Information Architecture and Search

INFO 200

Part I

Joseph Janes
Associate Professor, Information School

Information Architecture & Search 1

agenda

- search, and what makes it work: structured search
  - data modeling & encoding
  - database management systems

information behaviors

information **use, seeking, retrieval, organization, encountering, etc**

all imply **seeking** or **finding** at some level (so does information **destruction, censorship** for that matter)

- so **what enables search?** what makes search work, makes it possible?

(and, by the way, all of what is to come are information behaviors too)

search, and what makes it work

for what?

information objects

fair enough - what are information objects?

Web pages, books & published materials, tweets
words, facts
people, organizations
sounds, images, moving images, objects
basically anything

different objects, different searches

the search for each of these is enabled differently, with common features

an example: General Education Requirement Course Search

search

search is the **matching** of **representations** in a **database** by means of an **algorithm**

(representations = **metadata**)

this implies those representations, databases, algorithms have been somehow **created, defined, decided on** (all information behaviors)

and **structured**

so this “simple” instance is referred to as **structured search**

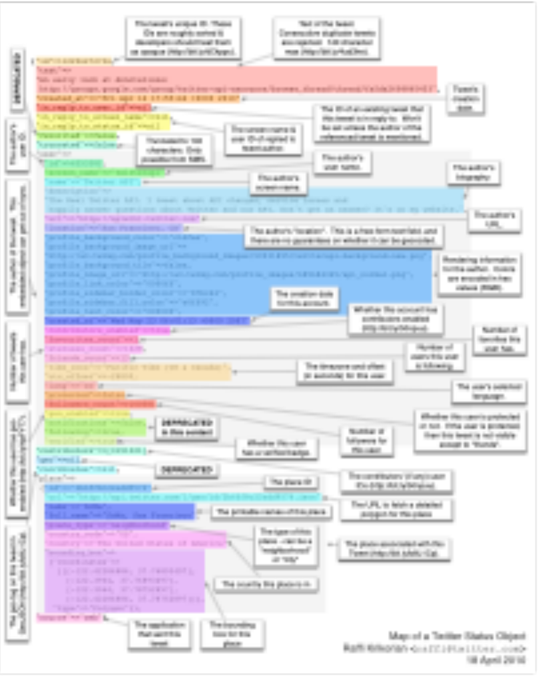
more examples:

UW Faculty/Staff/Student Directory - Seattle

Olympedia

no metadata, no search

these are all examples of **information systems** - so where is **power**?



DEPRECATED
The author's screen name
The author's biography
The author's URL
DEPRECATED
The place ID
The printable names of this place
The country this place is in
The application that sent this tweet
The bounding box for this place

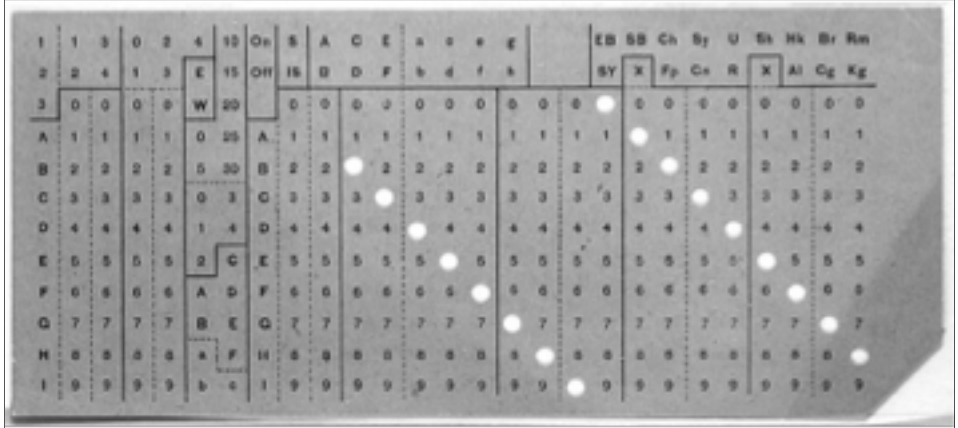
Map of a Twitter Status Object
Raffi Krikorian
18 April 2010

