UTF-8 Migration

24.10.2024

Grischa Pfister

Andreas Menrath

HMS analytical software

Agenda

- Was ist ein Encoding?
- K Funktionen
- SAS Migrationsprojekt aufsetzen
- Datasets konvertieren
- Code migrieren
- Ausblick

Was ist ein Encoding

Ist das Kunst oder kann das weg?

HMS
analytical software

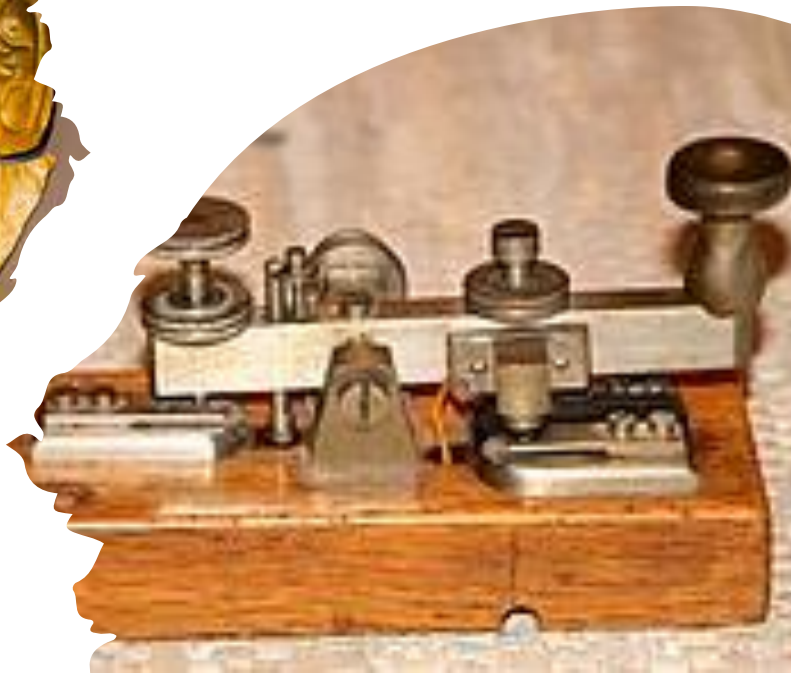
Kurze Geschichte des Encodings

Wikipedia

- Startet im Trojanischen Krieg / Agamemmnon

- Kommt über die britische Marine (Flaggenalphabet)

- Zu Morse-Code



Kurze Geschichte des Encodings

- Computer speichern binäre Zahlenfolgen
  → müssen in lesbare Zeichen umkodiert werden
- ASCII
  - 7Bit Darstellung (von 00 bis 7F) – 128 Zeichen
  - Der Rest: Sprachspezifische Umlaute, Sonderzeichen

