Unicode CLDR a standardizacija

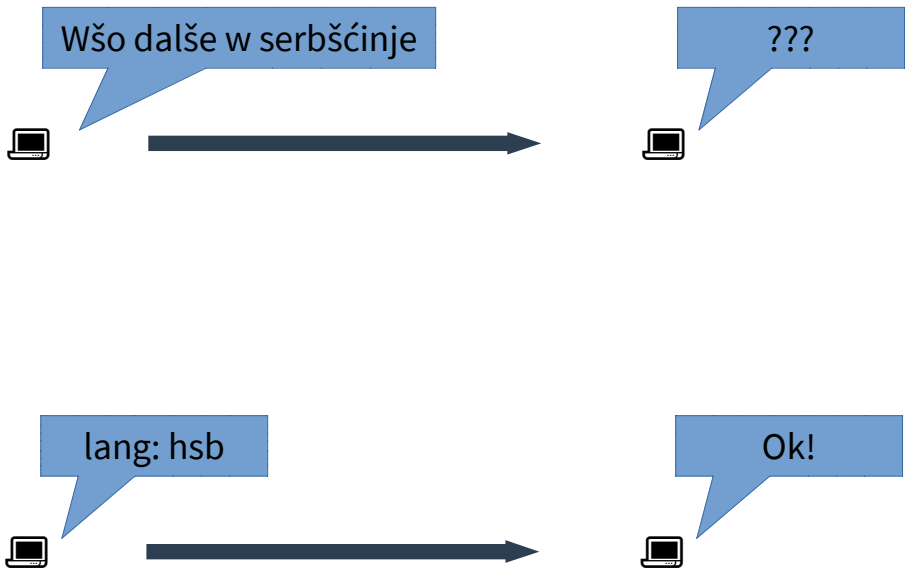
za widźomnosć serbšćiny w digitalnym swěće

Wito Bejmak
konferenca k digitalnej strategiji
Budyšin, 30.11.2019

Što je standardizacija w digitalnym swěće?

- **mjezynarodnje připóznate rjadowanja**
  - přez mjezynarodnje normowace organizacije (n.př. ISO)
- **zhromadne postajenja, regularije a terminologije**
- **je trjeba za interoperabilitu, kompatibilitu a trajnosć:**
  - potajkim kmanosć, zo rozdźělne techniske systemy zhromadnje dźěłaja, so zrozumja a daty wuměnić móža a to tež přez dlěši časowy horicont

Mašiny so dojednaja



Přikłady za standardy

- ISO 639, BCP 47
- Unicode (abo ISO 10646)
- Unicode Locale Data => Unicode CLDR
- šěrši kontekst – terminologije (koděrowane)

ISO 639 – kody za rěče

- mjezynarodna norma za definowanje koda za rěče
- kak: kode + pomjenowanje rěče
- za serbsku rěč wosebje ISO 639-2 (tak mjenowany 3-alpha-code)
- Nastaće ISO 639-2: „The list was largely based on the MARC Code List for Languages, which has been in wide use since 1968.“

  https://www.loc.gov/standards/iso639-2/faq.html

‘Serbske stawizny‘ w ISO 639

Sorbian languages [wen]
UF Lusatian Sorbian languages
Wendic languages
Collective code for:
Upper Sorbian

- **Serbske knihi na ameriskich uniwersitnych bibliotekach**
  - Archibald G. Coolidge (ameriski historikar a dipl. we Wienje) dari 1895 Harvardowej uniwersiće wjazane Časopisy Maćicy Serbskeje
  - Kónc 1960-tych lět katalogizowanje do computerowych systemow: kode: **wen** (USMARC code list for languages)
- **wen: 1989 do noweje ISO 639-2 (přewzate z USMARC code list)**
- **dsb a hsb: 2003 přiwzaće do ISO 639-2**
  - Rozeznawanje hornjo- a delnjoserbšćiny jako hesło hižo 1986 w LoC
- **wen: jako rěčna skupina přidatnje do 639-5**
- **dsb a hsb: tež do ISO 639-3 (etnologiska lisćina)**

