Otvoreno računarstvo

3. Otvorenost zapisa podataka

Creative Commons

Otvoreno računarstvo

3. Otvorenost zapisa podataka

- **Uvod**
- Binarni zapisi
- Prikazi znakova u računalu

- Postoji li uopće ikakva formalna specifikacija zapisa?
  - Format čini prejednostavan da bi ga dokumentirao
  - Format ionako neće koristiti drugi programi / programeri

- Ako postoji formalna specifikacija, je li dostupna/javna/otvorena?
  - Čuvanje intelektualnog vlasništva
  - Očekivanje prihoda od prodaje specifikacije formata ili licencija za korištenje
  - Ograničavanje konkurencije da uvozi podatke u svoja rješenja
  - Dugoročno vezanje korisnika za vlastito rješenje

Klasifikacije zapisa podataka

- Otvoreni – zatvoreni
  - npr. ODT - DOC
- Opće namjene – specijalizirani
  - npr. SGML - SVG
- Podaci – podaci + metapodaci
  - npr. TXT – HTML
- Tekstni – binarni
  - npr. TXT - JPEG

Otvoreno računarstvo

3. Otvorenost zapisa podataka

- Uvod
- **Binarni zapisi**
- Prikazi znakova u računalu

Binarni zapis

- Računalu čitljiv oblik zapisa podataka
- Niz bitova organiziranih u oktete
  - Slike, zvuk, filmovi, programi
  - Tekst ?!
- Sadržaj binarnih zapisa
  - Jednostavan/plošni
  - Složeni (zaglavlja, sadržaj)

Kako zapisivati podatke - **prenosivost**

- Ako na jednoj platformi zapišemo tip podatka *int*, što ćemo pročitati na drugoj?
  - Big ili little endian?
  - Raspon vrijednosti / duljina zapisa?
  - NBC, jedinični ili dvojni komplement?
- A što ako zapišemo *float*?
  - IEEE 754? ANSI C, C++
- Zapisivanje strukture podataka (mem. područja)?
  - Poravnavanje riječi?
- Primjer: GNU Gnulib
  - https://www.gnu.org/software/gnulib/manual/
  - Target platforms
  - Portability guidelines

Otvoreno računarstvo

3. Otvorenost zapisa podataka

- Uvod
- Binarni zapisi
- **Prikazi znakova u računalu**

Tekstni zapisi

- Slojevitost pristupa tekstnim zapisima:
  - Sloj prikaza znakova
  - Sloj zapisa podataka
- I ovi se zapisi sastoje od niza okteta
  - organizirani po retcima uz posebne znakove za kraj retka
  - okteti ograničeni na ljudima čitljive znakove i manji broj posebnih znakova
- Ljudima u cijelosti čitljiv oblik zapisa podataka
- Mogu sadržavati meta-podatke
  - također ljudima čitljivi
- Sadržaj odvojen od prikaza/reprezentacije

```
1 {\rtf1\ansi\ansicpg1250\deff0\nouicompat\def
  lang1050{\fonttbl{\f0\fnil\fcharset238
  Calibri;}{\f1\fnil\fcharset0 Calibri;}}
2 {\colortbl ;\red255\green0\blue0;}
3 {\*\generator Riched20
  10.0.19041}\viewkind4\uc1
4 \pard\sa200\sl276\slmult1\f0\fs22 Pozdrav
  \b svima \b0 od \cf1\ul
  mene!\cf0\ulnone\f1\par
  }
  NUL
```

Pozdrav **svima** od mene!

