بسم اللہ الرحمن الرحیم

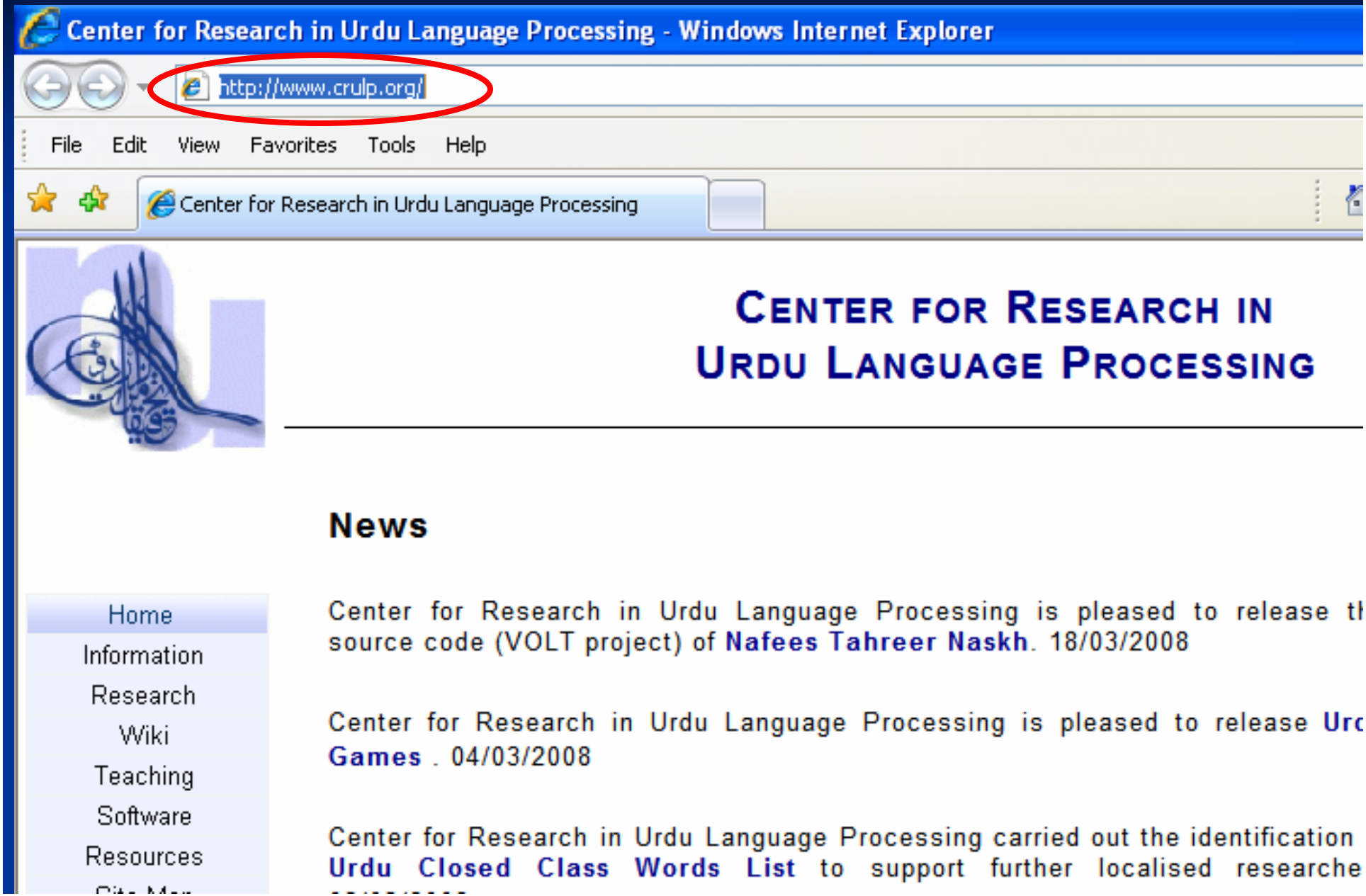
# پاکستانی زبانوں میں انٹرنیٹ کا پتہ ۔

# Domain Name in Pakistani Languages

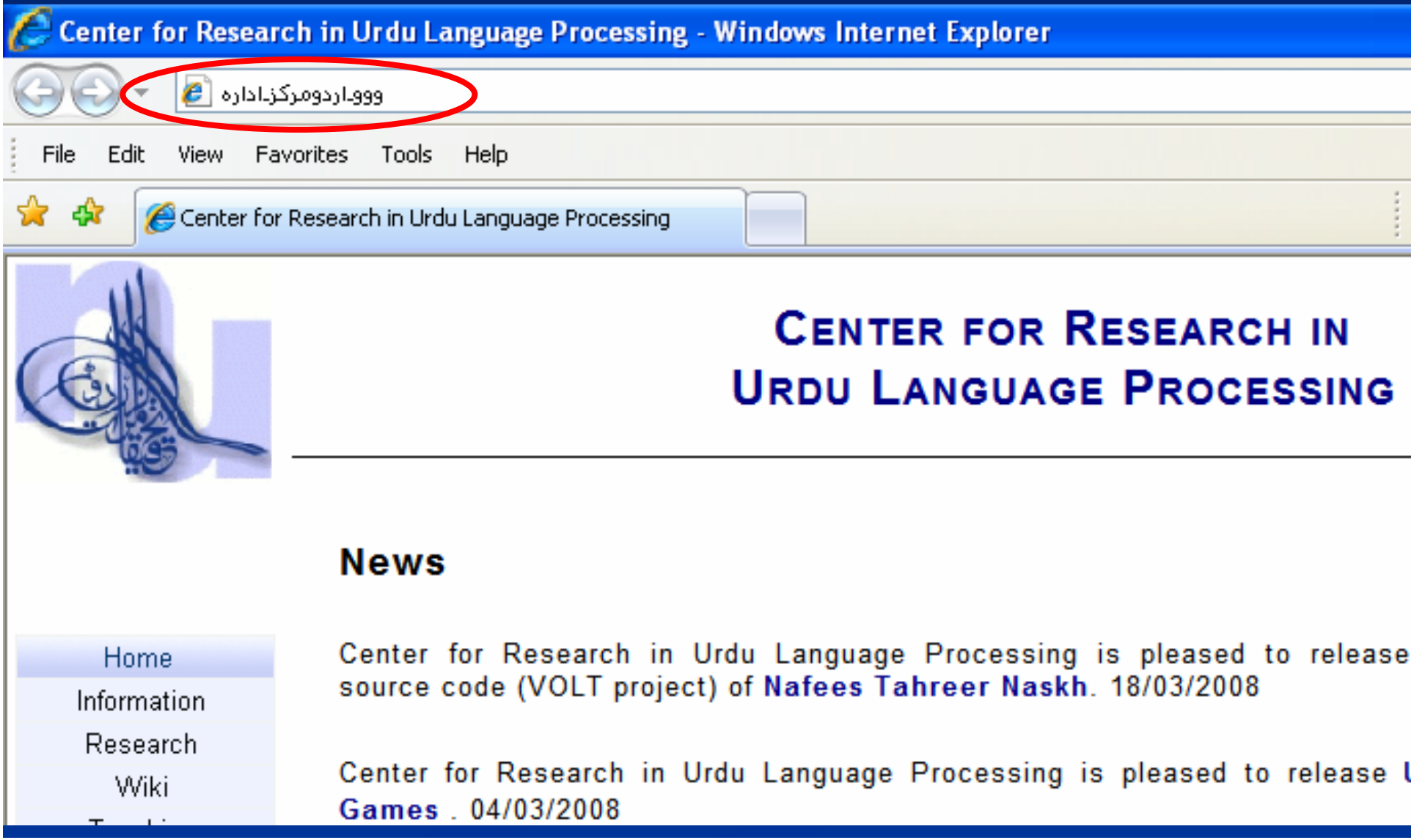
سرمد حسین، نیرہ کرامت، عرفان منصور

نیشنل یونیورسٹی آف کمپیوٹر اینڈ ایمرجنگ سائنسز، فاسٹ

# Domain Name


Center for Research in Urdu Language Processing - Windows Internet Explorer
http://www.crulp.org/
File Edit View Favorites Tools Help
Center for Research in Urdu Language Processing
CENTER FOR RESEARCH IN
URDU LANGUAGE PROCESSING
News
Home
Information
Research
Wiki
Teaching
Software
Resources
Center for Research in Urdu Language Processing is pleased to release th
source code (VOLT project) of Nafees Tahreer Naskh. 18/03/2008
Center for Research in Urdu Language Processing is pleased to release Urc
Games . 04/03/2008
Center for Research in Urdu Language Processing carried out the identification
Urdu Closed Class Words List to support further localised researche

# Internationalized Domain Name



# What letters of Pakistani Languages should be allowed in the Internationalized Domain Names (IDNs)?

- For each language?
- Collectively?

# Morning Session

- Background: Unicode
- Internationalized Domain Names (IDNs)
- Issues and challenges related to Arabic IDNs
- Sample (tentative) solution for Urdu language

# Afternoon Session

- Sample language tables for the following languages
  - Balochi
  - Pashto
  - Punjabi
  - Seraiki
  - Sindhi
  - Torwali
- Collective Issues for multiple languages

# Background: Unicode

- Everything in the computers is represented as numbers
- Initially ASCII encoding:
  - A → 65
  - B → 66 …
- Only supported Latin script, primarily English
- Other encodings developed for other languages, but cumbersome to develop separate encoding for each language of the world

# Unicode

- Thus effort started to develop Universal encoding or **UNIcode**
- Unicode Consortium develops the Unicode standard
- Covers almost all writing systems in current use today
- First version '*The Unicode Standard 1.0*' published in 1991