- ASCII – EBCDIC – andere

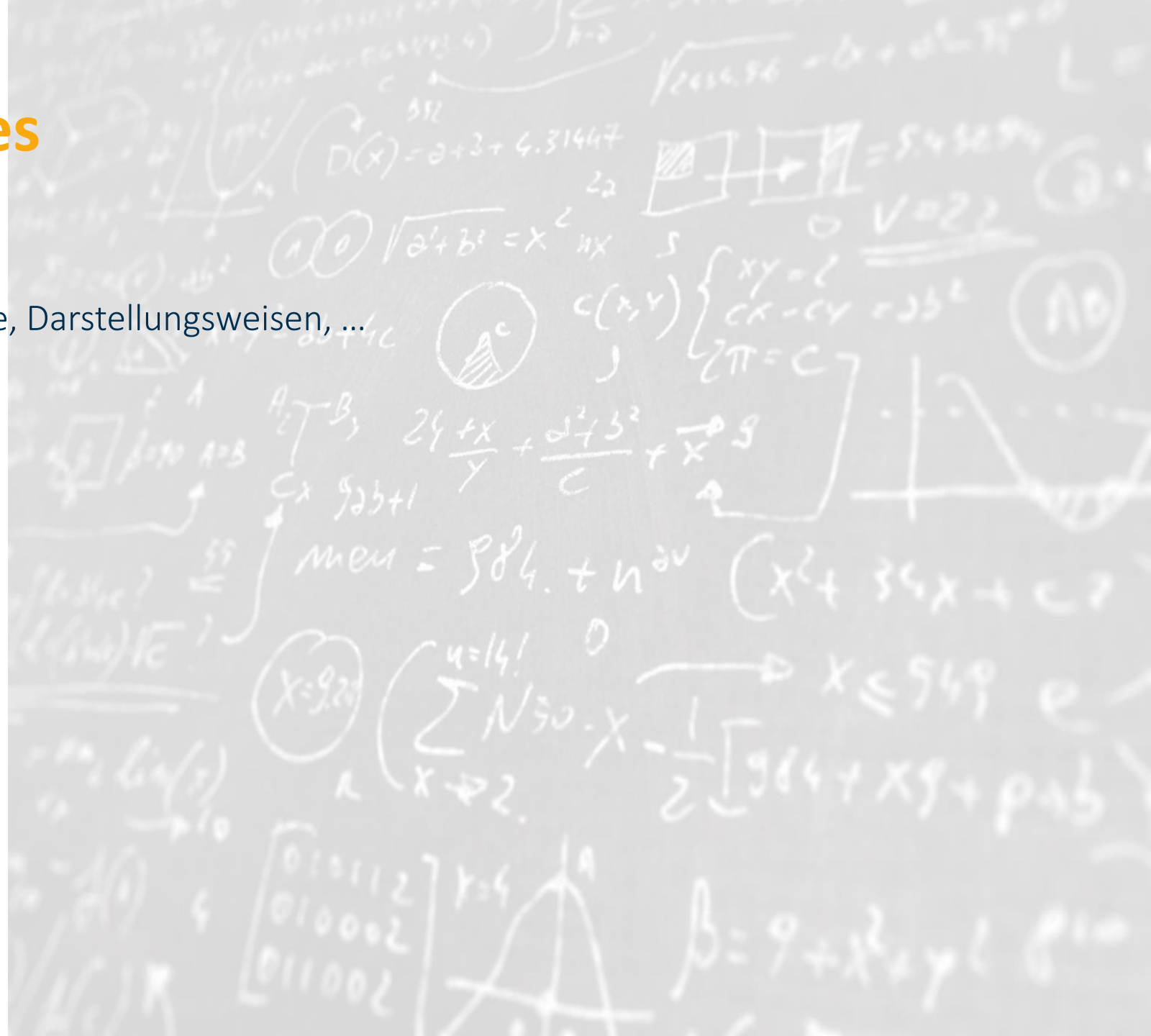
- DOS
  - Codepage 437: USA
  - Codepage 850: Europa

Buchstaben als 7-Bit-Code

ASCII	Dez	Hex	Binär
A	65	41	(0)100 0001
B	66	42	(0)100 0010
C	67	43	(0)100 0011
…	…	…	…
Z	90	5A	(0)101 1010

Kurze Geschichte des Encodings

- Überraschung
  - Es gibt mehr Sprachen, Umlaute, Darstellungsweisen, …
- Unicode (=ISO 10646)
- UTF-8
- UTF-16
- UTF-16[BE,LE]
- …



Kurze Geschichte des Encodings

…
„Zur Erleichterung für den verwirrten Leser sei noch angemerkt,
dass die weitaus meisten Texte in einer der drei Unicode-encodings UTF-8, UTF-16BE oder UTF-16LE gespeichert sind,
was den Umgang mit Texten wesentlich erleichtert.“

Quelle: https://de.wikipedia.org/wiki/Zeichenkodierung

SAS und Encodings

V6
- Proc Trantab

ASCII table 1:

```
     0 1 2 3 4 5 6 7 8 9 A B C D E F
00 '000102030405060708090A0B0C0D0E0F'x
10 '101112131415161718191A1B1C1D1E1F'x
20 '202122232425262728292A2B2C2D2E2F'x
30 '303132333435363738393A3B3C3D3E3F'x
40 '404142434445464748494A4B4C4D4E4F'x
50 '505152535455565758595A5B5C5D5E5F'x
60 '606162636465666768696A6B6C6D6E6F'x
70 '707172737475767778797A7B7C7D7E7F'x
80 '808182838485868788898A8B8C8D8E8F'x
90 '909192939495969798999A9B9C9D9E9F'x
A0 'A0A1A2A3A4A5A6A7A8A9AAABACADAEAF'x
B0 'B0B1B2B3B4B5B6B7B8B9BABBBCBDBEBF'x
C0 'C0C1C2C3C4C5C6C7C8C9CACBCCCDCECF'x
D0 'D0D1D2D3D4D5D6D7D8D9DADBDCDDDEDF'x
E0 'E0E1E2E3E4E5E6E7E8E9EAEBECEDEEEF'x
F0 'F0F1F2F3F4F5F6F7F8F9FAFBFCFDFEFF'x
```

