

String.Latin und die Identifikation von Personen

eine Standortbestimmung

9. XÖV-Konferenz
15./16. September 2016, Bremen
WORKSHOP 3

Bernd Kappenberg

bernd.kappenberg@gmx.de

Definitionen

- **Zeichensatz** (character set, character repertoire): Ein Vorrat an Zeichen (Buchstaben, Zahlen, Zeichensetzung, Diakritika...).
- Davon zu unterscheiden ist die **Kodierung**, bei der die Zeichen jeweils durch einen numerischen Wert codiert sind und sich daher für die digitale Speicherung und Übertragung von Daten eignen.
- **Schriftart** (typeface): Grafisches Design eines Zeichensatzes. Es ordnet die abstrakten numerischen Werte bestimmten Formen (Glyphen) und einer oder mehreren **Schriftgrößen** (fonts) zu.

Beispiel

ISO/IEC 10646 / Unicode: 8 bis 32 Bit

UTF-8 ist die sparsamste Kodierung für lateinische Zeichen.

(Es gibt auch UTF-16 / UCS-2, UTF-32 / UCS-4.)

ISO-8859-1

(Latin-1)

8 Bit



ISO-8859-15

(Latin-9)

8 Bit

ASCII:

7 Bit

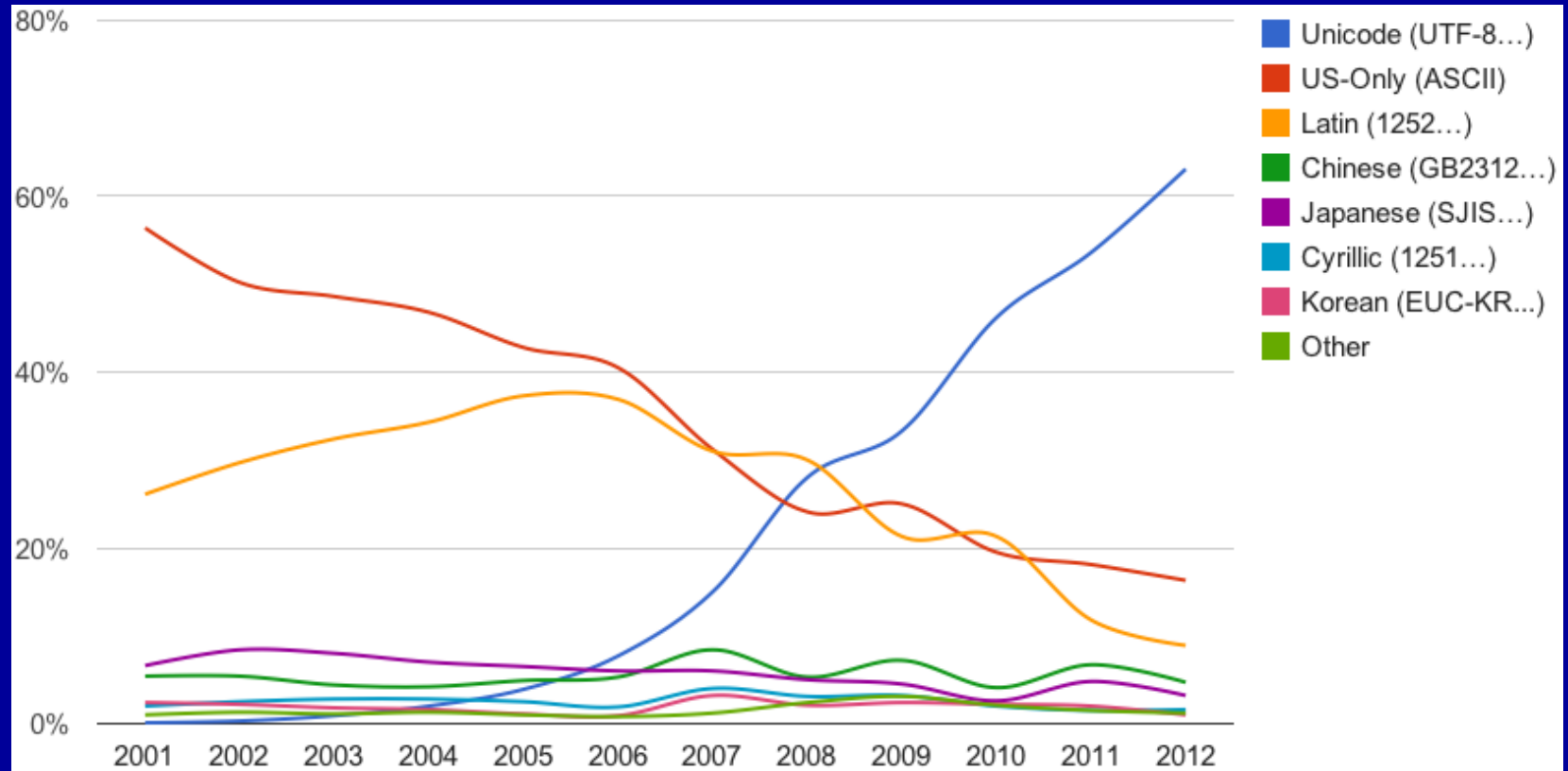
Kritischer Faktor: Zeichensatz

- Benötigt für String.Latin: 481 Zeichen (nächste Version \approx 600)
- Kodierung basiert auf Fernschreiber, nicht EDV!

