

Session 29 – EndUser – April 15, 2004

Coded Character Sets

A Technical Primer for Librarians

Michael Doran, Systems Librarian
University of Texas at Arlington

Writing Systems

Composed of characters...

- Letters of an alphabet
- Numbers
- Punctuation
- Special symbols
- Modifying marks (diacritics)
- Ideographs

EndUser 2004 - Session 29

Computers

Many uses

- Mathematical calculations
- Textual processing

Data storage and transmission

- Ones and zeros (1,0)
(i.e. binary digits, or bits)
- Numerical code

EndUser 2004 - Session 29

Writing Systems

Composed of characters

- Letters of an alphabet
- Numbers
- Punctuation
- Special symbols
- Modifying marks (diacritics)
- Ideographs

Computers

Many uses

- Mathematical calculations
- Textual processing

Data storage

- Ones and zeros (1,0)
(i.e. binary digits, or bits)
- Numerical code

Coded Character Sets

EndUser 2004 - Session 29

7-bit Code Matrix

16 columns x 8 rows = 128 cells

Each cell in the array contains a number (expressed in hexadecimal).

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F

["Binary inside"](#)

EndUser 2004 - Session 29

Number Systems

Base	Name	Example value
2	Binary	01111010
8	Octal	172
10	Decimal	122
16	Hexadecimal	7A

``

EndUser 2004 - Session 29

ASCII

- Solved one problem – it became a widely adopted standard for sharing data
- However... ASCII was not very useful for the non-English speaking world
- Fortunately, there became less need for a parity bit, thus freeing up the “eighth bit” for additional characters and leading to the creation of 8-bit character sets

EndUser 2004 - Session 29

Bit Combinations

	Examples	Possible
1-bit	<u>0</u>	$2^1 = 2$
2-bit	<u>01</u>	$2^2 = 4$
3-bit	<u>101</u>	$2^3 = 8$
4-bit	<u>0110</u>	$2^4 = 16$
5-bit	<u>10100</u>	$2^5 = 32$
6-bit	<u>001011</u>	$2^6 = 64$
7-bit	<u>1100110</u>	$2^7 = 128$
8-bit	<u>10011010</u>	$2^8 = 256$

EndUser 2004 - Session 29

8-bit Code Matrix

8-bit => 2⁸ bit combinations => 256 code points

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

8-bit Code Matrix

- Ready for characters -

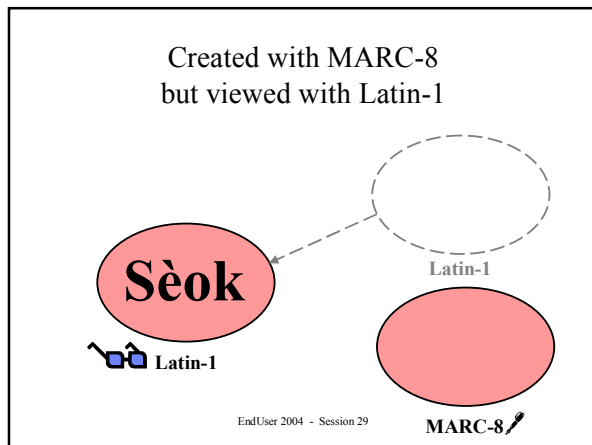
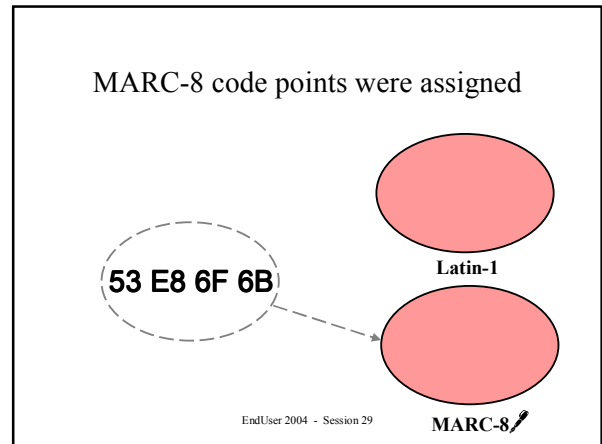
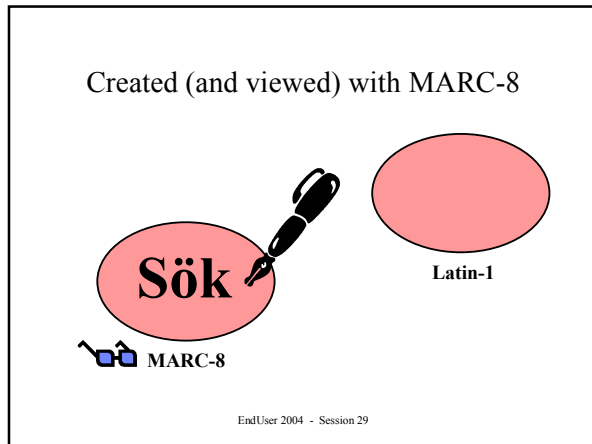
Hmmm, we might want to keep ASCII.