ITALIAN table 1:

```
     0 1 2 3 4 5 6 7 8 9 A B C D E F
00 '000102030405060708090A0B0C0D0E0F'x
10 '101112131415161718191A1B1C1D1E1F'x
20 '253346645D65634450516167302D3837'x
30 '74787A7C7E7F8081828332316B6C6D35'x
40 '5A8596989CA0AAADAFB1BBBDBFC1C3C7'x
50 'D8DADCDEE3E7F1F3F5F7FD5262533C2A'x
60 '3B8495979B9FA9ACAEB0BABCBEC0C2C6'x
70 'D7D9DBDDE2E6F0F2F4F6FC546F554026'x
80 '5E2045AB4B3927283D66E049D621FF22'x
90 '2347484C4D422E2F3F29DF4AD524FEFB'x
A0 '73345C5F5B6070563E58864E6E2C592B'x
B0 '71687B7D3A7257414379C84F75767736'x
C0 '8A888C92908E949AA4A2A6A8B5B3B7B9'x
D0 '9EC5CCCACED2D06AD4EBE9EDEFF9E5E1'x
E0 '89878B918F8D9399A3A1A5A7B4B2B6B8'x
F0 '9DC4CBC9CDD1CF69D3EAE8ECEEF8E4FA'
```

V8
- National Language Support (NLS)
- Locale de_DE

Viya
- UTF-8

K Funktionen ?!

Was ist das schon wieder?

HMS
analytical software

SAS K Functions – warum?

Zur Laufzeit fragt sich SAS:

- single-byte character set (SBCS)?
- double-byte character set (DBCS)?
- multi-byte character set (MBCS)?

Viele SAS Funktionen nehmen implizit an, dass es sich um SBCS handelt

!!! und liefern falsche Ergebnisse wenn diese Annahme nicht erfüllt ist !!!

Lösung: k-Funktionen können mit allen Varianten umgehen

Have a Comprehensive understanding of SAS® K functions: https://support.sas.com/resources/papers/proceedings18/1902-2018.pdf
Internationalization Compatibility for SAS String Functions: https://documentation.sas.com/doc/en/pgmsascdc/v_049/nlsref/p1pca7vwjjwucin178l8qddjn0gi.htm

SAS K Functions Beispiel

```
data test;
     text = "ä";
     laenge_sbcs = length(text);
     laenge_mbcs = klength(text);
run;
```

	text	laenge_sbcs	laenge_mbcs
1	ä	2	1

SAS K Functions

KCOMPARE Function
KCOMPRESS Function
KCOUNT Function
KCOUNTC Function
KCOUNTW Function
KCOUNTX Function
KCVT Function
KFIND Function
KFINDC Function
KFINDW Function
KINDEX Function
KINDEXB Function
KINDEXCB Function
KINDEXC Function
KLEFT Function

KLENGTH Function
KLOWCASE Function
KPROPCASE Function
KPROPCHAR Function
KPROPDATA Function
KREVERSE Function
KRIGHT Function
KSCAN Function
KSCANX Function
KSTRCAT Function
KSTRIP Function
KSUBSTR Function
KSUBSTRB Function
KSUBSTRN Function
KTRANSLATE Function

KTRIM Function
KTRUNCATE Function
KUPCASE Function
KUPDATE Function
KUPDATEB Function
KUPDATES Function
KVERIFYB Function
KVERIFY Function

https://documentation.sas.com/doc/en/pgmsascdc/v_049/nlsref/p1pca7vwjjwucin178l8qddjn0gi.htm

SAS K Makros

Macro Functions by Category

The following table provides brief descriptions of the SAS NLS macro functions. For more information, see the NLS entry for each macro function.

