What is text?
What is text analysis?
What is distant reading?

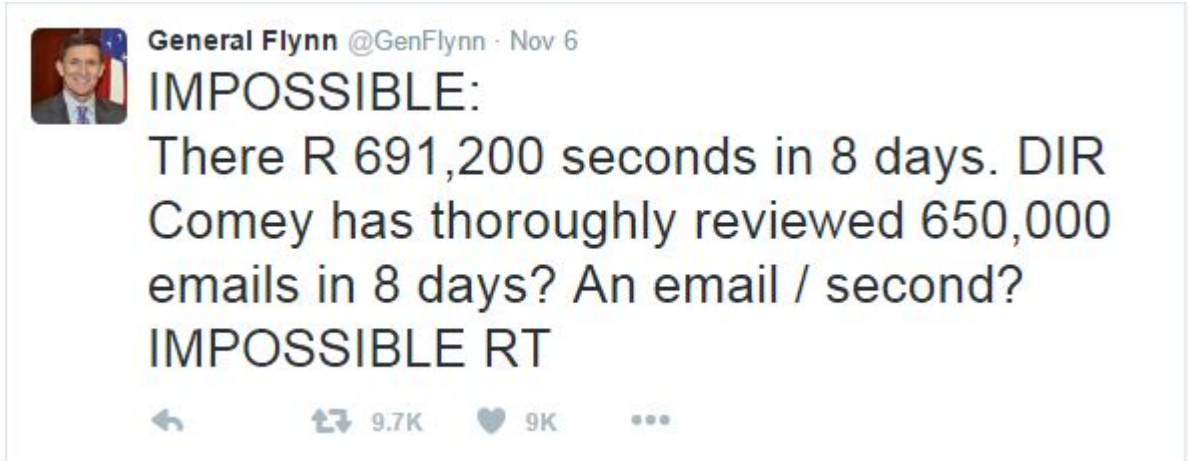
Christopher Ohge, Martin Steer

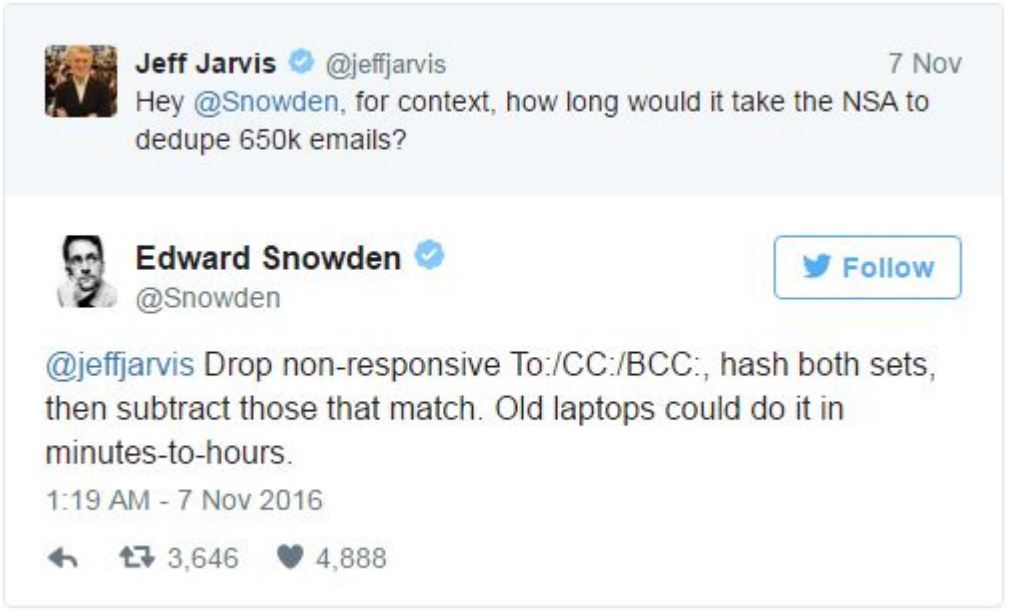
Riga Technical University, September 2019

Plan for today

- Layers of representation
- Plain text
- Text analysis: A Survey of Principles, Tools, and New Ways of Reading
- Distant reading: a new way of assessing history
- Hathi-Trust Corpus extracted features

Claim: You can’t review 650,000 emails in eight days.





Step 1: Search for following patterns:
To: Hillary Clinton
CC: Hillary Clinton
BCC: Hillary Clinton
From: Hillary Clinton
Step 2: De-duplicate copies from previous set.
Step 3: Examine remaining emails.

Plain text is best?

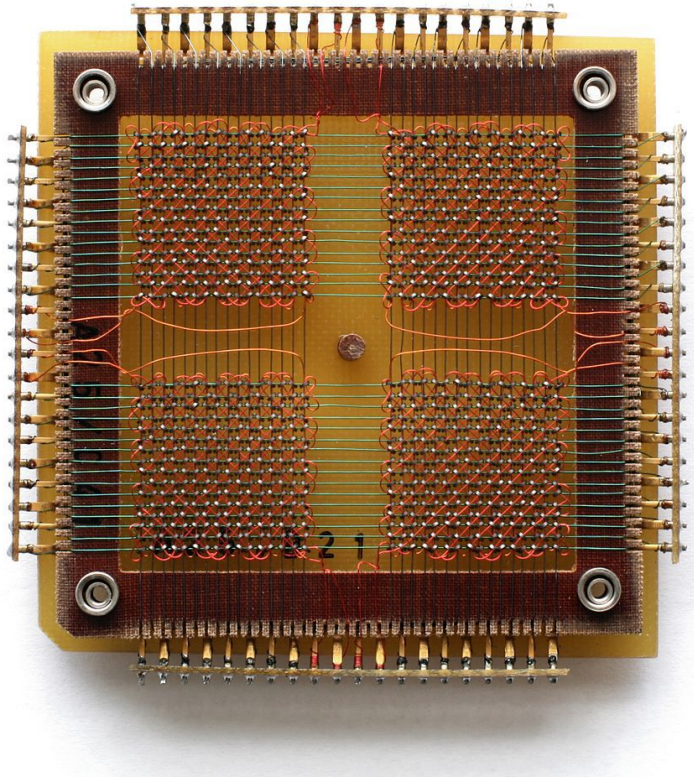
Plain text is best?

- Ladder of abstraction ==>

- Storage
  ||
  ||
  \/
- Applications

Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



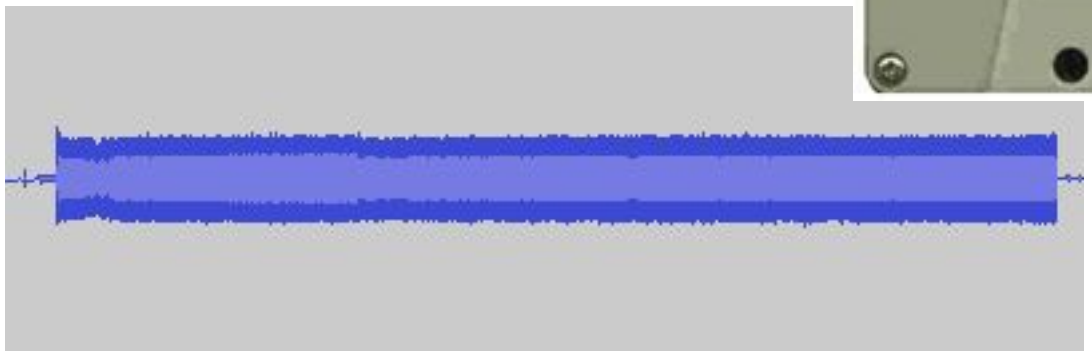
Ladder of abstraction

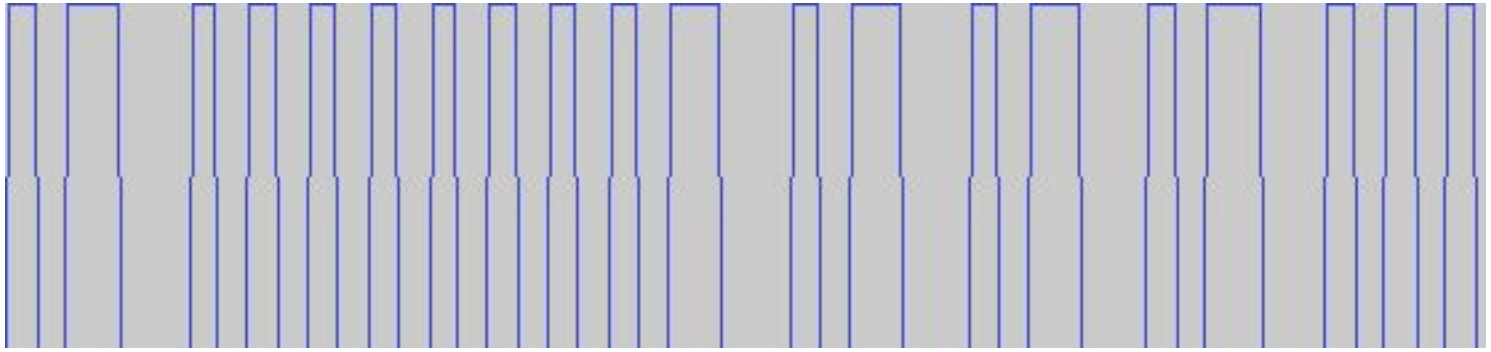
- Storage
- Encoding
- Formats
- Schemas
- Applications

https://en.wikipedia.org/wiki/Videodisc
https://flashbak.com/blank-vhs-cassette-packaging-design-trends-art-402545/

Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications

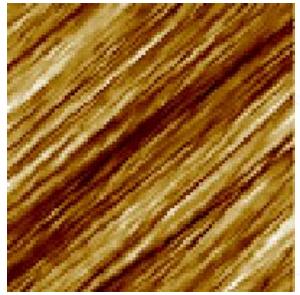


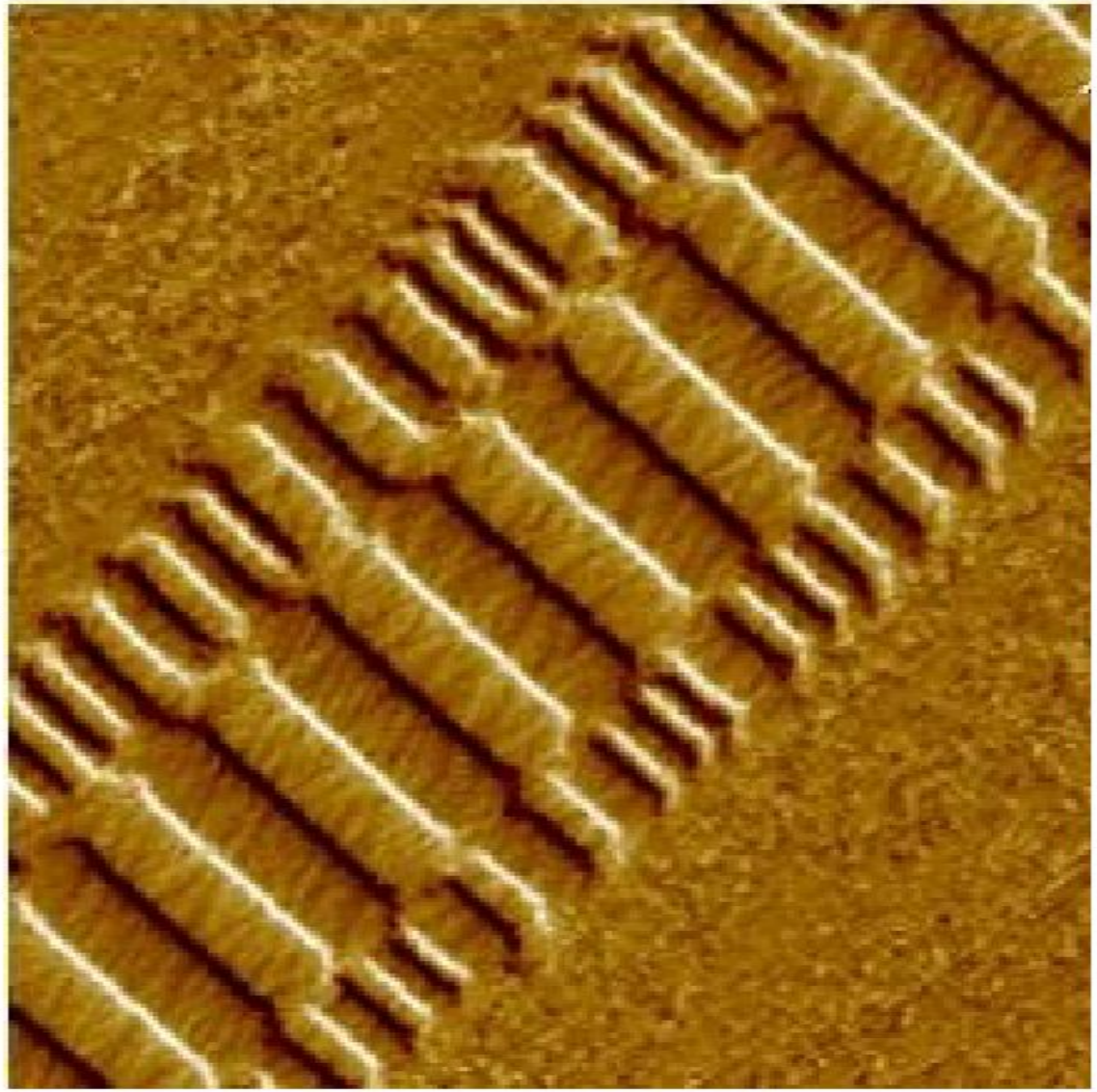


https://www.pagetable.com/?p=32

Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications

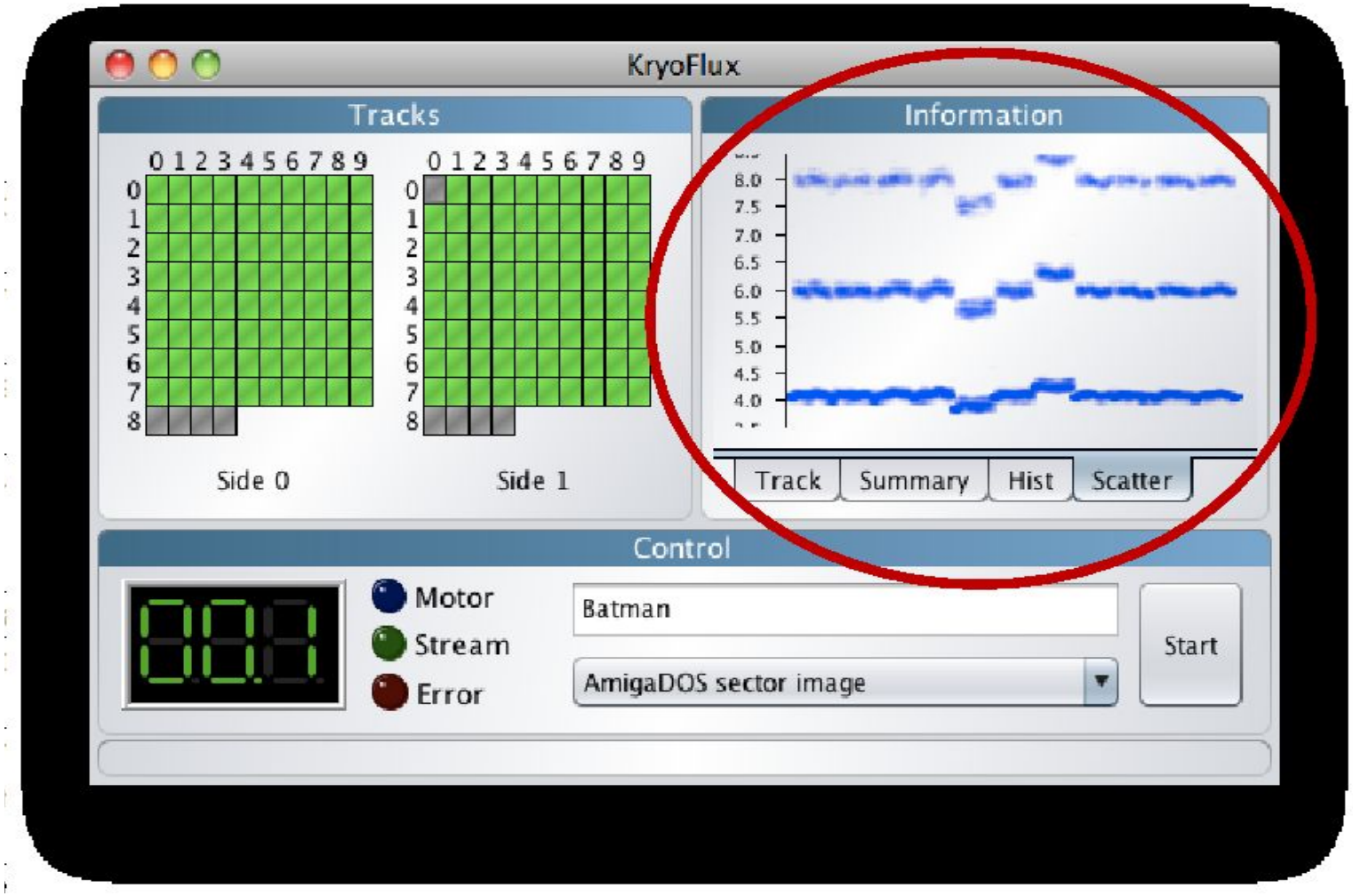




Magnetic Force Microscopy - https://www.engr.sjsu.edu/WofMatE/Mat%27sChar2.htm

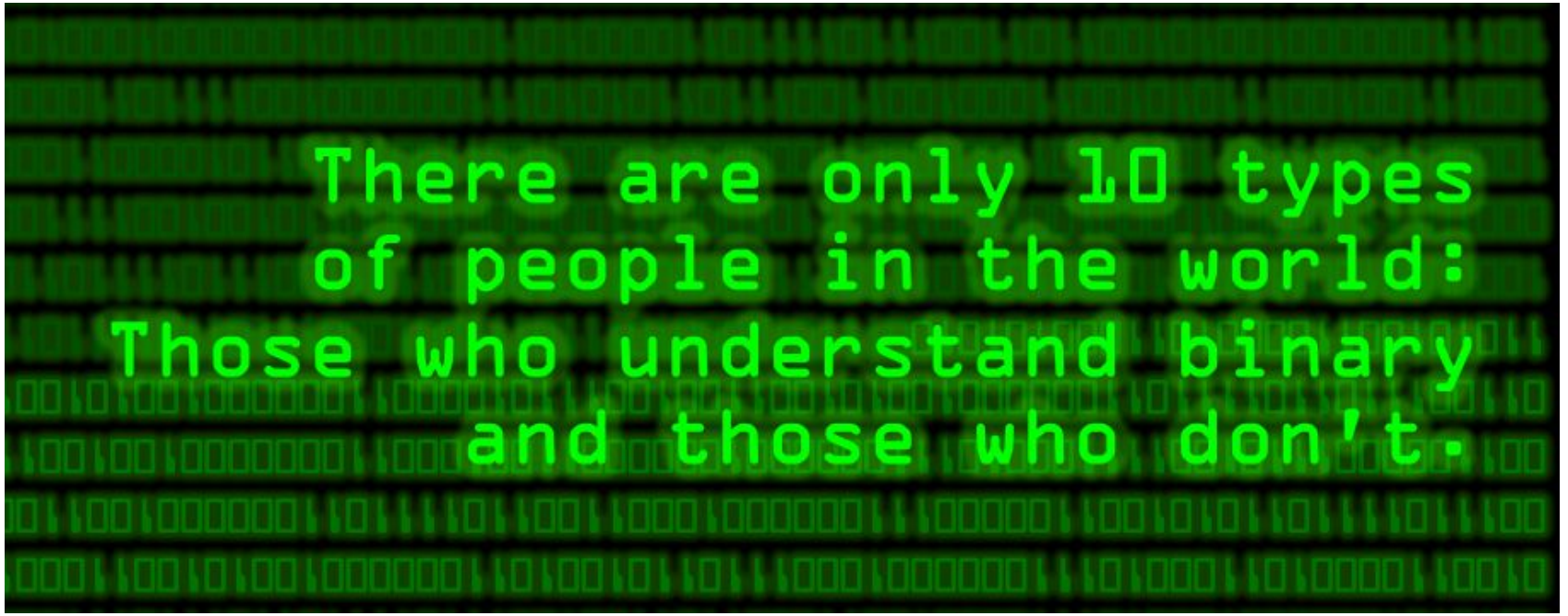
Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications

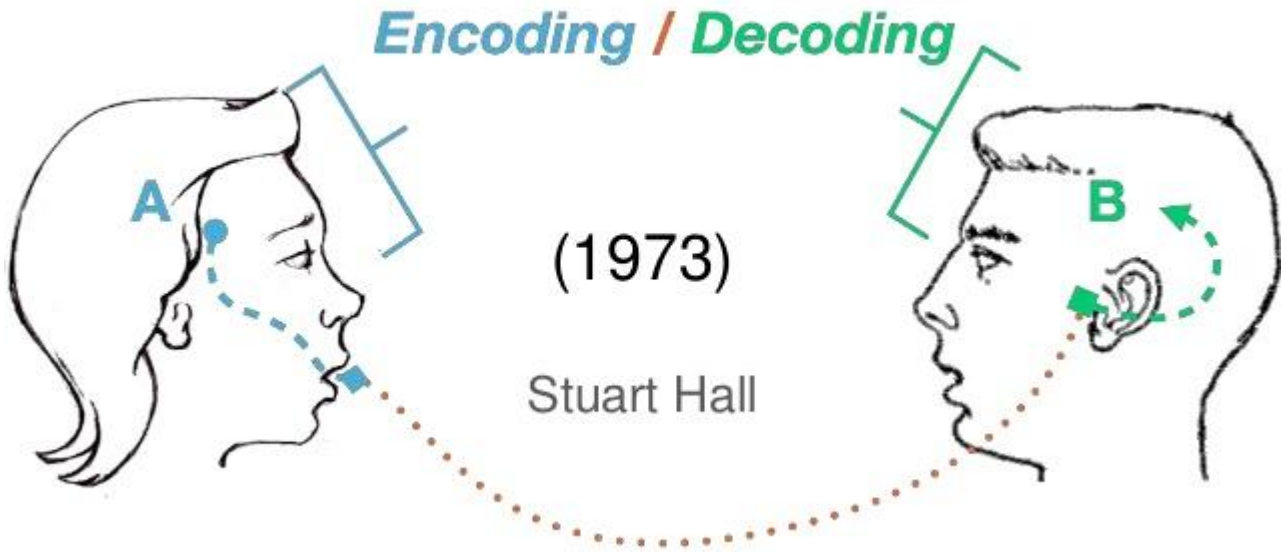
“Obsolete power corrupts obsoletely.” - Ted Nelson

