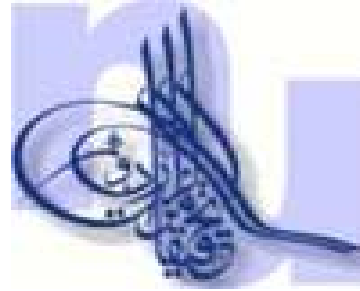


بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ

پاکستانی زبانوں میں انٹرنیٹ کا پتہ -

Domain Names in Pakistani Languages

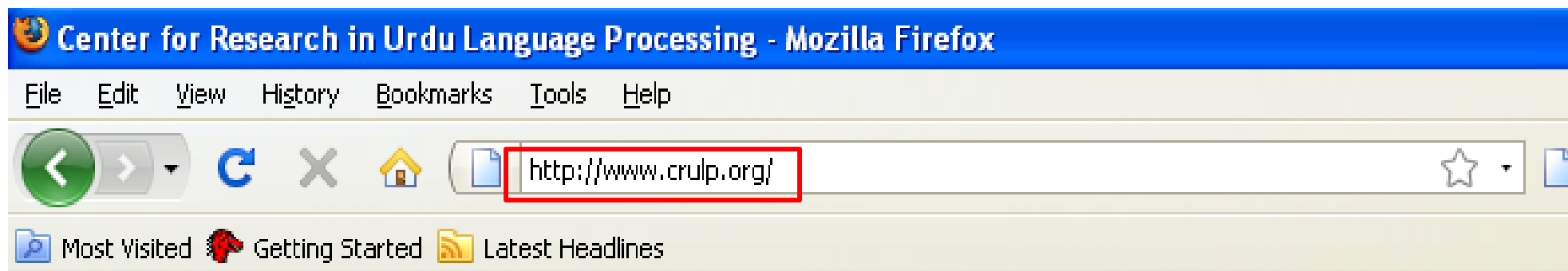


سرمد حسین اور رابعہ سرہندی

نیشنل یونیورسٹی آف کمپیوٹر اینڈ امرنگ سائنسز فاسٹ

# Domain name

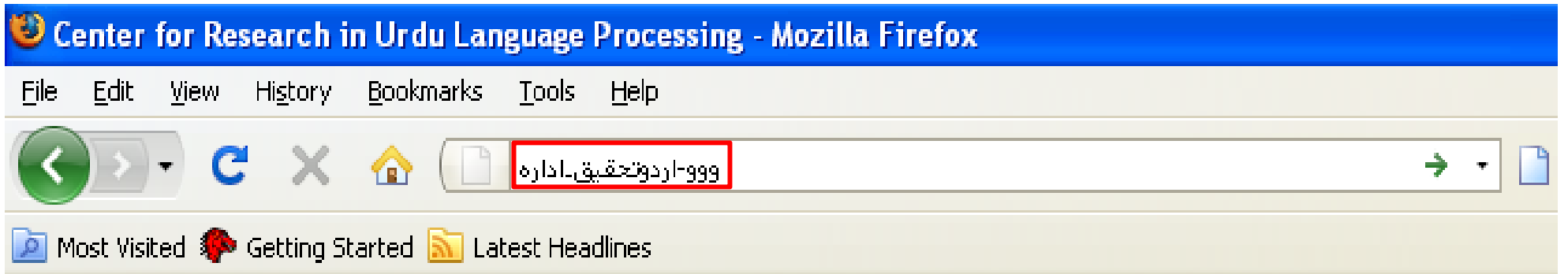
- Domain name is the address of the web page on which the content is located



**CENTER FOR RESEARCH IN  
URDU LANGUAGE PROCESSING**

# Internationalized Domain Name (IDN)

- Domain name or address of the web page in local language is called an IDN
- Based on the Unicode standard



**CENTER FOR RESEARCH IN  
URDU LANGUAGE PROCESSING**

# Morning Session

- Introduction to the Unicode standard
- Introduction to Internationalized Domain Names
- Issues related to IDNs for Pakistani languages

# Afternoon Session

- Exercises and Recommendations
  - Character status revision at script level
  - Resolving confusability of characters
  - Additional composed characters
  - Digits and Mixing
  - Single vs. multiple language tables
  - Character and Label separator
  - ccTLD string
  - gTLD translations

# Background: Unicode

- Everything in 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 **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
- 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

- 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.

