



Citation test..

Authors: Stefan Eilers
Submitted: 23. August 2021
Published: 23. August 2021
Volume: 1
Issue: 3
Affiliation: dddd
Languages: German
Keywords: key word ist das
Categories: Drug Research
DOI: 10.17160/josha.1.3.20

Abstract:

Abstract with no content

JOSHA

josha.org

**Journal of Science,
Humanities and Arts**

JOSHA is a service that helps scholars, researchers, and students discover, use, and build upon a wide range of content

Einführung des einheitlichen Zeichensatzes: Ergebnisse der AG ID □ Algorithmus und ihre Bedeutung

Dr. Fabian Büttner

MSI Unternehmensberatung

Koordinierungsstelle für IT-Standards (KoSIT)

4. XÖV-Anwenderkonferenz

1. Ausgangslage
2. Der Weg zum Ziel – Arbeitsweise
3. Das Ergebnis – Regelungen zur Identifikation
4. Effektivität
5. Ausblick

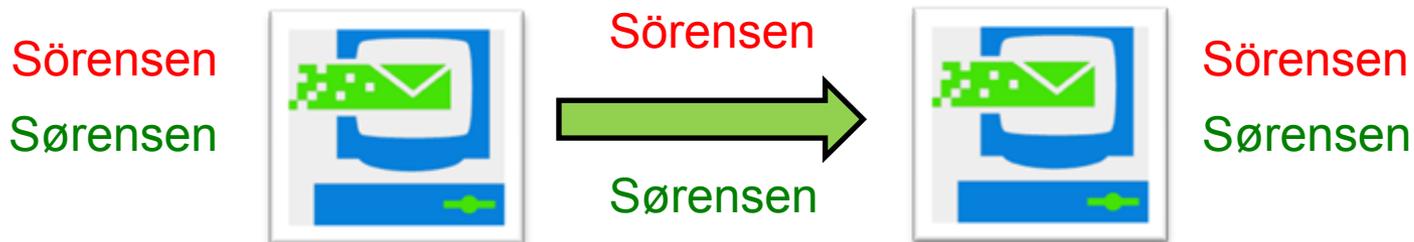
- Fachvorgabe für (elektronisches) Personenstandswesen
 - Bei gleicher Schrift sind Familiennamen und Vornamen buchstabengetreu unverändert wiederzugeben.
 - Die diakritischen Zeichen sind ebenfalls wiederzugeben, selbst wenn die Sprache, in der die Eintragung vorgenommen werden soll, solche Zeichen nicht kennt.
- Umsetzung obiger Vorgabe in Deutschland
 - Beurkundungsdaten sind in lateinischer Schrift zu erfassen, diakritische Zeichen sind unverändert wiederzugeben. Dabei ist der Zeichensatz nach ISO/IEC 10646 (Unicode) zu Grunde zu legen.
- Zusätzliche Anforderung
 - Namen von Personen sollen in den Registern des Melde- und des Personenstandswesen identisch dargestellt werden.

- Der AK I beschließt am 7. 10. 2008 für das Meldewesen:
 - Einführung ISO/IEC 10646 in UTF 8-Kodierung zum 1. 5. 2011
Grundsatzbeschluss zur Umstellung auf Unicode
- Der AK I bittet die KoSIT:
 - die Entwicklung einer Aufstellung des zulässigen Zeichenumfangs
Tabelle der Lateinischen Zeichen in Unicode
 - in Verbindung mit eine einheitlichen Lösung für die Altdaten
Entwicklung eines „Identifikationsalgorithmus“
 - in das Projekt DOL - Standardisierung einzubringen.
Fachunabhängiger Standard:
Zuständigkeit des IT-Planungsrat (ex: KoopA-ADV)

Herausgabe des Standards
Lateinische Zeichen in Unicode (Version 1.1.0)
zum 30. 9. 2011 durch die KoSIT im Auftrag des IT-PLR

- Ausgangslage: Es wird für einen längeren Zeitraum ein Nebeneinander von „alter und „neuer“ Schreibweise geben:
 1. Nicht alle Verfahren im Verbund werden umgestellt.
 2. Bestandsdaten werden in der Regel nicht korrigiert.
- Problematik: Bei Datenübermittlungen sollen alle Datensätze (weiterhin) eindeutig zugeordnet werden können.





