Overview of elementary standards

Language and locale coding

Character encoding

What do we need?

- Identification of linguistic communities
  - Librarian, technological, linguistic perspectives
- Identification of ‘locales’
  - Communities with cultural and linguistic specificities
- Identification of writing systems
  - How a language is represented in written form (from stone to computers)

Languages

- ISO 639:1988, *Code for the representation of names of languages. Part 1: Alpha-2 codes*
  - Two-letter language symbols
- ISO 639-2: *Code for the representation of names of languages. Part 2: Alpha-3 codes*
  - Three-letter language symbols

    en/eng = English

    fr/fra = French (français)

    es/esp = Spanish (español)

    de/deu = German (Deutsch)

Note: lowercase letters for language codes

Maintenance agency:

http://www.iso.org/iso/en/prods-services/iso3166ma/index.html

Countries

- ISO 3166: *Code for the representation of names of countries*
    - Two-letter country symbols

  GB = Great Britain, US = United States, FR = France, RO = Romania

Note: uppercase letters for country codes

- Combining languages and countries:

  fr FR = French French, fr CA = Canadian French

Difficulties

- Regional variants
  - ? Towards an extended codification of places (which granularity)
- A limited language repertoire
  - A lot of “peripheral” languages are not registered
    - Cf. Ethnologue http://www.sil.org

Representing written languages

Definitions, history
and current situation

Basic definitions

- Character repertoire
  - Set of distinct characters, defined independently of any coding or ordering rule/procedure
  - Each character is defined by a name and a reference shape
  - Rem.: distinct characters may be associated with the same shape
    - A: Latin capital A, Cyrillic capital A, Greek capital A

Basic definitions (cont.)

- Character code
  - One to one (bijection) association between a character repertoire and a set of positive integers
  - Hence, notion of **position**
    - Presentation of characters in a table

Basic definitions (cont.)

- Character encoding
  - Method (algorithm) to represent in electronic form (as a sequence of bytes) of a character code
  - By definition: a process which should be independent from the character code and the character repertoire
  - Simple case (When the code is defined within [0-256])
    - The integer code is associated to its standard representation as a byte

Example

- Character repertoire
    - “a”, “!”, “ä”, “‰”
- Character codes
    - ISO 10646
        - 97, 33, 228, 8240
- Encoding
    - As two bytes
        - 0 97, 0 33, 0 228, 32 48

Difficulties

- Charset/character set
  - Ambiguous term that designates globally the character repertoire, codes and/or encoding
  - E.g.: used in MIME headers
- Language
  - Often (but wrongly) associated with the choice of a repertoire (e.g. web browsers)
    - E.g.: Bulgarian can be represented in Cyrillic or Latin characters
- Fonts
  - Impose constraints on the representation of characters
  - Subordinated to the prior choice of a repertoire

Some archaeology…

- ASCII - American Standard Code for Information Interchange
  - Combines repertoire, codes and encoding
  - The ASCII code also contains control characters
    - E.g. CR, LF, ESC, TAB
  - Repertoire

```
 ! " # $ % & ' ( ) * + , - . /
0 1 2 3 4 5 6 7 8 9 : ; < = > ?
@ A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z [ \ ] ^ _
` a b c d e f g h i j k l m n o
P q r s t u v w x y z { | } ~
```

ASCII : definitions

- Character codes
  - One to one association of a number from 32 (“ ”) to 126 (“~”) following the order in the preceding table
  - Positions from 0 to 31, as well as 127are kept for « standardizes » control characters
- Character encoding
  - Codes are represented by their standard byte representation
  - No specific use is made of codes between 128 and 255 (parity)

From a standardization point of view

- United states (US-ASCII)
  - ANSI X3.4-1986
- International (ISO/IEC JTC1/SC2/WG3)
  - ISO 646
    - Introduces flexibility for some positions in the code
      - # $ ^ ` ~
    - Some positions are kept for “national usage”
      - @ [\ ] { | }
    - IRV (1991 edition): International Reference Version = US-ASCII

Next step…

- ISO Latin 1, alias ISO 8859-1
  - One member in a family of standards (ISO 8859)
  - Defines:
    - A character repertoire
      - Alphabet latin n° 1 (ISO Latin 1)
    - The corresponding codes
      - Where ASCII is seen as a sub-set
    - Encoding
      - Same as ASCII (byte encoding of integers from 0 to 255)