Kodierung:

- | | | | |
|--------------------|-------|-------------|----------|
| • 1870 Baudot-Code | 5 Bit | 58 Zeichen* | zu wenig |
| • 1963 ASCII | 7 Bit | 128 Zeichen | zu wenig |
| • 1980er (diverse) | 8 Bit | 256 Zeichen | zu wenig |
-
- | | | | |
|----------------|--------|-----------------------|-----|
| • 1991 Unicode | 16 Bit | 128,172 Zeichen (9.0) | OK? |
|----------------|--------|-----------------------|-----|

Unicode im Web



European Scripts	African Scripts	South Asian Scripts	Indo-European Scripts
Armenian	Adlam	Ahom	Balinese
Armenian Ligatures	Bamum	Bengali and Assamese	Batak
Caucasian Albanian	Bamum Supplement	Bhaiksuki	Buginese
Cypriot Syllabary	Bassa Vah	Brahmi	Buhaya
Cyrillic	Coptic	Chakma	Hanunoo
Cyrillic Supplement	Coptic in Greek block	Devanagari	Javanese
Cyrillic Extended-A	Coptic Epact Numbers	Devanagari Extended	Rejang
Cyrillic Extended-B	Egyptian Hieroglyphs (1MB)	Grantha	Sundanese
Cyrillic Extended-C	Ethiopic	Gujarati	Sundanese Supplement
Elbasan	Ethiopic Supplement	Gurmukhi	Tagalog
Georgian	Ethiopic Extended	Kaithi	Tagbanwa
Georgian Supplement	Ethiopic Extended-A	Kannada	East Asian
Glagolitic	Mende Kikakui	Kharoshthi	Bopomofo
Glagolitic Supplement	Meroitic	Khojki	Bopomofo Extended
Gothic	Meroitic Cursive	Khudawadi	CJK
Greek	Meroitic Hieroglyphs	Lepcha	CJK Extension A
Greek Extended	N'Ko	Limbu	CJK Extension B
Ancient Greek Numbers	Osmanya	Mahajani	CJK Extension C
Latin	Tifinagh	Malayalam	CJK Extension D
Basic Latin (ASCII)	Vai	Meetei Mayek	CJK Extension E
Latin-1 Supplement	Middle Eastern Scripts	Meetei Mayek Extensions	(see Middle Eastern Scripts)
Latin Extended-A	Anatolian Hieroglyphs	Modi	CJK
Latin Extended-B	Arabic	Mro	CJK Extension F
Latin Extended-C	Arabic Supplement	Multani	CJK
Latin Extended-D	Arabic Extended-A	Newa	CJK Extension G
Latin Extended-E	Arabic Presentation Forms-A	Oi Chiki	CJK Extension H
Latin Extended Additional	Arabic Presentation Forms-B	Oriya (Odia)	Idiosyncratic
Latin Ligatures	Aramaic, Imperial	Saurashtra	Hangul
Fullwidth Latin Letters	Avestan	Sharada	Hangul Supplement
IPA Extensions	Carian	Siddham	Hangul Extended
Phonetic Extensions	Cuneiform (1MB)	Sinhala	Hangul Extended Supplement
Phonetic Extensions Supplement	Cuneiform Numbers and Punctuation	Sinhala Archaic Numbers	Hangul Extended Supplement 2
Limban		Sans-Samran	

Kritischer Faktor: Zeichensatz

- Benötigt für String.Latin: 481 Zeichen (nächste Version \approx 600)
- Kodierung basiert auf Fernschreiber, nicht EDV!

Kodierung:

- | | | | |
|--------------------|-------|-------------|----------|
| • 1870 Baudot-Code | 5 Bit | 58 Zeichen* | zu wenig |
| • 1963 ASCII | 7 Bit | 128 Zeichen | zu wenig |
| • 1980er (diverse) | 8 Bit | 256 Zeichen | zu wenig |

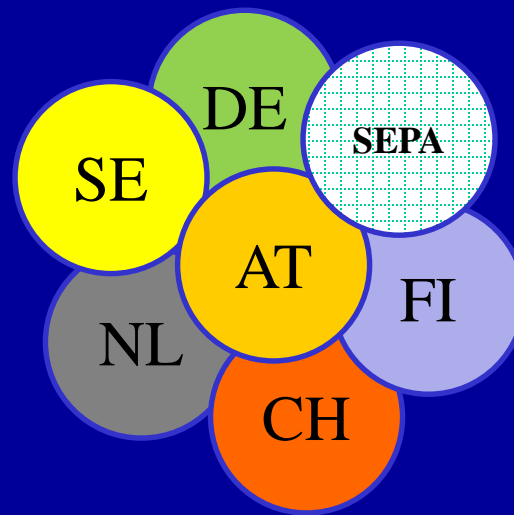
Lücke! = **Unicode**, aber reduziert auf lateinische Zeichen