Prenosivost tekstnih zapisa

- Prenosivost
  - Puno veća od binarnog oblika zapisa
  - Problemi novog retka, različitih zapisa znakova ...
  - Koriste se za posredne/univerzalne formate zapisa
  - Manja gustoća zapisa podataka u odnosu na binarne

Prikaz znakova u računalu

- Računala razumiju samo 0 i 1
- Znakovi (grafemi) – skupovi bitova
- Koji skup bitova označava koje slovo?
- Kako će sva računala i svi programi znati ispravno protumačiti bitove?
  - **kôdna stranica** (*code page*)
- Dvije vrste znakova:
  - ispisivi
  - kontrolni (neispisivi)

ASCII

- *American Standard Code for Information Interchange* (1963)
- ANSI standard
- **7-bitni** zapis -> 2^7 = **128** različitih znakova
- Najviši, 8. bit: paritet ili "0"
- 95 ispisivih znakova
- 33 kontrolna znaka

Char	Dec	Oct	Hex	Char	Dec	Oct	Hex	Char	Dec	Oct	Hex	Char	Dec	Oct	Hex
(nul)	0	0000	0x00	(sp)	32	0040	0x20	@	64	0100	0x40	`	96	0140	0x60
(soh)	1	0001	0x01	!	33	0041	0x21	A	65	0101	0x41	a	97	0141	0x61
(stx)	2	0002	0x02	"	34	0042	0x22	B	66	0102	0x42	b	98	0142	0x62
(etx)	3	0003	0x03	#	35	0043	0x23	C	67	0103	0x43	c	99	0143	0x63
(eot)	4	0004	0x04	$	36	0044	0x24	D	68	0104	0x44	d	100	0144	0x64
(enq)	5	0005	0x05	%	37	0045	0x25	E	69	0105	0x45	e	101	0145	0x65
(ack)	6	0006	0x06	&	38	0046	0x26	F	70	0106	0x46	f	102	0146	0x66
(bel)	7	0007	0x07	'	39	0047	0x27	G	71	0107	0x47	g	103	0147	0x67
(bs)	8	0010	0x08	(	40	0050	0x28	H	72	0110	0x48	h	104	0150	0x68
(ht)	9	0011	0x09	)	41	0051	0x29	I	73	0111	0x49	i	105	0151	0x69
(nl)	10	0012	0x0a	*	42	0052	0x2a	J	74	0112	0x4a	j	106	0152	0x6a
(vt)	11	0013	0x0b	+	43	0053	0x2b	K	75	0113	0x4b	k	107	0153	0x6b
(np)	12	0014	0x0c	,	44	0054	0x2c	L	76	0114	0x4c	l	108	0154	0x6c
(cr)	13	0015	0x0d	-	45	0055	0x2d	M	77	0115	0x4d	m	109	0155	0x6d
(so)	14	0016	0x0e	.	46	0056	0x2e	N	78	0116	0x4e	n	110	0156	0x6e
(si)	15	0017	0x0f	/	47	0057	0x2f	O	79	0117	0x4f	o	111	0157	0x6f
(dle)	16	0020	0x10	0	48	0060	0x30	P	80	0120	0x50	p	112	0160	0x70
(dc1)	17	0021	0x11	1	49	0061	0x31	Q	81	0121	0x51	q	113	0161	0x71
(dc2)	18	0022	0x12	2	50	0062	0x32	R	82	0122	0x52	r	114	0162	0x72
(dc3)	19	0023	0x13	3	51	0063	0x33	S	83	0123	0x53	s	115	0163	0x73
(dc4)	20	0024	0x14	4	52	0064	0x34	T	84	0124	0x54	t	116	0164	0x74
(nak)	21	0025	0x15	5	53	0065	0x35	U	85	0125	0x55	u	117	0165	0x75
(syn)	22	0026	0x16	6	54	0066	0x36	V	86	0126	0x56	v	118	0166	0x76
(etb)	23	0027	0x17	7	55	0067	0x37	W	87	0127	0x57	w	119	0167	0x77
(can)	24	0030	0x18	8	56	0070	0x38	X	88	0130	0x58	x	120	0170	0x78
(em)	25	0031	0x19	9	57	0071	0x39	Y	89	0131	0x59	y	121	0171	0x79
(sub)	26	0032	0x1a	:	58	0072	0x3a	Z	90	0132	0x5a	z	122	0172	0x7a
(esc)	27	0033	0x1b	;	59	0073	0x3b	[	91	0133	0x5b	{	123	0173	0x7b
(fs)	28	0034	0x1c	<	60	0074	0x3c	\	92	0134	0x5c	|	124	0174	0x7c
(gs)	29	0035	0x1d	=	61	0075	0x3d	]	93	0135	0x5d	}	125	0175	0x7d
(rs)	30	0036	0x1e	>	62	0076	0x3e	^	94	0136	0x5e	~	126	0176	0x7e
(us)	31	0037	0x1f	?	63	0077	0x3f	_	95	0137	0x5f	(del)	127	0177	0x7f