ISO 8859-1

- Additional characters
  - Codes from 160 to 255

    ¡ ¢ £ € ¥ | § ¨ © ª « ¬ – ® ¯
    ° ± 2 3 ´ µ ¶ · ¸ 1 ° » * * * ¿
    À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï
    ‹ Ñ Ò Ó Ô Õ Ö x Ø Ù Ú Û Ü † ﬁ ß
    à á â ã ä å æ ç è é ê ë ì í î ï
    › ñ ò ó ô õ ö ÷ ø ù ú û ü ‡ ﬂ ÿ

  - Rem.:
    - Positions from 128 to 159 are kept for control characters
      - E.g. Windows code page 1252, `windows-1252`
    - Code 160: no-break space

The rest of the family

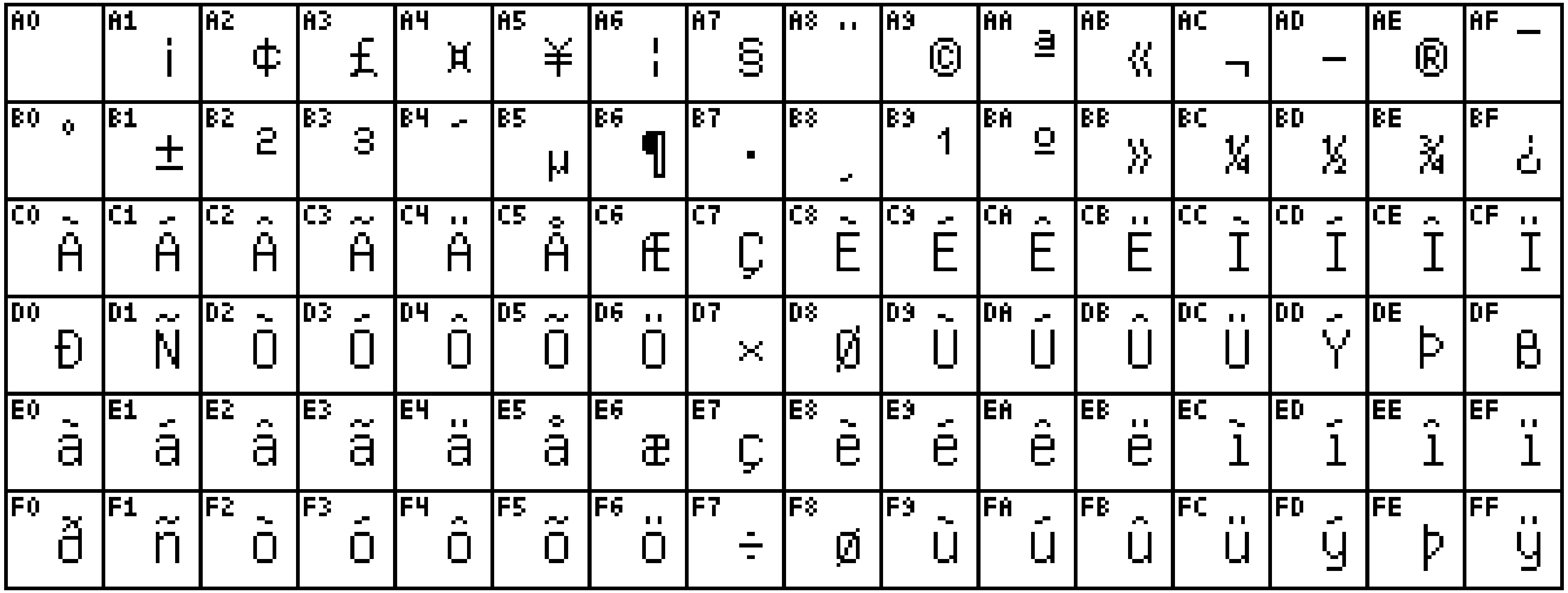
- ISO 8859 from a wider perspective
  - The same principles as those of ISO 8859-1 are used to describe other repertoires
  - ISO 8859-2 (ISO Latin 2)
    - Slavic languages from centre and eastern Europe
  - ISO 8859-15 (ISO Latin 9)
    - € !
  - Etc.