The technology associated with interpreting the representation at each of the layers can change or become less available

https://en.wikipedia.org/wiki/Ted_Nelson

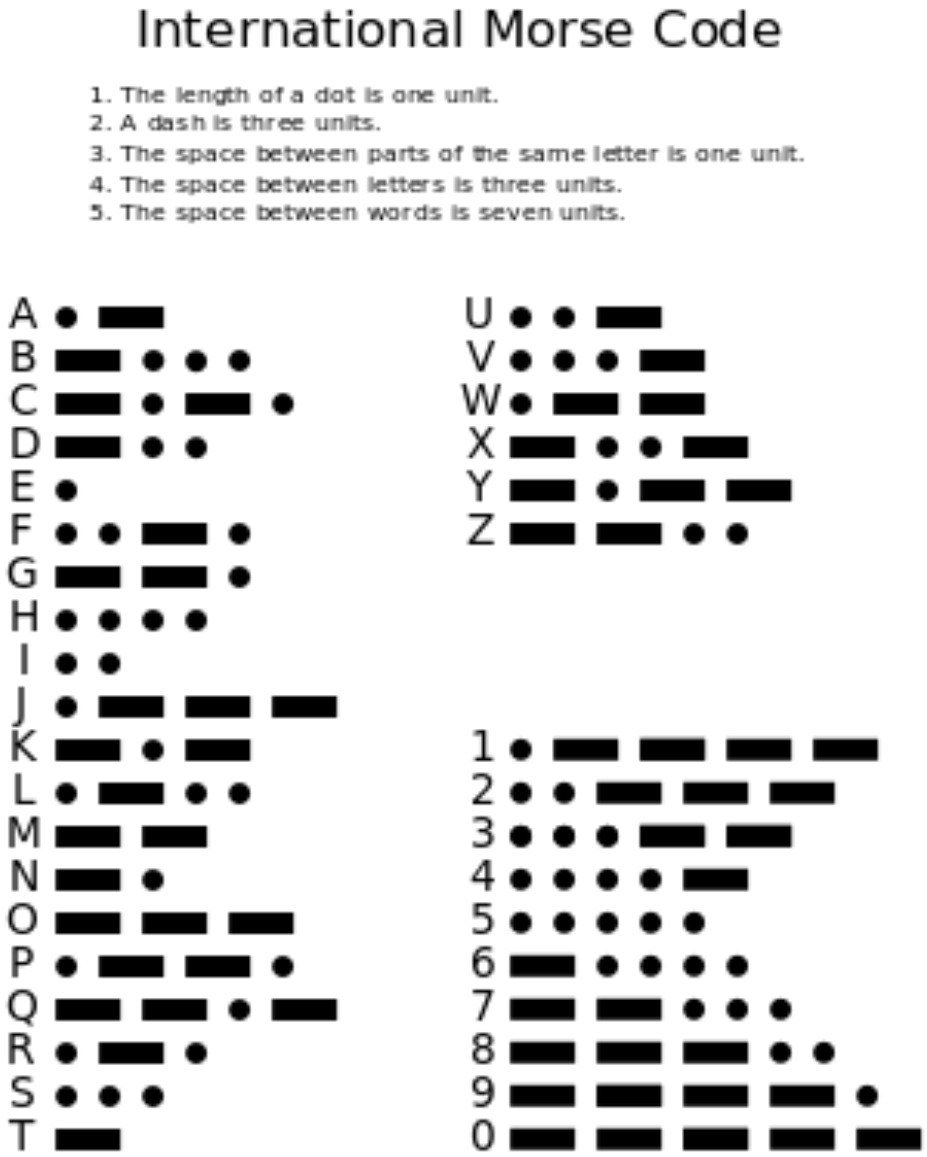
Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications

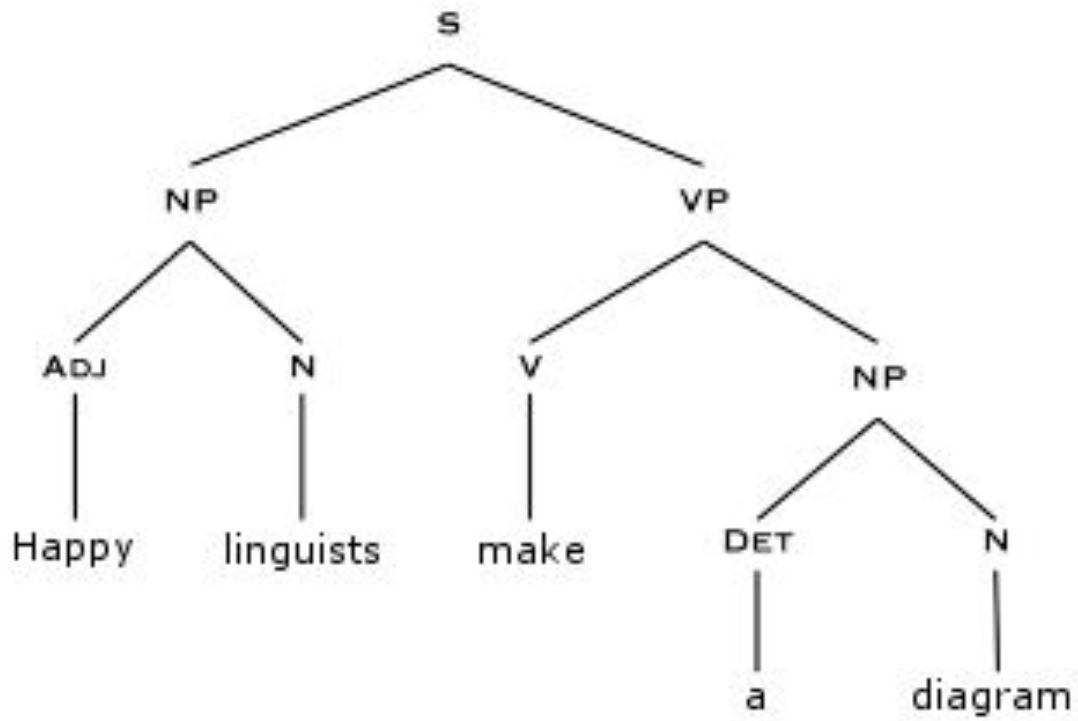


Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications

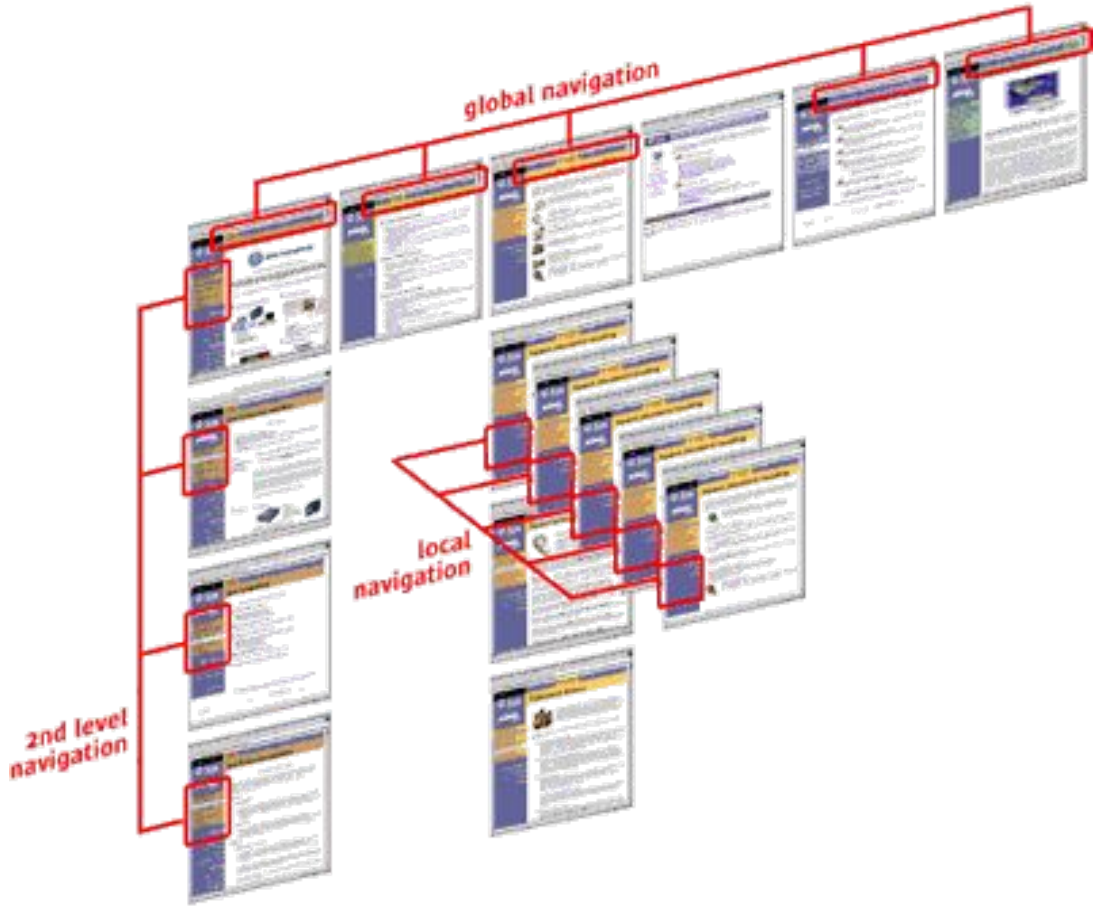
Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



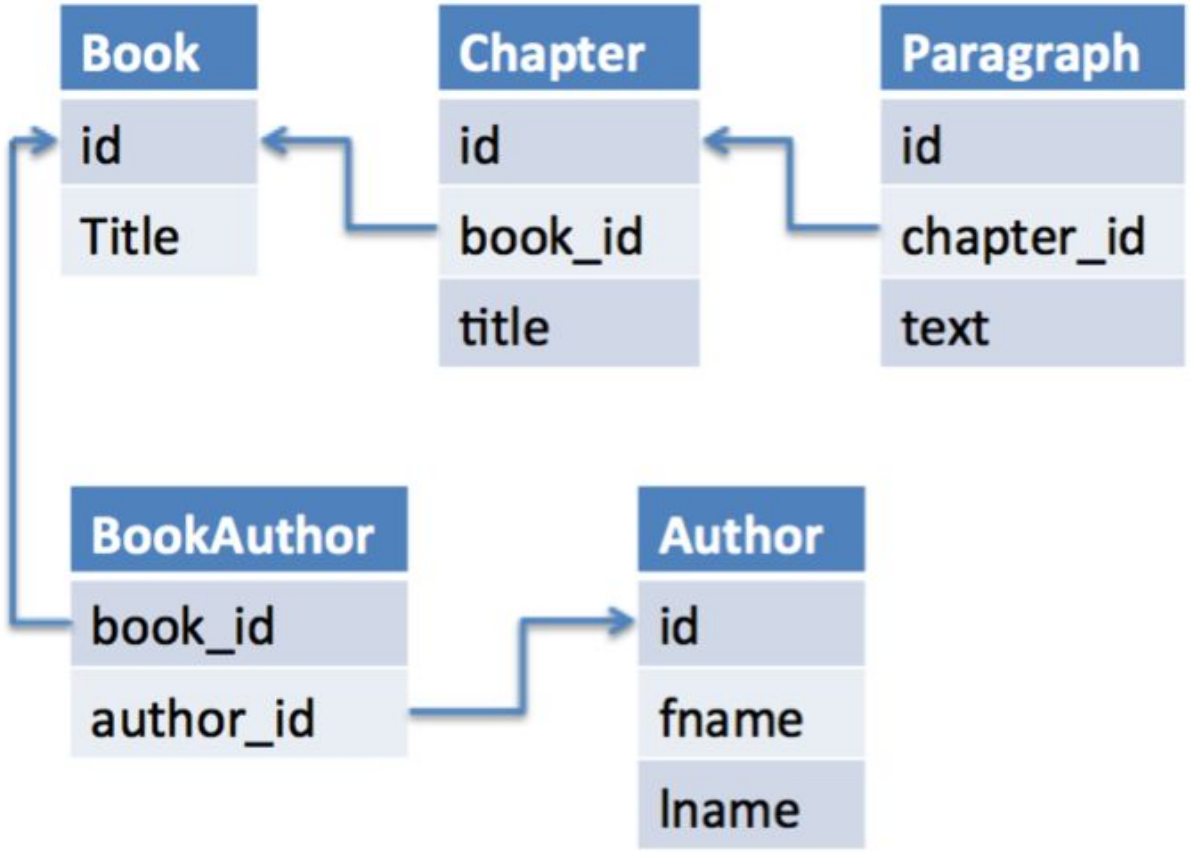
Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



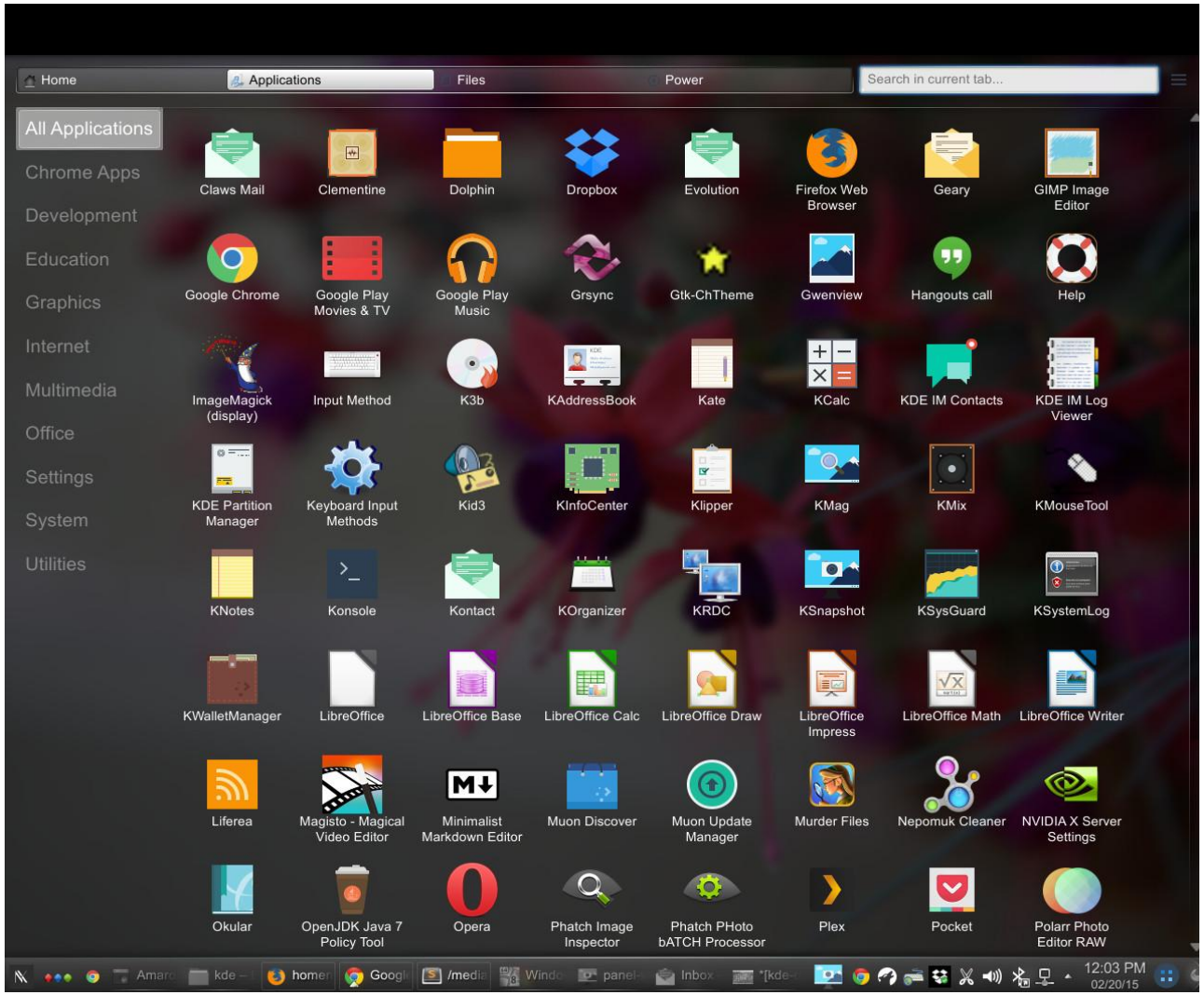
Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



Ladder of abstraction

- Storage
- Encoding
- Formats
- Schemas
- Applications



Ladder of (text) abstraction

- Storage =>
- Encoding =>
- Formats =>
- Schemas =>
- Applications =>

Ladder of (text) abstraction

- Storage => Bits, Bytes, Binary
- Encoding => Coding schemes, character sets
- Formats => XML, JSON, TSV, Word, Excel, Zip, Text
- Schemas => HTML, RDF, TEI
- Applications => MS Word, MS Excel, MS Access, Text editors

Ladder of (text) abstraction

- Storage => Bits, Bytes, Binary
- Encoding => Coding schemes, character sets
- Formats => XML, JSON, TSV, Word, Excel, Zip, Text
- Schemas => HTML, RDF, TEI
- Applications => MS Word, MS Excel, MS Access, Text editors

Working with text methods and tools

Ladder of (text) abstraction

- Storage => B
- Encoding => C
- Formats => X
- Schemas => H
- Applications => M

ter sets

Excel, Zip, Text

Ac

Wo

ods

"Obsolete power corrupts obsoletely." - Ted Nelson

The technology associated with interpreting the representation at each of the layers can change or become less available

https://en.wikipedia.org/wiki/Ted_Nelson

Storage

- Bits
- Bytes
- Binary
- Text

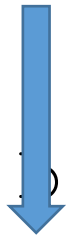
Storage

- Bits
- Bytes
- Binary
- Text

```
0 or 1
01100010
01100010 01101001 01110100 01110011
b    i    t    s
```

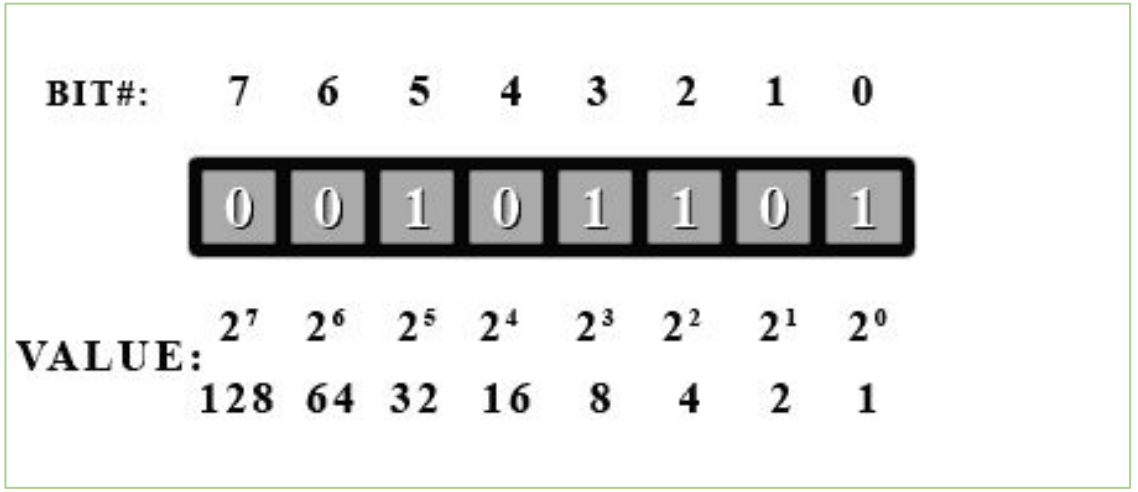
Storage

- Bits
- Bytes
- Binary
- Decimal
- Octal
- Hexadecimal
- Text



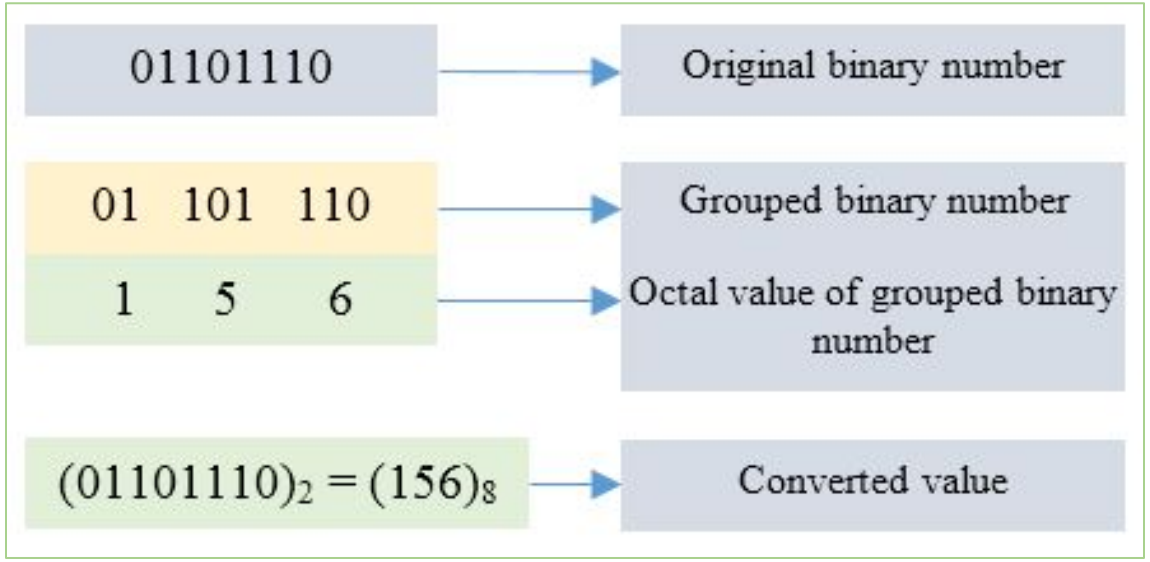
```
01100010 01101001 01110100 01110011

   ???

b    i     t     s
```