Kontrolni znakovi

- Primjeri:
  - prelazak u novi red, povratak na početak reda, tabulator, zvono, backspace, escape
- Problem: **višeznačnost** na različitim platformama :-(
- Primjer:
  - prelazak u **"novi red"**
  - prisjetite se - programski jezik C: \r\n

- **CR** – Carriage Return – pomicanje na početak reda
- **LF** – Line Feed – spuštanje za jedan redak

CR	Commodore, Mac OS (do v.9)
LF	Unix, Linux i slični sustavi
CR+LF	MS-DOS, Windows

  - Uporaba CR+LF za Internet protokole, ponekad se tolerira samo LF

Ispisivi znakovi

- Znamenke, slova, znakovi
- 7 znakova za akcente
  - mogu se kombinirati sa slovima, ovisno o programskoj podršci
- Brojevi: 0011 + BCD vrijednost
  - prisjetite se: **Binary Coded Decimal**
- Slova:
  - abecedni poredak
  - razlika između velikog i malog u jednom bitu
  - jednostavno sortiranje, pretvorbe

ASCII problemi

- Premalo znakova (127)
- Potrebni dodatni znakovi za latinicu
- Potrebna dodatna pisma
  - ćirilica, glagoljica, grčko pismo...

- Rješenje: uvođenje različitih proširenja za različite jezike

IBM PC

- Proširenje ASCII-ja
  - IBM-ov zapis
- Korištenje u MS-DOS-u

- **8-bitni** zapis
  - Prvih 128 znakova je jednako kao ASCII

- Verzije po regijama/pismima
  - CP 850 – Latin I - Western European
  - **CP 852 – Latin II - Eastern European**

MS Windows ANSI

- Temeljen na ANSI prijedlozima, nikad normiran!
- 8-bitni zapis
  - **1252** – West European Latin
  - **1250** – East Eureopean Latin

- Pitanje za programere: **podržati ili ne?**
  - Ne -> nije moguće raditi s ovakvim dokumentima
  - Ne -> pogreške pri radu programa
  - Da -> tada je bila de-facto norma

ISO/IEC 8859

- Alias ISO 8859
- Nadogradnja ASCII-a (kompatibilnost!)
  - **8-bitni** zapis -> **256** znakova
- Podijeljen u numerirane dijelove:
  - ISO 8859-1 ... ISO 8859-16
  - svaki dio prilagođen određenom pismu ili regiji

Neke kôdne stranice ISO 8859

Oznaka	Naziv	Opis
ISO 8859-1	Latin-1 Western European	većina zapadnoeuropskih zemalja
ISO 8859-2	Latin-2 Eastern European	srednjeistočna Europa koja koristi latinicu **(Hrvatska!)**
ISO 8859-5	Latin/Cyrillic	slavenski jezici koji koriste ćirilicu
ISO 8859-15	Latin-9	nadogradnja ISO 8859-1 (dodani znakovi €, Œ, Ÿ ...) potpuni francuski, finski, estonski
ISO 8859-16	Latin-10 South-Eastern European	Srednjeistočna Europa **(Hrvatska!)** uz finski, njemački, francuski... € znak

Problemi

- Kako pisati jedan dokument koji u sebi sadrži više jezika/pisama?
- Kako pisati strana imena u poruci e-pošte raspodijeljenom timu?
- Kako podržati azijske jezike, koji sadrže i po nekoliko tisuća različitih znakova?