- Current version '*The Unicode Standard 5.1*' published in April 2008

# Unicode

- European scripts
  - Latin, Greek, Cyrillic, Armenian, Georgian, IPA
- Bidirectional (Middle Eastern) scripts
  - Hebrew, Arabic, Syriac, Thaana
- Indic (Indian and Southeast Asian) scripts
  - Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Thai, Lao, Khmer, Myanmar, Tibetan, Philippine
- East Asian scripts
  - Chinese (Han) characters, Japanese (Hiragana and Katakana), Korean (Hangul), Yi

# Unicode

- Other modern scripts
  - Mongolian, Ethiopic, Cherokee, Canadian Aboriginal
- Historical scripts
  - Runic, Ogham, Old Italic, Gothic, Deseret
- Punctuation and symbols
  - Numerals, math symbols, scientific symbols, arrows, blocks, geometric shapes, Braille, musical notation, etc.

# Unicode is SCRIPT based

- One code per character per script
  - To avoid duplication of codes of same letter used by multiple scripts
  - For example:
    - The character code U+06A9 ک is same in Urdu, Sindhi, Pashto, Punjabi, Farsi, …
- Different code blocks reserved for different scripts
- For Arabic script 0600, 0601, …, 06FE, 06FF

# Characters Semantics

- The Unicode standard includes an extensive database that specifies a large number of *character properties*, including:
  - Name
  - Type (e.g., letter, digit, punctuation mark)
  - Decomposition
  - Case and case mappings (for cased letters)
  - Numeric value (for digits and numerals)
  - Combining class (for combining characters)
  - Cursive joining behavior

# Unicode

- Adopted by industry leaders as Apple, HP, IBM, Microsoft, etc.
- Supported in many platforms including Java, Linux and Microsoft Windows, etc.
- Supported by many internationalized applications including Open Office, Firefox, Thunderbird, Microsoft Office, etc.

# Unicode is the basis for Internationalized Domain Names

# Morning Session

- ✓ Background: Unicode
- Internationalized Domain Names (IDNs)
- Issues and challenges related to Arabic IDNs
- Sample (tentative solution) for Urdu language