Control Functions C0																
ASCII Graphic Characters G0	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Control Functions C1																
Graphic Characters G1																

Latin-1 Character Set

See also: [Latin-1 Code Chart](#)

Control Functions C0																
ASCII Graphic Characters G0	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Control Functions C1																
Graphic Characters G1	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ



Real-life “lens” example

1.) Created and viewed via the Voyager cataloging module:

2.) Viewed via SQL*Plus from the Voyager server command line:

```
SQL> select title from bib_text where bib_id = '978602';
```

TITLE

Andrê Le Nôtre : gardener to the sun king /

EndUser 2004 - Session 29

- ### The Great Escape(s) “MARC-8 Stratagem #2”
- The use of non-spacing graphic characters is fine for representing Latin languages and “adequate” for romanized languages, but...
 - It would be nice to have the option of representing non-Latin languages in their native character sets, but...
 - There’s no room in an 8-bit code matrix, so....
- EndUser 2004 - Session 29

Escape to an alternate character set

...blah, blah, εύρηκα! blah, blah...

```
... 62 6C 61 68 2C 20 62
6C 61 68 2C 20 1B 29 53
E6 A2 F9 F5 EA ED E1 21
1B 29 21 45 20 62 6C 61
68 2C 20 62 6C 61 68...
```

EndUser 2004 - Session 29

MARC-8 Default

See also: [MARC-8 Default Code Chart](#)

Control Functions C0															
ASCII	0	1	2	3	4	5	6	7	8	9	:	<	=	>	?
Graphic Characters G0	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	{		}	~
Control Functions C1															
ANSEL	̀	ˆ	˜	˘	˙	˚	˛	˜	˘	˙	˚	˛	˜	˘	˙
Graphic Characters G1	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı
	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı

MARC-8 with Alternate G1

An escape sequence of hex "1B 29 53" designates Greek as the G1 graphic character set.

Control Functions C0															
ASCII	0	1	2	3	4	5	6	7	8	9	:	<	=	>	?
Graphic Characters G0	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	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	{		}	~
Control Functions C1															
Greek	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο
Graphic Characters G1	π	ρ	σ	ς	τ	υ	φ	χ	ψ	ω	␣	␣	␣	␣	␣
	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣
	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣
	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣

Escape to an alternate character set
- illuminated -

...blah, blah, εύρηκα! blah, blah...

```

... 62 6C 61 68 2C 20 62
6C 61 68 2C 20 1B 29 53
E6 A2 F9 F5 EA ED E1 21
1B 29 21 45 20 62 6C 61
68 2C 20 62 6C 61 68...

```

EndUser 2004 - Session 29

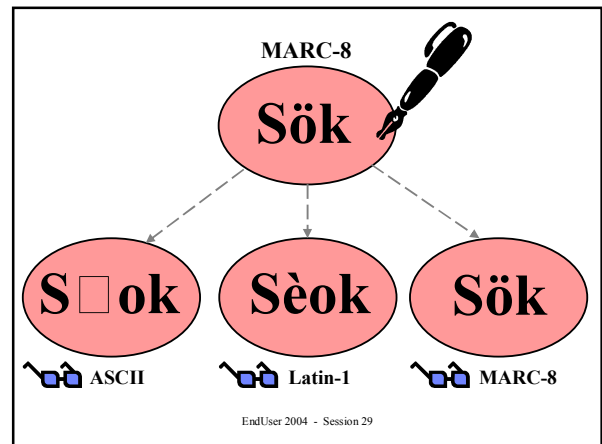
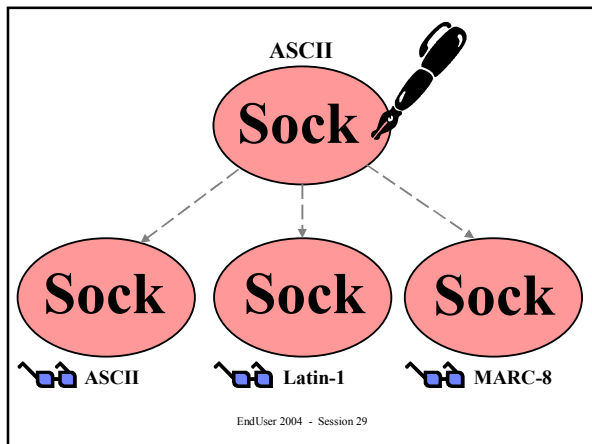
- ### MARC-8 Default and Alternate Character Sets
- “The Majors”