Jacquard Loom

1804

Hollerith punch card

1889

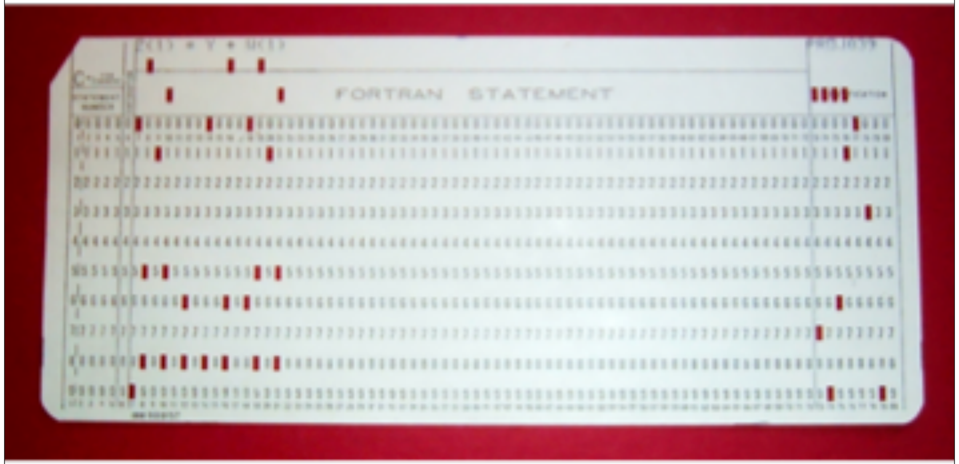


US Census schedule

1940

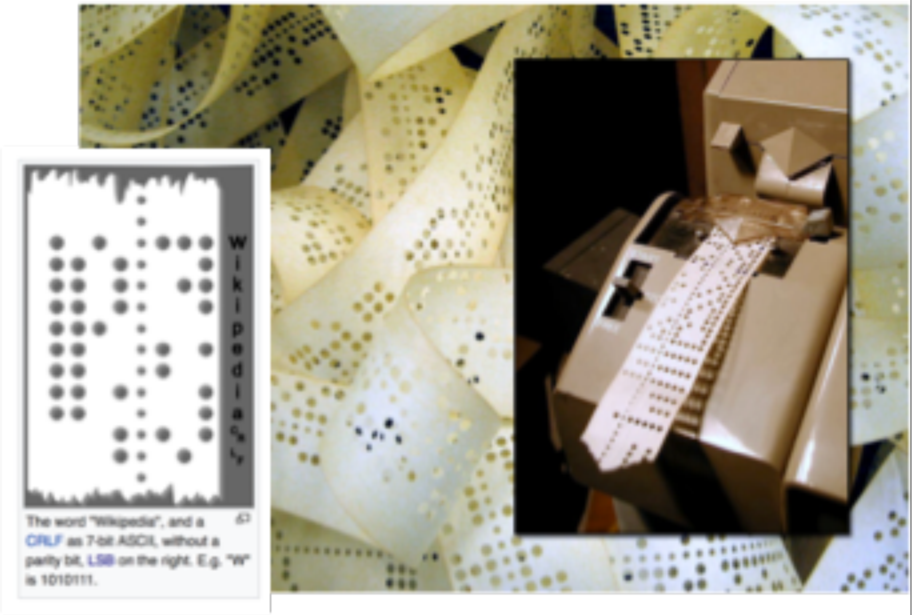
IBM punch card

1928

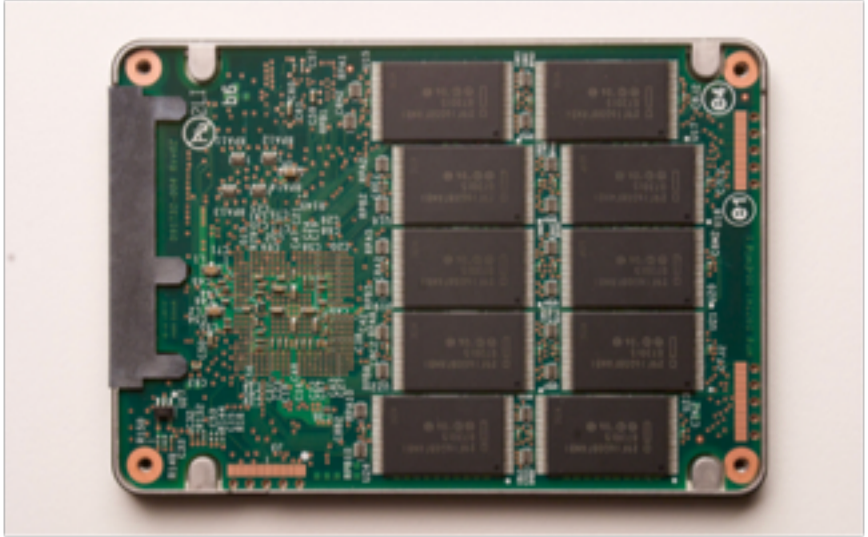


punch paper tape

orig. 1725 looms; computer/communication by 1944



The word "Wikipedia", and a CRLF as 7-bit ASCII, without a parity bit, LSB on the right. E.g. "W" is 1010111.