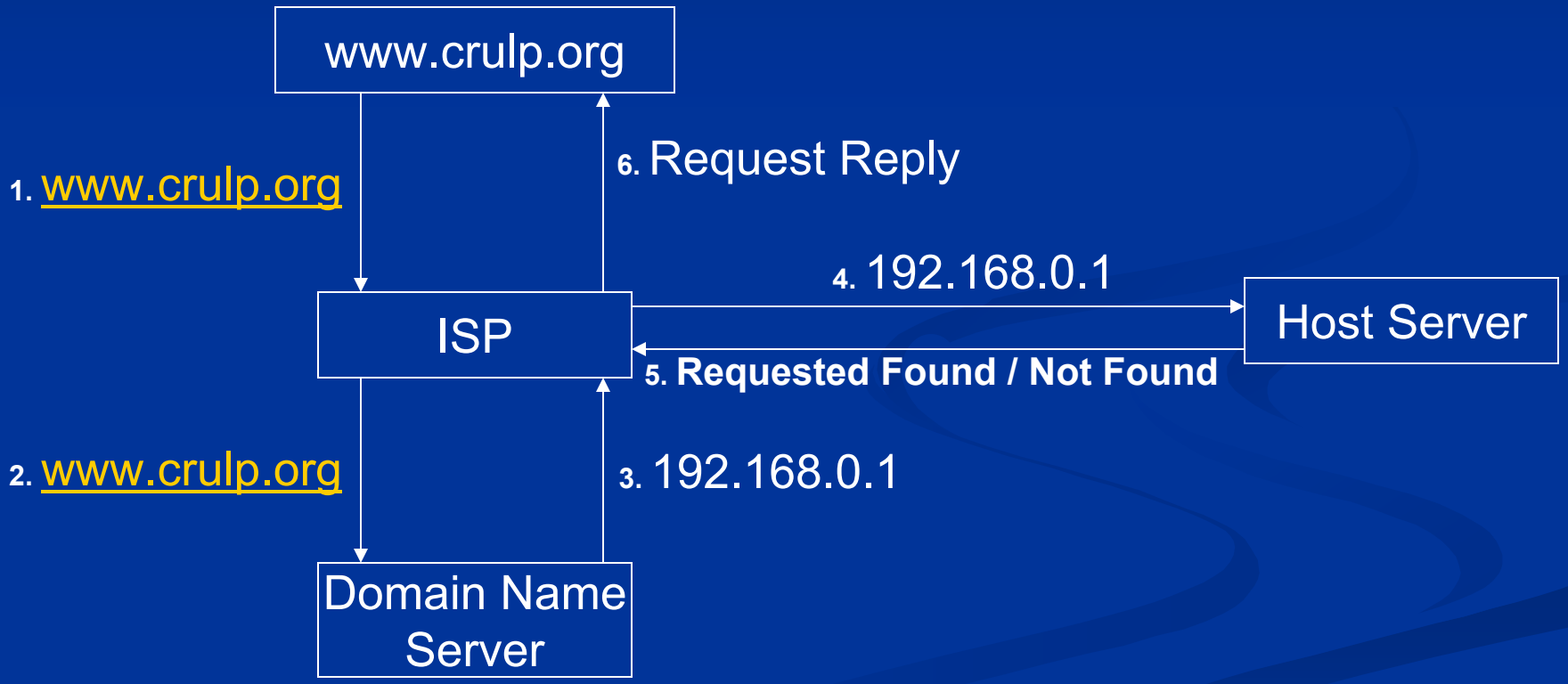
# Domain Name System (DNS)

# Domain Name System (DNS)

- Domain name is the address of a website which is used to access it

  e.g. www.crulp.org

# Domain Name System (DNS)




www.crulp.org = 192.168.0.1

# Need of IDNs

- Domain name system (DNS) is in ASCII, i.e. Latin script
- Makes it difficult to access internet for people who do not understand English or Latin script

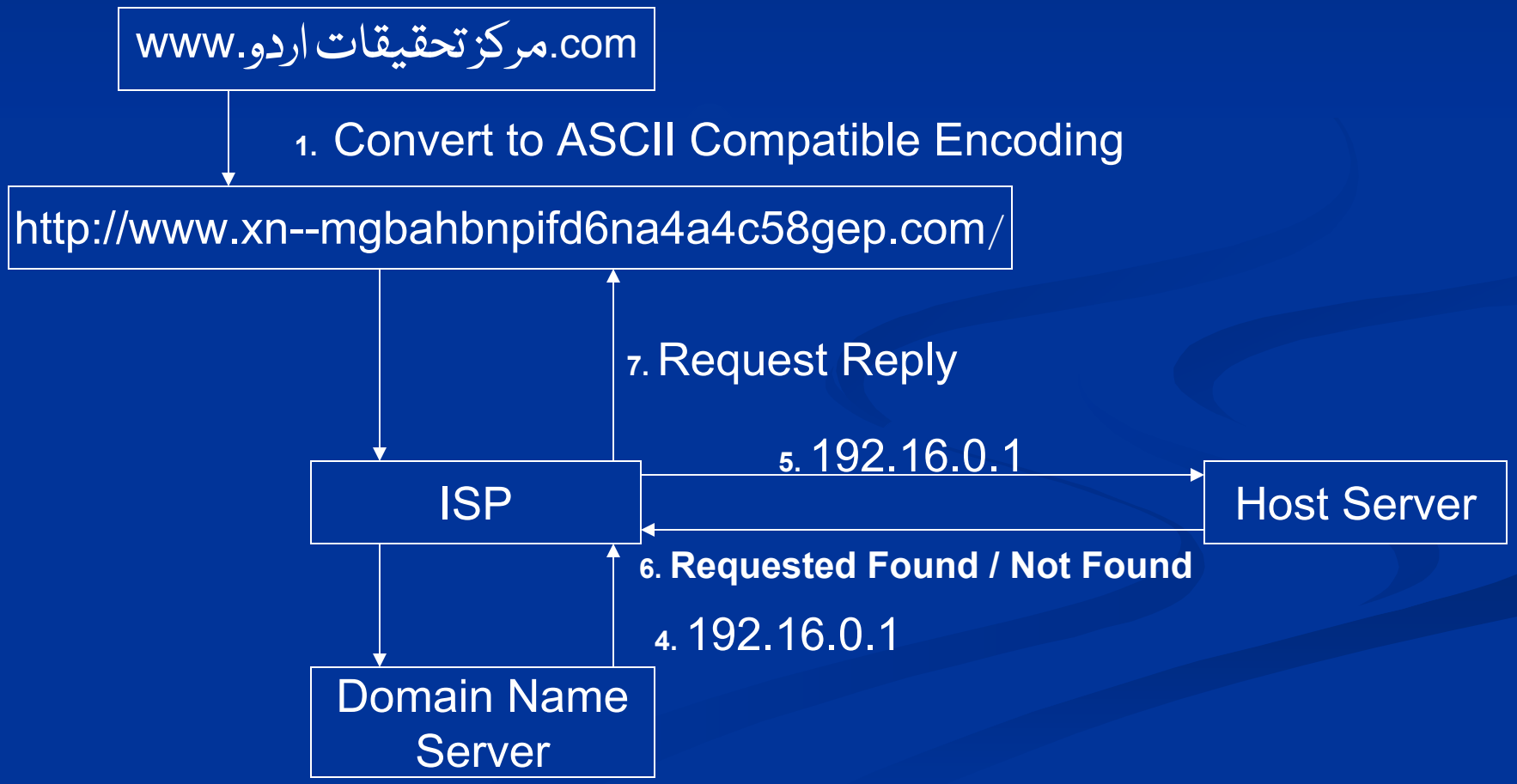
# IDNs

- Basic reason for that is the internet addresses map into 7-bit ASCII standard
- We can not change the overall existing system.
- The solution is to add layer that works on top of existing system
- IDN is any domain name consisting of labels which can be converted to ASCII format
- Initial set of protocols defined in 2003