- Tražimo **jedinstveno** rješenje!

Unicode

- **Unicode NIJE kôdna stranica!**
- **Ujedinjavanje regionalnih norma** u jednu
- Svaki znak – jedna numerička vrijednost, kôdna točka (*code point*)
- Oznaka *U+numerička_vrijednost*
- Potencijalno ~1,1 milijun znakova
- Trenutno zauzeto: ~10% prostora :-)
  - u verziji Unicode 13.0 (ožujak 2020.) – ~144 000 znakova
  - 154 pisma
  - skupovi simbola i *emojija*

Unicode - svojstva

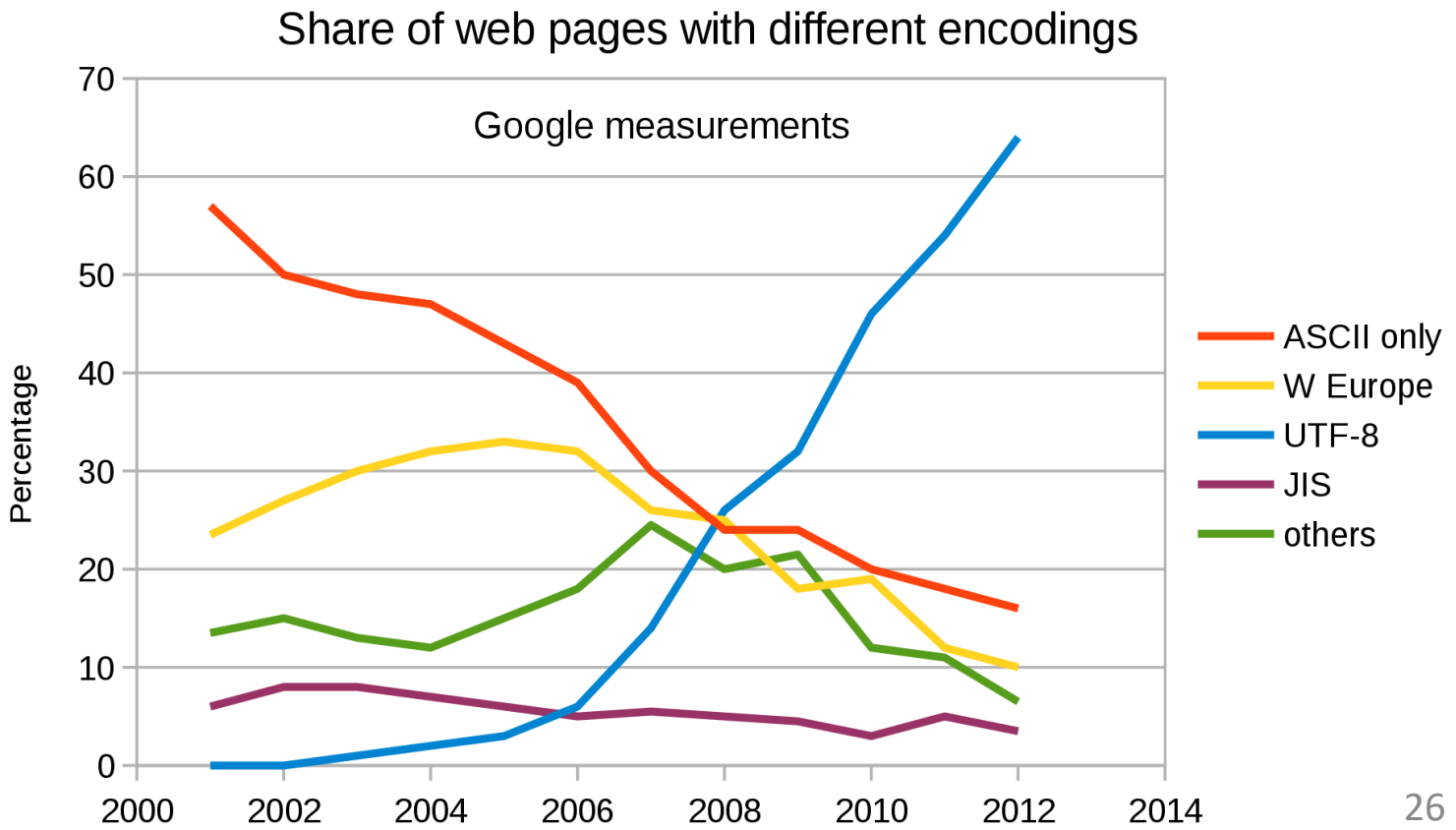
- Mapiranje prvih 256 znakova na **numeričke oznake identično s ISO 8859-1**
- Znakovi podijeljeni u "ravnine" (*planes*)
- Gotovo svi često korišteni znakovi su u prvih 64K numeričkih oznaka
  - *Basic Multilingual Plane* - **BMP**
- Višestruko pojavljivanje nekih znakova
  - lakša konverzija
- Ocrtava normu ISO/IEC 10646