# 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 is SCRIPT based

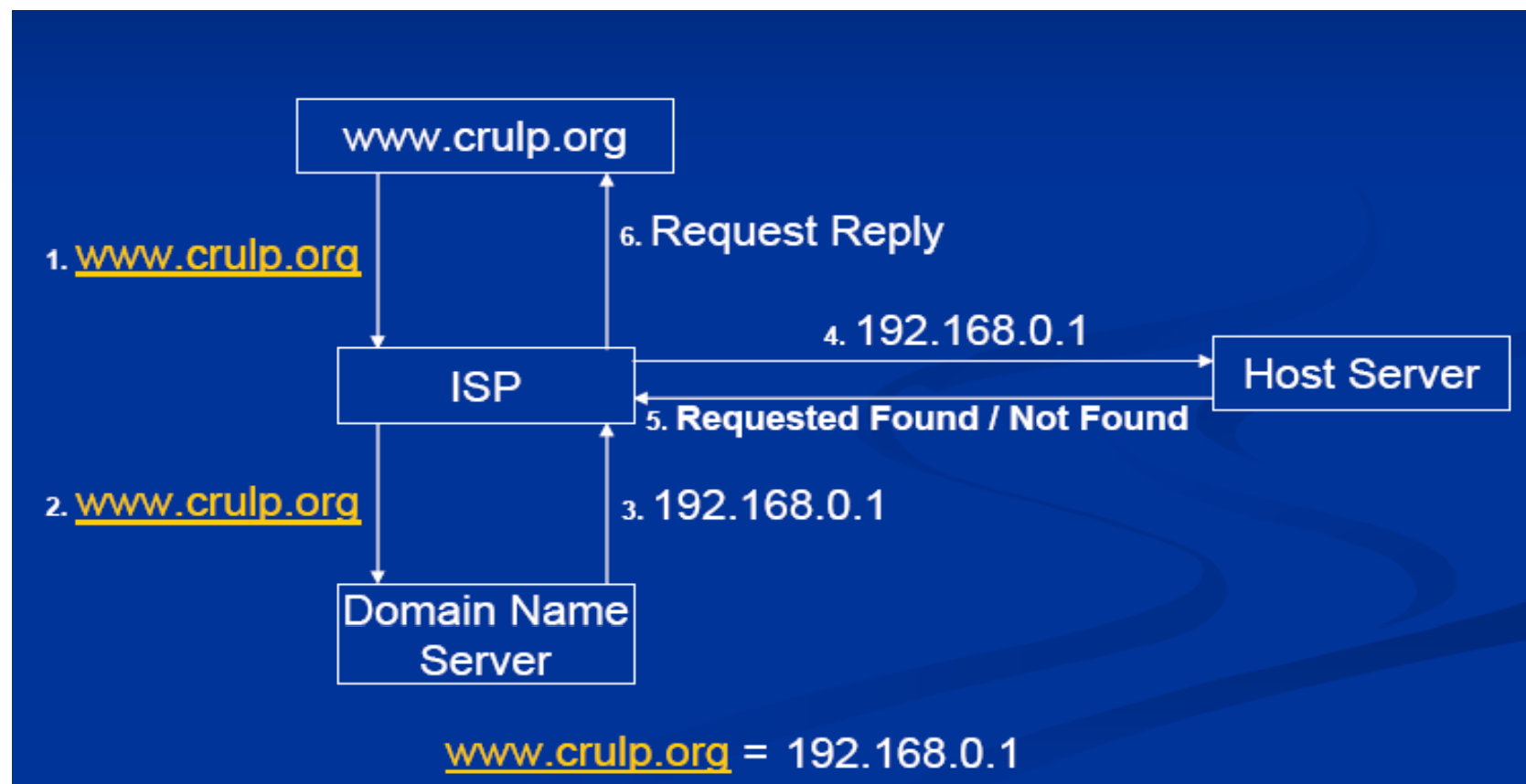
- One code per character per script
  - To avoid duplication of same letter used by multiple languages
  - For example:  
The character code 06A9 ﷥ is same in Urdu, Sindhi, Pashto, Punjabi, Farsi, etc.
- Different code blocks reserved for different scripts
- For Arabic script
  - 0600, 0601, ..., 06FE, 06FF
  - 0750...077F

