Unicode CLDR und Standardisierung

für die Sichtbarkeit der sorbischen Sprache in der digitalen Welt

Wito Böhmak

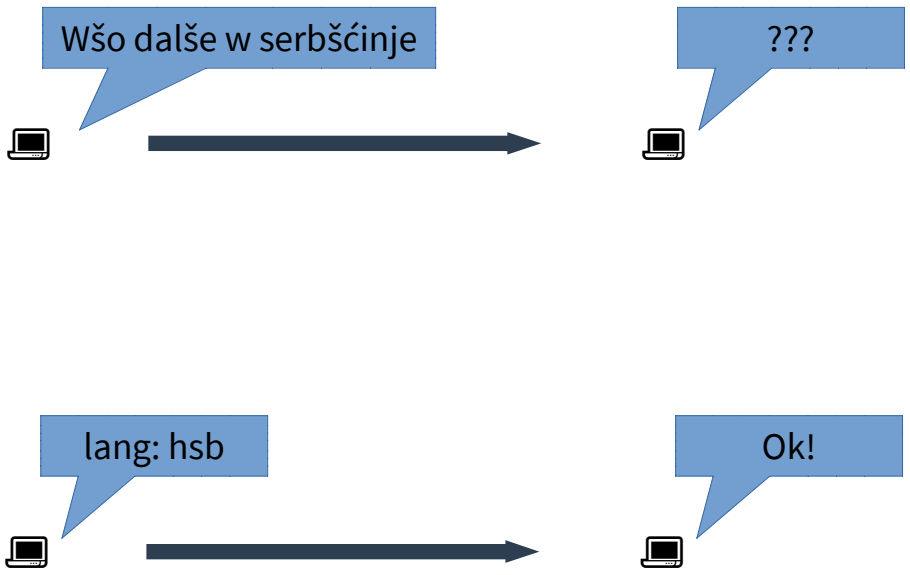
Konferenz zur Digitalstrategie

Bautzen, 30.11.2019

Was bedeutet Standardisierung in der digitalen Welt?

- **International anerkannte Regelungen**
  - durch internationale Normierungsgremien (z.B. ISO)
- **gemeinsame Festlegungen, Regularien und Terminologien**
- **ist unverzichtbar für Interoperabilität, Kompatibilität und Nachhaltigkeit:**
  - Die Fähigkeit unterschiedlicher technischer Systeme zusammenzuarbeiten, sich zu verstehen und Daten austauschen zu können, und das über einen längeren Zeitraum

Maschinen können sich einigen:



Beispiele für Standards

- ISO 639, BCP 47
- Unicode (oder ISO 10646)
- Unicode Locale Data => Unicode CLDR
- Breiterer Kontext – Terminologien (kodiert)

ISO 639 – Codes für Sprachen

- Internationale Norm zur Definition von Sprachcodes
- Wie: Code + Benennung der Sprache
- Für die sorbische Sprache besonders ISO 639-2 (sogenannter 3-alpha-code)
- Entstehung von 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

‘Sorbische Geschichte‘ in ISO 639

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

- **Sorbische Bücher in amerikanischen Universitätsbibliotheken**
  - Archibald Gary Coolidge (amerikanischer Historiker und Diplomat in Wien) schenkte 1895 der Harvard Universität gebundene Ausgaben der Zeitschrift Časopisy Maćicy Serbskeje
  - Ende der 1960-er Jahre begann die Katalogisierung in Computersysteme: Sprachcode: **wen** (USMARC code list for languages)
- **wen: 1989 Aufnahme in die neue ISO 639-2 (übernommen aus der USMARC code list)**
- **dsb und hsb: 2003 Aufnahme in ISO 639-2**
  - Unterscheidung von Ober- und Niedersorbisch als Schlagwort schon 1986 in der LoC
- **wen: als Sprachfamilie zusätzliche Aufnahme in 639-5**
- **dsb und hsb: auch in die ISO 639-3 (Ethnologische Liste)**