The whole family…

ISO 8859-1, Latin alphabet No. 1, Western", "West European"
ISO 8859-2, Latin alphabet No. 2, "Central European", "East European"
ISO 8859-3, Latin alphabet No. 3, "South European"; "Maltese & Esperanto"
ISO 8859-4, Latin alphabet No. 4, "North European"
ISO 8859-5, Latin/Cyrillic alphabet, (for Slavic languages)
ISO 8859-6, Latin/Arabic alphabet (for the Arabic language)
ISO 8859-7, Latin/Greek alphabet (for modern Greek)
ISO 8859-8, Latin/Hebrew alphabet (for Hebrew and Yiddish)
ISO 8859-9, Latin alphabet No. 5, "Turkish"
ISO 8859-10, Latin alphabet No. 6, "Nordic" (Sámi, Inuit, Icelandic)
ISO 8859-11, Latin/Thai alphabet, (for the Thai language; draft)
(Part 12 has not been defined.)
ISO 8859-13, Latin alphabet No. 7, Baltic Rim
ISO 8859-14, Latin alphabet No. 8, Celtic
ISO 8859-15, Latin alphabet No. 9, "euro"
ISO 8859-16, Latin alphabet No. 10, for a collection of languages

ISO 8859 tables

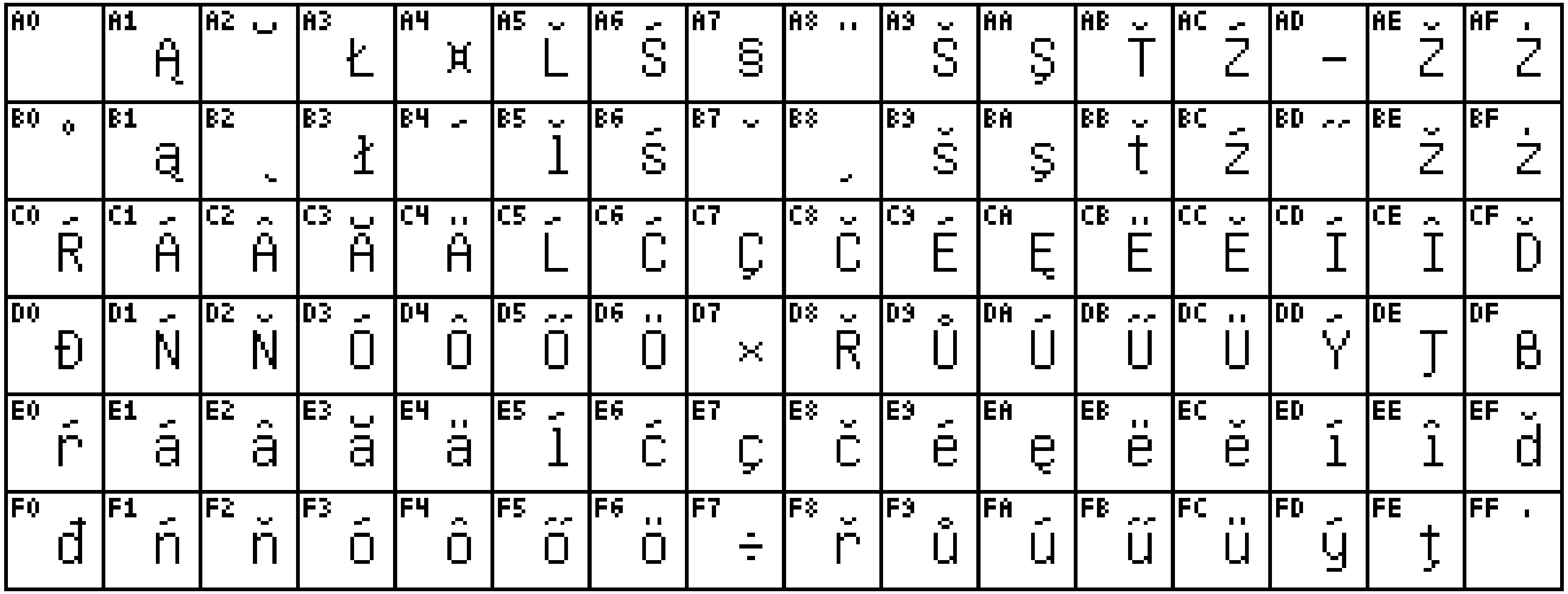
- ISO 8859-1
  - fr, es, Catalan (ca), Basque (eu), pt, it, Albanian (sq), Rhaeto-Romanic (rm), nl, de, da, sv, no, fi, Faroese (fo), Icelandic (is), Irish (ga), Scottish (gd), and en