# IDNs

- A layer that takes the address in local languages and converts that into ASCII format
- DNS continues to resolve ASCII format addresses
- IDNs may be resolved at the User's computer

# International Domain Name in Applications (IDNA)



# IDNA 200X

- Some Issues observed in the original IDNA2003 protocol
  - Dependence on Unicode ver. 3.2
  - Hardcoded language specific separators
  - …
- Decision to revise the original standard taken in 2006
- New standard, IDNA 200X currently under development

# IDNA 200X

- Assigns values to all Unicode Character Database (UCD) on the basis of Unicode properties
  - VALID (or allowed)
  - DISALLOWED
  - CONTEXTO or CONTEXTJ (depends on the context)

# Morning Session

- ✓ Background: Unicode
- ✓ Internationalized Domain Names (IDNs)
- Arabic IDNs
- Sample (tentative) solution for Urdu language

# Arabic Script

- Arabic script is the second largest script after Latin script.
- It is used for writing Arabic, Urdu, Persian, Baluchi, Pashto, Sindhi and many other languages across Pakistan and the world

# Arabic Script

- Arabic script is defined from:
  - U+0600 to U+06FF
  - U+0750 to U+077F
  - U+FB50 to U+FDFF (Obsolete presentation forms)
  - U+FE70 to U+FEFF (Obsolete presentation forms except U+FDFx sequence)

# Arabic Script

- Cursive script
  - Shape of each letter may have four different shapes depending on its position (isolated, initial, medial or final)
- Written from right to left
  - But Numerals written left to right

# Arabic Script

- Diacritics (optionally) used for vowels
- Stretched shapes used for text justification
- Shapes of letters highly context sensitive

# Positional Shapes of Different Letters

| Isolated | Initial | Medial | Final |
|---|---|---|---|
| ب | با | کبا | کب |
| چ | چا | کچا | کچ |
| و | NA | NA | کو |

# Issues in Arabic Script Encoding

- Similar character shapes across the languages that creates confusions e.g. Urdu character ی and Pashto character ی have similar shapes
- These different shapes are used as distinct letters in different languages. In Sindhi these are two different characters: ک (U+06AA), and ک (U+06A9)

# Confusable Variants of Different Characters

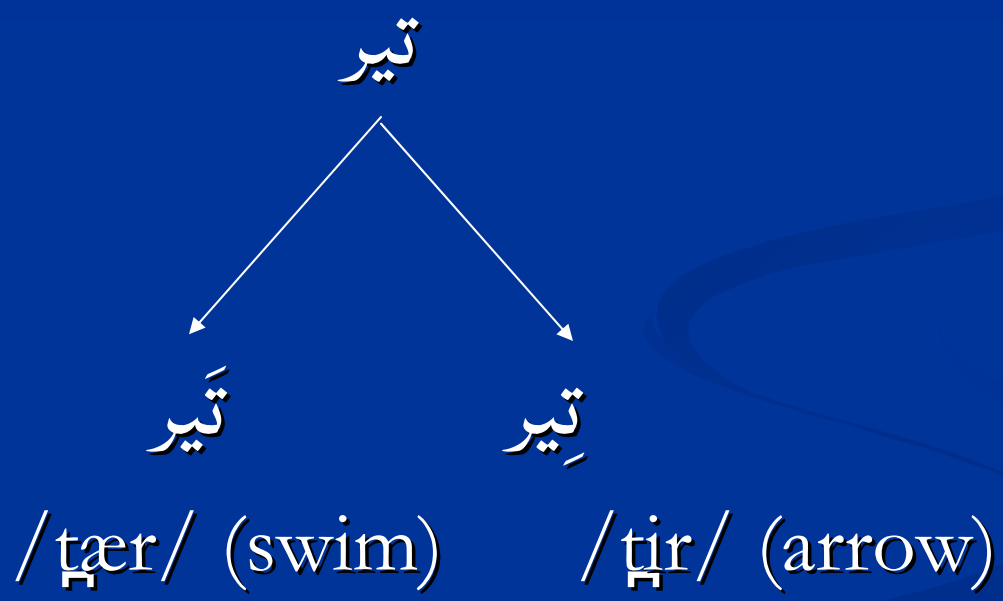
| Unicode | Isolated form | Initial form | Medial form | Final form | Remarks |
|---|---|---|---|---|---|
| U+0643 | ك | كـ | ـكـ | ـك | ARABIC LETTER KAF |
| U+06A9 | ک | کـ | ـکـ | ـک | ARABIC LETTER KEHEH (used for Persian and Urdu) |
| U+06AA | ڪ | ڪـ | ـڪـ ـا | ـڪ | ARABIC LETTER SWASH KAF (used for Sindhi) |

# Confusable Variants of Different Characters

| Unicode | Isolated form | Initial form | Medial form | Final form | Remarks |
|---|---|---|---|---|---|
| U+064A | ي | يـ | ـيـ | ـي | ARABIC LETTER YEH |
| U+06CC | ی | یـ | ـیـ | ـی | ARABIC LETTER FARSI YEH (Arabic, Persian, Urdu) |
| U+0649 | ى | | ـىـ | ى | ARABIC LETTER ALEF MAKSURA |

# Optional Diacritics

- Words normally written without diacritics, e.g. in Urdu:



# Space

- No concept of space between words in Urdu
- Need a separator character
  - words may assume wrong shapes without a separator e.g. دس دن will be displayed erroneously دسدن without separator
- One solution is Zero With Non Joiner (ZWNJ) but users unfamiliar with it.

# Bidirectionality

- مرکز تحقیقات اردو کی بنیاد مارچ 2001 کو رکھی گئی۔

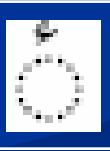
# Normalization

- There are characters that can be typed in more than one way, e.g.