	060	061	062	063	064	065	066	067	068	069	06A	06B	06C	06D	06E	06F	075	076	077	
0	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
1	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰	۲۱	۲۲	۲۳	۲۴	۲۵	۲۶	۲۷	۲۸	۲۹
2	۲۰	۲۱	۲۲	۲۳	۲۴	۲۵	۲۶	۲۷	۲۸	۲۹	۳۰	۳۱	۳۲	۳۳	۳۴	۳۵	۳۶	۳۷	۳۸	۳۹
3	۳۰	۳۱	۳۲	۳۳	۳۴	۳۵	۳۶	۳۷	۳۸	۳۹	۴۰	۴۱	۴۲	۴۳	۴۴	۴۵	۴۶	۴۷	۴۸	۴۹
4	۴۰	۴۱	۴۲	۴۳	۴۴	۴۵	۴۶	۴۷	۴۸	۴۹	۵۰	۵۱	۵۲	۵۳	۵۴	۵۵	۵۶	۵۷	۵۸	۵۹
5	۵۰	۵۱	۵۲	۵۳	۵۴	۵۵	۵۶	۵۷	۵۸	۵۹	۶۰	۶۱	۶۲	۶۳	۶۴	۶۵	۶۶	۶۷	۶۸	۶۹
6	۶۰	۶۱	۶۲	۶۳	۶۴	۶۵	۶۶	۶۷	۶۸	۶۹	۷۰	۷۱	۷۲	۷۳	۷۴	۷۵	۷۶	۷۷	۷۸	۷۹
7	۷۰	۷۱	۷۲	۷۳	۷۴	۷۵	۷۶	۷۷	۷۸	۷۹	۸۰	۸۱	۸۲	۸۳	۸۴	۸۵	۸۶	۸۷	۸۸	۸۹
8	۸۰	۸۱	۸۲	۸۳	۸۴	۸۵	۸۶	۸۷	۸۸	۸۹	۹۰	۹۱	۹۲	۹۳	۹۴	۹۵	۹۶	۹۷	۹۸	۹۹
9	۹۰	۹۱	۹۲	۹۳	۹۴	۹۵	۹۶	۹۷	۹۸	۹۹	۱۰۰	۱۰۱	۱۰۲	۱۰۳	۱۰۴	۱۰۵	۱۰۶	۱۰۷	۱۰۸	۱۰۹
A	۱۰۰	۱۰۱	۱۰۲	۱۰۳	۱۰۴	۱۰۵	۱۰۶	۱۰۷	۱۰۸	۱۰۹	۱۱۰	۱۱۱	۱۱۲	۱۱۳	۱۱۴	۱۱۵	۱۱۶	۱۱۷	۱۱۸	۱۱۹
B	۱۱۰	۱۱۱	۱۱۲	۱۱۳	۱۱۴	۱۱۵	۱۱۶	۱۱۷	۱۱۸	۱۱۹	۱۲۰	۱۲۱	۱۲۲	۱۲۳	۱۲۴	۱۲۵	۱۲۶	۱۲۷	۱۲۸	۱۲۹
C	۱۲۰	۱۲۱	۱۲۲	۱۲۳	۱۲۴	۱۲۵	۱۲۶	۱۲۷	۱۲۸	۱۲۹	۱۳۰	۱۳۱	۱۳۲	۱۳۳	۱۳۴	۱۳۵	۱۳۶	۱۳۷	۱۳۸	۱۳۹
D	۱۳۰	۱۳۱	۱۳۲	۱۳۳	۱۳۴	۱۳۵	۱۳۶	۱۳۷	۱۳۸	۱۳۹	۱۴۰	۱۴۱	۱۴۲	۱۴۳	۱۴۴	۱۴۵	۱۴۶	۱۴۷	۱۴۸	۱۴۹
E	۱۴۰	۱۴۱	۱۴۲	۱۴۳	۱۴۴	۱۴۵	۱۴۶	۱۴۷	۱۴۸	۱۴۹	۱۵۰	۱۵۱	۱۵۲	۱۵۳	۱۵۴	۱۵۵	۱۵۶	۱۵۷	۱۵۸	۱۵۹
F	۱۵۰	۱۵۱	۱۵۲	۱۵۳	۱۵۴	۱۵۵	۱۵۶	۱۵۷	۱۵۸	۱۵۹	۱۶۰	۱۶۱	۱۶۲	۱۶۳	۱۶۴	۱۶۵	۱۶۶	۱۶۷	۱۶۸	۱۶۹