A0	A1 ¡	A2 ¢	A3 £	A4 ¤	A5 ¥	A6 ¦	A7 §	A8 ¨	A9 ©	AA ª	AB «	AC ¬	AD -	AE ®	AF ¯
B0 °	B1 ±	B2 ²	B3 ³	B4 ´	B5 µ	B6 ¶	B7 ·	B8 ¸	B9 ¹	BA º	BB »	BC ¼	BD ½	BE ¾	BF ¿
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
D0 Ð	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Õ	D6 Ö	D7 ×	D8 Ø	D9 Ù	DA Ú	DB Û	DC Ü	DD Ý	DE Þ	DF ß
E0 à	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
F0 ð	F1 ñ	F2 ò	F3 ó	F4 ô	F5 õ	F6 ö	F7 ÷	F8 ø	F9 ù	FA ú	FB û	FC ü	FD ý	FE þ	FF ÿ

ISO 8859 tables

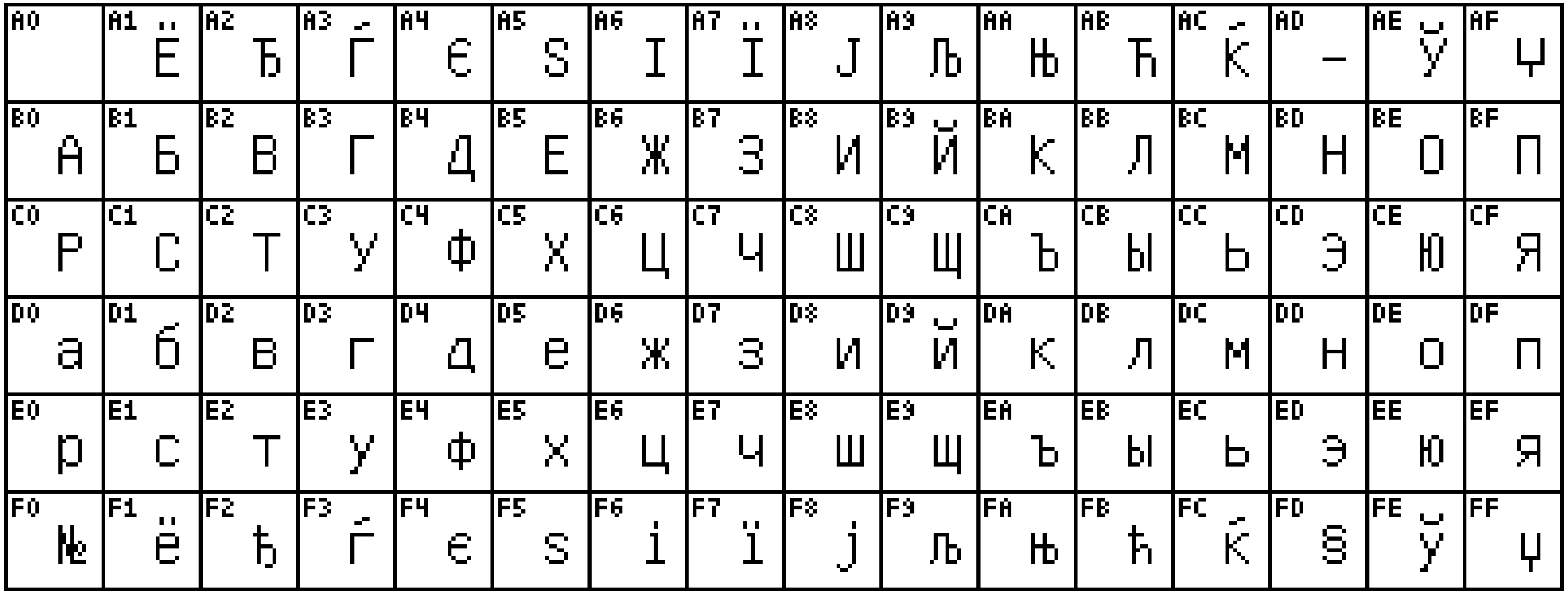
- ISO 8859-2 (Latin 2)
  - Czech (cs), Hungarian (hu), Polish (pl), Romanian (ro), Croatian (hr), Slovak (sk), Slovenian (sl), Sorbian



