



Data Formats

- Reading: Chapter 4 (Data Formats)



References

- UTF-8 Code table
<http://www.utf8-chartable.de/>



Data Formats

- Computers
 - Process and store all forms of data in binary format
- Data formats:
 - Specifications for converting data into computer-usable form
 - Define the different ways human data may be represented, stored and processed by a computer

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

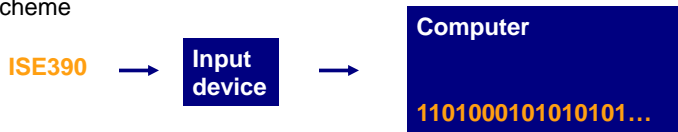
4-3



Sources of Data

- Binary input
 - Begins as discrete input
 - Example: keyboard input such as *ISE390*
 - Keyboard generates a binary number code for each key
- Analog
 - Continuous data such as sound or images
 - Requires hardware to convert data into binary numbers

Figure 3.1 with this color scheme



Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-4

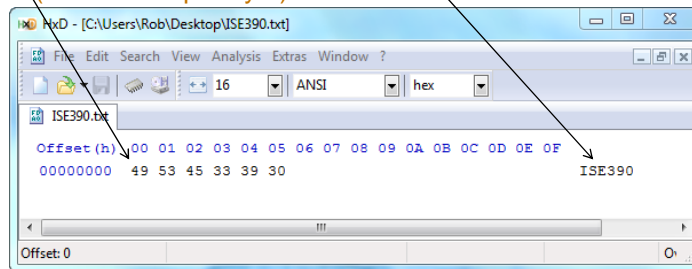


How is ISE390 Stored

- HxD shows a file computer layout
- ISE390.txt contains the string ISE390

Hex representation
(1 character per byte)

Character representation



Is the hex ordering the same as alphabetic ordering?

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

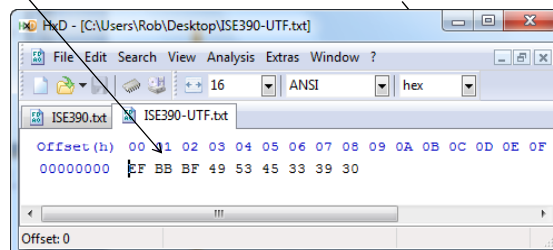
4-5



How is ISE390 Stored in UTF

- ISE390-UTF.txt contains the string ISE390 stored in UTF-8


Compare this to the ASCII representation



What are the first 3 bytes?

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly


4-6



Common Data Representations

Type of Data	Standard(s)
Alphanumeric	Unicode, ASCII, EDCDIC
Image (bitmapped) UTF-8 is a multibyte encoding for Unicode	<ul style="list-style-type: none"> ▪ GIF (graphical image format) ▪ TIF (tagged image file format) ▪ PNG (portable network graphics)
Image (object)	PostScript, JPEG, SWF (Macromedia Flash), SVG
Outline graphics and fonts	PostScript, TrueType
Sound	WAV, AVI, MP3, MIDI, WMA
Page description	PDF (Adobe Portable Document Format), HTML, XML
Video	Quicktime, MPEG-2, RealVideo, WMV

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly 4-7



Internal Data Representation

- Reflects the
 - Complexity of input source
 - Type of processing required
- Trade-offs
 - Accuracy and resolution
 - Simple photo vs. painting in an art book
 - Compactness (storage and transmission)
 - More data required for improved accuracy and resolution
 - *Compression* represents data in a more compact form
 - *Metadata*: data that describes or interprets the meaning of data
 - Ease of manipulation:
 - Processing simple audio vs. high-fidelity sound
 - Standardization
 - *Proprietary formats* for storing and processing data (WordPerfect vs. Word)
 - De facto standards: proprietary standards based on general user acceptance (PDF)

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly 4-8



Data Types: Numeric

- Used for mathematical manipulation
 - Add, subtract, multiply, divide
- Types
 - Integer (whole number)
 - Real (contains a decimal point)
- Covered later in the course

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-9



Data Types: Alphanumeric

- Alphanumeric:
 - Characters: *b T*
 - Number digits: *7 9*
 - Punctuation marks: *! ;*
 - Special-purpose characters: *\$ &*
- Numeric characters vs. numbers
 - Both entered as ordinary characters
 - Computer converts into numbers for calculation
 - Example: int in Java
 - Treated as characters if processed as text
 - Examples: Phone numbers, ZIP codes

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-10



Alphanumeric Codes

- Encoding - bits that represent characters
- Value of binary number representing character corresponds to placement in the alphabet
- Compactness vs. expressiveness (efficient storage vs. total number of characters)
- Consistency among devices
- Internet exchange of documents includes a statement by server of the document encoding

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-11



Representing Characters

- ASCII – dated, but still a widely used coding scheme
- Unicode: developed for worldwide use
- Special purpose (e.g., EBCDIC)