Storage



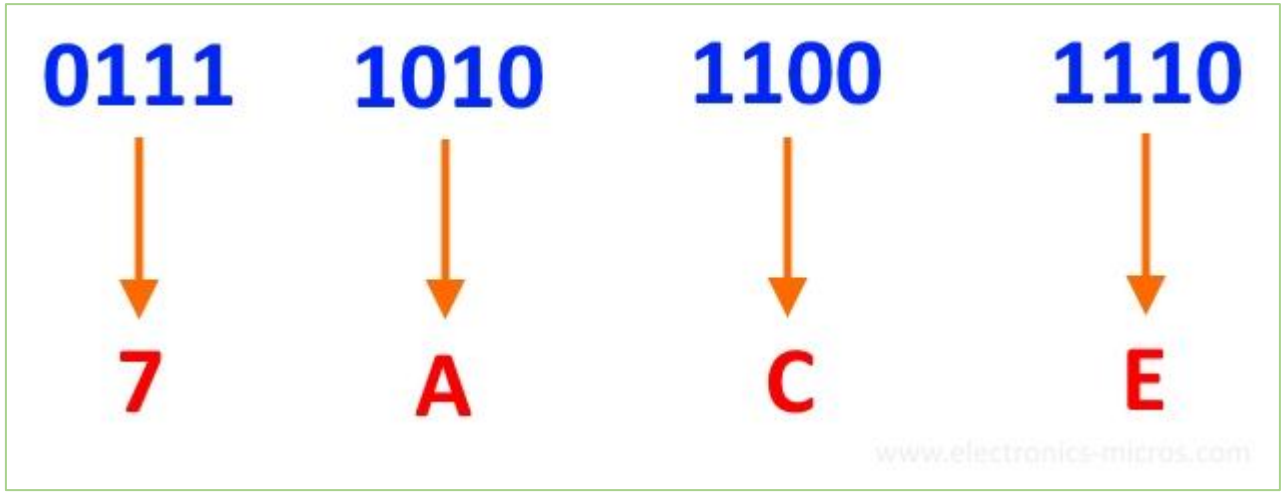
• Bits	0 or 1			
• Bytes	01100010			
• Binary	01100010	01101001	01110100	01110011
• Decimal	98	105	116	115
• Octal	142	151	164	163
• Hexadecimal	62	69	74	73
• Text	b	i	t	s

Storage



- Bits `0 or 1`
- Bytes `01100010`
- **Binary** `01100010 01101001 01110100 01110011`
- Decimal `98 105 116 115`
- **Octal** `142 151 164 163`
- Hexadecimal `62 69 74 73`
- Text `b i t s`

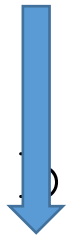
Storage



• Bits	0 or 1			
• Bytes	01100010			
• Binary	01100010	01101001	01110100	01110011
• Decimal	98	105	116	115
• Octal	142	151	164	163
• Hexadecimal	62	69	74	73
• Text	b	i	t	s

Storage

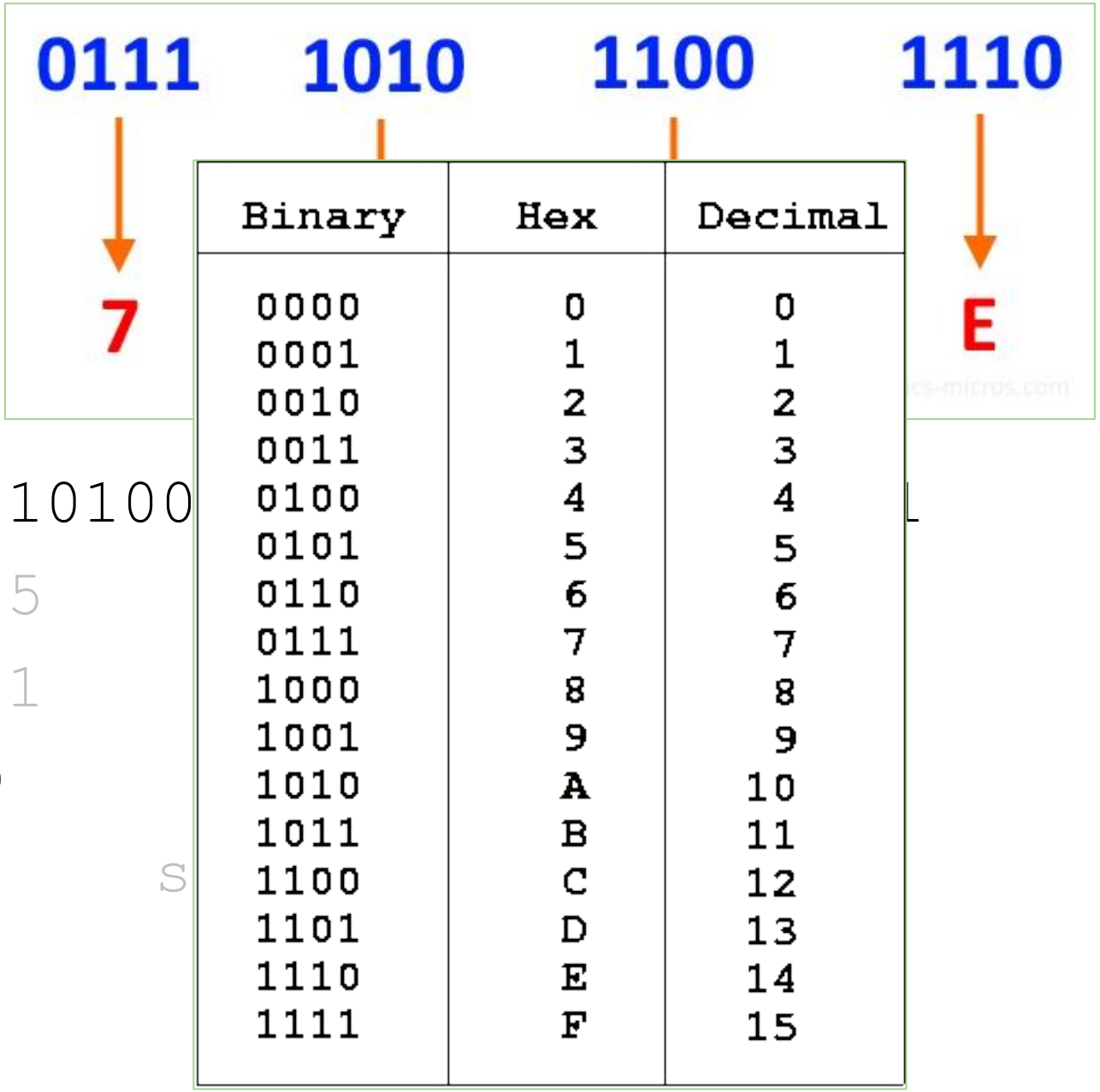
- Bits
- Bytes
- Binary
- Decimal
- Octal
- Hexadecimal
- Text



```
01100010 01101001 01110100 01110011

   ???

b    i     t     s
```

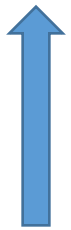
Storage

- Bits `0 or 1`
- Bytes `01100010`
- Binary `01100010 0110100`
- Decimal `98 105`
- Octal `142 151`
- Hexadecimal `62 69`
- Text `b i t s`



Binary	Hex	Decimal
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0110	6	6
0111	7	7
1000	8	8
1001	9	9
1010	A	10
1011	B	11
1100	C	12
1101	D	13
1110	E	14
1111	F	15

Encoding



Storage

Encoding

- Coding schemes
- Character sets

Encoding

- Coding schemes
- Character sets

- Definitions
  - encode – to convert into a coded form

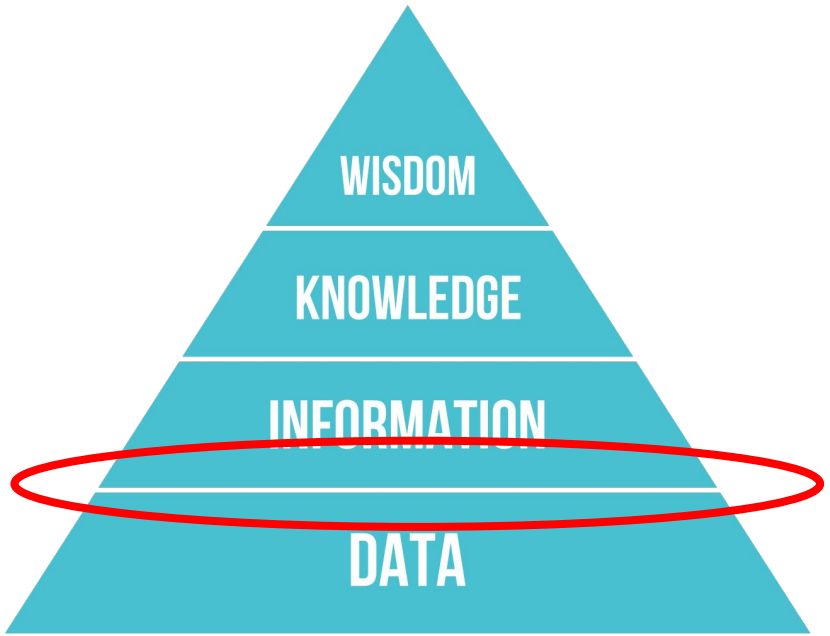
Encoding

- Coding schemes
- Character sets

- Definitions
  - encode – to convert into a coded form
  - code – a system of words, letters, figures or symbols used to represent others

Encoding

- Coding schemes
- Character sets



- Definitions
  - encode – to convert into a coded form
  - code – a system of words, letters, figures or symbols used to represent others
  - Naming systems and structures!

Encoding

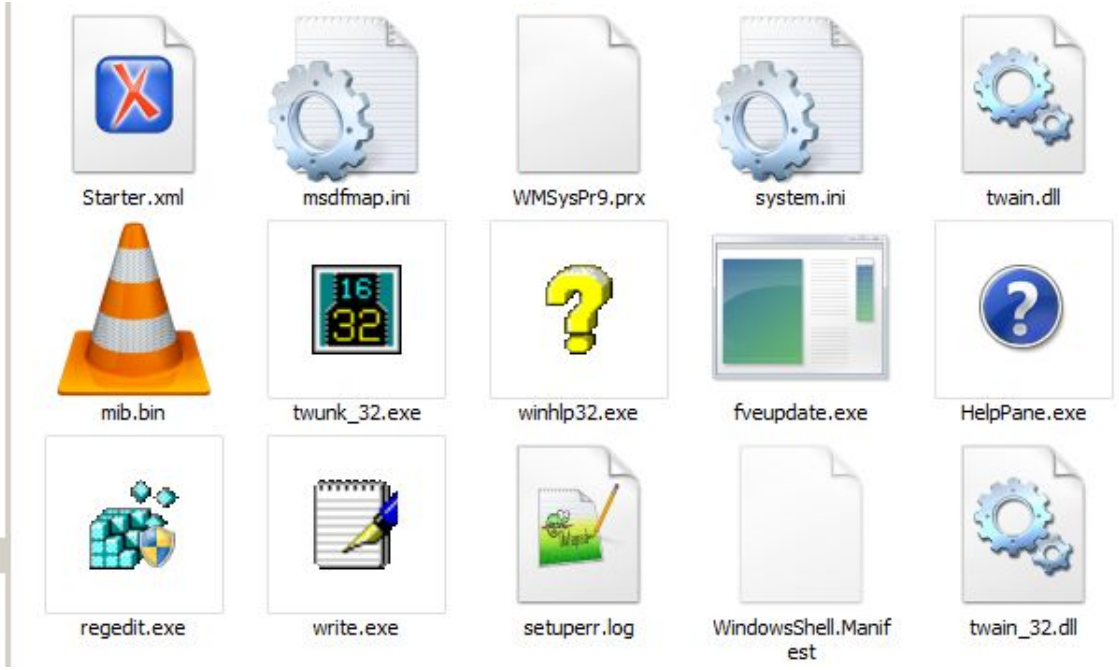
- Coding schemes
- Character sets

Binary	Hex	Decimal
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0110	6	6
0111	7	7
1000	8	8
1001	9	9
1010	A	10
1011	B	11
1100	C	12
1101	D	13

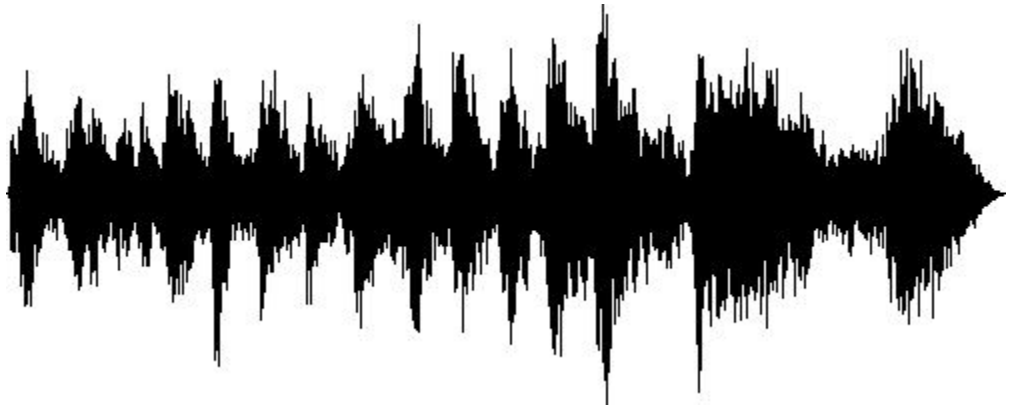
- Schemes
  - Numeric - Binary, Decimal, Octal, Hexadecimal

Encoding

- Coding schemes
- Character sets

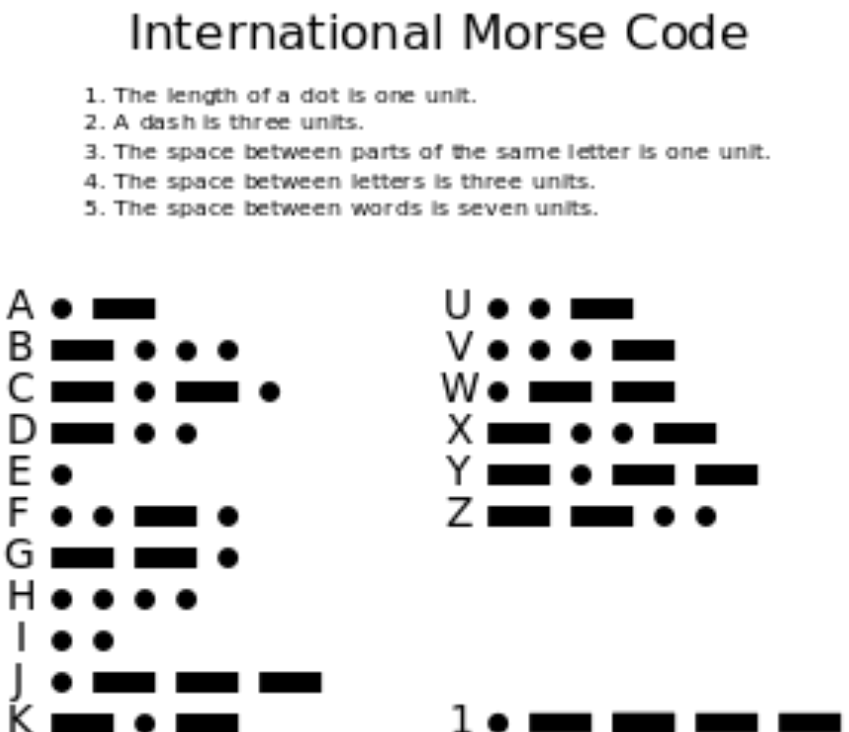


- Schemes
  - Numeric - Binary, Decimal, Octal, Hexadecimal
  - Binary – MP3, AVI, EXE



Encoding

- Coding schemes
- Character sets



- Schemes
  - Numeric - Binary, Decimal, Octal, Hexadecimal
  - Binary – MP3, AVI, EXE
  - Character – Braille, Morse code, ASCII, Unicode, ISO 8859-6 (Arabic)

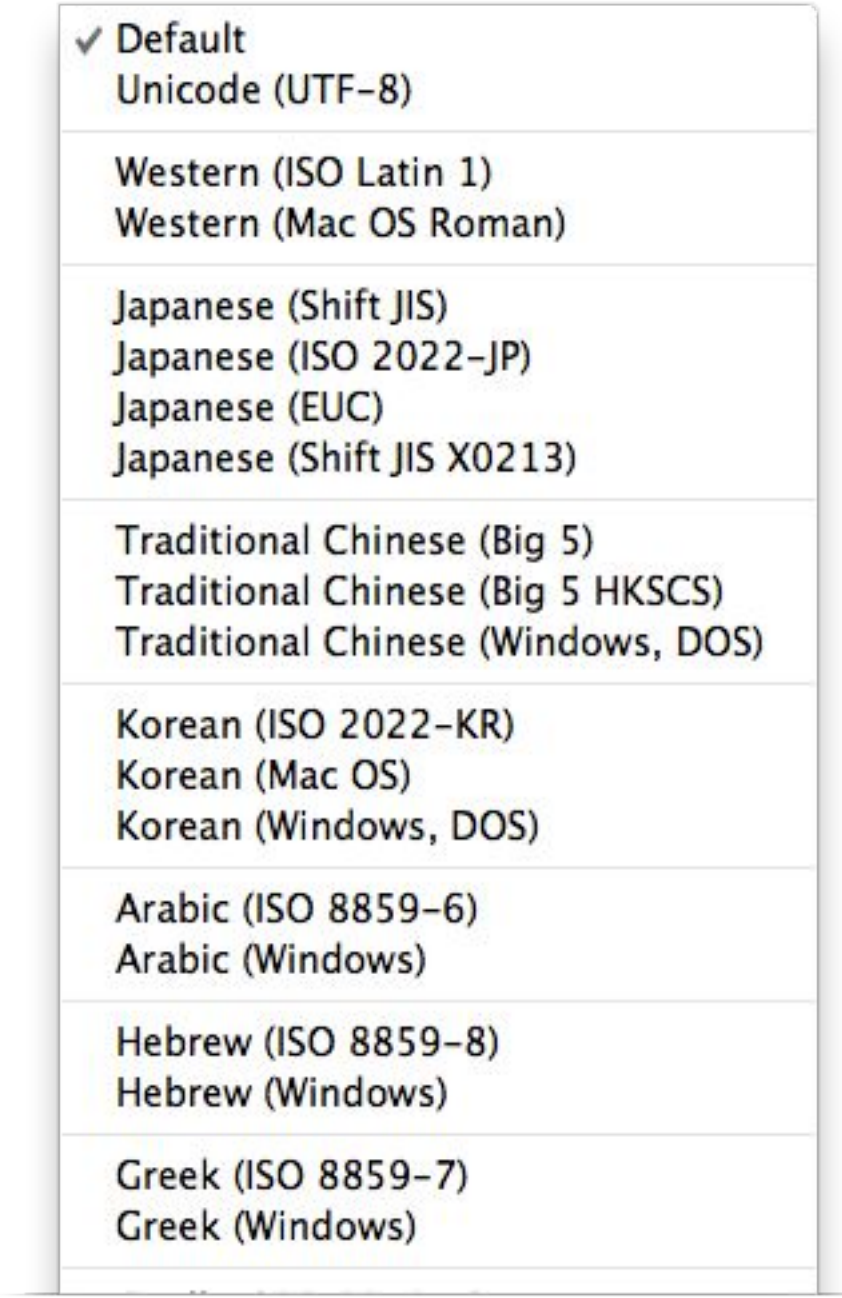
USASCII code chart

b_7 b_6 b_5 → / Bits					0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 ↓	b_3 ↓	b_2 ↓	b_1 ↓	Column → / Row ↓	0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

Encoding

- Coding schemes
- Character sets

https://en.wikipedia.org/wiki/Character_encoding#Common_character_encodings



“ÉGÉìÉRÅ[ÉfÉBÉìÉOÇÕìÔÇµÇ≠Ç»Ç¢”

"エンコーディングは難しくない"

"Unicode replacement character"  (U+FFFD)

bits	encoding	characters
11000100 01000010	Windows Latin 1	ÄB
11000100 01000010	Mac Roman	ƒB
11000100 01000010	GB18030	腂

http://kunststube.net/encoding/

bits	encoding	characters
11000100 01000010	Windows Latin 1	ÄB
11000100 01000010	Mac Roman	*f*B
11000100 01000010	GB18030	腂

characters	encoding	bits
Føö	Windows Latin 1	01000110 11111000 11110110
Føö	Mac Roman	01000110 10111111 10011010
Føö	UTF-8	01000110 11000011 10111000 11000011 10110110

http://kunststube.net/encoding/

Encoding

- Coding schemes
- Character sets
  - ASCII
  - Unicode

Encoding

- Coding schemes
- Character sets
  - ASCII
  - Unicode

- Uses 7-bits
- 128 characters (‘code points’)
- 33 non-printing control characters
- 95 printable characters
  (incl. space and line break)

- Extended-ASCII uses 1-byte (8-bits) for 256 code points

ASCII TABLE

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]

Encoding

- Coding schemes
- Character sets
  - ASCII
  - Unicode

- Multi-byte character encodings (UTF-8, UTF-16, UTF-32)
- 1,114,112 ‘code points’
- 135 modern and historic scripts

  and symbols

  9,731 stands for 
- Uses Code charts to organise

Encoding – different glyphs

and symbols

9,731 stands for 

- Uses Code charts to organise

and symbols

9,731 stands for 

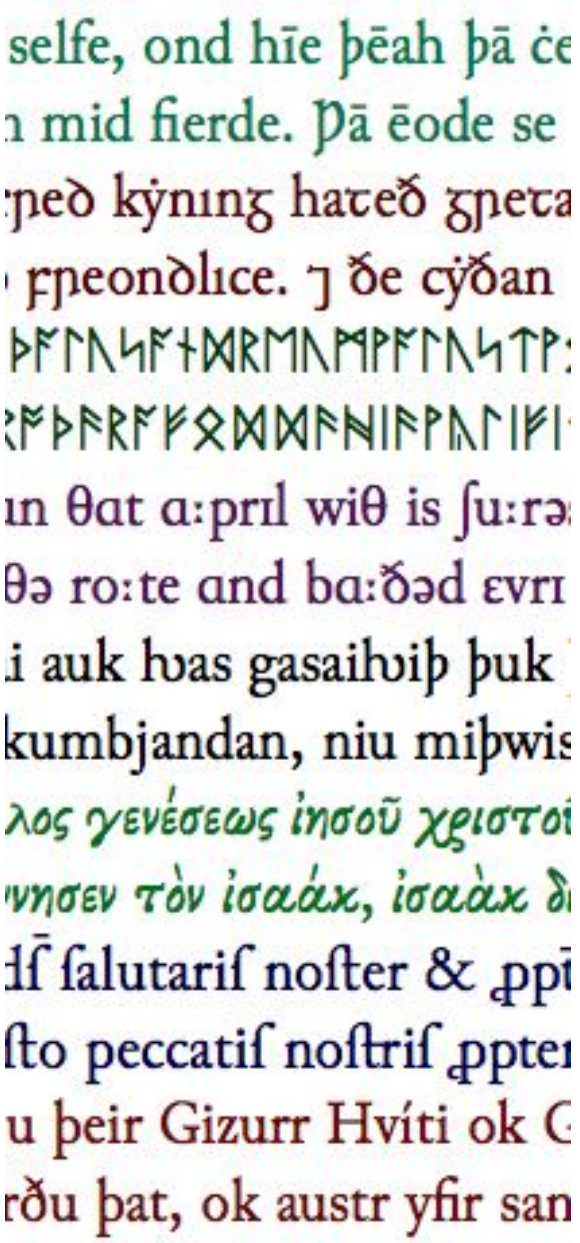
- Uses Code charts to organise

Encoding – different glyphs

Fonts!