A0	A1 Ą	A2 ˘	A3 Ł	A4 ¤	A5 Ľ	A6 Ś	A7 §	A8 ¨	A9 Š	AA Ş	AB Ť	AC Ź	AD -	AE Ž	AF Ż
B0 °	B1 ą	B2 ˛	B3 ł	B4 ´	B5 ľ	B6 ś	B7 ˇ	B8 ¸	B9 š	BA ş	BB ť	BC ź	BD ˝	BE ž	BF ż
C0 Ŕ	C1 Á	C2 Â	C3 Ă	C4 Ä	C5 Ĺ	C6 Ć	C7 Ç	C8 Č	C9 É	CA Ę	CB Ë	CC Ě	CD Í	CE Î	CF Ď
D0 Đ	D1 Ń	D2 Ň	D3 Ó	D4 Ô	D5 Ő	D6 Ö	D7 ×	D8 Ř	D9 Ů	DA Ú	DB Ű	DC Ü	DD Ý	DE Ţ	DF ß
E0 ŕ	E1 á	E2 â	E3 ă	E4 ä	E5 ĺ	E6 ć	E7 ç	E8 č	E9 é	EA ę	EB ë	EC ě	ED í	EE î	EF ď
F0 đ	F1 ń	F2 ň	F3 ó	F4 ô	F5 ő	F6 ö	F7 ÷	F8 ř	F9 ů	FA ú	FB ű	FC ü	FD ý	FE ţ	FF ˙

ISO 8859 tables

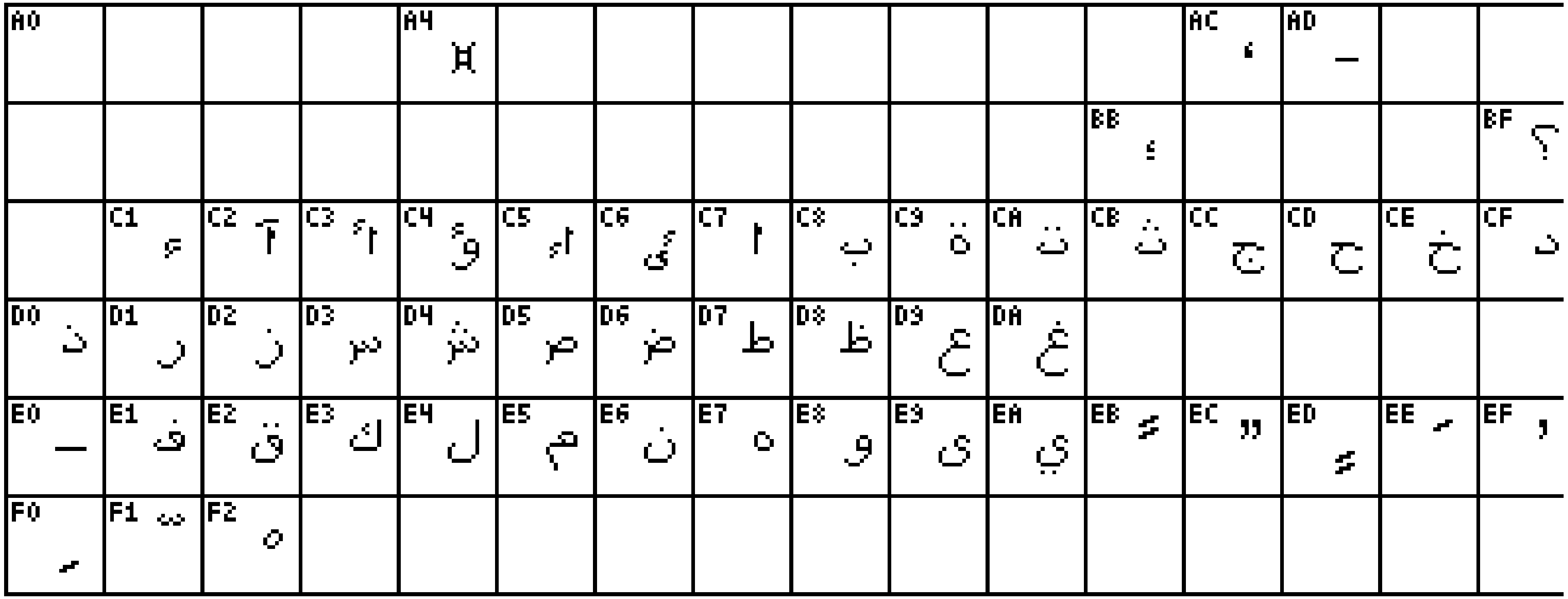
- ISO 8859-5 (Cyrillik)
  - Bulgarian (bg), Byelorussian (be), Macedonian (mk), Russian (ru), Serbian (sr)



