioned in RFC 4690

When will it be ready?

- "With in 6 months"
- Seriously: Request to people to write code based on the new standards. Last round of very careful review. Should go to official IETF review process during 2009. Last(?) wg meeting at IETF in San Francisco (March 23-27).
- Mailing list: idna-update@alvestrand.no

Patrik Fältström

paf@cisco.com