# **Unicode** is the basis for Internationalized Domain Names

# Domain Name System (DNS)

- Domain name is the address of a website in the internet space which is used to access its contents from another machine

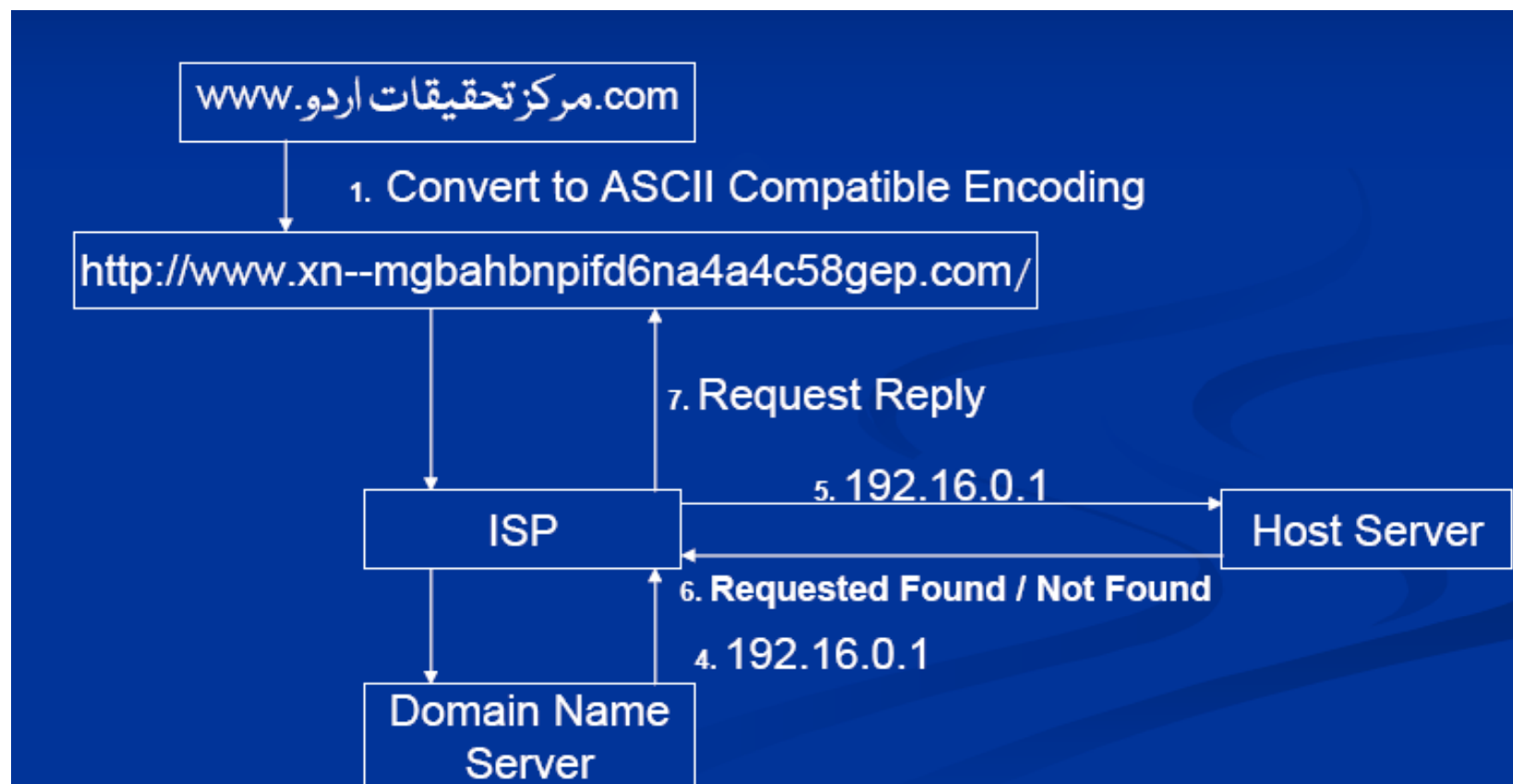


# Need of IDNs

- Current DNS is based on 7-bit ASCII standard, only supporting abc...xyz, 012...89, and ‘-’
- Makes it difficult to access Internet for people who do not understand English or Latin script
- We cannot change the overall existing system as it can break the internet
- The solution is to add layer that works on top of existing system
- IDN implements a mechanism which supports domain name in any language which can be converted to ASCII format and use the existing internet framework
- Initial set of protocols defined in 2003, called IDNA2003