  U+0622 (آ) = U+0627 (ا) + U+0653

- We have to normalize these characters

# Normalization

| Composed Form | Decomposed Form |
|---|---|
| U+0622 (آ) | U+0627 (ا) + U+0653 |
| U+0623 (أ) | U+0627 (ا) + U+0654 |
| U+0624 (ؤ) | U+0648 (و) + U+0653 |
| U+0625 (إ) | U+0627 (ا) + U+0655 |
| U+0626 (ئ) | U+064A (ي) + U+0654  |
| U+0675 (ٵ) | U+0627 (ا) + U+0674 |

# Normalization

| Composed Form | Decomposed Form |
|---|---|
| U+0676 (ٶ) | U+0648 (و) + U+0674 |
| U+0677 (ٷ) | U+06C7 (ۇ) + U+0674 |
| U+0678 (ٸ) | U+064A (ي) +U+0674 |
| U+06C0 (ۀ) | U+06D5 (ە ) + U+0654 |
| U+06C2 (ۂ) | U+06C1 (ہ) + U+0654 |
| U+06D3 (ۓ) | U+06D2 (ے) + U+0654 |

# Confusable Characters

| | | |
|---|---|---|
| ك (U+0643) | ک (U+06A9) | ک (U+06AA) |
| ی (U+06CC) | ی (U+0649) | ۍ (U+06CD) |
| ه (U+0647) | ہ (U+06C1) | ە (U+06D5) |

# Confusable Characters…

| | |
|---|---|
| ٠ (U+0660) | ۰ (U+06F0) |
| ١ (U+0661) | ۱ (U+06F1) |
| ٢ (U+0662) | ۲ (U+06F2) |
| ٣ (U+0663) | ۳ (U+06F3) |
| ٥ (U+0665) | ۵ (U+06F5) |
| ٧ (U+0667) | ۷ (U+06F7) |
| ٨ (U+0668) | ۸ (U+06F8) |
| ٩ (U+0669) | ۹ (U+06F9) |
| أ (U+0623) | ٱ (U+0672) |
| إ (U+0625) | ٳ (U+0673) |

# Morning Session

- ✓ Background: Unicode
- ✓ Internationalized Domain Names (IDNs)
- ✓ Arabic IDNs
- ■ Sample (tentative) solution for Urdu language

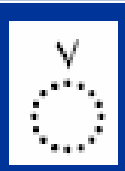
# Urdu IDNs

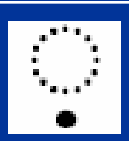
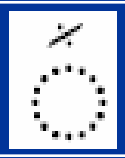
- The following are CONTEXTO by IDNA200X but are not recommended for Urdu

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 0600 | ARABIC NUMBER SIGN | CONTEXTO | NO |
| | 0601 | ARABIC SIGN SANAH | CONTEXTO | NO |
| | 0602 | ARABIC FOOTNOTE MARKER | CONTEXTO | NO |
| | 0603 | ARABIC SIGN SAFHA | CONTEXTO | NO |
| | 06DD | END OF AYAH | CONTEXTO | NO |

# Urdu IDNs

- The following are PVALID by IDNA200X but these are not recommended for Urdu

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ٱ | 0615 | ARABIC SMALL HIGH TAH | PVALID | NO |
| ـ | 0640 | ARABIC TATWEEL | PVALID | NO |
| ٗ | 0657 | ARABIC INVERTED DAMMA | PVALID | NO |
| | 0659 | ARABIC ZWARAKAY | PVALID | NO |
|  | 065A | ARABIC VOWEL SIGN SMALL V ABOVE | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 065B | ARABIC VOWEL SIGN INVERTED SMALL V ABOVE | PVALID | NO |
|  | 065C | ARABIC VOWEL SIGN DOT BELOW | PVALID | NO |
| | 065D | ARABIC REVERSED DAMMA | PVALID | NO |
|  | 065E | ARABIC FATHA WITH TWO DOTS | PVALID | NO |
| ٱ | 0671 | ARABIC LETTER ALEF WASLA | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| أ | 0672 | ARABIC LETTER ALEF WITH WAVY HAMZA ABOVE | PVALID | NO |
| إ | 0673 | ARABIC LETTER ALEF WITH WAVY HAMZA BELOW | PVALID | NO |
| ء | 0674 | ARABIC LETTER HIGH HAMZA | PVALID | NO |
| ٹ | 067A | ARABIC LETTER TTEHEH | PVALID | NO |
| ٻ | 067B | ARABIC LETTER BEEH | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ٹ | 067C | ARABIC LETTER TEH WITH RING | PVALID | NO |
| ٹ | 067D | ARABIC LETTER TEH WITH THREE DOTS ABOVE DOWN | PVALID | NO |
| ٹ | 067F | ARABIC LETTER TEHEH | PVALID | NO |
| ٹ | 0680 | ARABIC LETTER BEHEH | PVALID | NO |
| ٹ | 0681 | ARABIC LETTER HAH WITH HAMZA ABOVE | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڂ | 0682 | ARABIC LETTER HAH WITH TWO DOTS VERTICAL ABO | PVALID | NO |
| ڃ | 0683 | ARABIC LETTER NYEH | PVALID | NO |
| ڄ | 0684 | ARABIC LETTER DYEH | PVALID | NO |
| څ | 0685 | ARABIC LETTER HAH WITH THREE DOTS ABOVE | PVALID | NO |
| ڇ | 0687 | ARABIC LETTER TCHEHEH | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڈ | 0689 | ARABIC LETTER DAL WITH RING | PVALID | NO |
| ڊ | 068A | ARABIC LETTER DAL WITH DOT BELOW | PVALID | NO |
| ڋ | 068B | ARABIC LETTER DAL WITH DOT BELOW AND SMALL T | PVALID | NO |
| ڌ | 068C | ARABIC LETTER DAHAL | PVALID | NO |
| ڍ | 068D | ARABIC LETTER DDAHAL | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڈ | 068E | ARABIC LETTER DUL | PVALID | NO |
| ډ | 068F | ARABIC LETTER DAL WITH THREE DOTS ABOVE DOWN | PVALID | NO |
| ڊ | 0690 | ARABIC LETTER DAL WITH FOUR DOTS ABOVE | PVALID | NO |
| ڑ | 0692 | ARABIC LETTER REH WITH SMALL V | PVALID | NO |
| ړ | 0693 | ARABIC LETTER REH WITH RING | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڔ | 0694 | ARABIC LETTER REH WITH DOT BELOW | PVALID | NO |
| ڕ | 0695 | ARABIC LETTER REH WITH SMALL V BELOW | PVALID | NO |
| ږ | 0696 | ARABIC LETTER REH WITH DOT BELOW AND DOT ABO | PVALID | NO |
| ڗ | 0697 | ARABIC LETTER REH WITH TWO DOTS ABOVE | PVALID | NO |
| ڙ | 0699 | ARABIC LETTER REH WITH FOUR DOTS ABOVE | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ښ | 069A | ARABIC LETTER SEEN WITH DOT BELOW AND DOT AB | PVALID | NO |
| ڀ | 069B | ARABIC LETTER SEEN WITH THREE DOTS BELOW | PVALID | NO |
| ڃ | 069C | ARABIC LETTER SEEN WITH THREE DOTS BELOW AND | PVALID | NO |
| ڇ | 069D | ARABIC LETTER SAD WITH TWO DOTS BELOW | PVALID | NO |
| ڈ | 069E | ARABIC LETTER SAD WITH THREE DOTS ABOVE | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ظ | 069F | ARABIC LETTER TAH WITH THREE DOTS ABOVE | PVALID | NO |
| ۼ | 06A0 | ARABIC LETTER AIN WITH THREE DOTS ABOVE | PVALID | NO |
| ف | 06A1 | ARABIC LETTER DOTLESS FEH | PVALID | NO |
| ڤ | 06A2 | ARABIC LETTER FEH WITH DOT MOVED BELOW | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڤ | 06A3 | ARABIC LETTER FEH WITH DOT BELOW | PVALID | NO |
| ڤ | 06A4 | ARABIC LETTER VEH | PVALID | NO |
| ڥ | 06A5 | ARABIC LETTER FEH WITH THREE DOTS BELOW | PVALID | NO |
| ڦ | 06A6 | ARABIC LETTER PEHEH | PVALID | NO |
| ڧ | 06A7 | ARABIC LETTER QAF WITH DOT ABOVE | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڧ | 06A8 | ARABIC LETTER QAF WITH THREE DOTS ABOVE | PVALID | NO |
| ک | 06AA | ARABIC LETTER SWASH KAF | PVALID | NO |
| ګ | 06AB | ARABIC LETTER KAF WITH RING | PVALID | NO |
| ڭ | 06AC | ARABIC LETTER KAF WITH DOT ABOVE | PVALID | NO |
| ڮ | 06AD | ARABIC LETTER NG | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڭ | 06AE | ARABIC LETTER KAF WITH THREE DOTS BELOW | PVALID | NO |
| گ | 06B0 | ARABIC LETTER GAF WITH RING | PVALID | NO |
| گ | 06B1 | ARABIC LETTER NGOEH | PVALID | NO |
| گ | 06B2 | ARABIC LETTER GAF WITH TWO DOTS BELOW | PVALID | NO |