‘Sorbische Geschichte‘ in 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

Endgültiger Stand in 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‘ ist ein Code zur Identifikation natürlicher Sprachen
- Basiert auf anderen Standards
- Kombiniert Sprachcode (ISO 639) und Land (ISO 3166-1) und weitere (ISO 15924 Schriftsystem)

  en-us en-gb

  hsb hsb-de

  dsb dsb-de

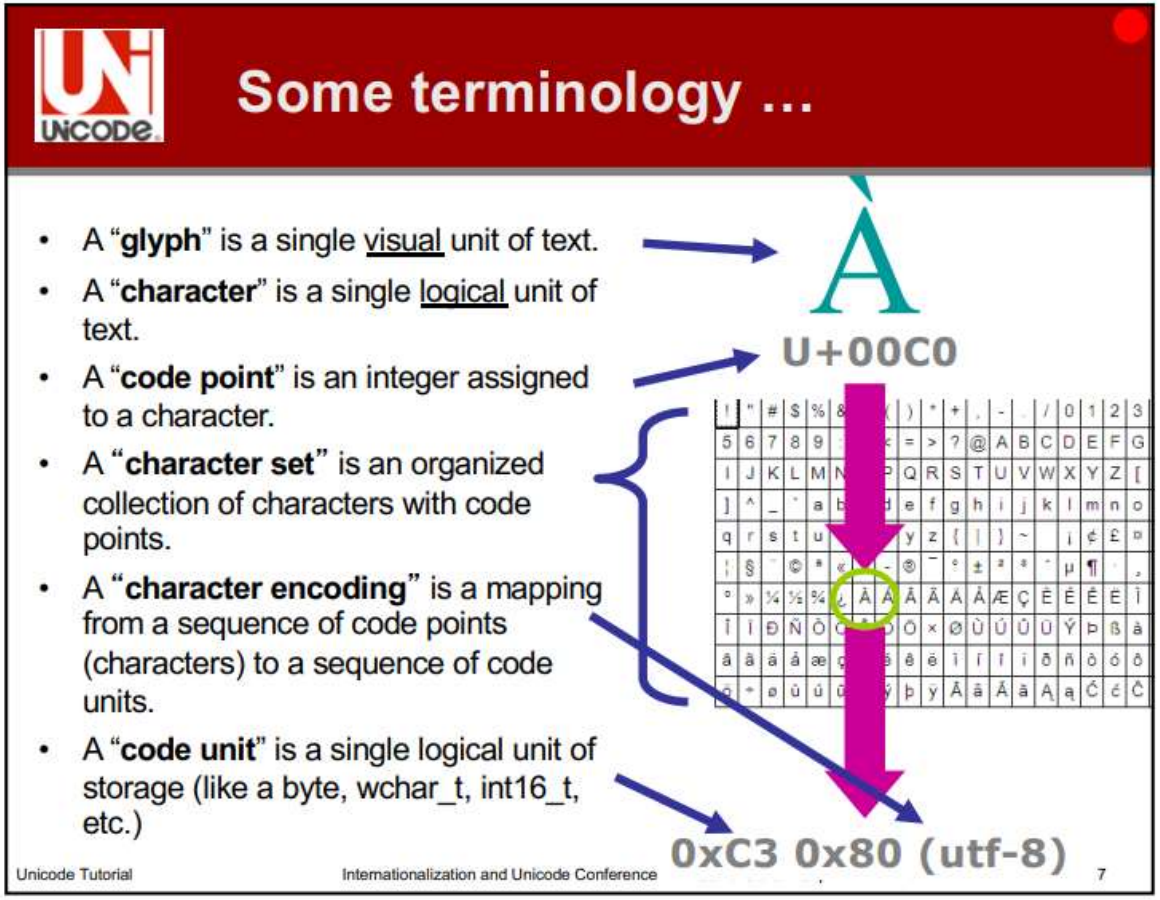
Unicode-Zeichensatz

- **definiert Schriftzeichen mit zugeordneter fester Zahl (sogenannter Code Point)**
  - Unicode Zeichen: eindeutiger Name und Code Point
  - Definiert nicht Layout/Font/Glyph
- **z.B.: 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 ist eine spezielle Kodierungsform (1-4 Byte)**
  - *UTF-8 (hex) 0xC4 0x9B wenn ein Programm das ě falsch interpretiert und anzeigt: Ä›*