Category	Language Elements	Description
DBCS	%KCMPRES Function	Compresses multiple blanks and removes leading and trailing blanks.
	%KINDEX Macro Function	Returns the position of the first character of a string.
	%KLEFT Macro Function	Left-aligns an argument by removing leading blanks.
	%KLENGTH Macro Function	Returns the length of a string.
	%KSCAN Macro Function	Search for a word that is specified by its position in a string.
	%KSUBSTR Macro Function	Produce a substring of a character string.
	%KUPCASE Macro Function	Convert values to uppercase.
	%QKCMPRES Function	Compresses multiple blanks and removes leading and trailing blanks.
	%QKLEFT Macro Function	Left-aligns an argument by removing leading blanks.
	%QKSCAN Macro Function	Search for a word that is specified by its position in a string.
	%QKSUBSTR Macro Function	Produce a substring of a character string.
	%QKUPCASE Macro Function	Convert values to uppercase.

https://documentation.sas.com/doc/en/pgmsascdc/v_049/nlsref/n02vld3869thbgn1r9ozgh77o9jm.htm

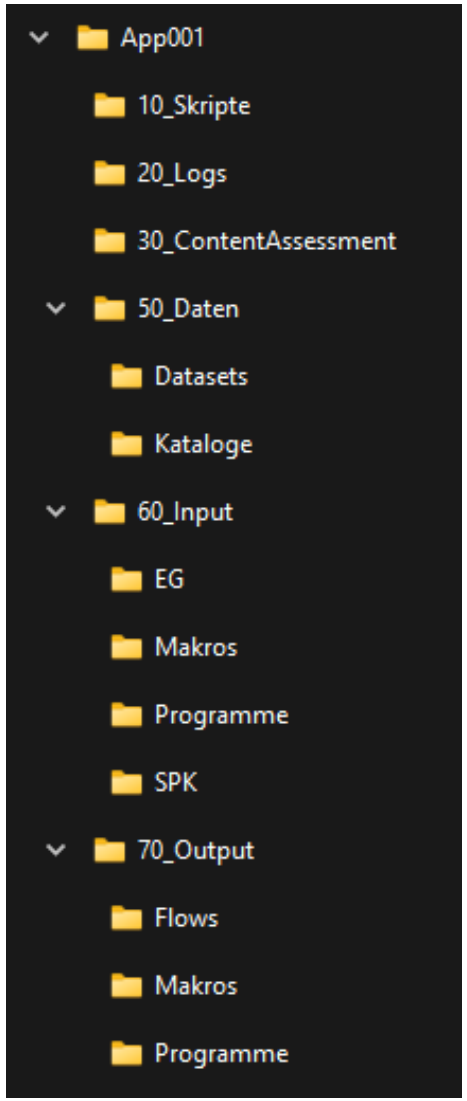
SAS Migrationsprojekt aufsetzen

HMS
analytical software

SAS Migration - Projektablauf je Applikation

- Lokale Verzeichnisse & Git einrichten
- SAS Content Assessment?
- Daten + Code migrieren
- Code deployen nach Viya
- Tests durchführen

Beispiel für Projektordnerstruktur



Praxis-Tipps:

- möglichst viel automatisieren
- Versionsverwaltung nutzen
- .gitignore verwenden
- auch Artefakte vor der Migration versionieren.
  →Änderungen nachvollziehen können
- Git Prozesse (Pull Requests, Pipelines) nutzen:
  - Code Review
  - Freigabeprozesse
  - Tests

.gitignore:

```
20_Logs/
30_ContentAssessment/
50_Daten/
```

Datasets konvertieren

HMS
analytical software

SAS Datasets migrieren

Daten nach Viya kopieren
(komprimieren, übertragen, entpacken, prüfen, dokumentieren)

Dataset-Konvertierung nach UTF-8

Indizes neu erstellen

SAS Datasets migrieren

The CONTENTS Procedure

Data Set Name	SASHELP.CLASS	Observations	19
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	06.08.2020 03:16:10	Observation Length	40
Last Modified	06.08.2020 03:16:10	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Student Data		
Data Representation	WINDOWS_64		
Encoding	us-ascii ASCII (ANSI)		