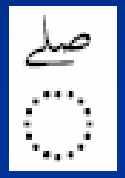
| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڳ | 06B3 | ARABIC LETTER GUEH | PVALID | NO |
| ڴ | 06B4 | ARABIC LETTER GAF WITH THREE DOTS ABOVE | PVALID | NO |
| ڵ | 06B5 | ARABIC LETTER LAM WITH SMALL V | PVALID | NO |
| ڶ | 06B6 | ARABIC LETTER LAM WITH DOT ABOVE | PVALID | NO |
| ڷ | 06B7 | ARABIC LETTER LAM WITH THREE DOTS ABOVE | PVALID | NO |

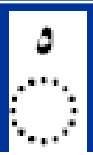
| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڀ | 06B8 | ARABIC LETTER LAM WITH THREE DOTS BELOW | PVALID | NO |
| ڹ | 06B9 | ARABIC LETTER NOON WITH DOT BELOW | PVALID | NO |
| ڻ | 06BB | ARABIC LETTER RNOON | PVALID | NO |
| ڼ | 06BC | ARABIC LETTER NOON WITH RING | PVALID | NO |

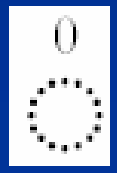
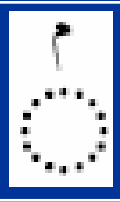
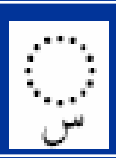
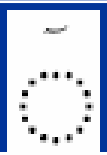
| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ں | 06BD | ARABIC LETTER NOON WITH THREE DOTS ABOVE | PVALID | NO |
| ڿ | 06BF | ARABIC LETTER TCHEH WITH DOT ABOVE | PVALID | NO |
| ۄ | 06C4 | ARABIC LETTER WAW WITH RING | PVALID | NO |
| ۅ | 06C5 | ARABIC LETTER KIRGHIZ OE | PVALID | NO |
| ۆ | 06C6 | ARABIC LETTER OE | PVALID | NO |

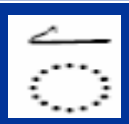
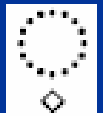
| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ۇ | 06C7 | ARABIC LETTER U | PVALID | NO |
| ۈ | 06C8 | ARABIC LETTER YU | PVALID | NO |
| ۉ | 06C9 | ARABIC LETTER KIRGHIZ YU | PVALID | NO |
| ۊ | 06CA | ARABIC LETTER WAW WITH TWO DOTS ABOVE | PVALID | NO |
| ۋ | 06CB | ARABIC LETTER VE | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ۍ | 06CD | ARABIC LETTER YEH WITH TAIL | PVALID | NO |
| ێ | 06CE | ARABIC LETTER YEH WITH SMALL V | PVALID | NO |
| ۏ | 06CF | ARABIC LETTER WAW WITH DOT ABOVE | PVALID | NO |
| ې | 06D0 | ARABIC LETTER E | PVALID | NO |
| ۑ | 06D1 | ARABIC LETTER YEH WITH THREE DOTS BELOW | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ە | 06D5 | ARABIC LETTER AE | PVALID | NO |
|  | 06D6 | ARABIC SMALL HIGH LIGATURE SAD WITH LAM WITH | PVALID | NO |
| | 06D7 | ARABIC SMALL HIGH LIGATURE QAF WITH LAM WITH | PVALID | NO |
| | 06D8 | ARABIC SMALL HIGH MEEM INITIAL FORM | PVALID | NO |
| | 06D9 | ARABIC SMALL HIGH LAM ALEF | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 06DA | ARABIC SMALL HIGH JEEM | PVALID | NO |
| | 06DB | ARABIC SMALL HIGH THREE DOTS | PVALID | NO |
| | 06DC | ARABIC SMALL HIGH SEEN | PVALID | NO |
|  | 06DF | ARABIC SMALL HIGH ROUNDED ZERO | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
|  | 06E0 | ARABIC SMALL HIGH UPRIGHT RECTANGULAR ZERO | PVALID | NO |
| | 06E1 | ARABIC SMALL HIGH DOTLESS HEAD OF KHAH | PVALID | NO |
|  | 06E2 | ARABIC SMALL HIGH MEEM ISOLATED FORM | PVALID | NO |
|  | 06E3 | ARABIC SMALL LOW SEEN | PVALID | NO |
|  | 06E4 | ARABIC SMALL HIGH MADDA | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 06E5 | ARABIC SMALL WAW | PVALID | NO |
| | 06E6 | ARABIC SMALL YEH | PVALID | NO |
|  | 06E7 | ARABIC SMALL HIGH YEH | PVALID | NO |
| | 06E8 | ARABIC SMALL HIGH NOON | PVALID | NO |
|  | 06EA | ARABIC EMPTY CENTRE LOW STOP | PVALID | NO |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 06EB | ARABIC EMPTY CENTRE HIGH STOP | PVALID | NO |
| | 06EC | ARABIC ROUNDED HIGH STOP WITH FILLED CENTRE | PVALID | NO |
| | 06ED | ARABIC SMALL LOW MEEM | PVALID | NO |
| | 06EE | ARABIC LETTER DAL WITH INVERTED V | PVALID | NO |
| | 06EF | ARABIC LETTER REH WITH | PVALID | NO |
| | 06FF | ARABIC LETTER HEH WITH INVERTED V | PVALID | NO |