‘Serbske stawizny‘ w ISO 639-2

ISO 639-2 Code	English name of Language	French name of Language	Date Added or Changed	Category of Change
tlh	Klingon; tlhIngan-Hol	klingon	2004-02-24	Add
byn	Blin; Bilin	blin; bilen	2003-10-27	Add
jbo	Lojban	lojban	2003-09-02	Add
dsb	Lower Sorbian	bas-sorabe	2003-09-01	Add
hsb	Upper Sorbian	haut-sorabe	2003-09-01	Add
csb	Kashubian	kachoube	2003-05-19	Add

Doskónčny staw w ISO 639-2

ISO 639-2 Code	ISO 639-1 Code	English name of Language	French name of Language	German name of Language
dsb		Lower Sorbian	bas-sorabe	Niedersorbisch
hsb		Upper Sorbian	haut-sorabe	Obersorbisch
wen		Sorbian languages	sorabes, langues	Sorbisch (Andere)

https://www.loc.gov/standards/iso639-2/php/code_list.php

BCP 47

- Internet Engineering Task Force (IETF) Best Current Practice (BCP) 47
- ‘IETF BCP 47 language tag‘ je kode za identifikaciju přirodnych rěčow
- bazěruje na druhich standardach
- kombinuje kode za rěč (ISO 639) a kraj (ISO 3166-1) a dalše (ISO 15924 za pismo)

  en-us en-gb

  hsb hsb-de

  dsb dsb-de

Znamješkowa sadźba Unicode

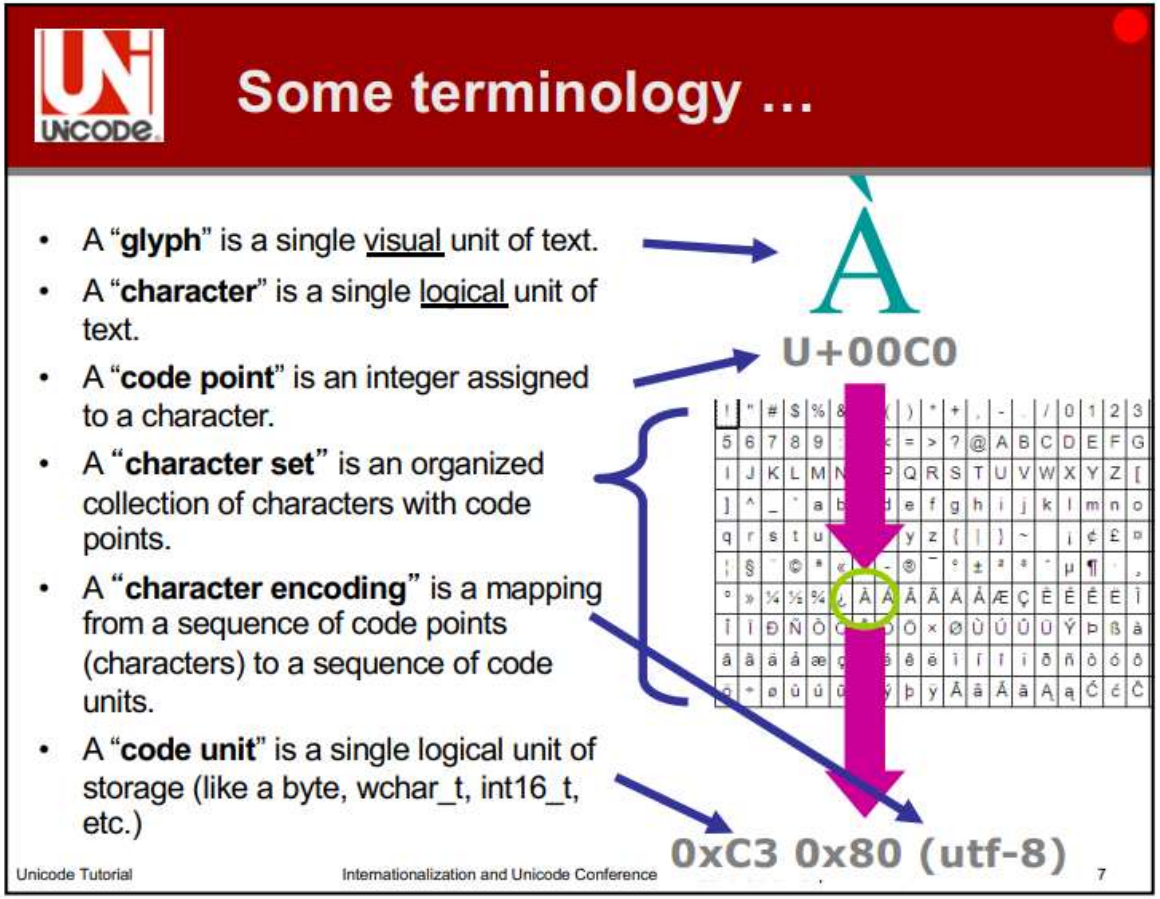
- **definuje znamješka (Schriftzeichen) z krutej ličbu (tak mjenowany code point)**
  - Unicode znamješko: jednozmyslne mjeno a code point
  - njedefinuje layout/font/glyph
- **n.př.: LATIN SMALL LETTER E WITH CARON (U+011B)**
  - => **ě**
  - Block Latin Extended-A
  - CategoryLetter, Lowercase
  - Decomposition LATIN SMALL LETTER E (U+0065) COMBINING CARON (U+030C)
  - Upper case U+011A
- **UTF-8 je specielna koděrowanska forma (1-4 bytow)**