    UTF-16 (hex) 0x011B

Unicode

LATIN CAPITAL LETTER A WITH GRAVE



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

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

- **Für die öffentliche Verwaltung der Bundesrepublik**
- **Alle lateinischen Buchstaben des Unicode => sogenannter String Latin**
- **Koordinierungsstelle für IT-Standards (2012) (Bund + Länder):**

  „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: gemeinsame Datenbank für Lokalisierungs-Daten**
- **Was sind Lokalisierungs-Daten?**
  - Datensatz zur Definition besonderer Kriterien einer Sprache-Region
  - Beschreibung sprachlicher und lokaler Besonderheiten (bzgl. der ausgewählten Sprache auf dem Rechner oder mobilen Gerät)
  - Daten die das Verhalten der Hardware oder Software (auf den Nutzer) in Abhängigkeit der ausgewählten Sprache und Region beschreiben, die kulturellen Besonderheiten beachtend
  - z. B. Datumsformat und vieles mehr

Unicode CLDR definiert für Sprachen:

- Formate und Muster für Datum und Zeit
- Formate und Muster für Zahlen und Währungen
- Maßeinheiten
  - z.B. ein amerikanisches Maß für Volumen: cup - šalka
- Kollation: Sortier-, Such und Matching-Regeln
- Benennungen von Sprachen, Territorien, Schriftsystemen, Zeitzonen, Währungen
- Emoji Zeichen (Code und definierte Benennungen)

Unicode Common Locale Data Repository

- **Ist ein spezielles Projekt des Unicode Konsortiums**
  - Unicode Standard selbst definiert keine Lokalisierungs-Daten
- **CLDR ist Open-Source**
- **Daten in XML (LDML) und JSON Format**
- **Klarer Prozess und Tools für die Datenerfassung**
  - Es wird gesammelt, abgestimmt, Probleme gelöst und publiziert
- **Große Community von Mitwirkenden und Nutzern**

Unicode CLDR – Hauptnutzer

- 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 definiert Emojis

	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

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

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

Unicode CLDR – die sorbischen Sprachen

- **beide sorbischen Sprachen wurden 2015, in der Version CLDR 27 erfasst**

  (W. Böhmak, sprachwiss. Unterstützung und Terminologie durch S. Wölke für hsb und F. Kaulfürst für dsb)