# Urdu Language Table For IDNs

- Following characters are DISALLOWED by IDNA 200X and are not required

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 060B | AFGHANI SIGN | DISALLOWED | N0 |
| | 060C | ARABIC COMMA | DISALLOWED | N0 |
| | 060D | ARABIC DATE SEPARATOR | DISALLOWED | N0 |
| | 060E | ARABIC POETIC VERSE SIGN | DISALLOWED | N0 |
| | 060F | ARABIC SIGN MISRA | DISALLOWED | N0 |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 061B | ARABIC SEMICOLON | DISALLOWED | N0 |
| | 061E | ARABIC TRIPLE DOT PUNCTUATION MARK | DISALLOWED | N0 |
| | 061F | ARABIC QUESTION MARK | DISALLOWED | N0 |
| | 066A | ARABIC PERCENT SIGN | DISALLOWED | N0 |
| | 066B | ARABIC DECIMAL SEPARATOR | DISALLOWED | N0 |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ٬ | 066C | ARABIC THOUSANDS SEPARATOR | DISALLOWED | N0 |
| ٭ | 066D | ARABIC FIVE POINTED STAR | DISALLOWED | N0 |
| | 0675 | ARABIC LETTER HIGH HAMZA ALEF | DISALLOWED | N0 |
| | 0676 | ARABIC LETTER HIGH HAMZA WAW | DISALLOWED | N0 |
| | 0677 | ARABIC LETTER U WITH HAMZA ABOVE | DISALLOWED | N0 |
| | 0678 | ARABIC LETTER HIGH HAMZA YEH | DISALLOWED | N0 |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ۔ | 06D4 | ARABIC LETTER HAH WITH HAMZA ABOVE | DISALLOWED | N0 |
| | 06DE | ARABIC START OF RUB EL HIZB | DISALLOWED | N0 |
| | 06E9 | ARABIC PLACE OF SAJDAH | DISALLOWED | N0 |
| ؏ | 06FD | ARABIC SIGN SINDHI AMPERSAND | DISALLOWED | N0 |
| ﷲ | FDF2 | ARABIC LIGATURE ALLAH ISOLATED FORM | DISALLOWED | N0 |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| اكبر | FDF3 | ARABIC LIGATURE AKBAR ISOLATED FORM | DISALLOWED | N0 |
| محمد | FDF4 | ARABIC LIGATURE MOHAMMAD ISOLATED FORM | DISALLOWED | N0 |
| | FDF5 | ARABIC LIGATURE SALAM ISOLATED FORM | DISALLOWED | N0 |
| رسول | FDF6 | ARABIC LIGATURE RASOUL ISOLATED FORM | DISALLOWED | N0 |
| عليه | FDF7 | ARABIC LIGATURE ALAYHE ISOLATED FORM | DISALLOWED | N0 |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | FDF8 | ARABIC LIGATURE WASALLAM<br>ISOLATED FORM | DISALLOWED | N0 |
| | FDF9 | ARABIC LIGATURE SALLA ISOLATED<br>FORM | DISALLOWED | N0 |
| | FDFA | ARABIC LIGATURE SALLALLAHOU ALAYHE<br>WASALLAM | DISALLOWED | N0 |
| ﷻ | FDFB | ARABIC LIGATURE<br>JALLAJALALOUHOU | DISALLOWED | N0 |