# Internationalized Domain Name in Applications (IDNA)

- A layer that takes the address in local languages and converts that into ASCII format (using toASCII() )
- DNS continues to resolve ASCII format as usual





# 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
  - PROTOCOL VALID (or allowed)
  - DISALLOWED
  - CONTEXTO or CONTEXTJ (depends on the context of use)

# Morning Session

- ✓ Introduction to Unicode
- ✓ Internationalized domain names
- Issues related to IDNs for Pakistani languages











# Arabic Script

- Arabic script is the second largest script after Latin script
- It is used for writing Arabic, Urdu, Persian, Balochi, Pashto, Sindhi and many other languages across Pakistan and the world
- 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+FDx sequence)
  - New addition of dot-less characters and separate dots

# Arabic Script

- Cursive script
  - Shape of each letter may have four different shapes depending on its position (isolated, initial, medial or final)
- Bidirectional
  - Letters written from right to left
  - Numerals written left to right
- Diacritics (optionally) used for vowels
- Stretched shapes used for text justification
- Shapes of letters highly context sensitive

# Contextual Shapes of Different Letters

Isolated	Initial	Medial	Final
			
			
	NA	NA	

# Issues in Arabic Script Encoding

- Character status revision at script level
- Resolving confusability of characters
- Additional composed characters
- Digits and Mixing
- Single vs. multiple language tables
- Label separator
- ccTLD string
- gTLD translations

# Character Status Revision at Script Level

- Currently a formula using character properties determines which character is PVALID or DISALLOWED
- Some PVALID characters not used by any language and should be DISALLOWED
- ASIWG recommendations (Handout pg. 2)
  - Quranic marks
  - Formatting marks
- Do we agree for Pakistani languages?



# Confusability

- Visually similar character shapes create confusion
- Confusion can be due to initial, medial, final or isolated forms
- Different cases of confusability
  - Shape confusability
    - Exact shape confusion
    - Similar shape confusion
  - Composition confusability

## Exact Shape Confusion

- $ل + ك = كل$  looks same as  $ل + ك = كل$
- $چ + ل + ی (06CC) = چلی$  looks same as  
 $چ + ل + ی (0649) = چلی$
- $ی (06CC) + ا = یا$  looks same as  $ی$   
 $(064A) + ا = یا$

## Similar Shape Confusion

- Urdu character ۛ (06CC) and Pashto character ۛ (06CD)
- Sindhi ڪ (06AA) and Urdu ڪ (06A9)
  - ڪ vs. ڪ

## Composition Confusability

- There are characters that can be typed in more than one ways
  - U+0622 (اَ) =  
U+0627 (ا) + U+0653 (ّ)
- Although they look similar to the user, they translate to different ASCII codes

Composed Form	Decomposed Form	
U+0622 (اَ)	U+0627 (ا) + U+0653	
U+0623 (اِ)	U+0627 (ا) + U+0654	
U+0624 (اِو)	U+0648 (و) + U+0654	
U+0625 (اِء)	U+0627 (ا) + U+0655	
U+0626 (اِي)	U+064A (ي) + U+0654	
U+0675 (اِء)	U+0627 (ا) + U+0674	

# Solution and Problem

- Solution
  - Mapping for confusable shapes
    - For Urdu ۛ (0649) can be mapped to ۛ (06CC)
  - Normalization for composed forms
- Problem
  - Unicode does not provide mapping
    - Language dependent
  - Only partial normalization is provided in the Unicode standard onto pre-composed characters
    - Script dependent

# Issues in Arabic Script Encoding

- *Character status revision at script level*
- *Resolving confusability of characters*
- *Additional composed characters*
- Digits and Mixing
- Character and Label separator
- Single vs. multiple language tables
- ccTLD string
- gTLD translations