  UTF-8 (hex) 0xC4 0x9B hdyž program ě wopak čita a pokaza: Ä›

  UTF-16 (hex) 0x011B

Unicode

LATIN CAPITAL LETTER A WITH GRAVE



Žórło: https://www.unicodeconference.org/presentations-42/TS1T2-Cummings-McKenna-Texin.pdf

“LATEINISCHE ZEICHEN IN UNICODE“ (Datentyp String.Latin)

- **za zjawne zarjadnišća Němskeje**
- **wšě łaćonske znamješka z Unicoda => tak mjenowany String Latin**
- **Koordinierungsstelle für IT-Standards (2012) (zwjazk + kraje):**

  „Daher besteht der Bedarf, den Umfang von Unicode auf ein für die öffentliche Verwaltung Deutschlands geeignetes Maß einzuschränken. Das ‘geeignete Maß‘ ist das Lateinische Alphabet, denn dieses wird im Verwaltungshandeln und der Registerführung in Deutschland regelhaft zu Grunde gelegt. [...]

  Zusammengefasst ergibt sich aus den oben stehenden Ausführungen folgender Grundsatz für die Registerführung der öffentlichen Verwaltung Deutschlands:

  Daten sind in lateinischer Schrift zu erfassen, diakritische Zeichen sind unverändert wiederzugeben. Dabei ist der Zeichensatz nach ISO-/IEC 10646 (der Unicode Standard) zu Grunde zu legen.“

https://www.xoev.de/die_standards/lateinische_zeichen_in_unicode-4813

Unicode Common Locale Data Repository

- **CLDR: zhromadna datowa banka za ‘lokalne daty‘**
- **Što su ‘lokalne daty‘ ?**
  - sep datow za definowanje wosebitych kriterijow jedneje rěče-regiona
  - wopisanje rěčnych a lokalnych wosebitosćow (nastupajo wuzwoleneje rěče na ličaku abo mobilnym nastroju)
  - daty kotrež wopisuja zadźerženje hardware abo software (na wužiwarja) we wotwisnosći wužiwaneje rěče a regiony, wobkedźbujo kulturne wosebitosće
  - n. př. format datuma a wjele wjac

Unicode CLDR definuje za rěče:

- formaty a mustry za datum a čas
- formaty a mustry za ličby a za měny
- jednoty za měry
  - n. př. Jedna ameriska měra za wolumen: cup - šalka
- kolacija: prawidła za sortěrowanje, pytanje, runanje (matching)
- pomjenowanja za rěče, teritorije, pisma, časowe pasma, měny
- emoji znamješka (kode a definowane pomjenowanja)

Unicode Common Locale Data Repository

- **je wosebity projekt Unicode-konsorcija**
  - Unicode standard sam njedefinuje lokalne daty
- **CLDR je Open-Source**
- **daty w XML (LDML) a JSON formaće**
- **jasny proces a tooly za zapisanje datow**
  - wone so zběra, so hłosuje, so rozrisa a publikuje
- **wulka community podawarjow a wužiwarjow**

Unicode CLDR – Hłowni wužiwarjo

- Apple (macOS, iOS, watchOS, tvOS; Apple Mobile Device Support, iTunes for Windows)
- Google (Web Search, Chrome, Android, Adwords, Google Maps, Blogger, Google Analytics)
- Microsoft (Windows, Office, Visual Studio)
- Wikimedia Foundation (Wikipedia)