# Urdu IDNs

- Following are PVALID characters by IDNA200X decision and are required for Urdu

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 0610 | ARABIC SIGN SALLALLAHOU ALAYHE WASSALLAM | PVALID | YES |
| | 0611 | ARABIC SIGN ALAYHE ASSALLAM | PVALID | YES |
| | 0612 | ARABIC SIGN RAHMATULLAH ALAYHE | PVALID | YES |
| | 0613 | ARABIC SIGN RADI ALLAHOU ANHU | PVALID | YES |
| | 0614 | ARABIC SIGN TAKHALLUS | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ء | 0621 | ARABIC LETTER HAMZA | PVALID | YES |
| آ | 0622 | ARABIC LETTER ALEF WITH MADDA ABOVE | PVALID | YES |
| أ | 0623 | ARABIC LETTER ALEF WITH HAMZA ABOVE | PVALID | YES |
| ؤ | 0624 | ARABIC LETTER WAW WITH HAMZA ABOVE | PVALID | YES |
| إ | 0625 | ARABIC LETTER ALEF WITH HAMZA BELOW | PVALID | YES (Variant of Base Character) |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ئ | 0626 | ARABIC LETTER YEH WITH HAMZA ABOVE | PVALID | YES (Variant of Base Character) |
| ا | 0627 | ARABIC LETTER ALEF | PVALID | YES |
| ب | 0628 | ARABIC LETTER BEH | PVALID | YES |
| ة | 0629 | ARABIC LETTER TEH MARBUTA | PVALID | YES (Variant of Base Character) |
| ت | 062A | ARABIC LETTER TEH | PVALID | YES |
| ث | 062B | ARABIC LETTER THEH | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ج | 062C | ARABIC LETTER JEEM | PVALID | YES |
| ح | 062D | ARABIC LETTER HAH | PVALID | YES |
| خ | 062E | ARABIC LETTER KHAH | PVALID | YES |
| د | 062F | ARABIC LETTER DAL | PVALID | YES |
| ذ | 0630 | ARABIC LETTER THAL | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ر | 0631 | ARABIC LETTER REH | PVALID | YES |
| ز | 0632 | ARABIC LETTER ZAIN | PVALID | YES |
| س | 0633 | ARABIC LETTER SEEN | PVALID | YES |
| ش | 0634 | ARABIC LETTER SHEEN | PVALID | YES |
| ص | 0635 | ARABIC LETTER SAD | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ض | 0636 | ARABIC LETTER DAD | PVALID | YES |
| ط | 0637 | ARABIC LETTER TAH | PVALID | YES |
| ظ | 0638 | ARABIC LETTER ZAH | PVALID | YES |
| ع | 0639 | ARABIC LETTER AIN | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| غ | 063A | ARABIC LETTER GHAIN | PVALID | YES |
| ف | 0641 | ARABIC LETTER FEH | PVALID | YES |
| ق | 0642 | ARABIC LETTER QAF | PVALID | YES |
| ك | 0643 | ARABIC LETTER KAF | PVALID | YES (Variant of Base Character) |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ل | 0644 | ARABIC LETTER LAM | PVALID | YES |
| م | 0645 | ARABIC LETTER MEEM | PVALID | YES |
| ن | 0646 | ARABIC LETTER NOON | PVALID | YES |
| ه | 0647 | ARABIC LETTER HEH | PVALID | YES (Variant of Base Character) |
| و | 0648 | ARABIC LETTER WAW | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ى | 0649 | ARABIC LETTER ALEF MAKSURA | PVALID | YES (Variant of Base Character) |
| ي | 064A | ARABIC LETTER YEH | PVALID | YES (Variant of Base Character) |
| | 064B | ARABIC FATHATAN | PVALID | YES |
| | 064C | ARABIC DAMMATAN | PVALID | YES |
| | 064D | ARABIC KASRATAN | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 064E | ARABIC FATHA | PVALID | YES |
| | 064F | ARABIC DAMMA | PVALID | YES |
| | 0650 | ARABIC KASRA | PVALID | YES |
| | 0651 | ARABIC SHADDA | PVALID | YES |
| | 0652 | ARABIC SUKUN | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| | 0653 | ARABIC MADDAH ABOVE | PVALID | YES |
| | 0654 | ARABIC HAMZA ABOVE | PVALID | YES |
| | 0655 | ARABIC HAMZA BELOW | PVALID | YES |
| | 0656 | ARABIC SUBSCRIPT ALEF | PVALID | YES |
| | 0658 | ARABIC MARK NOON GHUNNA | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ٠ | 0660 | ARABIC-INDIC DIGIT ZERO | PVALID | YES (Variant of Base Character) |
| ١ | 0661 | ARABIC-INDIC DIGIT ONE | PVALID | YES (Variant of Base Character) |
| ٢ | 0662 | ARABIC-INDIC DIGIT TWO | PVALID | YES (Variant of Base Character) |
| ٣ | 0663 | ARABIC-INDIC DIGIT THREE | PVALID | YES (Variant of Base Character) |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ٤ | 0664 | ARABIC-INDIC DIGIT FOUR | PVALID | YES (Variant of Base Character) |
| ٥ | 0665 | ARABIC-INDIC DIGIT FIVE | PVALID | YES (Variant of Base Character) |
| ٦ | 0666 | ARABIC-INDIC DIGIT SIX | PVALID | YES (Variant of Base Character) |
| ٧ | 0667 | ARABIC-INDIC DIGIT SEVEN | PVALID | YES (Variant of Base Character) |
| ٨ | 0668 | ARABIC-INDIC DIGIT EIGHT | PVALID | YES (Variant of Base Character) |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ٩ | 0669 | ARABIC-INDIC DIGIT NINE | PVALID | YES (Variant of Base Character) |
| | 0670 | ARABIC LETTER SUPERSCRIPT ALEF | PVALID | YES |
| ٹ | 0679 | ARABIC LETTER TTEH | PVALID | YES |
| پ | 067E | ARABIC LETTER PEH | PVALID | YES |
| چ | 0686 | ARABIC LETTER TCHEH | PVALID | YES |
| ڈ | 0688 | ARABIC LETTER DDAL | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ڑ | 0691 | ARABIC LETTER RREH | PVALID | YES |
| ژ | 0698 | ARABIC LETTER JEH | PVALID | YES |
| ک | 06A9 | ARABIC LETTER KEHEH | PVALID | YES |
| گ | 06AF | ARABIC LETTER GAF | PVALID | YES |
| ں | 06BA | ARABIC LETTER NOON GHUNNA | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ھ | 06BE | ARABIC LETTER HEH DOACHASHMEE | PVALID | YES |
| ۃ | 06C0 | ARABIC LETTER HEH WITH YEH ABOVE | PVALID | YES (Variant of Base character) |
| ہ | 06C1 | ARABIC LETTER HEH GOA | PVALID | YES |
| ۂ | 06C2 | ARABIC LETTER HEH GOAL WITH HAMZA ABOVE | PVALID | YES |
| ۃ | 06C3 | ARABIC LETTER TEH MARBUTA GOAL | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ی | 06CC | ARABIC LETTER FARSI YEH | PVALID | YES |
| ے | 06D2 | ARABIC LETTER YEH BARREE | PVALID | YES |
| ۓ | 06D3 | ARABIC LETTER YEH BARREE WITH HAMZA ABOVE | PVALID | YES |
| ۰ | 06F0 | EXTENDED ARABIC-INDIC DIGIT ZERO | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ۱ | 06F1 | EXTENDED ARABIC-INDIC DIGIT ONE | PVALID | YES |
| ۲ | 06F2 | EXTENDED ARABIC-INDIC DIGIT TWO | PVALID | YES |
| ۳ | 06F3 | EXTENDED ARABIC-INDIC DIGIT THREE | PVALID | YES |
| ۴ | 06F4 | EXTENDED ARABIC-INDIC DIGIT FOUR | PVALID | YES |
| ۵ | 06F5 | EXTENDED ARABIC-INDIC DIGIT FIVE | PVALID | YES |

| Character | Unicode | Description | Current status in IDNA 200x | Recommen-dation |
|---|---|---|---|---|
| ۶ | 06F6 | EXTENDED ARABIC-INDIC DIGIT SIX | PVALID | YES |
| ۷ | 06F7 | EXTENDED ARABIC-INDIC DIGIT SEVEN | PVALID | YES |
| ۸ | 06F8 | EXTENDED ARABIC-INDIC DIGIT EIGHT | PVALID | YES |
| ۹ | 06F9 | EXTENDED ARABIC-INDIC DIGIT NINE | PVALID | YES |

# Afternoon Session

- Sample language tables for the following languages
  - Balochi
  - Pashto
  - Punjabi
  - Seraiki
  - Sindhi
  - Urdu
  - Torwali
- Collective Issues for multiple languages

# Collective Issues

- Separator
- Diacritics
- Honorifics
- Confusable characters
  - Kaf
  - Yay
  - Hay
  - Gol Tay
  - Others?
- Digits
- Space/ZWNJ
- Blocking vs. Bundling