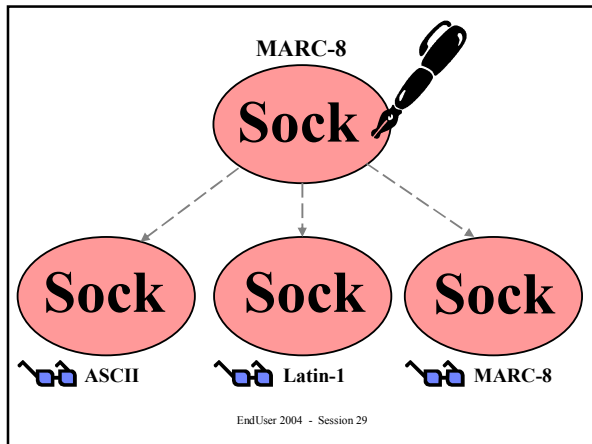
 - Basic Latin
 - Extended Latin
 - Basic Arabic
 - Extended Arabic
 - Basic Cyrillic
 - Extended Cyrillic
 - Greek
 - Hebrew
 - East Asian*

“The Minors”

 - Greek Symbols
 - Subscripts
 - Superscripts

[Associated Standards](#)
- EndUser 2004 - Session 29





Mid-Session Review ☺

- Text is created (using a character set)
 - Each character is assigned a code point (i.e. number)
 - It's the numerical code that is stored
- That textual data can be shared
 - Numerical code is transferred (plus metadata?)
- Can be viewed with same or different charset
 - Same charset that created data => good thing
 - Different charset => bad thing (probably)
- Why bad? The numerical code hasn't changed
 - The same code point represents different characters in different coded character sets

EndUser 2004 - Session 29

Changing the numerical code

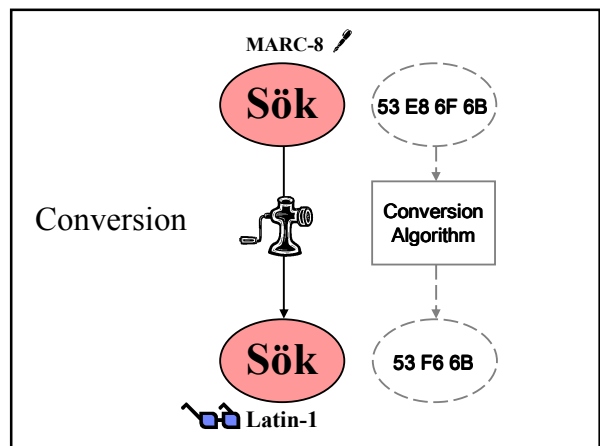
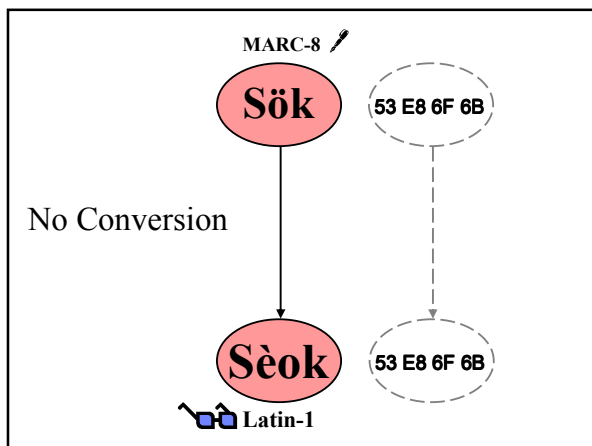
- Character set conversion
 - Permanently (convert the source file)
 - On-the-fly (convert a *copy* of the source file)
- Why would you want/need to convert?
 - Data integrity (shouldn't mix charsets within a single "container" - e.g. database)
 - The original character set isn't available to recipients of your textual data
 - Trade up to a newer, better model