Junicode (short for Junius-Unicode) is a TrueType/OpenType font for medievalists with extensive coverage of the Latin Unicode ranges, plus Runic and Gothic.

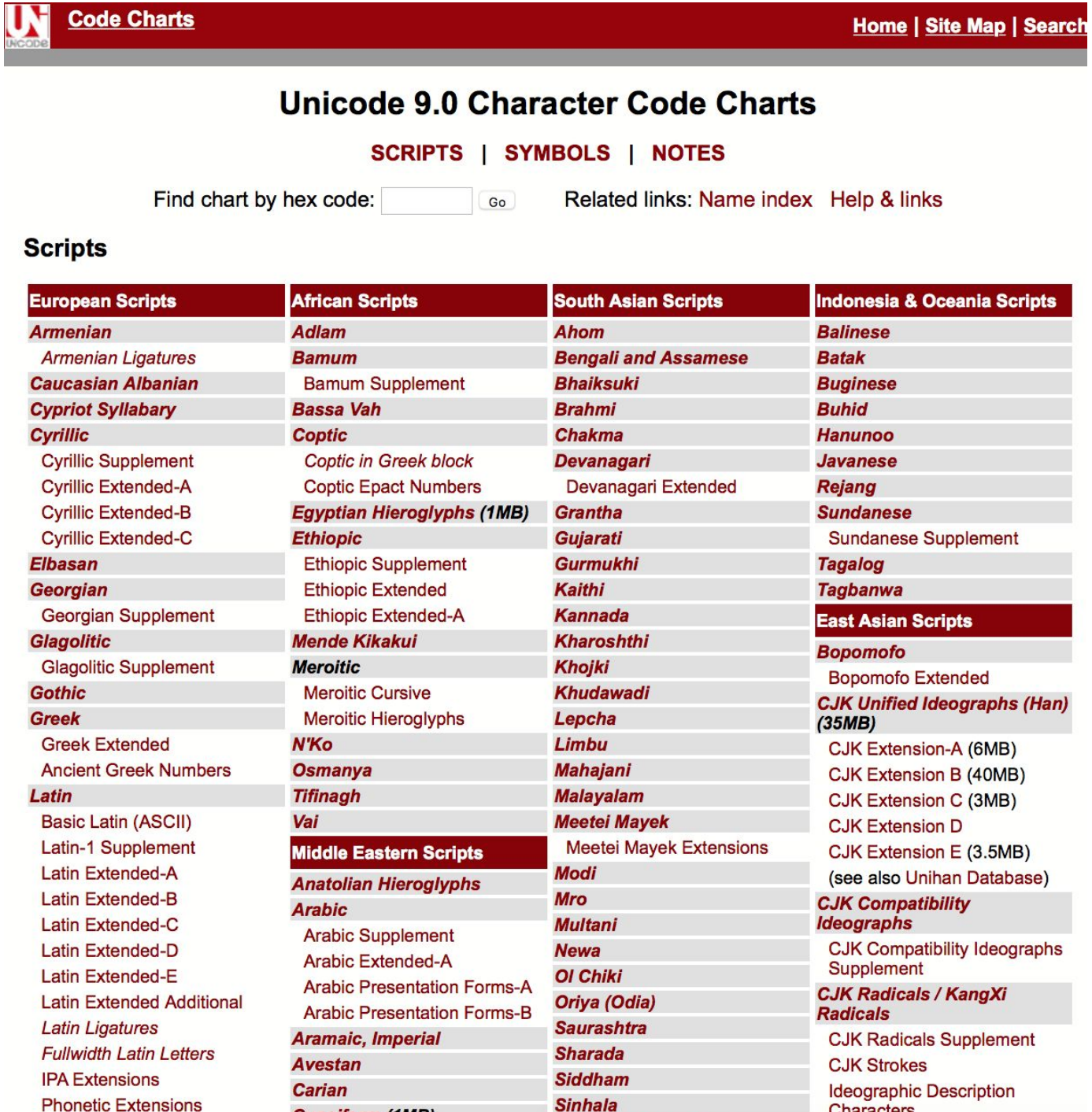
http://junicode.sourceforge.net



Encoding

- Coding schemes
- Character sets
  - ASCII
  - Unicode

http://unicode.org/charts/



Code Charts

Home | Site Map | Search

Unicode 9.0 Character Code Charts

SCRIPTS | SYMBOLS | NOTES

Find chart by hex code: Go Related links: Name index Help & links

Scripts

European Scripts

- Armenian
- Armenian Ligatures
- Caucasian Albanian
- Cypriot Syllabary
- Cyrillic
- Cyrillic Supplement
- Cyrillic Extended-A
- Cyrillic Extended-B
- Cyrillic Extended-C
- Elbasan
- Georgian
- Georgian Supplement
- Glagolitic
- Glagolitic Supplement
- Gothic
- Greek
- Greek Extended
- Ancient Greek Numbers
- Latin
- Basic Latin (ASCII)
- Latin-1 Supplement
- Latin Extended-A
- Latin Extended-B
- Latin Extended-C
- Latin Extended-D
- Latin Extended-E
- Latin Extended Additional
- Latin Ligatures
- Fullwidth Latin Letters
- IPA Extensions
- Phonetic Extensions

African Scripts

- Adlam
- Bamum
- Bamum Supplement
- Bassa Vah
- Coptic
- Coptic in Greek block
- Coptic Epact Numbers
- Egyptian Hieroglyphs (1MB)
- Ethiopic
- Ethiopic Supplement
- Ethiopic Extended
- Ethiopic Extended-A
- Mende Kikakui
- Meroitic
- Meroitic Cursive
- Meroitic Hieroglyphs
- N'Ko
- Osmanya
- Tifinagh
- Vai

Middle Eastern Scripts

- Anatolian Hieroglyphs
- Arabic
- Arabic Supplement
- Arabic Extended-A
- Arabic Presentation Forms-A
- Arabic Presentation Forms-B
- Aramaic, Imperial
- Avestan
- Carian

South Asian Scripts

- Ahom
- Bengali and Assamese
- Bhaiksuki
- Brahmi
- Chakma
- Devanagari
- Devanagari Extended
- Grantha
- Gujarati
- Gurmukhi
- Kaithi
- Kannada
- Kharoshthi
- Khojki
- Khudawadi
- Lepcha
- Limbu
- Mahajani
- Malayalam
- Meetei Mayek
- Meetei Mayek Extensions
- Modi
- Mro
- Multani
- Newa
- Ol Chiki
- Oriya (Odia)
- Saurashtra
- Sharada
- Siddham
- Sinhala

Indonesia & Oceania Scripts

- Balinese
- Batak
- Buginese
- Buhid
- Hanunoo
- Javanese
- Rejang
- Sundanese
- Sundanese Supplement
- Tagalog
- Tagbanwa

East Asian Scripts

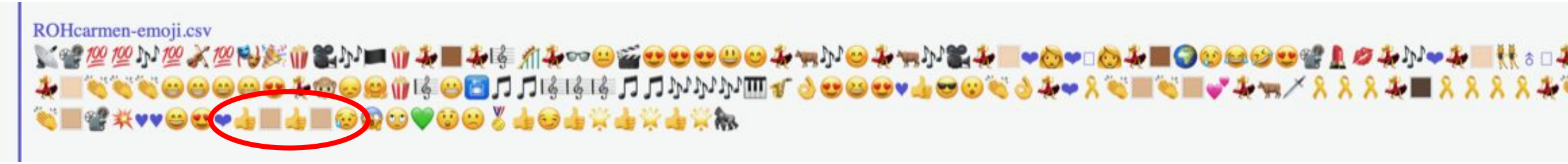
- Bopomofo
- Bopomofo Extended
- CJK Unified Ideographs (Han) (35MB)
- CJK Extension-A (6MB)
- CJK Extension B (40MB)
- CJK Extension C (3MB)
- CJK Extension D
- CJK Extension E (3.5MB)
- (see also Unihan Database)
- CJK Compatibility Ideographs
- CJK Compatibility Ideographs Supplement
- CJK Radicals / KangXi Radicals
- CJK Radicals Supplement
- CJK Strokes
- Ideographic Description Characters

What Every Programmer Absolutely, Positively Needs To Know About Encodings And Character Sets To Work With Text

If you are dealing with text in a computer, you need to know about encodings. Period. Yes, even if you are just sending emails. Even if you are just *receiving* emails. You don't need to understand every last detail, but you must at least know what this whole "encoding" thing is about. And the good news first: while the topic *can* get messy and confusing, the basic idea is really, really simple.

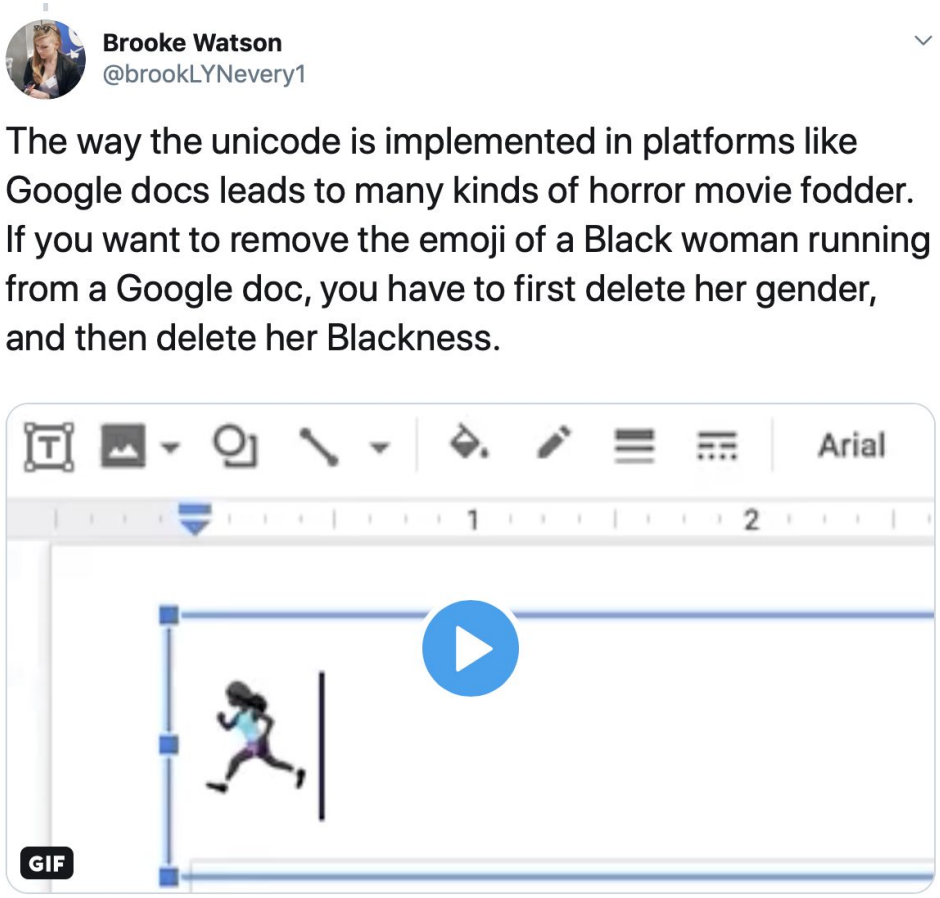
David C. Zentgraf, http://kunststube.net/encoding/

MS Outlook can corrupt multi-byte emoji



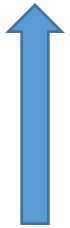
Could this encoding error be considered racist?

Multi-byte Unicode implementations



https://twitter.com/brooklynevery1/status/1167409916899934209 [Accessed 2019-09-17]

Formats



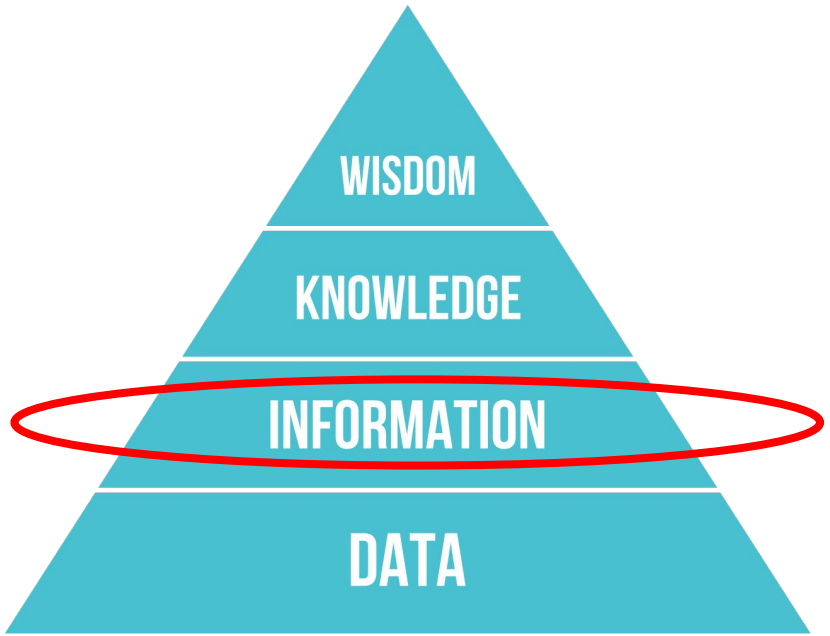
Encoding

Formats

- Information
- Informative

Formats

- Information
- Informative



Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text

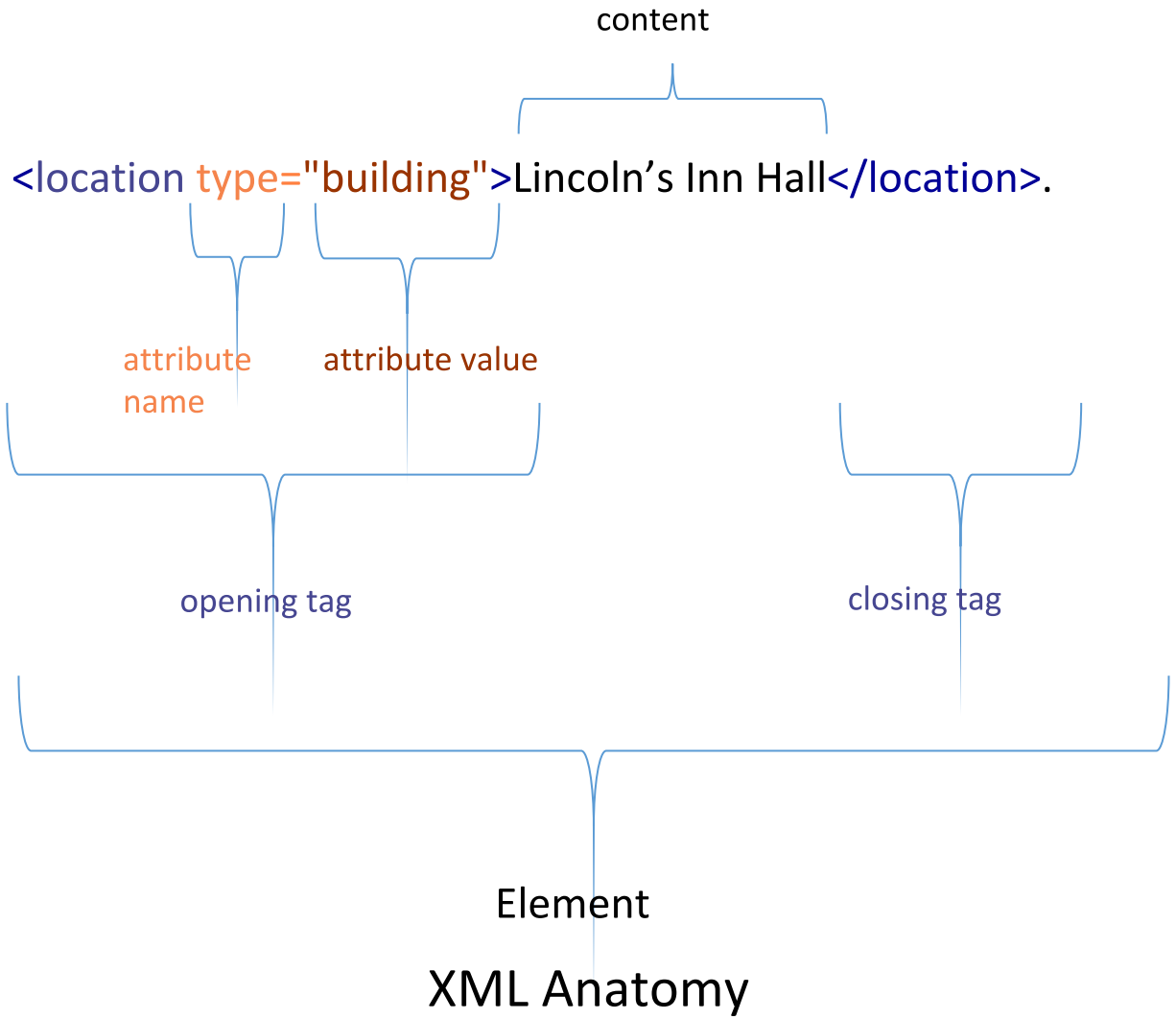
Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text

```
<publications>
 <publication>
  <pub_id>808</pub_id>
  <comp_id>70396</comp_id>
  <pagescans>
   <page order="1" number="vii" filename="scans/0808_0007_roman.png" title="Page vii"/>
   <page order="2" number="viii" filename="scans/0808_0008_roman.png" title="Page viii"/>
   <page order="3" number="ix" filename="scans/0808_0009_roman.png" title="Page ix"/>
   <page order="4" number="x" filename="scans/0808_0010_roman.png" title="Page x"/>
   <page order="5" number="xi" filename="scans/0808_0011_roman.png" title="Page xi"/>
   <page order="6" number="xii" filename="scans/0808_0012_roman.png" title="Page xii"/>
   <page order="7" number="xiii" filename="scans/0808_0013_roman.png" title="Page xiii"/>
   <page order="8" number="xiv" filename="scans/0808_0014_roman.png" title="Page xiv"/>
   <page order="9" number="xv" filename="scans/0808_0015_roman.png" title="Page xv"/>
   <page order="10" number="xvi" filename="scans/0808_0016_roman.png" title="Page xvi"/>
  </pagescans>
 </publication>
 <publication>
  <pub_id>808</pub_id>
  <comp_id>70397</comp_id>
  <pagescans>
   <page order="1" number="1" filename="scans/0808_0001_arabic.png" title="Page 1"/>
   <page order="2" number="2" filename="scans/0808_0002_arabic.png" title="Page 2"/>
  </pagescans>
 </publication>
```

Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text



XML Anatomy

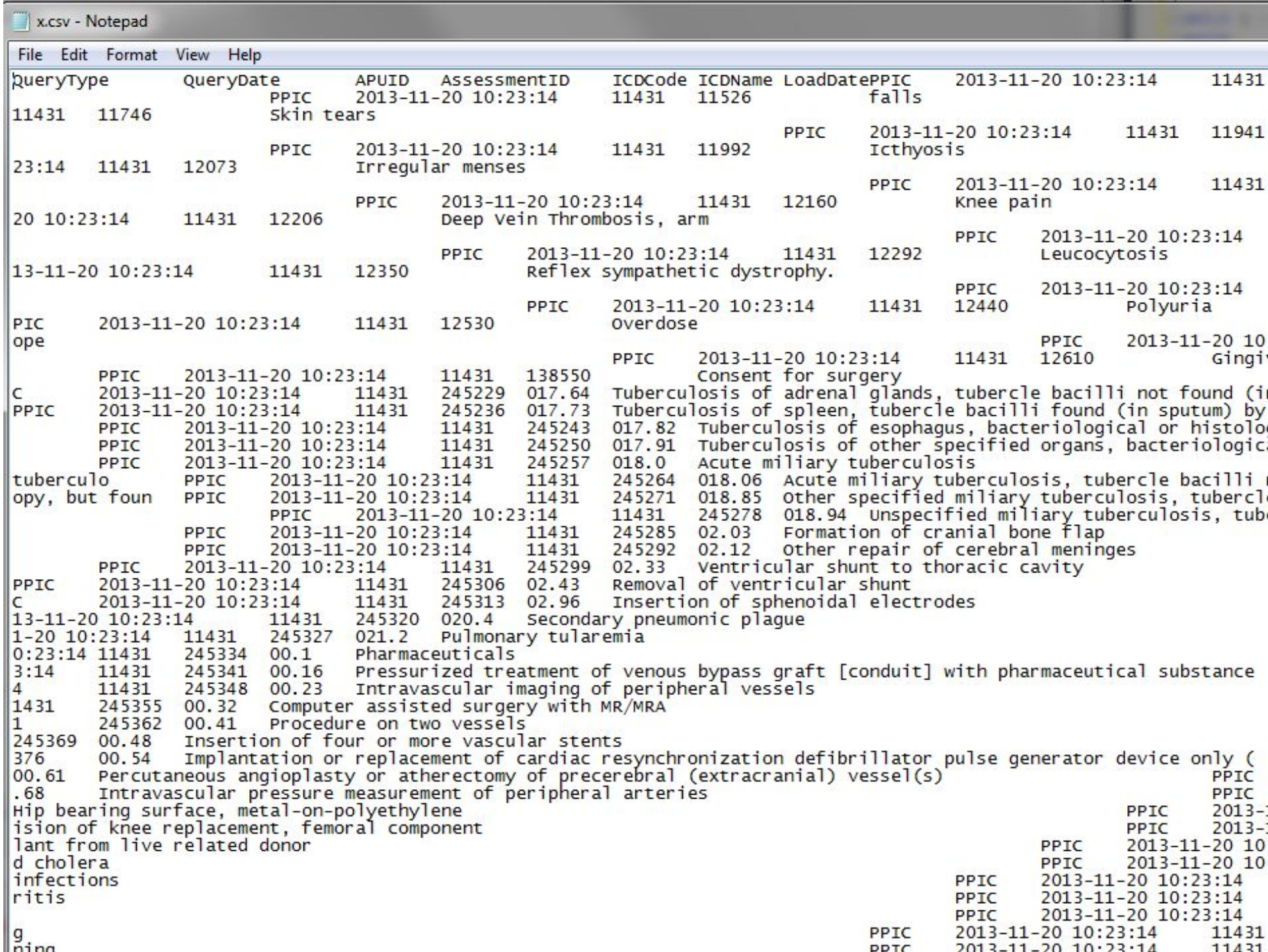
Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text

```json
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```

Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text

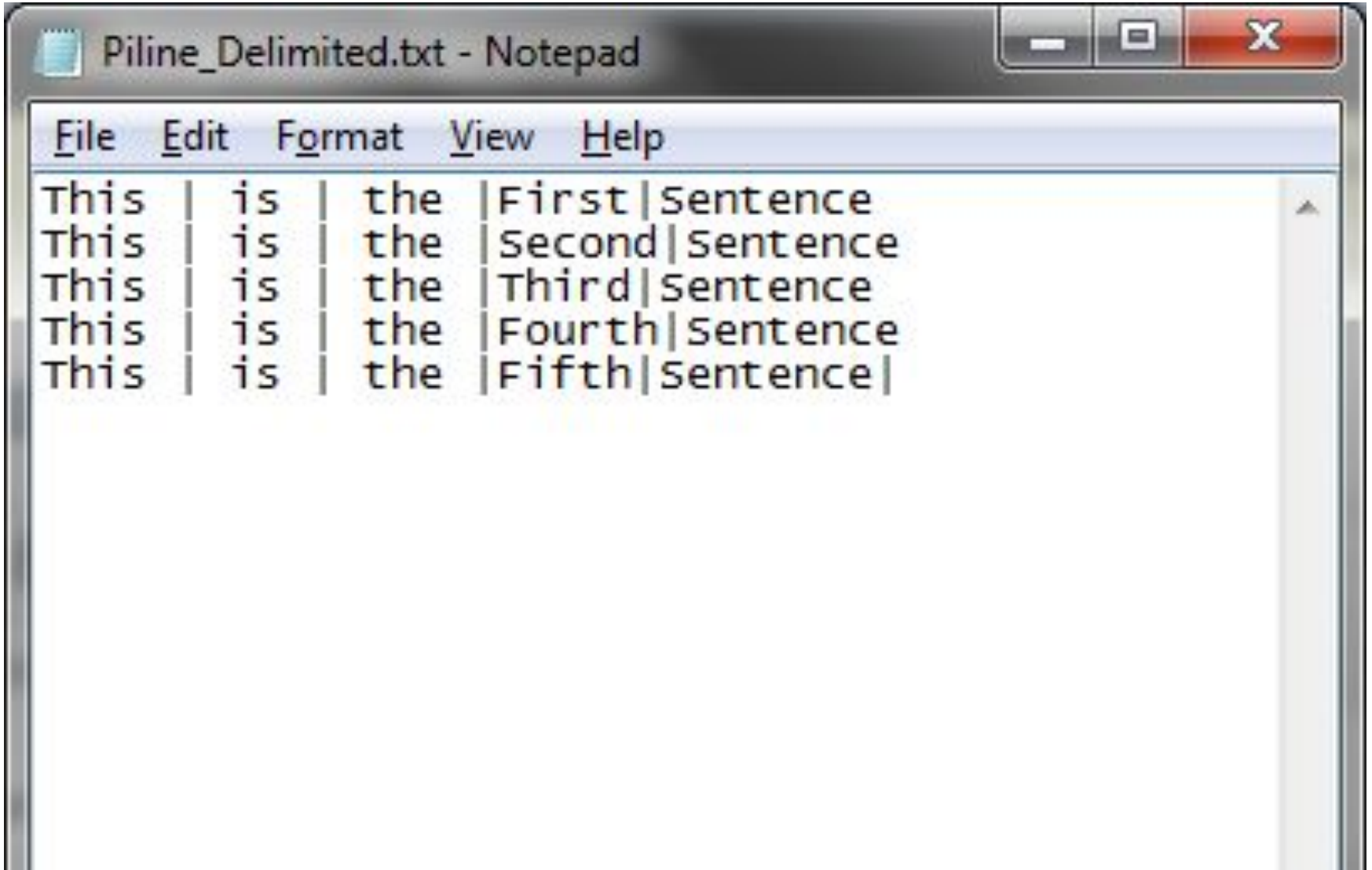

x.csv - Notepad

File Edit Format View Help

```
QueryType       QueryDate       APUID   AssessmentID    ICDCode ICDName LoadDatePPIC    2013-11-20 10:23:14     11431
                        PPIC    2013-11-20 10:23:14     11431   11526           falls
11431   11746           Skin tears
                                                                        PPIC    2013-11-20 10:23:14     11431   11941
                        PPIC    2013-11-20 10:23:14     11431   11992           Icthyosis
23:14   11431   12073           Irregular menses
                                                                        PPIC    2013-11-20 10:23:14     11431
                        PPIC    2013-11-20 10:23:14     11431   12160           Knee pain
20 10:23:14     11431   12206           Deep Vein Thrombosis, arm
                                                                        PPIC    2013-11-20 10:23:14
                                PPIC    2013-11-20 10:23:14     11431   12292           Leucocytosis
13-11-20 10:23:14       11431   12350           Reflex sympathetic dystrophy.
                                                                        PPIC    2013-11-20 10:23:14
                                        PPIC    2013-11-20 10:23:14     11431   12440           Polyuria
PIC     2013-11-20 10:23:14     11431   12530           Overdose
ope
                                                                                PPIC    2013-11-20 10
                                        PPIC    2013-11-20 10:23:14     11431   12610           Gingi
        PPIC    2013-11-20 10:23:14     11431   138550          Consent for surgery
C       2013-11-20 10:23:14     11431   245229  017.64  Tuberculosis of adrenal glands, tubercle bacilli not found (i
PPIC    2013-11-20 10:23:14     11431   245236  017.73  Tuberculosis of spleen, tubercle bacilli found (in sputum) by
        PPIC    2013-11-20 10:23:14     11431   245243  017.82  Tuberculosis of esophagus, bacteriological or histolo
        PPIC    2013-11-20 10:23:14     11431   245250  017.91  Tuberculosis of other specified organs, bacteriologic
        PPIC    2013-11-20 10:23:14     11431   245257  018.0   Acute miliary tuberculosis
tuberculo       PPIC    2013-11-20 10:23:14     11431   245264  018.06  Acute miliary tuberculosis, tubercle bacilli
opy, but foun   PPIC    2013-11-20 10:23:14     11431   245271  018.85  Other specified miliary tuberculosis, tubercl
                PPIC    2013-11-20 10:23:14     11431   245278  018.94  Unspecified miliary tuberculosis, tub
                PPIC    2013-11-20 10:23:14     11431   245285  02.03   Formation of cranial bone flap
                PPIC    2013-11-20 10:23:14     11431   245292  02.12   Other repair of cerebral meninges
        PPIC    2013-11-20 10:23:14     11431   245299  02.33   Ventricular shunt to thoracic cavity
PPIC    2013-11-20 10:23:14     11431   245306  02.43   Removal of ventricular shunt
C       2013-11-20 10:23:14     11431   245313  02.96   Insertion of sphenoidal electrodes
13-11-20 10:23:14       11431   245320  020.4   Secondary pneumonic plague
1-20 10:23:14   11431   245327  021.2   Pulmonary tularemia
0:23:14 11431   245334  00.1    Pharmaceuticals
3:14    11431   245341  00.16   Pressurized treatment of venous bypass graft [conduit] with pharmaceutical substance
4       11431   245348  00.23   Intravascular imaging of peripheral vessels
1431    245355  00.32   Computer assisted surgery with MR/MRA
1       245362  00.41   Procedure on two vessels
245369  00.48   Insertion of four or more vascular stents
376     00.54   Implantation or replacement of cardiac resynchronization defibrillator pulse generator device only (
00.61   Percutaneous angioplasty or atherectomy of precerebral (extracranial) vessel(s)                        PPIC
.68     Intravascular pressure measurement of peripheral arteries                                               PPIC
Hip bearing surface, metal-on-polyethylene                                                              PPIC    2013-
ision of knee replacement, femoral component                                                            PPIC    2013-
lant from live related donor                                                                    PPIC    2013-11-20 10
d cholera                                                                                       PPIC    2013-11-20 10
infections                                                                              PPIC    2013-11-20 10:23:14
ritis                                                                                   PPIC    2013-11-20 10:23:14
                                                                                        PPIC    2013-11-20 10:23:14
g                                                                               PPIC    2013-11-20 10:23:14     11431
ning                                                                            PPIC    2013-11-20 10:23:14     11431
```

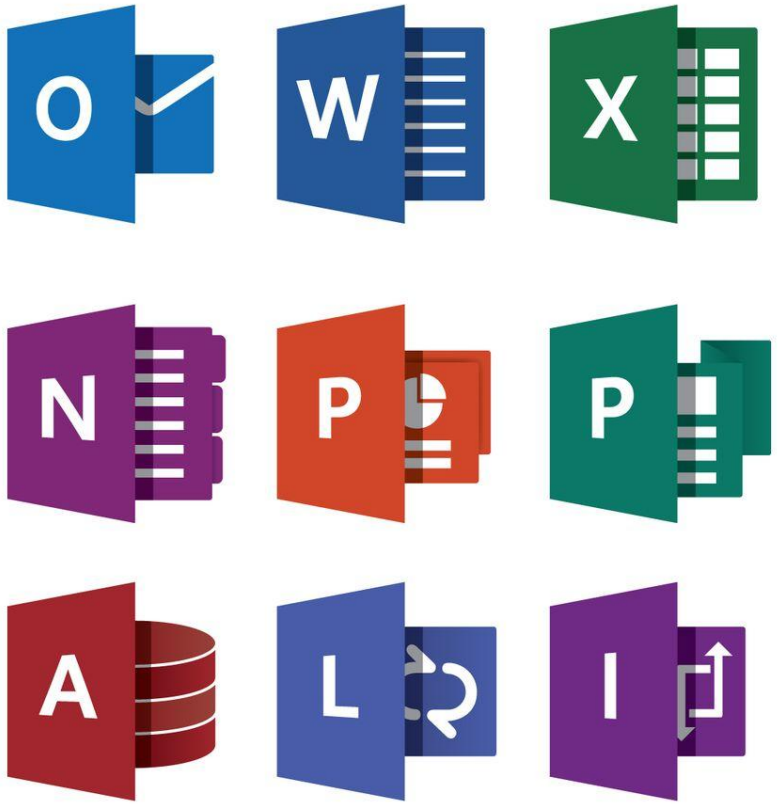
Formats

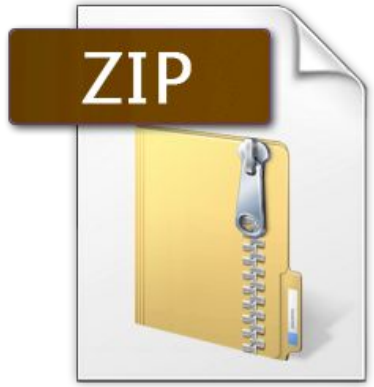
- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text



Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text





Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text

Proprietary formats
Binary encoding

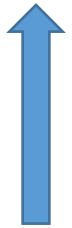
Formats

- XML
- JSON
- TSV
- Word
- Excel
- PDF
- Zip
- Text

Open formats
Text encoding

demo (head and tail)

Schema



Formats

Schema

- Data

schema

/ˈskiːmə/

noun
noun: **schema**; plural noun: **schemata**; plural noun: **schemas**

1. *technical*
   a representation of a plan or theory in the form of an outline or model.
   "a schema of scientific reasoning"
2. LOGIC
   a syllogistic figure.
3. (in Kantian philosophy) a conception of what is common to all members of a class; a general or essential type or form.

Origin

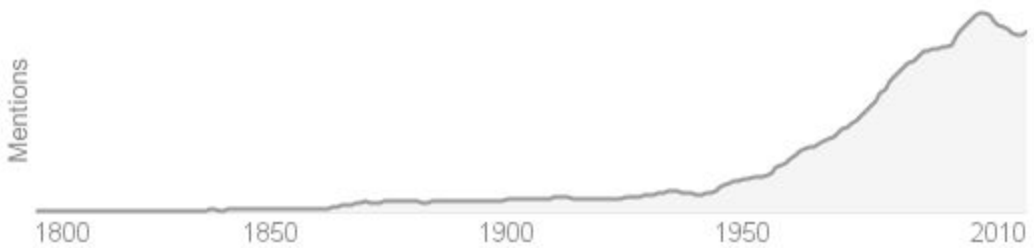
GREEK

skhēma → schema
form, figure — *late 18th century*

late 18th century (as a term in philosophy): from Greek *skhēma* 'form, figure'.

Translate schema to Choose language

Use over time for: schema

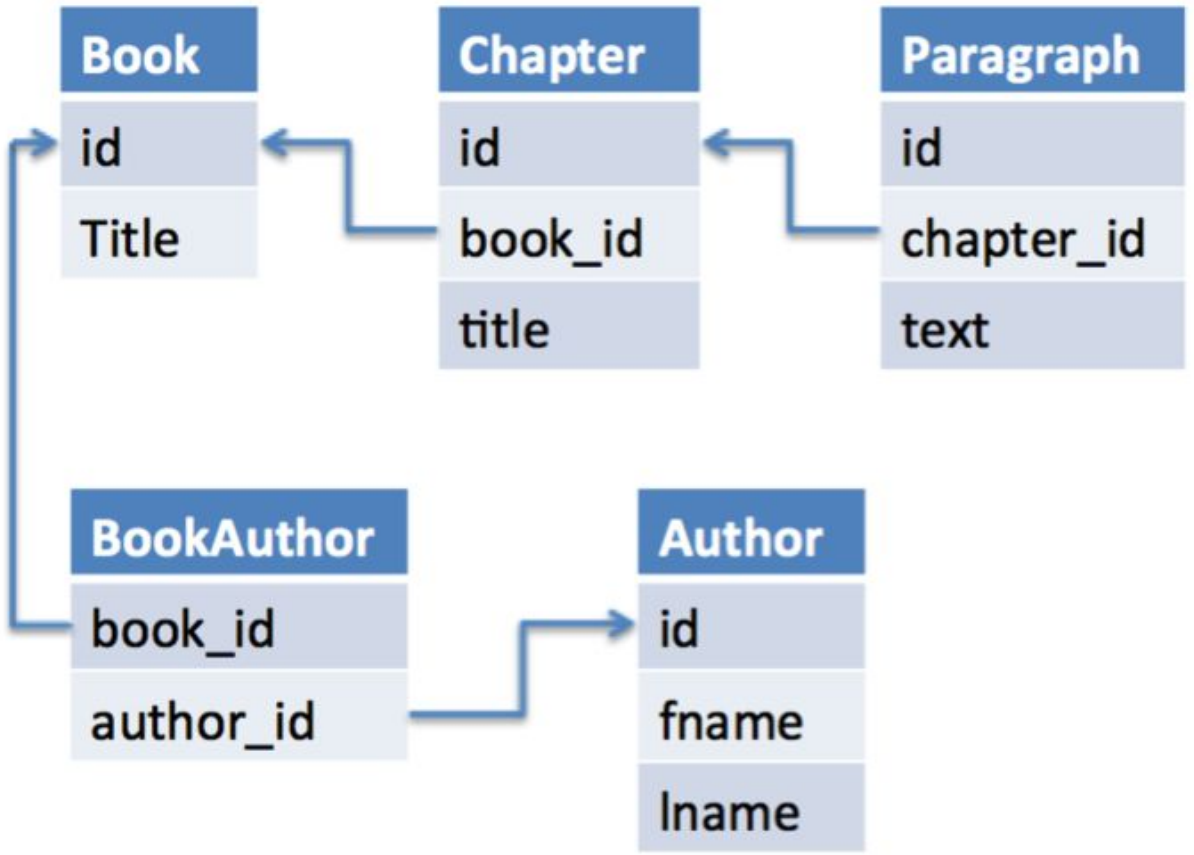


Schema

- Data
- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

Schema

- Data
- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values



Schema (text)

- HTML
- RDF
- TEI

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

What is HTML?

HTML is the Web's core language for creating content for everyone to use anywhere.

```
<!DOCTYPE html>
<html>
<title>Story</title>
<h1>My Story</h1>
<p>One upon a time,
  …</p>
</html>
```

Fig 1. HTML source code

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

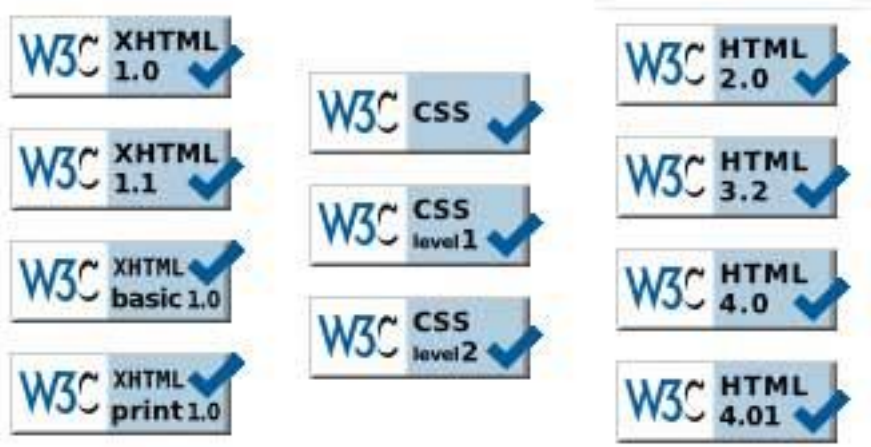
```
<!DOCTYPE html>
<html>
<title>Story</title>
<h1>My Story</h1>
<p>One upon a time,
  …</p>
</html>
```

Fig 1. HTML source code

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

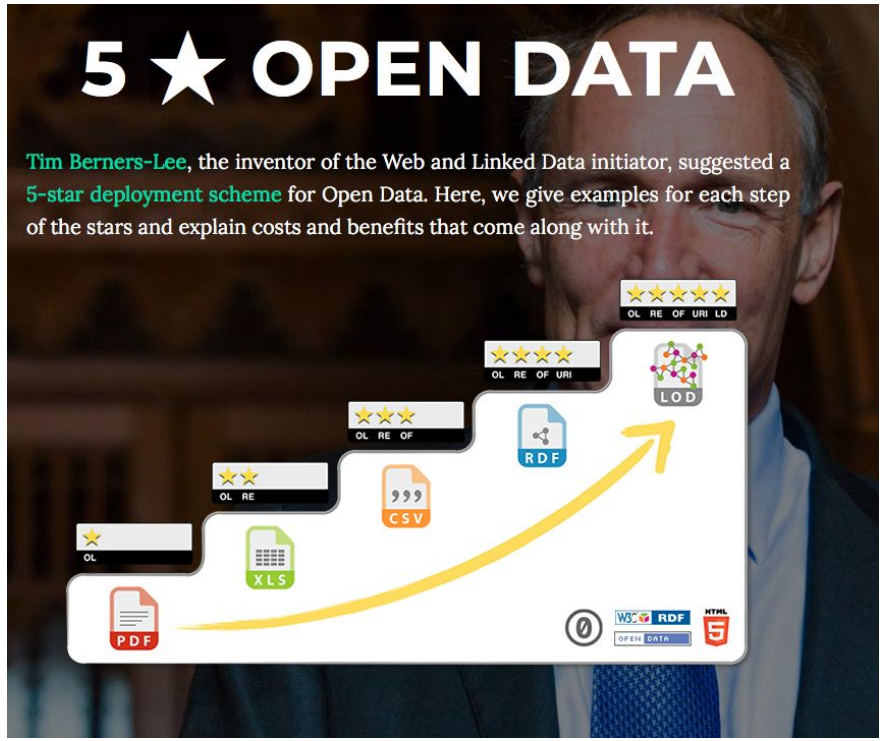


https://www.w3.org

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values



In standard N-Triples format, this RDF can be written as:

```
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#fullName> "Eric Miller" .
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#mailbox>
<mailto:e.miller123(at)example> .
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#personalTitle> "Dr." .
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2000/10/swap/pim/contact#Person> .
```

Equivalently, it can be written in standard Turtle (syntax) format as:

```
@prefix eric:    <http://www.w3.org/People/EM/contact#> .
@prefix contact: <http://www.w3.org/2000/10/swap/pim/contact#> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

eric:me contact:fullName "Eric Miller" .
eric:me contact:mailbox <mailto:e.miller123(at)example> .
eric:me contact:personalTitle "Dr." .
eric:me rdf:type contact:Person .
```

Or, it can be written in RDF/XML format as:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#" xmlns:eric="http://www.w3.org/People/EM/contact#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:mailbox rdf:resource="mailto:e.miller123(at)example"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:personalTitle>Dr.</contact:personalTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <rdf:type rdf:resource="http://www.w3.org/2000/10/swap/pim/contact#Person"/>
  </rdf:Description>
</rdf:RDF>
```