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-12



ASCII

- Developed by ANSI (American National Standards Institute)
- Represents
 - Latin alphabet, Arabic numerals, standard punctuation characters
 - Plus small set of accents and other European special characters
- ASCII
 - 7-bit code: 128 characters

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-13



ASCII Reference Table

MSD \ LSD	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P		p
1	SOH	DC1	!	1	A	Q	a	W
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACJ	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(8	H	X	h	x
9	HT	EM)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[k	{
C	FF	FS	,	<	L	\	l	
D	CR	GS	-	=	M]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

74₁₆
111 0100

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-14



Unicode

- Standard for consistent encoding
- Represents more than 109,000 characters
- Can be implemented by different character encodings (e.g., UTF-8)
- UTF-8 encoding uses variable representation
- ASCII Latin-I subset of Unicode
 - Values 0 to 255 in Unicode table
- Multilingual: defines codes for
 - Nearly every character-based alphabet
 - Large set of ideographs for Chinese, Japanese and Korean
 - Composite characters for vowels and syllabic clusters required by some languages

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-15



Editing and Viewing Hex Codes

- HxD – Editing hex codes of a txt file
<http://www.editpadlite.com/>
- EditPad Lite – viewing characters in various code representations
[download.cnet.com/HxD-Hex-Editor/
3000-2352_4-10891068.html](http://download.cnet.com/HxD-Hex-Editor/3000-2352_4-10891068.html)

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-16



Are We on Track?

- Create a document containing your name followed by the special characters represented in UTF-8 with the hex numbers
 - 24 Examples in Wikipedia UTF-8 page
 - C2 A2
 - E2 82 AC
 - F0 A4 AD A2
- Use HxD and EditPad

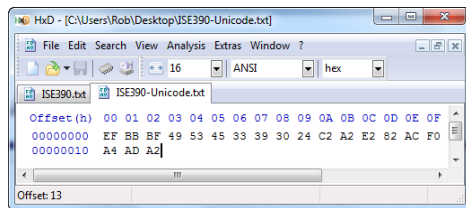
Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-17

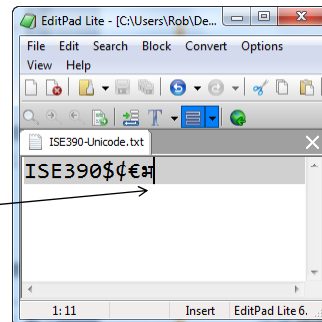


Were We on Track?

- Results



Is the last character correct?



Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-18



Collating Sequence

- Alphabetic sorting if software handles mixed upper- and lowercase codes
- In ASCII, numbers collate first;
- ASCII collating sequence for string of characters

Letters						Numeric Characters				
Adam	A	d	a	m		1	011	0001		
Adamian	A	d	a	m	i	a	n			
Adams	A	d	a	m	s					
						2	011	0010		
						12	011	0001	011	0010

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-19



2 Classes of Codes

- *Printing* characters
 - Produced on the screen or printer
- *Control* characters
 - Control position of output on screen or printer
 - VT: vertical tab
 - LF: Line feed
 - Cause action to occur
 - BEL: bell rings
 - DEL: delete current character
 - Communicate status between computer and I/O device
 - ESC: provides extensions by changing the meaning of a specified number of contiguous following characters

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-20



Keyboard Input

- *Scan code*
 - Two different scan codes on keyboard
 - One generated when key is struck and another when key is released
 - Converted to Unicode, ASCII or EBCDIC by software in terminal or PC
- Advantage
 - Easily adapted to different languages or keyboard layout
 - Separate scan codes for key press/release for multiple key combinations
 - Examples: shift and control keys

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-21



Other Alphanumeric Input

- *OCR* (optical character reader)
 - Scans text and inputs it as character data
 - Used to read specially encoded characters
 - Example: magnetically printed check numbers
- *Bar Code Readers*
 - Used in applications that require fast, accurate and repetitive input with minimal employee training
 - Examples: supermarket checkout counters and inventory control
- *Magnetic stripe reader*: alphanumeric data from credit cards
- *RFID*: store and transmit data between RFID tags and computers
- *Voice*
 - Digitized audio recording common but conversion to alphanumeric data difficult
 - Requires knowledge of sound patterns in a language (*phonemes*) plus rules for pronunciation, grammar, and syntax

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-22



Are We on Track

- Create another short text document in a more expressive format (e.g., WordPad)
- Open the document in HxD.
- What do you see?