- **ca. 10 000 Feldeinträge**
- **Erfassung mittels des Survey Tools**

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

- **Überblick 2015:**

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

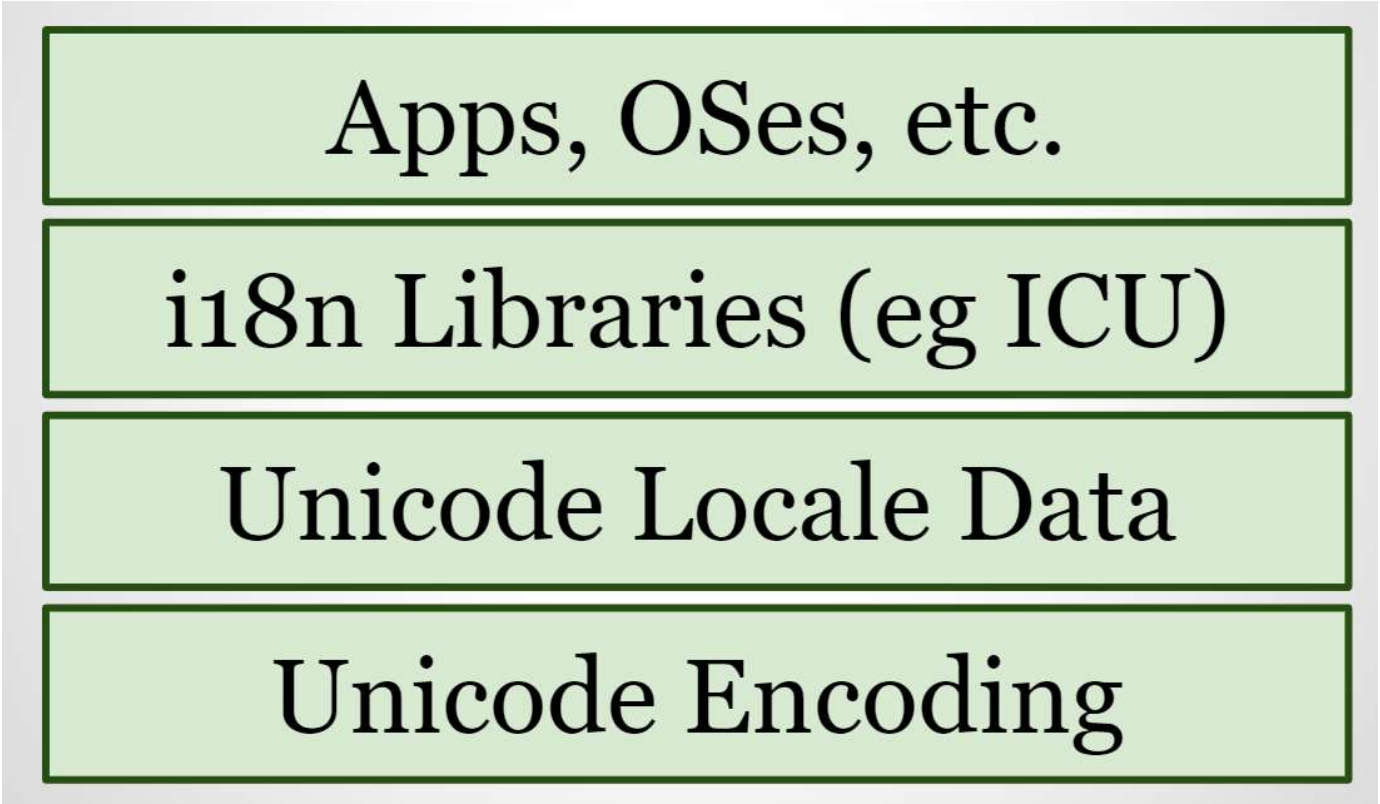
Unicode CLDR – ein Beispiel

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)



Quelle: http://cldr.unicode.org/index

ICU (International Components for Unicode)

- Programmbibliothek (Software) in Java und C
- Läuft auf 2 Milliarden Geräten
- Verwendet in Apple iOS OSX, Google Android, Microsoft Windows
- Sorbisch ist integriert seit 2015 ab Version ICU 55

  (und damit ist Sorbisch millionenfach verteilt) 

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

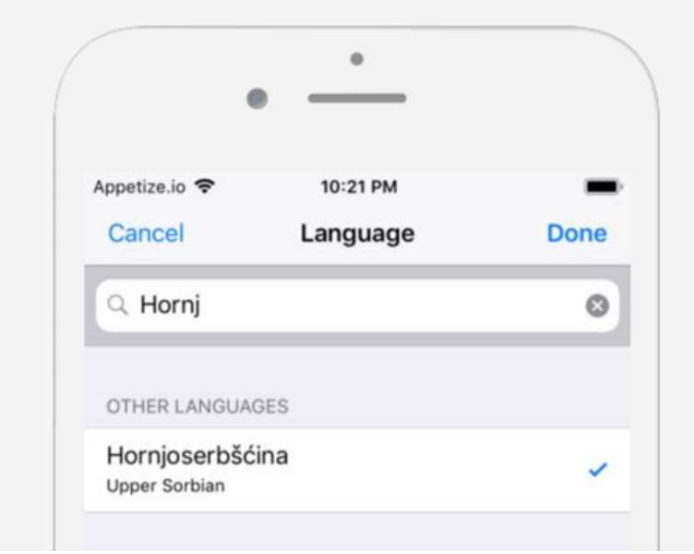
Unicode CLDR Locale Coverage Level – unterschiedliche Umfänge

- **Maß für Datenqualität – Hersteller werten das aus**
- **Modern / Moderate / Basic / Core**
- **dsb/hsb im Jahr 2015 (CLDR 27):**
  - 99 % Modern ansonsten 100%
  - https://www.unicode.org/cldr/charts/27/supplemental/locale_coverage.html#ccp
- **dsb und hsb im Jahr 2019 (CLDR 36):**
  - dsb/hsb: 48.8% (abs. 3,997) 84.1% 96.8% 100.0%