Schema

- HTML
- RDF
- TEI https://en.wikipedia.org/wiki/Text_Encoding_Initiative

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

Prose tags [edit]

TEI allows texts to be marked up syntactically at any level of granularity, or mixtur
sentences (s) and clauses (cl).[8]

```
<s>
 <cl>It was about the beginning of September, 1664,
 <cl>that I, among the rest of my neighbours,
      heard in ordinary discourse
  <cl>that the plague was returned again to Holland; </cl
  </cl>
 </cl>
 <cl>for it had been very violent there, and particularly
    Amsterdam and Rotterdam, in the year 1663, </cl>
 <cl>whither, <cl>they say,</cl> it was brought,
 <cl>some said</cl> from Italy, others from the Levant, a
 <cl>which were brought home by their Turkey fleet;</cl>
 </cl>
 <cl>others said it was brought from Candia;
    others from Cyprus. </cl>
</s>
<s>
 <cl>It mattered not <cl>from whence it came;</cl>
 </cl>
 <cl>but all agreed <cl>it was come into Holland again.</
 </cl>
</s>
```

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values

Verse [edit]

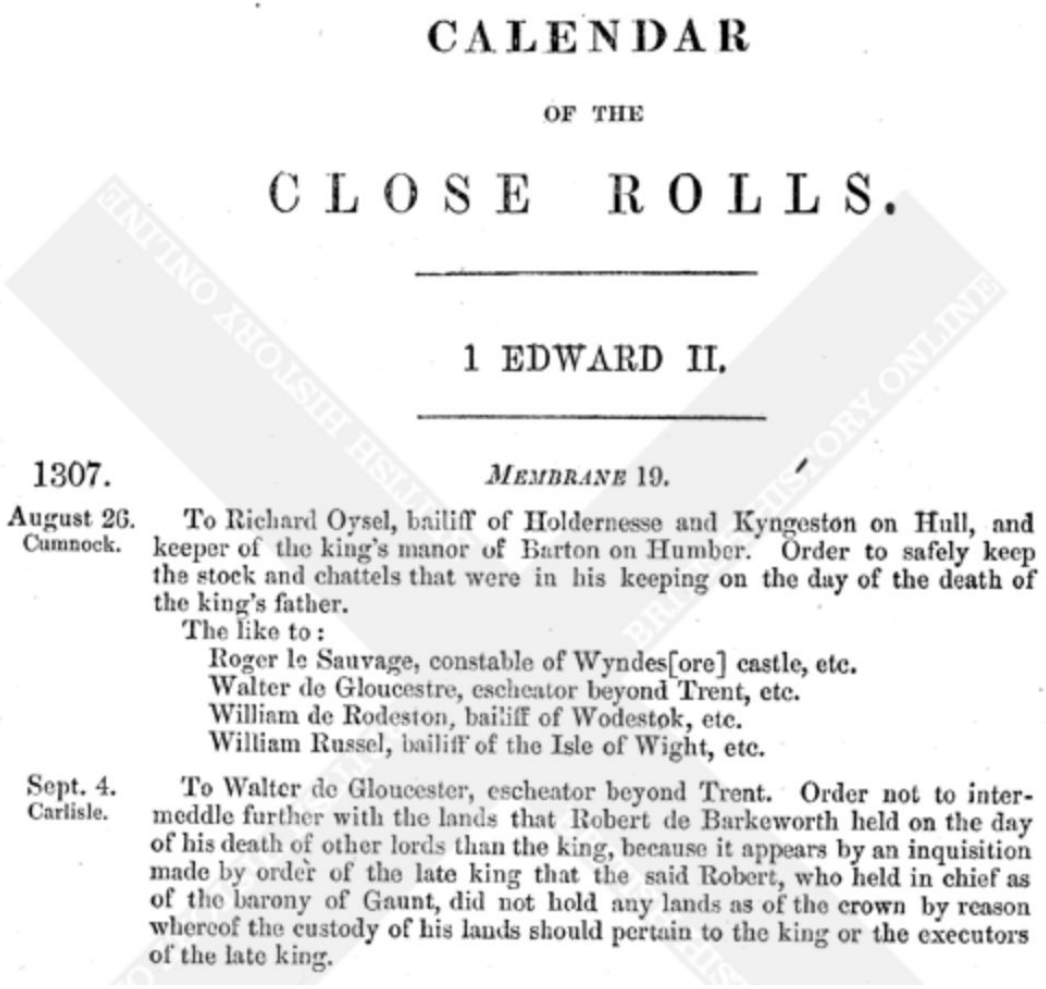
TEI has tags for marking up verse. This example (taken from the French translation

```
<div type="sonnet">
 <lg type="quatrain">
  <l>Les amoureux fervents et les savants austères</l>
  <l> Aiment également, dans leur mûre saison,</l>
  <l> Les chats puissants et doux, orgueil de la maison,</l
  <l> Qui comme eux sont frileux et comme eux sédentaires.<
 </lg>
 <lg type="quatrain">
  <l>Amis de la science et de la volupté</l>
  <l> Ils cherchent le silence et l'horreur des ténèbres ;<
  <l> L'Érèbe les eût pris pour ses coursiers funèbres,</l>
  <l> S'ils pouvaient au servage incliner leur fierté.</l>
 </lg>
 <lg type="tercet">
  <l>Ils prennent en songeant les nobles attitudes</l>
  <l>Des grands sphinx allongés au fond des solitudes,</l>
  <l>Qui semblent s'endormir dans un rêve sans fin ;</l>
 </lg>
 <lg type="tercet">
  <l>Leurs reins féconds sont pleins d'étincelles magiques,
  <l> Et des parcelles d'or, ainsi qu'un sable fin,</l>
  <l>Étoilent vaguement leurs prunelles mystiques.</l>
 </lg>
</div>
```

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values
  - Syntax, Marks, Typographical structures



CALENDAR

OF THE

CLOSE ROLLS.

1 EDWARD II.

1307. *Membrane* 19.

August 26. Cumnock. To Richard Oysel, bailiff of Holdernesse and Kyngeston on Hull, and keeper of the king's manor of Barton on Humber. Order to safely keep the stock and chattels that were in his keeping on the day of the death of the king's father.

The like to:

Roger le Sauvage, constable of Wyndes[ore] castle, etc.
Walter de Gloucestre, escheator beyond Trent, etc.
William de Rodeston, bailiff of Wodestok, etc.
William Russel, bailiff of the Isle of Wight, etc.

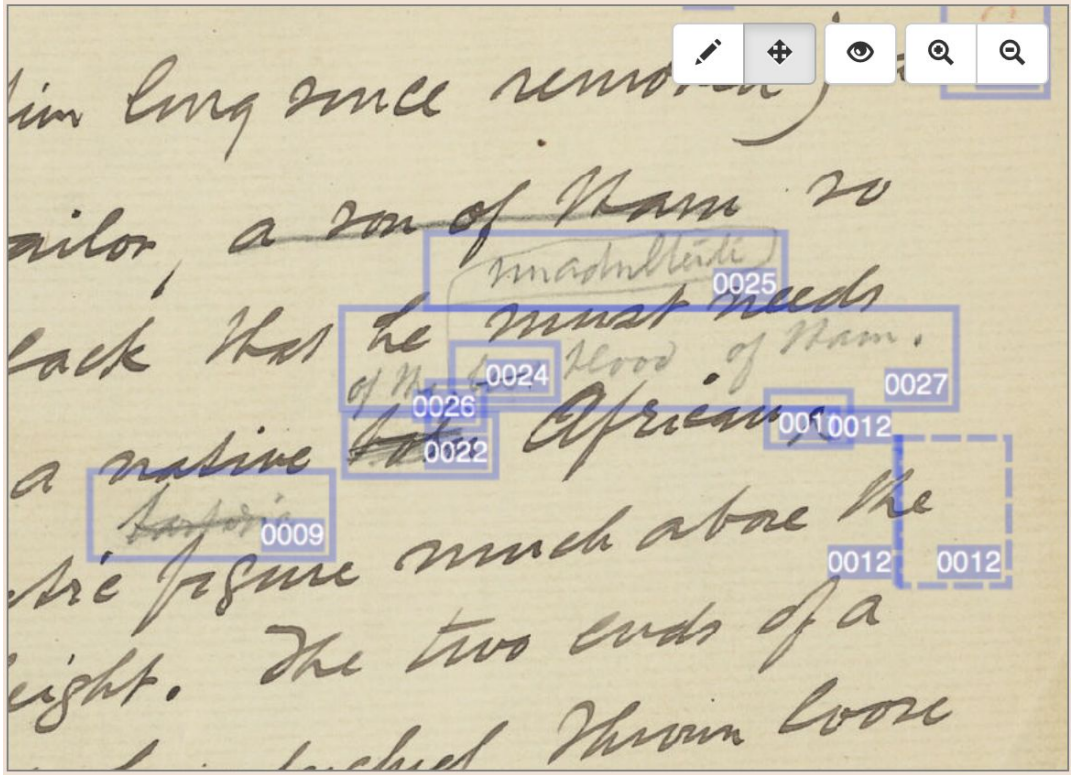
Sept. 4. Carlisle. To Walter de Gloucester, escheator beyond Trent. Order not to intermeddle further with the lands that Robert de Barkeworth held on the day of his death of other lords than the king, because it appears by an inquisition made by order of the late king that the said Robert, who held in chief as of the barony of Gaunt, did not hold any lands as of the crown by reason whereof the custody of his lands should pertain to the king or the executors of the late king.

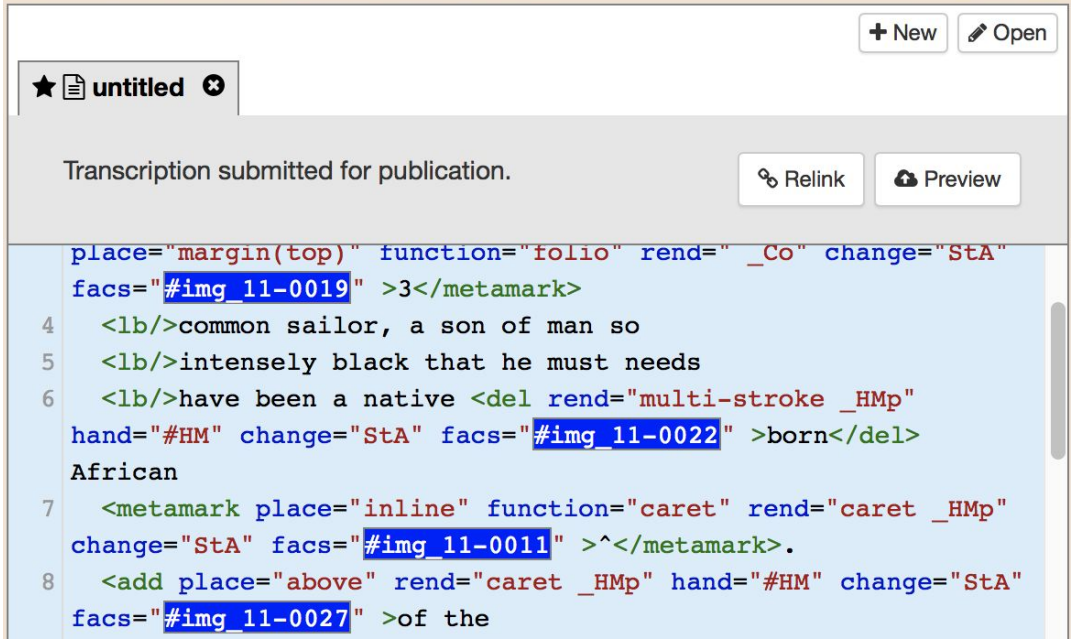
http://www.british-history.ac.uk/cal-close-rolls/edw2/vol1/p1

Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values
  - Syntax, Content

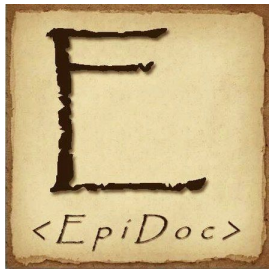


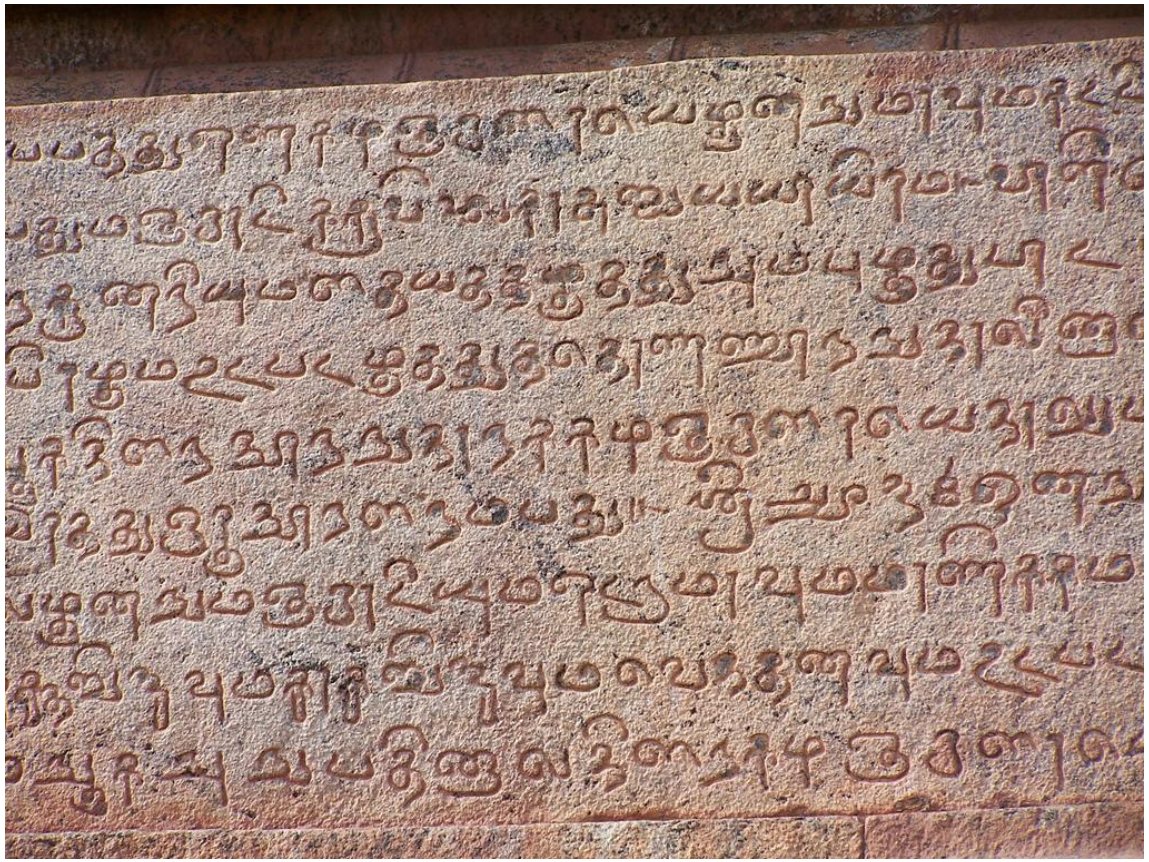


Schema

- HTML
- RDF
- TEI

- Standards for representing
  - Structures, Relationships
  - Entities, Attributes, Values
  - Syntax, Content
  - + encoding





https://en.wikipedia.org/wiki/Linguistics#/media/File:Ancient_Tamil_Script.jpg

Code Charts

Home | Site Map | Search

Unicode 9.0 Character Code Charts

http://www.unicode.org/charts/PDF/U0B80.pdf

SCRIPTS | SYMBOLS | NOTES

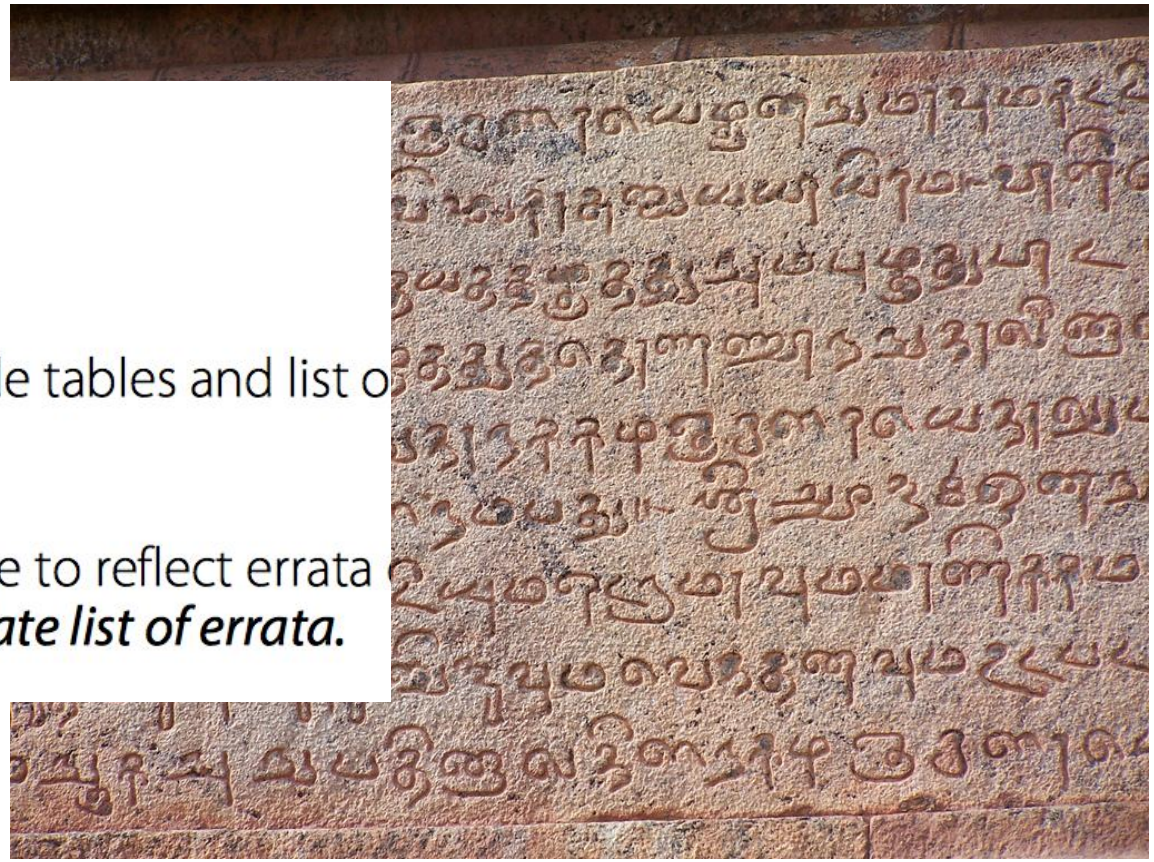
Tamil

Range: 0B80–0BFF

This file contains an excerpt from the character code tables and list o
The Unicode Standard, Version 9.0

This file may be changed at any time without notice to reflect errata
See http://www.unicode.org/errata/ for an up-to-date list of errata.

- Syntax, Content
- + encoding



https://en.wikipedia.org/wiki/Linguistics#/media/File:Ancient_Tamil_Script.jpg

Code Charts

Home | Site Map | Search

Unicode 9.0 Character Code Charts

http://www.unicode.org/charts/PDF/U0B80.pdf

SCRIPTS | SYMBOLS | NOTES

Tamil

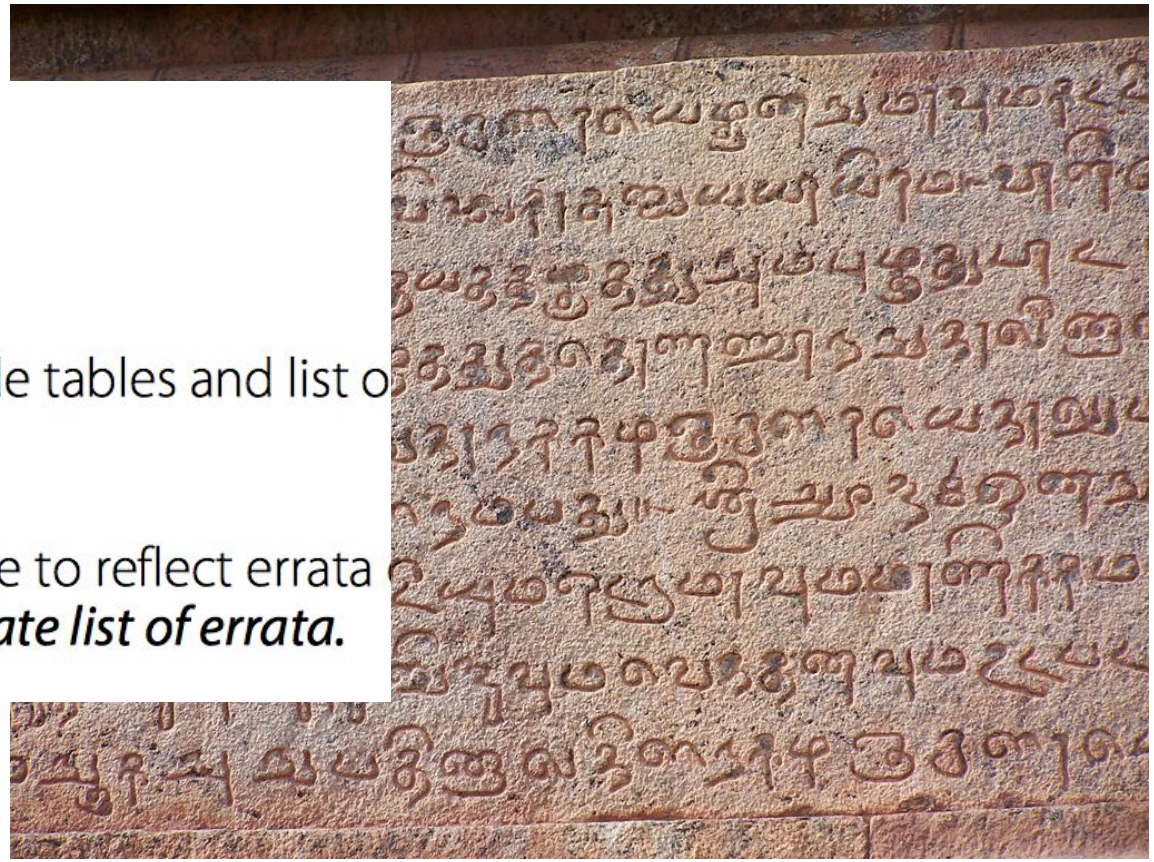
not ancient

Range: 0B80–0BFF

This file contains an excerpt from the character code tables and list o
The Unicode Standard, Version 9.0

This file may be changed at any time without notice to reflect errata
See http://www.unicode.org/errata/ for an up-to-date list of errata.