The CONTENTS Procedure

Data Set Name	COPYOUT.CLASS	Observations	19
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	08/05/2020 21:16:10	Observation Length	40
Last Modified	08/05/2020 21:16:10	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Student Data		
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64, LINUX_POWER_64		
Encoding	utf-8 Unicode (UTF-8)		

```
82   proc print data=test.class; run;
NOTE: Data file TEST.CLASS.DATA is in a format that is native to another host, or the file encoding does not match the session
      encoding. Cross Environment Data Access will be used, which might require additional CPU resources and might reduce
      performance.
```

```
libname copyin cvp "/data/app001/";
libname copyout "/data/app001/utf8";

proc migrate in=copyin out=copyout;
run;
```

```
1   libname copyin cvp "/data/app001/";
2   libname copyout "/data/app001/utf8";
3
4   proc datasets library=copyin;
5       copy out=copyout noclone;
6   run;
7
```

Migrating Data to UTF-8 for the SAS® Viya ® Platform:
https://documentation.sas.com/api/collections/pgmsascdc/v_023/docsets/viyadatamig/content/viyadatamig.pdf?locale=enBeschreibt

Code migrieren

HMS analytical software

SAS Code migrieren

Code extrahieren (EG, STP, DI Job, …)

Clean up: UTF8, Zeilenumbrüche (Windows zu Linux/Unix)
(optional: Tabs to Spaces, Trailing Blanks, …)

Code Ersetzungen

Datenquellen & Pfade/Verzeichnisse ersetzen

Manuelle Prüfung und ggfs. Nacharbeiten

K-Funktionen: Suchen und Ersetzen per Skript

Was kann schon schief gehen?

- auch Code in Kommentaren wird ersetzt
- Programme mehrfach überarbeiten: k-Funktionen bekommen noch ein zusätzliches k, z.B. kcount() wird zu kkcount()
- Nur nach Schlüsselwörtern ersetzen ist gefährlich
  → „length“ könnte z.B. auch Teil von Variablenname sein
  → „left“ kommt z.B. auch als Schlüsselwort in SQL LEFT JOIN vor
  → bessere Regel: nach Funktionsname muss eine öffnende Klammer kommen (irgendwann)
- Ausnahmen z.B. für PROC SQL: SELECT COUNT(*)

➔Ersetzungsregeln mit komplexen regulären Ausdrücken automatisieren auch den Großteil der Problemfälle

Manuelle Problemfälle und Nacharbeiten

- SUBSTR(left of =) ersetzen:

```
data x;
  set sashelp.class;
  substr(name,1,1) = 'x';
run;
```

→SAS verweist auf KUPDATE() und KUPDATEB()

- strip() maskiert Makrovariablen, kstrip nicht.
  →strip(&myvar.) wird zu kstrip(%bquote(&myvar.))

Ausblick

HMS
analytical software

Da wäre noch mehr (Out of Scope)

Was es nicht in die Präsentation geschafft hat:

- SAS/AF
- Compiled Macros
- Formate und Kataloge
- EG to Flow
- Daten in SAS OLAP- oder SPDS- Server
- Länge von Character-Feldern in Datenbanken
- Teststrategien

Unser Fazit

andreas
endlich Emojis 😉 😍 🔥 🚀

grischa
Schei Encoding

```
data fazit;
   andreas = "endlich Emojis 😉😍🔥🚀";
   grischa = "Schei Encoding";
run;
```

Fragen?

Bing Image Creator:
image with questionmarks in style of dali

HMS

analytical software

Bildnachweise

Flaggenalphabet
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.boot.de%2Fde%2FMedia_News%2Fboot.radar%2FThema_Ausr%25C3%25BCstung%2FRatgeber%2FDas_internationale_Flaggenalphabet&psig=AOvVaw3wPJ7ccaRVVIDTOwhRBdUD&ust=1728740444917000&source=images&cd=vfe&opi=89978449&ved=0CBEQjRxqFwoTCMDv55i6hokDFQAAAAAdAAAAABAE

Morsecode
- https://de.wikipedia.org/wiki/Datei:Morsetaste.jpg

Maske Agamemmnon

Buchstaben als 7-BIT-Code
- https://de.wikipedia.org/wiki/American_Standard_Code_for_Information_Interchange

Schei? Encoding
- https://www.amazon.de/Schei-Encoding-Computerfreaks-gehandelter-Bio-Baumwolle/dp/B07C1PLNJQ