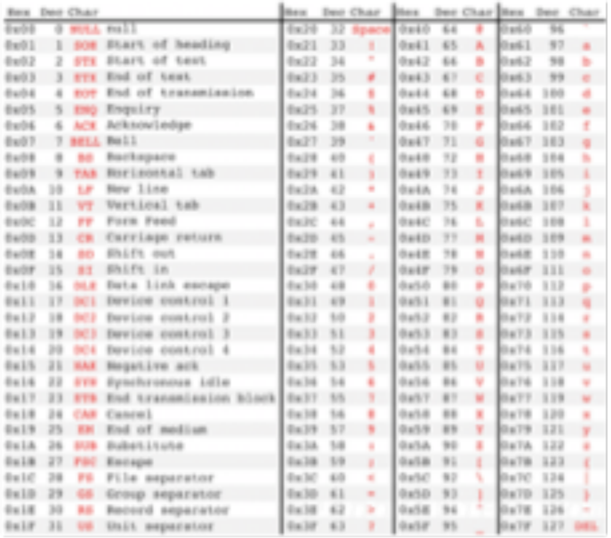


If computers store only patterns of bits, how do we reliably encode **text** into files so that multiple programs can display that text again?

character sets

Define a mapping between patterns of bits and characters

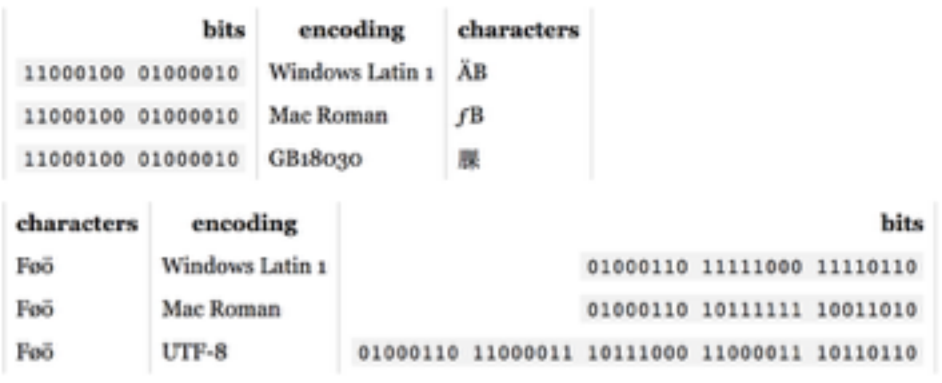
Contain decisions that may have significant social ramifications