- Syntax, Content
- + encoding



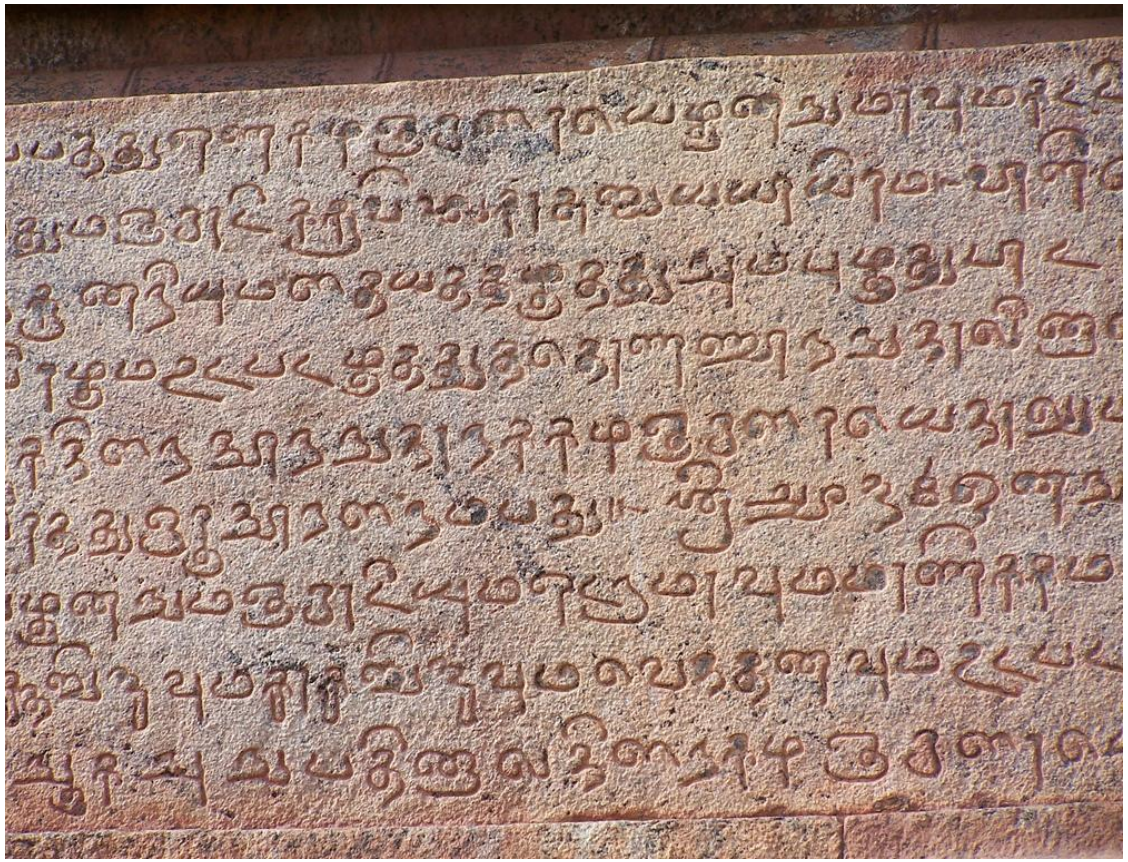
https://en.wikipedia.org/wiki/Linguistics#/media/File:Ancient_Tamil_Script.jpg

Code Charts

Home | Site Map | Search

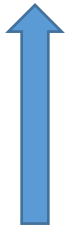
	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ 0B90		ர 0BB0	◌ 0BC0	ஓ 0BD0		ய 0BF0
1				ற 0BB1	◌ி 0BC1			ர 0BF1
2	◌ 0B82	ஒ 0B92		ல 0BB2	◌ீ 0BC2			ற 0BF2
3	◌◌ 0B83	ஓ 0B93	ண 0BA3	ள 0BB3				ல 0BF3
4		ஔ 0B94	த 0BA4	ழ 0BB4				ள 0BF4
5	அ 0B85	க 0B95		வ 0BB5				வு 0BF5

http://www.unicode.org/charts/PDF/U0B80.pdf



//en.wikipedia.org/wiki/Linguistics#/media/File:Ancient_Tamil_Script.jpg

Applications



Schema
Formats
Encodings
Storage

Applications

- Word file => MS Word
- Excel file => MS Excel
- Zip file => winzip, gzip
- Database => Oracle MySQL

- Text file => ?

Applications

- Word file => MS Word
- Excel file => MS Excel
- Zip file => winzip, gzip
- Database => Oracle MySQL

- Text file => 86 Free Software
  45 Proprietary

  Any format
  Any schema

Applications

- Word file
- Excel file
- Zip file
- MySQL Database

- Text file

“Obsolete power corrupts obsoletely.” - Ted Nelson

The technology associated with interpreting the representation at each of the layers can change or become less available

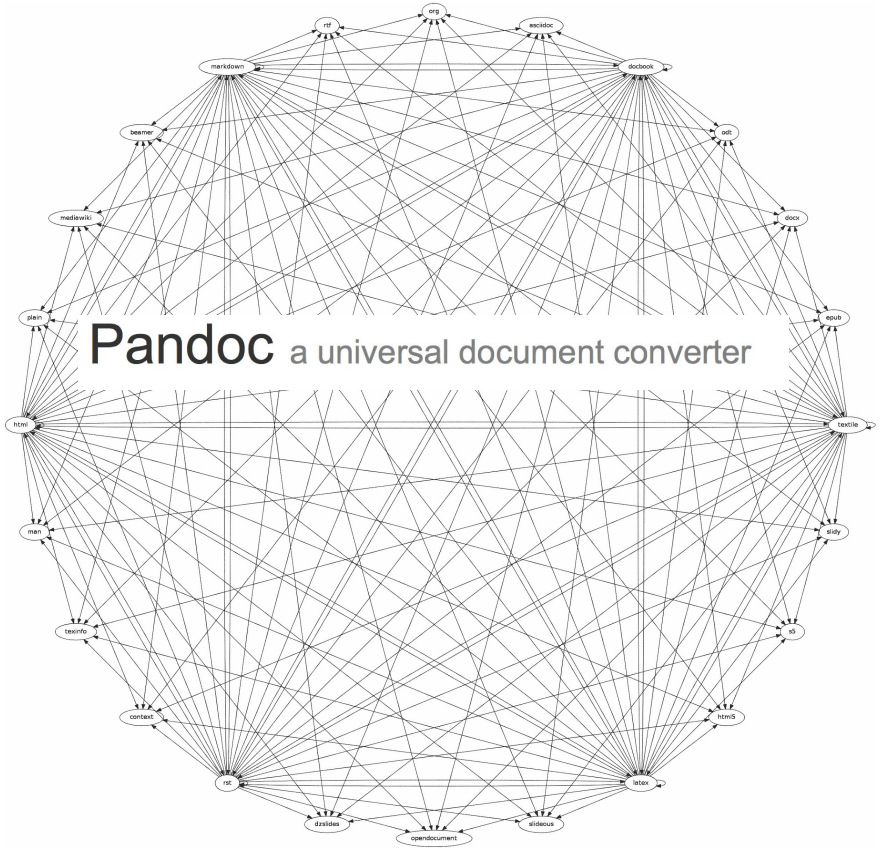
https://en.wikipedia.org/wiki/Ted_Nelson

Applications + Transformations

- Word file => MS Word
- Excel file => MS Excel
- Zip file => winzip, gzip
- Database => Oracle MySQL

- Text file => 86 Free Software
  45 Proprietary

  Any format
  Any schema



Applications + Standards

- British Museum's
  History of the world in 100 objects
  podcast
- The S'hia 'Alam'
  - A lavishly gilded ceremonial sword which represented the standards which were carried during conflict
  - Rules of engagement.

https://www.britishmuseum.org/explore/a_history_of_the_world.aspx
https://sites.google.com/site/100objectsbritishmuseum/home/shi-a-religious-parade-standard---steel-alam-from-iran

Applications + Standards

- British Museum Room 33
  - China and South Asia
    The Sir Joseph Hotung Gallery

- Museum label:
  - “This alam is covered in Islamic religious inscriptions, giving it talismanic power” (AD 680)

Applications + Standards + Transformations

Realtime machine translation
of English -> Chinese

Text Analysis:

A Survey of Principles, Tools, and New Ways of Reading

What is text analysis?

Using computational tools and/or programs to analyse text data of varying sizes by yielding quantitative results and making arguments about historical trends or form, style, content, and context.

This can be broken down further:

Statistical analysis: counting words, calculating word pairings (collocates) and ngrams (groups of two or three words or more co-occurring words), average word and sentence length, mean word usage; organising words that occur too frequently to be studied one by one.

Corpus analysis: searching and pattern recognition across multiple texts.

Linguistic analysis: parts-of-speech tagging, lexical variety and uniqueness, topic modeling, sentiment analysis, stylometry.

Network analysis: mapping connections between metadata and textual data.

Various other visualisations (graphs, maps and trees) for explanatory force.

A variety of calculations (simple character counts, mean word usage, hapax percentages) can help to identify repetition, brevity, and unique word clusters.

Anthony Kenny discusses the “mysterious veneration” that some literary scholars have for single and rare word occurrences, when “the rate of occurrence of a dull common word in a text may be a much more significant feature” (*Computation of Style* [1982], 67–68).

Similarly, **John F. Burrows**, in *Computation into Criticism*, bases his analysis on the 30 most common words in Jane Austen’s novels, with less attention to unique words.

Distant reading and interpretation

Franco Moretti: “Quantitative research provides a type which is ideally independent of interpretations … it provides *data*, not interpretation” (*Graphs, Maps, Tress* [Verso, 2007]).

History: shifting the gaze from extraordinary people and events to everyday facts. What literature can be found in large mass of facts?

“Abstraction is not an end in itself, but a way to widen the domain of the literary historian, and enrich its internal problematic” (Moretti, 2).

Distance is a new kind of knowledge—a model

Burrows (2004) argues that the styles of authors come from common words (i. e., articles and prepositions).

“[T]he real value of studying the common words rests on the fact that they constitute the underlying fabric of a text, a barely visible web that gives shape to whatever is being said … The principal point of interest is neither a single stitch, a single thread, nor even a single color but the overall effect. Such effects are best seen, moreover, when different pieces are put side by side.”

The value of experimentation, and the possibilities for new modes of reading. Burrows:“computer-assisted textual analysis can be of value in many different sorts of literary inquiry, helping to resolve some questions, to carry others forward, and to open entirely new ones.”

David Hoover (2013): “the computer’s greatest strengths are in storing, counting, comparing, sorting, and performing statistical analysis. This makes computer-assisted textual analysis especially appropriate and effective for investigating textual differences and similarities” (see “Textual Analysis” <https://dlsanthology.mla.hcommons.org/textual-analysis/>).

Possible directions of travel:

- Testing a hunch, hypothesis, or thesis about an author, text, passage, genre, or period
- Testing the claims of a critical work
- Investigating how and the extent to which authors differentiate the voices of characters or narrators in a work
- Investigating shifts in style and how they change over time
- Investigating the history of an important word, concept, or group of words or concepts over a long time span
- Studying the effects of genre conventions
- Investigating claim of authorship

Overconfidence and the problem of validation

Adam Hammond suggests that one of the failures of distant reading comes from its lack of discoveries and tendency to over-validate its tools. When unique results are generated, they often cannot be verified, in his estimation.
("The double bind of validation: distant reading and the digital humanities' 'trough of disillusionment.'" *Literature Compass* 14.8 (August 2017): e12402.)

Nan Z. Da: distant reading encourages reductive tautological thinking.
Lacks an intermediary scale of meaningful relations between the macro and the micro; ("The Computational Case against Computational Literary Studies", *Critical Inquiry* 45.3 (Spring 2019): 601-639).

Some principles

1. **Relative frequency**: this comes from John Stuart Mill's **principle of concomitant variation**, which states that if an antecedent circumstance is observed to change proportionally with the occurrence of a phenomenon, it is probably the cause of that phenomenon.

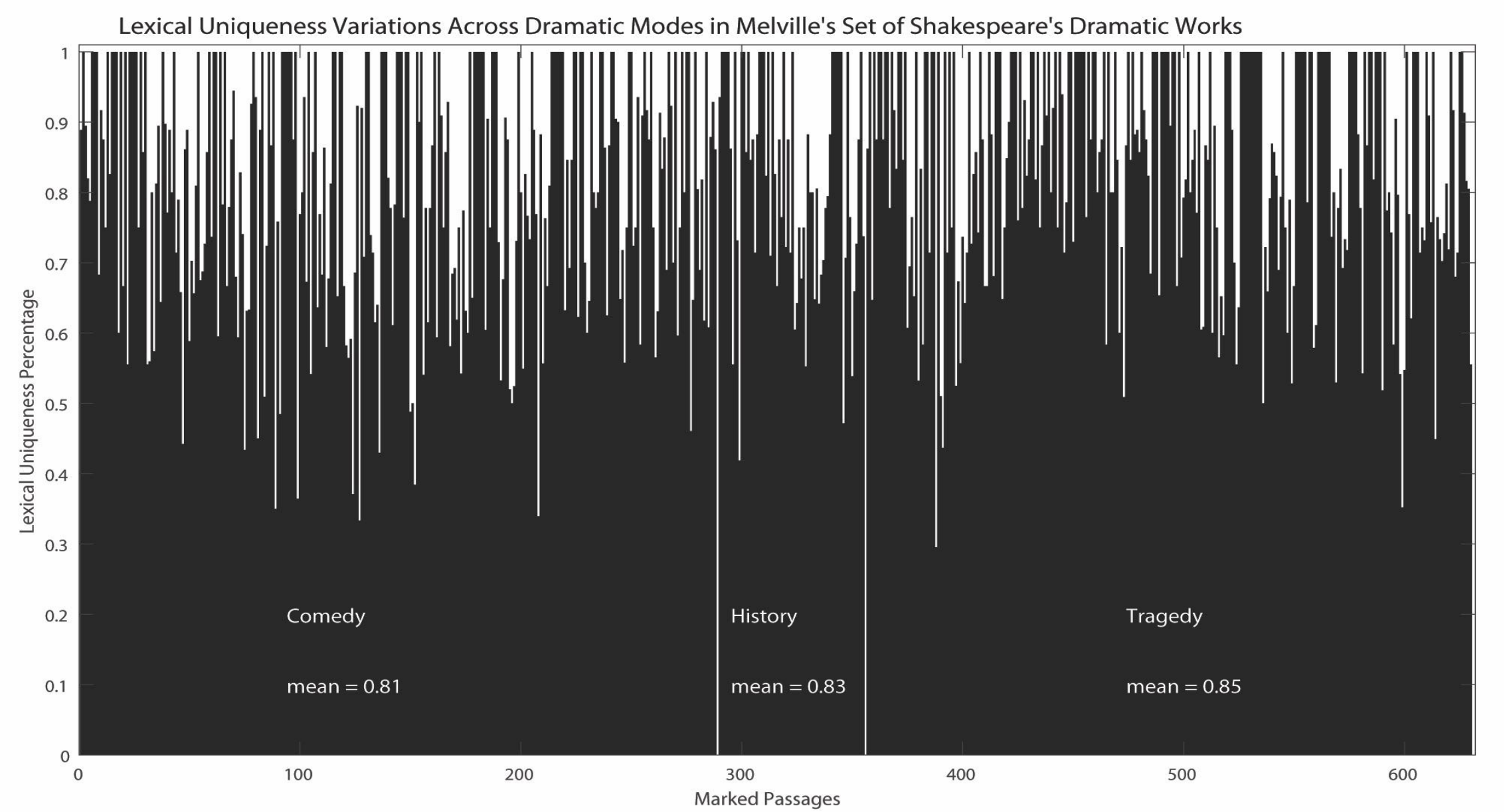
Effects are typically proportional to their causes. In the case of machine learning and word frequencies, this means that texts with similar variables tend to be similar and oftentimes causally connected.

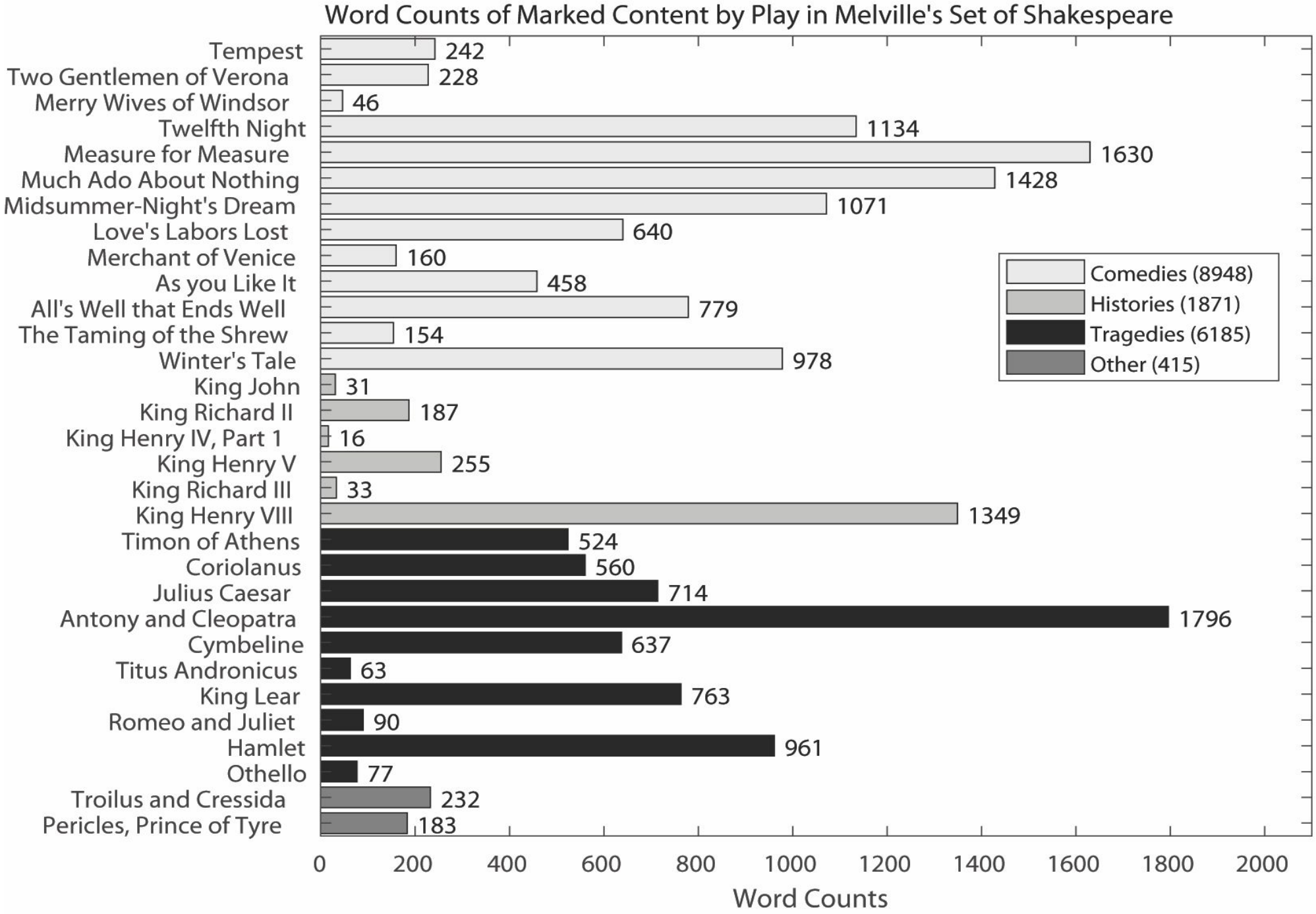
2. **Inductive inference and probability**: creating general hypotheses from specific instances; *yet*, understanding that the results are only probabilistic (i.e., only as good as the quality and comprehensiveness of the data). Justified true belief rather than mere fact.

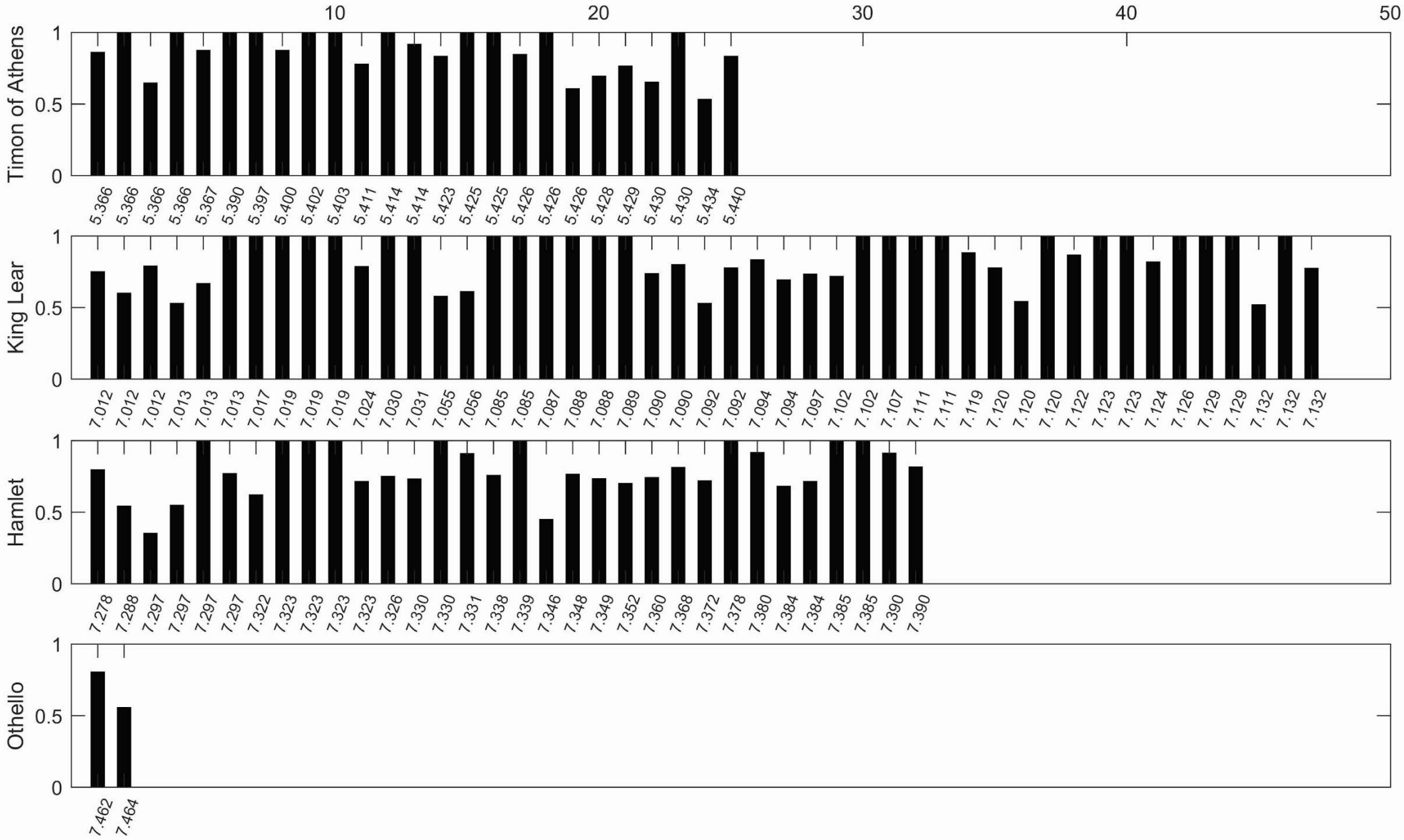
3. **Close reading and critical thinking**. Looking closer at peripheral results, disaffinity and exceptions; letting the research questions guide the exploration; and use skepticism.