Sender		Empfänger	
String.Latin-fähig?	Bestandsdaten in transliterierter Form?	String.Latin-fähig?	Bestandsdaten in transliterierter Form?
●		●	
●	●	●	
●		●	●
	(●)	●	
●			(●)

ICAO 9303
([...] Machine-readable
Official Travel Documents)

Die Beteiligten hatten ähnliche Anforderungen:

- ⇒ eine klare und praxisnahe Regelung zur Identifikation
- ⇒ keine Vorgabe eines Suchalgorithmus
- ⇒ Klare Vorgaben zur Umschlüsselung

DIN 5007-2
Schriftzeichen
Sortierung

Die betrachteten Normen stimmen bzgl. der Transliteration von diakritischen Zeichen, Ligaturen und Grundbuchstaben außer A – Z weitestgehend überein.

DIN 31638
Ordnungsregeln

Für die „Abweichler“ (Ä Ö Ü Å Ö Ü Ø ð 3)
Vorzug für ICAO und DIN (und Vorgaben der PG Standard)

EN 13710
(European Character Folding)

⇒ dies entspricht dem Handeln in der Praxis (in D.)

Unicode TR #39
(Unicode Character Folding)

⇒ und dem „Delta for German“ der EN

PG Standard

„Identifikationsalgorithmus“

1. Definition der **minimalen Treffermenge**, die ein Identifikationsverfahren liefern muss.
2. Tabelle zur **Umschlüsselung** von String.Latin in andere Datensätze
 - Zwei Zeichenketten gelten bzgl. der Identifikation als gleich, wenn ihre **Suchform** gleich ist (Mindestanforderung)
 - Bildung der Suchform:
 1. Ligaturen werden in die einzelnen Zeichen aufgelöst
 2. diakritische Zeichen werden auf ihre Basiszeichen zurückgeführt (Sonderbehandlung für einige Zeichen)
 3. Zeichen, die nicht A – Z sind, werden auf A – Z umgesetzt
 4. es wird in Großschreibung gewandelt
 - Die Definition der Umschlüsselung erfolgt analog der Schritte 1 – 3

	Datensatz (empfangen)	Datensatz 2 (im Register)
	Noel Schmidt-Strauß	Noël Schmidt-Strauß
1. Ligaturen auflösen	Noel Schmidt-Strau ss	Noël Schmidt-Strau ss
2. Diakritika entfernen	Noel Schmidt-Strauss	Noel Schmidt-Strauss
3. Basiszeichen ersetzen	Noel Schmidt-Strauss	Noel Schmidt-Strauss
4. Großschreibung	NOEL SCHMIDT- STRAUSS	NOEL SCHMIDT- STRAUSS

Daher: Datensätze sind als identisch anzusehen (Treffer)

diese Vorgabe sagt nichts aus über:

Schmidt-Strauß / Schmidt

Jean-Pierre / Jeanpierre / Jean Pierre

- Diese Vorgaben sind konform zur gängigen Praxis der Transliteration
 - Absicherung durch zwei anonymisierte, nicht repräsentative Auswertungen von Standesamtsregistern
- sie sind mit vertretbarem Aufwand in bestehenden Verfahren umsetzbar
- sie sind kein Garant für eine 100%ige Trefferquote
- sie berücksichtigen keine abweichende Schreibweisen aus anderen Gründen: phonetische Suchen („Meyer – Meier“), Transkription



Die Tabelle

Vorgaben für Identifikationsverfahren - Abschlussbericht (GESCHÜTZT) - Adobe Reader