Unicode - kodiranja

- Kako zapisati Unicode kôdne točke svakog znaka (*code point*)?
- 3 vrste:
  - UTF-**8**
  - UTF-**16**
  - UTF-**32**

UTF-8

- **Varijabilna** dužina (**1, 2, 3** ili **4** bajta)
- 1 B: Prvih 128: zapis identičan ASCII-ju
- 2 B: Ostali često korišteni znakovi (Hrvatska!)
- 3 B: Ostali znakovi iz BMP
- 4 B: Znakovi iz ostalih ravnina
- **Najviše raširen**
- Korištenje:
  - XML, e-pošta
  - Web stranice (>95% u 2020.)
  - Unix/Linux



UTF-16

- **Varijabilna** dužina (**2** ili **4** bajta)
  - 2B: Gotovo svi često korišteni znakovi
  - 4B: Ostatak
- Problem: kojim se redom šalju bajtovi?
  - little/big endian -> UTF-16LE, UTF-16BE
  - UTF-16 (BOM – Byte Order Mark - na početku)
- Korištenje
  - interna reprezentacija znakova
    - Windows NT/2000/XP/CE <- od 2019. se preporuča UTF-8, no i dalje postoje problemi
      - Unicode in Microsoft Windows
    - Java i .NET programska okruženja

UTF-32

- **Fiksna** dužina (**4** bajta)
- Trenutno je vrlo rijetko pojavljivanje znakova za koje su doista potrebna 4 bajta
- Rijetko korišten

- Kolika je duljina ovog dokumenta u oktetima?
  - Ovisi ;) <ž/>
  - ASCII: ne može se zapisati!
  - ISO-8859-1 ne može se zapisati!
  - ISO-8859-2: 4 okteta `3C BE 2F 3E`
  - UTF-8: 5 okteta `3C C5 BE 2F 3E`
  - UTF-8 (BOM): 8 okteta `EF BB BF 3C C5 BE 2F 3E`
  - UTF-16: 10 okteta `FF FE 3C 00 7E 01 2F 00 3E 00`

Primjer pretvaranja *Unicode* kôdne točke u UTF-8...

Koji je Unicode code point za znak ž? 0x017E (U-017E)

Pravila:

BOM -> UTF16(BE)= FE FF, UTF16(LE)=FF FE,
BOM -> UTF8=EF BB BF (ne treba biti prisutan kod UTF8)

Algoritam konverzije, *code point* -> UTF-8:

U+000000-U+00007F (0xxxxxxx)	0xxxxxxx
U+000080-U+0007FF (00000yyy xxxxxxxx)	110yyyxx 10xxxxxx
U+000800-U+00FFFF (yyyyyyyy xxxxxxxx)	1110yyyy 10yyyyxx 10xxxxxx
U+010000-U+10FFFF (000zzzzz yyyyyyyy xxxxxxxx)	11110zzz 10zzyyyy 10yyyyxx 10xxxxxx