Unicode CLDR definuje Emoji

	1F606	***grinning squinting face** | face | laugh | mouth | satisfied | smile	***straoiseog ag gáire le súile dúnta** | béal ar oscailt | meangadh gáire | súile dúnta go dlúth	***aodann le gàire, beul fosgailte ⁊ sùilean dùinte** | aodann | beul | fiamh-ghàire | fosgailte | gàire | sàsaichte
	1F605	***grinning face with sweat** | cold | face | open | smile | sweat	***straoiseog ag gáire le fuarallas** | béal ar oscailt | fuarallas | meangadh gáire	***aodann le gàire a' cur fallas** | aodann | fallas | fiamh-ghàire | fosgailte | fuar
	1F923	***rolling on the floor laughing** | face | floor | laugh | rolling	***sna trithí gáire** | aghaidh | gáire | sna tríthí | urlár	***a' ruidhleadh air an làr a' gàireachdainn** | aodann | gàire | gàireachdainn | làr | roladh | ruidhleadh

Žórło: https://www.unicodeconference.org/presentations-42/S1T2-Loomis-Scherer.pdf

https://unicode.org/cldr/charts/latest/annotations/slavic.html

Unicode CLDR – serbskej rěči

- **wobej serbskej rěči buchu zapisani w lěće 2015, do wersije CLDR 27**

  (W. Bejmak, rěčespytna podpěra a terminologija S. Wölkowa za hsb a F. Kaulfürst za dsb)