PROJEKTGRUPPE STANDARD DES AK I DER IMK

Vorgaben für Identifikationsverfahren - Abschlussbericht

Code-point (hex)	Character	Unicode Name	Category	Suchform	DOS CP437	DOS CP850	EBCDIC CP1141	EBCDIC CP500	ISO 8859-1	ISO 8859-15	LA8 Passport	WINDOWS CP1252
0009		CHARACTER TABULATION	OTHER	0009 ()	09 ()	09 ()	05 ()	05 ()	09 ()	09 ()	n/a	09 ()
000A		LINE FEED	OTHER	000A ()	0A ()	0A ()	25 ()	25 ()	0A ()	0A ()	000A ()	0A ()
000D		CARRIAGE RETURN	OTHER	000D ()	0D ()	0D ()	0D ()	0D ()	0D ()	0D ()	n/a	0D ()
0020		SPACE	SEPARATOR	0020 ()	20 ()	20 ()	40 ()	40 ()	20 ()	20 ()	0020 ()	20 ()
0021	!	EXCLAMATION MARK	PUNCTUATION	0021 (!)	21 (!)	21 (!)	4F (!)	4F (!)	21 (!)	21 (!)	0021 (!)	21 (!)
0022	"	QUOTATION MARK	PUNCTUATION	0022 (")	22 (")	22 (")	7F (")	7F (")	22 (")	22 (")	0022 (")	22 (")
0023	#	NUMBER SIGN	PUNCTUATION	0023 (#)	23 (#)	23 (#)	7B (#)	7B (#)	23 (#)	23 (#)	0023 (#)	23 (#)
0024	\$	DOLLAR SIGN	SYMBOL	0024 (\$)	24 (\$)	24 (\$)	5B (\$)	5B (\$)	24 (\$)	24 (\$)	0024 (\$)	24 (\$)
0025	%	PERCENT SIGN	PUNCTUATION	0025 (%)	25 (%)	25 (%)	6C (%)	6C (%)	25 (%)	25 (%)	0025 (%)	25 (%)
0026	&	AMPERSAND	PUNCTUATION	0026 (&)	26 (&)	26 (&)	50 (&)	50 (&)	26 (&)	26 (&)	0026 (&)	26 (&)
0027	'	APOSTROPHE	PUNCTUATION	0027 (')	27 (')	27 (')	7D (')	7D (')	27 (')	27 (')	0027 (')	27 (')
0028	(LEFT PARENTHESIS	PUNCTUATION	0028 (()	28 (()	28 (()	4D (()	4D (()	28 (()	28 (()	0028 (()	28 (()
0029)	RIGHT PARENTHESIS	PUNCTUATION	0029 ())	29 ())	29 ())	5D ())	5D ())	29 ())	29 ())	0029 ())	29 ())
002A	*	ASTERISK	PUNCTUATION	002A (*)	2A (*)	2A (*)	5C (*)	5C (*)	2A (*)	2A (*)	002A (*)	2A (*)
002B	+	PLUS SIGN	SYMBOL	002B (+)	2B (+)	2B (+)	4E (+)	4E (+)	2B (+)	2B (+)	002B (+)	2B (+)
002C	,	COMMA	PUNCTUATION	002C (,)	2C (,)	2C (,)	6B (,)	6B (,)	2C (,)	2C (,)	002C (,)	2C (,)
002D	-	HYPHEN-MINUS	PUNCTUATION	002D (-)	2D (-)	2D (-)	60 (-)	60 (-)	2D (-)	2D (-)	002D (-)	2D (-)
002E	.	FULL STOP	PUNCTUATION	002E (.)	2E (.)	2E (.)	4B (.)	4B (.)	2E (.)	2E (.)	002E (.)	2E (.)
002F	/	SOLIDUS	PUNCTUATION	002F (/)	2F (/)	2F (/)	61 (/)	61 (/)	2F (/)	2F (/)	002F (/)	2F (/)
0030	0	DIGIT ZERO	NUMBER	0030 (0)	30 (0)	30 (0)	F0 (0)	F0 (0)	30 (0)	30 (0)	0030 (0)	30 (0)
0031	1	DIGIT ONE	NUMBER	0031 (1)	31 (1)	31 (1)	F1 (1)	F1 (1)	31 (1)	31 (1)	0031 (1)	31 (1)
0032	2	DIGIT TWO	NUMBER	0032 (2)	32 (2)	32 (2)	F2 (2)	F2 (2)	32 (2)	32 (2)	0032 (2)	32 (2)
0033	3	DIGIT THREE	NUMBER	0033 (3)	33 (3)	33 (3)	F3 (3)	F3 (3)	33 (3)	33 (3)	0033 (3)	33 (3)

Seite 17 von 49

- Die Herausgabe des Standard durch die KoSIT kann nur eine Zwischenlösung sein
 - Auf Grund des Handlungsbedarfes in der Innenverwaltung
- Standardisierung auf EU-Ebene angestrebt

2. Targets dates for the overall action

1	Duration of the action	S+ 36 months
2	Interim progress report	S+18 months
3	Final report	S+39 months

- Ermittlung eines geeigneten Formats zum Austausch von Geburtsdaten
- Zunächst Technical Specification, später als Europäische Norm
 - Registrierung des 'Kern-Zeichenvorrats' als Untermenge von UCS bzw. ISO/IEC 10646 (UNICODE)



Wo finde ich weitere Informationen?

Unter www.xoev.de:

- Der Standard „Lateinische Zeichen in Unicode 1.1“
 - Zeichensatzdatei inklusive Umschlüsselungstabellen in XML
 - Empfehlung zur Identifikation und Umschlüsselung („ID-Algorithmus“)

Ansprechpartner:

Hannes Weber

Frank Steimke