Hex	Dec	Char		Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char
0x00	0	NULL	null	0x20	32	Space	0x40	64	@	0x60	96	`
0x01	1	SOH	Start of heading	0x21	33	!	0x41	65	A	0x61	97	a
0x02	2	STX	Start of text	0x22	34	"	0x42	66	B	0x62	98	b
0x03	3	ETX	End of text	0x23	35	#	0x43	67	C	0x63	99	c
0x04	4	EOT	End of transmission	0x24	36	$	0x44	68	D	0x64	100	d
0x05	5	ENQ	Enquiry	0x25	37	%	0x45	69	E	0x65	101	e
0x06	6	ACK	Acknowledge	0x26	38	&	0x46	70	F	0x66	102	f
0x07	7	BELL	Bell	0x27	39	'	0x47	71	G	0x67	103	g
0x08	8	BS	Backspace	0x28	40	(	0x48	72	H	0x68	104	h
0x09	9	TAB	Horizontal tab	0x29	41	)	0x49	73	I	0x69	105	i
0x0A	10	LF	New line	0x2A	42	*	0x4A	74	J	0x6A	106	j
0x0B	11	VT	Vertical tab	0x2B	43	+	0x4B	75	K	0x6B	107	k
0x0C	12	FF	Form Feed	0x2C	44	,	0x4C	76	L	0x6C	108	l
0x0D	13	CR	Carriage return	0x2D	45	-	0x4D	77	M	0x6D	109	m
0x0E	14	SO	Shift out	0x2E	46	.	0x4E	78	N	0x6E	110	n
0x0F	15	SI	Shift in	0x2F	47	/	0x4F	79	O	0x6F	111	o
0x10	16	DLE	Data link escape	0x30	48	0	0x50	80	P	0x70	112	p
0x11	17	DC1	Device control 1	0x31	49	1	0x51	81	Q	0x71	113	q
0x12	18	DC2	Device control 2	0x32	50	2	0x52	82	R	0x72	114	r
0x13	19	DC3	Device control 3	0x33	51	3	0x53	83	S	0x73	115	s
0x14	20	DC4	Device control 4	0x34	52	4	0x54	84	T	0x74	116	t
0x15	21	NAK	Negative ack	0x35	53	5	0x55	85	U	0x75	117	u
0x16	22	SYN	Synchronous idle	0x36	54	6	0x56	86	V	0x76	118	v
0x17	23	ETB	End transmission block	0x37	55	7	0x57	87	W	0x77	119	w
0x18	24	CAN	Cancel	0x38	56	8	0x58	88	X	0x78	120	x
0x19	25	EM	End of medium	0x39	57	9	0x59	89	Y	0x79	121	y
0x1A	26	SUB	Substitute	0x3A	58	:	0x5A	90	Z	0x7A	122	z
0x1B	27	ESC	Escape	0x3B	59	;	0x5B	91	[	0x7B	123	{
0x1C	28	FS	File separator	0x3C	60	<	0x5C	92	\	0x7C	124	|
0x1D	29	GS	Group separator	0x3D	61	=	0x5D	93	]	0x7D	125	}
0x1E	30	RS	Record separator	0x3E	62	>	0x5E	94	^	0x7E	126	~
0x1F	31	US	Unit separator	0x3F	63	?	0x5F	95	_	0x7F	127	DEL

ASCII Character Set (1963)
7 bits per character = 128 possible characters

multiple character sets cause serious problems



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

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

Unicode (1991)

16 bits per character

65,536 possible "code points" (characters)

http://unicode-table.com

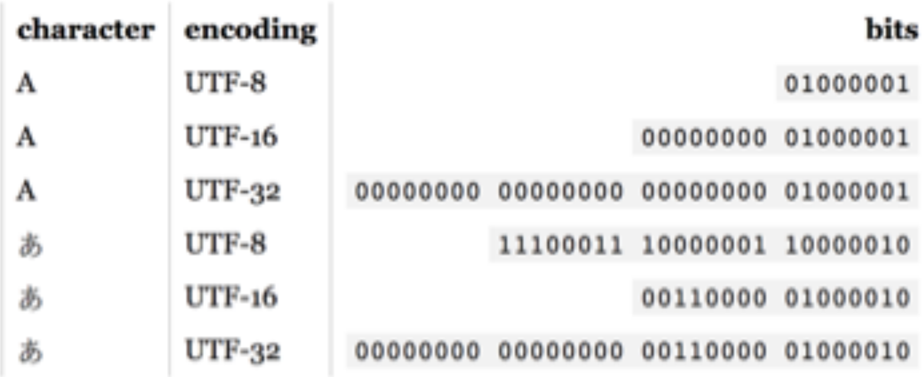
multiplane Unicode

32 bits per-character = 4.2 billion code points

Multiple encoding strategies:

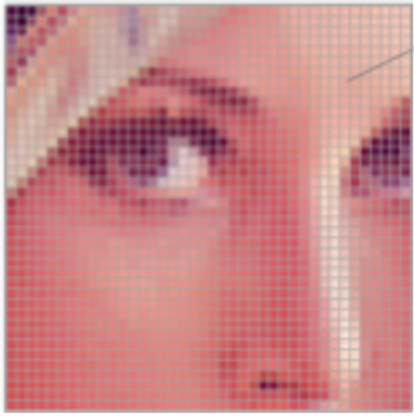
UTF-32	32 bits for every character (UNIX)
UTF-16	16 bits for low chars; 32 for high (Java, .Net)
UTF-8	8 to 32 bits, depending on char (the web, Go)

unicode encoding strategies



character	encoding	bits
A	UTF-8	01000001
A	UTF-16	00000000 01000001
A	UTF-32	00000000 00000000 00000000 01000001
あ	UTF-8	11100011 10000001 10000010
あ	UTF-16	00110000 01000010
あ	UTF-32	00000000 00000000 00110000 01000010

raster image encoding

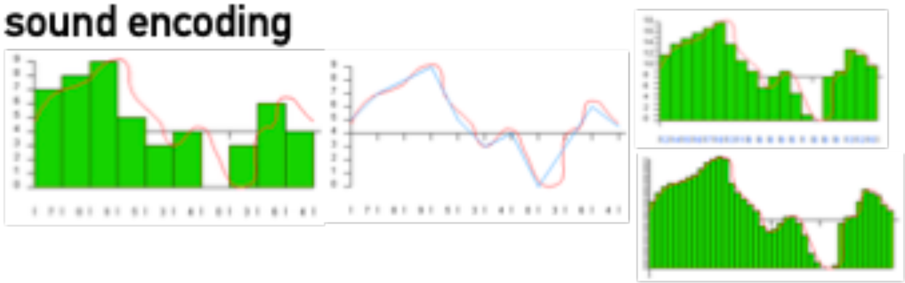


Red: 233
Green: 157
Blue: 144

3 numbers per pixel, each 0-255

24 bits per pixel

sound encoding



Bit depth is the number of **bits** available for each **sample**. The higher the bit depth, the higher the quality of the audio. Bit depth is usually 16 bits on a CD and 24 bits on a DVD.

A bit depth of 16 has a resolution of 65,536 possible values (ranging from 0 to 65,535), and a bit depth of 24 has over 16 million possible values (ranging from 0 to 16,777, 216).

The bit rate is calculated using the formula:

Frequency × bit depth × channels = bit rate

A typical, uncompressed high-quality audio file has a **sample rate** of 44,100 **samples** per second, a bit depth of 16 bits per sample and 2 channels of stereo audio. The bit rate for this file would be:

44,100 samples per second × 16 bits per sample × 2 channels = 1,411,200 bits per second (or 1,411.2 kbps)

A four-minute (240 second) song at this bit rate would create a file size of:

14,411,200 × 240 = 338,688,000 bits (or 40.37 megabytes)

https://electronics.howstuffworks.com/analog-digital3.htm
https://www.bbc.com/bitesize/guides/z7vc7ty/revision/1

representation

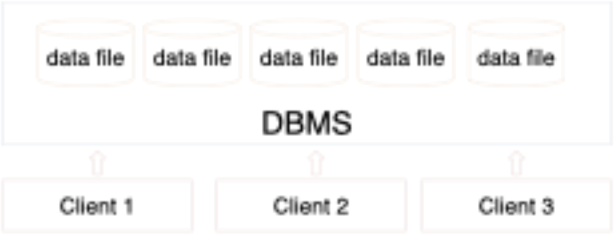
anything stored, digitally or otherwise, has to be **represented** in a structured, standardized way - so that it can be accessed and found

("representation" has multiple meanings, yes?)

we've seen representing text, sound, images, moving images - now another kind at a higher conceptual level

Database Management System (DBMS)

A software process that allows clients to define a data structure; add data to that structure; navigate, tabulate, update, and delete those data; maintain data security and integrity; and automatically recover from failure



relational data modeling

The process of designing a structure capable of holding the data the system needs to track, while avoiding redundancy and ensuring data integrity