- **ca. 10 000 polow/zapiskow**
- **zapisanje přez Survey Tool**

  https://st.unicode.org/cldr-apps/v#/hsb//

- **přehlad 2015:**

  http://www.unicode.org/cldr/charts/27/summary/hsb.html

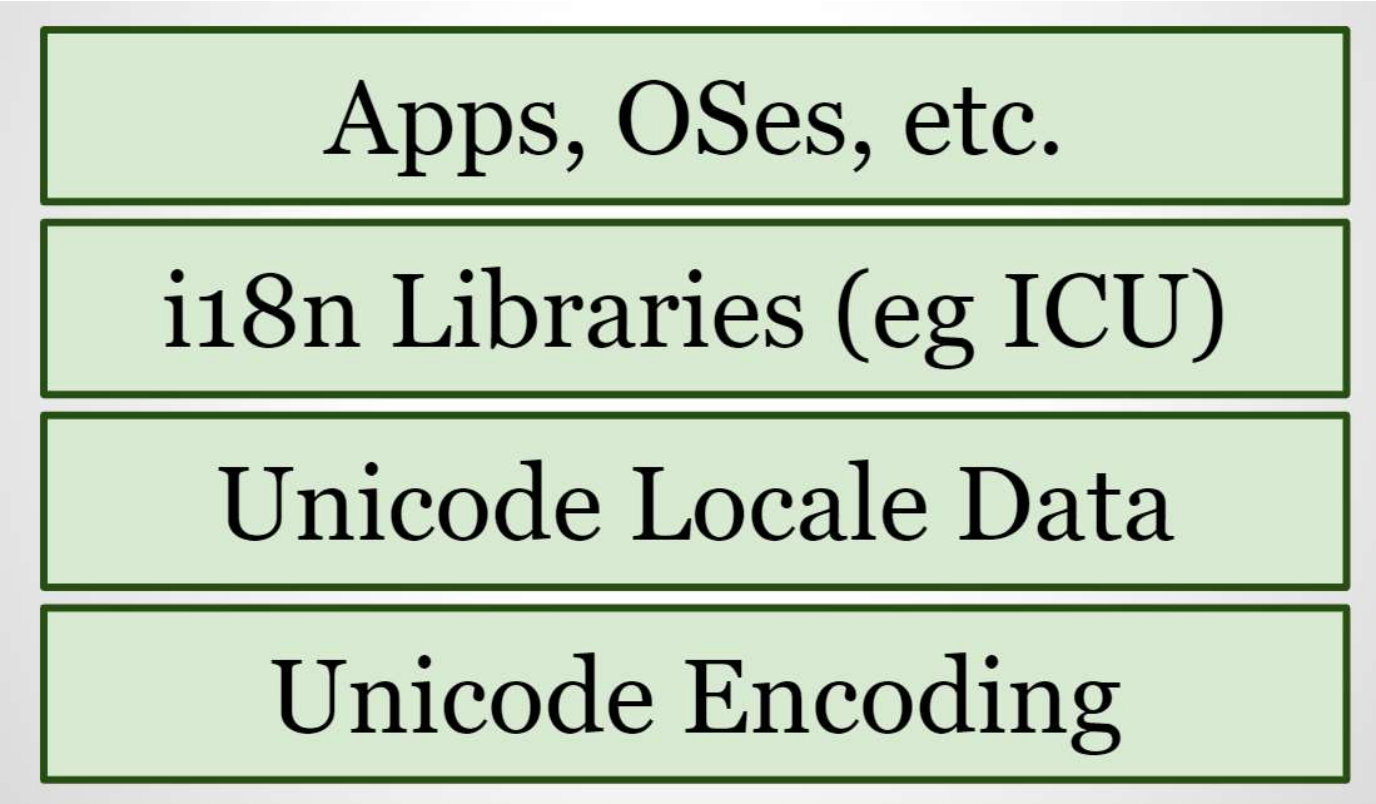
Unicode CLDR - přikład

Cup				
long-displayName	cups		✔	šalki ☆
long-one	{0} cup	ⓔ	✔	{0} šalka ☆
long-two	{0} c	ⓔ	✔	{0} šalce ☆
long-few	{0} c	ⓔ	✔	{0} šalki ☆
long-other	{0} cups	ⓔ	✔	{0} šalkow ☆
short-displayName	cups		✔	š. ☆

https://st.unicode.org/cldr-apps/v#/hsb/Volume/

https://github.com/unicode-org/cldr/blob/master/common/collation/hsb.xml

ICU (International Components for Unicode)



Žórło: http://cldr.unicode.org/index

ICU (International Components for Unicode)

- programowa biblioteka (software) w Java a C
- běži na 2 miliardach nastrojach
- so wužije w Apple ios OSX, Google Android, Microsoft Windows
- serbšćina z lěta 2015 wot wersije ICU 55 zapřijata

  (a z tym serbšćina na milionach nastrojach) 

- http://demo.icu-project.org/icu-bin/icudemos

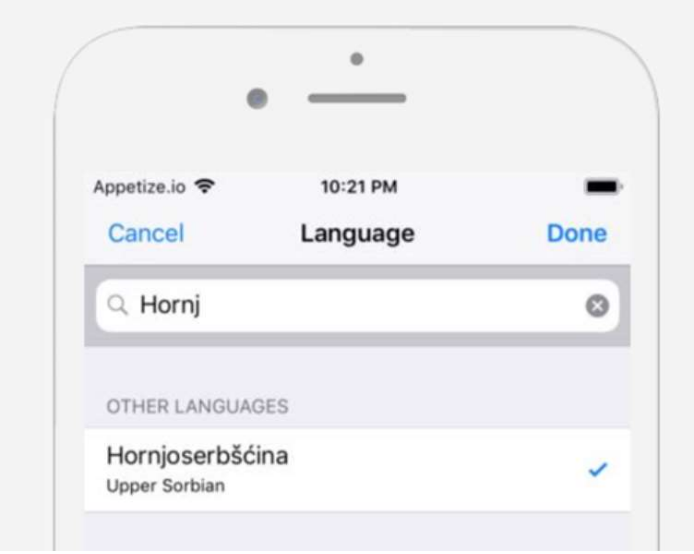
Unicode CLDR Locale Coverage Level – rozdźělne wobjimy