A0	A1 Ё	A2 Ђ	A3 Ѓ	A4 Є	A5 Ѕ	A6 І	A7 Ї	A8 Ј	A9 Љ	AA Њ	AB Ћ	AC Ќ	AD -	AE Ў	AF Џ
B0 А	B1 Б	B2 В	B3 Г	B4 Д	B5 Е	B6 Ж	B7 З	B8 И	B9 Й	BA К	BB Л	BC М	BD Н	BE О	BF П
C0 Р	C1 С	C2 Т	C3 У	C4 Ф	C5 Х	C6 Ц	C7 Ч	C8 Ш	C9 Щ	CA Ъ	CB Ы	CC Ь	CD Э	CE Ю	CF Я
D0 а	D1 б	D2 в	D3 г	D4 д	D5 е	D6 ж	D7 з	D8 и	D9 й	DA к	DB л	DC м	DD н	DE о	DF п
E0 р	E1 с	E2 т	E3 у	E4 ф	E5 х	E6 ц	E7 ч	E8 ш	E9 щ	EA ъ	EB ы	EC ь	ED э	EE ю	EF я
F0 №	F1 ё	F2 ђ	F3 ѓ	F4 є	F5 ѕ	F6 і	F7 ї	F8 ј	F9 љ	FA њ	FB ћ	FC ќ	FD §	FE ў	FF џ

ISO 8859 tables

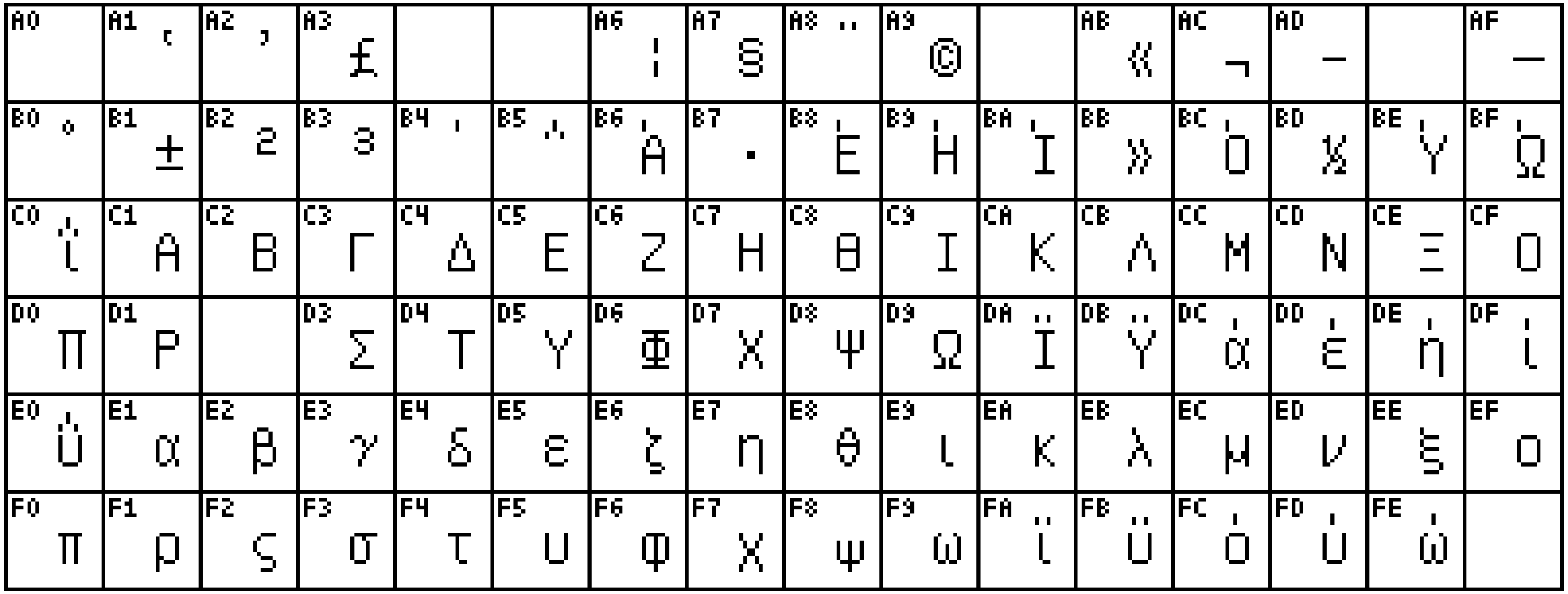
- ISO-8859-6 (Arabic - ar)
  - Characters are missing for Perse (fa) and Urdu (ur) in Pakistan