- | | | | |
|----------------|--------|-----------------------|----------|
| • 1991 Unicode | 16 Bit | 128,172 Zeichen (9.0) | zu viel! |
|----------------|--------|-----------------------|----------|



Kritischer Faktor: Zeichensatz

- Die „Latein-Lücke“ wird bisher gefüllt mit nicht vollständig kompatiblen, meist nationalen Eigenkreationen (immer Unicode-basiert):



Kritischer Faktor: Zeichensatz

Status der Latein-Lückenfüller (1):

- NL (Beschluss Standardschreibweise Personendaten): rechtlich verpflichtend für die allgemeine Verwaltung (1993)
- SE (e-Namen): Richtlinie für die elektronische Verwaltung; bisher *nicht* bindend (2006)
- AT (Handbuch diakritische Zeichen): Richtlinie; *de facto* verpflichtend für Pässe und Personalausweise (Personenstandsgesetz; Meldegesetz nur durch Normverweis) (2006)
- DE (String.Latin): rechtlich verpflichtend für Pässe, Personalausweise, Meldedaten... (2012)

Kritischer Faktor: Zeichensatz

Status der Latein-Lückenfüller (2):

- FI: ...?
- CH (eCH-0011 Datenstandard Personendaten; eCH-0018 XML Best Practices): Unicode (2014) bei Polizei, Diplomaten, aber Zivilstand nur Latin-9, Migration: Latin-1
- UK: nur Aa-Zz (ICAO 9303)

Kritischer Faktor: Zeichensatz

Status der Latein-Lückenfüller (3):

- SEPA-Zeichensatz (alle Diakritika der Länder aus dem Euro-Zahlungsverkehrsraum): nur Vorschlag; bis jetzt ist SEPA (ISO 20022) zwar Unicode, aber *de facto* eingeschränkt auf Aa-Zz.
- Deutsche Kreditwirtschaft ist für Erweiterung!
- Bedenkenträger #1:
 - „Unbekannte Entwicklungskosten für interne Systeme“ (immer noch 7-bit ASCII & EBCDIC??)

12. JANUAR 2011

NEUER PERSONALAUSWEIS

Gute Zeichen, schlechte Zeichen

Von SEBASTIAN MICHAEL BRAUNS



Bringt das System zum Absturz: Der Ausweis von Bundesinnenminister Thomas de Maizière.

Foto: dpa

- Bedenkenträger #2:
 - „Transaktions-Screening (Geldwäsche, Terrorfinanzierung)?“ – kann einfach gelöst werden durch *interne* Repräsentation ohne Diakritika (z.B. gemäß ICAO 9303):



Vinxxas

P<ESPVINXXAS<VINXXAS<<MARIA<MERCEDES<<-

Kritischer Faktor: Zeichensatz

- Frage: Kann man daraus einen einheitlichen „EU-Zeichensatz“ auf Unicode-Basis machen?



EU?

Ist eine EU-Vorgabe möglich?

- Vorteil: in allen 28 Mitgliedstaaten einheitlich
- Zuständigkeit der Union gegeben?
 - AEU-Vertrag Art. 18: Verbot von Diskriminierung aus Gründen der Staatsangehörigkeit
 - Grundrechtecharta
 - Art. 4.2 (h), 170, 171: Transeuropäische Netze
- **DIN/CEN-Vorschlag:** “Define the summary repertoire for use in name writing in European public registers, especially in the light of current and potential future legal requirements.” (2011 abgelehnt)

ISA² (Interoperabilitätsstandards für die Administration)

- Vorschlag (Call for proposals): 26/06/2017 bis 28/08/2017
 - Andere EU-Länder mit ins Boot nehmen!
 - Muss mehr politisch sein als 2011! (Stichwort Flüchtlinge, EESSI)
 - Einheitlicher lateinischer Zeichensatz plus Suchalgorithmen, schönere Schriftart, automatische Transkription...
 - Evaluation (Work programme preparation): 09/2017 bis 03/2018
 - Annahme (Europäische Kommission): 03/2018
 - Durchführung (Work programme implementation): 04/2018 bis 12/2020
- http://ec.europa.eu/isa/isa2/index_en.htm

Vielen Dank für Ihre
Aufmerksamkeit



Setting Signs for Europe – Why Diacritics Matter for European Integration (ibidem 2015: ISBN 978-3-8382-0663-9)

www.sonderzeichen.com



Zeichen setzen für Europa – Der Gebrauch europäischer lateinischer Sonderzeichen in der deutschen Öffentlichkeit (ibidem 2012: ISBN 978-3-89821-749-1)



PS: Was ich gern von Ihnen wüsste: **Wie suchen Sie nach Namen?**

- Suchalgorithmen
 - englisch: Soundex, Phonix, Metaphone
 - deutsch: Kölner Phonetik, PHONEM, Phonet
 - osteuropäisch: Daitch-Mokotoff
 - international: IPA, SAMPA
 - Hamming-Distanz, N-Gram, Damerau-Levenshtein...
 - Wildcard für Buchstaben mit Diakritika
 - Laufzeitverhalten?
- Identifikationsalgorithmus?