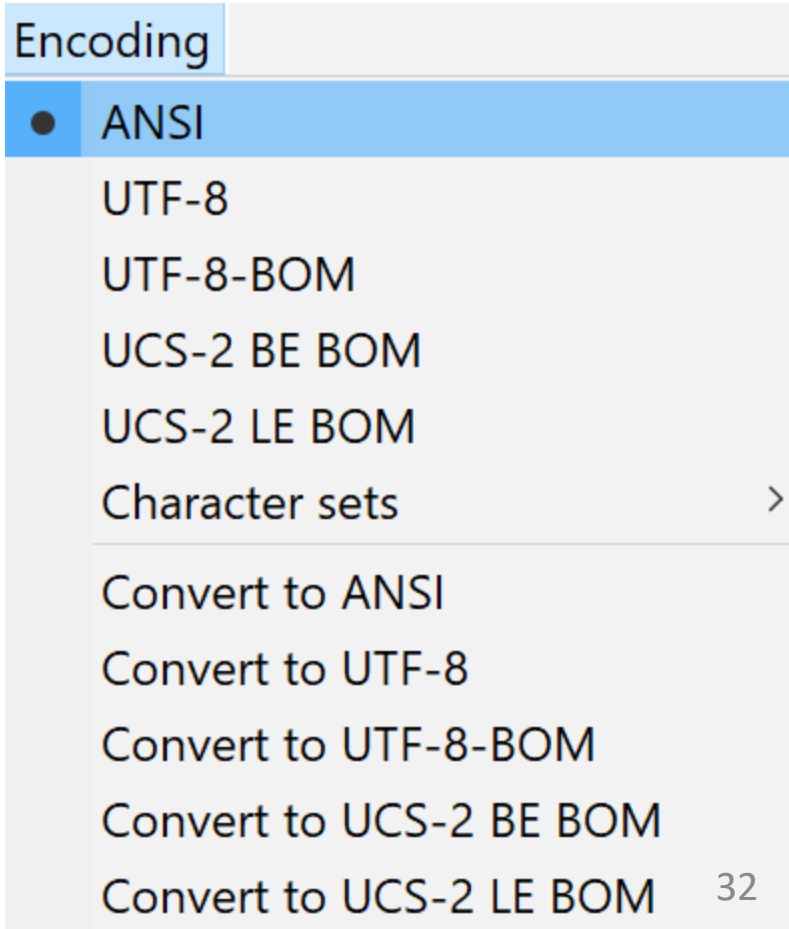
Razlike među UTF-kôdnim stranicama

- http://unicode.org/faq/utf_bom.html#gen6

Name	UTF-8	UTF-16	UTF-16BE	UTF-16LE	UTF-32	UTF-32BE	UTF-32LE
Smallest code point	0000	0000	0000	0000	0000	0000	0000
Largest code point	10FFFF	10FFFF	10FFFF	10FFFF	10FFFF	10FFFF	10FFFF
Code unit size	8 bits	16 bits	16 bits	16 bits	32 bits	32 bits	32 bits
Byte order	N/A	<BOM>	big-endian	little-endian	<BOM>	big-endian	little-endian
Fewest bytes per character	1	2	2	2	4	4	4
Most bytes per character	4	4	4	4	4	4	4

Kako prepoznati kôdnu stranicu?

- Oznaka na početku datoteke
- Ručni odabir u programu
- Web, e-mail: oznaka u zaglavlju



There Ain’t No Such Thing As Plain Text

Joel Spolsky

A gdje smo mi?