Resulting structure is called a **schema**

Visually designed and represented in an **Entity-Relationship Diagram (ERD)**

entity

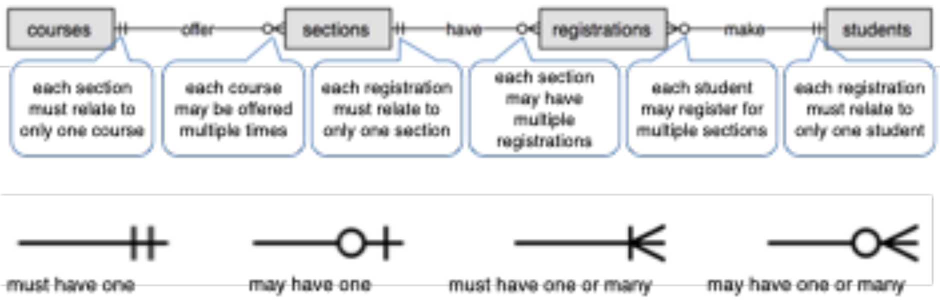
A person, place, thing, or concept included in a system

Look for the **core concepts** as people talk about their data, as well relationships that have their own data

"I want to track the courses we can teach, the sections of those courses we offer, the students who register for those sections, and what their final grades were."

relationships

How do entities relate, and how many instances can they relate to?



each section must relate to only one course

each course may be offered multiple times

each registration must relate to only one section

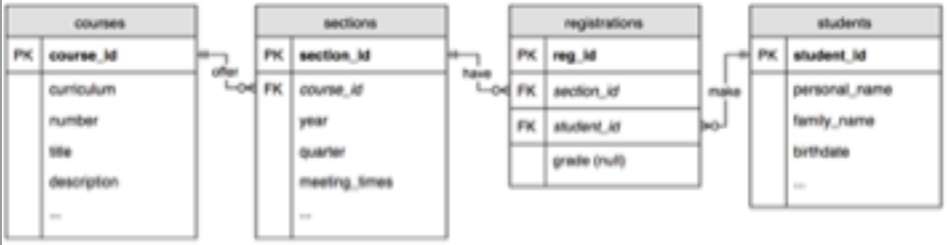
each section may have multiple registrations

each student may register for multiple sections

each registration must relate to only one student

attributes

What do you want to track about each entity?



courses	
PK	**course_id**
	curriculum
	number
	title
	description
	...

sections	
PK	**section_id**
FK	*course_id*
	year
	quarter
	meeting_times
	...

registrations	
PK	**reg_id**
FK	*section_id*
FK	*student_id*
	grade (null)

students	
PK	**student_id**
	personal_name
	family_name
	birthdate
	...

PK = Primary Key = Unique record identifier
FK = Foreign Key = PK value in related table
other examples of keys?

attributes with (null) allow missing data (null values)

representation

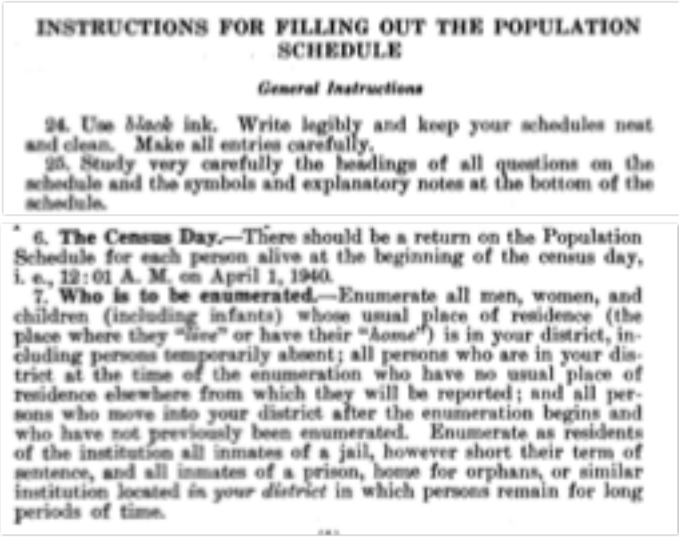
anything stored, digitally or otherwise, has to be **represented** in a structured, standardized way - so that it can be accessed and found

("representation" has multiple meanings, yes?)