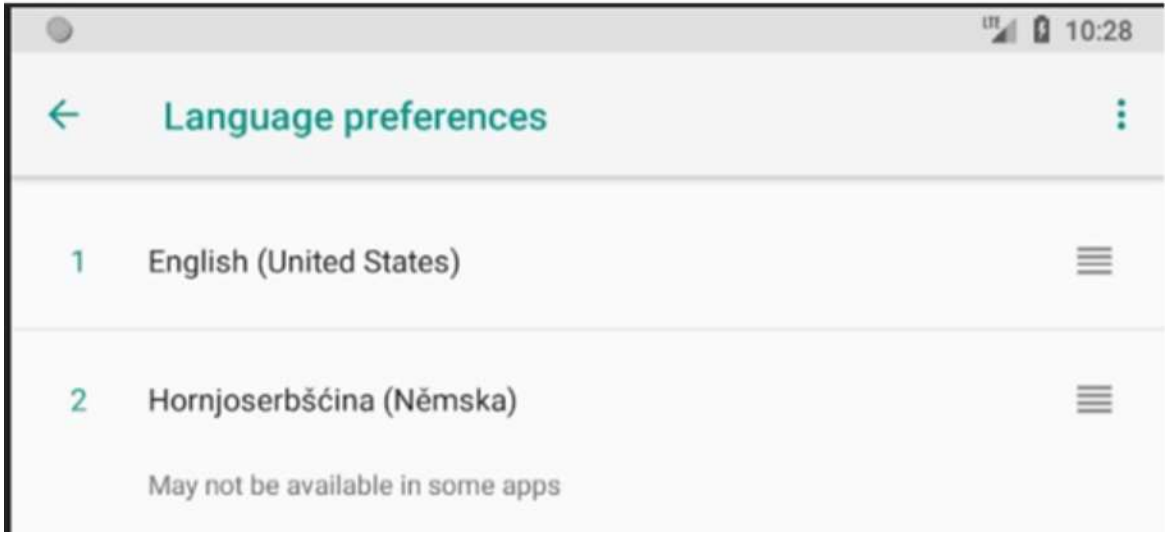
- měritko za kwalitu datow – firmy to wuhódnoća
- Modern / Moderate / Basic / Core
- dsb/hsb w lěće 2015 (CLDR 27):
  - 99 % Modern hewak wšitko 100%
  - https://www.unicode.org/cldr/charts/27/supplemental/locale_coverage.html#ccp
- dsb a hsb w lěće 2019 (CLDR 36):
  - dsb/hsb: 48.8% (abs. 3,997) 84.1% 96.8% 100.0%

    (Smy dale pódla?)

Serbšćina w Unicode CLDR – Što je wunošk?

- **Serbska rěč z tym w Android a iOS !**
  - ale pola Android wot firmow wotwisne





iPhone 8 ios 12.2

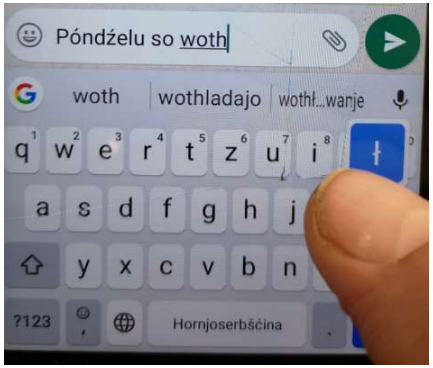
Nexus 7 Android 8.1

Standardizacija – Terminologije(koděrowane)

- za Mozilla projekty (Firefox, Thunderbird)
  - Hlej http://sorbzilla.de
- za Wordpress
  - http://hsb.wordpress.org
- za LibreOffice
  - https://hsb.libreoffice.org/
- za Wikipediju
- ...

Gboard – tastatura a słowne wudospołnjenja

- **daty z hsb-wikipedije, CLDR za „widźomnosć“ serbšćiny**



„How Gboard is helping European languages in the digital age [...]

Beyond the 24 official languages of the European Union, Gboard supports many other languages, like Welsh, Corsican, Luxembourgish, Sicilian, Scottish Gaelic, Upper Sorbian, Northern Sami, Manx, and more [...]“ (26.9.2018)

https://www.blog.google/around-the-globe/google-europe/how-gboard-helping-european-languages-digital-age/

Wuhlad – standardizacija – Unicode CLDR

- **trajny přewod wostanje trěbny**
- **přez zapisanje w Unicode CLDR je serbšćina daloko widźomna**
- **z tym je baza do přichoda położena, za podpěru a wužiwanje serbšćiny też na nowym polu kumštneje inteligency a Deep Learning**
  - móže zajim pola třećich zbudźić serbšćinu w nowych wužiwanjach podpěrać
- **Je struktura Unicode CLDR projekta jedyn muster za nas?**
  - rozdźělni sponsorojo
  - community a fachowcy
  - sobudźěło a kwalitu zaručacy workflow
  - swobodnje wužiwajomne daty

Wutrobny dźak!