EndUser 2004 - Session 29

Examples of charset conversion

- Importing MARC records into Voyager
 - MARC-8/OCLC/RLIN/VRLIN/Latin-1 => VRLIN
 - Why? Data integrity within the Oracle database
- Exporting MARC records out of Voyager
 - VRLIN => MARC-8/OCLC/RLIN/VRLIN/Latin-1
 - Why? VRLIN isn't widely available to recipients of data
- Displaying MARC record data in WebVoyage
 - VRLIN => Latin-1
 - Why? The Internet world isn't hip to bibliographic character set standards and Latin-1 was/is in wide use
 - Charset conversion pitfall: [1000's of chars vs. 256]

EndUser 2004 - Session 29



Examples of charset conversion (continued)

- Oracle SQL query via ODBC driver
 - Database character set => ODBC character set setting
 - Why? Who knows! It's a feature.
- Voyager with Unicode upgrade
 - VRLIN => Unicode UTF-8
 - Why? Unicode is better (trading up)

EndUser 2004 - Session 29

Real-life (bad) conversion example “André Le Nôtre” in MARC-8

1.) Viewed via SQL*Plus from the Voyager server command line:

```
SQL> select title from bib_text where bib_id = '978602';
TITLE
-----
Andr e Le N otre : gardener to the sun king /
```

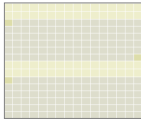
2.) Viewed via SQL*Plus from a PC client:

```
SQL> select title from bib_text where bib_id = '978602';
TITLE
-----
Andrbe Le Ncotre : gardener to the sun king /
```

EndUser 2004 - Session 29

Unicode

- Unicode is a coded character set that endeavors to provide a unique code point for every character in every language
- 16-bit encoding (2^{16}) => 65,536 code points



x 256

EndUser 2004 - Session 29

MARC 21 Unicode Environment

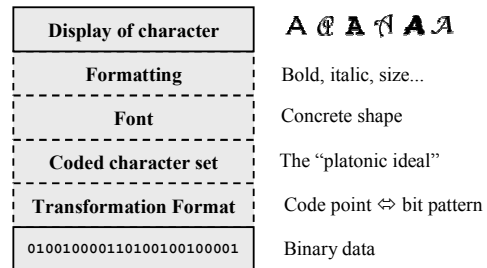
- The MARC 21 Unicode environment is simply the MARC-8 character repertoire translated into the Unicode equivalent code points.
 - Continues the use of non-spacing graphic characters for diacritics – precomposed versions of Unicode characters are not valid
 - Only the default and alternate character sets previously defined in MARC-8 are valid

EndUser 2004 - Session 29

Final Review ☺

Challenge	Solution	New Challenge
Reconciling writing systems & computer data storage	Coded character sets	Computer manufacturers used proprietary character sets
Sharing computer data	Interchange standards (e.g. ASCII & Latin-1)	Limited character repertoires
Encoding bibliographic citations for collections in multiple languages	MARC-8 standards	The non-library world doesn't use these standards
Overcoming system incompatibilities	Conversion to Unicode	MARC 21 does not provide for <i>full</i> implementation of Unicode.

Coded Character Sets ... are just one layer of the cake



EndUser 2004 - Session 29

Character Set “Negotiation”

- Usually transparent within a “closed” system
- Metadata is important within a client-server environment
 - Internal to word processing document
 - HTML content meta tag
 - XML declaration
 - MIME (Multipurpose Internet Mail Extensions)
 - Database transaction

EndUser 2004 - Session 29

HTML metatag

```
<html><head>  
<title>New Books List - User List</title>  
<meta http-equiv="Content-Type"  
      content="text/html; charset=ISO-8859-1">  
</head>
```

XML metatag

```
<?xml version="1.0" encoding="UTF-8"?>  
<!DOCTYPE html  
      PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"  
      "DTD/xhtml1-transitional.dtd">
```

Email header

```
From: <AHIGHSMI@lib-gw.tamu.edu>  
To: <doran@uta.edu>  
Subject: character sets  
Mime-Version: 1.0  
Content-Type: text/plain; charset=US-ASCII  
Content-Transfer-Encoding: 7bit
```

Communications Breakdown

- Problems can occur when...
 - No metadata is present
 - Metadata is overridden
 - The source character set is not available
 - A bad conversion takes place

EndUser 2004 - Session 29

That's all folks!



EndUser 2004 - Session 29