# Digit sets in Arabic

	ASCII		ARABIC-INDIC		EXTENDED ARABIC-INDIC
0	U+0300	◦	U+0660	◦	U+06F0
1	U+0301	۱	U+0661	۱	U+06F1
2	U+0302	۲	U+0662	۲	U+06F2
3	U+0303	۳	U+0663	۳	U+06F3
4	U+0304	٤	U+0664	٤/٤	U+06F4
5	U+0305	٥	U+0665	٥	U+06F5
6	U+0306	٦	U+0666	٦/٦	U+06F6
7	U+0307	٧	U+0667	٧/٧	U+06F7
8	U+0308	٨	U+0668	٨	U+06F8
9	U+0309	٩	U+0669	٩	U+06F9



# Mixing Digit Cases

## 1. Two sets are mixed

- [www.اردو.com](http://www.اردو.com)
- [www.123اردو.com](http://www.123اردو.com)
- [www.۱۲۳اردو.com](http://www.۱۲۳اردو.com)
- [www.۱۲3اردو.com](http://www.۱۲3اردو.com)

## 2. No mixing of digits

- [www.اردو.com](http://www.اردو.com)
- [www.123اردو.com](http://www.123اردو.com)
- [www.۱۲۳اردو.com](http://www.۱۲۳اردو.com)

# Mixing Digits

- Mixing digits
  - A large number of domain names can be generated
  - Many of the labels generated are linguistically incorrect
  - Users may perceive mixed digit labels similar to non-mixed ones; potential for spoofing/confusion
- No mixing
  - Number of domain names limited
  - Some languages may require mixing for complete representation of words

# Mixing Digits

- Two of these digit blocks used by Pakistani languages
  - ASCII and Extended Arabic-Indic
- Which set is required in IDNs by the language?
- Is mixing of both types of digits allowed?

# Character Separator

- Need a character separator for proper shaping in Urdu
  - Words may assume wrong shapes without a separator e.g. دس دن will be displayed erroneously دسدن without a separator
- Space not allowed in domain names
- Zero Width Non Joiner (ZWNJ)
  - But users unfamiliar with it
  - Not available on conventional keyboards
- Any alternate Solution?

# Label separator

- Pakistani languages use +06D4 (-) as label separator
- Standard ASCII names in DNS use 002E (.) as separator
- Using dash for Pakistani languages
  - Pros: Keyboard switching not required
  - Cons: Mapping has to be standardized for web browsers and other applications
- Using dot
  - Pros: Part of the existing Internet standard; no mapping is needed
  - Cons: Keyboard switching required
- What should be label separator?

## Keeping in view the issues discussed so far...

- Language tables can be constructed in two ways
  - One table for each Pakistani language
  - Single table for all languages
- Both have advantages and disadvantages

# Single Language Table

- All languages represented in one table
- Lists needed and not needed characters for all languages in single table
  - Easier to maintain
  - New languages can be added conveniently
  - But, how to deal with additional confusability? May compromise complete language being expressed

# Multiple Language Tables

- One table for each Pakistani language.
  - For e.g. Baluchi, Pashto, Punjabi, Saraiki, Sindhi, Torwali
  - List each language's character-set separately
  - Confusability is limited and can be addressed without compromising language expression
  - But, difficult to maintain
  - And difficult to upgrade develop separate table for each of the 66+ languages of Pakistan



# ccTLD String

- Candidate Country-Code Top-Level Domain string

- پاکستان

- پاک

- ووو۔ اردو مرکز۔ ادارہ۔ پاکستان؟

- ووو۔ اردو مرکز۔ ادارہ۔ پاک؟

# gTLD Translations

gTLD String	gTLD Abbrev.	Urdu
ARPA	arpa	انٹرنیٹ
COMPANY	com	کمپنی
EDUCATION	edu	تعلیم
GOVERNMENT	gov	حکومت
MILITARY	mil	فوج
ORGANIZATION	org	ادارہ
INTERNATIONAL	int	عالمی

gTLD String	gTLD Abbrev.	Urdu
NET	net	نیٹ
INFORMATION	info	اطلاعات
MEDIA	media	میڈیا
NAME	name	نام
BUSINESS	biz	کاروبار
AEROSPACE	aero	فضائیات
PROFESSIONAL	pro	پروفیشنل
MUSEUM	museum	میوزیم
Employment Related	jobs	ملازمت
Travel agents/ Airlines	travel	سیاحت