A0				A4 ¤								AC ،	AD –		
											BB ؛				BF ؟
	C1 ء	C2 آ	C3 أ	C4 ؤ	C5 إ	C6 ئ	C7 ا	C8 ب	C9 ة	CA ت	CB ث	CC ج	CD ح	CE خ	CF د
D0 ذ	D1 ر	D2 ز	D3 س	D4 ش	D5 ص	D6 ض	D7 ط	D8 ظ	D9 ع	DA غ					
E0 ـ	E1 ف	E2 ق	E3 ك	E4 ل	E5 م	E6 ن	E7 ه	E8 و	E9 ى	EA ي	EB ً	EC ٌ	ED ٍ	EE َ	EF ُ
F0 ِ	F1 ّ	F2 ْ													

ISO 8859 tables

- ISO 8859-7 (Greek - el)

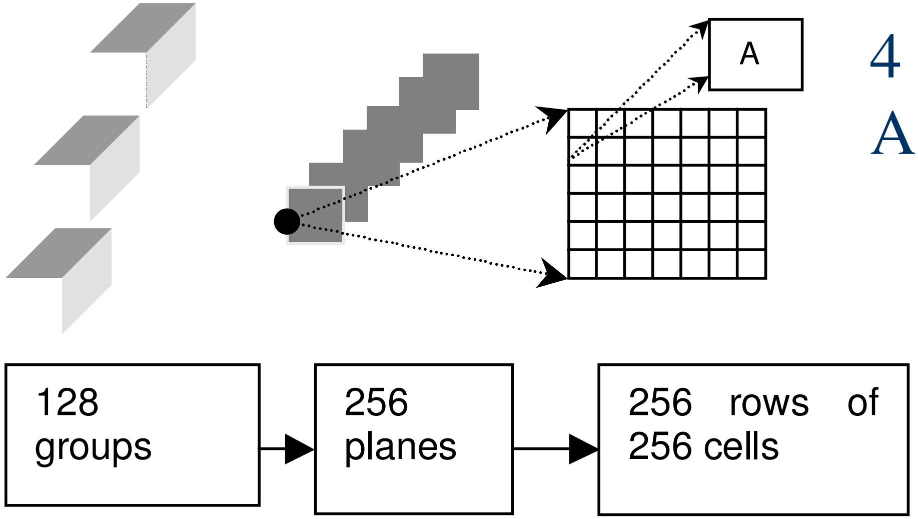


A0	A1 ʽ	A2 ʼ	A3 £			A6 ¦	A7 §	A8 ¨	A9 ©		AB «	AC ¬	AD –		AF ―
B0 °	B1 ±	B2 ²	B3 ³	B4 ΄	B5 ΅	B6 Ά	B7 ·	B8 Έ	B9 Ή	BA Ί	BB »	BC Ό	BD ½	BE Ύ	BF Ώ
C0 ΐ	C1 Α	C2 Β	C3 Γ	C4 Δ	C5 Ε	C6 Ζ	C7 Η	C8 Θ	C9 Ι	CA Κ	CB Λ	CC Μ	CD Ν	CE Ξ	CF Ο
D0 Π	D1 Ρ		D3 Σ	D4 Τ	D5 Υ	D6 Φ	D7 Χ	D8 Ψ	D9 Ω	DA Ϊ	DB Ϋ	DC ά	DD έ	DE ή	DF ί
E0 ΰ	E1 α	E2 β	E3 γ	E4 δ	E5 ε	E6 ζ	E7 η	E8 θ	E9 ι	EA κ	EB λ	EC μ	ED ν	EE ξ	EF ο
F0 π	F1 ρ	F2 ς	F3 σ	F4 τ	F5 υ	F6 φ	F7 χ	F8 ψ	F9 ω	FA ϊ	FB ϋ	FC ό	FD ύ	FE ώ	

Towards a universal representation of characters

- ISO/IEC 10646 (UCS)
  - An international standard
  - UCS: Universal Character Set
  - An extensible character repertoire associated to a code
  - Underlying abstract model
- Unicode
  - An industry consortium standard
  - Defines a character repertoire and a code made compatible with that of ISO 10646
    - Provides additional constraints on character usage