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-23



Image Data

- Photographs, figures, icons, drawings, charts and graphs
- Two approaches:
 - *Bitmap* or *raster images* of photos and paintings with continuous variation
 - *Object* or *vector images* composed of *graphical objects* like lines and curves defined geometrically (e.g., SVG)
- Differences include:
 - Quality of the image
 - Storage space required
 - Time to transmit
 - Ease of modification

We will discuss image display devices later in the course

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-24



Bitmap Images

- Used for realistic images with continuous variations in shading, color, shape and texture
 - Examples:
 - Scanned photos
 - Clip art generated by a *paint* program
- Preferred when image contains large amount of detail and processing requirements are fairly simple
- Input devices:
 - Scanners
 - Digital cameras and video capture devices
 - Graphical input devices like mice and pens
- Managed by *photo editing software* or *paint software*
 - Editing tools to make tedious bit by bit process easier

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-25



Bitmap Images

- Each individual *pixel* (*pi(x)cture element*) in a graphic stored as a binary number
 - Pixel: A small area with associated coordinate location
 - Example: each point in the happy face is represented by an RGB binary string (24 bit most common)

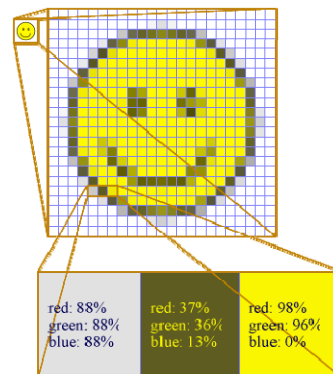


Image from knowledgerush.com

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-26



Bitmap Display

- Monochrome: black or white
 - 1 bit per pixel
- Gray scale: black, white or 254 shades of gray
 - 1 byte per pixel
- Color graphics: 16 colors, 256 colors, or 24-bit true color (16.7 million colors)
 - 4, 8, and 24 bits respectively

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-27



Storing Bitmap Images

- Frequently large files
 - Example: 600 rows of 800 pixels with 1 byte for each of 3 colors → ~1.5MB file
- File size affected by
 - *Resolution* (the number of pixels)
 - Amount of detail affecting clarity and sharpness of an image
 - Levels: number of bits for displaying shades of gray or multiple colors
 - *Palette*: color translation table that uses a code for each pixel rather than actual color value
 - Data compression Think of a color palette as the number of colors that can be displayed simultaneously

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-28



GIF (Graphics Interchange Format)

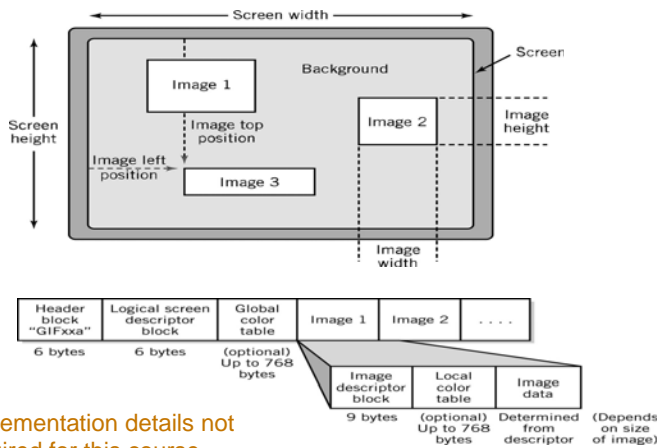
- First developed by CompuServe in 1987
- GIF89a enabled animated images
 - allows images to be displayed sequentially at fixed time sequences
- Color limitation: 256
- Image compressed by LZW (Lempel-Zif-Welch) algorithm
- Preferred for line drawings, clip art and pictures with large blocks of solid color
- *Lossless compression*

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-29



GIF (Graphics Interchange Format)



Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-30



JPEG (Joint Photographers Expert Group)

- Allows more than 16 million colors
- Suitable for highly detailed photographs and paintings
- Employs *lossy compression* algorithm that
 - Discards data to decrease file size and transmission speed
 - May reduce image resolution, tends to distort sharp lines

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-31




Are We on Track?


- Install ImageAnalyzer
[download.cnet.com/Image-Analyzer/
3000-2192_4-10429018.html](http://download.cnet.com/Image-Analyzer/3000-2192_4-10429018.html)
- Open an image
- Select Operations/Show image information
 - What are the image dimensions?
 - How many colors are used?
 - What is the file size?
 - What is the format?
- Open a very different image and compare results

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-32




Were We on Track?