4. Distant reading: Making broad claims about historical text data by analysing samples of huge data sets.

Close reading with computers:
Analysing Herman Melville's reading of Shakespeare's plays

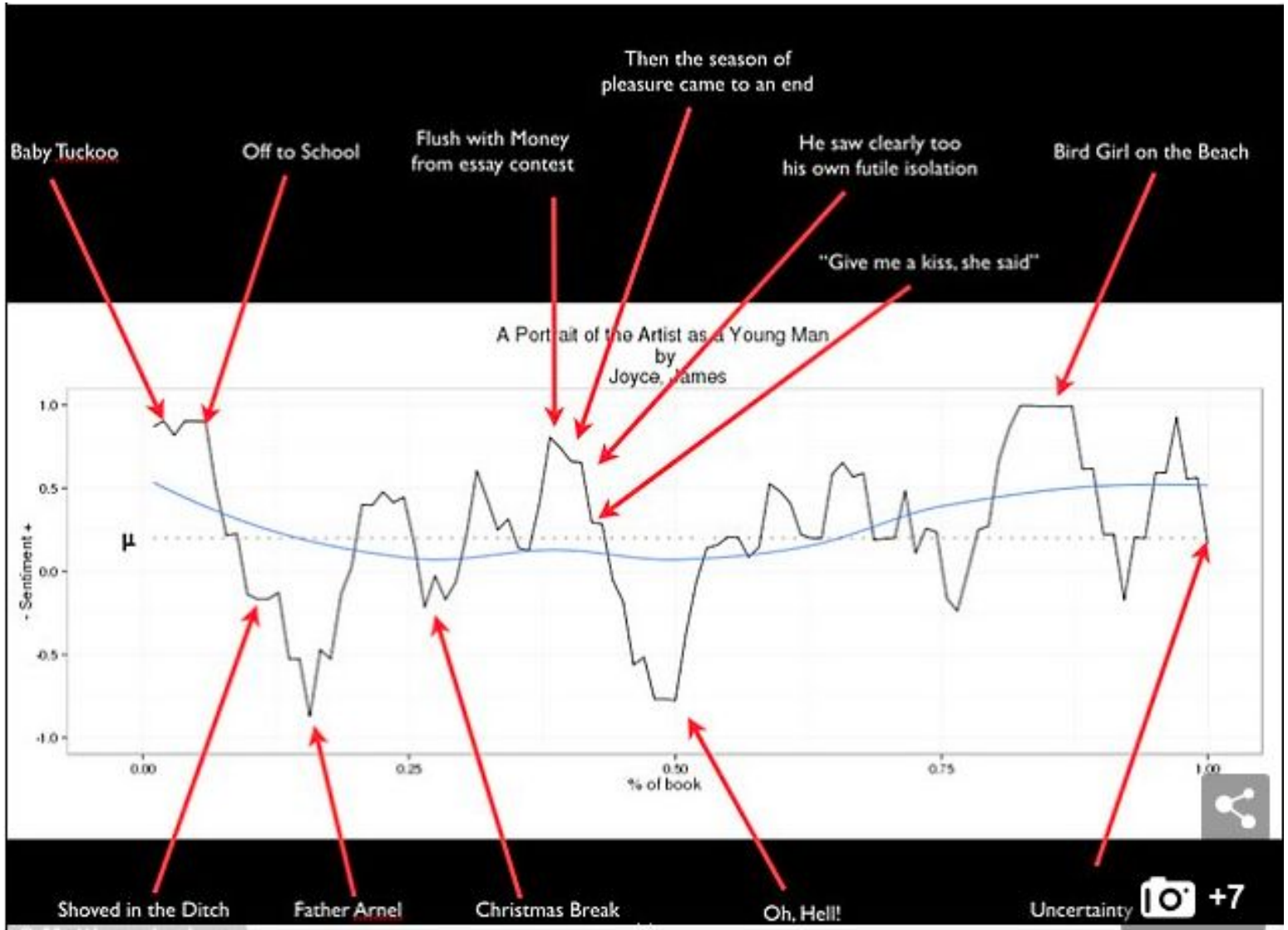



Word Counts of Marked Content by Play in Melville's Set of Shakespeare
Tempest
Two Gentlemen of Verona
Merry Wives of Windsor
Twelfth Night
Measure for Measure
Much Ado About Nothing
Midsummer-Night's Dream
Love's Labors Lost
Merchant of Venice
As you Like It
All's Well that Ends Well
The Taming of the Shrew
Winter's Tale
King John
King Richard II
King Henry IV, Part 1
King Henry V
King Richard III
King Henry VIII
Timon of Athens
Coriolanus
Julius Caesar
Antony and Cleopatra
Cymbeline
Titus Andronicus
King Lear
Romeo and Juliet
Hamlet
Othello
Troilus and Cressida
Pericles, Prince of Tyre
242
228
46
1134
1630
1428
1071
640
160
458
779
154
978
31
187
16
255
33
1349
524
560
714
1796
637
63
763
90
961
77
232
183
Comedies (8948)
Histories (1871)
Tragedies (6185)
Other (415)
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Word Counts

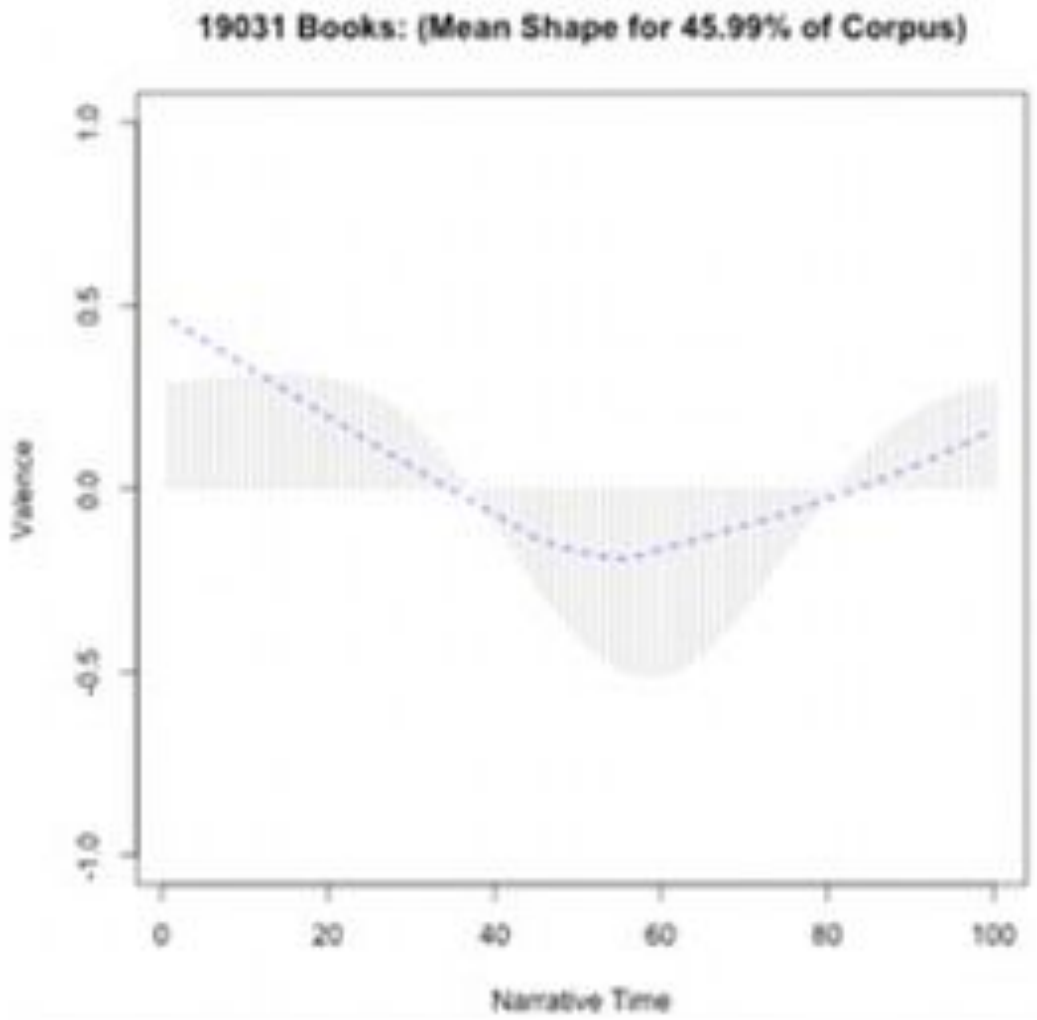

Lexical Uniqueness Values for each Marked Passage in Melville's Marginalia
to Timon of Athens, King Lear, Hamlet, and Othello
10
20
30
40
50
Timon of Athens
King Lear
Hamlet
Othello
1
0.5
0
5.366 5.366 5.366 5.366 5.367 5.390 5.397 5.400 5.402 5.403 5.411 5.414 5.414 5.423 5.425 5.425 5.426 5.426 5.426 5.428 5.429 5.430 5.430 5.434 5.440
7.012 7.012 7.012 7.013 7.013 7.013 7.017 7.019 7.019 7.019 7.024 7.030 7.031 7.055 7.056 7.085 7.085 7.087 7.088 7.088 7.089 7.090 7.090 7.092 7.092 7.094 7.094 7.097 7.102 7.102 7.107 7.111 7.111 7.119 7.120 7.120 7.120 7.122 7.123 7.123 7.124 7.126 7.129 7.129 7.132 7.132 7.132
7.278 7.288 7.297 7.297 7.297 7.297 7.322 7.323 7.323 7.323 7.323 7.326 7.330 7.330 7.331 7.338 7.339 7.346 7.348 7.349 7.352 7.360 7.368 7.372 7.378 7.380 7.384 7.384 7.385 7.385 7.390 7.390
7.462 7.464

Distant Reading:

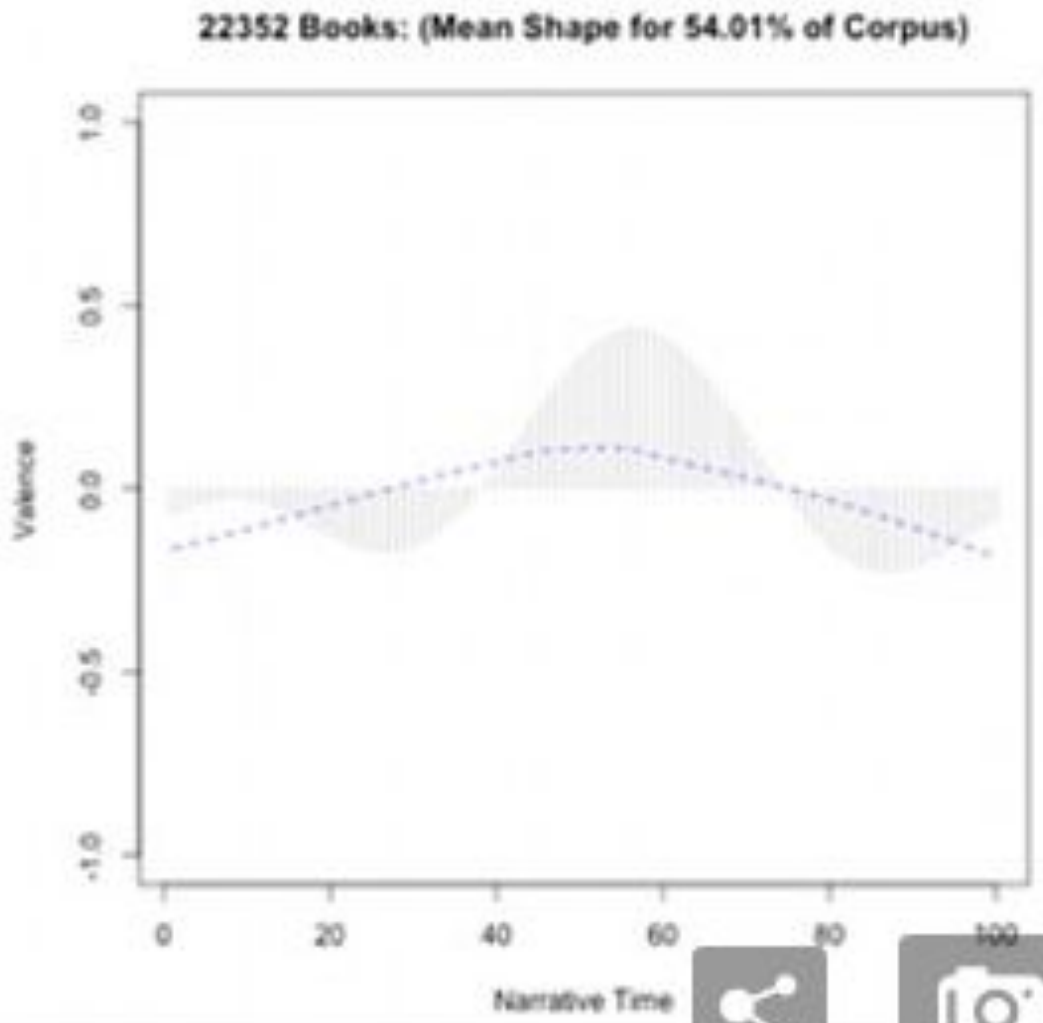
“Professor who analysed 40,000 novels claims there are just SIX possible storylines” (*Daily Mail,* 26 February 2015).

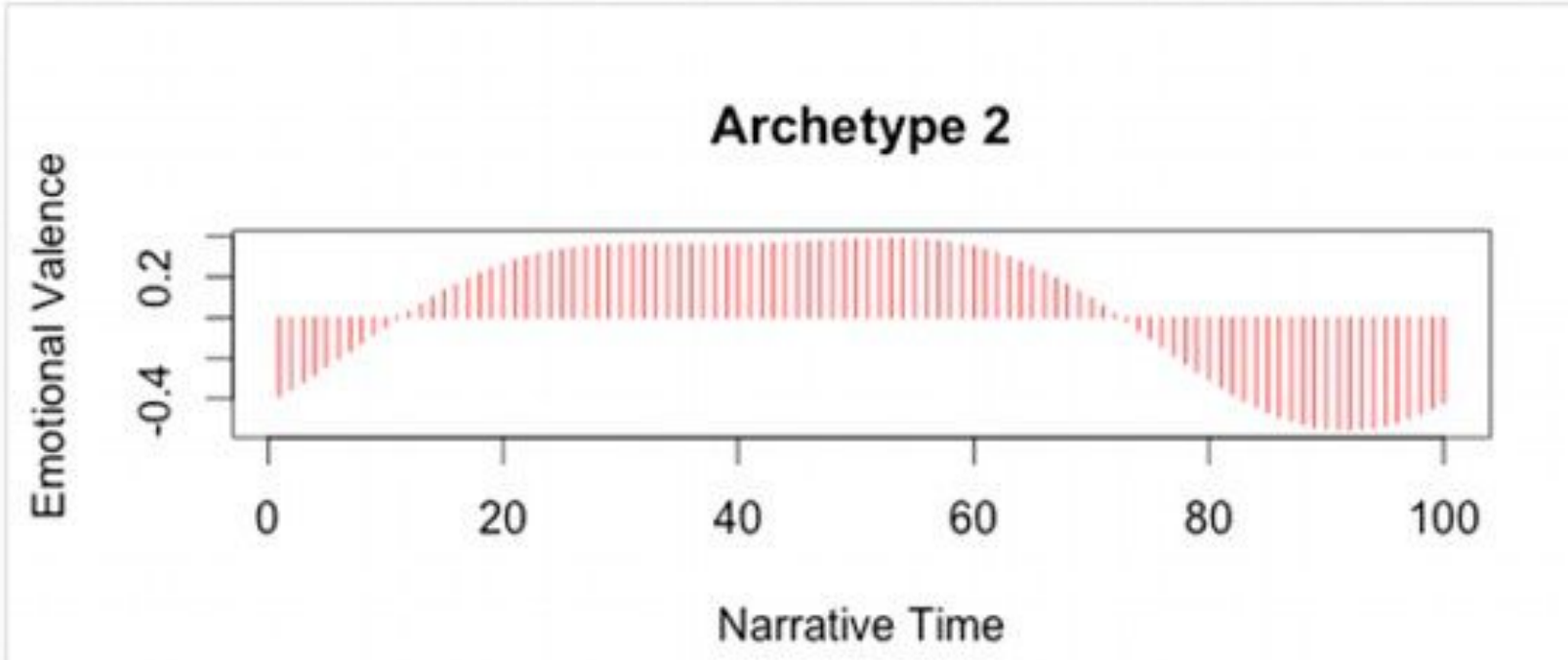


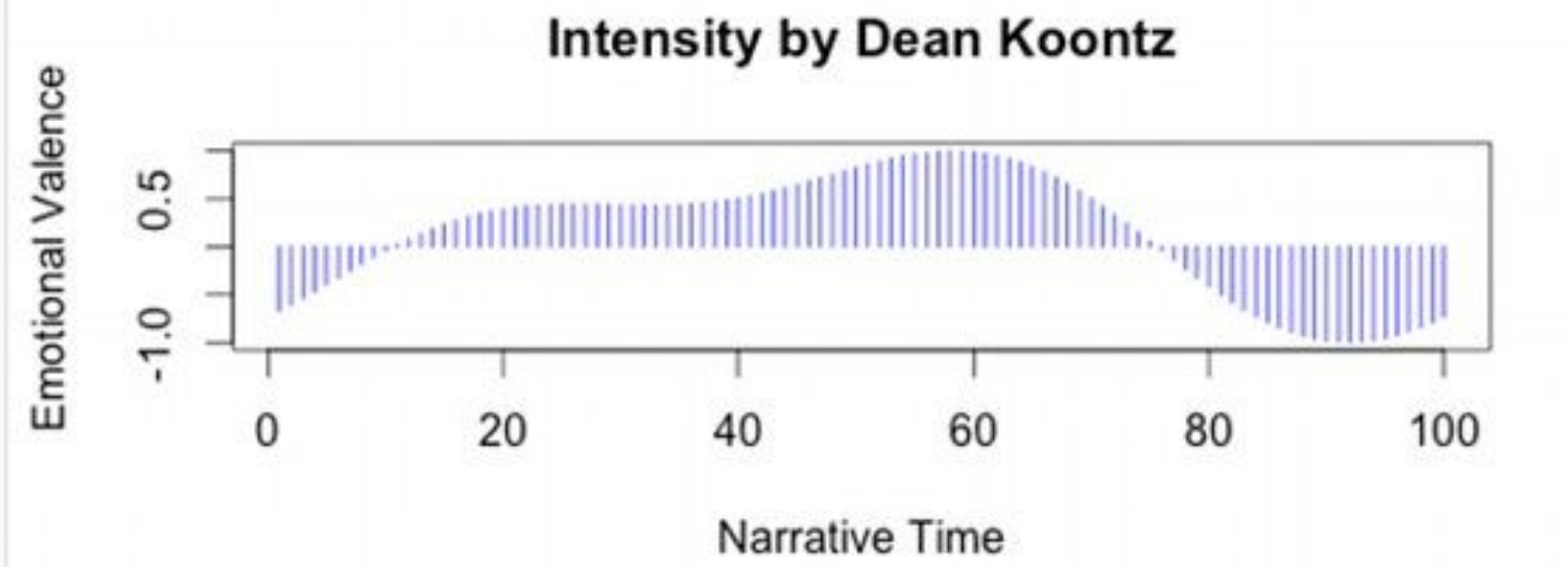
"Man in Hole"



"Man on Hill"




Archetype 2
Emotional Valence
-0.4
0.2
0
20
40
60
80
100
Narrative Time


Intensity by Dean Koontz
Emotional Valence
-1.0
0.5
0
20
40
60
80
100
Narrative Time

Computer-assisted versus computational

Computer-assisted: using out-of-the-box tools to generate results.

Pros: fast results, good interfaces and visualisations.

Cons: lack of control over features and inability to manipulate data; hard to validate and replicate results.

Computational: using programming languages to generate results.

Pros: you have as much control as you can muster; complete customisation.

Cons: requires some knowledge of programming.

Options for performing text analyses

- Voyant Tools <https://voyant-tools.org/>
- AntConc <http://www.laurenceanthony.net/software/antconc/>
- Language databases, such as the Historical Thesaurus of English <https://ht.ac.uk/>.
- Text database tools, such as Hathi Trust Bookworm and Google nGram searches.
- Programming Language: R or Python (we'll demo R because that is what I know) <https://www.r-project.org/>