Structure of ISO/IEC 10646



A

128 groups

256 planes

256 rows of 256 cells

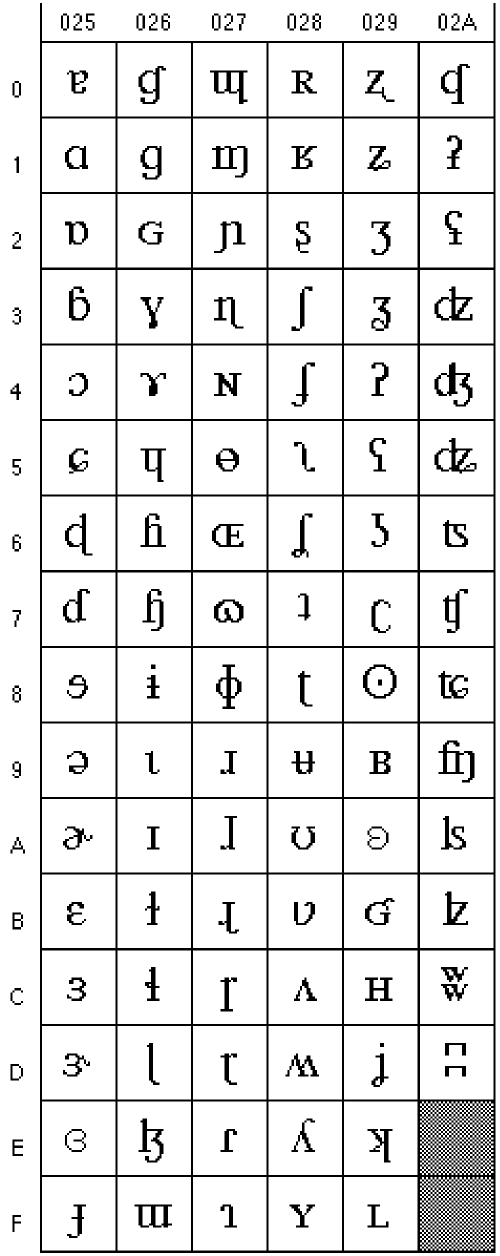
4 byte encoding
A = 00 00 00 41

Structure of ISO/IEC 10646 (cont.)

- The character code is identified by:
  - Group - Plane - Row - Cell
- BMP - Basic Multilingual Plane
  - Group = 0, Plane = 0
  - Corresponds to a two byte encoding seen as four zones

A	**alphabets, symbols, phonetic section of CJK, hangul...**	**0000 à 4DFF**	*19903 positions*
I	**Unified representations of ideograms (CJK)**	**4E00 à 9EFF**	*20992 positions*
O	**Reserved for future use**	**A000 à DFFF**	*16384 positions*
R	**Private use, compatibility zone, arabic special forms =restricted use section**	**E000 à FFFD**	*8190 positions*

Example : IPA (International Phonetic Alphabet) U+0250..U+02AF



	025	026	027	028	029	02A
0	ɐ	ɠ	ɰ	ʀ	ʐ	ʠ
1	ɑ	ɡ	ɱ	ʁ	ʑ	ʡ
2	ɒ	ɢ	ɲ	ʂ	ʒ	ʢ
3	ɓ	ɣ	ɳ	ʃ	ʓ	ʣ
4	ɔ	ɤ	ɴ	ʄ	ʔ	ʤ
5	ɕ	ɥ	ɵ	ʅ	ʕ	ʥ
6	ɖ	ɦ	ɶ	ʆ	ʖ	ʦ
7	ɗ	ɧ	ɷ	ʇ	ʗ	ʧ
8	ɘ	ɨ	ɸ	ʈ	ʘ	ʨ
9	ə	ɩ	ɹ	ʉ	ʙ	ʩ
A	ɚ	ɪ	ɺ	ʊ	ʚ	ʪ
B	ɛ	ɫ	ɻ	ʋ	ʛ	ʫ
C	ɜ	ɬ	ɼ	ʌ	ʜ	ʬ
D	ɝ	ɭ	ɽ	ʍ	ʝ	ʭ
E	ɞ	ɮ	ɾ	ʎ	ʞ	
F	ɟ	ɯ	ɿ	ʏ	ʟ	

Encodings (for BMP/Unicode)

- Reference encoding
  - UCS-2
    - Representation of characters as a sequences of two bytes
- Alternative
  - UTF-8
    - Codes below 128 are represented as one byte (7 bits, cf. ASCII codes)
    - Other codes are represented as a sequence of 2 to 6 bytes (belonging to [128,255])

Summary

- We are close to a stable picture for character representation
  - 30 years to acheive this!
- General idea of the standardisation process
  - Combines:
    - Identification of existing practices
    - Abstraction to cope for additional needs

Sources

- Unicode Technical Report #17: Character Encoding Model
- Korpela, Jukka 2001. A tutorial on character code issues