Information

File name: CountyCork.jpg
 Image dimensions: 1920 x 1200
 Pixel format: 24 bit
 Total number of pixels: 2,304,000
 Number of colors used: 104,811
 -- EXIF --
 File size: 820kb
 File date: 3/23/2007
 Photo date: 2006:11:08 01:35:57
 Make (Model): NIKON CORPORATION (NIKON D70s)
 Exposure time: 1/40 sec
 Focal length: 18.00 mm
 FNumber: f8.0
 Flash: No

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly
4-33



Object Images

- Composed of lines and shapes in various colors
- Computer translates geometric formulas to create the graphic
- Storage space depends on image complexity
 - number of instructions to create lines, shapes, fill patterns
- Suitable for high quality images in varying resolutions (similar to TrueType fonts)
- Cannot be displayed or printed directly

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly
4-34



Video Images

- Require massive amount of data
 - Video camera producing full screen 640 x 480 pixel true color image at 30 frames/sec → 27.65 MB of data/sec
 - 1-minute film clip → 1.6 GB storage
- Options for reducing file size (or streaming rate): decrease resolution and/or reduce frame rate
- Method depends on how video delivered to users
 - *Streaming video*: video displayed as it is downloaded from the Web server (generally compressed)
 - Local data (file on DVD or downloaded onto system) for higher quality

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-35



Audio Data

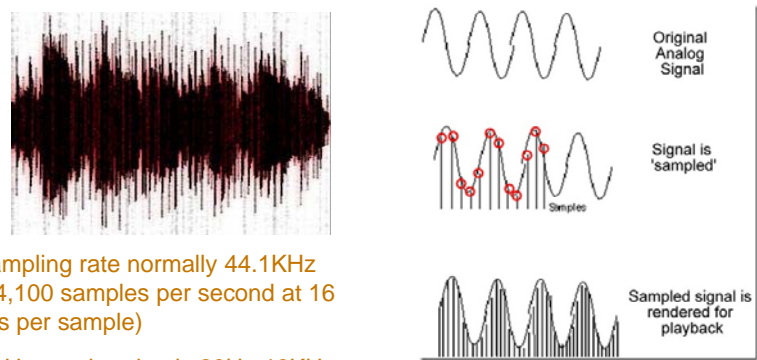
- Transmission requirements less demanding than those for video
- Waveform audio: digital representation of sound
- Analog sound converted to digital values by Analog-to-Digital (A-to-D) converter
- In a computer, the digital audio data is converted to analog signals with a digital-to-analog (DAC) converter
- Analog output signals connected to amplifiers and/or headphones
- Sound processing integrated into computers

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

4-36

Waveform Audio

- Waveform -a graph that charts minute changes in air pressure as sound waves propagate



Sampling rate normally 44.1KHz
(44,100 samples per second at 16 bits per sample)

Human hearing is 20Hz-16KHz


Images from Michael Hanley Consulting and Dan Cragan Music Services

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly 4-37

Data Compression

- Compression:** recoding data so that it requires fewer bytes of storage space.
- Compression ratio:** the amount file is shrunk
- Lossless:** inverse algorithm restores data to exact original form
 - Examples: GIF, TIFF
- Lossy:** trades off data degradation for file size and download speed
 - Much higher compression ratios, often 10 to 1
 - Example: JPEG
 - Common in multimedia

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly 4-38




Audio Formats

Lots of different digital audio formats

- Uncompressed (~10MB/minute) Note the different measures used
 - AIFF – primarily Apple
 - WAV – primarily Windows lossless
 - CDA – music CD
- Compressed, but lossless (~ 5MB/minute)
 - FLAC – Free Lossless Audio Codec
 - ALAC – Apple Lossless Audio Codec

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly 4-39



Audio Formats (continued)

- Lossy (~1MB/minute)
 - MP3
 - AAC (Apple) – better compression than MP3 (~25%)
 - WMA – Windows Media Audio (not in wide use)
 - Vorbis – free open-source format (inferior quality)

AAC is the standard format for iPod, iTunes, iPhone, etc.)

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly 4-40



Internal Computer Data Format

- All data stored as binary numbers
- Interpreted based on
 - Operations computer can perform
 - Data types supported by programming language used to create application

The next major topic is the internal representation of numerical data



Java Primitives

Java Primitive Data Types (8)

Type	Contains	Default	Size	Range
boolean	true or false	false	1 bit	NA
char	Unicode character unsigned	\u0000	16 bits or 2 bytes	0 to 2 ¹⁶ -1 or \u0000 to \uFFFF
byte	Signed integer	0	8 bit or 1 byte	-2 ⁷ to 2 ⁷ -1 or -128 to 127
short	Signed integer	0	16 bit or 2 bytes	-2 ¹⁵ to 2 ¹⁵ -1 or -32768 to 32767
int	Signed integer	0	32 bit or 4 bytes	-2 ³¹ to 2 ³¹ -1 or -2147483648 to 2147483647
long	Signed integer	0	64 bit or 8 bytes	-2 ⁶³ to 2 ⁶³ -1 or -9223372036854775808 to 9223372036854775807
float	IEEE 754 floating point single-precision	0.0f	32 bit or 4 bytes	±1.4E-45 to ±3.4028235E+38
double	IEEE 754 floating point double-precision	0.0	64 bit or 8 bytes	±439E-324 to ±1.7976931348623157E+308

From Javacamp