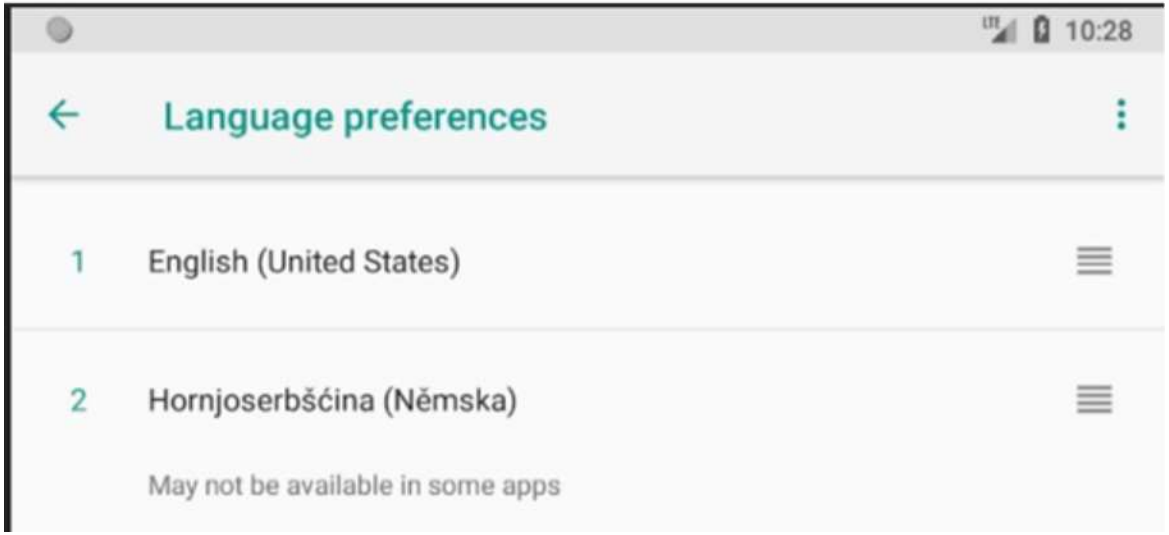
    (Sind wir weiter dabei?)

Sorbisch im Unicode CLDR – Was hat´s gebracht?

- **Sorbische Sprache dadurch im Android und iOS!**
  - Aber bei Android von den Herstellern abhängig



iPhone 8 ios 12.2



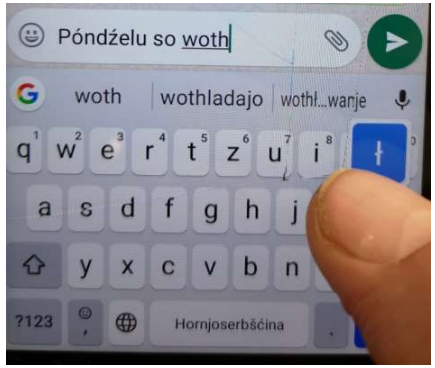
Nexus 7 Android 8.1

Standardisierung – Terminologien(kodiert)

- **Mozilla Projekte (Firefox, Thunderbird)**
  - siehe http://sorbzilla.de
- **für Wordpress**
  - http://hsb.wordpress.org
- **für LibreOffice**
  - https://hsb.libreoffice.org/
- **für Wikipedia**
- **...**

Gboard – Tastatur und Wort-Vervollständigung

- **Daten aus der hsb-Wikipedia, CLDR für „Sichtbarkeit“ von Sorbisch**



„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/

Ausblick – Standardisierung – Unicode CLDR

- **Dauerhafte Mitarbeit ist notwendig**
- **Durch Erfassung in der Unicode CLDR Datenbank ist Sorbisch weithin sichtbar**
- **Damit ist Basis gelegt für die Unterstützung und Nutzung von Sorbisch auch auf den neuen Feldern der Künstlichen Intelligenz und des Deep Learnings**
  - Damit kann das Interesse bei Dritten geweckt werden Sorbisch in neuen Anwendungen zu unterstützen
- **Ist die Struktur des Unicode CLDR Projekts ein Muster für uns?**
  - Verschiedene Sponsoren
  - Community und Fachleute
  - Mitarbeit und Qualität sichernder Workflow
  - Frei zugängliche Daten

Wutrobny dźak!