US Census schedule

1940

1940 US census instructions



INSTRUCTIONS FOR FILLING OUT THE POPULATION SCHEDULE

General Instructions

24. Use *black* ink. Write legibly and keep your schedules neat and clean. Make all entries carefully.

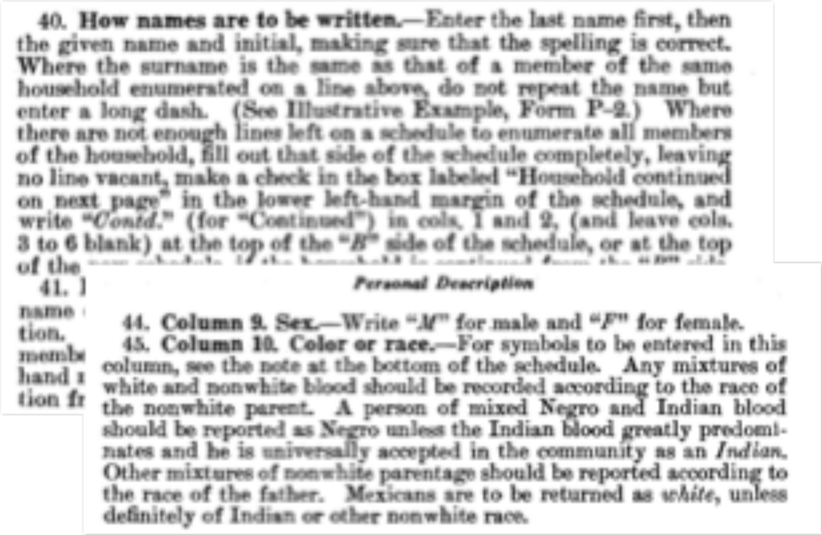
25. Study very carefully the headings of all questions on the schedule and the symbols and explanatory notes at the bottom of the schedule.

6. **The Census Day.**—There should be a return on the Population Schedule for each person alive at the beginning of the census day, i. e., 12:01 A. M. on April 1, 1940.

7. **Who is to be enumerated.**—Enumerate all men, women, and children (including infants) whose usual place of residence (the place where they "*live*" or have their "*home*") is in your district, including persons temporarily absent; all persons who are in your district at the time of the enumeration who have no usual place of residence elsewhere from which they will be reported; and all persons who move into your district after the enumeration begins and who have not previously been enumerated. Enumerate as residents of the institution all inmates of a jail, however short their term of sentence, and all inmates of a prison, home for orphans, or similar institution located *in your district* in which persons remain for long periods of time.

https://1940census.archives.gov/

1940 US census instructions



40. **How names are to be written.**—Enter the last name first, then the given name and initial, making sure that the spelling is correct. Where the surname is the same as that of a member of the same household enumerated on a line above, do not repeat the name but enter a long dash. (See Illustrative Example, Form P-2.) Where there are not enough lines left on a schedule to enumerate all members of the household, fill out that side of the schedule completely, leaving no line vacant, make a check in the box labeled "Household continued on next page" in the lower left-hand margin of the schedule, and write "*Contd.*" (for "Continued") in cols. 1 and 2, (and leave cols. 3 to 6 blank) at the top of the "*B*" side of the schedule, or at the top of the

41. I
name
tion.
membe
hand
tion fr

Personal Description

44. **Column 9. Sex.**—Write "*M*" for male and "*F*" for female.

45. **Column 10. Color or race.**—For symbols to be entered in this column, see the note at the bottom of the schedule. Any mixtures of white and nonwhite blood should be recorded according to the race of the nonwhite parent. A person of mixed Negro and Indian blood should be reported as Negro unless the Indian blood greatly predominates and he is universally accepted in the community as an *Indian*. Other mixtures of nonwhite parentage should be reported according to the race of the father. Mexicans are to be returned as *white*, unless definitely of Indian or other nonwhite race.

https://1940census.archives.gov/

Information Architecture and Search

INFO 200

Part I

Joseph Janes
Associate Professor, Information School