- Mnogi još uvijek koriste ISO 8859-2
  - mnogi problemi s interoperabilnošću
- UTF-8
  - najbolji dugoročni izbor

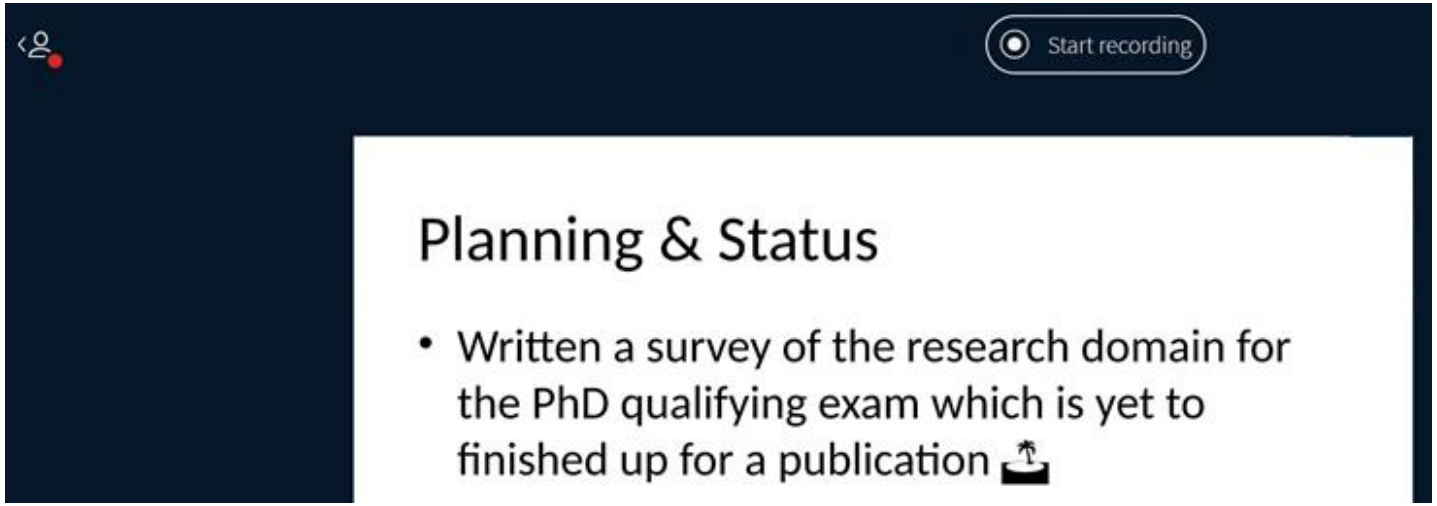
- Nemojte misliti da ovo nikada nećete vidjeti! :-)

Prikazano u / Napisano u	CP852	windows-1250	iso-8859-2	utf-8
CP852	čćšđž ČĆŠĐŽ	ź†çĐ§ ¬Źć Ń¦	��çĐ§ Ź�ć ŃŚ	���Ч ���Ѧ
windows-1250	ŕŠÜ× ╚ĂŐđÄ	čćšđž ČĆŠĐŽ	čć�đ ČĆ�Đ	��Ъ ЂDў
iso-8859-2	ŕŠ╡ż ╚Ăęđ«	čćąđľ ČĆ©Đ®	čćšđž ČĆŠĐŽ	��Ъ ЂΣЮ
utf-8	–Ź—ç┼і—Ĺ┼ż —î—ć┼á—É┼Ż	ĂŤĂ‡Ĺ˘Ä·ĹľĂŚĂ†Ĺ Ă€Ĺ˝	Ă€Ă€ĹĄĂ€Ĺž Ă€Ă€Ĺ Ă€Ĺ˝	čćšđž ČĆŠĐŽ

Što kada stvari krenu naopako?

Planning & Status

- Written a survey of the research domain for the PhD qualifying exam which is yet to finished up for a publication ☺



Ima li toga još?

- Ima :-)
- UTF-9 i UTF-18
  - April Fool's Day RFC dokument
  - tehnički izvediv :-)
- Postoji još mnogo kôdnih stranica
  - nama manje važnih

Otvoreno računarstvo

3. Otvorenost zapisa podataka

- Uvod
- Binarni zapisi
- Prikazi znakova u računalu

Korišten *CreativeCommons* sadržaj

- Chris55 - Own work; Usage of the main encodings on the web from 2001 to 2012 as recorded by Google, CC BY-SA 4.0