

# Introduction to Data and Encoding

## What is a digital file?



[Wikipedia: Filename extension](#)

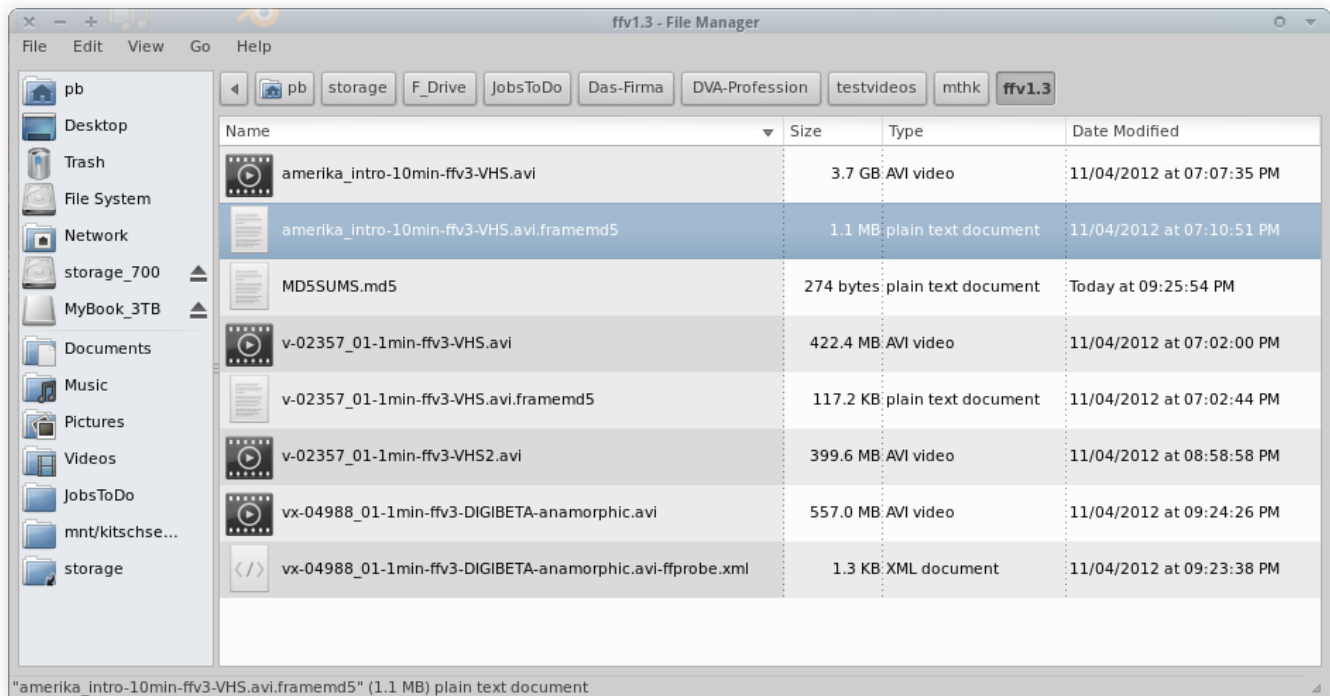
Most people identify what a file is, according to its filename - and the filetype according to its suffix after the “.” dot.

If all is well, usually a quick and sane choice, but there's more...

# Understanding digital objects

- **Bit:**  
A single binary digit (0/1)
- **Byte:**  
A unit: 8 bits (half = **Nibble**)
- **File:**  
Stored segment or block of information available to a computer program
- **File system:**  
A mechanism for controlling and organizing bytes into structure (files/folders) for storage and retrieval
- **File Format:**  
A standard way that information is encoded in a computer file.

# Identifying files



## Speaker notes

What can you say about these files?

These file properties (filename, date/time, size, ownership, access rights, flags) can often be used to say something/more about a digital object, therefore it's good to consider preserving this layer of information too.

For example, when documenting the original state of externally acquired collections/objects. More about this in the metadata session...

# The Filesystem

- Filename
- Date/time
- Filesize
- File extension
- Path
- Access rights

## Speaker notes

Without a filesystem, data on a storage device is just a long string of numbers... No beginning, no end, no structure, no folders, no files. Just bytes!

If your filesystem is broken, you can't access your data - although the "data" is actually exactly where it was. Untouched. But there's no "map" to find where to go, and where a file starts or ends.

**What is Data?**

**Everything's a number**

# ASCII TABLE

<https://commons.wikimedia.org/wiki/File:ASCII-Table-wide.svg>

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

## Speaker notes

Each byte in a file is a number. Depending on the encoding, each number maps to a certain character. This table shows a common character encoding: “[ASCII](#)” ([American Standard Code for Information Interchange](#))

This view also shows the *hexadecimal* (short “hex”) value which is more common and better to view data as, than decimal.

# Character encoding

- [ASCII \(7 Bit\)](#)
- [CP437](#)
- [ISO-8859-1](#)
- [ISO-8859-7 \(Latin & Greek\)](#)
- ...

See: [Character sets, encodings, and Unicode \(By Nick Gammon\)](#)

## Speaker notes

Classic “code pages” work fine for the language/region they are designed for. Mixing characters from different languages is a problem with this approach though!

Mis-interpreting a character by applying the wrong codepage is the reason for encoding errors. For practical and history reasons, the ASCII set is usually mapped compatible across all codepages.

# Unicode

*“Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world’s writing systems.”*

– [Wikipedia: Unicode](#)

## Mixing languages

Лорем ипсум долор сит амет  
側經意責家方家閉討店暖育田庁載社  
पढाए हिंदी रहारूप अनुवाद कार्यलय  
국민경제의 발전을 위한 중요정책의  
旅口京青利セムレ弱改フヨス  
غينيا واستمر العصبة ضرب قد. وباءت

[UTF-8 encoding table](#)



# Unicode Symbols

- U+1F973 🍷
- U+262F ☯
- U+1F643 😬

[Emoji List](#)

## Comments?

## Questions?